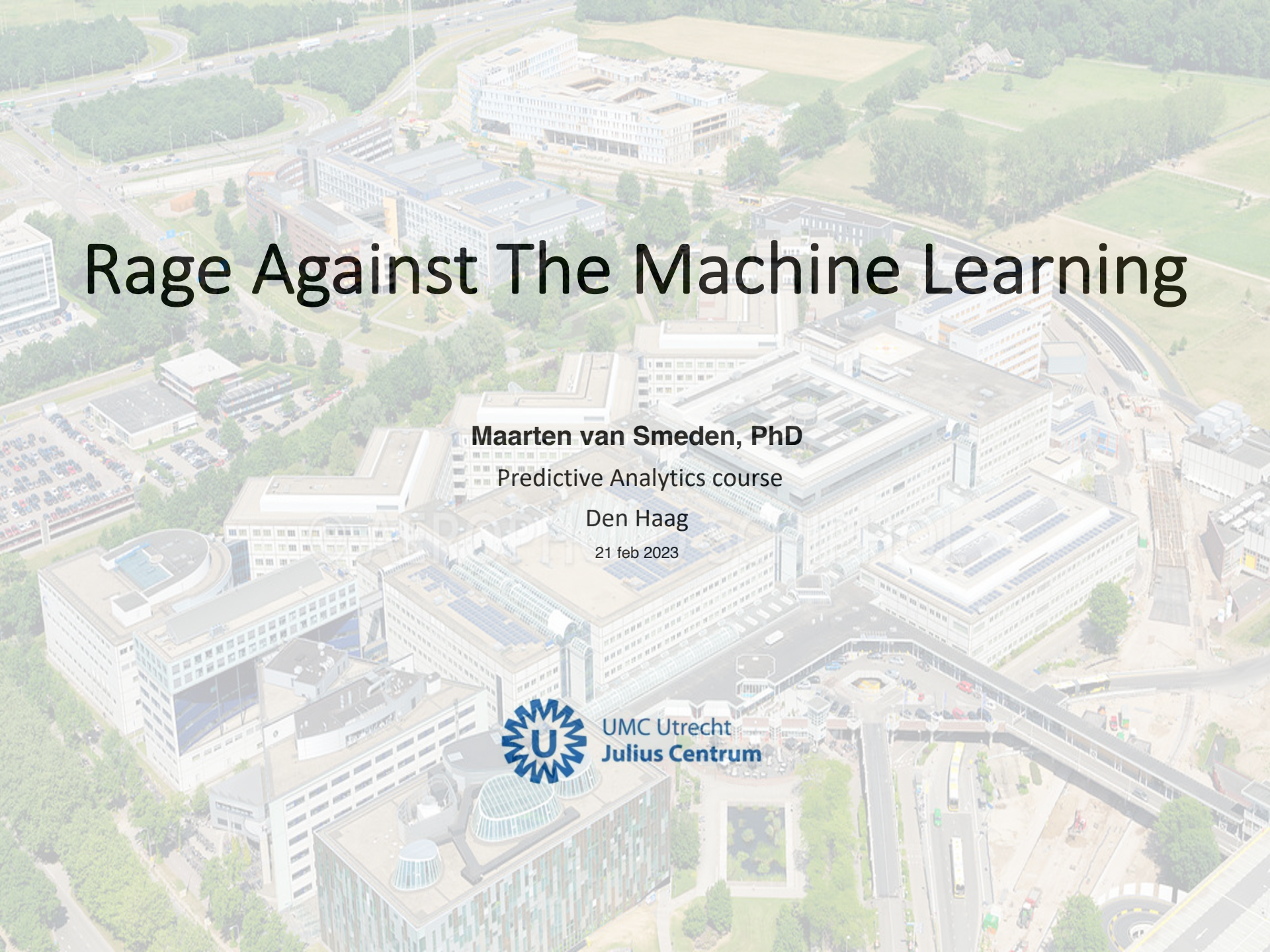


Rage Against The Machine Learning

Maarten van Smeden, PhD

Predictive Analytics course

Den Haag

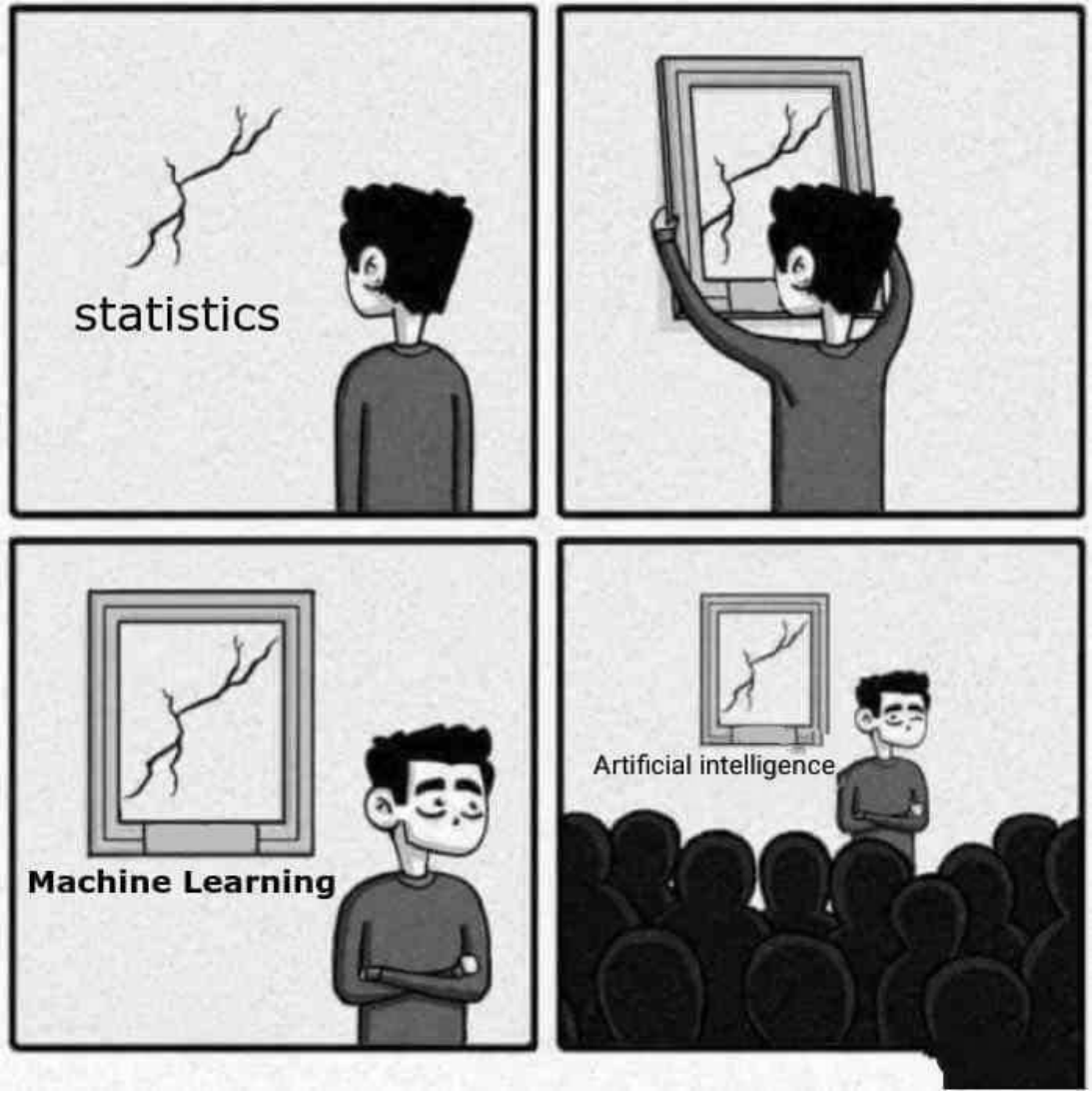
21 feb 2023

UMC Utrecht
Julius Centrum

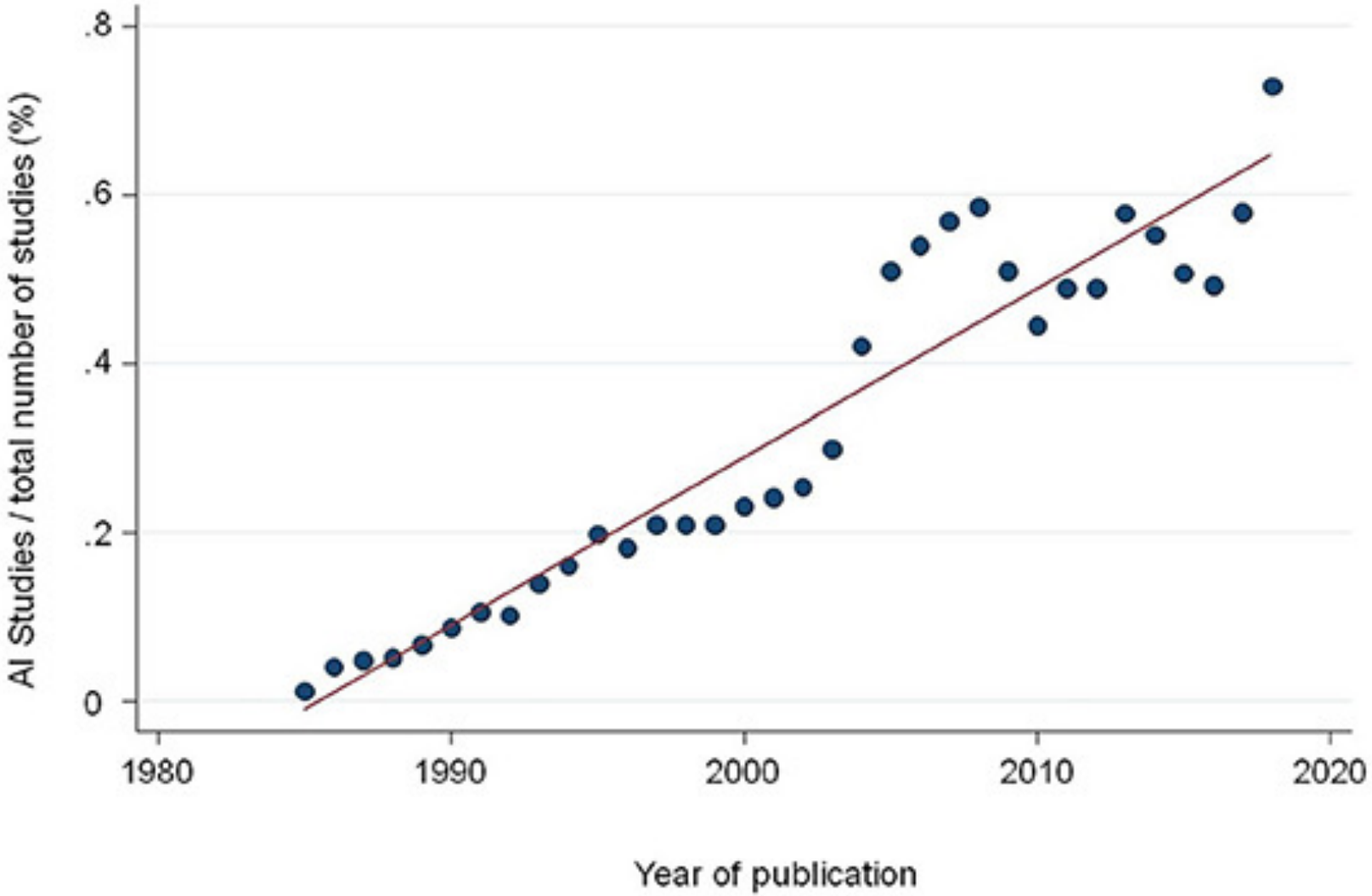
Terminology

In medical research, “artificial intelligence” usually just means “machine learning” or “algorithm”

UMC Utrecht



UMC Utrecht



Proportion of studies indexed in Medline with the Medical Subject Heading (MeSH) term “Artificial Intelligence” divided by the total number of publications per year.

UMC Utrecht

Reviewer #2

used in this paper. Second, since the prediction performance of logistic regression models is often inferior to those of powerful machine learning algorithms such as random forest or boosting, focussing logistic regression models only can be boring. The detailed comments are given below.

Cross Validated

Home

Questions

Tags

Users

Unanswered

What do statisticians do that can't be automated?

Asked 8 years ago Active 1 year, 11 months ago Viewed 6k times

26

Will software eventually make statisticians obsolete? What is done that can't be programmed into a computer?

machine-learning dataset careers

8

share cite improve this question

edited Feb 10 '12 at 19:54

community wiki
3 revs, 2 users 86%
Adam

UMC Utrecht

JNM The Journal of NUCLEAR MEDICINE

HOT TOPICS

Machine Learning Will Change Medicine

Michael Forsting

Essen University Hospital, University of Essen-Duisburg, Essen, Germany

UMC Utrecht

Doctors about to be replaced by hospital AI systems offering better diagnosis and less arrogance

09/12/2017 / By Jhoanna Robinson

Framingham, Massachusetts-based market intelligence provider IDC Health Insights, in its recently published report on artificial intelligence and cognitive computing adoption in the Asia/Pacific titled *IDC Peerscape: Cognitive/AI Practices for Healthcare in Asia/Pacific (Excluding Japan)*, stated the best possible healthcare solutions that hospitals and health insurance companies all around the Asia-Pacific countries should adopt.

UMC Utrecht

Tech company business model

UMC Utrecht

Tech company business model

Opinion

Big Data and Machine Learning in Health Care

Nearly all aspects of modern life are in some way being changed by big data and machine learning. Netflix knows what movies people like to watch and Google knows what people want to know based on their search histories. Indeed, Google has recently begun to replace much of its existing non–machine learning technology with machine learning algorithms, and there is great optimism that these techniques can provide similar improvements across many sectors.

spectrum (#19 in the
evidence-based clinical
model of this sort, and s
ist on the machine le
Figure). On the extrem
spectrum would be he
not directly involve the
itly derived from data (

Suppose a new ca

From Netflix to Heart Attacks: Collaborative Filtering in Medical Datasets

ABSTRACT

Recommender systems are widely used to provide users with personalized suggestions for products or services. These systems typically rely on collaborative filtering (CF) to make automated predictions about the interests of a user, by col-
ny users. CF tech-
nd can be used on

are used by a number of different commercial organizations, including Amazon [14], Google [9], Netflix [7], TiVo [2] and Yahoo! [18].

Most of these recommender systems are based on collaborative filtering (CF), which uses past user behavior and preferences (such as product ratings) to make automated predictions about the interests of a user. CF techniques

Multicenter Comparison of Machine Learning Methods and Conventional Regression for Predicting Clinical Deterioration on the Wards

An estimated 2.5 quintillion bytes of data are generated every day, and the volume, velocity, and variety of information has led to the popularization of the term "big data (1)." Companies like Google, Amazon, and Netflix have leveraged big data in concert with complex algorithms to improve predictions of human behavior and events (2). These algorithms, known as machine learning in computer science, are flexible techniques designed to learn and generalize from

Predictor Variables

Age, time since ward admission, number of previous ICU stays,
vital signs, and
lytes, creatinin
utilized as prec
the electronic l
University of (
the NorthShor

Big Data And New Knowledge In Medicine: The Thinking, Training, And Tools Needed For A Learning Health System

Netflix, the popular entertainment company, is known for making useful movie suggestions to its customers. In 2006 the company embarked on a project to further improve its ability to predict which movies its customers would like.[1] Through an open competition, Netflix offered a $1 million prize to the group that most improved on Netflix's traditional approach, which was based on conventional

18,000 movie titles by almost 500,000 people. The winning teams not only focused on how each person rated movies but also, importantly, discovered that an individual's ratings were influenced by factors such as whether the person ranks many movies at a time (which tended to accentuate positive or negative preferences) or by the overall popularity of a movie across raters at a particular point in time. Ultimately, the winners produced an algorithm that increased the

Other success stories

ARTICLE

doi:10.1038/nature16961

Mastering the game of Go with deep neural networks and tree search

David Silver[1]*, Aja Huang[1]*, Chris J. Maddison[1], Arthur Guez[1], Laurent Sifre[1], George van den Driessche[1], Julian Schrittwieser[1], Ioannis Antonoglou[1], Veda Panneershelvam[1], Marc Lanctot[1], Sander Dieleman[1], Dominik Grewe[1], John Nham[2], Nal Kalchbrenner[1], Ilya Suts[2], Timothy Lillicrap[1], Madeleine Leach[1], Koray Kavukcuoglu[1], Thore Graepel[1] & Demis Hassabis[1]

The game of Go has long been viewe
enormous search space and the diffic
to computer Go that uses 'value netw
neural networks are trained by a novel
learning from games of self-play. W
of-the-art Monte Carlo tree search pr
new search algorithm that combines M
our program AlphaGo achieved a 99.8
champion by 5 games to 0. This is the
full-sized game of Go, a feat previous

US & WORLD | TECH | ARTIFICIAL INTELLIGENCE

Former Go champion beaten by DeepMind retires after declaring AI invincible

'Even if I become the number one, there is an entity that cannot be defeated'

By James Vincent | Nov 27, 2019, 8:42am EST

IBM Watson winning Jeopardy! (2011)

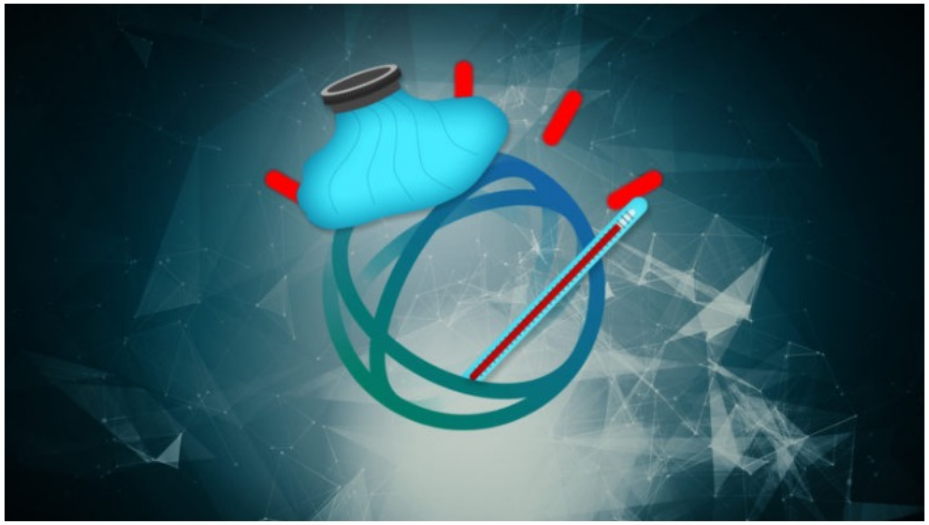
IBM Watson for oncology

EXCLUSIVE

STAT+

IBM's Watson supercomputer recommended 'unsafe and incorrect' cancer treatments, internal documents show

By CASEY ROSS @caseymross *and* IKE SWETLITZ / JULY 25, 2018



ALEX HOGAN/STAT

Internal IBM documents show that its Watson supercomputer often spit out erroneous cancer treatment advice and that company medical specialists and customers identified "multiple examples of unsafe and incorrect treatment recommendations" as IBM was promoting the product to hospitals and physicians around the world.

UMC Utrecht

Machine learning everywhere

Articles

Development and validation of the automated imaging differentiation in parkinsonism (AID-P): a multicentre machine learning study

Derek B Archer, Justin T Bricker, Winston T Chu, Roxana G Burciu, Johanna L McCracken, Song Lai, Stephen A Coombes, Ruogu Fang, Angelos Barmpoutis, Daniel M Corcos, Ajay S Kurani, Trina Mitchell, Mieniecia L Black, Ellen Herschel, Tanya Simuni, Todd B Parrish, Cynthia Comella, Tao Xie, Klaus Seppi, Nicolaas I Bohnen, Martijn LTM Müller, Roger L Albin, Florian Krismer, Guangwei Du, Mechelle M Lewis, Xuemei Huang, Hong Li, Ofer Pasternak, Nikolaus R McFarland, Michael S Okun, David E Vaillancourt

Summary

Background Development of valid, non-invasive biomarkers for park
aimed to assess whether non-invasive diffusion-weighted MRI can

JAMA Network Open

JACC: CARDIOVASCULAR INTERVENTIONS
© 2019 PUBLISHED BY ELSEVIER ON BEHALF OF THE AMERICAN COLLEGE OF CARDIOLOGY FOUNDATION
VOL. 12, NO. 14, 2019

Leveraging Machine Learning Techniques to Forecast Patient Prognosis After Percutaneous Coronary Intervention

Chad J. Zack, MD, MS,[a,*] Conor Senecal, MD,[b,*] Yaron Kinar, PhD,[c] Yaakov Metzger, MD, PhD,[c] Yoav Bar-Sinai, MS,[c] R. Jay Widmer, MD, PhD,[b] Ryan Lennon, MS,[d] Mandeep Singh, MD, MPH,[b] Malcolm R. Bell, MD,[b] Amir Lerman, MD,[b] Rajiv Gulati, MD, PhD[b]

Original Investigation | Substance Use and Addiction

Identifying Smoking Environments From Images of Daily Life With Deep Learning

Matthew M. Engelhard, MD, PhD; Jason A. Oliver, PhD; Ricardo Henao, PhD; Matt Hallyburton, BA; Lawrence E. Carin, PhD; Cynthia Conklin, PhD; F. Joseph McClernon, PhD

nature medicine
PERSPECTIVE
https://doi.org/10.1038/s41591-019-0548-6

Do no harm: a roadmap for responsible machine learning for health care

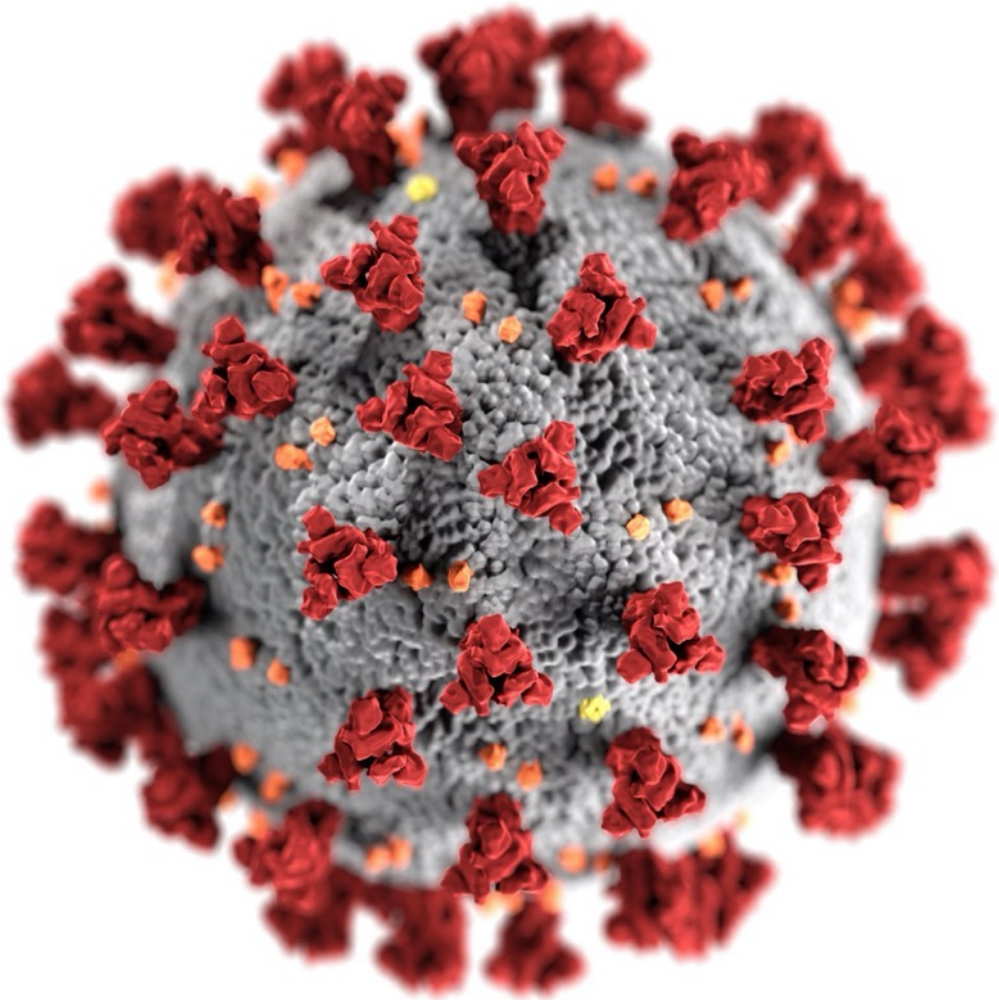
Jenna Wiens[1,19*], Suchi Saria[2,3,4,19], Mark Sendak[5], Marzyeh Ghassemi[6,7,8], Vincent X. Liu[9], Finale Doshi-Velez[10], Kenneth Jung[11], Katherine Heller[12,13], David Kale[14], Mohammed Saeed[15], Pilar N. Ossorio[16], Sonoo Thadaney-Israni[17] and Anna Goldenberg[6,8,18,19*]

JCO® Clinical Cancer Informatics
An American Society of Clinical Oncology Journal

ORIGINAL REPORT

Open Source Infrastructure for Health Care Data Integration and Machine Learning Analyses

Veli-Matti Isoviita, MD[1]; Liina Salminen, MD[2,3]; Jimmy Azar, PhD[1]; Rainer Lehtonen, PhD[1]; Pia Roering, MSc[3]; Olli Carpén, MD, PhD[1,3]; ...



UMC Utrecht

"As of today, we have deployed the system in 16 hospitals, and it is performing over 1,300 screenings per day"
MedRxiv pre-print only, 23 March 2020,
doi.org/10.1101/2020.03.19.20039354

European Commission › Strategy › Shaping Europe's digital future › News ›

Shaping Europe's digital future

NEWS ARTICLE | 19 May 2020

Using AI to fast and effectively diagnose COVID-19 in hospitals

The European Commission will invest in the use of Artificial Intelligence to speed up the diagnosis of COVID-19 and improve the future treatment of patients. A software developed to assist the work of medical staff by analysing images of pulmonary infections is introduced in 10 hospitals across Europe.

About Artificial intelligence

- Policies
- Blog posts
- News

Infervision's AI is in Italy Helping to Battle COVID-19

PRESS RELEASE UPDATED: MAR 23, 2020

ROME, March 20, 2020 (Newswire.com) - COVID-19 is spreading, with European countries already declaring a pandemic. The World Health Organization has declared Europe as the new 'epicenter' for COVID-19. Italy announced a full lock-down on March 10. Due to the spreading of COVID-19, Italian medical institutions are facing tremendous pressure as patient numbers surge. Meanwhile, issues over long turnaround times for PCT testing and limited availabilities of the kit are concerning. Using CT images will help with the screening of COVID-19.

New tool could 'help UK doctors spot high-risk Covid patients in seconds'

Study claims risk calculator will help clinicians with expected influx of patients this autumn

- Coronavirus - latest updates
- See all our coronavirus coverage

▲ The calculator was tested in a hospitalised elderly population, so is not applicable for use within the community. Photograph: Murdo MacLeod/The Guardian

score of three or less. The tool, which is easily accessible on a smartphone or computer, takes seconds to generate a score and is expected to be rolled out in the NHS this week.

UMC Utrecht

Covid: Extra 1.7m vulnerable added to shielding list

By Nick Triggle
Health correspondent

1 day ago | Comments

Coronavirus pandemic

There is to be a large expansion of the number of people being asked to shield in England.

An extra 1.7 million people are expected to be added to the 2.3 million already on the list.

Half of the group have not yet been vaccinated so will now be prioritised urgently by their local GPs.

It comes after a **new model** was developed that takes into account extra factors rather than just health.

This calculation includes things such as ethnicity, deprivation (by postcode) and weight to work out a person's risk of becoming seriously ill if they were to catch Covid.

It also looks at age, underlying health issues and prescribed medications.

Prof Andrew Hayward, a member of the New and Emerging Respiratory Virus Threats Advisory Group (Nervtag), which has been involved in the modelling, said it considered a "combination of factors" such as age, ethnicity and chronic illness and put them together to reach a score.

FDA APPROVED

CDS algorithm that predicts COVID-19 complications receives Emergency Use Authorization from FDA

Dascena's COViage system uses demographic and vital-sign data pulled from a COVID-19 patients' EHR to calculate their risk of hemodynamic instability or respiratory failure.

By Dave Muoio | October 06, 2020 | 03:05 pm

SHARE 60

FDA APPROVED

Using artificial intelligence to improve the outcome of critically ill patients

CLEW's ICU solution – CLEWICU, delivers customizable real-time clinical optimization, actionable predictive clinical analytics and patient risk stratification. The platform utilizes the full range of available patient data to provide continuous predictions based on sophisticated machine learning algorithms and models. The solution enables early identification and intervention and patient context prioritization. The CLEWICU system interfaces with existing EMR systems and medical devices and can be deployed either on-premises or in the cloud.

Maximizing scarce ICU resources

Healthcare reforms are affecting the finances of hospitals, the workloads of healthcare providers, and the treatment of patients. These reforms are driving an increased focus on improving clinical value by identifying high risk patients within the ICU, prioritizing treatment based on patient acuity, and reducing total length of stay and iatrogenesis.

UMC Utrecht

Living review (update 4)

RESEARCH

FAST TRACK

Prediction models for diagnosis and prognosis of covid-19: systematic review and critical appraisal

Laure Wynants,[1,2] Ben Van Calster,[2,3] Gary S Collins,[4,5] Richard D Riley,[6] Georg Heinze,[7] Ewoud Schuit,[8,9] Elena Albu,[2] Banafsheh Arshi,[1] Vanesa Bellou,[10] Marc M J Bonten,[8,11] Darren L Dahly,[12,13] Johanna A Damen,[8,9] Thomas P A Debray,[8,14] Valentijn M T de Jong,[8,9] Maarten De Vos,[2,15] Paula Dhiman,[4,5] Joie Ensor,[6] Shan Gao,[2] Maria C Haller,[7,16] Michael O Harhay,[17,18] Liesbet Henckaerts,[19,20] Pauline Heus,[8,9] Jeroen Hoogland,[8] Mohammed Hudda,[21] Kevin Jenniskens,[8,9] Michael Kammer,[7,22] Nina Kreuzberger,[23] Anna Lohmann,[24] Brooke Levis,[6] Kim Luijken,[24] Jie Ma,[5] Glen P Martin,[25] David J McLernon,[26] Constanza L Andaur Navarro,[8,9] Johannes B Reitsma,[8,9] Jamie C Sergeant,[27,28] Chunhu Shi,[29] Nicole Skoetz,[22] Luc J M Smits,[1] Kym I E Snell,[6] Matthew Sperrin,[30] René Spijker,[8,9,31] Ewout W Steyerberg,[3] Toshihiko Takada,[8,32] Ioanna Tzoulaki,[10,33] Sander M J van Kuijk,[34] Bas C T van Bussel,[1,35] Iwan C C van der Horst,[35] Kelly Reeve,[36] Florien S van Royen,[8] Jan Y Verbakel,[37,38] Christine Wallisch,[7,39,40] Jack Wilkinson,[24] Robert Wolff,[41] Lotty Hooft,[8,9] Karel G M Moons,[8,9] Maarten van Smeden[8]

UMC Utrecht

doi: 10.1136/bmj.m1328

Living review (update 4)

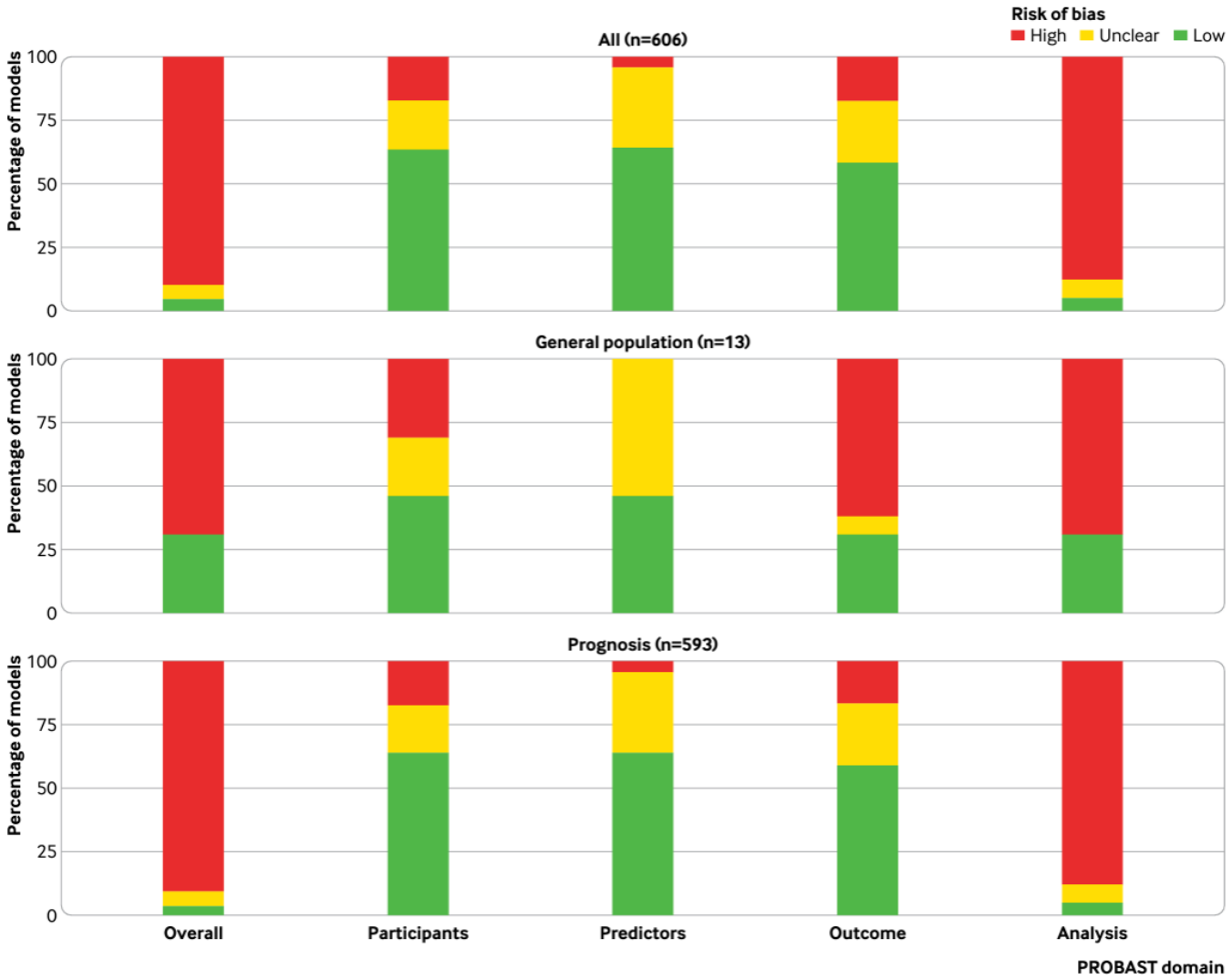


Fig 2 | PROBAST (prediction model risk of bias assessment tool) risk of bias for all included models combined (n=606) and broken down per type of analysis

doi: 10.1136/bmj.m1328

Systematic evaluation and external validation of 22 prognostic models among hospitalised adults with COVID-19: an observational cohort study

Rishi K. Gupta [1,2], Michael Marks [2,3], Thomas H.A. Samuels[2], Akish Luintel[2], Tommy Rampling[2], Humayra Chowdhury[2], Matteo Quartagno[4], Arjun Nair[2], Marc Lipman [5], Ibrahim Abubakar [1], Maarten van Smeden [6], Wai Keong Wong[2], Bryan Williams[7,8] and Mahdad Noursadeghi [2,9], on behalf of The UCLH COVID-19 Reporting Group[10]

@ERSpublications
Oxygen saturation on room air and patient age are strong predictors of deterioration and mortality, respectively, among hospitalised adults with COVID-19. None of the 22 prognostic models evaluated in this study adds incremental value to these univariable predictors. https://bit.ly/2Hg24TO

UMC Utrecht

Clinical prediction models for mortality in patients with covid-19: external validation and individual participant data meta-analysis

Valentijn M T de Jong,[1,2,3] Rebecca Z Rousset,[1] Neftalí Eduardo Antonio-Villa,[4,5] Arnoldus G Buenen,[6,7] Ben Van Calster,[8,9,10] Omar Yaxmehen Bello-Chavolla,[4] Nigel J Brunskill,[11,12] Vasa Curcin,[13] Johanna A A Damen,[1,2] Carlos A Fermín-Martínez,[4,5] Luisa Fernández-Chirino,[4,14] Davide Ferrari,[13,15] Robert C Free,[16,17] Rishi K Gupta,[18] Pranabashis Haldar,[16,17,19] Pontus Hedberg,[20,21] Steven Kwasi Korang,[22] Steef Kurstjens,[23] Ron Kusters,[23,24] Rupert W Major,[11,25] Lauren Maxwell,[26] Rajeshwari Nair,[27,28] Pontus Naucler,[20,21] Tri-Long Nguyen,[1,29,30] Mahdad Noursadeghi,[31] Rossana Rosa,[32] Felipe Soares,[33] Toshihiko Takada,[1,34] Florien S van Royen,[1] Maarten van Smeden,[1] Laure Wynants,[7,35] Martin Modrák,[36] on behalf of the CovidRetro collaboration, Folkert W Asselbergs,[37,38,39] Marijke Linschoten,[37] on behalf of CAPACITY-COVID consortium, Karel G M Moons,[1,2] Thomas P A Debray[1,2]

UMC Utrecht

doi: 10.1136/bmj-2021-069881

Clinical prediction models for mortality in patients with covid-19:

Results Datasets included 27 clusters from 18 different countries and contained data on 46 914patients. The pooled estimates ranged from 0.67 to 0.80 (C statistic), 0.22 to 1.22 (calibration slope), and 0.18 to 2.59 (O:E ratio) and were prone to substantial between study heterogeneity. The 4C Mortality Score by Knight et al (pooled C statistic 0.80, 95% confidence interval 0.75 to 0.84, 95% prediction interval 0.72 to 0.86) and clinical model by Wang et al (0.77, 0.73 to 0.80, 0.63 to 0.87) had the highest discriminative ability. On average, 29% fewer deaths were observed than predicted by the 4C Mortality Score (pooled O:E 0.71, 95% confidence interval 0.45 to 1.11, 95% prediction interval 0.21 to 2.39), 35% fewer than predicted by the Wang clinical model (0.65, 0.52 to 0.82, 0.23 to 1.89), and 4% fewer than predicted by Xie et al's model (0.96, 0.59 to 1.55, 0.21 to 4.28).

Conclusion The prognostic value of the included models varied greatly between the data sources. Although the Knight 4C Mortality Score and Wang clinical model appeared most promising, recalibration (intercept and slope updates) is needed before implementation in routine care.

UMC Utrecht

doi: 10.1136/bmj-2021-069881

nature machine intelligence

ANALYSIS

https://doi.org/10.1038/s42256-021-00307-0

Check for updates

OPEN

Common pitfalls and recommendations for using machine learning to detect and prognosticate for COVID-19 using chest radiographs and CT scans

Michael Roberts[1,2]✉, Derek Driggs[1], Matthew Thorpe[3], Julian Gilbey[1], Michael Yeung[4], Stephan Ursprung[4,5], Angelica I. Aviles-Rivero[1], Christian Etmann[1], Cathal McCague[4,5], Lucian Beer[4], Jonathan R. Weir-McCall[4,6], Zhongzhao Teng[4], Effrossyni Gkrania-Klotsas[7], AIX-COVNET*, James H. F. Rudd[8,36], Evis Sala[4,5,36] and Carola-Bibiane Schönlieb[1,36]

Machine learning methods offer great promise for fast and accurate detection and prognostication of coronavirus disease 2019 (COVID-19) from standard-of-care chest radiographs (CXR) and chest computed tomography (CT) images. Many articles have been published in 2020 describing new machine learning-based models for both of these tasks, but it is unclear which are of potential clinical utility. In this systematic review, we consider all published papers and preprints, for the period from 1 January 2020 to 3 October 2020, which describe new machine learning models for the diagnosis or prognosis of COVID-19 from CXR or CT images. All manuscripts uploaded to bioRxiv, medRxiv and arXiv along with all entries in EMBASE and MEDLINE in this timeframe are considered. Our search identified 2,212 studies, of which 415 were included after initial screening and, after quality screening, 62 studies were included in this systematic review. Our review finds that none of the models identified are of potential clinical use due to methodological flaws and/or underlying biases. This is a major weakness, given the urgency with which validated COVID-19 models are needed. To address this, we give many recommendations which, if followed, will solve these issues and lead to higher-quality model development and well-documented manuscripts.

UMC Utrecht

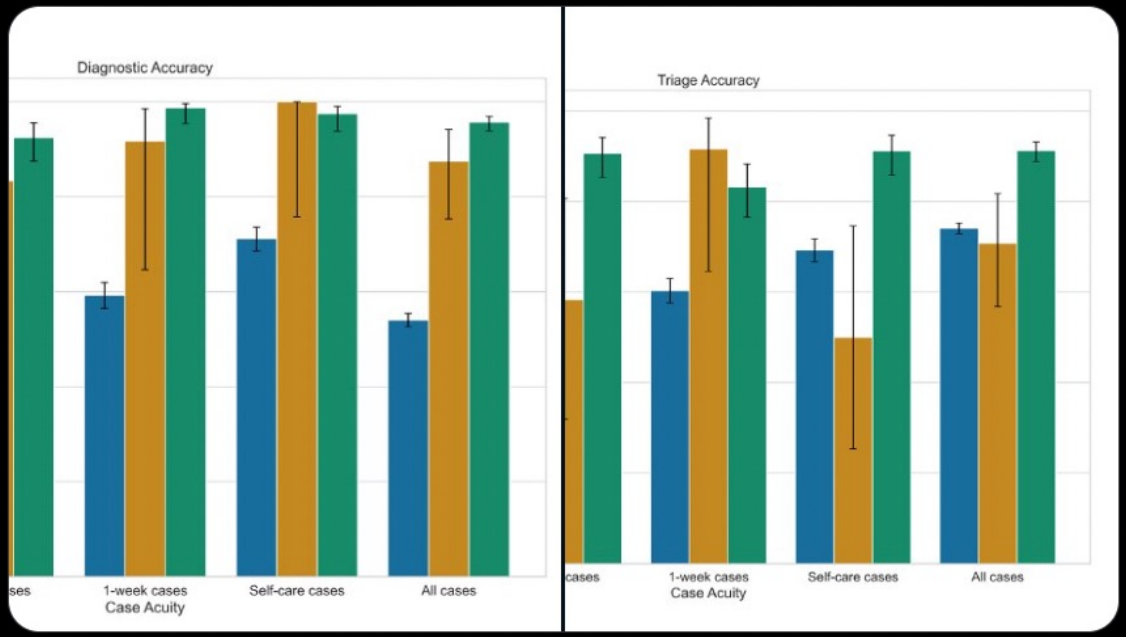
ChatGPT

Andrew Beam
@AndrewLBeam

New preprint!

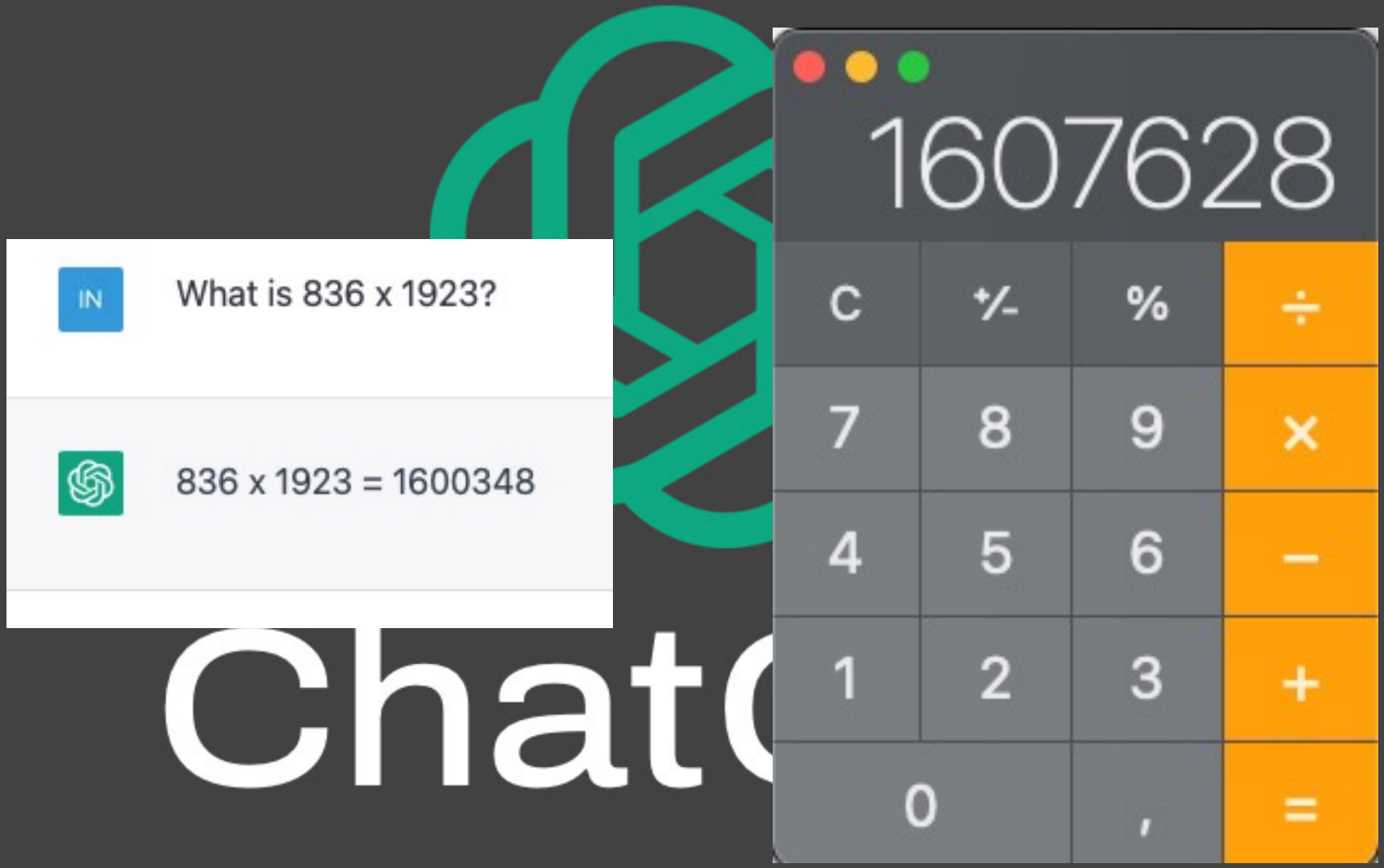
Q: How accurate is GPT-3 at predicting Dx and triage advice compared to Harvard docs and to an average person Googling their symptoms?

For Dx, GPT-3 is much more accurate than your average person and almost as accurate as the Harvard docs! For triage, not so much.



7:42 pm · 1 Feb 2023 · **1,266** Views

https://twitter.com/AndrewLBeam/status/1620855064033382401?s=20&t=VO9_LdFFCj_wcwIQLvKcIQ



Source: Ilse Kant (UMC Utrecht)

what are these
machine learning methods?

MIT Technology Review

Artificial Intelligence / Machine Learning

What is machine learning?

Machine-learning algorithms find and apply patterns in data. And they pretty much run the world.

by **Karen Hao**

Nov 17, 2018

UMC Utrecht

“Everything is an ML method”

Scott H. Hawley
@drscotthawley

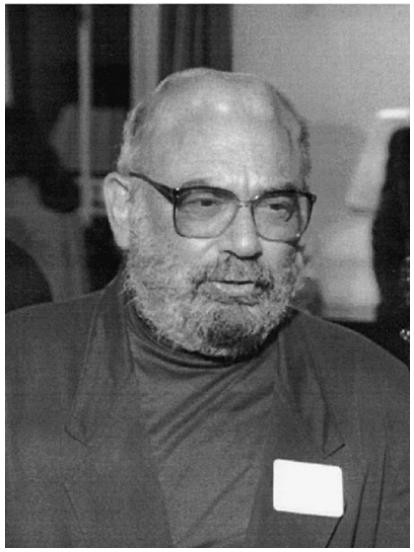
Replying to @JuliaHCox, @mikarv and @GSCollins

Logistic regression IS machine learning.

4:17 pm · 17 Feb 2019 · Twitter for iPhone

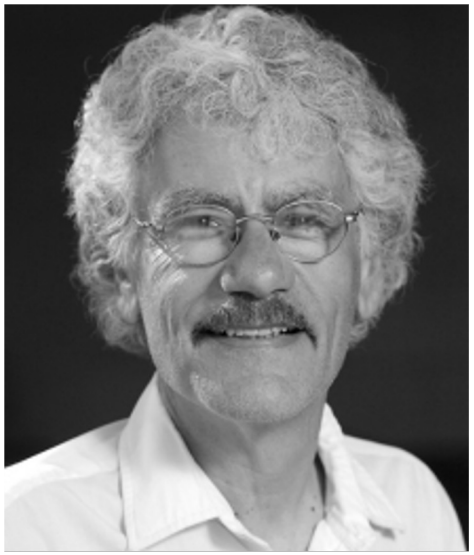
UMC Utrecht

“ML methods come from computer science”



Leo Breiman

CART, random forest



Jerome H Friedman

Gradient boosting

Trevor Hastie

Elements of statistical learning

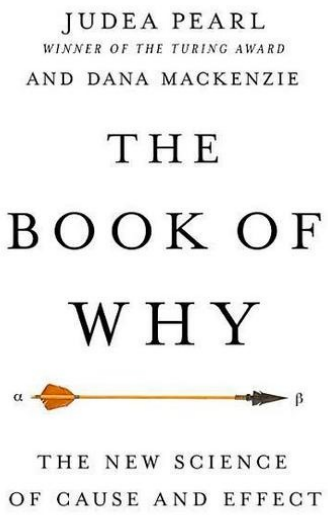
Education

UMC Utrecht

“ML methods for prediction, statistics for explaining”

ML and causal inference, small selection[1]

- Superlearner (e.g. van der Laan)
- High dimensional propensity scores (e.g. Schneeweiss)
- The book of why (Pearl)



UMC Utrecht

[1]See further: Kreiff and Diaz Ordaz; https://bit.ly/2m1eYdK

Two cultures

Statistical Science
2001, Vol. 16, No. 3, 199–231

Statistical Modeling: The Two Cultures

Leo Breiman

Abstract. There are two cultures in the use of statistical modeling to

UMC Utrecht

Language

Statistical modeling	Machine learning/AI
Estimating a model/Fitting	Learning
Prediction/Regression	Supervised learning
Latent variable modeling	Unsupervised learning
Case/Data point	Example/Instance
Sensitivity	Recall
Positive predictive value	Precision
Independent variable/Covariate	Feature
Dependent variable	Target
Response	Label
Parameters	Weights
Log likelihood	Loss
Structural equation model	Gaussian Bayesian network
Model for a categorical dependent variable	Classifier
Model for a continuous dependent variable	Regression
Model	Network, Graphs
Multinomial regression	Softmax
Prediction error	Error
Prediction of the sampling error	Variance
Average prediction error	Bias
Test set performance	Generalization
Contingency table	Confusion matrix
Criterion variable, reference test, gold standard	Ground truth
Overfitting	Overfitting
Measurement invariance	Transfer learning
Measurement error	Noise
Measurement error model (correction)	Noise aware machine learning
Measurement error model (estimation)	Inverse model
Deviance/Chi-square	Perplexity

UMC Utrecht

Machine learning: large grant = $1,000,000

Statistics: large grant = $50,000

UMC Utrecht

ML refers to a culture, not to methods

Distinguishing between statistics and machine learning

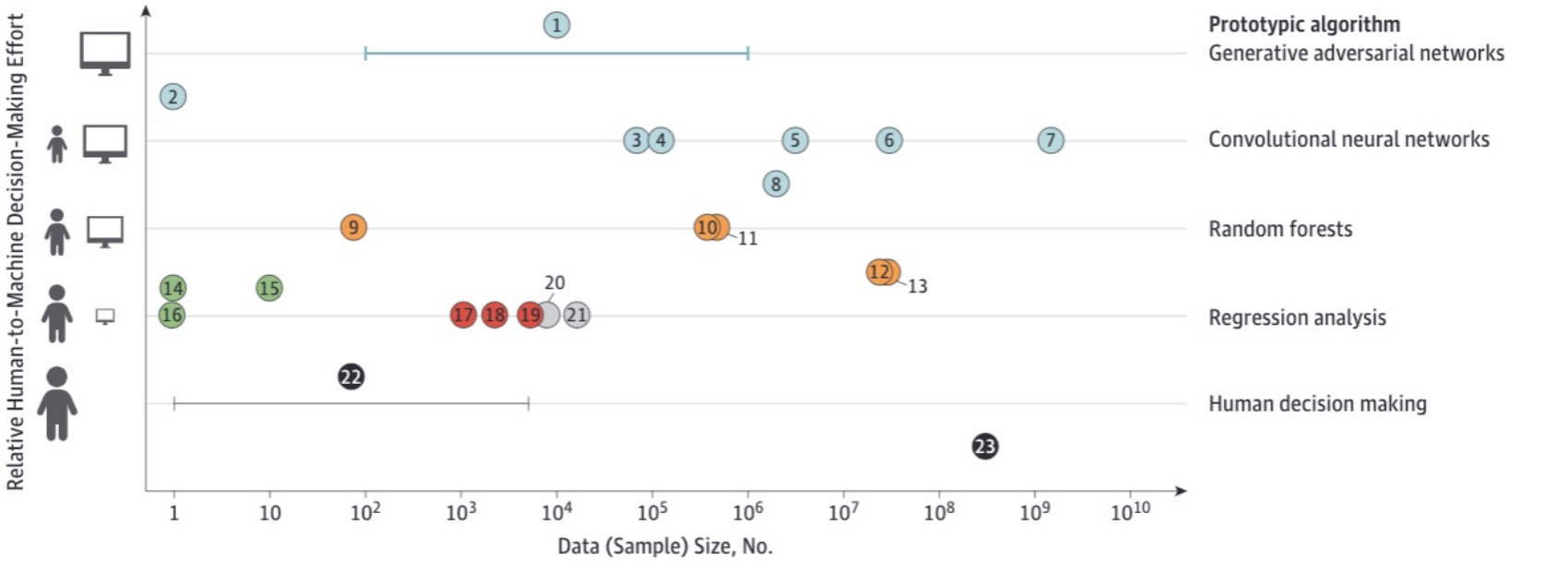
- Substantial **overlap methods** used by both cultures
- Substantial **overlap analysis goals**
- Attempts to separate the two frequently result in **disagreement**

Pragmatic approach:

I'll use "ML" to refer to models roughly outside of the traditional regression types of analysis: decision trees (and descendants), SVMs, neural networks (including Deep learning), boosting etc.

UMC Utrecht

Figure. The Axes of Machine Learning and Big Data



Deep learning

1. Generative adversarial networks (2014)
2. Google AlphaGo Zero (2017)
3. ATM check readers (1998)
4. Google diabetic retinopathy (2016)
5. ImageNet computer vision models (2012-2017)
6. Google AlphaGo (2015)
7. Facebook Photo Tagger (2015)
8. Prediction of 1-y all-cause mortality (2017)

Classic machine learning

9. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling (2002)
10. EHR-based CV risk prediction (2017)
11. Netflix Prize winner (2006)
12. Google Search (1998)
13. Amazon product recommendation (2003)

Expert AI systems

14. MYCIN (1975)
15. CASNET (1982)
16. DXplain (1986)

Risk calculators

17. CHA_2DS_2-VASc Score for atrial fibrillation stroke risk (2017)
18. MELD end-stage liver disease risk score (2001)
19. Framingham CV risk score (1998)

Randomized Clinical Trials

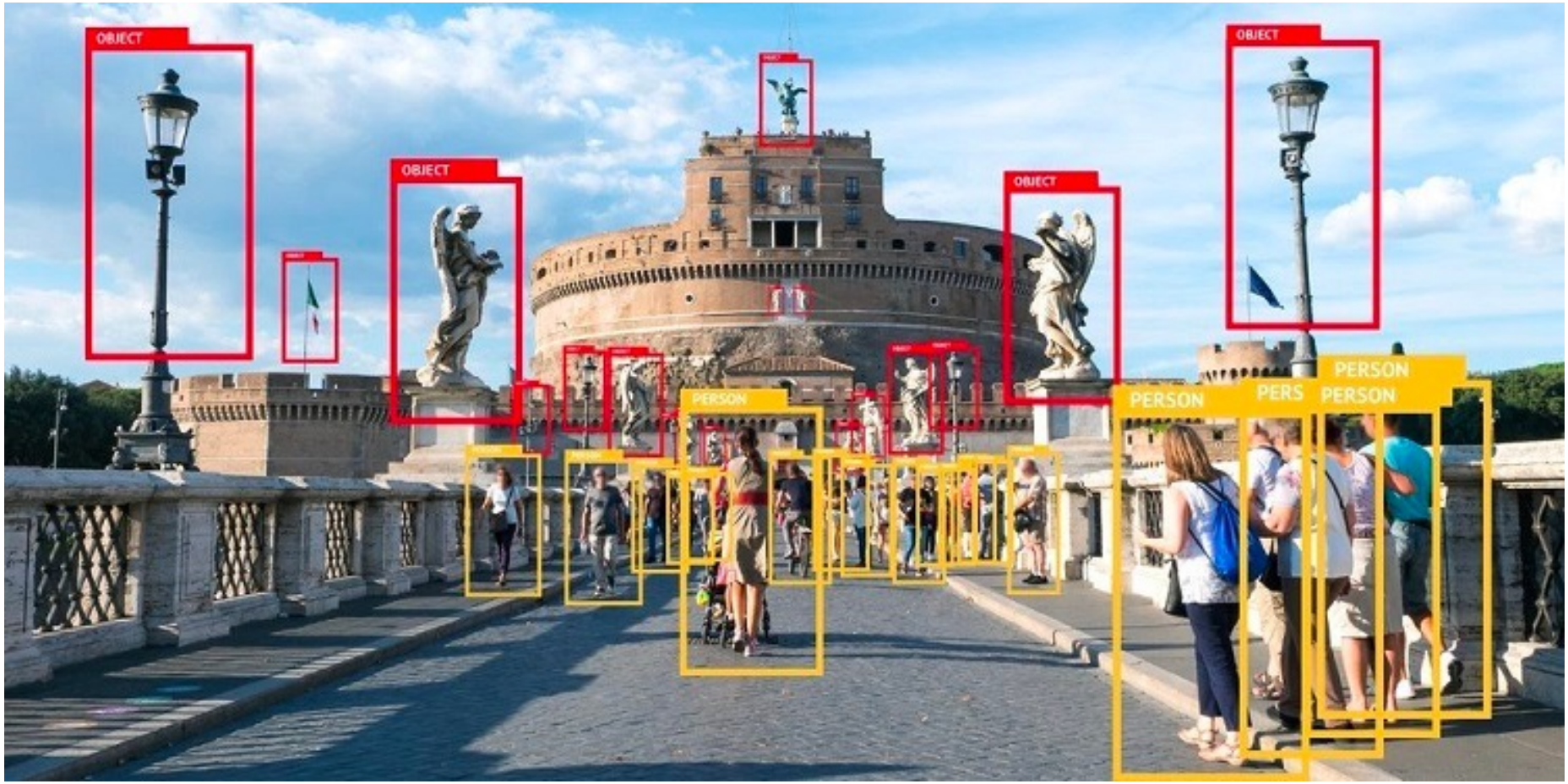
20. Celecoxib vs nonsteroidal anti-inflammatory drugs for osteoarthritis and rheumatoid arthritis (2002)
21. Use of estrogen plus progestin in healthy postmenopausal women (2002)

Other

22. Clinical wisdom
23. Mortality rate estimates from US Census (2010)

UMC Utrecht

Examples where “ML” has done well



UMC Utrecht

Example: retinal disease

Research

JAMA | Original Investigation | INNOVATIONS IN HEALTH CARE DELIVERY

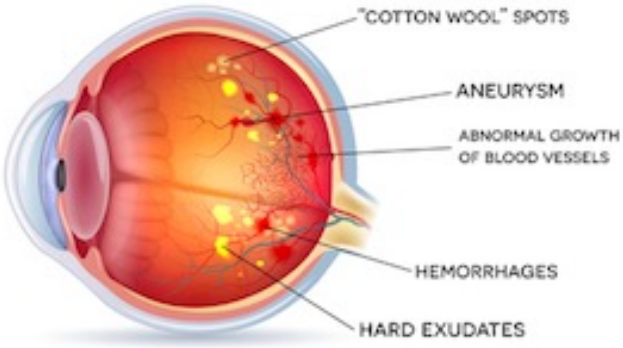
Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs

Varun Gulshan, PhD; Lily Peng, MD, PhD; Marc Coram, PhD; Martin C. Stumpe, PhD; Derek Wu, BS; Arunachalam Narayanaswamy, PhD; Subhashini Venugopalan, MS; Kasumi Widner, MS; Tom Madams, MEng; Jorge Cuadros, OD, PhD; Ramasamy Kim, OD, DNB; Rajiv Raman, MS, DNB; Philip C. Nelson, BS; Jessica L. Mega, MD, MPH; Dale R. Webster, PhD

Deep learning (= Neural network)

- 128,000 images
- Transfer learning (preinitialization)
- Sensitivity and specificity > .90
  - Estimated from training data

Diabetic retinopathy



UMC Utrecht

Gulshan et al, JAMA, 2016, 10.1001/jama.2016.17216; Picture retinopathy: https://bit.ly/2kB3X2w

Example: lymph node metastases

JAMA | Original Investigation

Diagnostic Assessment of Deep Learning Algorithms for Detection of Lymph Node Metastases in Women With Breast Cancer

Babak Ehteshami Bejnordi, MS; Mitko Veta, PhD; Paul Johannes van Diest, MD, PhD; Bram van Ginneken, PhD; Nico Karssemeijer, PhD; Geert Litjens, PhD; Jeroen A. W. M. van der Laak, PhD; and the CAMELYON16 Consortium

Deep learning competition

But:

- 390 teams signed up, 23 submitted
- "Only" 270 images for training
- Test AUC range: 0.56 to 0.99

	Task 1: Metastasis Identification	Task 2: Metastases Classification
Codename[b]	FROC Score (95% CI)[c]	AUC (95% CI)[c]
HMS and MIT II	0.807 (0.732-0.889)	0.994 (0.983-0.999)
HMS and MGH III	0.760 (0.692-0.857)	0.976 (0.941-0.999
HMS and MGH I	0.596 (0.578-0.734)	0.964 (0.928-0.989)
VISILAB II	0.116 (0.063-0.177)	0.651 (0.549-0.742)
Anonymous I	0.097 (0.049-0.158)	0.628 (0.530-0.717)
Laboratoire d'Imagerie Biomédicale I	0.120 (0.079-0.182)	0.556 (0.434-0.654)

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices

Michael D. Abràmoff[1,2,3,4], Philip T. Lavin[5], Michele Birch[6], Nilay Shah[7] and James C. Folk[1,2,3]

The AI system exceeded all pre-specified superiority endpoints at sensitivity of 87.2% (95% CI, 81.8–91.2%) (>85%), specificity of 90.7% (95% CI, 88.3–92.7%) (>82.5%), and imageability rate of 96.1% (95% CI, 94.6–97.3%), demonstrating AI's ability to bring specialty-level diagnostics to primary care settings.

FDA Homepage

Search Menu

IN THIS SECTION

← Press Announcements

FDA NEWS RELEASE

FDA permits marketing of artificial intelligence-based device to detect certain diabetes-related eye problems

Share Tweet Email

For Immediate Release:

April 11, 2018

Español

The U.S. Food and Drug Administration today permitted marketing of the first medical device to use artificial intelligence to detect greater than a mild level of the eye disease diabetic retinopathy in adults who have diabetes.

UMC Utrecht

PLOS MEDICINE

RESEARCH ARTICLE

Computer-aided X-ray screening for tuberculosis and HIV testing among adults with cough in Malawi (the PROSPECT study): A randomised trial and cost-effectiveness analysis

Peter MacPherson[1,2,3]*, Emily L. Webb[4], Wala Kamchedzera[2], Elizabeth Joekes[1], Gugu Mjoli[5], David G. Lalloo[1], Titus H. Divala[2,3,6], Augustine T. Choko[1,2], Rachael M. Burke[2,3], Hendramoorthy Maheswaran[7], Madhukar Pai[8], S. Bertel Squire[1], Marriott Nliwasa[2,6], Elizabeth L. Corbett[2,3]

1 Department of Clinical Sciences, Liverpool School of Tropical Medicine, Liverpool, United Kingdom, **2** Malawi-Liverpool-Wellcome Trust Clinical Research Programme, Blantyre, Malawi, **3** Clinical Research Department, London School of Hygiene and Tropical Medicine, London, United Kingdom, **4** MRC Tropical Epidemiology Group, London School of Hygiene and Tropical Medicine, London, United Kingdom, **5** Department of Radiology, Chris Hani Baragwanath Hospital, Soweto, South Africa, **6** Helse Nord TB Initiative, College of Medicine, University of Malawi, Blantyre, Malawi, **7** Department of Public Health and Policy, University of Liverpool, Liverpool, United Kingdom, **8** McGill International TB Centre, McGill University, Montreal, Canada

Check for updates

OPEN ACCESS

Primary outcome: time to TB treatment.

Time to TB treatment lowered from a median of 11 days in standard of care to 1 day with computer aided X-ray screening

UMC Utrecht

Radiology

ORIGINAL RESEARCH • PEDIATRIC IMAGING

Artificial Intelligence Algorithm Improves Radiologist Performance in Skeletal Age Assessment: A Prospective Multicenter Randomized Controlled Trial

David K. Eng, MS • Nishith B. Khandwala, MS • Jin Long, PhD • Nancy R. Fefferman, MD • Shailee V. Lala, MD • Naomi A. Strubel, MD • Sarah S. Milla, MD • Ross W. Filice, MD • Susan E. Sharp, MD • Alexander J. Towbin, MD • Michael L. Francavilla, MD • Summer L. Kaplan, MD • Kirsten Ecklund, MD • Sanjay P. Prabhu, MD • Brian J. Dillon, MD • Brian M. Everist, MD • Christopher G. Anton, MD • Mark E. Bittman, MD • Rebecca Dennis, DO • David B. Larson, MD, MBA • Jayne M. Seekins, DO • Cicero T. Silva, MD • Arash R. Zandieh, MD • Curtis P. Langlotz, MD, PhD , • Matthew P. Lungren, MD, MPH • Safwan S. Halabi, MD

Materials and Methods: In this prospective randomized controlled trial, the accuracy of skeletal age assessment on hand radiograph examinations was performed with (n = 792) and without (n = 739) the AI algorithm as a diagnostic aid. For examinations with the AI algorithm, the radiologist was shown the AI interpretation as part of their routine clinical work and was permitted to accept or modify it. Hand radiographs were interpreted by 93 radiologists from six centers. The primary efficacy outcome was the mean absolute difference between the skeletal age dictated into the radiologists' signed report and the average interpretation of a panel of four radiologists not using a diagnostic aid. The secondary outcome was the interpretation time. A linear mixed-effects regression model with random center- and radiologist-level effects was used to compare the two experimental groups.

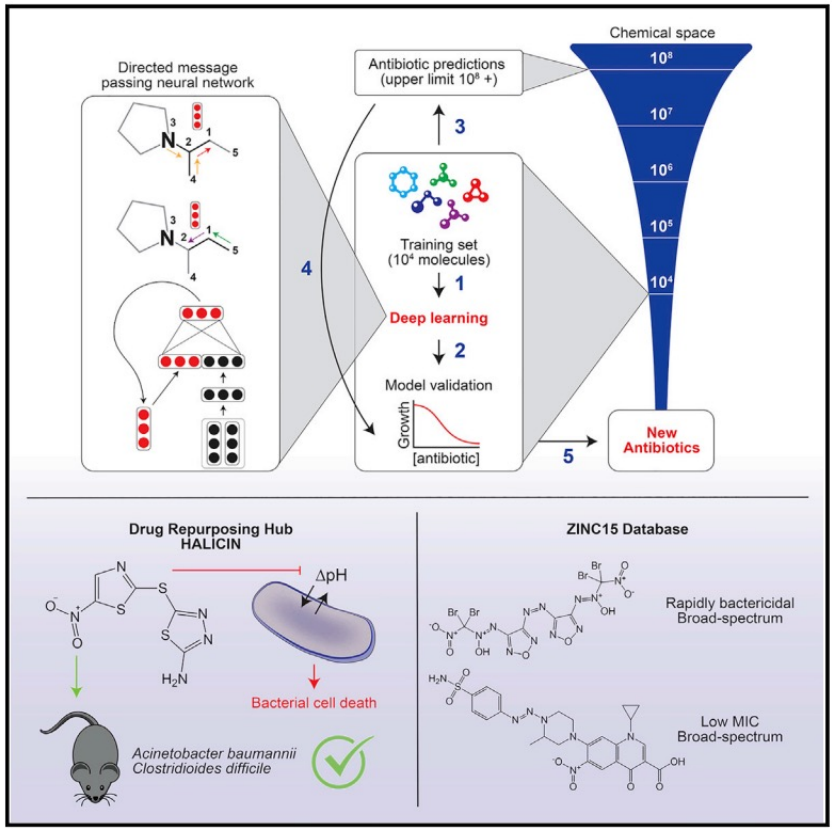
Results: Overall mean absolute difference was lower when radiologists used the AI algorithm compared with when they did not (5.36 months vs 5.95 months; P = .04). The proportions at which the absolute difference exceeded 12 months (9.3% vs 13.0%, P = .02) and 24 months (0.5% vs 1.8%, P = .02) were lower with the AI algorithm than without it. Median radiologist interpretation time was lower with the AI algorithm than without it (102 seconds vs 142 seconds, P = .001).

UMC Utrecht

Cell

A Deep Learning Approach to Antibiotic Discovery

Graphical Abstract



Authors

Jonathan M. Stokes, Kevin Yang, Kyle Swanson, ..., Tommi S. Jaakkola, Regina Barzilay, James J. Collins

Correspondence

regina@csail.mit.edu (R.B.), jimjc@mit.edu (J.J.C.)

In Brief

A trained deep neural network predicts antibiotic activity in molecules that are structurally different from known antibiotics, among which Halicin exhibits efficacy against broad-spectrum bacterial infections in mice.

Highlights

- A deep learning model is trained to predict antibiotics based on structure

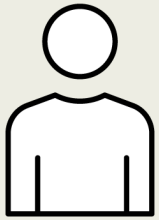
10.1016/j.cell.2020.01.021

JAMA Oncology

RCT: Long-term Effect of Machine Learning–Triggered Behavioral Nudges on Serious Illness Conversations and End-of-Life Outcomes Among Patients With Cancer

PATIENT ENCOUNTERS

18 762 Men, 22 259 Women



Patients with cancer without a prior documented serious illness conversation (SIC)

Mean age, 62 y

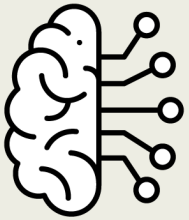
SETTINGS / LOCATIONS

9 Oncology clinics in Pennsylvania

INTERVENTION

41 021 Patient encounters with 20 506 patients

12 356 Usual care



28 665 Machine learning–based intervention

Peer-comparison emails, machine learning–based high-risk patient lists, and opt-out text prompts

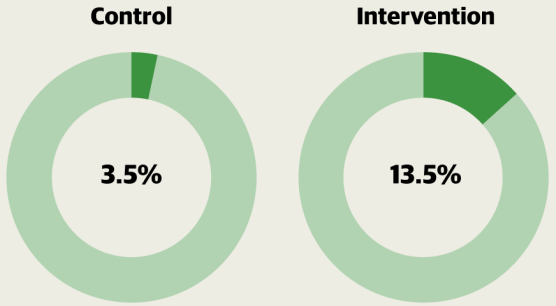
PRIMARY OUTCOME

Rates of SICs with all oncology patients and those with a high mortality risk between control vs intervention periods

FINDINGS

The machine learning–based behavioral intervention was associated with significantly more SICs with outpatients with cancer; a greater increase was seen for patients at high risk for death

Rates of SIC in control vs intervention groups among high-risk patients



Adjusted odds ratio (aOR) of SIC

All patients vs control: aOR, 2.09 (95% CI, 1.53-2.87)
Patients at high risk vs control: aOR, 2.62 (95% CI, 1.84-3.72)

Manz CR, Zhang Y, Chen K, et al. Long-term effect of machine learning-triggered behavioral nudges on serious illness conversations and end-of-life outcomes among patients with cancer: a randomized clinical trial. *JAMA Oncol*. Published online January 12, 2023. doi:10.1001/jamaoncol.2022.6303

UMC Utrecht

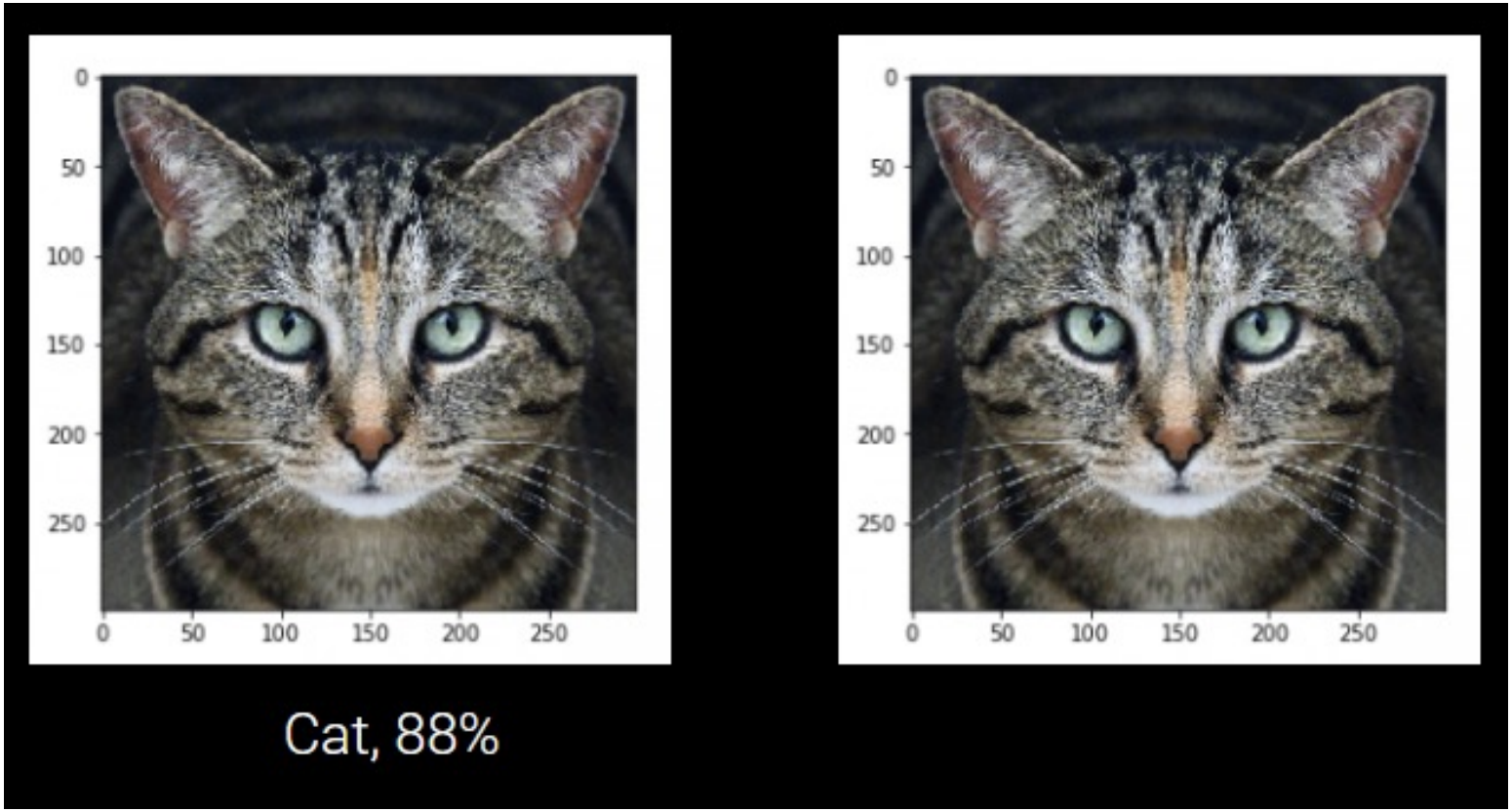
Examples where “ML” has done poorly

Granny Smith	85.6%
iPod	0.4%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.1%

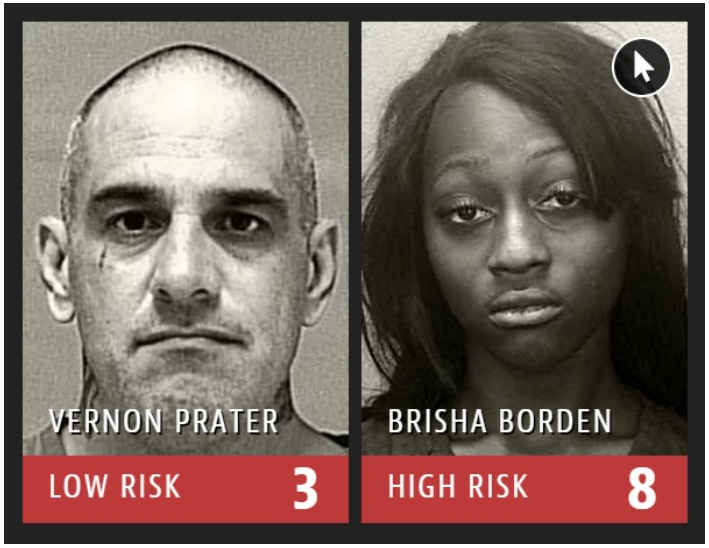
Granny Smith	0.1%
iPod	99.7%
library	0.0%
pizza	0.0%
toaster	0.0%
dough	0.0%

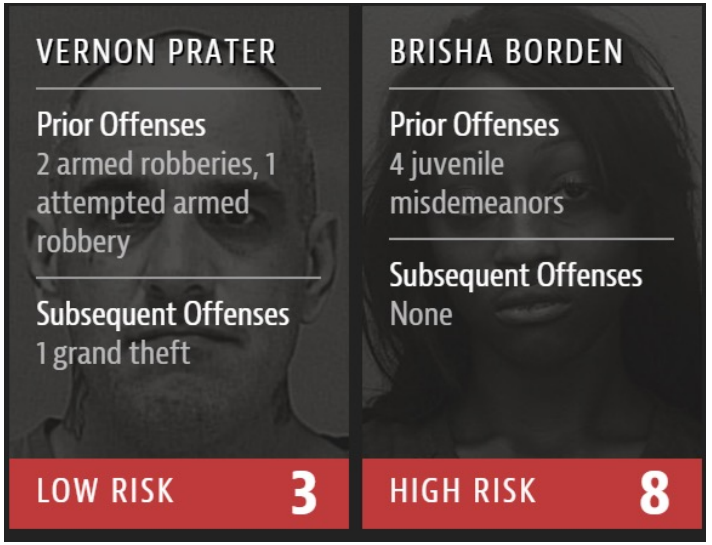
UMC Utrecht

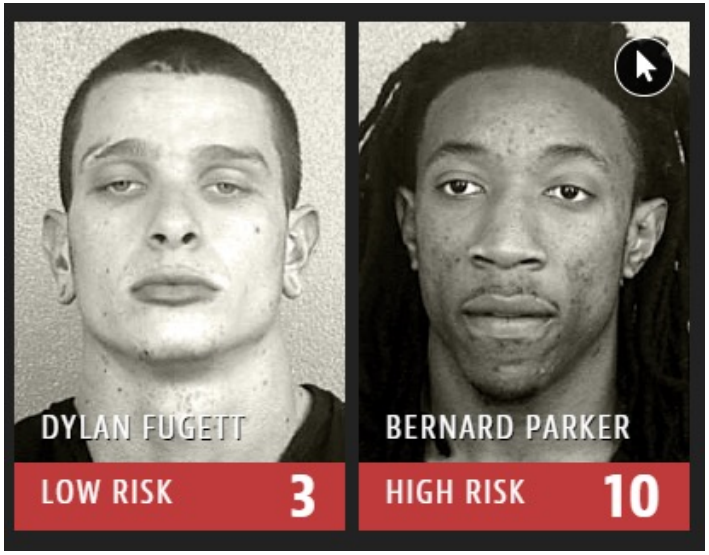
Adversarial examples



Recidivism Algorithm







UMC Utrecht

Skin cancer and rulers

Dermatologist-level classification of skin cancer with deep neural networks

Andre Esteva[1]*, Brett Kuprel[1]*, Roberto A. Novoa[2,3], Justin Ko[2], Susan M. Swetter[2,4], Helen M. Blau[5] & Sebastian Thrun[6]

Skin cancer, the most common human malignancy[1–3], is primarily diagnosed visually, beginning with an initial clinical screening and followed potentially by dermoscopic analysis, a biopsy and histopathological examination. Automated classification of skin lesions using images is a challenging task owing to the fine-grained variability in the appearance of skin lesions. Deep convolutional neural networks (CNNs)[4,5] show potential for general and highly variable tasks across many fine-grained object categories[6–11]. Here we demonstrate classification of skin lesions using a single CNN, trained end-to-end from images directly, using only pixels and disease labels as inputs. We train a CNN using a dataset of 129,450 clinical images—two orders of magnitude larger than previous datasets[12]—consisting of 2,032 different diseases. We test its performance against 21 board-certified dermatologists on

images (for example, smartphone images) exhibit variability in factors such as zoom, angle and lighting, making classification substantially more challenging[23,24]. We overcome this challenge by using a data-driven approach—1.41 million pre-training and training images make classification robust to photographic variability. Many previous techniques require extensive preprocessing, lesion segmentation and extraction of domain-specific visual features before classification. By contrast, our system requires no hand-crafted features; it is trained end-to-end directly from image labels and raw pixels, with a single network for both photographic and dermoscopic images. The existing body of work uses small datasets of typically less than a thousand images of skin lesions[16,18,19], which, as a result, do not generalize well to new images. We demonstrate generalizable classification with a new dermatologist-labelled dataset of 129,450 clinical images, including

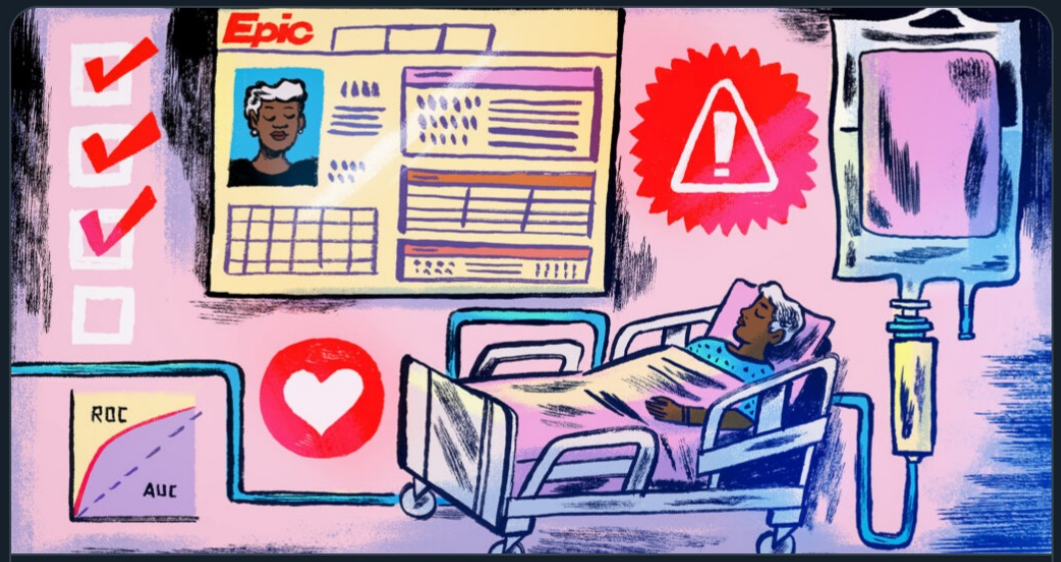
He and his colleagues had one such problem in their their study with rulers. When dermatologists are looking at a lesion that they think might be a tumor, they'll break out a ruler—the type you might have used in grade school—to take an accurate measurement of its size. Dermatologists tend to do this only for lesions that are a cause for concern. So in the set of biopsy images, if an image had a ruler in it, the algorithm was more likely to call a tumor malignant, because the presence of a ruler correlated with an increased likelihood a lesion was cancerous. Unfortunately, as Novoa emphasizes, the algorithm doesn't know why that correlation makes sense, so it could easily misinterpret a random ruler sighting as grounds to diagnose cancer.

Casey Ross
@caseymross

Epic's algorithm to provide early warning of sepsis uses active antibiotic orders as a predictor. That makes it accurate in a computer lab, but unlikely to offer advance notice to clinicians in the real world.

My latest for @statnews:

statnews.com/2021/09/27/epi... via @statnews



Epic's sepsis algorithm struggles in the real world. Its variables may be why

A STAT investigation found Epic did not fully examine the real-world impact of its sepsis model, which uses several demographic variables.

statnews.com

3:52 pm · 27 Sep 2021 · Twitter for iPhone

Predicting mortality – the conclusion

RESEARCH ARTICLE

Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease

Conclusion

We have demonstrated that machine learning approaches on routine EHR data can achieve comparable or better performance than expert-selected, imputed data in manually optimised models for risk prediction. Our comparison of a range of machine learning algorithms found that elastic net regression performed best, with cross-validated variable selection based on log-rank tests enabling Cox models and random forests to achieve comparable performance.

UMC Utrecht

PlosOne, 2018, DOI: 10.1371/journal.pone.0202344

Predicting mortality – the results

RESEARCH ARTICLE

Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease

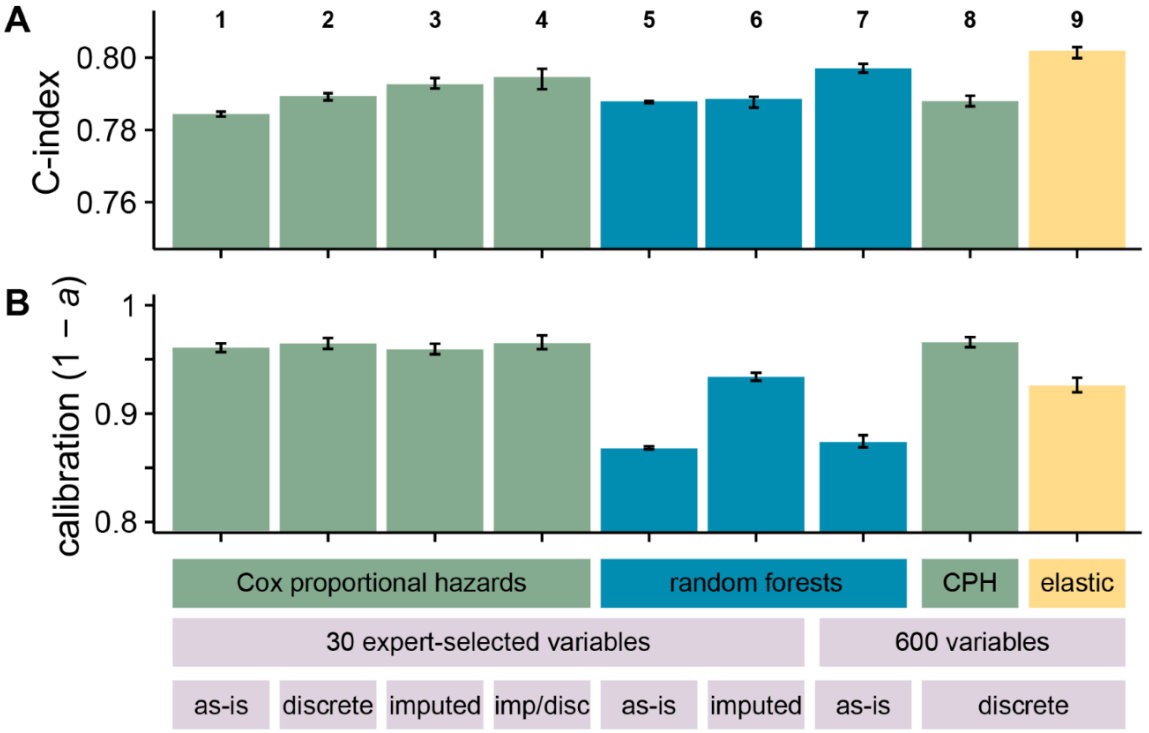


Fig 1. Overall discrimination and calibration performance for the different models and datasets used. (A) shows discrimination (C-

UMC Utrecht

PlosOne, 2018, DOI: 10.1371/journal.pone.0202344

Predicting mortality – the media

RESEARCH ARTICLE

Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease

THE FRANCIS CRICK INSTITUTE

RESEARCH CAREERS AND STUDY PARTNERSHIPS WHAT'S ON NEWS ABOUT US

AI beats doctors at predicting heart disease deaths

4 SEPTEMBER 2018 HUMAN BIOLOGY HEALTH AND AGEING NEWS

AI NEWS RESEARCH —

Artificial Intelligence beats doctors at predicting heart disease deaths

BY SHACK15 - 5 SEPTEMBER, 2018

ScienceDaily

Your source for the latest research news

SD Health Tech Enviro Society Quirky

Science News *from research organizations*

AI beats doctors at predicting heart disease deaths

Date: September 4, 2018

UMC Utrecht

PlosOne, 2018, DOI: 10.1371/journal.pone.0202344; https://bit.ly/2Q6H41R; https://bit.ly/2m3RLrn

HYPE!

Systematic review clinical prediction models

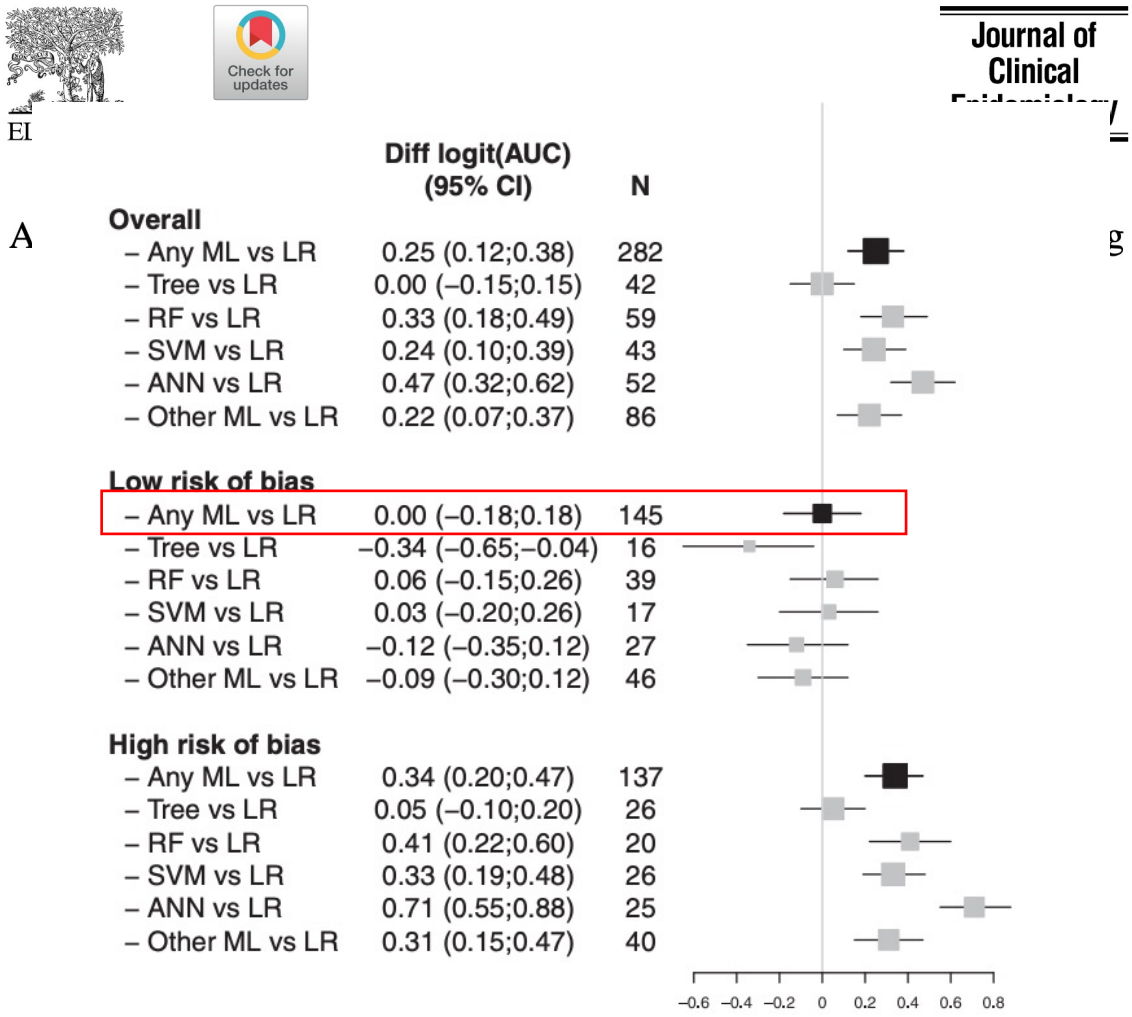


Fig. 4. Differences in discriminative ability between LR and ML models, overall and according to risk of bias (n = 282 comparisons).

Sources of prediction error

$$Y = f(x) + \varepsilon$$

For a model k the **expected test prediction** error is:

$$\sigma^2 + \text{bias}^2\left(\hat{f}_k(x)\right) + \text{var}\left(\hat{f}_k(x)\right)$$

Irreducible error

$\approx$

What we don't model

Mean squared prediction error

$\approx$

How we model

$$(\text{with } E(\varepsilon) = 0, \text{var}(\varepsilon) = \sigma^2, \text{values in } x \text{ are not random})$$

UMC Utrecht

See equation 2.46 in Hastie et al., the elements of statistical learning, https://stanford.io/2voWjra

Sources of prediction error

In words, two main components for error in predictions are:

- **Mean squared predictor error**
  - Under control of the modeler

$$Y = f(x) + \varepsilon$$

For a model k the **expected test prediction** error is:

$$\sigma^2 + \text{bias}^2\left(\hat{f}_k(x)\right) + \text{var}\left(\hat{f}_k(x)\right)$$

Irreducible error $\approx$ **What** we don't model

Mean squared prediction error $\approx$ **How** we model

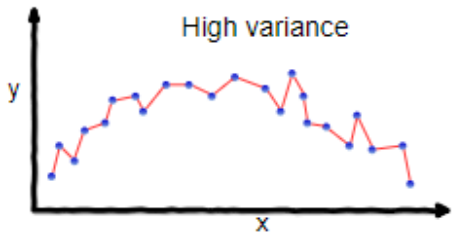
$$(\text{with } \mathrm{E}(\varepsilon) = 0, \text{var}(\varepsilon) = \sigma^2, \text{values in } x \text{ are not random})$$

See equation 2.46 in Hastie et al., the elements of statistical learning, https://stanford.io/2voWjra

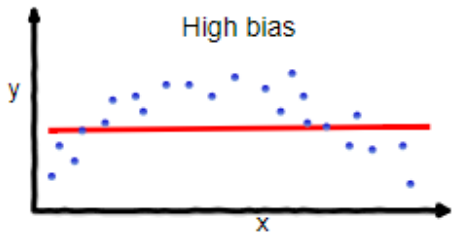
Sources of prediction error

In words, two main components for error in predictions are:

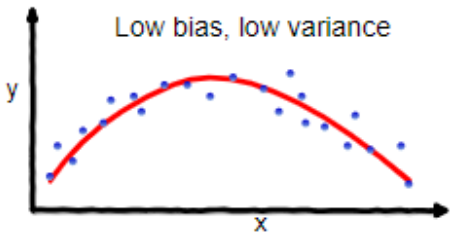
- **Mean squared predictor error**
  - Under control of the modeler



overfitting



underfitting



”just right”

See equation 2.46 in Hastie et al., the elements of statistical learning, https://stanford.io/2voWjra

Sources of prediction error

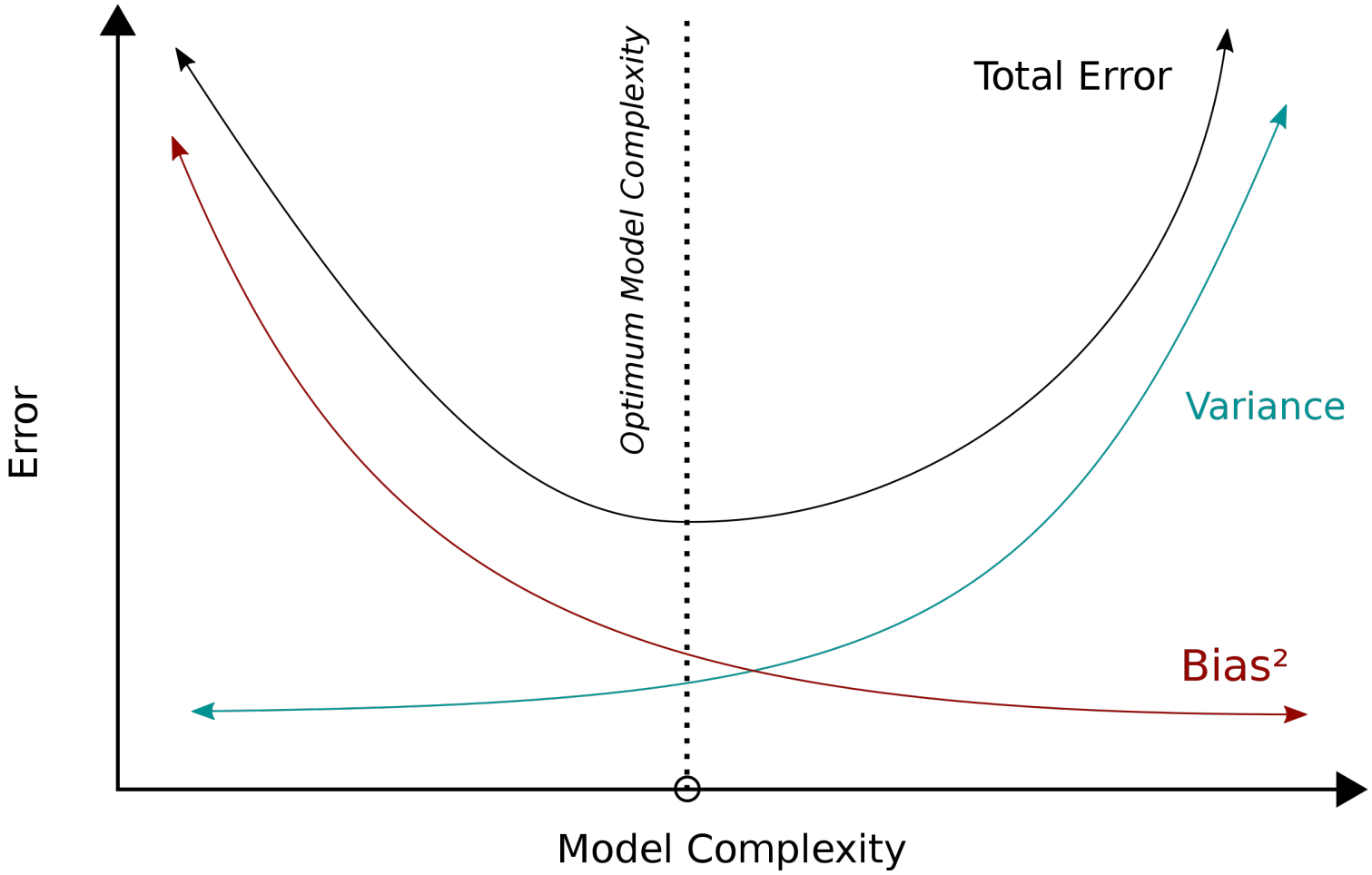
In words, two main components for error in predictions are:

- **Mean squared predictor error**
  - Under control of the modeler
- **Irreducible error**
  - Not under direct control of the modeler

See equation 2.46 in Hastie et al., the elements of statistical learning, https://stanford.io/2voWjra

UMC Utrecht

Bias-variance trade-off



UMC Utrecht

Irreducible error is often large

- Health and lack thereof complex to measure ('no gold standard')
- Predictors of diseases are often **imperfectly and partly measured**
- We often **don't know all the causal mechanisms** at play
  - much easier to predict if you know the causal mechanisms!
- "Prediction is very difficult, especially if it's about the future!" (Niels Bohr might have said this first)

UMC Utrecht

What can we do to reduce “irreducible” error?

- **Changing the information**

- Prognostication by text mining electronic health records
  - e.g. predicting life expectancy
    https://bit.ly/2k8Ao8e

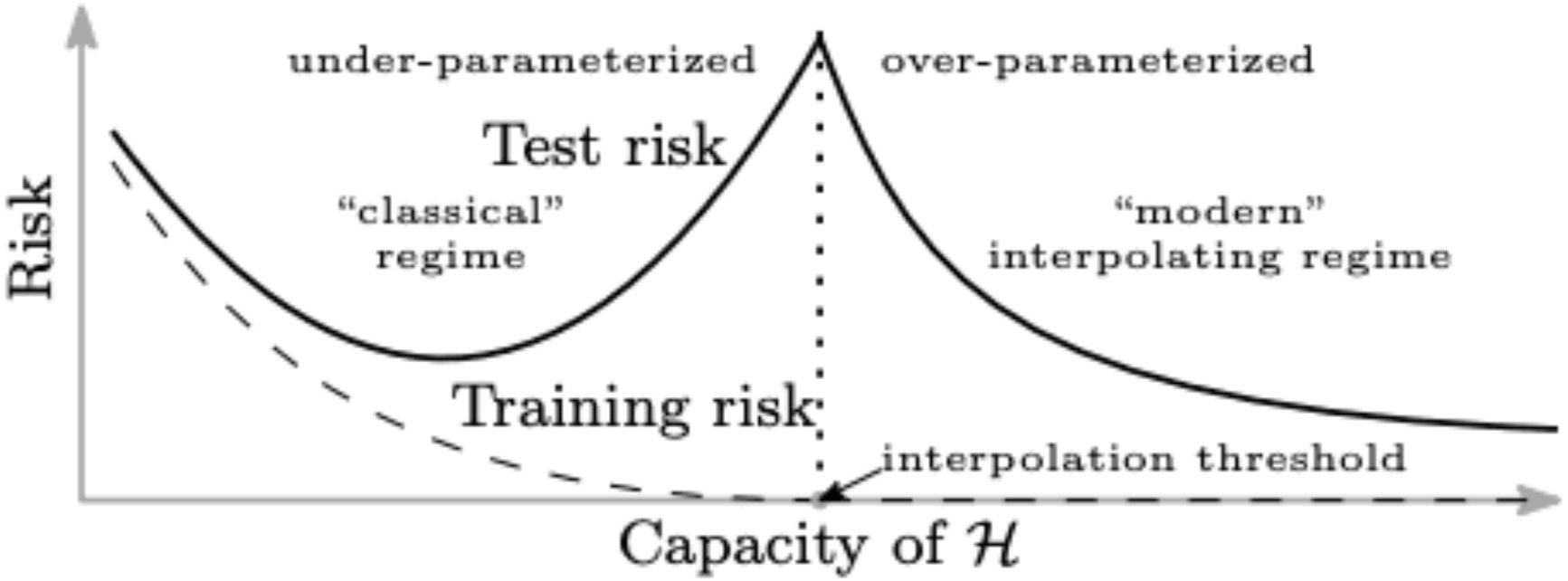
- Analyzing social media posts
  - e.g. pharmacovigilance, adverse events monitoring via Twitter posts
    https://bit.ly/2m0KKrg

- Speech signal processing
  - e.g. Parkinson‟s disease,
    https://bit.ly/2v3ZdHR

- Medical imaging

UMC Utrecht

Bias-variance trade-off revisited: double descent



But…

Statistical Science
2006, Vol. 21, No. 1, 1–14
DOI 10.1214/088342306000000060

Classifier Technology and the Illusion of Progress

David J. Hand

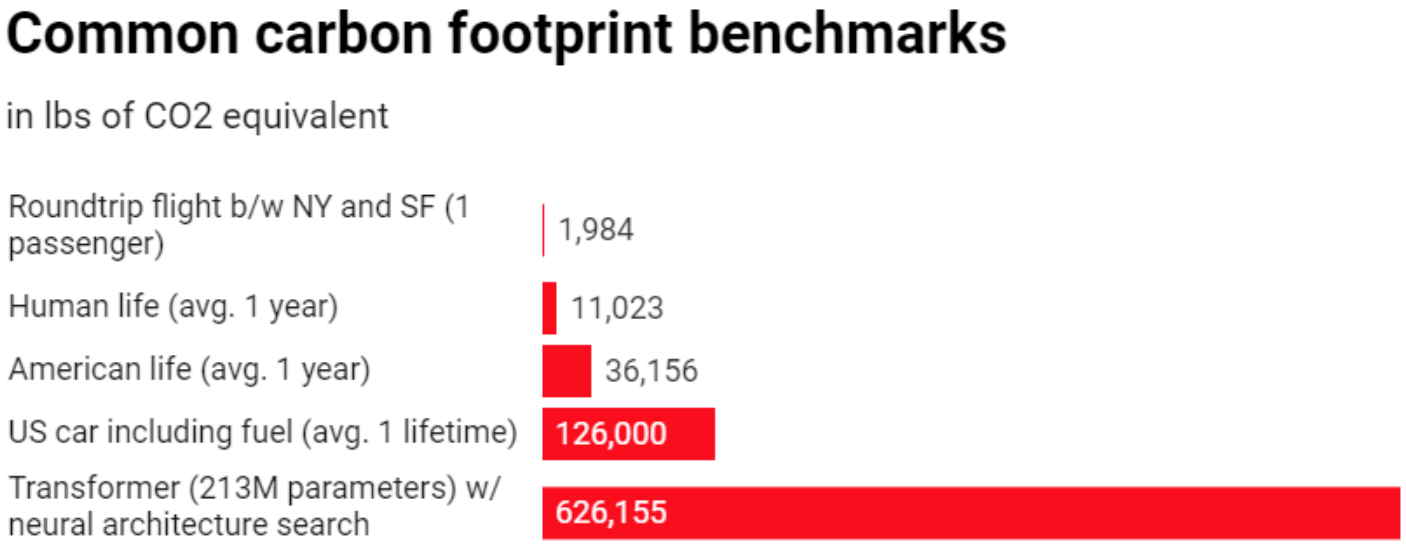
later stages. Furthermore, if one looks at the historical development of classification methods, then the earlier approaches involve relatively simple structures (e.g., the linear forms of linear or logistic discriminant analysis), while more recent approaches involve more complicated structures (e.g., the decision surfaces of neural networks or support vector machines). It follows that the simple approaches will have led to greater improvement in predictive performance than the later approaches which are necessarily trying to improve on the predictive performance obtained by the simpler earlier methods. Put another way, there is a law of diminishing returns.

UMC Utrecht

Flexible algorithms are data hungry

UMC Utrecht

Flexible algorithms are energy hungry



The costs of running (cloud computing) the Transformer algorithm are estimated at

UMC Utrecht

npj | Digital Medicine

www.nature.com/npjdigitalmed

ARTICLE OPEN

Check for updates

Impact of a deep learning assistant on the histopathologic classification of liver cancer

Amirhossein Kiani[1,6], Bora Uyumazturk[1,6], Pranav Rajpurkar[1,6], Alex Wang[1], Rebecca Gao[2], Erik Jones[1], Yifan Yu[1], Curtis P. Langlotz[3,4], Robyn L. Ball[3], Thomas J. Montine[3,5], Brock A. Martin[5], Gerald J. Berry[5], Michael G. Ozawa[5], Florette K. Hazard[5], Ryanne A. Brown[5], Simon B. Chen[5], Mona Wood[5], Libby S. Allard[5], Lourdes Ylagan[5], Andrew Y. Ng[1,7]✉ and Jeanne Shen[3,5,7]✉

Table 3. Impact of assistance on diagnostic accuracy under different conditions[a].

	Assistance (11 pathologists)	Assistance (9 pathologists)	Model correct (11 pathologists)	Model incorrect (11 pathologists)
OR (95% CI)	1.281 (0.882, 1.862)	1.499 (1.007, 2.230)	4.289 (2.360, 7.794)	0.253 (0.126, 0.507)
p-value	0.184	0.045	0.000	0.000

[a]The results of mixed-effect logistic regression analyses evaluating the impact of assistance on diagnostic accuracy are presented as odds ratios (OR) for pathologist diagnostic accuracy, with 95% confidence intervals (95% CI) and *p*-values from likelihood ratio testing (a two-tailed $p \leq 0.05$ was considered statistically significant).

UMC Utrecht

Algorithm based medicine

- Algorithms are high maintenance
  - Developed models need **repeated testing and updating** to remain useful **over time and place**
  - Many new barriers: black box **proprietary algorithms**, computing costs

- **Regulation** and **quality control** of algorithms
  - Algorithms need testing, preferably in **experimental fashion**

UMC Utrecht

← **Tweet**

Hugh Harvey
@DrHughHarvey

Some say data is the new oil. Others say the algorithm is the gold.

Actually, it's the labels.

Quality labels are hard to come by, expensive to generate, and without them, safe, accurate algorithmic performance isn't possible.

8:53 pm · 19 Feb 2020 · Twitter Web App

UMC Utrecht

Old statistics wine in new machine learning bottles?

Lots of…

- **Hype**
- **Rebranding** traditional analysis as ML and AI
- Methodological **reinventions**
- **Traditional issues** such as low sample size, lack of adequate validation, poor reporting

Also, real developments in…

- **Methods** and architectures, allowing for modeling (unstructured) data that could previously not easily be used
- **Software**
- **Computing** power
- **Clinical trials** showing benefit of AI assistance

UMC Utrecht

From research AI model to implemented AI application, innovation is ….

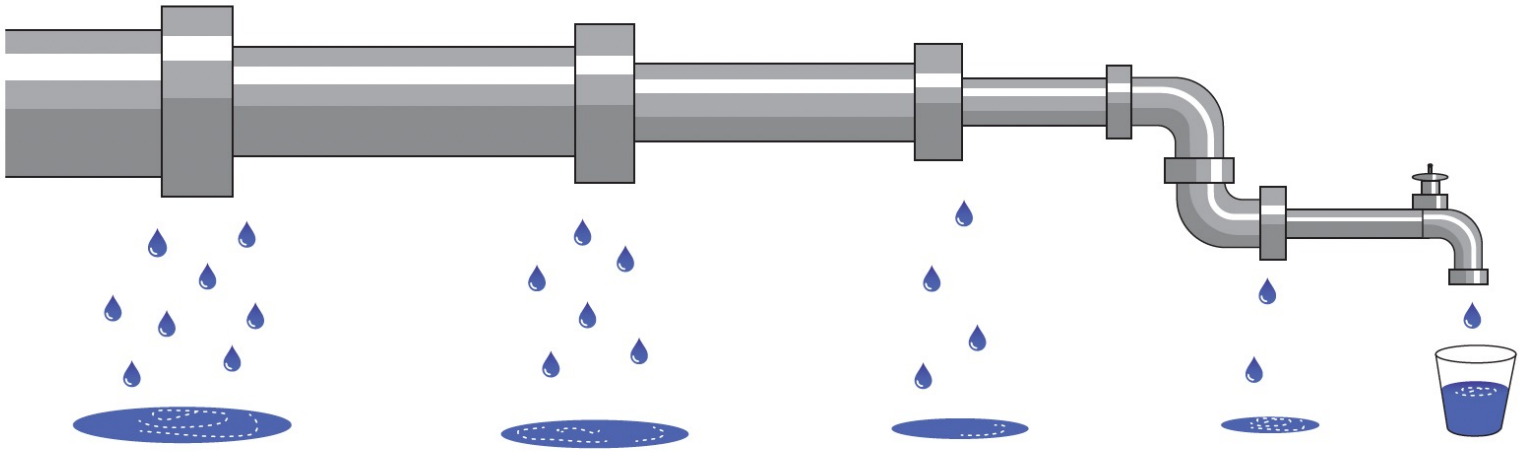
3000	100	10	2	1
Ideas	Explorations	well defined projects	Launches	Succes

UMC Utrecht

Source: Ilse Kant (UMC Utrecht), adapted from doi: 10.1080/08956308.1997.11671126


AI/ML MODELS USED IN PRACTICE
AI/ML MODELS THAT WILL NEVER BE USED IN PRACTICE

Pipeline of algorithmic medicine failure

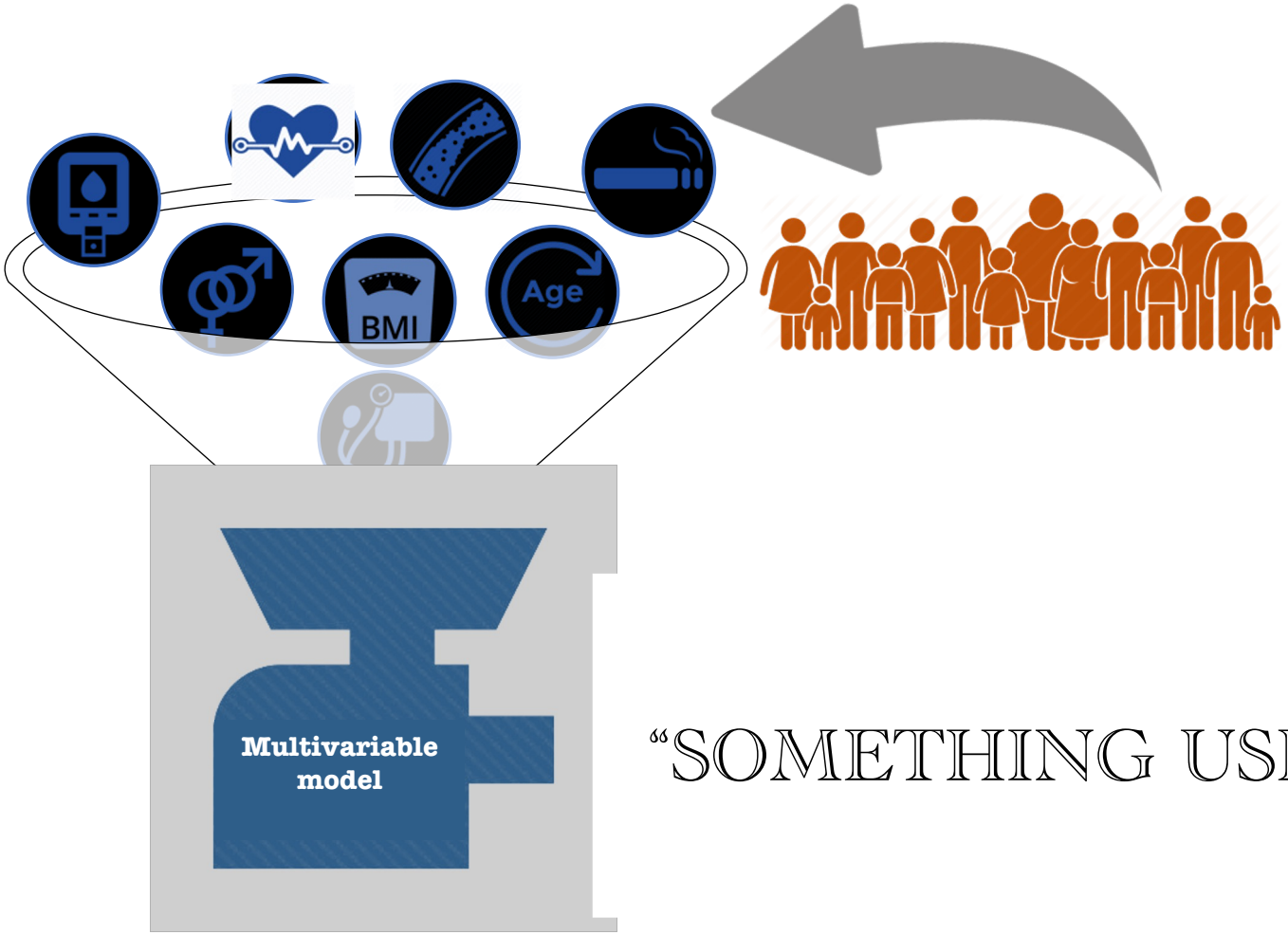


Not fit for purpose	No validation	No implementation	Not adopted
Developed on wrong patient population	Lack of data or incentive to pursue validation studies	No impact on decision making or patient (health) outcomes	Prediction (perceived as) not useful
Expensive or non-available predictors	Incompletely reported prediction model	No software developed to implement and use the model	Predictions not trusted
Time intensive to use model	Poorly developed or overfitted model	Requirements for adherence to (medical device) regulations	Model not transparent enough, or no tools available to enhance its use in practice
Outcome measured unreliably	Proprietary model code	Cost(-effectiveness) of using proprietary model	Model (perceived as) outdated

FIGURE 1 Leaky prognostic model adoption pipeline. Examples of reasons for failed prediction model adoption in clinical practice.

UMC Utrecht

Utopia



"SOMETHING USEFUL"

AI/ML models are…

UMC Utrecht

AI/ML models are…

- Expensive
- Not one-size-fits-all
- Many alternatives usually available
- Need crash testing (“impact”)
- Require regular MOT (“validation”)
- Require regular maintenance (”updating”)
- Require people to be trained how to operate them
- Can be dangerous when wrongly used
- Regulations apply

UMC Utrecht

npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review

Anne A. H. de Hond[1,2,3,8]✉, Artuur M. Leeuwenberg[4,8]✉, Lotty Hooft[4,5], Ilse M. J. Kant[1,2,3], Steven W. J. Nijman[4], Hendrikus J. A. van Os[2,6], Jiska J. Aardoom[6,7], Thomas P. A. Debray[4], Ewoud Schuit[4], Maarten van Smeden[4], Johannes B. Reitsma[4], Ewout W. Steyerberg[2,3], Niels H. Chavannes[6,7] and Karel G. M. Moons[4]

While the opportunities of ML and AI in healthcare are promising, the growth of complex data-driven prediction models requires careful quality and applicability assessment before they are applied and disseminated in daily practice. This scoping review aimed to identify actionable guidance for those closely involved in AI-based prediction model (AIPM) development, evaluation and implementation including software engineers, data scientists, and healthcare professionals and to identify potential gaps in this guidance. We performed a scoping review of the relevant literature providing guidance or quality criteria regarding the development, evaluation, and implementation of AIPMs using a comprehensive multi-stage screening strategy. PubMed, Web of Science, and the ACM Digital Library were searched, and AI experts were consulted. Topics were extracted from the identified literature and summarized across the six phases at the core of this review: (1) data preparation, (2) AIPM development, (3) AIPM validation, (4) software development, (5) AIPM impact assessment, and (6) AIPM implementation into daily healthcare practice. From 2683 unique hits, 72 relevant guidance documents were identified. Substantial guidance was found for data preparation, AIPM development and AIPM validation (phases 1–3), while later phases clearly have received less attention (software development, impact assessment and implementation) in the scientific literature. The six phases of the AIPM development, evaluation and implementation cycle provide a framework for responsible introduction of AI-based prediction models in healthcare. Additional domain and technology specific research may be necessary and more practical experience with implementing AIPMs is needed to support further guidance.

npj Digital Medicine (2022)5:2 ; https://doi.org/10.1038/s41746-021-00549-7

UMC Utrecht

Step 2: from review to national guideline

npj | Digital Medicine

www.nature.com/npjdigitalmed

REVIEW ARTICLE OPEN

Check for updates

Guidelines and quality criteria for artificial intelligence-based prediction models in healthcare: a scoping review

Anne A. H. de Hond[1,2,3,8✉], Artuur M. Leeuwenberg[4,8✉], Lotty Hooft[4,5], Ilse M. J. Kant[1,2,3], Steven W. J. Nijman[4], Hendrikus J. A. van Os[2,6], Jiska J. Aardoom[6,7], Thomas P. A. Debray[4], Ewoud Schuit[4], Maarten van Smeden[4], Johannes B. Reitsma[4], Ewout W. Steyerberg[2,3], Niels H. Chavannes[6,7] and Karel G. M. Moons[4]

Ministerie van Volksgezondheid, Welzijn en Sport

heid > Documenten >

Guideline for high-quality diagnostic and prognostic applications of AI in healthcare

This guideline provides a description of what the work field considers good professional conduct in the development, testing and implementation of an Artificial Intelligence Prediction Algorithm (AIPA) in the medical sector, including public healthcare.

Download 'Guideline for high-quality diagnostic and prognostic applications of AI in healthcare'

PDF document | 80 pagina's | 713 kB
Publicatie | 28-12-2021

The Dutch version of the guideline AI in healthcare.

www.leidraad-ai.nl

UMC Utrecht

The guideline for diagnostic/prognostic applications

- What the **healthcare field** considers **good professional conduct** in the development, testing and implementation of **AI-based prediction models** in the **medical sector**, including public healthcare.

- **Starting point**: stakeholder opinions and review

- Use of the guideline can (hopefully) **improve quality and lower costs of healthcare**

- Guideline is **not legally binding**

UMC Utrecht

https://www.leidraad-ai.nl/

Welkom bij

Leidraad kwaliteit AI in de zorg

Leidraad voor kwalitatieve diagnostische en prognostische toepassingen van AI in de zorg

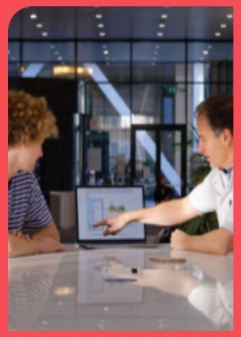
Waardevolle AI

Hoe beoordeel je voorspellende artificiële intelligentie (AI) algoritmen voor gezondheid en zorg op kwaliteit en effectiviteit? De Leidraad kwaliteit AI in de zorg geeft u een overzicht van de belangrijkste eisen en aanbevelingen per fase, van ontwikkeling tot implementatie. Via de online cursus komt u snel meer te weten.

De volledige Leidraad kwaliteit AI in de zorg bekijken? U kunt de pdf hier downloaden (81 pagina's). *The English version of the guideline AI in healthcare can be downloaded here.*

Lees meer berichten over de leidraad op datavoorgezondheid.nl

Vragen/opmerkingen? Deze kunt u delen via de LinkedIn-pagina Leidraad kwaliteit AI in de zorg.



Doe de online cursus

De online leeromgeving helpt u in korte tijd op weg de leidraad beter te begrijpen en toe te kunnen passen. De modules zijn zeer toegankelijk. U kunt direct aan de slag.

Start

Check uw kennis

Wilt u na de online cursus of het lezen van de volledige Leidraad kwaliteit AI in de zorg kijken of u de stof hebt begrepen? Doe dan de korte toets en ontvang een certificaat als bewijs van deelname.

Start

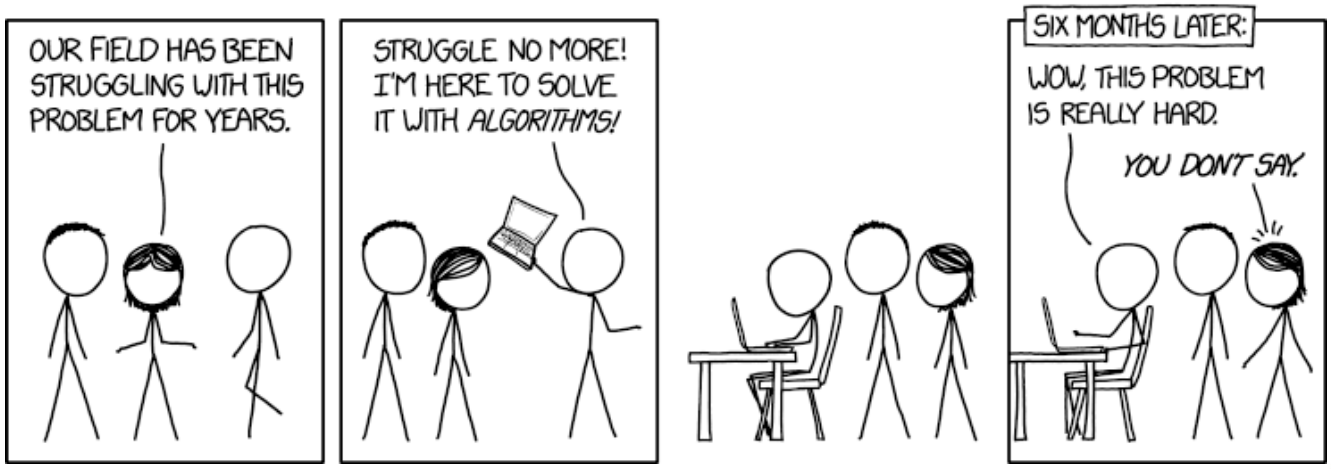
Voor en door het zorgveld

Een brede groep experts, vertegenwoordigers van (koepel)organisaties en betrokkenen hebben gewerkt aan de Leidraad kwaliteit AI in de zorg. In deze video een aantal van hen aan het woord.

UMC Utrecht

Email: M.vanSmeden@umcutrecht.nl

Twitter: @MaartenvSmeden



UMC Utrecht