Visceral Medicine

Review Article

Visc Med 2020;36:443–449
DOI: 10.1159/000511930

Received: May 27, 2020
Accepted: September 30, 2020
Published online: October 12, 2020

Artificial Intelligence in Medicine: Chances and Challenges for Wide Clinical Adoption

Julian Varghese[a]

[a] Institute of Medical Informatics, University of Münster, Münster, Germany

Keywords
Clinical decision support · Artificial intelligence · Precision medicine · Machine learning · Neural networks · Deep learning

Abstract
Background: Artificial intelligence (AI) applications that utilize machine learning are on the rise in clinical research and provide highly promising applications in specific use cases. However, wide clinical adoption remains far off. This review reflects on common barriers and current solution approaches. ***Summary:*** Key challenges are abbreviated as the RISE criteria: Regulatory aspects, Interpretability, interoperability, and the need for Structured data and Evidence. As reoccurring barriers of AI adoption, these concepts are delineated and complemented by points to consider and possible solutions for effective and safe use of AI applications. ***Key Messages:*** There is a fraction of AI applications with proven clinical benefits and regulatory approval. Many new promising systems are the subject of current research but share common issues for wide clinical adoption. The RISE criteria can support preparation for challenges and pitfalls when designing or introducing AI applications into clinical practice.

Introduction

Digital transformation has affected all areas of society. In healthcare, computer systems are not only designed to support documentation and administrative tasks but expected to efficiently assist health professionals in complex clinical situations. The technical term under which these types of systems are classified is clinical decision support systems (CDSS). Mostly referred to as computerized systems, they aim to support clinical decision-making and utilize individual patient characteristics to provide health-related recommendations [1].

Looking back at history, the term artificial intelligence (AI) was coined by John McCarthy [2] and research in AI began in the 1940s and 1950s [3]. Expectations were quickly evolving that computers would mimic any complex human task, but eventually it became clear that those expectations were too high and computational resources too low. As a consequence, 2 major so-called "AI winters" occurred from the 1970s to the 1980s [4] and around the 1990s [5]. Over the last 10 years there has been renewed interest, which is facilitated by the availability of large annotated datasets and modern graphical processing units that train and test deeper neural networks more efficiently [6]. Although AI encompasses all types of algorithms that can mimic intelligent decision-making, nowadays it refers to systems going beyond simple rule-based systems and deals with machine learning approaches, which include deep learning [7]. Regarding clinical research, AI is evaluated with a high diagnostic accuracy, e.g., dermatologist level classification of skin cancer classification [8] or cardiologist level detection of arrhythmia [6]. Even though only a small number of AI systems have been tested in prospective clinical settings and received regulatory approval as medical devices, the number of approvals is increasing. In the domain of ophthalmology, the IDx-DR system detects diabetic retinopathy based on fundus im-

karger@karger.com
www.karger.com/vis

Karger
OPEN ACCESS

Julian Varghese
Institute of Medical Informatics, University of Münster
Albert-Schweitzer-Campus 1/Gebäude A11
DE–48149 Münster (Germany)
julian.varghese@uni-muenster.de

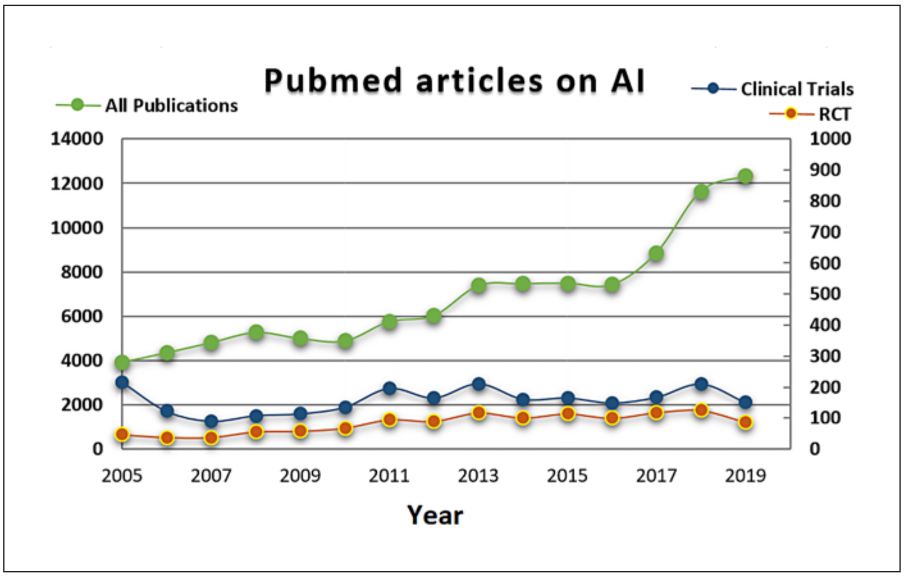

Fig. 1. Articles listed on PubMed matching the MeSH terms "clinical decision support" or "artificial intelligence" in the abstract (green line) and those listed as clinical trials (blue lines) and randomized controlled trials (RCT; yellow line). Timeframe: 2005–2019. The search query is available in the Appendix.

aging [9]. The system showed a high sensitivity and specificity and led to one of the first FDA approvals [10] of an AI system for use by healthcare providers as an autonomous diagnostic system [11]. The AI system Viz.ai Contact was approved by the FDA for CT scans and submits text messages to alert specialists if the system has identified significant vessel blockage [12]. A more commonly known system in daily life is the ECG app in Apple's Series 4 smart watch for atrial fibrillation detection, which has received FDA approval [13] and CE mark with clearance in the European Economic Area [14].

Large-scale implementation and wide clinical adoption are still not established, which raises questions regarding real-world evidence and regulatory or sociotechnical barriers. Focusing on hospital-based settings, this review continues with current challenges and potential solutions for wider clinical adoption. It elaborates on requirements drawn by reoccurring patterns from own CDSS-related research and a previous systematic review on CDSS in hospital care [15]. These requirements are abbreviated as the RISE (Regulatory aspects, Interpretability, Interoperability, Structured Data and Evidence) criteria and will be the recurrent theme in this paper. Within this framework, current AI challenges and possible solutions are presented. To complement important and recent findings, a nonsystematic literature review was conducted on PubMed and Google Scholar from 2010 using the terms "clinical decision support," "precision medicine," "machine learning," "deep learning," and "artificial intelligence." The RISE criteria shall serve as a highly generic framework for understanding inherent challenges and pitfalls of medical AI applications for wide adoption regardless of the specific clinical discipline. This review does not focus on aspects of social or AI-specific usability and acceptability. However, awareness should be raised for communication interfaces, which are currently in a premature phase but could improve doctor-patient and clinic-patient communication [16], e.g., to enable machine-based medical history taking within the scope of a virtual doctor [17].

Regulatory and Evidence

Any AI-driven software or CDSS that aims to have an impact on clinical decision making and is used as such in an existing clinical workflow fulfils the definition of software as a medical device [18, 19]. As such, it needs to be approved for clinical routine. Similar to the approval of medicinal agents, for a new software system to be cleared as medical device it must be validated for secure use and effectivity regarding the intended purpose. Essentially, this approval requires initial evidence, e.g., through a literature review of similar systems, and continues with controlled – ideally multicenter – clinical trials of the actual system as the next level of evidence. According to a systematic review in the field of cancer applications [20], the majority of AI studies have a rather theoretical and retrospective scope and miss translation into this next level of strict clinical trials. Even recently, this finding can be replicated by conducting a simple PubMed search on AI-based articles and filtering for articles associated with clinical trials (Fig. 1). While articles published in the context of AI are rapidly increasing and might mention hope and potential solutions, the number of clinical trials associated with AI remains unchanged. On the one hand, these frequencies do not represent specific evaluation results and could suffer from publication bias [21, 22]. On the other hand, it may very

DOI: 10.1159/000511930

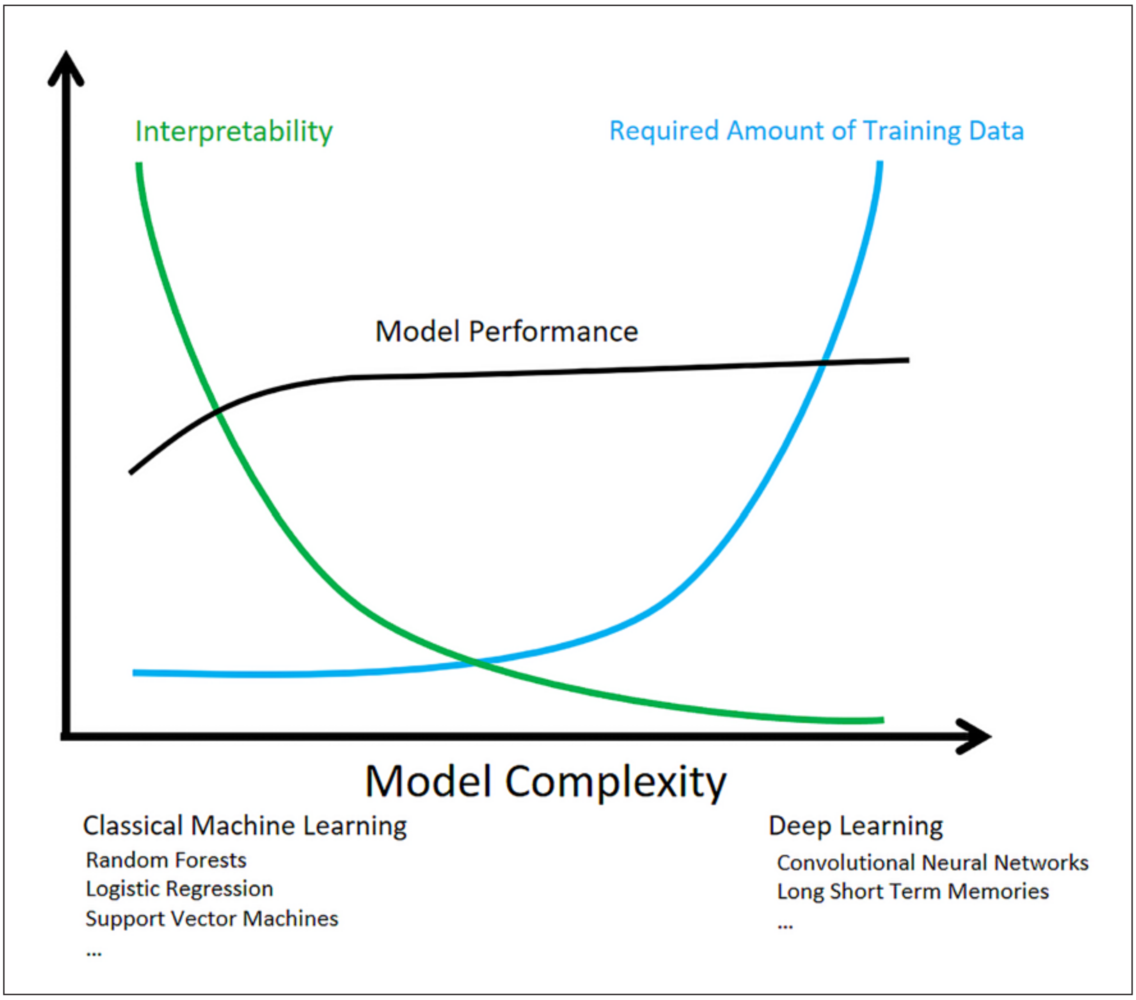

Fig. 2. Machine learning model complexity and possible effects on interpretability, model performance, and the required amount of training data, especially under high-dimensional data as imaging data.

well be that only a small number of promising AI applications show a high potential and therefore only a small number can be financially supported to reach a cost-intensive phase of clinical trials and continuous quality management as a medical device.

CDSS implementation into a clinical workflow requires scientific evidence and regulatory approval. Ideally, one can utilize a system that fulfils both, i.e., high evidence through clinical trials and medical device approval. As this is not the case when introducing new use cases, one should start with evidence based on similar systems in the literature showing positive effects in high-quality studies. After training and testing with experimental or retrospective data, a prospective clinical trial – ideally randomized and at different sites– should be designed in close collaboration with clinical principal investigators, study design experts, statisticians, and technical AI experts. This would also be a key preparatory step to obtain medical device approval, which in turn is necessary for wide clinical adoption.

Interpretability

This review uses the term “interpretability” synonymously with algorithmic explicability. Figure 2 illustrates common observations between a high model complexity and both performance and interpretability. While this figure is a considerable simplification regarding the large number of different machine learning techniques, it provides a basic trade-off and awareness for using or combining different techniques.

It is noteworthy that any well-established machine learning technique could outperform other techniques in specific use cases and data peculiarities, which is often referred to as the “no-free-lunch theorem” [23]. In essence, classical machine learning techniques and expert-based feature engineering provide high interpretability and can outperform deep learning techniques in use cases with low-dimensional data (e.g., analysis of structured questionnaires or a limited set of lab values) or a limited number of training samples.

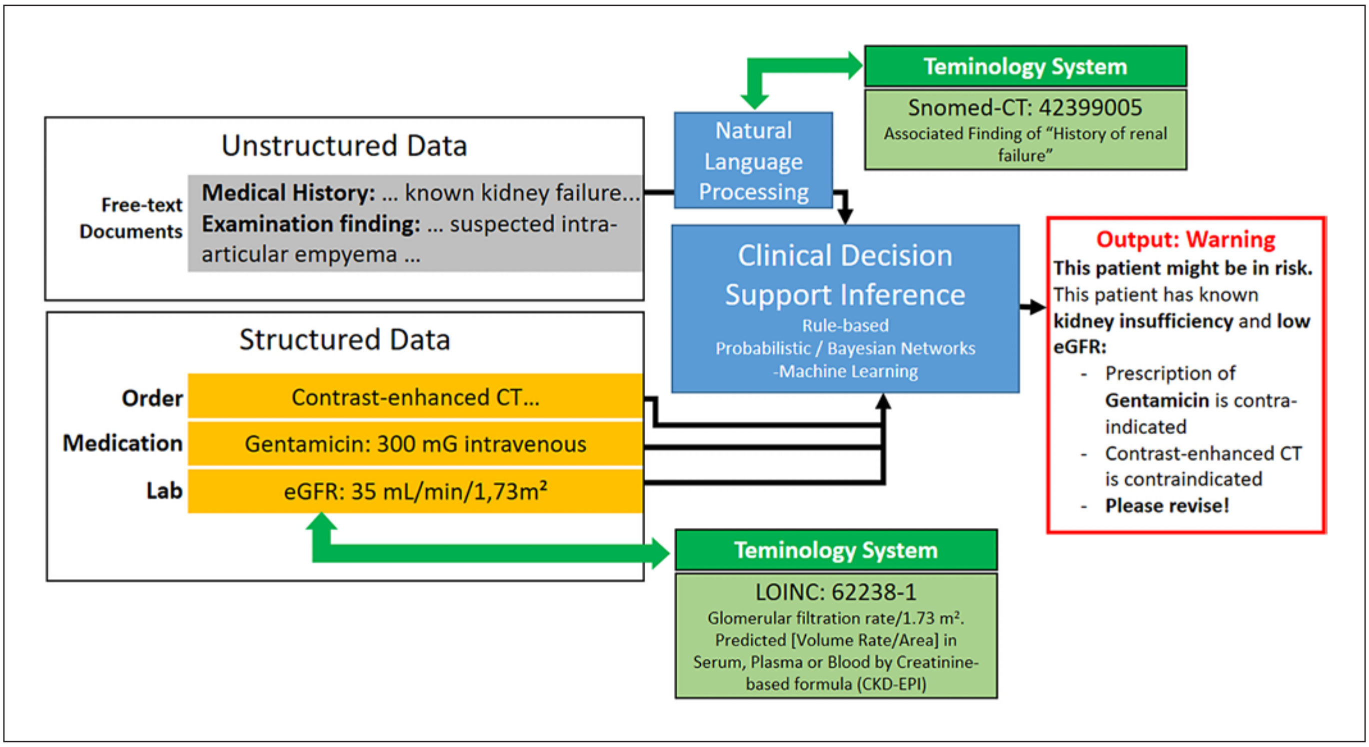
Teminology System
Snomed-CT: 42399005
Associated Finding of "History of renal failure"
Natural Language Processing
Unstructured Data
Free-text Documents
Medical History: ... known kidney failure...
Examination finding: ... suspected intra-articular empyema ...
Clinical Decision Support Inference
Rule-based
Probabilistic / Bayesian Networks
-Machine Learning
Output: Warning
This patient might be in risk.
This patient has known kidney insufficiency and low eGFR:
- Prescription of Gentamicin is contra-indicated
- Contrast-enhanced CT is contraindicated
- Please revise!
Structured Data
Order
Contrast-enhanced CT...
Medication
Gentamicin: 300 mG intravenous
Lab
eGFR: 35 mL/min/1,73m²
Teminology System
LOINC: 62238-1
Glomerular filtration rate/1.73 m². Predicted [Volume Rate/Area] in Serum, Plasma or Blood by Creatinine-based formula (CKD-EPI)

Fig. 3. Clinical decision support in the domain of nephrotoxicity executed on unstructured and structured data. A major challenge is semantic interoperability is finding the necessary information (e.g., eGFR and gentamycin prescription) in different hospital information systems despite the different local identifiers and naming ambiguities in the natural language.

Deep learning architectures are complex neural networks with a series of several hidden layers and can provide exceptionally high performances in image-based analyses, but they come at the cost of a low interpretability of their training and decision making process. This aspect is of particularly high relevance in the critical field of medicine, where AI model outputs are accepted more often when enriched with human-reasonable explanations [24]. This implies a critical discussion between AI experts and health providers in order to also consider low-complexity models with sufficient performance but higher interpretability as, e.g., random forests or gradient boosted decision trees [25], and logistic regression [26] or case-based reasoning with k-nearest neighborhood [27] or probabilistic inferences through bayesian networks [28]. If, however, a model with a high complexity is chosen due to the use case-specific environment and its superior performance, it is recommended to coimplement interpretability-increasing measures to flatten the steep fall of the interpretability of deep learning models (Fig. 2). These measures include visualization of hidden layers, permutation/sensitivity analyses and transformation to more interpretable models, and studying information gain of input features with domain experts [29, 30]. These techniques can enrich the decision of AI models with meaningful text-based explanations tailored to the current case.

Interoperability

Interoperability is the ability of 2 systems, techniques or organizations to work together and communicate efficiently without restrictions. The ISO-13606 differentiates between syntactic interoperability – the ability to communicate with same data or messaging formats and structures – and semantic interoperability, which is additional on top of syntactic interoperability and requires preservation of the relevant meaning of the content being communicated. As a third important level of interoperability, which is explicitly mentioned in the European interoperability framework [31], one should mention organizational interoperability, which requires harmonization of workflows and process management as part of the decision making process. While there exist various adopted standards as for syntactic interoperability in health care, e.g., HL-7 messaging V2 or HL-7 FHIR, and openEHR archetypes, none of these provide fine-grained coding of highly specific medical variables (e.g., lab variables, imaging findings, and symptoms). Although some of the aforementioned standards provide basic semantics, expressive semantic coding can only be incorporated by using well-established reference terminologies, such as LOINC and Snomed-CT [32]. Missing semantic coding is a root cause for long-term data integration or migration approaches [33].

DOI: 10.1159/000511930

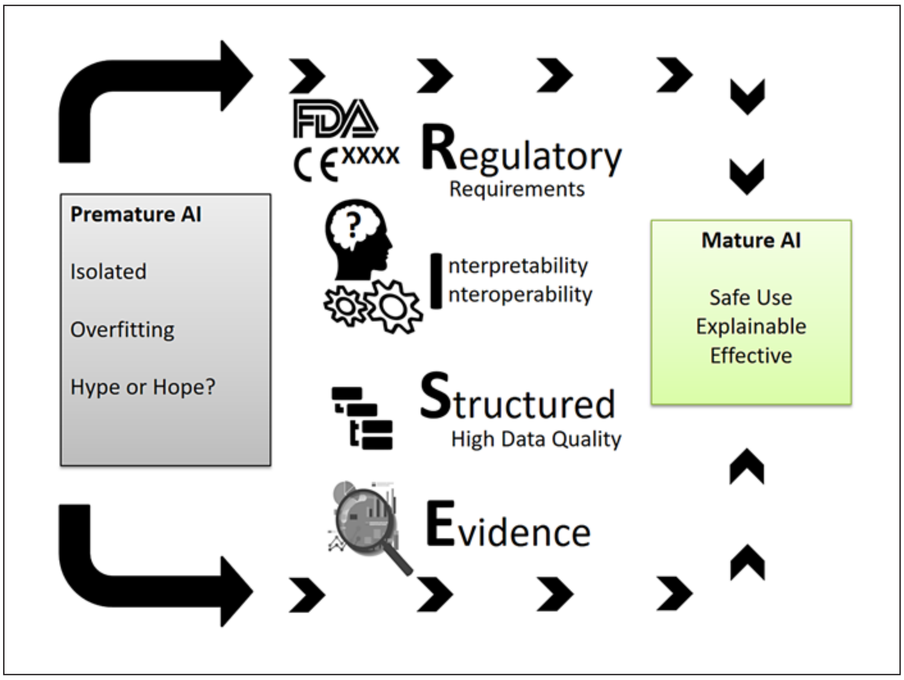

Fig. 4. RISE criteria to facilitate mature and wide adoption of AI applications in medicine.

Figure 3 illustrates a CDSS for nephrotoxicity checks within a hospital setting. Typically, input is available through structured data with clearly defined data elements and unstructured data from free-text documents within the electronic medical record or hospital information system. One of the fundamental problems is to enable the CDSS inference engine to work for different sites. That is, the use-case-relevant input variables, such as "kidney failure," "eGFR," and "gentamicin prescription," are defined with locally different identifiers naming conventions and/or data structures. This problem is well-known as the "Curley brace" problem [34] in the rule-based CDSS community, but it is just another example of semantic interoperability issues. To ease local adaptions and make the inference of this CDSS work in different future sites, relevant medical variables should be annotated with reference terminologies as Snomed-CT or LOINC and mapped to the local variable definitions and permissible values and using the Unified Code for Units of Measure [35]. Though semantic coding is a time-consuming process, it facilitates data analyses and future data interfaces [36].

Structured Data and High Data Quality

Though Figure 3 shows that valuable information could be available in free text, natural language processing (NLP) faces major challenges for accurately extracting relevant medical concepts and their clinical context [37]. Compared to structured data elements, free-text information suffers more from syntactic and semantic ambiguity and therefore it is well known that promising NLP techniques such as IBM Watson that work well for quiz-like free-texts as in Jeopardy may fail in medical applications [38]. Therefore, the input of AI-based technologies should not solely rely on NLP techniques to generate structured data but it should also be founded on primary structured data, such as data generated by laboratory, imaging, medication, or other clinical findings using computerized data entry forms. The majority of the promising use cases from the introduction section were cleared for medical usage and provided a high performance using images, which are a special case of highly structured data. There are 2 further characteristics, besides being structured, that explain why images prove to be well-suited for AI-based pattern recognition. The content contains accurate information on spatial relationships (e.g., between neighboring pixels) and is highly compositional (e.g., pixels constitute edges and edges build shapes, which in turn can build pathological structures). Spatial relations and compositionality are natively modelled by deep learning architectures as multiple layers in convolutional neural networks [39–41], which in turn were inspired by the neurobiology of the visual cortex. In general, medical use cases, in which images represent a significant part of diagnostics, provide a great potential for accurate pattern recognition. As for all data analyses, data quality indicators, such as data availability, completeness, correctness, and plausibility, should be well investigated [42, 43]. In addition, the training and test data should be representative and provide natural diversity within the targeted use case to prevent biased learning [44]. Apart from data quality, data quantity is of particular relevance for the training machine learning algorithms [45] and is addressed in Figure 2. It is likely that there is a lack of training and test data or an uneven balance of classes. Although primary data generation is the ideal way to increase training and test data, it might be too costly if not impossible. Therefore, procedures such as data augmentation [46–48] or transfer learning [49, 50] to leverage knowledge from existing data sources could still improve AI models with a reasonable effort.

Conclusion

AI has been applied in different medical disciplines, but wide clinical adoptions with regulatory approval remain limited to specific use cases. This review provides an overview of use-case-independent key challenges (Fig. 4) and their potential solutions regarding regulatory aspects, interpretability, interoperability, structured data, data quality, and evidence. As innovative and emerging data analyses tools, clinical applications should be prepared by meeting these challenges ideally in the design phase or before clinical introduction. By doing this, effectiveness, interpretability, and safe use are going to be facilitated to enable wide clinical adoption.

Conflict of Interest Statement

The author has no conflict of interests to declare.

Funding Sources

None.

Author Contributions

J.V. designed the concept and wrote the manuscript.

References

1 Kawamoto K, Houlihan CA, Balas EA, Lobach DF. Improving clinical practice using clinical decision support systems: a systematic review of trials to identify features critical to success. BMJ. 2005 Apr;330(7494):765.

2 Cukier K. Ready for robots: how to think about the future of AI. Foreign Aff. 2019;98:192.

3 Ranschaert ER, Morozov S, Algra PR. Artificial intelligence in medical imaging: opportunities, applications and risks. New York: Springer; 2019. 369 p.

4 Crevier DA. The tumultuous history of the search for artificial intelligence. New York: Basic Books; 1993.

5 McCorduck P, Cfe C. Machines who think: A personal inquiry into the history and prospects of artificial intelligence. Boca Raton: CRC Press; 2004. 599 p.

6 Jang H, Park A, Jung K. Neural network implementation using cuda and openmp. In: 2008 digital image computing: Techniques and applications. Washington: IEEE Computer Society; 2008. p. 155–61.

7 Hollis KF, Soualmia LF, Séroussi B. Artificial intelligence in health informatics: Hype or reality? Yearb Med Inform. 2019 Aug;28(1):3–4.

8 Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017 Feb;542(7639):115–8.

9 Abràmoff MD, Lou Y, Erginay A, Clarida W, Amelon R, Folk JC, et al. Improved Automated Detection of Diabetic Retinopathy on a Publicly Available Dataset Through Integration of Deep Learning. Invest Ophthalmol Vis Sci. 2016 Oct;57(13):5200–6.

10 Padhy SK, Takkar B, Chawla R, Kumar A. Artificial intelligence in diabetic retinopathy: A natural step to the future. Indian J Ophthalmol. 2019 Jul;67(7):1004–9.

11 Abràmoff MD, Lavin PT, Birch M, Shah N, Folk JC. Pivotal trial of an autonomous AI-based diagnostic system for detection of diabetic retinopathy in primary care offices. NPJ Digit Med. 2018 Aug;1(1):39.

12 Bluemke DA. Radiology in 2018: Are You Working with AI or Being Replaced by AI? Radiology. 2018 May;287(2):365–6.

13 Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019 Jan;25(1):44–56.

14 ECG app and irregular rhythm notification on Apple Watch available today across Europe and Hong Kong – Apple [Internet]. [cited 2020 Mar 13]. Available from: https://www.apple.com/newsroom/2019/03/ecg-app-and-irregular-rhythm-notification-on-apple-watch-available-today-across-europe-and-hong-kong/.

15 Varghese J, Kleine M, Gessner SI, Sandmann S, Dugas M. Effects of computerized decision support system implementations on patient outcomes in inpatient care. J Am Med Inform Assoc. 2018 May 1;25(5):593–602.

16 Nadarzynski T, Miles O, Cowie A, Ridge D. Acceptability of artificial intelligence (AI)-led chatbot services in healthcare: A mixed-methods study. Digit Health. 2019 Aug 21;5:2055207619871808.

17 Spänig S, Emberger-Klein A, Sowa JP, Canbay A, Menrad K, Heider D. The virtual doctor: An interactive clinical-decision-support system based on deep learning for non-invasive prediction of diabetes. Artif Intell Med. 2019 Sep;100:101706.

18 Becker K, Lipprandt M, Röhrig R, Neumuth T. Digital health: Software as a medical device in focus of the medical device regulation (MDR). Digit Health. 2019;61(5–6):211–218.

19 Yaeger KA, Martini M, Yaniv G, Oermann EK, Costa AB. United States regulatory approval of medical devices and software applications enhanced by artificial intelligence. Health Policy Technol. 2019 Jun;8(2):192–7.

20 Lisboa PJ, Taktak AF. The use of artificial neural networks in decision support in cancer: a systematic review. Neural Netw. 2006 May;19(4):408–15.

21 Easterbrook PJ, Berlin JA, Gopalan R, Matthews DR. Publication bias in clinical research. Lancet. 1991 Apr;337(8746):867–72.

22 Bright TJ, Wong A, Dhurjati R, Bristow E, Bastian L, Coeytaux RR, et al. Effect of clinical decision-support systems: a systematic review. Ann Intern Med. 2012 Jul;157(1):29–43.

23 Wolpert DH. The supervised learning no-free-lunch theorems: Soft computing and industry. New York: Springer; 2002. p. 25–42.

24 Vellido A, Martin-Guerrero JD, Lisboa PJG. Making machine learning models interpretable. Proceedings of the European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning; 2012 Apr 25–27; Bruges, Belgium.

25 Chen T, Guestrin C. Xgboost: A scalable tree boosting system. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2016 Aug; San Francisco, USA.

26 Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002 Oct-Dec;35(5-6):352–9.

DOI: 10.1159/000511930

27 Peterson LE. K-nearest neighbor. Scholarpedia. 2009;4(2):1883.

28 Lucas P. Expert knowledge and its role in learning bayesian networks in medicine: an appraisal. In: Quaglini S, Barahona P, Andreassen S, editors. Artificial intelligence in medicine. Berlin: Springer; 2001. p. 156–66.

29 Bastani O, Kim C, Bastani H. Interpretability via Model Extraction. 2017;arXiv preprint (arXiv:1706.09773).

30 Samek W, Wiegand T, Müller KR. Explainable Artificial Intelligence. 2017;arXiv preprint (arXiv:1708.08296).

31 Vernadat FB. Technical, semantic and organizational issues of enterprise interoperability and networking. Annu Rev Contr. 2010; 34(1):139–44.

32 Bodenreider O, Cornet R, Vreeman DJ. Recent developments in clinical terminologies: SNOMED CT, LOINC, and RxNorm. Yearb Med Inform. 2018 Aug;27(1):129–39.

33 Dugas M. Missing semantic annotation in databases. The root cause for data integration and migration problems in information systems. Methods Inf Med. 2014;53(6):516–7.

34 Samwald M, Fehre K, de Bruin J, Adlassnig KP. The Arden Syntax standard for clinical decision support: experiences and directions. J Biomed Inform. 2012 Aug;45(4):711–8.

35 Gansel X, Mary M, van Belkum A. Semantic data interoperability, digital medicine, and e-health in infectious disease management: a review. Eur J Clin Microbiol Infect Dis. 2019 Jun;38(6):1023–34.

36 Varghese J, Sandmann S, Dugas M. Web-based information infrastructure increases the interrater reliability of medical coders: quasi-experimental study. J Med Internet Res. 2018 Oct;20(10):e274.

37 Becker M, Böckmann B. Extraction of UMLS® concepts using apache cTAKES™ for German language. Stud Health Technol Inform. 2016; 223:71–6.

38 Strickland E. IBM Watson, heal thyself: how IBM overpromised and underdelivered on AI health care. IEEE Spectr. 2019;56(4):24–31.

39 LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015 May;521(7553):436–44.

40 Poggio T, Mhaskar H, Rosasco L, Miranda B, Liao Q. Why and when can deep-but not shallow-networks avoid the curse of dimensionality: a review. Int J Autom Comput. 2017; 14(5):503–19.

41 Lu L, Zheng Y, Carneiro G, Yang L. Deep learning and convolutional neural networks for medical image computing. New York: Springer; 2017.

42 Arts DG, De Keizer NF, Scheffer GJ. Defining and improving data quality in medical registries: a literature review, case study, and generic framework. J Am Med Inform Assoc. 2002 Nov-Dec;9(6):600–11.

43 Weiskopf NG, Weng C. Methods and dimensions of electronic health record data quality assessment: enabling reuse for clinical research. J Am Med Inform Assoc. 2013 Jan; 20(1):144–51.

44 Le Berre C, Sandborn WJ, Aridhi S, Devignes MD, Fournier L, Smaïl-Tabbone M, et al. Application of artificial intelligence to gastroenterology and hepatology. Gastroenterology. 2020 Jan;158(1):76–94.e2.

45 Goodfellow I, Bengio Y, Courville A. Deep learning. Cambridge: MIT Press; 2016.

46 Mikołajczyk A, Grochowski M. Data augmentation for improving deep learning in image classification problem. In: 2018 International Interdisciplinary PhD Workshop (IIPhDW). 2018. p. 117–22.

47 Perez L, Wang J. The Effectiveness of Data Augmentation in Image Classification using Deep Learning. 2017 Dec 13;arXiv preprint(arXiv:1712.04621).

48 Lu X, Zheng B, Velivelli A, Zhai C. Enhancing text categorization with semantic-enriched representation and training data augmentation. J Am Med Inform Assoc. 2006 Sep-Oct; 13(5):526–35.

49 Dai W, Yang Q, Xue GR, Yu Y. Boosting for transfer learning. New York: ACM; 2007.

50 Torrey L, Shavlik J. Transfer learning: Handbook of research on machine learning applications and trends – algorithms, methods, and techniques. Hershy: IGI Global; 2010. p. 242–64.