Kelly et al. BMC Medicine (2019) 17:195
https://doi.org/10.1186/s12916-019-1426-2

BMC Medicine

OPINION Open Access

Key challenges for delivering clinical impact with artificial intelligence

Christopher J. Kelly[1*], Alan Karthikesalingam[1], Mustafa Suleyman[2], Greg Corrado[3] and Dominic King[1]

Abstract

Background: Artificial intelligence (AI) research in healthcare is accelerating rapidly, with potential applications being demonstrated across various domains of medicine. However, there are currently limited examples of such techniques being successfully deployed into clinical practice. This article explores the main challenges and limitations of AI in healthcare, and considers the steps required to translate these potentially transformative technologies from research to clinical practice.

Main body: Key challenges for the translation of AI systems in healthcare include those intrinsic to the science of machine learning, logistical difficulties in implementation, and consideration of the barriers to adoption as well as of the necessary sociocultural or pathway changes. Robust peer-reviewed clinical evaluation as part of randomised controlled trials should be viewed as the gold standard for evidence generation, but conducting these in practice may not always be appropriate or feasible. Performance metrics should aim to capture real clinical applicability and be understandable to intended users. Regulation that balances the pace of innovation with the potential for harm, alongside thoughtful post-market surveillance, is required to ensure that patients are not exposed to dangerous interventions nor deprived of access to beneficial innovations. Mechanisms to enable direct comparisons of AI systems must be developed, including the use of independent, local and representative test sets. Developers of AI algorithms must be vigilant to potential dangers, including dataset shift, accidental fitting of confounders, unintended discriminatory bias, the challenges of generalisation to new populations, and the unintended negative consequences of new algorithms on health outcomes.

Conclusion: The safe and timely translation of AI research into clinically validated and appropriately regulated systems that can benefit everyone is challenging. Robust clinical evaluation, using metrics that are intuitive to clinicians and ideally go beyond measures of technical accuracy to include quality of care and patient outcomes, is essential. Further work is required (1) to identify themes of algorithmic bias and unfairness while developing mitigations to address these, (2) to reduce brittleness and improve generalisability, and (3) to develop methods for improved interpretability of machine learning predictions. If these goals can be achieved, the benefits for patients are likely to be transformational.

Keywords: Artificial intelligence, Machine learning, Algorithms, Translation, Evaluation, Regulation

Background

The exciting promise of artificial intelligence (AI) in healthcare has been widely reported, with potential applications across many different domains of medicine [1, 2]. This promise has been welcomed as healthcare systems globally struggle to deliver the 'quadruple aim', namely improving experience of care, improving the health of populations, reducing per capita costs of healthcare [3], and improving the work life of healthcare providers [4].

Nevertheless, the potential of AI in healthcare has not been realised to date, with limited existing reports of the clinical and cost benefits that have arisen from real-world use of AI algorithms in clinical practice. This article explores the main challenges and limitations of AI in healthcare, and considers the steps required to translate these potentially transformative technologies from research to clinical practice.

The potential of artificial intelligence in healthcare

A rapidly accelerating number of academic research studies have demonstrated the various applications of AI in healthcare, including algorithms for interpreting chest

* Correspondence: cjkelly@google.com
[1]Google Health, London, UK
Full list of author information is available at the end of the article

radiographs [5–9], detecting cancer in mammograms [10, 11], analysing computer tomography scans [12–15], identifying brain tumours on magnetic resonance images [16], and predicting development of Alzheimer's disease from positron emission tomography [17]. Applications have also been shown in pathology [18], identifying cancerous skin lesions [19–22], interpreting retinal imaging [23, 24], detecting arrhythmias [25, 26], and even identifying hyperkalaemia from electrocardiograms [27]. Furthermore, AI has aided in polyp detection from colonoscopy [28], improving genomics interpretation [29], identifying genetic conditions from facial appearance [30], and assessing embryo quality to maximise the success of in vitro fertilisation [31].

Analysis of the immense volume of data collected from electronic health records (EHRs) offers promise in extracting clinically relevant information and making diagnostic evaluations [32] as well as in providing real-time risk scores for transfer to intensive care [33], predicting in-hospital mortality, readmission risk, prolonged length of stay and discharge diagnoses [34], predicting future deterioration, including acute kidney injury [35], improving decision-making strategies, including weaning of mechanical ventilation [36] and management of sepsis [37], and learning treatment policies from observational data [38]. Proof-of-concept studies have aimed to improve the clinical workflow, including automatic extraction of semantic information from transcripts [39], recognising speech in doctor–patient conversations [40], predicting risk of failure to attend hospital appointments [41], and even summarising doctor–patient consultations [42].

Given this impressive array of studies, it is perhaps surprising that real world deployments of machine learning algorithms in clinical practice are rare. Despite this, we believe that AI will have a positive impact on many aspects of medicine. AI systems have the potential to reduce unwarranted variation in clinical practice, improve efficiency and prevent avoidable medical errors that will affect almost every patient during their lifetime [43]. By providing novel tools to support patients and augment healthcare staff, AI could enable better care delivered closer to the patient in the community. AI tools could assist patients in playing a greater role in managing their own health, primary care physicians by allowing them to confidently manage a greater range of complex disease, and specialists by offering superhuman diagnostic performance and disease management. Finally, through the detection of novel signals of disease that clinicians are unable to perceive, AI can extract novel insights from existing data. Examples include the identification of novel predictive features for breast cancer prognosis using stromal cells (rather than the cancer cells themselves) [44], predicting cardiovascular risk factors and sex from a fundus photograph [45], inferring blood flow in coronary arteries from cardiac computed tomography [46], detecting individuals with atrial fibrillation from ECG acquired during normal sinus rhythm [26], and using retinal imaging to assist an earlier diagnosis of dementia [47].

The challenge of translation to clinical practice

Retrospective versus prospective studies

While existing studies have encompassed very large numbers of patients with extensive benchmarking against expert performance, the vast majority of studies have been retrospective, meaning that they use historically labelled data to train and test algorithms. Only through prospective studies will we begin to understand the true utility of AI systems, as performance is likely to be worse when encountering real-world data that differ from that encountered in algorithm training. The limited number of prospective studies to date include diabetic retinopathy grading [48–50], detection of breast cancer metastases in sentinel lymph node biopsies [51, 52], wrist fracture detection [53], colonic polyp detection [28, 54], and detection of congenital cataracts [55]. Consumer technology is enabling enormous prospective studies, in relation to historical standards, through the use of wearables; for example, there is an ongoing study to detect atrial fibrillation in 419,093 consenting Apple watch owners [56].

Peer-reviewed randomised controlled trials as an evidence gold standard

As is common in the machine learning community, many studies have been published on preprint servers only and are not submitted to peer-reviewed journals. Peer-reviewed evidence will be important for the trust and adoption of AI within the wider medical community. There are very few randomised controlled trials (RCTs) of AI systems to date; these include an algorithm to detect childhood cataracts with promising performance in a small prospective study [55] but less accurate performance compared to senior clinicians in a diagnostic RCT [57]; a single-blind RCT that showed a significantly reduced blind-spot rate in esophagogastroduodenoscopy [58]; an open, non-blinded randomised trial of an automatic polyp detection algorithm for diagnostic colonoscopy demonstrating a significant increase in detection of diminutive adenomas and hyperplastic polyps [59]; a simulated prospective, double-blind RCT of an algorithm to detect acute neurologic events [60]; and an unmasked RCT of a system to provide automated interpretation of cardiotocographs in labour that found no improvement in clinical outcomes for mothers or babies [61]. The final study is a cautionary example of how higher accuracy enabled by AI systems does not necessarily result in better patient outcomes [61]. Future studies should aim to use clinical outcomes as trial endpoints to demonstrate

longer-term benefit, while recognising that algorithms are likely to result in changes of the sociocultural context or care pathways; this may necessitate more sophisticated approaches to evaluation [62].

High quality reporting of machine learning studies is critical. Only with full and clear reporting of information on all aspects of a diagnosis or prognosis model can risk of bias and potential usefulness of prediction models be adequately assessed. Machine learning studies should aim to follow best practice recommendations, such as the Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD), designed to assist the reporting of studies that develop, validate or update a prediction model for either diagnostic or prognostic purposes [63]. In addition, a new version of the TRIPOD statement that is specific to machine learning prediction algorithms (TRIPOD-ML) is in development and will focus on the introduction of machine learning prediction algorithms, establishing methodological and reporting standards for machine learning studies in healthcare [64].

Metrics often do not reflect clinical applicability

The term 'AI chasm' has been coined to reflect the fact that accuracy does not necessarily represent clinical efficacy [65]. Despite its universal use in machine learning studies, area under the curve of a receiver operating characteristic curve is not necessarily the best metric to represent clinical applicability [66] and is not easily understandable by many clinicians. As well as reporting sensitivity and specificity at a selected model operating point (required to turn the continuous model output into discrete decision categories), papers should include information about positive and negative predictive values. As no single measure captures all the desirable properties of a model, several measures are typically reported to summarise its performance. However, none of these measures ultimately reflect what is most important to patients, namely whether the use of the model results in a beneficial change in patient care [67].

Clinicians need to be able to understand how the proposed algorithms could improve patient care within a relatable workflow, yet most papers do not attempt to present such information; potential approaches to this have been suggested, including decision curve analysis, which aims to quantify the net benefit of using a model to guide subsequent actions [68]. To improve understanding, medical students and practising clinicians should be provided with an easily accessible AI curriculum to enable them to critically appraise, adopt and use AI tools safely in their practice.

Difficulty comparing different algorithms

The comparison of algorithms across studies in an objective manner is challenging due to each study's performance being reported using variable methodologies on different populations with different sample distributions and characteristics. To make fair comparisons, algorithms need to be subjected to comparison on the same independent test set that is representative of the target population, using the same performance metrics. Without this, clinicians will have difficulty in determining which algorithm is likely to perform best for their patients.

The curation of independent local test sets by each healthcare provider could be used to fairly compare the performance of the various available algorithms in a representative sample of their population. Such independent test sets should be constructed using an unenriched representative sample along with data that are explicitly not available to train algorithms. A supplementary local training dataset could be provided to allow fine tuning of algorithms prior to formal testing.

For researchers, comparison will become easier with the increasing availability of large, open datasets, allowing studies to benchmark their performance in a consistent manner.

Challenges related to machine learning science

AI algorithms have the potential to suffer from a host of shortcomings, including inapplicability outside of the training domain, bias and brittleness (tendency to be easily fooled) [69]. Important factors for consideration include dataset shift, accidentally fitting confounders rather than true signal, propagating unintentional biases in clinical practice, providing algorithms with interpretability, developing reliable measures of model confidence, and the challenge of generalisation to different populations.

Dataset shift

Particularly important for EHR algorithms, it is easy to ignore the fact that all input data are generated within a non-stationary environment with shifting patient populations, where clinical and operational practices evolve over time [70]. The introduction of a new predictive algorithm may cause changes in practice, resulting in a new distribution compared to that used to train the algorithm. Therefore, methods to identify drift and update models in response to deteriorating performance are critical. Mitigations to manage this effect include careful quantification of performance over time to proactively identify problems, alongside the likely requirement for periodical retraining. Data-driven testing procedures have been suggested to recommend the most appropriate updating method, from simple recalibration to full model retraining, in order to maintain performance over time [71].

Accidentally fitting confounders versus true signal

Machine learning algorithms will use whatever signals are available to achieve the best possible performance in the dataset used. This may include the exploitation of unknown confounders that may not be reliable, impairing the algorithm's ability to generalise to new datasets. For instance, in one classic example, a machine learning model did not learn the intrinsic difference between dogs and wolves, but instead learned that wolves are usually pictured standing on snow, while dogs usually appear on grass [72]. There are similar concerns in healthcare. In one study, an algorithm was more likely to classify a skin lesion as malignant if an image had a ruler in it because the presence of a ruler correlated with an increased likelihood of a cancerous lesion [19]. The presence of surgical skin markings have also been shown to falsely increase a deep learning model's melanoma probability scores and hence false positive rate [73]. In another study, hip fracture detection was found to be aided by confounders, including the scanner model and scans marked 'urgent' [74]. Another algorithm for detection of pneumonia on chest x-rays was able to accurately identify hospital equipment and department, learning an association between a portable x-ray machine and pneumonia [75]. Ongoing work is required to understand the specific features being learned by neural networks and will be critical for generalisation across multiple healthcare settings.

Challenges in generalisation to new populations and settings

The majority of AI systems are far from achieving reliable generalisability, let alone clinical applicability, for most types of medical data. A brittle model may have blind spots that can produce particularly bad decisions. Generalisation can be hard due to technical differences between sites (including differences in equipment, coding definitions, EHR systems, and laboratory equipment and assays) as well as variations in local clinical and administrative practices.

To overcome these issues, it is likely that a degree of site-specific training will be required to adapt an existing system for a new population, particularly for complex tasks like EHR predictions. Methods to detect out-of-distribution inputs and provide a reliable measure of model confidence will be important to prevent clinical decisions being made on inaccurate model outputs. For simpler tasks, including medical image classification, this problem may be less crucial and overcome by the curation of large, heterogenous, multi-centre datasets [14]. Generalisation of model operating points may also prove challenging across new populations, as illustrated in a recent study to detect abnormal chest radiographs, where specificity at a fixed operating point varied widely, from 0.566 to 1.000, across five independent datasets [5].

Proper assessment of real-world clinical performance and generalisation requires appropriately designed external validation involving testing of an AI system using adequately sized datasets collected from institutions other than those that provided the data for model training. This will ensure that all relevant variations in patient demographics and disease states of target patients in real-world clinical settings are adequately represented in the system where it will be applied [76]. This practice is currently rare in the literature and is of critical concern. A recent systematic review of studies that evaluated AI algorithms for the diagnostic analysis of medical imaging found that only 6% of 516 eligible published studies performed external validation [77].

Algorithmic bias

Intertwined with the issue of generalisability is that of discriminatory bias. Blind spots in machine learning can reflect the worst societal biases, with a risk of unintended or unknown accuracies in minority subgroups, and there is fear over the potential for amplifying biases present in the historical data [78]. Studies indicate that, in some current contexts, the downsides of AI systems disproportionately affect groups that are already disadvantaged by factors such as race, gender and socioeconomic background [79]. In medicine, examples include hospital mortality prediction algorithms with varying accuracy by ethnicity [80] and algorithms that can classify images of benign and malignant moles with accuracy similar to that of board-certified dermatologists [19, 81], but with underperformance on images of lesions in skin of colour due to training on open datasets of predominantly fair skinned patients. The latter is particularly concerning as patients with skin of colour already present with more advanced dermatological diseases and have lower survival rates than those with fair skin [82].

Algorithmic unfairness can be distilled into three components, namely (1) model bias (i.e. models selected to best represent the majority and not necessarily underrepresented groups), (2) model variance (due to inadequate data from minorities), and (3) outcome noise (the effect of a set of unobserved variables that potentially interacts with model predictions, avoidable by identifying subpopulations to measure additional variables) [80]. A greater awareness of these issues and empowering clinicians to participate critically in system design and development will help guide researchers to ensure that the correct steps are taken to quantify bias before deploying models. Algorithms should be designed with the global community in mind, and clinical validation should be performed using a representative population of the intended deployment population. Careful performance analysis by population subgroups should be performed, including age, ethnicity, sex, sociodemographic stratum

and location. Analysis to understand the impact of a new algorithm is particularly important, i.e. if the spectrum of disease detected using the AI system differs from current clinical practice, then the benefits and harms of detecting this different spectrum of disease must be evaluated. In mammography, this might be the detection of less severe ductal carcinoma in situ, potentially resulting in increased treatment with little benefit in outcomes. Prospective pilots within healthcare systems should be undertaken to understand the product characteristics and identify potential pitfalls in practical deployment.

Susceptibility to adversarial attack or manipulation

Algorithms have been shown to be susceptible to risk of adversarial attack. Although somewhat theoretical at present, an adversarial attack describes an otherwise-effective model that is susceptible to manipulation by inputs explicitly designed to fool them. For example, in one study, images of benign moles were misdiagnosed as malignant by adding adversarial noise or even just rotation [83].

Logistical difficulties in implementing AI systems

Many of the current challenges in translating AI algorithms to clinical practice are related to the fact that most healthcare data are not readily available for machine learning. Data are often siloed in a multitude of medical imaging archival systems, pathology systems, EHRs, electronic prescribing tools and insurance databases, which are very difficult to bring together. Adoption of unified data formats, such as Fast Healthcare Interoperability Resources [84], offer the potential for better aggregation of data, although improved interoperability does not necessarily fix the problem of inconsistent semantic coding in EHR data [85].

Achieving robust regulation and rigorous quality control

A fundamental component to achieving safe and effective deployment of AI algorithms is the development of the necessary regulatory frameworks. This poses a unique challenge given the current pace of innovation, significant risks involved and the potentially fluid nature of machine learning models. Proactive regulation will give confidence to clinicians and healthcare systems. Recent U.S. Food and Drug Administration guidance has begun developing a modern regulatory framework to make sure that safe and effective artificial intelligence devices can efficiently progress to patients [86].

It is also important to consider the regulatory impact of improvements and upgrades that providers of AI products are likely to develop throughout the life of the product. Some AI systems will be designed to improve over time, representing a challenge to traditional evaluation processes. Where AI learning is continuous, periodic system-wide updates following a full evaluation of clinical significance would be preferred, compared to continuous updates which may result in drift. The development of ongoing performance monitoring guidelines to continually calibrate models using human feedback will support the identification of performance deficits over time.

Human barriers to AI adoption in healthcare

Even with a highly effective algorithm that overcomes all of the above challenges, human barriers to adoption are substantial. In order to ensure that this technology can reach and benefit patients, it will be important to maintain a focus on clinical applicability and patient outcomes, advance methods for algorithmic interpretability, and achieve a better understanding of human–computer interactions.

Algorithmic interpretability is at an early stage but rapidly advancing

While AI approaches in medicine have yielded some impressive practical successes to date, their effectiveness is limited by their inability to 'explain' their decision-making in an understandable way [87]. Even if we understand the underlying mathematical principles of such models, it is difficult and often impossible to interrogate the inner workings of models to understand how and why it made a certain decision. This is potentially problematic for medical applications, where there is particular demand for approaches that are not only well-performing, but also trustworthy, transparent, interpretable and explainable [88].

Healthcare offers one of the strongest arguments in favour of explainability [88, 89]. Given the combination of the devastating consequences of unacceptable results, the high risk of unquantified bias that is difficult to identify a priori, and the recognised potential for models to use inappropriate confounding variables, explainability enables system verification. This improves experts' ability to recognise system errors, detect results based upon inappropriate reasoning, and identify the work required to remove bias. In addition, AI systems are trained using large numbers of examples and may detect patterns in data that are not accessible to humans. Interpretable systems may allow humans to extract this distilled knowledge in order to acquire new scientific insights. Finally, recent European Union General Data Protection Regulation legislation mandates a 'right to explanation' for algorithmically generated user-level predictions that have the potential to 'significantly affect' users; this suggests that there must be a possibility to make results re-traceable on demand [88].

At present, a trade-off exists between performance and explainability. The best performing models (e.g. deep learning) are often the least explainable, whereas models with poorer performance (e.g. linear regression, decision trees) are the most explainable. A key current limitation of deep learning models is that they have no explicit declarative knowledge representation, leading to considerable difficulty in generating the required explanation structures [90]. Machine learning methods that build upon a long history of research in traditional symbolic AI techniques to allow for encoding of semantics of data and the use of ontologies to guide the learning process may permit human experts to understand and retrace decision processes more effectively [91, 92]. One recent approach replaced end-to-end classification with a two-stage architecture comprising segmentation and classification, allowing the clinician to interrogate the segmentation map to understand the basis of the subsequent classification [24].

If 'black box' algorithms are to be used in healthcare, they need to be used with knowledge, judgement and responsibility. In the meantime, research into explainable AI and evaluation of interpretability is occurring at a rapid pace [93]. Explainable AI approaches are likely to facilitate faster adoption of AI systems into the clinical healthcare setting, and will help foster vital transparency and trust with their users.

Developing a better understanding of interaction between human and algorithm

We have a limited but growing understanding of how humans are affected by algorithms in clinical practice. Following the U. S. Food and Drug Administration approval of computer-aided diagnosis for mammography in the late 1990s, computer-aided diagnosis was found to significantly increase recall rate without improving outcomes [94]. Excessive warnings and alerts are known to result in alert fatigue [94, 95]. It has also been shown that humans assisted by AI performed better than either alone in a study of diabetic retinopathy screening [96, 97]. Techniques to more meaningfully represent medical knowledge, provide explanation and facilitate improved interaction with clinicians will only improve this performance further. We need to continue gaining a better understanding of the complex and evolving relationship between clinicians and human-centred AI tools in the live clinical environment [98].

Conclusion

Recent advances in artificial intelligence present an exciting opportunity to improve healthcare. However, the translation of research techniques to effective clinical deployment presents a new frontier for clinical and machine learning research. Robust, prospective clinical evaluation will be essential to ensure that AI systems are safe and effective, using clinically applicable performance metrics that go beyond measures of technical accuracy to include how AI affects the quality of care, the variability of healthcare professionals, the efficiency and productivity of clinical practice and, most importantly, patient outcomes. Independent datasets that are representative of future target populations should be curated to enable the comparison of different algorithms, while carefully evaluating for signs of potential bias and fitting to unintended confounders. Developers of AI tools must be cognisant of the potential unintended consequences of their algorithms and ensure that algorithms are designed with the global community in mind. Further work to improve the interpretability of algorithms and to understand human–algorithm interactions will be essential to their future adoption and safety supported by the development of thoughtful regulatory frameworks.

Abbreviations

AI: artificial intelligence; EHRs: electronic health records; RCT: randomised controlled trial; TRIPOD: Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis

Acknowledgements

Not applicable.

Authors' contributions

CK wrote the first draft. All authors contributed to the final manuscript. All authors read and approved the final manuscript.

Funding

Google LLC.

Availability of data and materials

Not applicable.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

All authors are employed by Google LLC.

Author details

[1]Google Health, London, UK. [2]DeepMind, London, UK. [3]Google Health, California, USA.

Received: 31 May 2019 Accepted: 16 September 2019
Published online: 29 October 2019

References

1. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44–56.
2. Esteva A, Robicquet A, Ramsundar B, Kuleshov V, DePristo M, Chou K, et al. A guide to deep learning in healthcare. Nat Med. 2019;25:24–9.
3. Berwick DM, Nolan TW, Whittington J. The triple aim: care, health, and cost. Health Aff. 2008;27:759–69. https://doi.org/10.1377/hlthaff.27.3.759.
4. Bodenheimer T, Sinsky C. From triple to quadruple aim: care of the patient requires care of the provider. Ann Fam Med. 2014;12:573–6.
5. Hwang EJ, Park S, Jin K-N, Kim JI, Choi SY, Lee JH, et al. Development and validation of a deep learning-based automated detection algorithm

for major thoracic diseases on chest radiographs. JAMA Netw Open. 2019;2:e191095.
6. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM. ChestX-Ray8: Hospital-Scale Chest X-Ray Database and Benchmarks on Weakly-Supervised Classification and Localization of Common Thorax Diseases. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2017. https://doi.org/10.1109/cvpr.2017.369.
7. Li Z, Wang C, Han M, Xue Y, Wei W, Li L-J, et al. Thoracic Disease Identification and Localization with Limited Supervision. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. p. 2018. https://doi.org/10.1109/cvpr.2018.00865.
8. Singh R, Kalra MK, Nitiwarangkul C, Patti JA, Homayounieh F, Padole A, et al. Deep learning in chest radiography: detection of findings and presence of change. PLoS One. 2018;13:e0204155. https://doi.org/10.1371/journal.pone.0204155.
9. Nam JG, Park S, Hwang EJ, Lee JH, Jin K-N, Lim KY, et al. Development and validation of deep learning–based automatic detection algorithm for malignant pulmonary nodules on chest radiographs. Radiology. 2019;290:218–28. https://doi.org/10.1148/radiol.2018180237.
10. Geras KJ, Wolfson S, Shen Y, Wu N, Gene Kim S, Kim E, et al. High-resolution breast cancer screening with multi-view deep convolutional neural networks. arXiv. 2017; https://arxiv.org/abs/1703.07047. Accessed 1 May 2019.
11. Wu N, Phang J, Park J, Shen Y, Huang Z, Zorin M, et al. Deep neural networks improve radiologists' performance in breast cancer screening. arXiv. 2019; https://arxiv.org/abs/1903.08297. Accessed 1 May 2019.
12. Hua K-L, Hsu C-H, Hidayati SC, Cheng W-H, Chen Y-J. Computer-aided classification of lung nodules on computed tomography images via deep learning technique. Onco Targets Ther. 2015;8:2015–22.
13. Yasaka K, Akai H, Abe O, Kiryu S. Deep learning with convolutional neural network for differentiation of liver masses at dynamic contrast-enhanced CT: a preliminary study. Radiology. 2018;286:887–96. https://doi.org/10.1148/radiol.2017170706.
14. Chilamkurthy S, Ghosh R, Tanamala S, Biviji M, Campeau NG, Venugopal VK, et al. Deep learning algorithms for detection of critical findings in head CT scans: a retrospective study. Lancet. 2018;392:2388–96.
15. Shadmi R, Mazo V, Bregman-Amitai O, Elnekave E. Fully-convolutional deep-learning based system for coronary calcium score prediction from non-contrast chest CT. 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018); 2018. https://doi.org/10.1109/isbi.2018.8363515.
16. Kamnitsas K, Ferrante E, Parisot S, Ledig C, Nori AV, Criminisi A, et al. DeepMedic for brain tumor segmentation. In: International Workshop on Brainlesion: Glioma, Multiple Sclerosis, Stroke and Traumatic Brain Injuries; 2016. p. 38–49. https://doi.org/10.1007/978-3-319-55524-9_14.
17. Ding Y, Sohn JH, Kawczynski MG, Trivedi H, Harnish R, Jenkins NW, et al. A deep learning model to predict a diagnosis of Alzheimer disease by using F-FDG PET of the brain. Radiology. 2019;290:456–64.
18. Chang HY, Jung CK, Woo JI, Lee S, Cho J, Kim SW, et al. Artificial intelligence in pathology. J Pathol Transl Med. 2019;53:1–12.
19. 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;542:115–8.
20. Haenssle HA, Fink C, Schneiderbauer R, Toberer F, Buhl T, Blum A, et al. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. Ann Oncol. 2018;29:1836–42.
21. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Invest Dermatol. 2018;138:1529–38.
22. Brinker TJ, Hekler A, Enk AH, Klode J, Hauschild A, Berking C, et al. Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. Eur J Cancer. 2019;113:47–54.
23. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316:2402–10.
24. De Fauw J, Ledsam JR, Romera-Paredes B, Nikolov S, Tomasev N, Blackwell S, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24:1342–50.
25. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25:65–9.
26. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. Lancet. 2019;394(10201):861–7. https://doi.org/10.1016/S0140-6736(19)31721-0.
27. Galloway CD, Valys AV, Shreibati JB, Treiman DL, Petterson FL, Gundotra VP, et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. JAMA Cardiol. 2019;4(5):428–36. https://doi.org/10.1001/jamacardio.2019.0640.
28. Wang P, Xiao X, Glissen Brown JR, Berzin TM, Tu M, Xiong F, et al. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. Nat Biomed Eng. 2018;2:741–8. https://doi.org/10.1038/s41551-018-0301-3.
29. Xu J, Yang P, Xue S, Sharma B, Sanchez-Martin M, Wang F, et al. Translating cancer genomics into precision medicine with artificial intelligence: applications, challenges and future perspectives. Hum Genet. 2019;138:109–24.
30. Gurovich Y, Hanani Y, Bar O, Nadav G, Fleischer N, Gelbman D, et al. Identifying facial phenotypes of genetic disorders using deep learning. Nat Med. 2019;25:60–4.
31. Khosravi P, Kazemi E, Zhan Q, Malmsten JE, Toschi M, Zisimopoulos P, et al. Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization. NPJ Digit Med. 2019;2:21. https://doi.org/10.1038/s41746-019-0096-y.
32. Liang H, Tsui BY, Ni H, Valentim CCS, Baxter SL, Liu G, et al. Evaluation and accurate diagnoses of pediatric diseases using artificial intelligence. Nat Med. 2019;25:433–8.
33. Escobar GJ, Turk BJ, Ragins A, Ha J, Hoberman B, LeVine SM, et al. Piloting electronic medical record-based early detection of inpatient deterioration in community hospitals. J Hosp Med. 2016;11(Suppl 1):S18–24.
34. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, et al. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1:18. https://doi.org/10.1038/s41746-018-0029-1.
35. Tomašev N, Glorot X, Rae JW, Zielinski M, Askham H, Saraiva A, et al. A clinically applicable approach to continuous prediction of future acute kidney injury. Nature. 2019;572:116–9.
36. Prasad N, Cheng L-F, Chivers C, Draugelis M, Engelhardt BE. A reinforcement learning approach to weaning of mechanical ventilation in intensive care units. arXiv. 2017; https://arxiv.org/abs/1704.06300. Accessed 1 May 2019.
37. Raghu A, Komorowski M, Ahmed I, Celi L, Szolovits P, Ghassemi M. Deep reinforcement learning for sepsis treatment. arXiv. 2017; https://arxiv.org/abs/1711.09602. Accessed 1 May 2019.
38. Gottesman O, Johansson F, Meier J, Dent J, Lee D, Srinivasan S, et al. Evaluating reinforcement learning algorithms in observational health settings. arXiv. 2018; https://arxiv.org/abs/1805.12298. Accessed 1 May 2019.
39. Kannan A, Chen K, Jaunzeikare D, Rajkomar A. Semi-supervised learning for information extraction from dialogue. Interspeech. 2018;2018:2077–81. https://doi.org/10.21437/interspeech.2018-1318.
40. Chiu C-C, Tripathi A, Chou K, Co C, Jaitly N, Jaunzeikare D, et al. Speech recognition for medical conversations. arXiv. 2017; https://arxiv.org/abs/1711.07274. Accessed 1 May 2019.
41. Nelson A, Herron D, Rees G, Nachev P. Predicting scheduled hospital attendance with artificial intelligence. NPJ Digit Med. 2019;2:26. https://doi.org/10.1038/s41746-019-0103-3.
42. Rajkomar A, Kannan A, Chen K, Vardoulakis L, Chou K, Cui C, et al. Automatically charting symptoms from patient-physician conversations using machine learning. JAMA Intern Med. 2019;179(6):836–8. https://doi.org/10.1001/jamainternmed.2018.8558.
43. McGlynn EA, McDonald KM, Cassel CK. Measurement is essential for improving diagnosis and reducing diagnostic error: a report from the institute of medicine. JAMA. 2015;314:2501–2.
44. Beck AH, Sangoi AR, Leung S, Marinelli RJ, Nielsen TO, van de Vijver MJ, et al. Systematic analysis of breast cancer morphology uncovers stromal features associated with survival. Sci Transl Med. 2011;3:108ra113.
45. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, et al. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2:158–64.
46. Zarins CK, Taylor CA, Min JK. Computed fractional flow reserve (FFTCT) derived from coronary CT angiography. J Cardiovasc Transl Res. 2013;6:708–14. https://doi.org/10.1007/s12265-013-9498-4.

47. Mutlu U, Colijn JM, Ikram MA, Bonnemaijer PWM, Licher S, Wolters FJ, et al. Association of retinal neurodegeneration on optical coherence tomography with dementia: a population-based study. JAMA Neurol. 2018;75:1256–63.
48. 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;1:39. https://doi.org/10.1038/s41746-018-0040-6.
49. Kanagasingam Y, Xiao D, Vignarajan J, Preetham A, Tay-Kearney M-L, Mehrotra A. Evaluation of artificial intelligence-based grading of diabetic retinopathy in primary care. JAMA Netw Open. 2018;1:e182665. https://doi.org/10.1001/jamanetworkopen.2018.2665.
50. Bellemo V, Lim ZW, Lim G, Nguyen QD, Xie Y, Yip MYT, et al. Artificial intelligence using deep learning to screen for referable and vision-threatening diabetic retinopathy in Africa: a clinical validation study. Lancet Digit Health. 2019;1:e35–44.
51. Liu Y, Kohlberger T, Norouzi M, Dahl GE, Smith JL, Mohtashamian A, et al. Artificial intelligence-based breast cancer nodal metastasis detection: insights into the black box for pathologists. Arch Pathol Lab Med. 2018;143(7):859–68. https://doi.org/10.5858/arpa.2018-0147-oa.
52. Steiner DF, MacDonald R, Liu Y, Truszkowski P, Hipp JD, Gammage C, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42:1636–46.
53. Lindsey R, Daluiski A, Chopra S, Lachapelle A, Mozer M, Sicular S, et al. Deep neural network improves fracture detection by clinicians. Proc Natl Acad Sci U S A. 2018;115:11591–6.
54. Mori Y, Kudo S-E, Misawa M, Saito Y, Ikematsu H, Hotta K, et al. Real-time use of artificial intelligence in identification of diminutive polyps during colonoscopy. Ann Intern Med. 2018;169:357. https://doi.org/10.7326/m18-0249.
55. Long E, Lin H, Liu Z, Wu X, Wang L, Jiang J, et al. An artificial intelligence platform for the multihospital collaborative management of congenital cataracts. Nat Biomed Eng. 2017;1:0024. https://doi.org/10.1038/s41551-016-0024.
56. Turakhia MP, Desai M, Hedlin H, Rajmane A, Talati N, Ferris T, et al. Rationale and design of a large-scale, app-based study to identify cardiac arrhythmias using a smartwatch: The Apple Heart Study. Am Heart J. 2019;207:66–75.
57. Lin H, Li R, Liu Z, Chen J, Yang Y, Chen H, et al. Diagnostic efficacy and therapeutic decision-making capacity of an artificial intelligence platform for childhood cataracts in eye clinics: a multicentre randomized controlled trial. EClinicalMedicine. 2019;9:52–9. https://doi.org/10.1016/j.eclinm.2019.03.001.
58. Wu L, Zhang J, Zhou W, An P, Shen L, Liu J, et al. Randomised controlled trial of WISENSE, a real-time quality improving system for monitoring blind spots during esophagogastroduodenoscopy. Gut. 2019. https://doi.org/10.1136/gutjnl-2018-317366.
59. Wang P, Berzin TM, Brown JRG, Bharadwaj S, Becq A, Xiao X, et al. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. Gut. 2019;68(10):1813–9. https://doi.org/10.1136/gutjnl-2018-317500.
60. Titano JJ, Badgeley M, Schefflein J, Pain M, Su A, Cai M, et al. Automated deep-neural-network surveillance of cranial images for acute neurologic events. Nat Med. 2018;24:1337–41.
61. Brocklehurst P, Field D, Greene K, Juszczak E, Keith R, Kenyon S, et al. Computerised interpretation of fetal heart rate during labour (INFANT): a randomised controlled trial. Lancet. 2017;389:1719–29. https://doi.org/10.1016/s0140-6736(17)30568-8.
62. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: an introduction to the new Medical Research Council guidance. In: Evidence-based Public Health: Effectiveness and Efficiency; 2009. p. 185–202. https://doi.org/10.1093/acprof:oso/9780199563623.003.012.
63. Collins GS, Reitsma JB, Altman DG, Moons KGM. Transparent Reporting of a Multivariable Prediction Model for Individual Prognosis or Diagnosis (TRIPOD). Circulation. 2015;131:211–9. https://doi.org/10.1161/circulationaha.114.014508.
64. Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. Lancet. 2019;393:1577–9.
65. Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. NPJ Digit Med. 2018;1:40. https://doi.org/10.1038/s41746-018-0048-y.
66. Saito T, Rehmsmeier M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. PLoS One. 2015;10:e0118432. https://doi.org/10.1371/journal.pone.0118432.
67. Shah NH, Milstein A, Bagley PhD SC. Making machine learning models clinically useful. JAMA. 2019. https://doi.org/10.1001/jama.2019.10306.
68. Vickers AJ, Cronin AM, Elkin EB, Gonen M. Extensions to decision curve analysis, a novel method for evaluating diagnostic tests, prediction models and molecular markers. BMC Med Inform Decis Mak. 2008;8:53.
69. Marcus G. Deep learning: a critical appraisal. arXiv. 2018; https://arxiv.org/abs/1801.00631. Accessed 1 May 2019.
70. Nestor B, McDermott MBA, Chauhan G, Naumann T, Hughes MC, Goldenberg A, et al. Rethinking clinical prediction: why machine learning must consider year of care and feature aggregation. In: Machine Learning for Health (ML4H): NeurIPS; 2018. https://arxiv.org/abs/1811.12583. Accessed 1 May 2019.
71. Davis SE, Greevy RA, Fonnesbeck C, Lasko TA, Walsh CG, Matheny ME. A nonparametric updating method to correct clinical prediction model drift. J Am Med Inform Assoc. 2019. https://doi.org/10.1093/jamia/ocz127.
72. Ribeiro M, Singh S, Guestrin C. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations; 2016. https://doi.org/10.18653/v1/n16-3020.
73. Winkler JK, Fink C, Toberer F, Enk A, Deinlein T, Hofmann-Wellenhof R, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. 2019. https://doi.org/10.1001/jamadermatol.2019.1735.
74. Badgeley MA, Zech JR, Oakden-Rayner L, Glicksberg BS, Liu M, Gale W, et al. Deep learning predicts hip fracture using confounding patient and healthcare variables. arXiv. 2018; https://arxiv.org/abs/1811.03695. Accessed 1 May 2019.
75. Zech JR, Badgeley MA, Liu M, Costa AB, Titano JJ, Oermann EK. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 2018;15:e1002683.
76. Debray TPA, Vergouwe Y, Koffijberg H, Nieboer D, Steyerberg EW, Moons KGM. A new framework to enhance the interpretation of external validation studies of clinical prediction models. J Clin Epidemiol. 2015;68:279–89.
77. Kim DW, Jang HY, Kim KW, Shin Y, Park SH. Design characteristics of studies reporting the performance of artificial intelligence algorithms for diagnostic analysis of medical images: results from recently published papers. Korean J Radiol. 2019;20:405–10.
78. Crawford K, Calo R. There is a blind spot in AI research. Nature. 2016;538:311–3.
79. Barocas S, Selbst AD. Big Data's Disparate Impact. 104 California Law Review 671; 2016. https://doi.org/10.2139/ssrn.2477899.
80. Chen IY, Johansson FD, Sontag D. Why Is My Classifier Discriminatory? In: 32nd Conference on Neural Information Processing Systems (NeurIPS). 2018. http://papers.nips.cc/paper/7613-why-is-my-classifier-discriminatory.pdf.
81. Haenssle HA, Fink C, Rosenberger A, Uhlmann L. Reply to "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists" by H. A. Haenssle et al. Ann Oncol. 2019. https://doi.org/10.1093/annonc/mdz015.
82. Ward-Peterson M, Acuña JM, Alkhalifah MK, Nasiri AM, Al-Akeel ES, Alkhaldi TM, et al. Association between race/ethnicity and survival of melanoma patients in the United States over 3 decades. Medicine. 2016;95:e3315. https://doi.org/10.1097/md.0000000000003315.
83. Finlayson SG, Bowers JD, Ito J, Zittrain JL, Beam AL, Kohane IS. Adversarial attacks on medical machine learning. Science. 2019;363:1287–9.
84. Mandel JC, Kreda DA, Mandl KD, Kohane IS, Ramoni RB. SMART on FHIR: a standards-based, interoperable apps platform for electronic health records. J Am Med Inform Assoc. 2016;23:899–908.
85. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PRO, Bernstam EV, et al. Caveats for the use of operational electronic health record data in comparative effectiveness research. Med Care. 2013;51(8 Suppl 3):S30–7.
86. Food and Drug Administration. Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD): FDA; 2019. https://www.regulations.gov/document?D=FDA-2019-N-1185-0001. Accessed 1 May 2019.
87. Core MG, Lane HC, van Lent M, Gomboc D, Solomon S, Rosenberg M. Building Explainable Artificial Intelligence Systems. IAAI'06 Proceedings of the 18th conference on Innovative Applications of Artificial Intelligence. Volume 2; 2006. p. 1766–73.
88. Holzinger A, Biemann C, Pattichis CS. What do we need to build explainable AI systems for the medical domain? arXiv. 2017; https://arxiv.org/abs/1712.09923. Accessed 1 May 2019.
89. Samek W, Wiegand T, Müller K-R. Explainable artificial intelligence: understanding, visualizing and interpreting deep learning models. arXiv. 2017; http://arxiv.org/abs/1708.08296. Accessed 1 May 2019.

90. Bologna G, Hayashi Y. Characterization of symbolic rules embedded in deep DIMLP networks: a challenge to transparency of deep learning. J Art Intel Soft Comput Res. 2017;7(4):265–86. https://doi.org/10.1515/jaiscr-2017-0019.
91. Fox J. A short account of Knowledge Engineering. Knowl Eng Rev. 1984;1:4–14. https://doi.org/10.1017/s0269888900000424.
92. Lacave C, Díez FJ. A review of explanation methods for Bayesian networks. Knowl Eng Rev. 2002;17:107–27. https://doi.org/10.1017/s026988890200019x.
93. Doshi-Velez F, Kim B. Towards a rigorous science of interpretable machine learning. arXiv. 2017; http://arxiv.org/abs/1702.08608. Accessed 1 May 2019.
94. Lehman CD, Wellman RD, Buist DSM, Kerlikowske K, Tosteson ANA, Miglioretti DL, et al. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. JAMA Intern Med. 2015;175:1828–37.
95. Phansalkar S, van der Sijs H, Tucker AD, Desai AA, Bell DS, Teich JM, et al. Drug-drug interactions that should be non-interruptive in order to reduce alert fatigue in electronic health records. J Am Med Inform Assoc. 2013;20:489–93.
96. Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, et al. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology. 2019;126:552–64.
97. Wang D, Khosla A, Gargeya R, Irshad H, Beck AH. Deep Learning for Identifying Metastatic Breast Cancer. 2016. http://arxiv.org/abs/1606.05718. Accessed 28 Aug 2019.
98. Google. People and AI Guidebook. https://pair.withgoogle.com/. Accessed 10 May 2019.

Publisher's Note