Framing the challenges of artificial intelligence in medicine

Kun-Hsing Yu, Isaac S Kohane

Biomedical Informatics, Harvard Medical School, Boston, Massachusetts, USA

Correspondence to
Dr Kun-Hsing Yu and Professor Isaac S Kohane, Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA;
Kun-Hsing_Yu@hms.harvard.edu,
isaac_kohane@harvard.edu

Received 4 July 2018
Revised 20 August 2018
Accepted 26 August 2018

To cite: Yu K-H, Kohane IS. *BMJ Qual Saf* Epub ahead of print: [please include Day Month Year]. doi:10.1136/bmjqs-2018-008551

On a clear January morning in Florida, a Tesla enthusiast and network entrepreneur was driving his new Tesla Model S on US Highway 27A, returning from a family trip. He had posted dozens of widely circulated YouTube tutorial videos on his vehicle and clearly understood many of the technical details of his car. That day, he let the vehicle run autonomously on Autopilot mode for 37 min, before it crashed into the trailer of a truck turning left. The Autopilot did not identify the white side of the trailer as a potential hazard, and the driver was killed, leaving his family and his high-tech business behind.[1] This tragedy is not a metaphor for artificial intelligence (AI) applications but an example of a long-recognised challenge in AI: the Frame Problem.[2] Although rarely appreciated in the scholarly and lay descriptions of the stunning recent successes of AI in medical applications, the Frame Problem and related AI challenges will have unintended harmful effects to the care of patients if not directly addressed.

With the recent advancement in machine learning algorithms, many medical tasks previously thought to require human expertise have been replicated by AI systems at or above the level of accuracy in human experts. These important demonstrations range from evaluating fundus retinography[3] and histopathology[4] to reading chest radiographs[5] and assessment of skin lesions.[6] These studies have encompassed very large numbers of patient cases and have been extensively benchmarked against clinicians. However, all these studies are retrospective in that they involve a collection of labelled cases against which the AI systems are trained and another collection against which they are tested or validated. So far, they have not entered into routine prospective use in the clinic where the Frame Problem will manifest itself most pathologically.

The Frame Problem was first introduced by computer scientists and cognitive science pioneers McCarthy and Hayes[2] in 1969 and revolved around the difficulty in identifying and updating a set of axioms to properly describe the environment for autonomous agents. To provide a medical example, let us define a worrisome chest X-ray as being one in which a shadow or a density appears that resembles those seen in lung cancer, pneumonia or various pulmonary pathologies. As in the recent successes, we are confident that an AI program can be trained with enough well-curated cases to give accuracies greater than 90% and better than the typical or even expert radiologist.[5] What if the X-ray technician leaves on patient Jill Doe the adhesive ECG lead connectors from her recent inpatient ECG. Will the AI program classify these circular medical artefacts as one of the known chest lesions? That false positive would soon become apparent and the AI engineers would include these circular ECG leads in the training sets and that error would be eliminated. What if the ECG leads superimposed part of a real shadow of a lung nodule such that the AI program would miss it? Presumably, after a few such cases where the nodule would become clinically obvious, the AI engineers would ensure that the training sets would have enough cases of such overlap to give adequate sensitivity and specificity in these instances. What if Jill Doe, despite the technician's warning, had placed her hand with a wedding ring on her chest? If no one except the AI program looks at the image, would automated classification dismiss the ring as a non-medical artefact or would it classify it incorrectly as a lesion? If the AI program is trained to recognise such non-medical artefacts, then how will it classify a toddler—Jane Doe's—chest X-ray if she comes in with stridor and shortness of breath? Specifically, with a ring visible in the image because Jane Doe had aspirated her mother's ring? In each of these cases, a reasonable response is that the program could be further trained and adjusted, or a human overseeing the process could use their common sense and experience to intervene. If the former, then the central question is how rapidly will training or adjusting reach acceptable performance? If the latter, how efficiently can a human oversee these programs with consistent vigilance while preserving the cost and accuracy benefits? Fortunately, these are empirically answerable questions. Unfortunately, they can only be answered through prospective trials if AI programs in medical care are to have the public and professional trust required for their full impact.

Trust in medical technology is closely related to its anticipated utility. Disruptions of the current clinical workflow will inevitably face inertia, and if the perception emerges that there are untoward consequences of a new technology then the barrier to any similar technologies will become next to insurmountable. The multidecade hiatus in gene therapy[7 8] and the retardant effect of the Theranos debacle on fingerprick blood diagnostics[9] are two more recent examples. This issue becomes especially complicated if the technology is complex or if the implementation details are proprietary, such that the general public or even domain experts cannot fully evaluate its efficacy and potential hazard based on the information they received. Moreover, even in cases of overwhelming public good such as in vaccination programs, lack of full disclosure and inadequate attention to patient education and autonomy can have dramatic negative consequences for the diffusion of helpful technologies.[10] The relevance here is that the Frame Problem and related issues will inevitably cause medical errors that will draw the attention of both the public and, at least in the USA, lawsuits against parties using, deploying or developing medical AI applications. The 'black-box' nature of many modern machine learning algorithms will further exacerbate the issue. High-profile examples of harmful or inadequate performance will bring extra scrutiny on the whole field and may retard the further development of even more robust AI systems.

Furthermore, data-driven AI algorithms are not immune from the 'garbage-in-garbage-out' rule. Machine learning algorithms are designed to identify the hidden patterns of the data and generate output predictions based on what they have seen in the past.[11] As many input data sets contain artefacts or biases, the models learnt from the data carry the biases and can potentially amplify them. For instance, electronic medical records and insurance claims data sets are records of patients' clinical courses and also a tool for healthcare providers to justify specific levels of reimbursement. Consequently, optimised machine learning models in healthcare can be confounded by their training data where the reimbursement strategies driving diagnostic coding are implicit and may not reflect a more objective clinical assessment. Additionally, AI systems could perpetuate racial bias since the biases exist in historical data,[12] resulting in partiality in the seemingly 'objective' computational methods. Ongoing data-driven controversies regarding the causes of poor health outcomes among disadvantaged populations (ie, differential access vs differential biological health risks)[13] illustrate just how difficult it is to avoid confounding in the analysis of observational health data. Paying a lot more attention to data quality and provenance, an expensive proposition, will go a long way to foster 'patient trust' in medical AI systems, and to avoid unethical medical performance, even if only by negligence.

Last but not least, even if an AI system is designed to advise human practitioners, rather than to carry out the actual diagnostic or treatment tasks, it may still result in detrimental unintended consequences, such as confirmatory bias and alert fatigue. A recent study showed that over-reliance on decision support systems resulted in increased false negative rate in radiology diagnoses, compared with the study scenario where the computer-aided diagnostic system was unavailable to the same group of radiologists.[14] Additionally, excessive warning information will result in alert fatigue,[15] and inexperienced practitioners may over-react to the warning messages. As such, AI developers need to pay attention to the clinical usage of automated systems, even if the systems only play an advisory role.

To address these challenges, researchers need to acknowledge and address the limitations in the current association-based machine learning paradigm and ensure quality control of the AI-based applications in various clinical settings and patient populations (table 1). For instance, the chest X-ray films with atypical feature statistics should be reviewed by radiologists to ensure that the obvious artefacts or unusual clinical contexts were adequately captured. In addition, prospective trials are needed to better understand the behaviour of AI systems in the real-world clinical settings, and continual calibration by human feedback is warranted to identify the development of emerging diseases as well as to examine the effectiveness of AI in recognising previously unclassified disease patterns. Due to the fact that AI technologies evolve at a fast pace and that machine learning models can update with additional pieces of information, regulatory bodies face a unique challenge in specifying trial requirements for regulatory approval.[16] To address this issue, the US Food and Drug Administration recently announced a pilot certification approach that inspects the AI developers, in addition to the product.[17] Detailed policies regarding the certification of developers are yet to be established. Furthermore, since many machine learning algorithms only focused on association identification, causal inference analyses are needed to characterise the causal relations underpinning the

Table 1 A list of prominent issues of medical artificial intelligence (AI) applications and potential solutions

Issue	Potential solution
The Frame Problem (the difficulty in identifying and updating a set of axioms to properly describe the environment for autonomous agents).	▶ Clinician review of input with atypical feature statistics. ▶ Rigorous prospective clinical trials in diverse patient populations. ▶ Continual calibration by human feedback.
Trust in the performance of the AI program.	▶ Disclosure of implementation details, nature of training sets and shortcomings of the AI systems. ▶ Develop interpretable machine learning models. ▶ Patient education.
Amplifying biases presented in the historical data.	▶ Ongoing acquisition of training data reflecting current practice and population characteristics. ▶ Identify confounders in the association-based models.
Clinical workflow disruption.	▶ Redesign workflow that enables AI assistance without encouraging clinician decision-making passivity or aggravating 'alert fatigue'.

observed associations,[18] thereby mitigating the issues of confounding, and provide more transparency to the machine learning models. These steps are required to ensure public trust in novel medical AI applications.

Our focus on the near-term limitations of association-driven AI does not excuse any myopia regarding the highly variable and sometimes woefully inadequate performance of human clinicians. The mortality cost from medical errors alone, not including suboptimal decisions, has been widely documented for decades.[19] Nonetheless, since modern machine learning algorithms perform complex mathematical transformations to the input data,[16] errors made by computational systems will require extra vigilance to detect and interpret.[20] These cryptic errors and biases in the AI black boxes may systematically harm numerous patients simultaneously and worsen health disparities at scale.[21] In addition, even a robust AI application can reduce efficiency and cause additional medical errors if not adequately integrated into the current clinical workflow.[20] A better workflow would allow human clinicians and AI applications to compensate for their different and complementary weaknesses and blind spots to best serve the interests of patient safety and clinical efficiency.

Building an intelligent automated entity to evaluate, diagnose and treat patients in research settings is arguably the easiest part of designing an end-to-end medical AI system. In the context of the hype and hopes surrounding emerging AI applications in medicine, we need to acknowledge the brittleness of these systems, the importance of defining the correct frameworks for their application, and ensure rigorous quality control, including human supervision, to avoid driving our patients on autopilot towards unexpected, unwanted and unhealthful outcomes.

Contributors K-HY conceptualised and drafted the manuscript. ISK revised the manuscript and supervised the work.

Funding This study was funded by Harvard University (Harvard Data Science Fellowship) and the National Institutes of Health (Grant Number: OT3OD025466).

Competing interests Harvard Medical School submitted a provisional patent application on digital pathology profiling to the United States Patent and Trademark Office (USPTO).

Patient consent Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

REFERENCES

1. Abrams R, Kurtz A, Brown J, 2016. Who died in self-driving accident, tested limits of his tesla: the New York Times. https://www.nytimes.com/2016/07/02/business/joshua-brown-technology-enthusiast-tested-the-limits-of-his-tesla.html (accessed 3 Aug 2018).
2. McCarthy J, Hayes PJ. Some philosophical problems from the standpoint of artificial intelligence. *Readings in Artificial Intelligence* 1969:431–50.
3. Gulshan V, Peng L, Coram M, *et al*. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. *JAMA* 2016;316:2402–10.
4. Yu KH, Berry GJ, Rubin DL, *et al*. Association of omics features with histopathology patterns in lung adenocarcinoma. *Cell Syst* 2017;5:620–7.
5. Wang X, Peng Y, Lu L. ChestX-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. *arXiv preprint arXiv* 2017:170502315.
6. Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
7. Sibbald B. Death but one unintended consequence of gene-therapy trial. *CMAJ* 2001;164:1612.
8. Wilson JM. Lessons learned from the gene therapy trial for ornithine transcarbamylase deficiency. *Mol Genet Metab* 2009;96:151–7.
9. Waltz E. After Theranos. *Nat Biotechnol* 2017;35:11–15.
10. Amin ANE, Parra MT, Kim-Farley R, *et al*. Ethical issues concerning vaccination requirements. *Public Health Rev* 2012;34:14.
11. Yu KH, Snyder M. Omics profiling in precision oncology. *Mol Cell Proteomics* 2016;15:2525–36.
12. Buranyi S, 2017. Rise of the racist robots – how AI is learning all our worst impulses the fuardian: the guardian. https://www.theguardian.com/inequality/2017/aug/08/rise-of-the-racist-robots-how-ai-is-learning-all-our-worst-impulses (accessed 27 Mar 2018).
13. Richardson LD, Norris M. Access to health and health care: how race and ethnicity matter. *Mt Sinai J Med* 2010;77:166–77.
14. Lehman CD, Wellman RD, Buist DS, *et al*. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA Intern Med* 2015;175:1828–37.
15. Phansalkar S, van der Sijs H, Tucker AD, *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.

16. Yu KH, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018:(accepted).
17. U.S. Food and Drug Administration, 2018. Digital health software precertification (Pre-Cert) program. https://www.fda.gov/MedicalDevices/DigitalHealth/DigitalHealthPreCertProgram/default.htm (accessed 2 Aug 2018).
18. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. *J Epidemiol Community Health* 2006;60:578–86.
19. Stelfox HT, Palmisani S, Scurlock C, *et al*. The "To Err is Human" report and the patient safety literature. *Qual Saf Health Care* 2006;15:174–8.
20. Ash JS, Berg M, Coiera E. Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *J Am Med Inform Assoc* 2004;11:104–12.
21. O'Neil C. *Weapons of math destruction: how big data increases inequality and threatens democracy*. New York City: Broadway Books, 2016.