PSNet
PATIENT SAFETY NETWORK

Emerging Safety Issues in Artificial Intelligence

July 1, 2019
Challen R. Emerging Safety Issues in Artificial Intelligence. PSNet [internet]. 2019.
https://psnet.ahrq.gov/perspective/emerging-safety-issues-artificial-intelligence

Perspective

Background

Since the beginning of digital health records, there has been interest in using computers to help clinicians deliver a safer and higher quality service. Medicine has a long history of artificial intelligence (AI) decision support systems.(1) Many rule-based AI decision support systems are in widespread use today, and others are still under active development.(2) More recently, however, AI research has focused on machine learning as a way to use past examples or experience to develop an internal model of a problem and use that model to make inferences about new patients.(3) Due to this focus within the current literature, the phrase AI has recently become almost synonymous with machine learning.

Future machine learning systems may make diagnostic predictions about new patients based on their data or develop strategies for managing patients by maximizing a beneficial clinical outcome.(4) Such systems may either be periodically retrained or may actively learn from previous decisions, and in the longer term autonomous decision-making systems are expected to be applied to specific clinical problems. This evolution may lead in the short, medium, and long term to a range of specific safety issues investigated in a recently published review (Figure).(5) However, in the short term, Food and Drug Administration (FDA) approval is emerging for machine learning systems that principally offer diagnostic support (3), and safety issues of these systems are the focus of this article.

Emerging Safety Issues

In traditional clinical decision support systems, rules are authored by experts and have a defined provenance, which is ideally evidence-based. Machine learning behavior, on the other hand, is dependent on the data on which it has been trained. When the patient is not adequately represented by the data that the machine learning system was trained on, the accuracy of the system will be compromised. Known as distributional shift, this phenomenon can be the result of unrepresentative training data, inadequate labeling of the patient's outcome, or inappropriate application of a machine learning system to a different patient

group ("out of sample"). These are summarized in the Figure but can, for example, be the result of different patient demographics, temporal changes (6), differences in the clinical stage of disease, inconsistencies defining a gold standard diagnosis (7), or even the result of differences in machines used to scan a patient.(8) The question that clinicians should ask is how can we be confident that a particular machine learning decision support system is appropriate for a given patient, without in-depth knowledge of the whole training process?

Machine learning training involves a process to optimize prediction accuracy, and increasing numbers of studies claim superhuman performance compared to clinicians. In one such example, the diagnostic performance of machine learning for classifying skin disease was demonstrably better than that of clinicians on artificial test cases across a range of benign and malignant skin disorders, but at the same time the machine learning system missed more malignant disease than clinicians did.(9) In cases such as these, the outcome severity must be considered when training a machine learning system (9), and this in turn must take into account the purpose for which the algorithm is intended, weighing both the benefit of detection and the potential harm caused by false positives and false negatives. Some algorithms, such as the Isabel diagnostic support system, include a "can't miss diagnoses" category to reprioritize diagnostic recommendations around this axis.

Apple has recently obtained FDA approval for an algorithm to use smart watches to detect atrial fibrillation, but "there are legitimate concerns that the widescale use of such an algorithm, particularly in the low-risk, young population who wear Apple Watches, will lead to a substantial number of false-positive atrial fibrillation diagnoses and prompt unnecessary medical evaluations."(3) Preliminary results of a large scale study on the Apple Watch were recently presented to the American College of Cardiology, from which it is too early to draw conclusions, but of the small number of atrial fibrillation alerts produced, only 34% were confirmed by follow-up electrocardiogram patch studies.(10) However, the principle guiding the development of the GRADE guidelines (11) is also of relevance here, and frequently not considered in machine learning studies: "If a test fails to improve patient-important outcomes, there is no reason to use it, whatever its accuracy."(12)

In an assessment of a rules-based diagnostic support system for electrocardiograms, Tsai and colleagues (13) neatly demonstrated that clinicians are influenced by the advice of expert systems, even when erroneous, and demonstrate the same automation bias (5) identified in the aviation industry (14), or in the context of autonomous vehicles.(15) In reviews of decision support in medicine (16) and in nonmedical contexts (14,15), there is limited evidence of deskilling, but what seems more apparent is that people trained in the presence of decision support fail to acquire the skills to operate without it.(15) This issue is significant, but not limited to AI in general or machine learning in particular. It is difficult to imagine a definitive diagnosis of valvular heart disease would be made in the absence of an echocardiogram. Unsurprisingly, auscultation skills are becoming less important in clinical practice, due in part to the availability of highly accurate and accessible technology.(17)

There are important differences in the application of machine learning technologies versus clinical tools such as point-of-care echocardiography. As learning systems, machine learning systems are fallible, just as human decision makers are, and they will inevitably make mistakes. They may also be inscrutable and operate like "black boxes" (5), in which it is impossible to assess the machine learning system's decision-

making process. It is generally assumed (3) that it will be a clinician's role to interpret the machine learning system's recommendation and take control (and responsibility) when it errs. However, in a rather unique and potentially self-fulfilling way, the very existence of machine learning systems and associated automation bias may prevent the development and refinement of the clinical skills required to safely oversee their fallible decisions.

Whatever the technology, eventually it may reach a point where an undeniably superhuman AI system is being supervised by a comparatively incompetent clinician and makes a mistake that causes harm. In this way, health care mirrors the experience of other high-risk technological systems, such as self-driving automobiles and advanced automation in aviation, in that a small number of high-profile accidents will cause the public to question the wisdom of automation. What responsibility, if any, should the clinician have in this scenario? Our view is that such a machine will not fit well into existing definitions of a medical device, but rather will need to operate within a regulatory framework similar to that of human decision-makers, both in terms of appropriate qualification, expected standard of practice, and performance review, but also in terms of assuming responsibility for its own errors. In this case, the AI system would need to be insured against error, just like a clinician. It will take a brave company to be the first to insure an AI system.

Conclusion

This piece has explored various practical and philosophical issues that could arise for adoption of machine learning, and more generally AI, systems in medicine. However, the severity of these issues depends critically on the AI application. Trials are underway for sepsis prediction algorithms (18), where the benefits are clear and can be expressed in patient-important outcomes, and where algorithmic monitoring of large volumes of data is complementary to the clinical skill of identifying the sick patient from their overall clinical picture. Clinicians and machine learning researchers need to focus on finding the "no-brainer" applications where machine learning techniques and computers' superior data-crunching power can be harnessed safely for clinical benefit. In our view, a key to this is a proper understanding of the role diagnostic tests play in a clinical setting.(11)

Studies reporting superhuman machine learning performance in a laboratory setting (summarized by Topol and colleagues [3]) foster mistrust among clinicians, who suspect that such performance will not be achievable in real-world settings. Some studies, on the other hand, concentrate on testing the decision making of machine learning systems in conjunction with the clinician and show how this can lead to improved performance (18,19) compared to the unaided clinician. These paint a much brighter picture of machine learning working in collaboration with clinicians, in real clinical settings, and open the interesting potential for machine learning to act as both a teaching and decision support tool. The comparison between the clinician supported by machine learning versus the clinician without machine learning should be part of the standard design of machine learning experiments.

Robert Challen, MA, MBBS EPSRC Centre for Predictive Modelling in Healthcare and Department of Mathematics, Living Systems Institute University of Exeter Exeter, Devon, UK Taunton and Somerset NHS Foundation Trust Taunton, Somerset, UK

Acknowledgements

The author gratefully acknowledges the financial support of the Engineering and Physical Sciences Research Council (EPSRC) via grant EP/N014391/1, and the NHS Global Digital Exemplar program. The author thanks Dr. Luke Gompels, CCIO, TSFT; Mr. Tom Edwards, CCIO, TSFT; Prof. Krasimira Tsaneva-Atanasova, EPSRC Centre for Predictive Modelling in Healthcare, University of Exeter; and Prof. Martin Pitt, PenCHORD, University of Exeter, for comments that greatly enhanced this article.

References

1. Miller RA. Medical diagnostic decision support systems—past, present, and future: a threaded bibliography and brief commentary. J Am Med Inform Assoc. 1994;1:8-27. [go to PubMed]

2. Fox J. Cognitive systems at the point of care: the CREDO program. J Biomed Inform. 2017;68:83-95. [go to PubMed]

3. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25:44-56. [go to PubMed]

4. Ravi D, Wong C, Deligianni F, et al. Deep learning for health informatics. IEEE J Biomed Health Inform. 2017;21:4-21. [go to PubMed]

5. Challen R, Denny J, Pitt M, Gompels L, Edwards T, Tsaneva-Atanasova K. Artificial intelligence, bias and clinical safety. BMJ Qual Saf. 2019;28:231-237. [go to PubMed]

6. Davis SE, Lasko TA, Chen G, Siew ED, Matheny ME. Calibration drift in regression and machine learning models for acute kidney injury. J Am Med Inform Assoc. 2017;24:1052-1061. [go to PubMed]

7. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv [cs.CV]. 2017. [Available at]

8. De Fauw J, Ledsam JR, Romera-Paredes B, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. Nat Med. 2018;24:1342-1350. [go to PubMed]

9. Esteva A, Kuprel B, Novoa RA, et al. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542:115-118. [go to PubMed]

10. Apple Heart Study gives glimpse into how wearable technology may help flag heart rhythm problems. American College of Cardiology. March 16, 2019. [Available at]

11. Schünemann HJ, Oxman AD, Brozek J, et al. Grading quality of evidence and strength of recommendations for diagnostic tests and strategies. BMJ. 2008;336:1106-1110. [go to PubMed]

12. Schünemann HJ, Oxman AD, Brozek J, et al. GRADE: assessing the quality of evidence for diagnostic recommendations. Evid Based Med. 2008;13:162-163. [go to PubMed]

13. Tsai TL, Fridsma DB, Gatti G. Computer decision support as a source of interpretation error: the case of electrocardiograms. J Am Med Inform Assoc. 2003;10:478-483. [go to PubMed]

14. Parasuraman R, Manzey DH. Complacency and bias in human use of automation: an attentional integration. Hum Factors. 2010;52:381-410. [go to PubMed]

15. Trösterer S, Gärtner M, Mirnig A, et al. You Never Forget How to Drive: Driver Skilling and Deskilling in the Advent of Autonomous Vehicles. Proceedings of the 8th International Conference on Automotive User Interfaces and Interactive Vehicular Applications. New York, NY: ACM; 2016:209-216. [Available at]

16. Ash JS, Sittig DF, Campbell EM, Guappone KP, Dykstra RH. Some unintended consequences of clinical decision support systems. AMIA Annu Symp Proc. 2007:26-30. [go to PubMed]

17. Roelandt JRTC. The decline of our physical examination skills: is echocardiography to blame? Eur Heart J Cardiovasc Imaging. 2014;15:249-252. [go to PubMed]

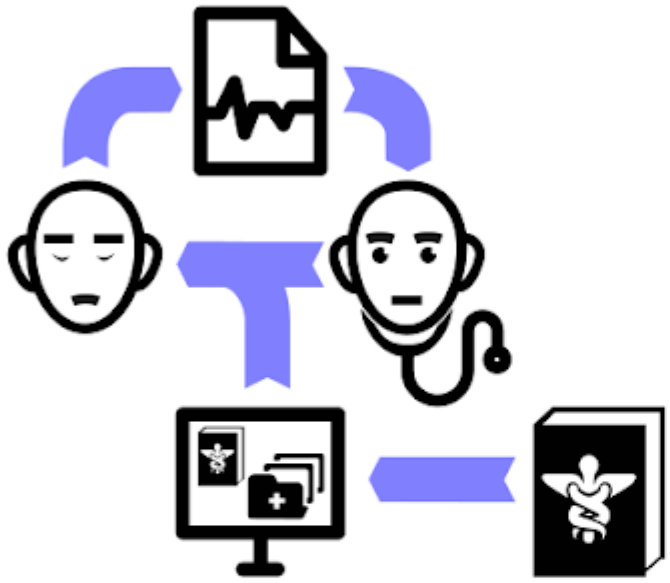
18. Mao Q, Jay M, Hoffman JL, et al. Multicentre validation of a sepsis prediction algorithm using only vital sign data in the emergency department, general ward and ICU. BMJ Open. 2018;8:e017833. [go to PubMed]

19. Steiner DF, MacDonald R, Liu Y, et al. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. Am J Surg Pathol. 2018;42:1636-1646. [go to PubMed]

Figure

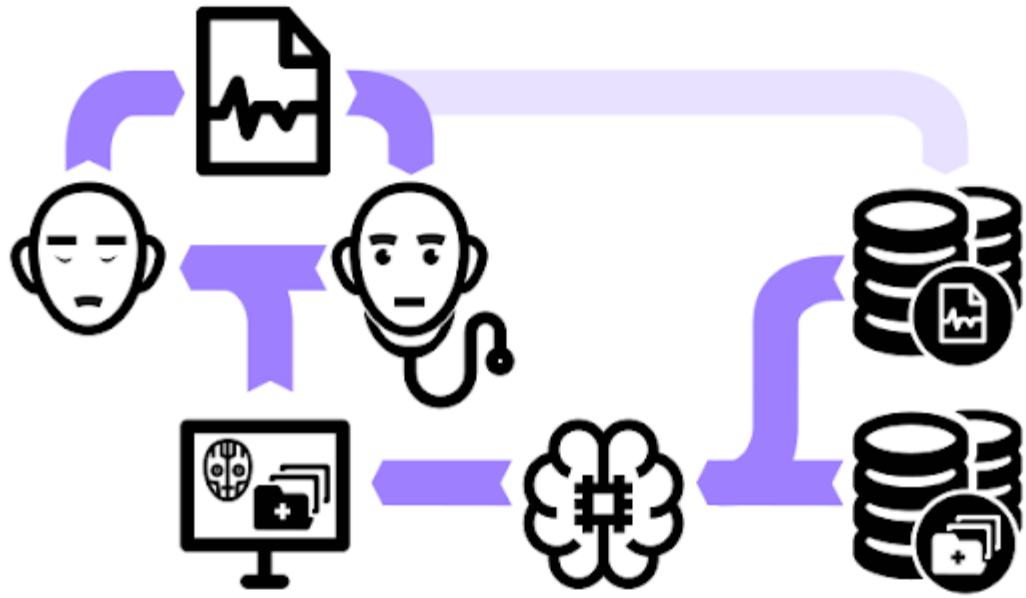
Figure. Trends in Machine Learning Research and Associated Short-, Medium-, and Long-Term Safety Risks.(5)

Rule based decision support systems (DSS) have been available in medicine for some time, e.g. ePrescribing. The DSS combines evidence-based rules with patient data to recommend a clinical action. These recommendations are acted on by a clinician.



DATA DRIVEN DECISION SUPPORT

Supervised machine learning (ML) systems use historical data and outcomes to find associations and make predictions. The clinician interprets guidance in the context of a patient.



ACTIVE LEARNING

More advanced AI systems can learn continuously from current and historical patient data. This feedback allows them to improve performance, but leads to a system which is dynamic