OPEN ACCESS

Bridging the implementation gap of machine learning in healthcare

Martin G Seneviratne ,[1] Nigam H Shah ,[1] Larry Chu[2]

[1]Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, California, USA
[2]Anesthesia, Stanford University, Stanford, California, USA

Correspondence to
Dr Nigam H Shah, 1265 Welch Road, Center for Biomedical Informatics Research, Stanford University School of Medicine, Stanford, CA 94305, USA; nigam@stanford.edu

MGS and NHS are joint first authors.

Accepted 13 December 2019
Published Online First 20 December 2019

To cite: Seneviratne MG, Shah NH, Chu L. *BMJ Innov* 2019;**6**:45–47.

Applications of machine learning on clinical data are now attaining levels of performance that match or exceed human clinicians.[1–3] Fields involving image interpretation—radiology, pathology and dermatology—have led the charge due to the power of convolutional neural networks, the existence of standard data formats and large data repositories. We have also seen powerful diagnostic and predictive algorithms built using a range of other data, including electronic health records (EHR), -omics, monitoring signals, insurance claims and patient-generated data.[4] The looming extinction of doctors has captured the public imagination, with editorials such as 'The AI Doctor Will See You Now'.[5] The prevailing view among experts is more balanced: that doctors who use artificial intelligence (AI) will replace those who do not.[6]

Amid such inflated expectations, the elephant in the room is the implementation gap of machine learning in healthcare.[7 8] Very few of these algorithms ever make it to the bedside; and even the most technology-literate academic medical centres are not routinely using AI in clinical workflows. A recent systematic review of deep learning applications using EHR data highlighted the need to focus on the last mile of implementation: 'for direct clinical impact, deployment and automation of deep learning models must be considered'.[9] The typical life-cycle of an algorithm remains: train on historical data, publish a good receiver-operator curve and then collect dust in the 'model graveyard'.

This begs the question: *if model performance is so promising, why is there such a chasm between development and deployment?* In order to bridge this implementation gap, our focus must shift away from optimising an area under the curve towards three more practical aspects of model design: actionability, safety and utility.

ACTIONABILITY

First, an algorithm must be clinically actionable—its output should be linked to some intervention by the clinician or patient. All too often, a sophisticated machine learning model is developed with excellent discriminative or predictive power, but without any clear follow-up action: should the patient be referred, should a medication be initiated or its dose modified, should serial imaging be performed for closer surveillance? By analogy, consider the simple rule-based risk scores that are being routinely used in practice, such as the Wells score for pulmonary embolism or the CHADS-VASC score for stroke assessment in atrial fibrillation. These scores are useful because there are accepted pathways on how to act in response to a certain value—'traffic-light' recommendations about whether to perform a pulmonary angiogram or whether to initiate anticoagulation. Machine learning tools can be seen as a fancier version of these traditional clinical scoring systems, and be similarly tied to clinical actions.[2] One illustration was a recent study by de Fauw *et al* using deep learning for interpretation of optical coherence tomography scans. The algorithm segmented the scan and classified between multiple different pathologies, and provided a simple recommendation back to the clinician: urgent, semiurgent, routine referral or observation.[10] User-experience design ought to be considered as a fundamental part of any health machine learning pipeline—the way to merge an algorithm into the 'socio-technical' milieu of the clinic.[11]

SAFETY

Patient safety must also become a foundational part of model design. The medical community is familiar with the rigorous regulatory process for vetting new pharmaceuticals and medical devices; however the safety of algorithms remains a significant concern for clinicians and patients

alike. This mistrust is often pinned on issues such as interpretability (the 'black box' problem of inscrutable deep learning algorithms[12]) and external validity (will an algorithm trained on external data apply here?). The underlying problem is the lack of empirical evidence to prospectively demonstrate the safety and efficacy of an algorithm in a real-world setting.[13] By comparison, consider all the commonly used medications where the underlying mechanism is incompletely understood (eg, lithium), but which have been shown to be safe and effective with empirical evidence.

In order for an algorithm to achieve widespread use, we need empirical validation and a plan for ongoing algorithmic and technical resilience—that is, surveillance of a model's calibration and performance over time, and robust infrastructure to ensure system uptime, error handling, and so on. It is critical for model developers to engage with regulatory bodies, including institutional review boards and federal organisations like the Food and Drug Administration, which has already begun to build a framework for assessing clinical algorithms.[14] It is also increasingly important to consider additional dimensions of patient safety, such as protecting against algorithmic bias (will certain ethnic or socioeconomic groups be systematically disadvantaged by an algorithm trained on historical prejudices[15]) and model brittleness (given the recent evidence showing adversarial attacks on deep neural networks[16]). While no algorithm is without risk, appropriate risk mitigation and support for 'clinician-in-the-loop' systems will accelerate the translation of algorithms into true clinical benefit.[17 18] This framework should also to include the 'patient-in-the-loop'—that is, soliciting patient feedback on the design of algorithm deployments.

UTILITY

The capstone to any machine learning project should be a cost utility assessment. If we compare the situation of working without the algorithm's help to working with it, taking into account the clinical and financial consequences of false positives and negatives, do we estimate a significant reduction in either overall morbidity or cost?

Consider, for example, an algorithm to screen an EHR for undiagnosed cases of a rare disease, such as familial hypercholesterolaemia.[19] The cost utility assessment must take into account the savings (both financial and clinical) associated with early detection, balanced against the cost of a false-positive case being unnecessarily investigated and the costs of deployment and maintenance of the algorithm. This utility assessment should be conducted early in any machine learning project and continuously revised as models are deployed.

CONCLUSION

Current machine learning frameworks have greatly streamlined the process of model training, such that the creation of clinical algorithms is increasingly commoditised. To realise the full potential of these algorithms in improving quality of care, we must shift our focus to implementation and the practical issues of actionability, safety and utility.

This implementation checklist must be considered from the point of problem selection. Table 1 describes five template problems based on existing implementation examples. These templates may help identify use cases where machine learning can add value in a real-world clinical environment. Moving forward, there will be much to learn from the rich field of implementation science, which has developed frameworks for the design of complex health service interventions.[20]

The prospect of AI in healthcare has been described as a Rorschach blot on which we cast our technological aspirations.[21] In order to transform this nebulous form into a solid reality, we must now focus on bridging the implementation gap and safely bringing algorithms to the bedside.

Twitter Martin G Seneviratne @martin_sen

Contributors MGS, NHS and LC all participated in the drafting of the manuscript. MGS and NHS are joint first authors.

Funding MGS was supported by the John Monash Scholarship.

Competing interests MGS is presently an employee of DeepMind Health. This paper was drafted prior to employment and represents personal views only.

Table 1 Template use cases for machine learning implementations

Template	Description	Examples
Rarity	Screening for rare conditions where there is a significant clinical and economic benefit from early intervention.	Screening an EHR for undiagnosed cardiac amyloidosis,[22] familial hypercholesterolaemia,[19] or hand-foot-and-mouth disease.[23]
Urgency	Reducing delays in diagnosis or treatment by flagging high-acuity cases or commencing initial management.	Reordering the radiologist worklist to prioritise intracranial haemorrhage.[24] Automated triage of emergency presentations.[25]
Quantity	Dealing with high patient throughput by increasing the speed of clinicians and/or automating routine clinical tasks.	Summarising historical notes and identifying relevant clinical data.[26] Automated quantification of cardiac volumes on MRI.[27]
Quality	Monitoring care delivery to ensure quality benchmarks are met or flag medical errors.	Ensuring patients with high mortality risk receive a palliative care referral at an appropriate point in their admission.[28] Double-reading medical imaging to identify missed lesions.[29]
Complexity	Extending the capabilities of clinicians with advanced diagnostic or treatment decisions on par or exceeding subspecialists.	Reinforcement learning for dynamic treatment regimens.[30 31] Facial recognition for identification of rare genetic syndromes.[32]

EHR, electronic health record.

Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

ORCID iDs
Martin G Seneviratne http://orcid.org/0000-0003-0435-3738
Nigam H Shah http://orcid.org/0000-0001-9385-7158

REFERENCES

1 Rajpurkar P, Hannun AY, Haghpanahi M, *et al*. Cardiologist-Level Arrhythmia Detection with Convolutional Neural Networks [Internet]. arXiv [cs.CV], 2017. Available: http://arxiv.org/abs/1707.01836
2 Esteva A, Kuprel B, Novoa RA, *et al*. Dermatologist-level classification of skin cancer with deep neural networks. *Nature* 2017;542:115–8.
3 Steele AJ, Denaxas SC, Shah AD, *et al*. Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease. *PLoS One* 2018;13:e0202344.
4 Yu K-H, Beam AL, Kohane IS. Artificial intelligence in healthcare. *Nat Biomed Eng* 2018;2:719–31.
5 Mims C. The AI doctor will see you now. WSJ online. wsj.com, 2018. Available: https://www.wsj.com/articles/the-ai-doctor-will-see-you-now-1526817600 [Accessed 2 Dec 2018].
6 Langlotz C. Radiologists who use AI will replace rads who don't. In: Center for Artificial Intelligence in Medicine & Imaging [Internet], 2017. Available: https://aimi.stanford.edu/about/news/rsna-2017-rads-who-use-ai-will-replace-rads-who-don-t
7 Chen JH, Asch SM. Machine learning and prediction in medicine — beyond the peak of inflated expectations. *N Engl J Med* 2017;376:2507–9.
8 Robbins R, Feuerstein A, Garde D, *et al*. He hunted for gold-standard research on AI in medicine — and didn't find much - STAT. In: STAT [Internet], 2019. Available: https://www.statnews.com/2019/02/14/artificial-intelligence-medicine-eric-topol/ [Accessed 3 Mar 2019].
9 Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc* 2018;25:1419–28.
10 De Fauw J, Ledsam JR, Romera-Paredes B, *et al*. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat Med* 2018;24:1342–50.
11 Sittig DF, Singh H. A new sociotechnical model for studying health information technology in complex adaptive healthcare systems. *Qual Saf Health Care* 2010;19 Suppl 3:i68–74.
12 Knight W. The dark secret at the heart of AI. MIT technology review. MIT technology review, 2017. Available: https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/ [Accessed 2 Dec 2018].
13 Keane PA, Topol EJ. With an eye to AI and autonomous diagnosis. npj digital medicine. *Nature Publishing Group* 2018;1.
14 Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) [Internet]. US Food & Drug Administration, 2019. Available: https://www.fda.gov/media/122535/download
15 Gianfrancesco MA, Tamang S, Yazdany J, *et al*. Potential biases in machine learning algorithms using electronic health record data. *JAMA Intern Med* 2018;178:1544–7.
16 Finlayson SG, Chung HW, Kohane IS, *et al*. Adversarial Attacks Against Medical Deep Learning Systems [Internet]. arXiv [cs.CR], 2018. Available: http://arxiv.org/abs/1804.05296
17 Verghese A, Shah NH, Harrington RA. What this computer needs is a physician: Humanism and artificial intelligence. *JAMA* 2018;319:19–20.
18 Coiera E. First compute no harm - The BMJ. In: The BMJ [Internet], 2017. Available: https://blogs.bmj.com/bmj/2017/07/19/enrico-coiera-et-al-first-compute-no-harm/ [Accessed 3 Mar 2019].
19 Banda JM, Sarraju A, Abbasi F, *et al*. Finding missed cases of familial hypercholesterolemia in health systems using machine learning. *npj Digital Medicine* 2019;2.
20 Wensing M, Grol R. Knowledge translation in health: how implementation science could contribute more. *BMC Med* 2019;17:88.
21 Coiera E. The fate of medicine in the time of AI. *The Lancet* 2018;392:2331–2.
22 Garg R, Dong S, Shah S, *et al*. A Bootstrap Machine Learning Approach to Identify Rare Disease Patients from Electronic Health Records [Internet]. *arXiv [cs.LG]* 2016.
23 Liu G, Xu Y, Wang X, *et al*. Developing a machine learning system for identification of severe hand, foot, and mouth disease from electronic medical record data. *Sci Rep* 2017;7:16341.
24 Arbabshirani MR, Fornwalt BK, Mongelluzzo GJ, *et al*. Advanced machine learning in action: identification of intracranial hemorrhage on computed tomography scans of the head with clinical workflow integration. *npj Digital Medicine* 2018;1.
25 Levin S, Toerper M, Hamrock E, *et al*. Machine-Learning-Based Electronic Triage More Accurately Differentiates Patients With Respect to Clinical Outcomes Compared With the Emergency Severity Index. *Ann Emerg Med* 2018;71:565–74.
26 Pivovarov R, Elhadad N. Automated methods for the summarization of electronic health records: table 1. *J Am Med Inform Assoc* 2015;22:938–47.10.1093/jamia/ocv032
27 Narang A, Mor-Avi V, Prado A, *et al*. Machine learning based automated dynamic quantification of left heart chamber volumes. *Eur Heart J Cardiovasc Imaging* 2019;20:541–9.
28 Avati A, Jung K, Harman S, *et al*. Improving Palliative Care with Deep Learning [Internet]. *arXiv [cs.CY]* 2017.
29 Venkatesh SS, Levenback BJ, Sultan LR, *et al*. Going beyond a first reader: a machine learning methodology for optimizing cost and performance in breast ultrasound diagnosis. *Ultrasound Med Biol* 2015;41:3148–62.
30 Komorowski M, Celi LA, Badawi O, *et al*. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nat Med* 2018;24:1716–20.
31 Nemati S, Ghassemi MM, Clifford GD. Optimal medication dosing from suboptimal clinical examples: a deep reinforcement learning approach. *Conf Proc IEEE Eng Med Biol Soc* 2016;2016:2978–81.
32 Gurovich Y, Hanani Y, Bar O, *et al*. DeepGestalt - Identifying Rare Genetic Syndromes Using Deep Learning [Internet]. *arXiv [cs.CV]* 2018.