Bridging the implementation gap of machine learning in healthcare

Martin G Seneviratne, Nigam H Shah, Larry Chu

Applications of machine learning on clinical data are now attaining levels of performance that match or exceed human clinicians. 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. The looming extinction of doctors has captured the public imagination, with editorials such as ‘The AI Doctor Will See You Now’. The prevailing view among experts is more balanced: that doctors who use artificial intelligence (AI) will replace those who do not.

Amid such inflated expectations, the elephant in the room is the implementation gap of machine learning in healthcare. 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’. 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.

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, semi-urgent, routine referral or observation. 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.

**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.
Consider, for example, an algorithm to screen an EHR for undiagnosed cases of a rare disease, such as familial hypercholesterolaemia. 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.

The prospect of AI in healthcare has been described as a Rorschach blot on which we cast our technological aspirations. 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.

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, familial hypercholesterolaemia, or hand-foot-and-mouth disease. |
| Urgency  | Reducing delays in diagnosis or treatment by flagging high-acuity cases or commencing initial management. | Reordering the radiologist worklist to prioritise intracranial haemorrhage. Automated triage of emergency presentations. |
| 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. Automated quantification of cardiac volumes on MRI. |
| 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. Double-reading medical imaging to identify missed lesions. |
| Complexity | Extending the capabilities of clinicians with advanced diagnostic or treatment decisions on par or exceeding subspecialists. | Reinforcement learning for dynamic treatment regimens. Facial recognition for identification of rare genetic syndromes. |

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.

COMMENTS
Patient consent for publication Not required.

Provenance and peer review Not commissioned; externally peer reviewed.

Open access This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: http://creativecommons.org/licenses/by-nc/4.0/.

ORCID iDs
Martin G Seneviratne http://orcid.org/0000-0003-0435-3738
Nigam 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 radiologists 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, Ash 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 in the Amazon. Then he hunted for gold — and found something else. STAT [Internet], 2017. 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, Ledsm 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, Topel 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/encecoiera-et-al-first-compute-no-harm/ [Accessed 3 Mar 2019].
19 Banda JM, Sarraja 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, Toopser 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 Pisvoyarov 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.CV] 2017.
29 Venkatess S, 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.