Bridging the implementation gap of machine learning in healthcare

Martin G Seneviratne ¹, ¹ Nigam H Shah ¹, ¹ Larry Chu²

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 action-able—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. ¹⁰ 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) 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. 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. 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) and model brittleness (given the recent evidence showing adversarial attacks on deep neural networks). 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.

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. 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.

**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>
<thead>
<tr>
<th>Table 1 Template use cases for machine learning implementations</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Template</strong></td>
</tr>
<tr>
<td>Rarity</td>
</tr>
<tr>
<td>Urgency</td>
</tr>
<tr>
<td>Quantity</td>
</tr>
<tr>
<td>Quality</td>
</tr>
<tr>
<td>Complexity</td>
</tr>
</tbody>
</table>

EHR, electronic health record.
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
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