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 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.

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 the input mammograms—devised to distinguish malignant from benign tissues—without the need for manual review by radiologists. This was possible because the algorithm was trained on and examined images that had been reviewed by experts, so it could learn the ‘socio-technical’ milieu of the clinical setting.

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.
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>
<thead>
<tr>
<th>Template</th>
<th>Description</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rarity</td>
<td>Screening for rare conditions where there is a significant clinical and economic benefit from early intervention.</td>
<td>Screening an EHR for undiagnosed cardiac amyloidosis,22 familial hypercholesterolaemia,19 or hand-foot-and-mouth disease.23</td>
</tr>
<tr>
<td>Urgency</td>
<td>Reducing delays in diagnosis or treatment by flagging high-acuity cases or commencing initial management.</td>
<td>Reordering the radiologist worklist to prioritise intracranial haemorrhage.14 Automated triage of emergency presentations.25</td>
</tr>
<tr>
<td>Quantity</td>
<td>Dealing with high patient throughput by increasing the speed of clinicians and/or automating routine clinical tasks.</td>
<td>Summarising historical notes and identifying relevant clinical data.26 Automated quantification of cardiac volumes on MRI.27</td>
</tr>
<tr>
<td>Quality</td>
<td>Monitoring care delivery to ensure quality benchmarks are met or flag medical errors.</td>
<td>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</td>
</tr>
<tr>
<td>Complexity</td>
<td>Extending the capabilities of clinicians with advanced diagnostic or treatment decisions on par or exceeding subspecialists.</td>
<td>Reinforcement learning for dynamic treatment regimens30 31 Facial recognition for identification of rare genetic syndromes.32</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