Improving the quality of machine learning in health applications and clinical research

Improving the quality of machine learning in health applications and clinical research

For machine learning developers, the use of prediction tools in real-world clinical settings can be a distant goal. Recently published guidelines for reporting clinical research that involves machine learning will help connect clinical and computer science communities, and realize the full potential of machine learning tools.

Bilal A. Mateen, James Liley, Alastair K. Denniston, Chris C. Holmes and Sebastian J. Vollmer

In the past decade, many impressive results have been reported for machine learning (ML) tools that are developed to assist in clinical decision making. However, translating these largely theoretical achievements to realistic settings remains a formidable challenge and requires active collaboration between ML researchers and healthcare experts. An important step towards facilitating this process is the extension of reporting guidelines in clinical and health sciences to incorporate ML and artificial intelligence (AI) approaches[1]. Recent examples are the SPIRIT-AI[2] and CONSORT-AI[3] checklists for reporting of clinical trials that involve AI methods, based on the original SPIRIT and CONSORT guidelines and drawing on pre-existing initiatives to support transparent reporting of ML model developments such as the TRIPOD[4] guidelines. Most recently there has been a focus on developing an extension for the reporting of diagnostic tests involving AI[1] — immediately relevant for clinical practice as seen in the current pandemic. While such checklists may appear cumbersome or restrictive, we argue in this Comment that they are essential to make full use of the promise of ML methods in healthcare and to facilitate eventual real-world applications. We invite the ML community to take an active role in the ongoing development of these guidelines.

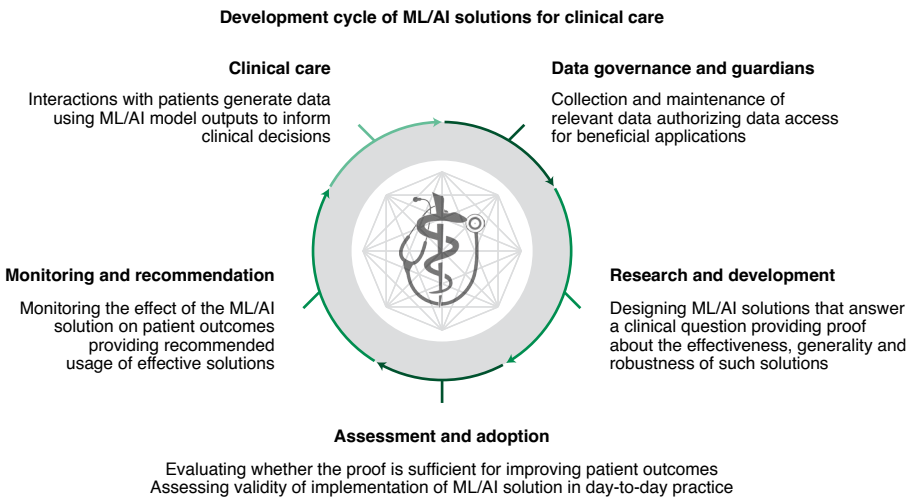

Fig. 1 | Development cycle of ML/AI solutions for clinical care. This figure is reproduced from a piece describing 20 questions that should be asked prior to initiating an MLH project[19]. It summarizes the stages relevant for and affected by ML methods and their adoption in clinical care. For most projects this process will not be directly from start to finish, rather requiring a continuous feedback loop between the different stakeholders responsible for the various tasks. The purpose of the illustration is to show the high-level steps between getting access to data, providing examinable proof that an algorithm indeed accomplishes the stated goals, and translating the tool into clinical practice. Note that the implementation of any new algorithm is likely to influence the data routinely collected, by assigning increased importance in those features that are used as model inputs, while generating new information (that is, the model outcomes), thus completing the development cycle, and initiating a new cycle.

Increasing value and reducing waste in ML research

Machine learning in healthcare (MLH) generally aims to predict some clinical outcome on the basis of multiple predictors. The potential of MLH is vast, with demonstrations of ML-based tools being able to achieve human-level or above diagnostic and prognostic capabilities having been described in almost every clinical specialty[5]. However, the number of ML tools adopted in clinical applications reflects only a fraction of the investment into the field as a whole, suggesting that most applications of MLH have not progressed beyond the initial publication[6]. On closer examination, there appears to be a tendency for ML researchers to stop once adequately accurate prediction (and hence novelty) has been demonstrated[6] with translation into the clinical practice left to interested domain experts[7]. This translation is difficult, with a high failure rate[6,8], but some of the potential difficulties may be readily addressable. Below we explore how the use of clinical reporting guidelines (and research frameworks more generally) can facilitate better translation and more clinically useful applications of MLH.

The role of reporting guidelines

One contributor to the complexity of translational research in MLH is ambiguous and incomplete reporting[9]. This issue is neither new nor intractable, but in the absence of complete information, it is difficult for end-users to objectively assess the strengths and weaknesses of any potential tool. This has been typified by reporting of predictive models during the COVID-19 pandemic: a recent systematic review[10] found that among the over 100 published prediction models for COVID-19 risk, all had shortcomings in reporting that precluded assessment of their clinical suitability. Paired demonstrations from

subsequent validation studies have found simple bedside measurements outperform many of these models[11]. It is likely that at least some research waste could be avoided by full and transparent reporting, which helps to avoid pursuing predictive models that are unlikely to reach clinical use. It seems reasonable to expect that as a minimum standard, model reporting should include an unambiguous description of the entire statistical modelling pipeline, with the details of cohort derivation, pre-processing procedures (for example, missing data handling) and model building. Notably, many of the reporting-specific issues originally noted in the non-ML literature continue to be identified in more recent MLH studies[12,13], suggesting that this is an issue that the community as a whole needs to address.

As well as potentially helping to avoid wasteful research, reporting guidelines serve an important purpose in clinical translation of research by identifying common elements of the research content that are necessary for verification of the method (usually mathematical details, model choices, and any new statistical methods), and those needed for reproducibility (generally data processing and analysis pipelines)[2–4]. Implicit in the requirement of reproducibility is a minimum expected standard for the model development pipeline and ongoing research. Although the literature on methodological best practices in prediction model building is vast, complicated even for those well versed in the underlying mathematical processes, and at times contradictory, the TRIPOD and forthcoming TRIPOD-ML guidelines provide a broad outline of reporting elements necessary (though not sufficient) for translatable MLH[14]. Outside of MLH, there is evidence to suggest that the introduction of reporting standards has generally led to improved confidence in study findings and enabled better decisions regarding interventions under evaluation[15].

No success without an appreciation of context

A second issue is the absence of a clear pathway to translation, one in which MLH advances can be efficiently and effectively evaluated to the point where clinicians, regulators and policymakers can be confident that they are ready for safe deployment into real-world health contexts. Extending a predictive model to a clinical setting is much more complex than developing a predictor in a reductionist research context. Examples of issues that need to be considered are how the MLH output — the prediction — will be integrated with other clinical information, and how it will be presented (for example, as a measure of probability or a direct translation into recommendations on selecting treatments). Addressing such issues is fundamentally necessary for the research to contribute to medical practice or at least to proceed to clinical trials. This requires researchers and end-users to work together to build a shared understanding within each community of the priorities and limitations of the other, thereby improving the ability to create ML solutions that make real-world impact. The important function of clinical ML frameworks is therefore to identify core issues of design and delivery that would be considered important for a high-quality study that is supporting the development of a clinical application[16].

For MLH developers, the potential use of a tool in a clinical context can be a distant goal, which is likely to contribute to the disconnect between the formulation of a prediction task and its corresponding clinical workflow. One way in which this often manifests is that without an awareness of the clinical context of the prediction task, it cannot be guaranteed that the experimental setting has a clinical homologue, in which case the research may be wasted. For example, there are many examples of computer-vision-based tools that seek to diagnose a single pathological entity in a radiograph[5]. However, radiologists are rarely interested solely in the diagnosis of a single pre-specified pathology in the image they are presented with. The technical contributions of developing a novel computer-vision-based approach or tool for diagnosing a single pathology are not wasteful but may be judged as such by clinicians due to lack of explicit accounting of how the research maps to the development cycle (Fig. 1). In general, while theoretical and technical contributions using clinical data to illustrate applicability are fundamental to the progress of the field, they are by nature different to attempts to create a prediction model for clinical practice.

Using the ends to justify the means

The purpose of research is not publication, but rather to improve the way in which systems and societies function. Research in MLH tends to focus on theoretical novelty and on improved methodological performance on benchmark datasets[17,18], which, while important, may be to the detriment of focus on application of methods. Indeed, we fundamentally lack case studies that demonstrate how the prospective use of MLH can substantially improve patient care. However, recent developments in the MLH literature may help address this issue. For example, the recent updates to the guidelines for randomized control trials and the development of diagnostic tools to explicitly address the nuances introduced by ML are in effect a checklist for developers[1,2]; they provide a roadmap of the expectations of any tool that may eventually make it into routine clinical practice, and hence are a valuable resource when designing a prediction modelling workflow. However, it is important to recognize that even for successfully translated work, reporting may fall short of the emerging ML reporting guidelines. These reporting guidelines are not benchmarks to retrospectively criticize such studies, but rather tools to provide a framework for study design, delivery and reporting to ensure that the medical community can derive the greatest value possible. Admittedly, in discussing the relevance of these tools to the ML community, we note that the uptake of the original versions of these frameworks in the MLH community has been low, which prompted the more recent AI-specific updates[14]. These updated guidelines should be seen as a starting point. Our hope is that by enfranchising the ML community and creating tools that are more suitable for ML-based interventions, groups will develop a sense of ownership of these frameworks, and thus the development of future iterations will serve as an opportunity for the MLH community, clinicians, patients, trialists, regulators and other stakeholders to work together to continue to develop and iterate these frameworks to be fit for purpose.

Conclusion

Culture change is difficult, but for ML as a field to have the impact that we believe is achievable, we must transition away from siloed working. MLH cannot exist separately in the clinical and mathematical communities if we are to realize its full potential. Reporting guidelines and MLH frameworks represent a key step in bridging these two (often) separate worlds. ❐

Bilal A. Mateen[1,2], James Liley[1,3], Alastair K. Denniston[4,5], Chris C. Holmes[1,6] and Sebastian J. Vollmer[1,7] ✉

[1]The Alan Turing Institute, London, UK. [2]Wellcome Trust, London, UK. [3]Medical Research Council Human Genetics Unit, Institute for Genetics and Molecular Medicine, University of Edinburgh, Edinburgh, UK. [4]Health Data Research UK, London, UK. [5]Department of Ophthalmology, University Hospitals Birmingham NHS Foundation Trust, Birmingham, UK. [6]Department of Statistics, University of Oxford, Oxford, UK. [7]Department of Statistics, University of Warwick, Coventry, UK.
✉*e-mail: svollmer@turing.ac.uk*

Published online: 2 October 2020
https://doi.org/10.1038/s42256-020-00239-1

References

1. Sounderajah, V. et al. *Nat. Med.* **26**, 807–808 (2020).
2. Liu, X. et al. *Nat. Med.* **26**, 1364–1374 (2020).
3. Cruz Rivera, S. et al. *Nat. Med.* **26**, 1351–1363 (2020).
4. Moons, K. G. M. et al. *Ann. Intern. Med.* **162**, W1–W73 (2015).
5. Topol, E. J. *Nat. Med.* **25**, 44–56 (2019).
6. Ben-Israel, D. et al. *Artif. Intell. Med.* **103**, 101785 (2020).
7. Chalmers, I. & Glasziou, P. *Lancet* **374**, 86–89 (2009).
8. Rajkomar, A., Dean, J. & Kohane, I. *N. Engl. J. Med.* **380**, 1347–1358 (2019).
9. Liu, X. et al. *Lancet Digit. Health* **1**, e271–e297 (2019).
10. Wynants, L. et al. *BMJ* **369**, m1328 (2020).
11. Gupta, R. K. et al. Preprint at https://doi.org/10.1101/2020.07.24.20149815 (2020).
12. Király, F. J., Mateen, B. & Sonabend, R. Preprint at https://arxiv.org/abs/1812.07519 (2018).
13. Freiman, J. A., Chalmers, T. C., Smith, H. & Kuebler, R. R. in *Medical Uses of Statistics* (eds Bailar, J. C. III & Mosteller, F.) 357–373 (NEJM, 1992).
14. Collins, G. S. & Moons, K. G. M. *Lancet* **393**, 1577–1579 (2019).
15. Glasziou, P. et al. *Lancet* **383**, 267–276 (2014).
16. Vollmer, S. et al. *BMJ* **368**, l6927 (2020).
17. Lipton, Z. C. & Steinhardt, J. Preprint at https://arxiv.org/abs/1807.03341 (2018).
18. Wagstaff, K. L. in *Proc. 29th Int. Conf. Machine Learning* 529–536 (ICML, 2012).
19. Vollmer, S. et al. Preprint at https://arxiv.org/abs/1812.10404 (2018).

Competing interests

The authors declare no competing interests.