Equity in essence: a call for operationalising fairness in machine learning for healthcare

Judy Wawira Gichoya,1,2 Liam G McCoy,3 Leo Anthony Celi,4,5,6 Marzyeh Ghassemi7,8,9

INTRODUCTION
Machine learning for healthcare (MLHC) is at the juncture of leaping from the pages of journals and conference proceedings to clinical implementation at the bedside. Succeeding in this endeavour requires the synthesis of insights from both the machine learning and healthcare domains, in order to ensure that the unique characteristics of MLHC are leveraged to maximise benefits and minimise risks. An important part of this effort is establishing and formalising processes and procedures for characterising these tools and assessing their performance. Meaningful progress in this direction can be found in recently developed guidelines for the development of MLHC models,1 guidelines for the design and reporting of MLHC clinical trials,2 3 and protocols for the regulatory assessment of MLHC tools.4 5

But while such guidelines and protocols engage extensively with relevant technical considerations, engagement with issues of fairness, bias and unintended disparate impact is lacking. Such issues have taken on a place of prominence in the broader ML community,6–9 with recent work highlighting issues such as racial disparities in the accuracy of facial recognition and gender classification software,6 10 gender bias in the output of natural language processing models11 12 and racial bias in algorithms for bail and criminal sentencing.13 MLHC is not immune to these concerns, as seen in disparate outcomes from algorithms for allocating healthcare resources,14 15 bias in language models developed on clinical notes16 and melanoma detection models developed primarily on images of light-coloured skin.17 Within this paper, we will examine the inclusion of fairness in recent guidelines for MLHC model reporting, clinical trials and regulatory approval. We highlight opportunities to ensure that fairness is made fundamental to MLHC, and examine ways how this can be operationalised for the MLHC context.

FAIRNESS AS AN AFTERTHOUGHT?

Model development and trial reporting guidelines
Several recent documents have attempted, with varying degrees of practical implication, to enumerate guiding principles for MLHC. Broadly, these documents do an excellent job of highlighting artificial intelligence (AI)-specific technical and operational concerns, such as how to handle human-AI interaction, or how to account for model performance errors. Yet as outlined in table 1, references to fairness are either conspicuously absent, made merely in passing, or relegated to supplemental discussion.

Notable examples are the recent the Standard Protocol Items: Recommendations for Interventional Trials-AI (SPIRIT-AI)2 and Consolidated Standards of Reporting Trials-AI (CONSORT-AI)3 extensions, which expand prominent guidelines for the design and reporting of AI clinical trials to include concerns relevant to AI. While the latter states in the discussion that ‘investigators should also be encouraged to explore differences in performance and error rates across population subgroups’,3 there is no more formal inclusion of the concept into the guideline itself. Similarly, the announcement papers for the upcoming Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis-Machine Learning (TRIPOD-ML)18 and Standards for Reporting of Diagnostic Accuracy Studies AI Extension (STARD-AI)19 guidelines for model reporting do not allude to these issues (though we wait in anticipation for their potential inclusion in the final versions of these guidelines). While recently published guidelines from the editors of...
OPERATIONALISING FAIRNESS IN MLHC PRACTICE

If fairness is an afterthought in the design and reporting of MLHC papers and trials, as well as regulatory processes, it is likely to remain an afterthought in the development and implementation of MLHC tools. If MLHC is going to prove effective for—and be trusted by—a diverse range of patients, fairness cannot be a post-hoc and after-the-fact consideration. Nor is it sufficient for fairness to be a vague abstraction of academic importance but ineffectual consequence. The present moment affords a tremendous opportunity to define MLHC such that fairness is integral, and to ensure that this commitment is reflected in model reporting guidelines, clinical trial guidelines and regulatory approaches. However, moving from vague commitments of fairness to practical and effective guidance is far from a trivial task. As work in the machine learning community has demonstrated, fairness has multiple definitions which can occasionally be incompatible and bias can arise from a complex range of sources. Operationalisation of fairness must be context-specific, and embeds the relevant values in a field. We call for concerted effort from the MLHC community, and in particular the groups responsible for the development and propagation of guidelines, to affirm a commitment to fairness in an explicit and operationalised fashion. Similarly, we call on the various regulatory agencies to establish clear minimum standards for AI fairness. In box 1, we highlight a non-exhaustive series of recommendations that are likely to be beneficial as the MLHC community engages in this endeavour.

Table 1 Fairness in recently released and upcoming guidelines

<table>
<thead>
<tr>
<th>Guideline</th>
<th>How is fairness included?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals</td>
<td>Discussion of the risk of unfairness is included in <a href="https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7167122/bcn/ccm-48-0623-s001.docx">https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7167122/bcn/ccm-48-0623-s001.docx</a> but not the main document.</td>
</tr>
<tr>
<td>TRIPOD-ML (Announcement Statement Only)</td>
<td>No explicit mention.</td>
</tr>
<tr>
<td>STARD-AI (Announcement Statement Only)</td>
<td>No explicit mention.</td>
</tr>
<tr>
<td>Checklist for Artificial Intelligence in Medical Imaging</td>
<td>Bias discussed, but not clearly in the context of fairness with respect to differential performance or impact between patient groups.</td>
</tr>
<tr>
<td>Clinical Trial Guidelines</td>
<td>Fairness is brought up in the discussion section but not included explicitly in any of the guideline checklist points.</td>
</tr>
<tr>
<td>CONSORT-AI Extension</td>
<td>No explicit mention.</td>
</tr>
<tr>
<td>CONSORT-AI, Consolidated Standards of Reporting Trials–Artificial Intelligence; SPIRIT-AI, Standard Protocol Items: Recommendations for Interventional Trials–Artificial Intelligence; STARD-AI, Standards for Reporting of Diagnostic Accuracy Studies–Artificial Intelligence; TRIPOD-ML, Transparent Reporting of a multivariable prediction model for Individual Prognosis or Diagnosis–Machine Learning.</td>
<td></td>
</tr>
</tbody>
</table>

Respiratory, sleep and critical care medicine journals engage with the concept in an exemplary fashion, the depth of their discussion is relegated to a supplementary segment of the paper.

Regulatory guidance

Broadly, the engagement of prominent regulatory bodies with MLHC remains at a preliminary stage, and engagement with fairness tends to be either minimal or vague. The Food and Drug Administration in the USA has made significant strides towards modernisation of its frameworks for the approval and regulation of software-based medical interventions, including MLHC tools. Their documents engage broadly with technical concerns, and criteria for effective clinical evaluation, but entirely lack discussion of fairness or the relationship between these tools and the broader health equity context. The Canadian Agency for Drugs and Technologies in Health has explicitly highlighted the need for fairness and bias to be considered, but further elaboration is lacking.

The work of the European Union on this topic remains at a broad stage. While their documents do make reference to principles of ‘diversity, non-discrimination and fairness’, they do so in a very broad manner without any clearly operationalised specifics. The engagement of the UK with MLHC is relatively advanced, with several prominent reports engaging with the topic and an explicit ‘Code of Conduct for Data-Driven Healthcare Technology’ from the Department of Health and Social Care that highlights the need for fairness. However, the specifics of this regulatory approach are still being decided, and no clear guidance has yet been put forth to clarify these principles in practice. MLHC as a whole would benefit from increased clarity and force in regulatory guidance from these major agencies.

Recommendations

- Engage members of the public and in particular members of marginalised communities in the process of determining acceptable fairness standards.
- Collect necessary data on vulnerable protected groups in order to perform audits of model function (eg, on race, gender).
- Analyse and report model performance for different intersectional subpopulations at risk of unfair outcomes.
- Establish target thresholds and maximum disparities for model function between groups.
- Be transparent regarding the specific definitions of fairness that are used in the evaluation of a machine learning for healthcare (MLHC) model.
- Explicitly evaluate for disparate treatment and disparate impact in MLHC clinical trials.
- Commit to postmarketing surveillance to assess the ongoing real-world impact of MLHC models.
CONCLUSION

Values are embedded throughout the MLHC pipeline, from the design of models, to the execution and reporting of trials, to the regulatory approval process. Guidelines hold significant power in defining what is worthy of emphasis. While fairness is essential to the impact and consequences of MLHC tools, the concept is often conspicuously absent or ineffectually vague in emerging guidelines. The field of machine MLHC has the opportunity at this juncture to render fairness integral to the identity field. We call on the MLHC community to commit to the project of operationalising fairness, and to emphasise fairness as a requirement in practice.

Author affiliations
1Department of Radiology & Imaging Sciences, Emory University, Atlanta, Georgia, USA
2Fogarty International Center, National Institutes of Health (NIH), Bethesda, Maryland, USA
3Temeny Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
4Laboratory for Computational Physiology, Harvard-MIT Division of Health Sciences and Technology, Cambridge, Massachusetts, USA
5Division of Pulmonary Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
6Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, USA
7Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
8Department of Medicine, University of Toronto, Toronto, Ontario, Canada
9Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada

Twitter Judy Wawira Gichoya @judywawira, Liam G McCoy @l iamgmc coy and Leo Anthony Cell @MITCriticalData

Contributors Initial conceptions and design: JWG, LGM, MG and LAC. Drafting of the paper: LGM, JWG, MG and LAC. Critical revision of the paper for important intellectual content: JWG, LGM, MG and LAC.

Funding Division of Electrical, Communications and Cyber Systems (1928481), National Institute of Biomedical Imaging and Bioengineering (EB017205).

Competing interests MG acts as an advisor to Radical Ventures in Toronto.

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 iD Leo Anthony Cell http://orcid.org/0000-0001-6712-6626

REFERENCES