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

BMJ Health & Care Informatics

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 Ghassemi[7,8,9]

To cite: Wawira Gichoya J, McCoy LG, Celi LA, *et al*. Equity in essence: a call for operationalising fairness in machine learning for healthcare. *BMJ Health Care Inform* 2021;**28**:e100289. doi:10.1136/bmjhci-2020-100289

Received 22 November 2020
Revised 07 February 2021
Accepted 09 February 2021

For numbered affiliations see end of article.

Correspondence to
Liam G McCoy;
liam.mccoy@mail.utoronto.ca

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 models[11 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 notes[16] 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-ML (TRIPOD-ML)[18] andStandards 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

Table 1 Fairness in recently released and upcoming guidelines

Guideline	How is fairness included?
Reporting guidelines	
Development and Reporting of Prediction Models: Guidance for Authors From Editors of Respiratory, Sleep, and Critical Care Journals[1]	Discussion of the risk of unfairness is included in https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7161722/bin/ccm-48-0623-s001.docx but not the main document.
TRIPOD-ML (Announcement Statement Only)[18]	No explicit mention.
STARD-AI (Announcement Statement Only)[19]	No explicit mention.
Checklist for Artificial Intelligence in Medical Imaging[31]	Bias discussed, but not clearly in the context of fairness with respect to differential performance or impact between patient groups.
Clinical Trial Guidelines	
CONSORT-AI Extension[3]	Fairness is brought up in the discussion section but not included explicitly in any of the guideline checklist points.
SPIRIT-AI Extension[2]	No explicit mention.

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.

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.[1]

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.[5] 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.[20] 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.[21]

The work of the European Union on this topic remains at a broad stage.[4] 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.[22 23] The engagement of the UK with MLHC is relatively advanced, with several prominent reports engaging with the topic,[24–26] and an explicit 'Code of Conduct for Data-Driven Healthcare Technology'[27] 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.[28] MLHC as a whole would benefit from increased clarity and force in regulatory guidance from these major agencies.[29]

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,[7] and bias can arise from a complex range of sources.[30] 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.

Box 1 Recommendations for operationalising fairness

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

Twitter Judy Wawira Gichoya @judywawira, Liam G McCoy @liamgmccoy and Leo Anthony Celi @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.

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

REFERENCES

1 Leisman DE, Harhay MO, Lederer DJ, *et al*. Development and reporting of prediction models: guidance for authors from editors of respiratory, sleep, and critical care journals. *Crit Care Med* 2020;48:623–33.
2 Cruz Rivera S, Liu X, Chan A-W, *et al*. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the SPIRIT-AI extension. *Nat Med* 2020;26:1351–63.
3 Liu X, Cruz Rivera S, Moher D, *et al*. Reporting guidelines for clinical trial reports for interventions involving artificial intelligence: the CONSORT-AI extension. *Nat Med* 2020;26:1364–74.
4 Cohen IG, Evgeniou T, Gerke S, *et al*. The European artificial intelligence strategy: implications and challenges for digital health. *Lancet Digit Health* 2020;2:e376–9.
5 FDA. Artificial intelligence and machine learning in software as a medical device, 2020. Available: https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device [Accessed 11 Oct 2020].
6 Buolamwini J, Gebru T. Gender shades: intersectional accuracy disparities in commercial gender classification. *Conference on Fairness, Accountability and Transparency*, 2018:77–91.
7 Gajane P, Pechenizkiy M. On formalizing fairness in prediction with machine learning. arXiv:171003184 [cs, stat], 2018. Available: http://arxiv.org/abs/1710.03184 [Accessed 20 Sept 2020].
8 et alMehrabi N, Morstatter F, Saxena N. A survey on bias and fairness in machine learning. arXiv:190809635 [cs], 2019. Available: http://arxiv.org/abs/1908.09635 [Accessed 11 Oct 2020].
9 De-Arteaga M, Romanov A, Wallach H. Bias in bios: a case study of semantic representation bias in a High-Stakes setting. *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19. Published online*, 2019:120–8.
10 Klare BF, Burge MJ, Klontz JC, *et al*. Face recognition performance: role of demographic information. *IEEE Transactions on Information Forensics and Security* 2012;7:1789–801.
11 Caliskan A, Bryson JJ, Narayanan A. Semantics derived automatically from language corpora contain human-like biases. *Science* 2017;356:183–6.
12 Bordia S, Bowman SR. Identifying and reducing gender bias in word-level language models. arXiv:190403035 [cs], 2019. Available: http://arxiv.org/abs/1904.03035 [Accessed 30 Jan 2021].
13 Huq AZ. Racial equity in algorithmic criminal justice. *Duke LJ* 2018;68:1043.
14 Obermeyer Z, Powers B, Vogeli C, *et al*. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019;366:447–53.
15 Benjamin R. Assessing risk, automating racism. *Science* 2019;366:421–2.
16 Zhang H, AX L, Abdalla M. Hurtful words: quantifying biases in clinical contextual word embeddings. *Proceedings of the ACM Conference on Health, Inference, and Learning. CHIL '20. Association for Computing Machinery*, 2020:110–20.
17 Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018;154:1247–8.
18 Collins GS, Moons KGM. Reporting of artificial intelligence prediction models. *Lancet* 2019;393:1577–9.
19 Sounderajah V, Ashrafian H, Aggarwal R, *et al*. Developing specific reporting guidelines for diagnostic accuracy studies assessing AI interventions: the STARD-AI steering group. *Nat Med* 2020;26:807–8.
20 Ferryman K. Addressing health disparities in the food and drug administration's artificial intelligence and machine learning regulatory framework. *J Am Med Inform Assoc* 2020;27:2016–9.
21 Mason A, Morrison A, Visintini S. *An overview of clinical applications of artificial intelligence*. Ottawa: CADTH, 2018.
22 Commission E. *COM(2019) 168 final: building trust in human-centric artificial intelligence*, 2019.
23 Commission E. *White paper on artificial intelligence–a European approach to excellence and trust*, 2020.
24 Tankelevitch L, Ahn A, Paterson R. *Advancing AI in the NHS*, 2018.
25 Fenech M, Strukelj N, Buston O. *Ethical, social, and political challenges of artificial intelligence in health*. London: Wellcome Trust Future Advocacy, 2018.
26 Topol E. *The Topol review: preparing the healthcare workforce to deliver the digital future*. Health Education England, 2019.
27 Department of Health and Social Care. Code of conduct for data-driven health and care technology, 2019. Available: https://www.gov.uk/government/publications/code-of-conduct-for-data-driven-health-and-care-technology/initial-code-of-conduct-for-data-driven-health-and-care-technology [Accessed 1 Aug 2020].
28 NHS. Regulating AI in health and care - Technology in the NHS. Available: https://healthtech.blog.gov.uk/2020/02/12/regulating-ai-in-health-and-care/ [Accessed 12 Oct 2020].
29 Parikh RB, Obermeyer Z, Navathe AS. Regulation of predictive analytics in medicine. *Science* 2019;363:810–2.
30 Suresh H, Guttag JV. A framework for understanding unintended consequences of machine learning. arXiv:190110002 [cs, stat], 2020. Available: http://arxiv.org/abs/1901.10002 [Accessed 20 Sept 2020].
31 Mongan J, Moy L, Kahn CE. Checklist for artificial intelligence in medical imaging (claim): a guide for authors and reviewers. *Radiology* 2020;2:e200029.