BMJ Health & Care Informatics

Ensuring machine learning for healthcare works for all

Liam G McCoy,[1,2] John D Banja,[3] Marzyeh Ghassemi,[4,5,6] Leo Anthony Celi [7,8,9]

To cite: McCoy LG, Banja JD, Ghassemi M, *et al*. Ensuring machine learning for healthcare works for all. *BMJ Health Care Inform* 2020;**27**:e100237. doi:10.1136/bmjhci-2020-100237

Received 09 September 2020
Revised 10 October 2020
Accepted 02 November 2020

Check for updates

© Author(s) (or their employer(s)) 2020. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

[1]Temerty Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada
[2]Institute of Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada
[3]Emory Center for Ethics, Emory University, Atlanta, Georgia, USA
[4]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada
[5]Department of Medicine, University of Toronto, Toronto, Ontario, Canada
[6]Vector Institute for Artificial Intelligence, Toronto, Ontario, Canada
[7]Laboratory for Computational Physiology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA
[8]Division of Pulmonary, Critical Care and Sleep Medicine, Beth Israel Deaconess Medical Center, Boston, Massachusetts, USA
[9]Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, Massachusetts, United States

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

INTRODUCTION

Machine learning, data science and artificial intelligence (AI) technology in healthcare (herein collectively referred to as machine learning for healthcare (MLHC)) is positioned to have substantial positive impacts on healthcare, enhancing progress in both the acquisition of healthcare knowledge and the implementation of this knowledge in numerous clinical contexts. However, there are concerns that have been identified with these technologies regarding their potential for negative impacts.[1–7] In particular that they may damage health equity by either introducing novel biases, or uncritically reproducing and magnifying existing systemic disparities. These concerns have led to a growth of scholarship on the intersection of ethics, AI and healthcare,[1–7] as well as significant restrictions on the use of patient data for MLHC research.[8 9]

Unfortunately, modern healthcare is already rife with treatments that fail to live up to evidentiary scrutiny,[10] while evidence behind their use is riddled with biases that further deepen health inequities.[11] Against this backdrop, it becomes clear that urgent and substantial change is needed, and that MLHC offers one of the most promising avenues toward achieving this end. Ethical concerns regarding the impact of this technology should be addressed and made foundational to the development of MLHC in meaningful ways. However, those concerns must not act to affect the field in a manner that perpetuates the structural inequalities that presently exist.

Through the conceptual lens of MLHC, this paper will explore various flaws of healthcare's current approaches to evidence, and the ways in which insufficient evidence and bias combine to lead to ineffective and even harmful care. We examine the potential for data science and AI technologies to address some of these issues, and we tackle commonly raised ethical concerns in this space. Ultimately, we provide a series of recommendations for reform in policies around MLHC which will facilitate the development of systems that provide a public benefit for all.

BIAS AND INSUFFICIENCY OF EVIDENCE IN HEALTHCARE

Many common interventions in healthcare are performed without good evidence to support them. A 2012 National Academy of Medicine report noted that high quality evidence is lacking or even non-existent for many clinical domains,[12] and a similar investigation from the UK's National Institute for Health and Care Excellence and the *BMJ* found that 50% of current treatments have unknown effectiveness, 10% are still in use despite being ineffective or even harmful and only 40% have some evidence for effectiveness.[13] As Prasad *et al* have found, studies that contradict previous research and lead to 'medical reversal' changes to practice standards are common—comprising up to 40% of papers that evaluated current standard of care in the *New England Journal of Medicine* from 2001 to 2010,[14] and many papers in *JAMA* and *The Lancet*.[10] It is clear that many interventions have insufficient evidence but continue to be adopted and propagated based on expert opinion typically backed by professional societies. Even when prospective randomised controlled trials are performed, they are subject to numerous opportunities for bias—and even outright conflict of interest—which can impact the quality and transferability of results.[15 16]

The burdens of medicine's failures in evidentiary quality and applicability are not borne equally.[11 17–19] The historical and ongoing omission in research of certain groups, including women and underserved populations, has skewed our understanding of health and disease.[11] The concerns that exist regarding the generation of algorithms on racially biased datasets[17] are unfortunately far from being new, but represent a continuation

BMJ Health Care Inform: first published as 10.1136/bmjhci-2020-100237 on 24 November 2020. Downloaded from http://informatics.bmj.com/ on September 25, 2021 by guest. Protected by copyright.

of a long-standing history of minority groups being underrepresented or entirely unrepresented in foundational clinical research.[11 18] The Framingham study, for example, generated its cardiovascular risk scores from an overwhelmingly white and male population, and has subsequently been inaccurate when uncritically used on black populations.[19] Similarly, women have been and continue to be heavily under-represented in clinical trials.[11 20 21] These problems extend to the global health context as well, as the trials used to inform clinical practice guidelines around the world tend to be conducted on a demographically restricted group of patients in high-income countries (mainly white males in the USA)[11] These issues are compounded by structural biases in medical education,[22] and the biases of the healthcare providers tasked with interpreting and implementing this medical knowledge in the clinical context.[23]

CAN MLHC HELP, OR WILL IT HARM?

The question is whether MLHC will help to remedy these shortcomings or exacerbate them. Models that are trained uncritically on databases embedded with societal biases and disparities will end up learning, amplifying and propagating those biases and disparities under the guise of algorithmic pseudo-objectivity.[2 17 24 25] Similarly, gaps in quality of care will be widened by the development and use of tools that are only beneficial to a certain population—such as a melanoma detection algorithm trained on a dataset containing mostly images of light toned skin.[26] Concerns also exist around patient privacy and safeguarding sensitive data (particularly for vulnerable groups such as HIV positive patients).[27] Finally, there are structural concerns related to the possibility that the information technology prerequisites for implementing MLHC will only be available to already privileged groups.[5 7]

Yet, and as recent scholarship has indicated, the potential for MLHC to counter biases in healthcare is considerable.[3 28] Data science methods can be used to audit healthcare datasets and processes, deriving insights and exposing implicit biases so they might be directly investigated and addressed.[1 3 29] While much has been made of the 'black box' characteristics of AI, it may be argued that human decision making in general is no more explainable.[30 31] This is particularly true in the context of the sort of implicit gender and racial biases that influence physicians' decisions but are unlikely to be consciously admitted.[23] As checklist studies in healthcare have demonstrated,[32] it may be possible to reduce these biases through the use of standardised prompts and clinical decision support tools that move clinical decisions closer to the data—and further from the biasing subjective evaluations. At the structural level, there is hope that AI will drive down the costs of care, increasing access for groups that have been traditionally underserved, and enabling greater levels of patient autonomy for self-management.[4 5]

Further, MLHC technologies may be able to address issues of disparity in the clinical research pipeline.[33] Improvements in the use and analysis of electronic health records and mobile health technology herald the possibility of mobilising massive amounts of healthcare data from across domestic and global populations. The prospect of using 'big data' (ie, large and comprehensive datasets involving many patient records) that better represents all patients for health research may hold promise for counteracting issues of evidentiary insufficiency and limitations. As shown by the 'All of Us' programme, biological information database initiatives can be specifically tailored toward the active inclusion of traditionally under-represented groups.[34] Recent progress in the ability to emulate a 'target trial' when no real trial exists may even enable scientists to regularly obtain real-world evidence and evolve insights about the effectiveness of treatments in groups absent from initial clinical trials.[35]

ENSURING MLHC WORKS FOR ALL

Despite this potential, MLHC is far from a magical solution, and should not be seen as such. Embracing it must not lead subsequently to the neglect of the role played by other structural factors such as economic inequities[36] and implicit physician bias.[23] No simple set of data-focused technical interventions alone can effectively deal with complex sociopolitical environments and structural inequity,[37] and simple 'race correction' methods can be deeply problematic.[38] The potential for 'big data' synthetic clinical trials, for example, must come as a supplement to and not a replacement for efforts to improve the diversity of clinical trial recruitment. Similarly, issues of structural bias must be acknowledged and addressed at all levels of the MLHC development pipeline,[17 39] from assessing the quality of the input data to ensuring adequate funding for the information technology needed to implement MLHC in underserved areas.

If MLHC is to be successful at reducing health disparities, it must reflect this function in its form. The troubling lack of diversity both in the field of AI[40] and in biomedical research generally[41] raises concerns about the perpetuation of biased perspectives in development, and the historical and ongoing flaws of healthcare and its research communities have led to distrust among minority communities.[42] The onus is on the MLHC community to rebuild this trust and embrace structural reform. Inclusion and active empowerment of members of marginalised communities is essential, and concepts around individual or collective data ownership and sovereignty[43] deserve further exploration.

At the same time, we must not forget the biases exerted by the status quo, which we cannot allow to slow the sort of progress that is necessary to address these problems. Problems evolving from the systematic exclusion of vulnerable populations from research will not be solved by the continued exclusion of these populations. While work

BMJ Health Care Inform: first published as 10.1136/bmjhci-2020-100237 on 24 November 2020. Downloaded from http://informatics.bmj.com/ on September 25, 2021 by guest. Protected by copyright.

Table 1 Areas of emphasis for ensuring machine learning for healthcare (MLHC) works for all

Area of emphasis	Recommendations
Ensure MLHC is equitable by design	► Develop pipelines for the promotion of diverse teams in all aspects of MLHC ► Ensure the inclusion of data from a broad range of groups, in a broad range of contexts ► Incorporate global partners to ensure health data science promotes global health equity.
Encourage public and open MLHC research	► Fund both direct MLHC research and research into ethical aspects of MLHC ► Harmonise ethical oversight between public and private research domains
Ensure adequate access to health information technology (IT) infrastructure	► Ensure all are included in the datasets by funding health data gathering infrastructure in underserved communities ► Develop MLHC products with an awareness of the broad range of health IT contexts for deployment
Ensure MLHC is clinically effective and impactful	► Ensure the presence of multidisciplinary teams that represent both clinical and data science perspectives ► Promote pathways for interdisciplinary training ► Hold MLHC innovations to the same standards as other healthcare interventions, including requirements for prospective validation and clear demonstration of impact
Audit MLHC on ethical metrics	► Mandate assessments of the performance of novel MLHC technology for impacts on marginalised and intersectional groups. ► Record the data necessary to perform these audits in an ongoing fashion
Mandate transparency in data collection, analysis and usage	► Build patient trust by ensuring that protocols for the collection, analysis and usage of data are transparent and open
Promote inclusive and interoperable data policy	► Ensure the existence of clear and ethical methods for ensuring the sharing of data between different sources while protecting patient rights and privacy ► Improve the standardisation of medical data generation and labelling across contexts ► Ensure that global partners are included, so that interoperability barriers do not hinder inclusive global collaboration

certainly must be done to ensure that minoritised patients do not need to be saved from MLHC research, work must also be done to remedy disparities and improve outcomes for minoritised patients through MLHC research.

The vigorous discussions surrounding ethical issues in MLHC must be translated into active efforts to construct the field from the ground up. Both the field itself and the outputs it creates must be ethical and equitable at their core, with these concerns rendered structurally integral rather than addressed post hoc. An emphasis is already growing throughout the field on the establishment of codes of conduct,[44] and practical procedures[6 33] for the ethical and equitable implementation of AI in healthcare. As outlined in table 1, we identify a number of critical areas of emphasis in the development of MLHC that fosters this vision. Just as the potential for problematic bias in MLHC has no single cause, the onus for achieving these recommendations does not fall on any single actor in the MLHC space. Open collaboration between universities, technology companies, ministries of health, regulators, patient advocates and individual clinicians and data scientists will be essential to its success.

CONCLUSION

The gaps in the medical knowledge system stem from the systematic exclusion of the majority of the world's population from health research. These gaps combined with implicit and explicit biases lead to suboptimal medical decision making which negatively impact health outcomes for everyone, but especially those in groups typically under-represented in health research.

Recent developments in machine learning and AI technologies hold some promise to address the issues with the generation of scientific evidence and human decision making. They also, however, have spurred concerns about their potential to maintain if not exacerbate these problems. These concerns must be aggressively addressed by adopting necessary structural reforms to ensure that the field is both equitable and ethical by design.

Claims of 'doing better' have, of course, come before in healthcare with respect to bias, and the burden is on MLHC as a field to grow in a fashion that is deserving of the hype it has received. MLHC is not a magic bullet, nor can it address issues of structural health inequity by itself, but its potential may be substantial. Healthcare is flawed, and it must be reformed so that it equitably benefits all. Effective and equitable machine learning, data science and AI will be an essential component of these efforts.

Twitter Liam G McCoy @liamgmccoy and Leo Anthony Celi @MITCriticalData

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

Funding LAC is funded by the National Institute of Health through the NIBIB R01 grant EB017205. JDB receives grant support from the Advanced Radiology Services Foundation in Grand Rapids, Michigan, USA.

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

BMJ Health Care Inform: first published as 10.1136/bmjhci-2020-100237 on 24 November 2020. Downloaded from http://informatics.bmj.com/ on September 25, 2021 by guest. Protected by copyright.

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 Celi http://orcid.org/0000-0001-6712-6626

REFERENCES

1 Zhang H, AX L, Abdalla M, *et al*. 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.
2 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.
3 Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019;21:167–79.
4 Nuffield Council on bioethics. Artificial intelligence (AI) in healthcare and research. In: *Bioethics Briefing note*, 2018: 1–8.
5 Centre for Data Ethics and Innovation. CDEI AI barometer, 2020. Available: https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf
6 Joshi I, Morley J. *Artificial intelligence: how to get it right. putting policy into practice for safe data-driven innovation in health and care*. NHSX, 2019.
7 Fenech M, Strukelj N, Buston O. *Ethical, social, and political challenges of artificial intelligence in health*. London: Wellcome Trust Future Advocacy, Published online 2018.
8 Ienca M, Ferretti A, Hurst S, *et al*. Considerations for ethics review of big data health research: a scoping review. *PLoS One* 2018;13:e0204937.
9 Loukides M, Mason H, Patil D. *Ethics and data science*. O'Reilly Media, Inc, 2018.
10 Herrera-Perez D, Haslam A, Crain T, *et al*. A comprehensive review of randomized clinical trials in three medical journals reveals 396 medical reversals. *Elife* 2019;8:e45183.
11 Oh SS, Galanter J, Thakur N, *et al*. Diversity in clinical and biomedical research: a promise yet to be fulfilled. *PLoS Med* 2015;12:e1001918.
12 Medicine I of, America C on the LHCS in. *Best care at lower cost: the path to continuously learning health care in America*. National Academies Press, 2013.
13 Smith QW, Street RL, Volk RJ, *et al*. Differing levels of clinical evidence: exploring communication challenges in shared decision making. Introduction. *Med Care Res Rev* 2013;70:3S–13.
14 Prasad V, Vandross A, Toomey C, *et al*. A decade of reversal: an analysis of 146 contradicted medical practices. *Mayo Clin Proc* 2013;88:790–8.
15 Gluud LL. Bias in clinical intervention research. *Am J Epidemiol* 2006;163:493–501.
16 Rothwell PM. Factors that can affect the external validity of randomised controlled trials. *PLoS Clin Trials* 2006;1:e9.
17 Geneviève LD, Martani A, Shaw D, Elger BS, *et al*. Structural racism in precision medicine: leaving no one behind. *BMC Med Ethics* 2020;21:17.
18 Buch B. Progress and collaboration on clinical trials. Available: https://www.fda.gov/news-events/fda-newsroom/fda-voices?feed=rss
19 Gijsberts CM, Groenewegen KA, Hoefer IE, *et al*. Race/Ethnic differences in the associations of the Framingham risk factors with carotid IMT and cardiovascular events. *PLoS One* 2015;10:e0132321.
20 Lippman A. *The inclusion of women in clinical trials: are we asking the right questions? women and health Protection=Action pour La protection de la santé des femmes*, 2006.
21 Yakerson A. Women in clinical trials: a review of policy development and health equity in the Canadian context. *Int J Equity Health* 2019;18:56.
22 Nolen L. How medical education is missing the bull's-eye. *N Engl J Med* 2020;382:2489–91.
23 Chapman EN, Kaatz A, Carnes M. Physicians and implicit bias: how doctors may unwittingly perpetuate health care disparities. *J Gen Intern Med* 2013;28:1504–10.
24 Osoba OA, Welser IVW. *An intelligence in our image: the risks of bias and errors in artificial intelligence*. Rand Corporation, 2017.
25 Seyyed-Kalantari L, Liu G, McDermott M, *et al*. CheXclusion: fairness gaps in deep chest X-ray classifiers. arXiv:200300827 [cs, eess, stat], 2020. Available: http://arxiv.org/abs/2003.00827 [Accessed 2 Sep 2020].
26 Phillips M, Marsden H, Jaffe W, *et al*. Assessment of accuracy of an artificial intelligence algorithm to detect melanoma in images of skin lesions. *JAMA Netw Open* 2019;2:e1913436.
27 Nissenbaum H, Patterson H. Biosensing in context: Health privacy in a connected world. In: *Quantified: biosensing technologies in everyday life*. 79, 2016.
28 Chen IY, Joshi S, Ghassemi M. Treating health disparities with artificial intelligence. *Nat Med* 2020;26:16–17.
29 Kleinberg J, Ludwig J, Mullainathan S, *et al*. Discrimination in the age of algorithms. *J Leg Anal* 2018;10:113–74.
30 Lipton ZC. The mythos of model interpretability. arXiv:160603490 [cs, stat], 2017. Available: http://arxiv.org/abs/1606.03490 [Accessed 3 Mar 2020].
31 Zerilli J, Knott A, Maclaurin J, *et al*. Transparency in algorithmic and human decision-making: is there a double standard? *Philos Technol* 2019;32:661–83.
32 Nordell J. Opinion | a fix for gender bias in health care? check. the new York times. Available: https://www.nytimes.com/2017/01/11/opinion/a-fix-for-gender-bias-in-health-care-check.html [Accessed 3 Mar 2020].
33 Crawford K, Whittaker M, Elish M, *et al*. The AI now report: the social and economic implications of artificial intelligence technologies in the near-term. *Report Prepared for the AI Now Public Symposium, Hosted by the White House and New York University's Information Law Institute*, 2016.
34 The All of Us Research Program Investigators. The "all of us" research program. *N Engl J Med* 2019.
35 Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;183:758–64.
36 Chokshi DA, Income CDA. Income, poverty, and health inequality. *JAMA* 2018;319:1312–3.
37 Society R. *Data management and use: governance in the 21st century—A British Academy and Royal Society project*, 2017.
38 Vyas DA, Eisenstein LG, Jones DS. Hidden in plain sight - reconsidering the use of race correction in clinical algorithms. *N Engl J Med* 2020;383:874–82.
39 Wiens J, Saria S, Sendak M, *et al*. Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019;25:1337–40.
40 West SM, Whittaker M, Crawford K. *Discriminating systems: gender, race, and power in AI*, 2019.
41 Drake MV. Diversity: boost diversity in biomedical research. *Nature* 2017;543:623.
42 Armstrong K, Ravenell KL, McMurphy S, *et al*. Racial/Ethnic differences in physician distrust in the United States. *Am J Public Health* 2007;97:1283–9.
43 Kukutai T, Taylor J. *Indigenous data Sovereignty: toward an agenda*. 38. Anu Press, 2016.
44 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].

BMJ Health Care Inform: first published as 10.1136/bmjhci-2020-100237 on 24 November 2020. Downloaded from http://informatics.bmj.com/ on September 25, 2021 by guest. Protected by copyright.