Journal of the American Medical Informatics Association, 27(12), 2020, 2024–2027
doi: 10.1093/jamia/ocaa085
Advance Access Publication Date: 25 June 2020
Perspective

Perspective

Patient safety and quality improvement: Ethical principles for a regulatory approach to bias in healthcare machine learning

Melissa D. McCradden,[1] Shalmali Joshi,[2] James A. Anderson,[1,3,4] Mjaye Mazwi,[5] Anna Goldenberg,[2,6,7,8] and Randi Zlotnik Shaul[1,9,10]

[1]Bioethics Department, The Hospital for Sick Children, Toronto, Ontario, Canada, [2]Vector Institute, Toronto, Ontario, Canada, [3]Institute for Health Policy, Management and Evaluation, University of Toronto, Toronto, Ontario, Canada, [4]Joint Centre for Bioethics, University of Toronto, Toronto, Ontario, Canada, [5]Department of Critical Care Medicine, The Hospital for Sick Children, Toronto, Ontario, Canada, [6]Genetics and Genome Biology, The Hospital for Sick Children, Peter Gilgan Centre for Research and Learning, Toronto, Ontario, Canada, [7]Department of Computer Science, University of Toronto, Toronto, Ontario, Canada, [8]CIFAR, Toronto, Ontario, Canada, [9]Department of Paediatrics, University of Toronto, Toronto, ON, Canada, and [10]Child Health Evaluative Sciences, The Hospital for Sick Children, Peter Gilgan Centre for Research, Toronto, Ontario, Canada

Corresponding Author: Melissa McCradden, PhD, Bioethics Department, 525 University Ave, Toronto, ON M5G 2L3, Canada; melissa.mccradden@sickkids.ca

Received 16 April 2020; Editorial Decision 17 April 2020; Accepted 1 May 2020

ABSTRACT

Accumulating evidence demonstrates the impact of bias that reflects social inequality on the performance of machine learning (ML) models in health care. Given their intended placement within healthcare decision making more broadly, ML tools require attention to adequately quantify the impact of bias and reduce its potential to exacerbate inequalities. We suggest that taking a patient safety and quality improvement approach to bias can support the quantification of bias-related effects on ML. Drawing from the ethical principles underpinning these approaches, we argue that patient safety and quality improvement lenses support the quantification of relevant performance metrics, in order to minimize harm while promoting accountability, justice, and transparency. We identify specific methods for operationalizing these principles with the goal of attending to bias to support better decision making in light of controllable and uncontrollable factors.

Key words: machine learning, systematic bias, healthcare delivery, patient safety, quality improvement

INTRODUCTION

It is now clear that healthcare inequalities are encoded right into the health data that serve as the substrate for healthcare machine learning (ML) research and model development. ML researchers have exposed the presence of pernicious bias (ie, differences in health conditions related to social inequality) within health data and their impact on model performance.[1–3] The persistence of healthcare inequalities poses an ethical threat to the core values of health institutions. As ML is adopted within medical practice, attention to how factors such as bias may impact the use of ML in vulnerable populations is imperative.

Despite long-standing recognition of disparities,[4] scientific knowledge about how social determinants of health drive disparate outcomes continues to evolve.[5,6] The effects of differential access, distrust of medical institutions, housing or food security, racialization, health insurance, and others can be replicated in ML models. The implications of imprudent incorporation of biased model predictions in clinical decision making can be troubling. Consider the case of a model intended to predict likelihood of mortality. When the model's error rate differs between racial groups,[1] what are the implications for end-of-life decision making on the basis of ML-assisted prognostic projections?

The problem of bias for healthcare ML

Recent evidence indicates that model performance discrepancies, as a function of protected identities (meaning characteristics protected by human and civil rights legislation [eg, race, gender]), can have significant health implications for vulnerable groups, potentially worsening health inequalities.[7,8] Automation of some elements of medical practice or clinical decision making poses particular risks given the potential to view such outputs as "objective" due to ML's "veneer of technical neutrality."[7]

Despite these perceptions, much of the ML research community is attuned to the fact that model performance may differ as a function of one's identity.[9–16] It is less clear how to deal with this challenge. Some differences are relevant and should be incorporated into models (eg, some age-related differences in patient outcomes). Others are related to pernicious bias (eg, impact of insurance status) and should be mitigated. These decisions turn on the nature of the clinical problem one is trying to address, our understanding of the role of identity related factors, and most importantly, the intended impact of the tool for patient care.[17]

The need for a regulatory approach to bias

Although it is important to note that bias is not a concern specific to ML,[18] ML may raise the stakes. Computerized outputs generally are perceived as objective and epistemically superior to the knowledge of a single human decision maker.[19] Emerging evidence suggests that when a clinician is uncertain, they may defer to computerized outputs.[20] Reliance on computational objectivity risks encoding and systematizing these biases, exacerbating their effects on marginalized populations.[7,8,15] When model errors are unevenly distributed then, so too are the risks.

We propose that a regulatory approach considers bias in ML as a patient safety and quality improvement issue with the aim of preventing unintended harms and augmenting the provision of healthcare delivery with respect to justice. Patient safety generally refers to the need to conduct assessments to guarantee a minimum standard of functioning with respect to minimizing preventable harms. Quality improvement is consistent with a learning healthcare system approach that aims to optimize the delivery of care to maximally benefit patients. Both lenses draw from broad, well accepted ethical commitments and apply these principles to individual cases.

ETHICAL PRINCIPLES UNDERLYING PATIENT SAFETY IN HEALTHCARE ML

The following ethical principles can serve as the foundation for regulation of healthcare ML as it concerns bias-related evaluations.

Nonmaleficence

Nonmaleficence underpins the obligation of moral agents (including clinicians, administrative decision makers, and others in health care) to adopt technologies that promote benefit and avoid harm to patients. This obligation can be operationalized through a risk-based classification system (eg, Food and Drug Administration's Oversight of Clinical Investigation),[21] whereby oversight is linked with the potential risks to patients. The overall assessment of risk will be similar to other clinical tools: (1) the likelihood of the error occurring (eg, false positive or false negative rates), (2) the impact on the patient, and (3) the extent to which the error would be detectable by the user.[21] Combined with the drafted guidance on software as a medical device, risk pertaining to the impact on the patient (list item 2) is modified by the state of the patient (critical, serious, nonserious) and significance of the information being provided by the tool (inform clinical management, drive clinical management, or treat or diagnose).[22]

Relevance

To promote thoughtful and proportionate approaches to regulation with respect to bias, relevance of bias-specific evaluations should be made according to the identified clinical problem. The patient safety lens suggests directing increased attention specifically to where harm is most likely (based on the data) rather than considering any particular group or population as one that is categorically more at risk.[23] This context-specific attention is important because it is not the simple case that the majority group receives the most accurate predictions, nor is it true that the error rate discrepancy between groups is similar for health conditions where inequalities are recognized.[1] Domain experts can support the identification of the particular bias-related considerations for a given ML model.[17]

Accountability

Patient safety requires accountability at multiple steps, including data collection, reporting at the statistical and clinical levels, and sustained oversight. Standardizing the collection and reporting of patient data[24] can promote consistency in the development, validation, and evaluation of ML models to support consistent standards as the basis for informing decision making. Reporting performance metrics according to relevant subgroups at both the statistical (ie, preclinical) and clinical levels provides the necessary data for hospital decision makers and regulators to make informed decisions about adopting models.

Transparency

Clear communication of model limitations and performance discrepancies with respect to protected identities is essential to mitigating harms to patients. Members of our group have called for transparency in reporting subgroup-specific model performance metrics during statistical and clinical validation steps.[17] Efforts to support explanation and interpretation of the overall model, and for intelligibility at the point of care, may consider incorporating information about subgroup-specific performance indicators.

Justice

Justice requires treating like cases alike, and different cases differently to promote fairnes. For ML, this principle entails ensuring that decisions are not made similarly for groups for whom the tool performs differently. It would be inappropriate, for example, to consider the result of skin biopsy as similarly determinative for 2 patients when one patient is more likely to have a higher false negative result.[14] Attending to variations in decisional accuracy across subgroups can promote justice by encouraging due diligence with respect to appropriate model use across these groups.

RECOMMENDATIONS

To operationalize patient safety and quality improvement, the following practices can be adopted by healthcare institutions and regulators who embrace these ethical principles for the delivery of ML-assisted health care.

Data collection

Whether bias is believed to influence health outcomes within a particular model or not, developers should maintain statistics on the characteristics of the population on which the model was developed. These variables should, at minimum, include the so-called protected characteristics under civil and human rights legislation (ie, gender, race, ethnicity, age, socioeconomic status).[24] These variables may or may not be included in the model itself but can support a post hoc evaluation of systematic differences in predictions. This information can be used to help researchers determine whether the model can generalize appropriately, whether a distinct model may be needed for some subgroups of patients, or to explore unappreciated causal factors that relate to subgroup differences.

Auditing and prospective evaluation

Model auditing is consistent with a focus on continuous quality improvement, and should collect and retain evidence that the ML-based decision-making tool is safe to use in the intended population. Audits should include (1) information on the reliability and validity of the target label, (2) evidence of sufficient representation of subgroups in the population for which the model is intended, (3) data collection errors stratified by subgroup, and (4) assessment of potentially confounding factors. In the event that such benefits are restricted to certain groups, recommendations regarding the subpopulations in which a tool may be safely used are essential. At a systemic level, care should be taken to ensure that benefits of predictive models do not unfairly accrue to privileged populations. Any reported disparity in outcome or treatment path determined by such models should accompany logged clinical justifications.

Models approved at the regulatory stage must also be evaluated locally. Performance of ML models is well recognized to vary across sites due to a number of factors[25]; thus, the need for local validation through a prospective, noninterventional silent period is apparent.[26] Techniques to investigate hidden stratification effects can reveal noncausative but correlative features that result in notable differences in performance accuracy[27]—such techniques may support identification of bias-related effects. Reporting of relevant subgroup differences in model performance is especially important for clinical trials involving AI. This information aids healthcare decision-makers in determining the suitability of a model for the population served by their institution, and for the clinician determining how much to rely on a model's output with respect to an individual patient.[17]

Practice guidance

Careful consideration must be given to determine how information about potential bias is included in the point-of-care interpretation of model outputs. Physicians have a fiduciary duty to continually act in their patient's best interests and obtain informed consent from capable patients or their surrogate decision makers, which includes offering them all relevant information to support decision making. Incorporating understanding of potential bias in communicating the model outputs to the patients can enhance trustworthiness. In collaboration with stakeholders, ML developers should consider relevant differences in model performance and identify users' needs in deciding what information to present to human decision makers.

Oversight

Oversight is central to any quality improvement effort. Model evaluations may be performed regularly to prevent the risk of model decay or the influence of feedback loops, which may worsen errors for specific populations over time.[28] Continued quality improvement efforts independent of bias alone should be conducted to ensure the performance of a model is maintained. These evaluations are particularly important as populations shift, practice changes occur, and new policies are implemented. Ensuring these evaluations keep track of subgroup-specific effects can continue to inform the model's ongoing use.

One of the key remaining questions facing ML is where oversight should occur, and by whom. At the present time, conversations about a feasible, long-term oversight strategy are in flux. In the long term, as these tools become ubiquitous across health care, the need for robust, consistent standards for local review will become more pronounced. In the interim, it is incumbent on ML researchers to retain records of model performance metrics and conduct evaluations. This means that there will have to be expertise on staff in hospitals where the tools are deployed that will attend to the fairness, efficacy, and safety of these algorithms. Clinicians can and should demand these evaluations when considering the adoption and use of models in clinical populations. Hospital decision makers should be critically evaluating these statistics to determine applicability to the population they serve.

CONCLUSION

The impact of social inequalities on health outcomes can be reflected in ML models, resulting in performance discrepancies that put some groups of patients at risk. The lenses of patient safety and continuous quality improvement provide ethical guidance that informs a model-specific evaluation of bias. By operationalizing the ethical principles that underlie these efforts, a regulatory approach that considers bias can better leverage ML to promote fairness in the delivery of healthcare.

AUTHOR CONTRIBUTIONS

Each author made substantial intellectual contributions to elements of the conception or recommendations in this work. MDM developed the initial draft and all authors revised it critically for important intellectual content. All authors gave final approval of the version to be published and agree to be accountable for all aspects of the work.

CONFLICT OF INTEREST STATEMENT

None declared.

REFERENCES

1. Chen IY, Szolovits P, Ghassemi M. Can AI help reduce disparities in general medical and mental health care? *AMA J Ethics* 2019; 21 (2): 167–79.
2. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In: proceedings of the ACM Conference on Health, Inference, and Learning; 2020: 110–20.
3. Pfohl S, Marafino B, Coulet A, Rodriguez F, Palaniappan L, Shah NH. Creating fair models of atherosclerotic cardiovascular disease risk. In proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society; 2019: 271–8.
4. Institute of Medicine Committee on Understanding and Eliminating Racial and Ethnic Disparities in Health Care; Smedley BD, Stith AY, Nelson

AR. *Unequal Treatment: Confronting Racial and Ethnic Disparities in Healthcare*. Washington, DC: National Academies Press; 2002.

5. Davidson KW, McGinn T. Screening for social determinants of health: the known and unknown. *JAMA* 2019; 322 (11): 1037–8.
6. Marmot M. Social determinants of health inequalities. *Lancet* 2005; 365 (9464): 1099–104.
7. Benjamin R. Assessing risk, automating racism. *Science* 2019; 366 (6464): 421–2.
8. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; 366 (6464): 447–53.
9. Buolamwini J, Gebru T. Gender shades: Intersectional accuracy disparities in commercial gender classification. In: conference on Fairness, Accountability and Transparency; 2018; 77–91.
10. Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. In proceedings of the 3rd Innovations in Theoretical Computer Science Conference; 2012; 214–26.
11. Angwin J, Larson J, Mattu S, Kirchner L. There's software used across the country to predict future criminals and it's biased against blacks. 2016. https://www. propublica. org/article/machine-bias-risk-assessments-in-criminal-sentencing Accessed May 20, 2020.
12. De-Arteaga M, Romanov A, Wallach H, *et al* Bias in bios: A case study of semantic representation bias in a high-stakes setting. In: proceedings of the Conference on Fairness, Accountability, and Transparency; 2019: 120–28.
13. Bolukbasi T, Chang KW, Zou JY, Saligrama V, Kalai AT. Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In: NIPS '16: Proceedings of the 30th International Conference on Neural Information Processing Systems; 2016: 4356–64.
14. Adamson AS, Smith A. Machine learning and health care disparities in dermatology. *JAMA Dermatol* 2018; 154 (11): 1247–8.
15. Zhang H, Lu AX, Abdalla M, McDermott M, Ghassemi M. Hurtful words: quantifying biases in clinical contextual word embeddings. In proceedings of the ACM Conference on Health, Inference, and Learning 2020; 110–20.
16. Garg N, Schiebinger L, Jurafsky D, Zou J. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proc Natl Acad Sci U S A* 2018; 115 (16): E3635–44.
17. McCradden MD, Joshi S, Mazwi M, Anderson JA. Ethical limitations of algorithmic fairness solutions in health care machine learning. *Lancet Digit Health* 2020; 2 (5): E221–3.
18. Johnstone M-J, Kanitsaki O. The neglect of racism as an ethical issue in health care. *J Immigr Minor Health* 2010; 12 (4): 489–95.
19. Cummings M. Automation bias in intelligent time critical decision support systems. In: AIAA 1st Intelligent Systems Technical Conference; 2004.
20. Kiani A, Uyumazturk B, Rajpurkar P, *et al* Impact of a deep learning assistant on the histopathologic classification of liver cancer. *NPJ Digit Med* 2020; 3 (1): 23.
21. U.S. Food and Drug Administration. Oversight of clinical investigations—a risk-based approach to monitoring. 2013. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/oversight-clinical-investigations-risk-based-approach-monitoring Accessed May 20, 2020.
22. US Food and Drug Administration. Clinical and patient decision support software: draft guidance for industry and Food and Drug Administration Staff. 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software Accessed May 20, 2020.
23. NHS England and NHS Improvement. *The NHS Patient Safety Strategy*. 2019. https://improvement.nhs.uk/resources/patient-safety-strategy/ Accessed May 20, 2020.
24. Hernandez-Broussard T, Bozkurt S, Ioannidis J. MINIMAI: MINimum Information for Medical AI-developing reporting standards for artificial intelligence solutions in healthcare.*J Am Med Inform Assoc*. doi: 10.1093/jamia/ocaa088.
25. Wiens J, Saria S, Sendak M, *et al* Do no harm: a roadmap for responsible machine learning for health care. *Nat Med* 2019; 25 (9): 1337–40.
26. Drysdale E, Dolatabadi E, Chivers C, *et al* White Paper: implementing AI in healthcare. Vector-SickKids Health AI Deployment Symposium. 2020. https://vectorinstitute.ai/wp-content/uploads/2020/03/implementing-ai-in-healthcare.pdf Accessed May 20, 2020.
27. Oakden-Rayner L, Dunnmon J, Carneiro G, Ré C. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. arXiv:1909.12475. 2019.
28. Ensign D, Friedler SA, Neville S, Scheidegger C, Venkatasubramanian S. Runaway feedback loops in predictive policing. arXiv:1706.09847. 2017.