Ethical limitations of algorithmic fairness solutions in health care machine learning

Artificial intelligence has exposed pernicious bias within health data that constitutes substantial ethical threat to the use of machine learning in medicine.[1,2] Solutions of algorithmic fairness have been developed to create neutral models: models designed to produce non-discriminatory predictions by constraining bias with respect to predicted outcomes for protected identities, such as race or gender.[3] These solutions can omit such variables from the model (widely regarded as ineffective and can increase discrimination), constrain it to ensure equal error rates across groups, derive outcomes that are independent of one's identity after controlling for the estimated risk of that outcome, or mathematically balance benefit and harm to all groups·

The temptation to engineer ethics into algorithm design is immense and industry is increasingly pushing these solutions. In the health-care space, where the stakes could be higher, clinicians will integrate these models into their care, trusting the issue of bias has been sufficiently managed within the model. However, even if well recognised technical challenges are set aside,[3,4] framing fairness as a purely technical problem solvable by the inclusion of more data or accurate computations is ethically problematic. We highlight challenges to the ethical and empirical efficacy of solutions of algorithmic fairness that show risks of relying too heavily on the so called veneer of technical neutrality,[5] which could exacerbate harms to vulnerable groups.

Historically, algorithmic fairness has not accounted for complex causal relationships between biological, environmental, and social factors that give rise to differences in medical conditions across protected identities. Social determinants of health play an important role, particularly for risk models. Social and structural factors affect health across multiple intersecting identities,[4] but the mechanism(s) by which social determinants affect health outcomes is not always well understood.

Additional complications flow from the reality that difference does not always entail inequity. In some instances, it is appropriate to incorporate differences between identities because there is a reasonable presumption of causation. For example, biological differences between genders can affect the efficacy of pharmacological compounds; incorporating these differences into prescribing practices does not make those prescriptions unjust. However, incorporating non-causative factors into recommendations can propagate unequal treatment by reifying extant inequities and exacerbating their effects. We should not allow models to promote different standards of care according to protected identities that do not have a causative association with the outcome. Nevertheless, in many cases it is difficult to distinguish between acknowledging difference and propagating discrimination.

Given the epistemic uncertainty surrounding the association between protected identities and health outcomes, the use of fairness solutions can create empirical challenges. Consider the case of heart attack symptoms among women.[6] The under-representation of women (particularly women of colour) in research of heart health is now well recognised as problematic and directly affected uneven improvements in treatment of heart attacks between women and men. By tailoring health solutions to the majority (ie, referent) group, we inevitably fall short of helping all patients. Many algorithmic fairness solutions, in effect, replicate this problem by trying to fit the non-referent groups to that of the referent,[7,8] ignoring heterogeneity and assuming that the latter represents a true underlying pattern.

Another concern is disconnection between the patient's clinical trajectory and the fair prediction. Consider the implications at the point-of-care, a model, corrected for fairness, will predict that a patient will respond to a treatment as a patient in the referent class would. What happens when that patient does not have the predicted response? This difference between an idealised model and non-ideal, real-world behaviour affects metrics of model performance (eg, specificity, sensitivity) and clinical utility in practice. Moreover, the model has made an ineffective recommendation that could have obscured more relevant interventions to help that patient. If clinicians and patients believe that the mode has been rendered neutral, then any discrepancies between model prediction and the patient's true clinical state might be impossible to interpret. The result would be to camouflage persistent

Panel: Recommendations for ethical approaches to issues of bias in health models of machine learning

Relying on neutral algorithms is problematic

Challenges cast doubt on whether any solution can adequately facilitate ethical goals. Resisting the tendency to view machine learning solutions as objective is essential to remaining patient-centered and preventing unintended harms.

Problem formulation can support improved models

Changing the way the problem is conceptualised and operationalised can reduce the effect of pernicious bias on outputs.[6] Clinicians have a key role in identifying actionable, clinical problems where the effects of bias are minimized, or causal knowledge exists to support algorithmic fairness solutions. Clinicians in non-medical sciences might have epistemic advantages to support such formulations.

Transparency is required surrounding model development and statistical validation

Standardisation of reporting for models of machine learning can promote transparency about training data and statistical validation which can help clinicians and health-care decision makers to determine the transferability of the model to their served population. Large discrepancies increase the risk of under-performance in under-represented patients. Group-specific performance metrics (eg, stratification by ethnicity, gender, or socioeconomic status) can inform considerations for patient safety in implementing a given model.

Initiating transparency at point-of-care

As the field of explainable or interpretable machine learning evolves, we suggest that, when sensitive attributes are used in problem formulation and could affect predictions, they should be accompanied by visibility of the top predictive features. Where predictions are affected by sensitive variables, prediction and rationale should be documented for the ensuing clinical decision to promote fair and transparent medical decision making. Communication with patients about the rationale is essential to shared decision making.

Transparency at the prediction level

Robust auditing processes are increasingly viewed as essential to proper oversight of machine learning tools. When bias is considered, auditing can promote reflexive understanding of how tools of machine learning can affect care management of vulnerable patients. Discrepancies in systematic prediction can become evident and used as evidence of the need for beneficial interventions and highlight large structural barriers that affect health inequalities.

Ethical decision making suggests engaging diverse knowledge sources

Ethical analysis should consider real-world consequences for affected groups, weigh benefits and risks of various approaches, and engage stakeholders to come to the most supportable conclusion. Therefore, analysis needs to focus on the downstream effects on patients rather than adopting the presumption that fairness is accomplished solely in the metrics of the system.

health inequalities. As such, fairness, operationalised by output metrics alone, is insufficient; real-world consequences should be carefully considered.

Bias and ineffective solutions of algorithmic fairness threaten the ethical obligation to avoid or minimise harms to patients (non-maleficence; panel). Non-maleficence demands that any new clinical tool should be assessed for patient safety. For health-care machine learning, safety should include awareness of model limitations with respect to protected identities and social determinants of health. Considerations of justice requires that implemented models do not exacerbate pernicious bias. There is a movement toward developing guidelines of standardised reporting for machine learning models of health care[9] and their prospective appraisal through clinical trials.[10] Appraisal is particularly important in determining the real-world implications for vulnerable patients when machine learning models are integrated into clinical decision making. Clinical trials are essential to providing a sense of the model's performance for clinicians to make informed decisions at the point-of-care through awareness of identity-related model limitations.

Some computations can promote justice through revealing unfairness and refining problem formulation. Obermeyer and colleagues[2] show how calibration can reveal unfairness in a seemingly neutral task through which choice of label can dictate how heavily bias is incorporated into predictions. It might be that no way exists to define a purely neutral problem; some clinical prediction tasks might be more susceptible to bias than others. Transparency at multiple points in the pipeline of machine learning including development, testing, and implementation stages can support interpretation of model outputs by relevant stakeholders (eg, researchers, clinicians, patients, and auditors). Combined with adequate documentation of outputs and ensuing decisions, these steps support a strong accountability framework for point-of-care machine learning tools with respect to safety and fairness to patients. Problem formulation with respect to bias will often be value-laden and ethically charged. Ethical decision making highlights the importance of converging knowledge sources to inform a given choice. Important stakeholders could include affected communities, cultural anthropologists, social scientists, and race and gender theorists.

Computations alone clearly cannot solve the bias problem, but they could be offered a place within a broader approach to addressing fairness aims in healthcare. Algorithmic fairness could be necessary to fix statistical limitations reflective of perniciously biased data, and we encourage this work. The worry is that suggesting these as solutions risks unintended harms.[5] Bias is not new; however, machine learning has potential to reveal bias, motivate change, and support ethical analysis while bringing this crucial conversation to a new audience. We are at a watershed moment in health care. Ethical considerations have rarely been so integral and essential to maximising success of a technology both empirically and clinically. The time is right to partake in thoughtful and collaborative engagement on the challenge of bias to bring about lasting change.

We declare no competing interests.

*Melissa D McCradden, Shalmali Joshi, Mjaye Mazwi, James A Anderson
melissa.mccradden@sickkids.ca

Department of Bioethics (MDM, JAA), Department of Critical Care Medicine (MM), The Hospital for Sick Children, Toronto, ON, Canada; and Vector Institute for Artificial Intelligence, Toronto, ON, Canada (SJ)

1 Char DS, Shah NH, Magnus D. Implementing machine learning in health care—addressing ethical challenges. *N Engl J Med* 2018; **378:** 981–83.

2 Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 2019; **366:** 447–53.

3 Corbett-Davies S, Goel S. The measure and mismeasure of fairness: a critical review of fair machine learning. https://5harad.com/papers/fair-ml.pdf (accessed March 16, 2020).

4 Marmot M. Social determinants of health inequalities. *Lancet* 2005; **365:** 1099–104.

5 Benjamin R. Assessing risk, automating racism. *Science* 2019; **366:** 421–22.

6 Wenger NK. Cardiovascular disease: the female heart is vulnerable. A call to action from the 10Q Report. *Clin Cardiol* 2012; **35:** 134–35.

7 Friedler SA, Scheidegger C, Venkatasubramanian S. On the (im)possibility of fairness. Sept 23, 2019. https://arxiv.org/pdf/1609.07236.pdf (accessed March 16, 2020).

8 Barocas S, Hardt M, Narayanan A. Fairness and machine learning: limitations and opportunities; 2018. Fairmlbook.org, 2019.

9 Collins GS, Reitsma JB, Altman DG, Moons KGM, TRIPOD Group. Transparent reporting of a multivariable prediction model for individual prognosis or diagnosis (TRIPOD): The TRIPOD Statement. *Ann Intern Med* 2015; **162:** 55–63.

10 The CONSORT-AI and SPIRIT-AI Steering Group. Reporting guidelines for clinical trials evaluating artificial intelligence interventions are needed. *Nat Med* 2019; **25:** 1467–68.