When Your Only Tool Is A Hammer

Ethical Limitations of Algorithmic Fairness Solutions in Healthcare Machine Learning

Melissa McCradden
The Hospital for Sick Children - Bioethics
Toronto, Ontario Canada
melissa.mccradden@sickkids.ca

Mjaye Mazwi
The Hospital for Sick Children – Critical Care
Toronto, Ontario Canada
mjaye.mazwi@sickkids.ca

Shalmali Joshi
Vector Institute for AI
Toronto, Ontario Canada
shalmali@vectorinstitute.ai

James A Anderson
The Hospital for Sick Children - Bioethics
Toronto, Ontario Canada
james.anderson@sickkids.ca

ABSTRACT

It is no longer a hypothetical worry that artificial intelligence - more specifically, machine learning (ML) - can propagate the effects of pernicious bias in healthcare. To address these problems, some have proposed the development of 'algorithmic fairness' solutions. The primary goal of these solutions is to constrain the effect of pernicious bias with respect to a given outcome of interest as a function of one's protected identity (i.e., characteristics generally protected by civil or human rights legislation. The technical limitations of these solutions have been well-characterized. Ethically, the problematic implication – of developers, potentially, and end users – is that by virtue of algorithmic fairness solutions a model can be rendered 'objective' (i.e., free from the influence of pernicious bias). The ostensible neutrality of these solutions may unintentionally prompt new consequences for vulnerable groups by obscuring downstream problems due to the persistence of real-world bias.

The main epistemic limitation of algorithmic fairness is that it assumes the relationship between the extent of bias's impact on a given health outcome and one's protected identity is mathematically quantifiable. The reality is that social and structural factors confluence in complex and unknown ways to produce health inequalities. Some of these are biologic in nature, and differences like these are directly relevant to predicting a health event and should be incorporated into the model's design. Others are reflective of prejudice, lack of access to healthcare, or implicit bias. Sometimes, there may be a combination. With respect to any specific task, it is difficult to untangle the complex relationships between potentially influential factors and which ones are 'fair' and which are not to inform their inclusion or mitigation in the model's design.

Empirically, when attempting to control for the effects of bias within an ML model we may unintentionally obfuscate persistent unfairness. Effectively the algorithmic solution creates a prediction that is based on relationships that do not map the real-world ones. As such, we risk seeing fairness in the predictions and mistaking them for having generated fair outcomes with respect to a given health condition. Moreover, given that the model's performance is assessed with respect to its ability to track true events, most notions of model performance would suffer. These discrepancies would only be evident longer term and not at the point-of-care where decisions must be made concerning the care management of patients. As such, 'fairness' as operationalized by output metrics alone is insufficient; the downstream, real-world consequences must be carefully considered.

Computations may be the hammer of ML, but they are likely not the answer for healthcare's bias problem. We do not wish to disparage such endeavours; they are valuable approaches to highlight inequalities in health. We merely wish to point out that the considerations above point to potential categories of problems that will not be well served by such solutions. Ethical analysis provides indispensable tools to engage in problem formulation, generate transparency, and scrutinize technologies, all with a focus on the real-world implications for patients affected by the solutions.

CCS CONCEPTS

• Social and professional topics → Computing / technology policy

KEYWORDS

Machine learning; algorithmic fairness; healthcare; medicine; bioethics; ethics; bias; racism; sexism; discrimination

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02.
https://doi.org/10.1145/3375627.3375824