Machine Learning, Health Disparities, and Causal Reasoning

In their current *Annals* article, Rajkomar and colleagues (1) warn us that the introduction of machine-learned predictive algorithms into medicine might inadvertently reinforce or create inequitable treatment of protected groups, for which the computer science community has adopted the terminology of "fairness." Others have noted the need for careful, ethical scrutiny of these models (2), and Rajkomar and colleagues add to those calls an elaborate taxonomy of pitfalls and an oversight structure to minimize the ethical harms.

The authors offer some useful constructs, many of which have direct counterparts in clinical research, epidemiology, and implementation science (such as "training–serving skew" for generalizability or transportability). In addition to requiring that care guided by algorithms not perpetuate past injustices, they remind us that we must ultimately judge this care by the same ethical standards that we would apply to any new health intervention. They also show that different methodologic definitions of fairness cannot always be satisfied simultaneously, potentially forcing difficult tradeoffs (3). In the article's table, the authors provide a list of process recommendations for each step in the development and application of machine-learned algorithms to maximize the chance that such tools will diminish health disparities, or at least not exacerbate them.

That said, the scheme presented to classify the potential distributional disparities created by machine-learning algorithms seems difficult to implement and may need some additional conceptual structure. Broadly speaking, 2 pathways exist by which clinical predictive algorithms may harm protected groups, which typically comprise persons who historically have been victims of social and health disparities. The first consists of the algorithms that predict outcomes less accurately for such persons than for the bulk of the population. This inequality may arise from many sources that the authors elucidate: Too few members of the protected group may have been studied (4); outcomes or predictors in such groups may have been measured less well than in the dominant study population; wider variation may exist among protected patients, leading to less precise estimates of a protected individual's risk; and developers or users of an algorithm may uncritically apply predictions for one group that were derived from another, with poorer predictive success.

A second, deeper, and more insidious pitfall may occur even for algorithms meticulously derived from an adequate number of protected persons. If the outcome is a surrogate for the true outcome of interest (such as police arrest for actual crime or rehospitalization for disease severity) and this surrogate is affected by social inequity, then those inequities could be used to predict and hence determine the future.

Since the time of Hume, so-called predictions have been understood to be merely future projections of relationships observed in the past. Why is this being brought to the fore only now, in the context of machine learning and social inequity? The answer perhaps is found in the emphasis that medicine has traditionally placed on developing these models in part to understand *why* things are happening, that is, causal inference. The Framingham model, one of the most durable in medicine, is a signal example. Although this model predicts the probability of cardiovascular disease, its main medical value has been in identifying modifiable causal factors, such as smoking, cholesterol levels, and blood pressure, preselected on the basis of prior knowledge. Research has shown that risk is indeed modified through changes in those risk factors.

In contrast, machine-learning algorithms traditionally have been developed and evaluated based only on predictive success, with the computer selecting and combining predictors; an understanding of why the model works has been distinctly secondary or absent. As Judea Pearl passionately argues in his *Book of Why* (5), why things happen—the causal relationships among inputs and outcomes—may elude machine-learning algorithms. Such algorithms might select discolored teeth as a better predictor of lung cancer than self-reported smoking status, which might be correct in some groups but because it is not a causal factor, may be useless in others, such as populations with different dietary patterns or dental care.

Imagine an algorithm designed to avoid overtreatment of cancer by withholding further treatment from patients who are predicted to die in a short interval. Let us suppose that those with historically poorer survival are from certain minority populations. If this survival were somehow biologically determined—for example, if resistance to therapy were related to the genetics of their tumors—then withdrawal of ineffective treatment might be appropriate. Differential outcomes by race or protected class are not ipso facto "unfair" if they arise from unalterable biologic differences. However, if poorer survival is the result of suboptimal treatment decisions for such groups in the past, then withholding such therapy would indeed be unfair. That distinction, however, requires understanding why survival was poor; failing to understand the underlying mechanism is bad science and may lead to poor predictions, poor decisions, and poor outcomes.

Another possibility is that poor survival, even with therapy, is poor because of factors like poor adherence, inability to afford transportation for follow-up visits, suboptimal home care, or poor nutrition. Understanding whether such factors affect prognosis—a major focus of health care disparity research—is critical to determining whether the algorithm provides ethically actionable predictions. This is the kind of knowledge that the oversight suggested by Rajkomar and colleagues would be expected to provide.

This article was published at Annals.org on 4 December 2018.

One might fairly say that requiring an understanding of the causal pathway between predictors and outcomes would vitiate the power of machine-learning techniques. However, medicine has a time-tested method for evaluating interventions whose causal mechanism we understand poorly: randomized trials. In the end, the ultimate test for any intervention is that more people are better off with than without it. To achieve distributive justice, we would add the requirement that some benefit accrue to all identifiable groups, particularly protected ones, and that harm to one subset is not being offset by benefits to another. This is the ultimate test of the "fairness" of algorithmically driven care, which also links to the goals of precision medicine. To achieve this, the only solution is to apply to artificial intelligence algorithms the very thing they are designed to supersede—human intelligence.

Steven N. Goodman, MD, MHS, PhD
Stanford University School of Medicine
Stanford, California

Sharad Goel, MS, PhD
Stanford University School of Engineering
Stanford, California

Mark R. Cullen, MD
Stanford University School of Medicine
Stanford, California

Disclosures: Authors have disclosed no conflicts of interest. Forms can be viewed at www.acponline.org/authors/icmje/ConflictOfInterestForms.do?msNum=M18-3297.

Corresponding Author: Steven N. Goodman, MD, MHS, PhD, Stanford University School of Medicine, 150 Governor's Lane, HRP/Redwood Building T265, Stanford, CA 94305; e-mail, Steve.goodman@stanford.edu.

Current author addresses are available at Annals.org.

Ann Intern Med. doi:10.7326/M18-3297

References

1. **Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH.** Ensuring fairness in machine learning to advance health equity. Ann Intern Med. 2018. [Epub ahead of print]. doi:10.7326/M18-1990
2. **Char DS, Shah NH, Magnus D.** Implementing machine learning in health care - addressing ethical challenges. N Engl J Med. 2018;378:981-983. [PMID: 29539284]
3. **Corbett-Davies S, Goel S.** The measure and mismeasure of fairness: a critical review of fair machine learning. Accessed at https://arxiv.org/abs/1808.00023 on 16 November 2018.
4. **Buolamwini J, Gebru T.** Gender shades: intersectional accuracy disparities in commercial gender classification. Proceedings of Machine Learning Research: Conference on Fairness, Accountability, and Transparency. 2018;81:1-15. Accessed at http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf on 26 November 2018.
5. **Pearl J, Mackenzie D.** The Book of Why: The New Science of Cause and Effect. New York: Basic Books; 2018.

Current Author Addresses: Dr. Goodman: Stanford University School of Medicine, 150 Governor's Lane, HRP/Redwood Building T265, Stanford, CA 94305.
Dr. Goel: Stanford School of Engineering, 475 Via Ortega, Stanford, CA 94305.
Dr. Cullen: Stanford Center for Population Health Sciences, 1070 Arastradero Road, Palo Alto, CA 94304.