Stanford University
Human-Centered
Artificial Intelligence

POLICY BRIEF
AUGUST 2022

SPECIAL HEALTHCARE SERIES

# Promoting Algorithmic Fairness in Clinical Risk Prediction

Stephen R. Pfohl, Agata Foryciarz, and Nigam H. Shah

**HEALTHCARE PROVIDERS AND MEDICAL PROFESSIONALS ARE INCREASINGLY USING MACHINE LEARNING TO ADVANCE HOW TREATMENT IS DELIVERED TO PATIENTS. From medical image analysis to a range of data processing functions, these machine learning applications will only continue to shape patient-care experiences and medical outcomes. Developers, doctors, patients, and policymakers are just some of the stakeholders grappling with these algorithmic uses.**

That said, there is a fundamental problem with machine learning in healthcare: We cannot assume developers are making strides to remedy bias and other fairness issues in a concerted manner. Discriminatory AI decision-making is concerning in any setting. This is especially pronounced in a clinical setting, where individuals' well-being and physical safety are on the line, and where medical professionals face life-or-death decisions every day.

Until now, the conversation about measuring algorithmic fairness in healthcare has focused on fairness itself—and has not fully taken into account how fairness

## KEY TAKEAWAYS

- We studied the trade-offs clinical predictive algorithms face between accuracy and fairness for outcomes like hospital mortality, prolonged stays in the hospital, and 30-day readmissions to the hospital. We found that techniques that make these programs more fair can degrade performance of the algorithm for everyone across the board.
- Making algorithmic fixes on the developer's side should only be one option considered to fix this. Policymakers should consider ways to incentivize model developers to engage in participatory design practices that incorporate perspectives from patient advocacy groups and civil society organizations.
- Algorithmic fixes may work in some contexts, but others may require policymakers to mandate that a human stays in the decision-making loop or the use of the algorithm may not be worthwhile at all.

techniques could impact clinical predictive models, which are often derived from large clinical datasets. Our new research, published in the *Journal of Biomedical Informatics*, seeks to ground this debate in evidence, and suggests the best way forward in developing fairer machine learning tools for a clinical setting.

We explicitly measure trade-offs in the fairness and performance of clinical predictive models. Using three large datasets spanning decades of health outcomes—such as hospital mortality, prolonged stays in the hospital, and 30-day readmissions to the hospital—our research compared these outcomes with three different notions of fairness across demographic groupings—such as race, ethnicity, sex, and age. In total, we find that improvements in algorithmic fairness, based on minimizing differences between demographic groups, cause lowered performance across multiple metrics. This exposes many challenges ahead in successfully mitigating bias in algorithms of the kind that has long plagued certain demographics within the United States.

Policymakers should recognize that there is no technical solution to address unfairness in clinical predictive models that does not decrease accuracy. Consequently, they should consider ways to incentivize responsible algorithm development alongside policies that address broader, structural healthcare inequities such as those caused by racism and socioeconomic inequality. The use of clinical predictive models must either be narrowly calibrated to a particular setting or constructed so that a human healthcare provider stays in the decision-making loop to ensure fair patient treatment. If machine learning models do not promote health equity, it may be appropriate to abstain from using an algorithm altogether.

# Introduction

There is much excitement about using machine learning and observational health data to guide clinical decision-making. Yet these tools have the potential to introduce and exacerbate health disparities for disadvantaged and underrepresented populations. Myriad realities contribute to this fact: inequity in historical and current patterns of care access and delivery, underrepresentation in clinical datasets, the use of biased or poorly calibrated statistical methods, and differences in the accessibility, usability, and effectiveness of predictive models across groups.

In response, considerable attention has been devoted to designing clinical predictive models to anticipate and proactively mitigate harm to the advancement of health equity, all while upholding ethical standards. Much of this research and the policy discussion surrounding it have centered on algorithmic fairness. In effect, algorithm fairness methods use mathematical formulas representative of an ideal state—such as equal error rates between male and female patients—and work to minimize deviations from that ideal. Researchers can then use this to audit predictive algorithms' problematic characteristics and promote transparency in their outputs.

Nonetheless, the debate about performance versus fairness in clinical algorithms tends to lose sight of the broader picture. A range of social, political, and economic factors contribute to massive health disparities in care access and quality. While algorithmic fairness techniques enable individuals to monitor and manipulate the outputs of predictive models, they are generally insufficient in and of themselves to mitigate the introduction or perpetuation of health disparities resulting from model-guided interventions.

Health disparities arise as a result of structural racism and related inequities in areas such as housing, education, employment, and criminal justice that affect healthcare access, utilization, and quality. These effects are further compounded by underrepresentation of the elderly, women, and ethnic minorities in clinical trials and cohort studies, as well as warped financial incentives in the healthcare system. At present, researchers often do not consider or explicitly discuss algorithmic fairness in this social, political, and economic context—a fact that is especially ironic when discussing definitions of group fairness. Instead, fairness criteria are defined primarily in technical terms related to a model's predictions, observed outcomes, and whether any of the patients in the data belong to a prespecified demographic group. What is more, the data is also evaluated on cohorts derived after the fact.

Stepping back, this means that algorithmic fairness criteria may be misleading, whether the observed outcome is a proxy for some underlying cause (e.g., structural racism in health pricing) or the predictive model is not appropriately contextualized in terms of the related interventions and policies. The conversation about algorithmic fairness thus erroneously confuses model performance and accrued benefits in health outcomes.

*While algorithmic fairness techniques enable individuals to monitor and manipulate the outputs of predictive models, they are generally insufficient in and of themselves to mitigate the introduction or perpetuation of health disparities resulting from model-guided interventions.*

# Research Outcomes

We conducted a large-scale study examining the trade-offs between algorithm performance and fairness criteria, specifically in clinical predictive models. Across 25 different combinations of datasets, clinical outcomes, and demographic attributes, we trained a series of predictive models to report on accuracy and group fairness metrics. Using different fairness metrics (e.g., conditional prediction parity, calibration, and cross-group ranking), we specified situations in which there was an imbalance within the social strata for the health outcome in question. The data was striking. Sometimes, proposed fairness methods made the algorithm's output "fairer" (by the definition used). But they also, in some cases, lowered the algorithm's across-the-board performance. In most cases, the original trained model produced unfair results. Predictions were better calibrated for some racial and ethnic groups than for others—or yielded different numbers of false positives and negatives. This, of course, is highly concerning in a clinical setting and could exacerbate harmful inequalities in the healthcare system.

To address these issues, we applied algorithmic fairness methods to the model and had some success. These changes improved performance: Error rates roughly equalized across groups, or predictions matched up better with outcomes in a narrowly defined scenario. But these fairness tweaks caused the model's overall predictive power to fall. Oftentimes, satisfying one notion of algorithmic fairness meant another would not be met. The possible fairness criteria did not all work in tandem.

Our research on fairness frameworks broadly yielded several findings. Common definitions of racial categories are entangled with historical and ongoing patterns of structural racism, and their continued use reinforces the idea of race as an accurate way to describe human variability, rather than a socially constructed taxonomy. Using these categorizations in algorithmic fairness work thus raises numerous ethical questions. Further, marginalized groups are often not well-represented by the attributes used to assess group fairness, including intersectional identities. Group fairness and algorithmic fairness criteria also address individual attributes (e.g., race, gender) independently and consequently treat them as abstract, interchangeable constructs—without awareness of meaningful contextual differences between them. For example, observed differences on the basis of race should be primarily interpreted as deriving from systemic and structurally racist factors. By contrast, those observed on the basis of sex could be erroneously attributed to clinically meaningful differences in human physiology.

# Policy Discussion

Striving for health equity requires designing policies that directly counteract the systemic factors underlying health disparities—primarily structural forms of racism and economic inequality. Algorithms are a bigger and bigger part of this landscape. Yet policymakers must realize the limits of, and problems with, current approaches to algorithmic fairness.

*...achieving algorithmic fairness should be defined in terms of the impact an algorithm-guided intervention has on individuals, groups, and status quo power structures...*

By considering only technical fixes to observable algorithmic properties, evaluating models through the group fairness framework ignores systemic issues and related second- and third-order effects on health disparities. It also ignores inequities in the data generation and measurement processes that inform how algorithms work. And it misses the causal framing and decision-making that, in practice, connects algorithmic predictions to clinical decisions—like when a medical professional decides how to use algorithmic outputs.

From a research and a policy standpoint, requiring predictive models to satisfy a notion of group fairness provides little more than a "veneer of neutrality." Constraining a model in a way that achieves group fairness is insufficient for, and may even work against,

promoting health equity via machine learning-guided interventions. To be clear, optimizing the satisfaction of fairness criteria can still be useful. But achieving algorithmic fairness should be defined in terms of the impact an algorithm-guided intervention has on individuals, groups, and status quo power structures that directly or indirectly perpetuate health disparities.

In light of these limitations, policymakers must consider ways to incentivize developers to use participatory design practices that explicitly incorporate a diverse set of stakeholder perspectives. For instance, this should include patient advocacy groups and civil rights organizations. Doing so will allow developers to better identify the mechanisms through which measurement error, bias, and historical inequities affect data collection, measurement, and problem formulation. Participatory design would also enable developers to better understand the relationship between those technical measurements and policy and other interventions in clinical settings.

All the while, it should not be assumed that algorithm-aided decision-making is always helpful in healthcare. If it is not practical to employ algorithm-aided decision-making responsibly, the better approach may be to abstain from using it altogether. For example, our recent study found that recalibrating risk models for subgroups—to better match outcomes—increased gaps between groups' false positive and negative rates. Using an equalized odds approach, meanwhile, to equalize error rates for all groups, better matched the guidelines we looked at yet created different error rates for patients in high-risk categories.

Pursuing this context-driven understanding of algorithm development and algorithmic fairness in healthcare could help developers overcome the limitations of the current group fairness framework. As AI increasingly influences decisions about our health and our lives, it is crucial that we work to accelerate these efforts to broaden the research and policy community's understanding of algorithmic fairness.

*As AI increasingly influences decisions about our health and our lives, it is crucial that we work to accelerate these efforts to broaden the research and policy community's understanding of algorithmic fairness.*

The original article, "**An Empirical Characterization of Fair Machine Learning for Clinical Risk Prediction**," can be accessed at: https://www.sciencedirect.com/science/article/pii/S1532046420302495?via%3Dihub.

Stanford University's Institute for Human-Centered Artificial Intelligence (HAI) applies rigorous analysis and research to pressing policy questions on artificial intelligence. A pillar of HAI is to inform policymakers, industry leaders, and civil society by disseminating scholarship to a wide audience. HAI is a nonpartisan research institute, representing a range of voices. The views expressed in this policy brief reflect the views of the authors. For further information, please contact **HAI-Policy@stanford.edu**.


**Stephen R. Pfohl** is a research scientist at Google and a recent Ph.D. graduate of Biomedical Informatics at Stanford University.

**Agata Foryciarz** is a Ph.D. candidate in Computer Science at Stanford University.

**Nigam H. Shah** is professor of medicine (biomedical informatics) and of biomedical data science at Stanford University. He serves as chief data scientist for Stanford Healthcare and is a faculty affiliate at HAI.



**Stanford University**
Human-Centered
Artificial Intelligence



**Stanford HAI:** Cordura Hall, 210 Panama Street, Stanford, CA 94305-1234

**T** 650.725.4537 **F** 650.123.4567 **E** HAI-Policy@stanford.edu **hai.stanford.edu**