AEA Papers and Proceedings 2021, 111: 37–42
https://doi.org/10.1257/pandp.20211078

NEW APPROACHES TO DETECTING DISCRIMINATION[‡]

On the Inequity of Predicting A While Hoping for B[†]

By Sendhil Mullainathan and Ziad Obermeyer*

One of the most influential papers in management is perfectly summarized by its title: "On the Folly of Rewarding A, While Hoping for B" (Kerr 1975). We incentivize teachers on test scores, and we get higher test scores, not necessarily more learning. We pay hospital systems to deliver treatments, and we get more utilization, not necessarily healthier patients. A similar maxim applies to prediction. We build algorithms to predict Y, and they will optimize that objective, no matter whether our objective was Y or Y^*. This discrepancy can be a major driver of algorithmic bias. We show this using examples from health care, where algorithms are already widely deployed and influence life and death decisions. But the forces we consider apply broadly, to a range of other important social sectors where algorithms are increasingly used.

The specific mechanism of bias we consider arises from label choice: the choice of a biased proxy as the algorithm's prediction target.[1] It is distinct from other mechanisms in the literature, for example when an algorithm trained on one group fails to generalize to another (e.g., pulse oximeter devices fail to detect dangerously low oxygen levels in darker skin). It is likewise unrelated to the decision to include or exclude race as a model predictor. Biased labels can induce bias even if race is excluded—as in the empirical example below. More extreme measures, such as excluding race and further ensuring that model predictions are orthogonal to race—also the case in our example below—will likewise fail to correct label choice bias. And conversely, inclusion of race as a predictor will not necessarily induce bias, unless the label is also biased.

We focus on label choice for two reasons. First, we have found that, relative to its importance as a source of bias, it is underappreciated. Second, despite the large-scale distortions it induces, it is difficult to detect. Unlike failures of model generalization, which are straightforward to check by comparing predictive performance across groups, exposing label choice bias requires more detective work. It requires understanding how algorithms influence decision-making in context and how structural biases affect measurement of the label.

I. An Empirical Example of Label Choice Bias

We illustrate the importance and the subtlety of label choice bias with an example of large-scale racial bias in an algorithm used to target extra help to patients with complex medical needs (Obermeyer et al. 2019). Nearly every major health system uses "high-risk care management" programs to help high-risk patients manage chronic illnesses. The goal is to prevent exacerbations of these illnesses and thereby reduce associated costs, e.g., from emergency and hospital utilization, making this a win-win for patients and the health system. Because program resources are themselves costly, algorithms have come into wide use for targeting programs to those who need them most. Industry estimates suggest that 150 million patients are screened every year by algorithms that trawl through

[‡]*Discussants:* Janet Currie, Princeton University; Max Tabord-Meehan, University of Chicago; Avi Feller, University of California-Berkeley.

* Mullainathan: Chicago Booth (sendhil@chicagobooth.edu); Obermeyer: UC Berkeley Public Health (zobermeyer@berkeley.edu), to whom correspondence should be addressed. We are grateful to our collaborators, Emily Bembeneck, Stephanie Eaneff, and Michael Stern, for their work with our partner organizations in health; to Stephen Lamb; and to the Center for Applied Artificial Intelligence at Chicago Booth for financial support.

[†]Go to https://doi.org/10.1257/pandp.20211078 to visit the article page for additional materials and author disclosure statement(s).

[1] We follow the convention in machine learning of denoting the dependent variable the label.

patients' health records to generate risk scores. In the setting we study, as in every other setting we have worked since, algorithm scores are used to screen out low-risk patients from consideration, prompt clinicians to consider enrolling others, and screen in the highest-risk patients.

How would we define algorithmic bias in this setting? Many measures of bias have been proposed, but these are inconsistent—two plausible-sounding measures can paint a contradictory picture of the extent of bias (Kleinberg, Mullainathan, and Raghavan 2016). We thus choose a measure that captures the consequences of bias for patients, in terms of who gets access to extra resources in light of their health needs. Denote the risk score S, the intervention it is used to help allocate T, patient i's health H, and a vector of (pre-treatment) patient covariates X, including indicator B for membership in a protected group (we consider, without loss of generality, a simple example with only one group). We consider a thought experiment with two patients, one Black and one white. If they have the same health, and thus the same health needs, does the algorithm score them similarly? Alternatively, if they have the same algorithm score, do they go on to have the same health needs? This definition of bias captures disparities in needs given equal treatment and is akin to "calibration": it compares $E\left[H \mid S, B = 0\right]$ to $E\left[H \mid S, B = 1\right]$. At the core of this definition is the fact that those with the same algorithm score have the same likelihood of getting extra help with their health needs. So they should have the same needs: an unbiased algorithm should assign the same score to patients with the same health needs, irrespective of race.

To measure health needs, we assembled data on a range of health outcomes over the year following algorithm score assignment and compared these for Black and white patients with similar scores. We found that Black patients went on to have far greater health needs than white patients. Figure 1 (analysis of data also appearing in Obermeyer et al. 2019, where links to replication data can be found) shows this for one important measure of health: the number of exacerbations of chronic illnesses (y-axis) was much higher for Black patients conditional on risk score (x-axis). The magnitude of bias was such that an unbiased algorithm would have increased the fraction of Black patients fast tracked into the program, from 17.7 to 46.5 percent. At 17.7 percent, Black patients were already overrepresented in the fast track relative to the base rate of 12.3 percent in this population. This highlights the need for meaningful measures of bias: targets based on identity criteria alone, like population representation, can understate (or overstate) its extent.

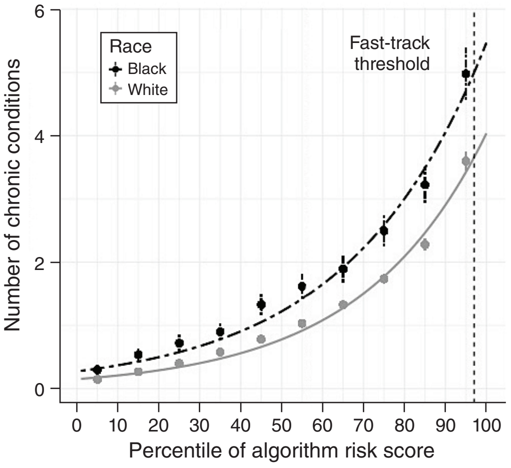

Figure 1. Exacerbations of Chronic Illness by Race (y-axis) versus Risk Score (x-axis; percentile)

II. Measurement Error and Biased Proxies

Where did the algorithm go wrong? One important clue can be found in where it went right: its performance for predicting health-care costs over the next year was accurate and unbiased. In fact, the algorithm was specifically trained to predict health-care costs, a target that is subtly but importantly different from the one articulated in the developer's promotional materials:[2] to "determine which individuals are in need of specialized intervention programs and which intervention programs are likely to have an impact on the quality of individuals' health."

An issue increasingly raised in the machine learning literature is how to translate semantic statements about an algorithm's goal into a well-defined objective function calculable in some dataset (Passi and Barocas 2019). The aforementioned text implies that the algorithm estimates individual-level treatment effects

[2] https://www.optum.com/content/dam/optum3/optum/en/resources/sell-sheet/impact-pro-sell-sheet.pdf.

of interventions on a patient's health. Using potential outcomes notation to capture treated and untreated health outcomes H_1 and H_0, such an ideal risk score could be written $S_i^* = E\left[H_{1i} - H_{0i} | X_i\right]$. But given the difficulty of this estimation problem, the algorithm's creators make a simplifying assumption: those with the greatest (untreated) health needs will have the greatest benefit from interventions. This turns a challenging causal inference task into a straightforward prediction problem (Kleinberg et al. 2015) and yields the simpler risk score, $S_i = E\left[Y_{0i} | X_i\right]$ (we will omit the subscripts and potential outcomes from now on), where Y is a proxy for future health-care needs.

But which proxy to use? "Health needs"—and even "health"—is a latent variable with no simple empirical definition. So the algorithm designers make another critical assumption: that future health-care costs C are a good proxy for health needs today. Cost is an appealing variable to use: it is clearly correlated with health needs, present in the large datasets owned by insurers, and available to researchers. It requires no laborious cleaning; missing values are zero. So it is comparatively easy for a data science team to produce an algorithm that predicts $S = E\left[C | X\right]$.

Of course, while costs and health needs are correlated, they are not the same:

$$C = H + \Delta.$$

As a long tradition of research has shown, costs vary widely across hospitals and geographies with similar health outcomes. That practice variation and overuse contributes to these trends is well established. But underuse also contributes: a patient's lack of knowledge that subtle squeezing in the chest can be a heart attack, lack of access to health care or insurance, or differential treatment by doctors.

In other words, despite being a reasonable proxy for health, cost is a biased one: the Δ term is not random with respect to socioeconomic and racial variables like B. Because of structural biases and differential treatment, Black patients with similar needs to white patients have long been known to have lower costs: $E\left[C | H, B = 1\right] < E\left[C \mid H, B = 0\right]$. So algorithmic risk scores based on $E\left[C | X\right]$ will build in a large negative bias for Black patients, because $E\left[\Delta | B = 1\right] < E\left[\Delta | B = 0\right]$.

More generally, whenever an algorithm's literal target Y differs from its true target Y^*—often because the true target is unmeasured—we can write that

$$Y = Y^* + \Delta.$$

Algorithms that predict Y will automate the error Δ along with the true—but mismeasured—signal Y^* (Mullainathan and Obermeyer 2017). This idea has two implications for algorithmic bias. First, algorithmic predictions can be more biased than the original variable they predict. If Δ is more predictable with X than Y^* is, predictions will be dominated by Δ, not Y^*. And since Δ is often an all-too-simple function of obvious socioeconomic and racial inequities, we might expect $\text{cov}\left(Y^*, B\right) < \text{cov}\left(\Delta, B\right)$. So, automating undesirable parts of a target variable will impact some groups more than others, creating label choice bias. Importantly, such predictions will appear accurate and show no bias when evaluated on traditional metrics of loss of the form $L\left(S, Y\right)$: Δ is just as much a part of the measured Y variable as Y^* is. Second, this bias can arise even if Y and Y^* are highly correlated. If the variances of Y and Y^* are large, only a small fraction of the total variance will be due to race. However, the predictable variance in Y due to B can be arbitrarily large.

III. Conflating Costs and Needs

Why did the algorithm's creators choose to predict costs rather than needs? An easy explanation is that the developer, and the health systems that purchase its software, care more about costs than patients' health. So our results could represent a "prediction externality," where society cares about health, but private companies care only about costs. In this setting, an algorithm that accurately predicts C allows companies to optimize purely on this metric and increases the size of the externality $\left(H - C\right)$.

We cannot say definitively how the decision was made. But our experience, as well as evidence from elsewhere in the health sector, points to a less nefarious—but perhaps more concerning—reason: a broader tendency to conflate health-care costs with health needs, well beyond the algorithm we study. In fact, the company that developed the algorithm was highly motivated to understand and solve the problem we identified. After we first noted bias in the algorithm, we initiated an (unpaid) collaboration with them to understand and correct the problem. All of our interactions with the technical team that developed the algorithm, and the executives who authorized and supported the work, indicated that their intention was not to optimize cost prediction at the expense of health. That collaboration led to a revised version of the algorithm with far less bias. The only change was that the new algorithm predicted outcomes related to health, e.g., exacerbations of chronic illnesses in the same dataset used to train the original algorithm, not cost alone. As a bonus, because health and cost are correlated, this revised algorithm still performed reasonably well in predicting costs. The intuition is that there are many possible functions, all producing correlated predictions but having different correlations with race (Obermeyer et al. 2019).

The enterprise of predicting cost, as opposed to health, goes well beyond one company's algorithm. Promotional materials for algorithms, including but not limited to the one we studied, are very transparent about the fact that they predict costs. A review of the ten most widely used algorithms for targeting care management, conducted by the Society of Actuaries in 2016, judged them explicitly by their accuracy for cost prediction. Two of these ten were developed by academic groups, one by the Centers for Medicare and Medicaid Services. These organizations did not catch the racial bias induced by cost prediction, nor did the health systems that purchased the tools—many of which have deep commitments to reducing health disparities. Nor did the physicians who had every opportunity to overrule the algorithm, nor did the millions of patients whose care was affected. We have replicated our finding of bias in another one of these widely used algorithms, this time deployed at a large health insurer: a nonprofit entity that employs a diverse team of full-time ethicists, who work with their clinicians and data scientists, and is advised by an ethics advisory group that actively incorporates input from employees, patients, and others. Despite this, a biased algorithm was in wide use for years.

We believe these biases go unsuspected and undetected because the conflation of costs with needs is common in health care. But it is just one instance of a deeper problem in many social sectors: making inferences about complex latent variables via biased proxies. In criminal justice, arrests and convictions are not the same as someone's propensity for crime. In education, test scores are not the same as teacher and student ability. In employment, interview and peer ratings are not the same as employee quality. Despite Goodhart's cautionary law, it is all too common for such measures to become targets. As algorithms begin to automate these targets, biases and other distortions can scale rapidly, unchecked by human judgment.

IV. Label Choice Bias Elsewhere in Health Care

Label choice bias is not limited to health algorithms: it also manifests in health policy, where the stakes can be even higher. The recent CARES Act for COVID-19 relief, for example, was passed by Congress to address two urgent needs: compensating health providers for the expenses of caring for patients with COVID-19 and offsetting revenue loss from reduced utilization due to the pandemic. These stated needs (Y^*) were then formulated in terms of measurable quantities (Y) in order to guide allocation of the $175 billion dollars of funding to health systems. The Y variable chosen: a provider's total revenue the year before COVID-19.

This label choice produced two distortions. First, many hard-hit areas received far less funding than they needed based on a range of other important measures of need: COVID-19 burden, baseline population illness, or area hospitals' financial distress. Other areas, largely spared by the pandemic and with well-resourced hospitals, received so much funding that some hospitals (e.g., Hospital Corporation of America, Kaiser Permanente) announced they were giving it back. It is unusual, to say the least, for hospitals to refund money to the federal government—a testament to the extent of the misallocation. Second, the degree to which regions were over- or underfunded was highly correlated with race, an instance of label choice bias. Comparing counties receiving the same amount of funding, Black counties had twice the COVID-19 burden, as shown in Figure 2 (analysis of data also appearing in Kakani et al. forthcoming, where links to replication data can be found). The decision to allocate funding proportional to hospital revenue or utilization is not unusual in health: Medicare programs, such as the Disproportionate Share and 340B drug rebate

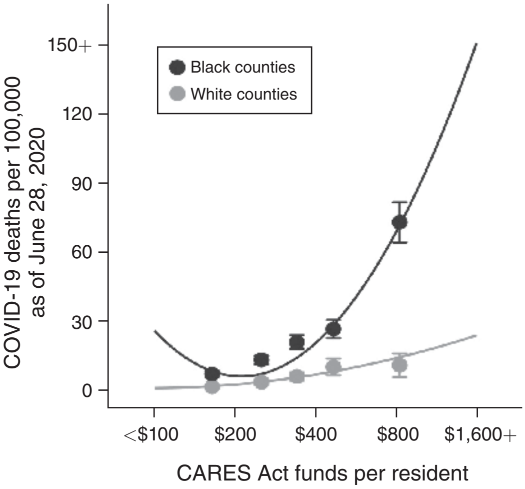

Figure 2. COVID-19 Deaths (y-axis) versus CARES Act Funding (x-axis)

programs, do the same, likely inducing similar biases in tens of billions of dollars of subsidies every year.

Several other instances of label choice bias have emerged from our collaborations with large organizations in health, including hospital systems, for- and nonprofit insurers, state and federal agencies, software companies, and others. As part of these partnerships, we conduct an inventory of algorithms deployed at each organization and grade each on its potential for bias. A key task is to articulate the exact target an algorithm is predicting and compare it with what an ideal target would be for a given decision. We highlight in Table 1 some particularly important and widespread examples of label choice bias that were brought to light through this process.

The first three examples relate to bias in triage tools. The Emergency Severity Index is used by triage nurses in emergency departments across the world. It assigns patients a score that dictates how long they can safely wait before seeing a doctor. The two prediction targets are the nurse's perception of the patient's acuity and expectation of how many "resources the patient is expected to consume." There are growing numbers of attempts to automate this score, which risk automating known disparities between groups in resource use (conditional on acuity), based on insurance, language, race, and access to care. Second, the "6-Clicks Mobility Score" measures a patient's objective mobility and ability to perform activities. It is used to guide planning for hospital discharge: patients with low mobility, who may be unable to care for themselves independently at home, are considered for placement in an assisted living or rehabilitation facility. But, of course, two patients with identical mobility may have very different abilities to care for themselves because of access to transportation, family support, home amenities, and other factors linked to income. Third, algorithms are commonly used to optimize outpatient clinic schedules by identifying patients who are likely to choose to skip a scheduled appointment. Slots are then double booked to maximize clinic productivity. But not all no-shows are voluntary: some patients fail to appear because of barriers to access, or even worsening health. The optimal decision for these no-shows, which are more common in disadvantaged groups, might not be to simply reallocate the clinic slot to another patient.

The last two examples concern bias in tools that predict onset of disease. A wide variety of algorithms are trained to identify patients who will go on to develop a disease, like diabetes or congestive heart failure, in order to target preventative care measures. These algorithms often predict the occurrence of International Classification of Disease (ICD) codes, which are appealing because they are widely available in the electronic records and claims data. But these codes, which are produced by transactions between health systems and insurers, are as much financial data as medical data. Predicting them can thus automate both components: signal for disease but also incentives to "up-code," which vary by hospitals' billing and coding resources or the insurance of the patients they serve. It can also automate underdiagnosis, particularly in patients who lack access to care or those who are misunderstood or dismissed by providers because of language, race, or education (Mullainathan and Obermeyer 2016). Finally, many machine vision algorithms take in a medical image, like an x-ray, and output a radiologist's interpretation of that x-ray. Because hospital datasets contain images paired with the radiologist's reports, the latter is again an appealing label based on its easy accessibility. But human interpretations are not ground truth. They too reflect varying hospital incentives to overcall findings and can also miss important findings in some patient groups. This was illustrated in a recent example where an algorithm

Table 1

	Ideal target: Y^*	Actual target: Y	Source of bias: $\mathrm{cov}(\Delta, B)$
Emergency Severity Index: emergency triage	Medical condition needing immediate attention	Nurse-rated acuity, "resources patient is expected to consume"	Resource consumption varies by race and insurance, conditional on acuity
6-Clicks Mobility Score: decisions about discharge destination	Inability to care for self at home without help	Physical measures of mobility and daily activities	Similar physical mobility scores have a larger impact on those lacking income
No-show prediction: clinic scheduling	Voluntary no-show to appointment	Any no-show to appointment	No-shows due to access barriers unequally distributed
Predicting disease onset: targeting preventive care	New disease onset (e.g., heart failure, kidney failure)	Provider-insurer transaction with ICD code for disease	$\Pr(Y \mid Y^*)$ varies by physician quality, hospital billing and coding, insurance, etc.
Kellgren-Lawrence grade: osteoarthritis on knee x-rays	Severity of knee osteoarthritis	Severity of osteoarthritis seen by radiologist on knee x-rays	Radiologists miss causes of knee pain affecting underserved groups

was trained to quantify the severity of knee osteoarthritis by predicting patients' pain ratings from their knee x-rays (Pierson et al. 2021). When this algorithm's severity measure was compared to the traditional severity measure, based on the radiologist's judgment, the algorithm explained a higher proportion of patients' symptoms. This new explanatory power was particularly helpful for Black patients: it reduced the Black-white gap in unexplained pain by nearly half.

V. Conclusions

Label choice is a major channel by which algorithms reproduce and scale up bias. But judicious label choice can also turn algorithms into a force for fighting bias. Finding the right labels to predict, of the hundreds or thousands of variables available in health datasets, is thus a particularly high-value activity—but one that is dramatically underresourced in current algorithm development pipelines. The applied empirical tool kit is likely to be useful in this regard. Understanding the data-generating process, acknowledging the nature of health data as both biological and social phenomena, and measuring and addressing disparities are all important considerations for those seeking to use algorithms for social good.

REFERENCES

Kakani, Pragya, Amitabh Chandra, Sendhil Mullainathan, and Ziad Obermeyer. Forthcoming. "Allocation of COVID-19 Relief Funding to Disproportionately Black Counties." *JAMA*. https://doi.org/10.1001/jama.2020.14978.

Kerr, Steven. 1975. "On the Folly of Rewarding A, While Hoping for B." *Academy of Management Journal* 18 (4): 769–783.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. 2015. "Prediction Policy Problems." *American Economic Review* 105 (5): 491–95.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2016. "Inherent Trade-Offs in the Fair Determination of Risk Scores." https://arxiv.org/abs/1609.05807.

Mullainathan, Sendhil, and Ziad Obermeyer. 2017. "Does Machine Learning Automate Moral Hazard and Error?" *American Economic Review* 107 (5): 476–80.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations." *Science* 366 (6464): 447–53.

Passi, Samir, and Solon Barocas. 2019. "Problem Formulation and Fairness." In *Proceedings of the Conference on Fairness, Accountability, and Transparency,* 39–48. New York: Association for Computing Machinery.

Pierson, Emma, David M. Cutler, Jure Leskovec, Sendhil Mullainathan, and Ziad Obermeyer. 2021. "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations." *Nature Medicine* 27 (1): 136–40.