RESEARCH ARTICLE

ECONOMICS

Dissecting racial bias in an algorithm used to manage the health of populations

Ziad Obermeyer[1,2]*, Brian Powers[3], Christine Vogeli[4], Sendhil Mullainathan[5]*†

Health systems rely on commercial prediction algorithms to identify and help patients with complex health needs. We show that a widely used algorithm, typical of this industry-wide approach and affecting millions of patients, exhibits significant racial bias: At a given risk score, Black patients are considerably sicker than White patients, as evidenced by signs of uncontrolled illnesses. Remedying this disparity would increase the percentage of Black patients receiving additional help from 17.7 to 46.5%. The bias arises because the algorithm predicts health care costs rather than illness, but unequal access to care means that we spend less money caring for Black patients than for White patients. Thus, despite health care cost appearing to be an effective proxy for health by some measures of predictive accuracy, large racial biases arise. We suggest that the choice of convenient, seemingly effective proxies for ground truth can be an important source of algorithmic bias in many contexts.

There is growing concern that algorithms may reproduce racial and gender disparities via the people building them or through the data used to train them (*1*–*3*). Empirical work is increasingly lending support to these concerns. For example, job search ads for highly paid positions are less likely to be presented to women (*4*), searches for distinctively Black-sounding names are more likely to trigger ads for arrest records (*5*), and image searches for professions such as CEO produce fewer images of women (*6*). Facial recognition systems increasingly used in law enforcement perform worse on recognizing faces of women and Black individuals (*7*, *8*), and natural language processing algorithms encode language in gendered ways (*9*).

Empirical investigations of algorithmic bias, though, have been hindered by a key constraint: Algorithms deployed on large scales are typically proprietary, making it difficult for independent researchers to dissect them. Instead, researchers must work "from the outside," often with great ingenuity, and resort to clever workarounds such as audit studies. Such efforts can document disparities, but understanding how and why they arise—much less figuring out what to do about them—is difficult without greater access to the algorithms themselves. Our understanding of a mechanism therefore typically relies on theory or exercises with researcher-created algorithms (*10*–*13*). Without an algorithm's training data, objective function, and prediction methodology, we can only guess as to the actual mechanisms for the important algorithmic disparities that arise.

In this study, we exploit a rich dataset that provides insight into a live, scaled algorithm deployed nationwide today. It is one of the largest and most typical examples of a class of commercial risk-prediction tools that, by industry estimates, are applied to roughly 200 million people in the United States each year. Large health systems and payers rely on this algorithm to target patients for "high-risk care management" programs. These programs seek to improve the care of patients with complex health needs by providing additional resources, including greater attention from trained providers, to help ensure that care is well coordinated. Most health systems use these programs as the cornerstone of population health management efforts, and they are widely considered effective at improving outcomes and satisfaction while reducing costs (*14*–*17*). Because the programs are themselves expensive—with costs going toward teams of dedicated nurses, extra primary care appointment slots, and other scarce resources—health systems rely extensively on algorithms to identify patients who will benefit the most (*18*, *19*).

Identifying patients who will derive the greatest benefit from these programs is a challenging causal inference problem that requires estimation of individual treatment effects. To solve this problem, health systems make a key assumption: Those with the greatest care needs will benefit the most from the program. Under this assumption, the targeting problem becomes a pure prediction policy problem (*20*). Developers then build algorithms that rely on past data to build a predictor of future health care needs.

Our dataset describes one such typical algorithm. It contains both the algorithm's predictions as well as the data needed to understand its inner workings: that is, the underlying ingredients used to form the algorithm (data, objective function, etc.) and links to a rich set of outcome data. Because we have the inputs, outputs, and eventual outcomes, our data allow us a rare opportunity to quantify racial disparities in algorithms and isolate the mechanisms by which they arise. It should be emphasized that this algorithm is not unique. Rather, it is emblematic of a generalized approach to risk prediction in the health sector, widely adopted by a range of for- and non-profit medical centers and governmental agencies (*21*).

Our analysis has implications beyond what we learn about this particular algorithm. First, the specific problem solved by this algorithm has analogies in many other sectors: The predicted risk of some future outcome (in our case, health care needs) is widely used to target policy interventions under the assumption that the treatment effect is monotonic in that risk, and the methods used to build the algorithm are standard. Mechanisms of bias uncovered in this study likely operate elsewhere. Second, even beyond our particular finding, we hope that this exercise illustrates the importance, and the large opportunity, of studying algorithmic bias in health care, not just as a model system but also in its own right. By any standard—e.g., number of lives affected, life-and-death consequences of the decision—health is one of the most important and widespread social sectors in which algorithms are already used at scale today, unbeknownst to many.

Data and analytic strategy

Working with a large academic hospital, we identified all primary care patients enrolled in risk-based contracts from 2013 to 2015. Our primary interest was in studying differences between White and Black patients. We formed race categories by using hospital records, which are based on patient self-reporting. Any patient who identified as Black was considered to be Black for the purpose of this analysis. Of the remaining patients, those who self-identified as races other than White (e.g., Hispanic) were so considered (data on these patients are presented in table S1 and fig. S1 in the supplementary materials). We considered all remaining patients to be White. This approach allowed us to study one particular racial difference of social and historical interest between patients who self-identified as Black and patients who self-identified as White without another race or ethnicity; it has the disadvantage of not allowing for the study of intersectional racial

[1]School of Public Health, University of California, Berkeley, Berkeley, CA, USA. [2]Department of Emergency Medicine, Brigham and Women's Hospital, Boston, MA, USA. [3]Department of Medicine, Brigham and Women's Hospital, Boston, MA, USA. [4]Mongan Institute Health Policy Center, Massachusetts General Hospital, Boston, MA, USA. [5]Booth School of Business, University of Chicago, Chicago, IL, USA.
*These authors contributed equally to this work.
†Corresponding author. Email: sendhil.mullainathan@chicagobooth.edu

and ethnic identities. Our main sample thus consisted of (i) 6079 patients who self-identified as Black and (ii) 43,539 patients who self-identified as White without another race or ethnicity, whom we observed over 11,929 and 88,080 patient-years, respectively (1 patient-year represents data collected for an individual patient in a calendar year). The sample was 71.2% enrolled in commercial insurance and 28.8% in Medicare; on average, 50.9 years old; and 63% female (Table 1).

For these patients, we obtained algorithmic risk scores generated for each patient-year. In the health system we studied, risk scores are generated for each patient during the enrollment period for the system's care management program. Patients above the 97th percentile are automatically identified for enrollment in the program. Those above the 55th percentile are referred to their primary care physician, who is provided with contextual data about the patients and asked to consider whether they would benefit from program enrollment.

Many existing metrics of algorithmic bias may apply to this scenario. Some definitions focus on calibration [i.e., whether the realized value of some variable of interest Y matches the risk score R (*2*, *22*, *23*)]; others on statistical parity of some decision D influenced by the algorithm (*10*); and still others on balance of average predictions, conditional on the realized outcome (*22*). Given this multiplicity and the growing recognition that not all conditions can be simultaneously satisfied (*3*, *10*, *22*), we focus on metrics most relevant to the real-world use of the algorithm, which are related to calibration bias [formally, comparing Blacks B and Whites W, $E[Y|R,W] = E[Y|R,B]$ indicates the absence of bias (here, E is the expectation operator)]. The algorithm's stated goal is to predict complex health needs for the purpose of targeting an intervention that manages those needs. Thus, we compare the algorithmic risk score for patient i in year t ($R_{i,t}$), formed on the basis of claims data $X_{i,(t-1)}$ from the prior year, to data on patients' realized health $H_{i,t}$, assessing how well the algorithmic risk score is calibrated across race for health outcomes $H_{i,t}$. We also ask how well the algorithm is calibrated for costs $C_{i,t}$.

To measure H, we link predictions to a wide range of outcomes in electronic health record data, including all diagnoses (in the form of International Classification of Diseases codes) as well as key quantitative laboratory studies and vital signs capturing the severity of chronic illnesses. To measure C, we link predictions to insurance claims data on utilization, including outpatient and emergency visits, hospitalizations, and health care costs. These data, and the rationale for the specific measures of H used in this study, are described in more detail in the supplementary materials.

Health disparities conditional on risk score

We begin by calculating an overall measure of health status, the number of active chronic conditions [or "comorbidity score," a metric used extensively in medical research (*24*) to provide a comprehensive view of a patient's health (*25*)] by race, conditional on algorithmic risk score. Fig. 1A shows that, at the same level of algorithm-predicted risk, Blacks have significantly more illness burden than Whites. We can quantify these differences by choosing one point on the x axis that corresponds to a very-high-risk group (e.g., patients at the 97th percentile of risk score, at which patients are auto-identified for program enrollment), where Blacks have 26.3% more chronic illnesses than Whites (4.8 versus 3.8 distinct conditions; $P < 0.001$).

What do these prediction differences mean for patients? Algorithm scores are a key input to decisions about future enrollment in a care coordination program. So as we might expect, with less-healthy Blacks scored at similar risk scores to more-healthy Whites, we find evidence

Table 1. Descriptive statistics on our sample, by race. BP, blood pressure; LDL, low-density lipoprotein.

	White	Black
n (patient-years)	88,080	11,929
n (patients)	43,539	6079
Demographics		
Age	51.3	48.6
Female (%)	62	69
Care management program		
Algorithm score (percentile)	50	52
Race composition of program (%)	81.8	18.2
Care utilization		
Actual cost	$7540	$8442
Hospitalizations	0.09	0.13
Hospital days	0.50	0.78
Emergency visits	0.19	0.35
Outpatient visits	4.94	4.31
Mean biomarker values		
HbA1c (%)	5.9	6.4
Systolic BP (mmHg)	126.6	130.3
Diastolic BP (mmHg)	75.5	75.7
Creatinine (mg/dl)	0.89	0.98
Hematocrit (%)	40.7	37.8
LDL (mg/dl)	103.4	103.0
Active chronic illnesses (comorbidities)		
Total number of active illnesses	1.20	1.90
Hypertension	0.29	0.44
Diabetes, uncomplicated	0.08	0.22
Arrythmia	0.09	0.08
Hypothyroid	0.09	0.05
Obesity	0.07	0.18
Pulmonary disease	0.07	0.11
Cancer	0.07	0.06
Depression	0.06	0.08
Anemia	0.05	0.10
Arthritis	0.04	0.04
Renal failure	0.03	0.07
Electrolyte disorder	0.03	0.05
Heart failure	0.03	0.05
Psychosis	0.03	0.05
Valvular disease	0.03	0.02
Stroke	0.02	0.03
Peripheral vascular disease	0.02	0.02
Diabetes, complicated	0.02	0.07
Heart attack	0.01	0.02
Liver disease	0.01	0.02

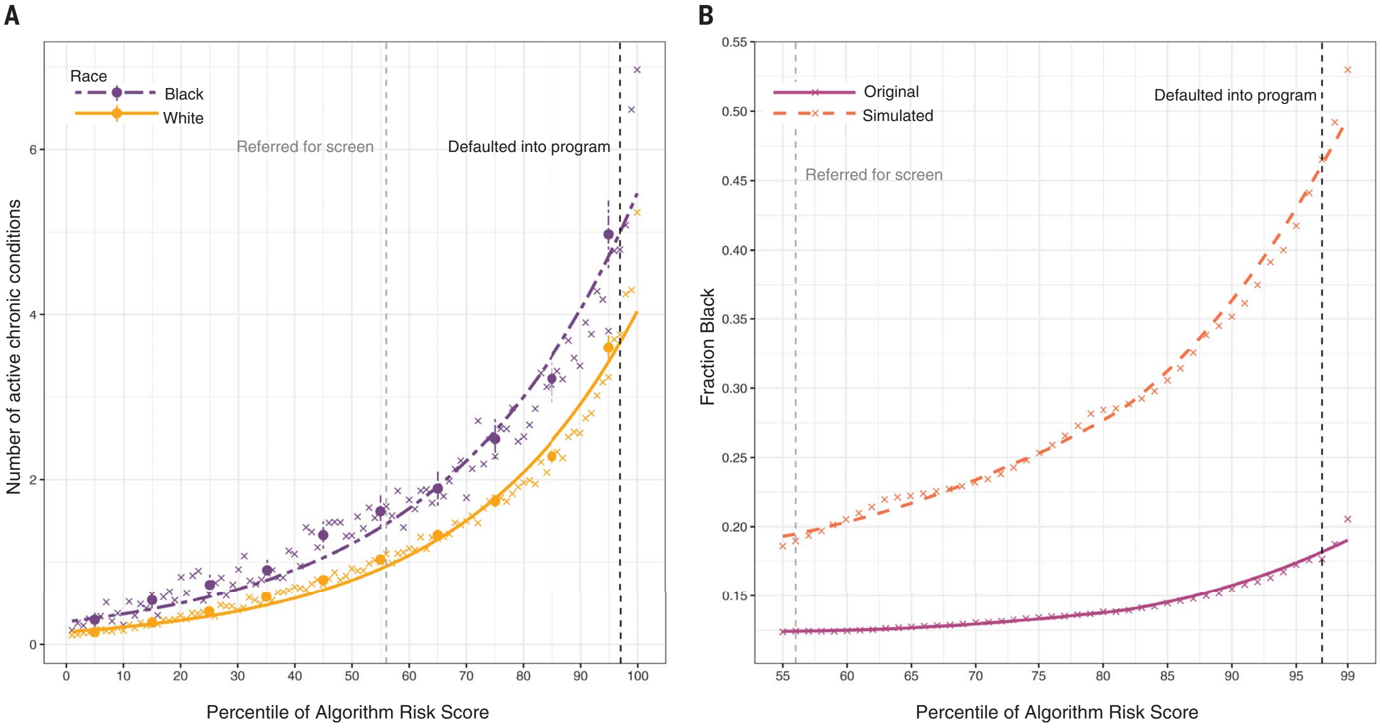

Fig. 1. Number of chronic illnesses versus algorithm-predicted risk, by race. (**A**) Mean number of chronic conditions by race, plotted against algorithm risk score. (**B**) Fraction of Black patients at or above a given risk score for the original algorithm ("original") and for a simulated scenario that removes algorithmic bias ("simulated": at each threshold of risk, defined at a given percentile on the *x* axis, healthier Whites above the threshold are replaced with less healthy Blacks below the threshold, until the marginal patient is equally healthy). The × symbols show risk percentiles by race; circles show risk deciles with 95% confidence intervals clustered by patient. The dashed vertical lines show the auto-identification threshold (the black line, which denotes the 97th percentile) and the screening threshold (the gray line, which denotes the 55th percentile).

of substantial disparities in program screening. We quantify this by simulating a counterfactual world with no gap in health conditional on risk. Specifically, at some risk threshold α, we identify the supramarginal White patient (i) with $R_i > \alpha$ and compare this patient's health to that of the inframarginal Black patient (j) with $R_j < \alpha$. If $H_i > H_j$, as measured by number of chronic medical conditions, we replace the (healthier, but supramarginal) White patient with the (sicker, but inframarginal) Black patient. We repeat this procedure until $H_i = H_j$, to simulate an algorithm with no predictive gap between Blacks and Whites. Fig. 1B shows the results: At all risk thresholds α above the 50th percentile, this procedure would increase the fraction of Black patients. For example, at α = 97th percentile, among those auto-identified for the program, the fraction of Black patients would rise from 17.7 to 46.5%.

We then turn to a more multidimensional picture of the complexity and severity of patients' health status, as measured by biomarkers that index the severity of the most common chronic illnesses in our sample (as shown in Table 1). This allows us to identify patients who might derive a great deal of benefit from care management programs—e.g., patients with severe diabetes who are at risk of catastrophic complications if they do not lower their blood sugar (*18*, *26*). (The materials and methods section describes several experiments to rule out a large effect of the program on these health measures in year *t*; had there been such an effect, we could not easily use the measures to assess the accuracy of the algorithm's predictions on health, because the program is allocated as a function of algorithm score.) Across all of these important markers of health needs—severity of diabetes, high blood pressure, renal failure, cholesterol, and anemia—we find that Blacks are substantially less healthy than Whites at any level of algorithm predictions, as shown in Fig. 2. Blacks have more-severe hypertension, diabetes, renal failure, and anemia, and higher cholesterol. The magnitudes of these differences are large: For example, differences in severity of hypertension (systolic pressure: 5.7 mmHg) and diabetes [glycated hemoglobin (HbA1c): 0.6%] imply differences in all-cause mortality of 7.6% (*27*) and 30% (*28*), respectively, calculated using data from clinical trials and longitudinal studies.

Mechanism of bias

An unusual aspect of our dataset is that we observe the algorithm's inputs and outputs as well as its objective function, providing us a unique window into the mechanisms by which bias arises. In our setting, the algorithm takes in a large set of raw insurance claims data $X_{i,t-1}$ (features) over the year $t - 1$: demographics (e.g., age, sex), insurance type, diagnosis and procedure codes, medications, and detailed costs. Notably, the algorithm specifically excludes race.

The algorithm uses these data to predict $Y_{i,t}$ (i.e., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote "costs" C_t) in year t as the label. Thus, the algorithm's prediction on health needs is, in fact, a prediction on health costs.

As a first check on this potential mechanism of bias, we calculate the distribution of realized costs C versus predicted costs R. By this metric, one could call the algorithm unbiased. Fig. 3A shows that, at every level of algorithm-predicted risk, Blacks and Whites have (roughly) the same costs the following year. In other words, the algorithm's predictions are well calibrated across races. For example, at the median risk score, Black patients had costs of $5147 versus $4995 for Whites (U.S. dollars); in the top 5% of algorithm-predicted risk, costs were $35,541 for Blacks versus $34,059 for Whites.

Because these programs are used to target patients with high costs, these results are largely inconsistent with algorithmic bias, as measured by calibration: Conditional on risk score, predictions do not favor Whites or Blacks anywhere in the risk distribution.

To summarize, we find substantial disparities in health conditional on risk but little disparity in costs. On the one hand, this is surprising: Health care costs and health needs are highly correlated, as sicker patients need and receive more care, on average. On the other hand, there are many opportunities for a wedge to creep in between needing health care and receiving health care—and crucially, we find that wedge to be correlated with race, as shown in Fig. 3B. At a given level of health (again measured by number of chronic illnesses), Blacks generate lower costs than Whites—on average, $1801 less per year, holding constant the number of chronic illnesses (or $1144 less, if we instead hold constant the specific individual illnesses that contribute to the sum). Table S2 also shows that Black patients generate very different kinds of costs: for example, fewer inpatient surgical and outpatient specialist costs, and more costs related to emergency visits and dialysis. These results suggest that the driving force behind the bias we detect is that Black patients generate lesser medical expenses, conditional on health, even when we account for specific comorbidities. As a result, accurate prediction of costs necessarily means being racially biased on health.

How might these disparities in cost arise? The literature broadly suggests two main potential channels. First, poor patients face substantial barriers to accessing health care, even when enrolled in insurance plans. Although the population we study is entirely insured, there are many other mechanisms by which poverty can lead to disparities in use of health care: geography and differential access to transportation, competing demands from jobs or child care, or knowledge of reasons to seek care (*29*–*31*). To the extent that race and socioeconomic status are correlated, these factors will differentially affect Black patients. Second, race could affect costs directly via several channels: direct ("taste-based") discrimination, changes to the doctor–patient relationship, or others. A recent trial randomly assigned Black patients to a Black or White primary care provider and found significantly higher uptake of recommended preventive care when the provider was Black (*32*). This is perhaps the most rigorous demonstration of this effect, and it fits with a larger literature on potential mechanisms by which race can affect health care directly. For example, it has long been documented that Black patients have reduced trust in the health care system (*33*), a fact that some studies trace to the revelations of the Tuskegee study and other adverse experiences (*34*). A substantial literature in psychology has documented physicians' differential perceptions of Black patients, in terms of intelligence, affiliation (*35*), or pain tolerance (*36*). Thus, whether it is communication, trust, or bias, something about the interactions of Black patients with the health care system itself leads to reduced use of health care. The collective effect of these many channels is to lower health spending substantially for Black patients, conditional on need—a finding that has been appreciated for at least two decades (*37*).

Problem formulation

Our findings highlight the importance of the choice of the label on which the algorithm is trained. On the one hand, the algorithm manufacturer's choice to predict future costs is reasonable: The program's goal, at least in part, is

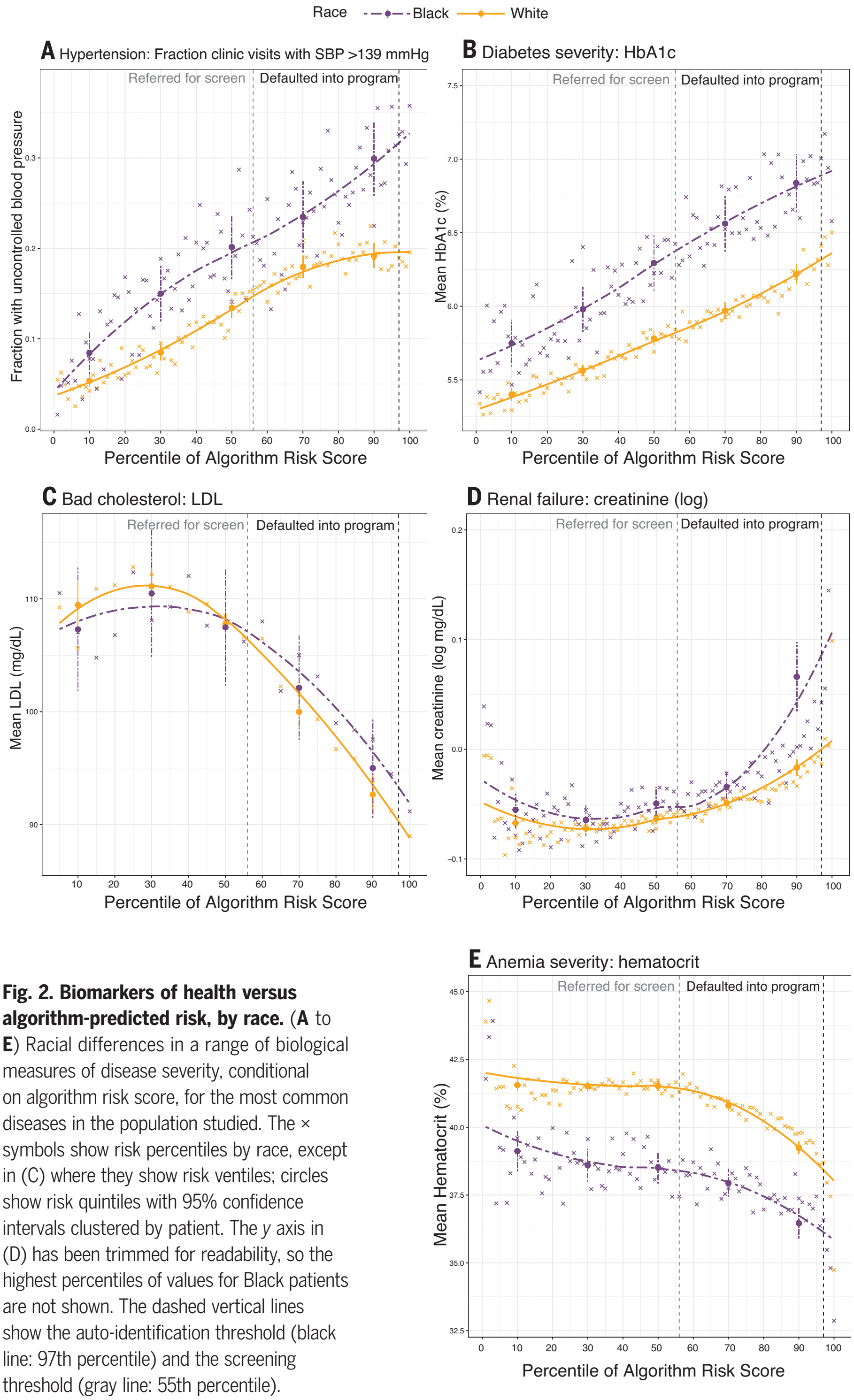

Fig. 2. Biomarkers of health versus algorithm-predicted risk, by race. (**A** to **E**) Racial differences in a range of biological measures of disease severity, conditional on algorithm risk score, for the most common diseases in the population studied. The × symbols show risk percentiles by race, except in (C) where they show risk ventiles; circles show risk quintiles with 95% confidence intervals clustered by patient. The *y* axis in (D) has been trimmed for readability, so the highest percentiles of values for Black patients are not shown. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile).

to reduce costs, and it stands to reason that patients with the greatest future costs could have the greatest benefit from the program. As noted in the supplementary materials, the manufacturer is not alone. Although the details of individual algorithms vary, the cost label reflects the industry-wide approach. For example, the Society of Actuaries's comprehensive evaluation of the 10 most widely used algorithms, including the particular algorithm we study, used cost prediction as its accuracy metric (*21*). As noted in the report, the enthusiasm for cost prediction is not restricted to industry: Similar algorithms are developed and used by non-profit hospitals, academic groups, and governmental agencies, and are often described in academic literature on targeting population health interventions (*18*, *19*).

On the other hand, future cost is by no means the only reasonable choice. For example, the evidence on care management programs shows that they do not operate to reduce costs globally. Rather, these programs primarily work to prevent acute health decompensations that lead to catastrophic health care utilization (indeed, they actually work to increase other categories of costs, such as primary care and home health assistance; see table S2). Thus avoidable future costs, i.e., those related to emergency visits and hospitalizations, could be a useful label to predict. Alternatively, rather than predicting costs at all, we could simply predict a measure of health; e.g., the number of active chronic health conditions. Because the program ultimately operates to improve the management of these conditions, patients with the most encounters related to them could also be a promising group on which to deploy preventative interventions.

The dilemma of which label to choose relates to a growing literature on "problem formulation" in data science: the task of turning an often amorphous concept we wish to predict into a concrete variable that can be predicted in a given dataset (*38*). Problems in health seem particularly challenging: Health is, by nature, holistic and multidimensional, and there is no single, precise way to measure it. Health care costs, though well measured and readily available in insurance claims data, are also the result of a complex aggregation process with a number of distortions due to structural inequality, incentives, and inefficiency. So although the choice of label is perhaps the single most important decision made in the development of a prediction algorithm, in our setting and in many others, there is often a confusingly large array of different options, each with its own profile of costs and benefits.

Experiments on label choice

Through a series of experiments with our dataset, we can gain some insight into how label choice affects both predictive performance and racial bias. We develop three new predictive algorithms, all trained in the same way, to predict the following outcomes: total cost in year *t* (this tailors cost predictions to our own dataset rather than the national training set), avoidable cost in year *t* (due to emergency visits and hospitalizations), and health in year *t* (measured by the number of chronic conditions that flare up in that year). We train all models in a random ⅔ training set and show all results only from the ⅓ holdout set. Furthermore, as with the original algorithm, we exclude race from the feature set (more details are in the materials and methods).

Table 2 shows the results of these experiments. The first finding is that all algorithms perform reasonably well for predicting not only the outcome on which they were trained but also the other outcomes: The concentration of realized outcomes in those at or above the 97th percentile is notably similar for all algorithms across all outcomes. The largest difference in performance across algorithms is seen for cost prediction: Of all costs in the holdout set, the fraction generated by those at or above the 97th percentile is 16.5% for the cost predictor versus 12.1% for the predictor

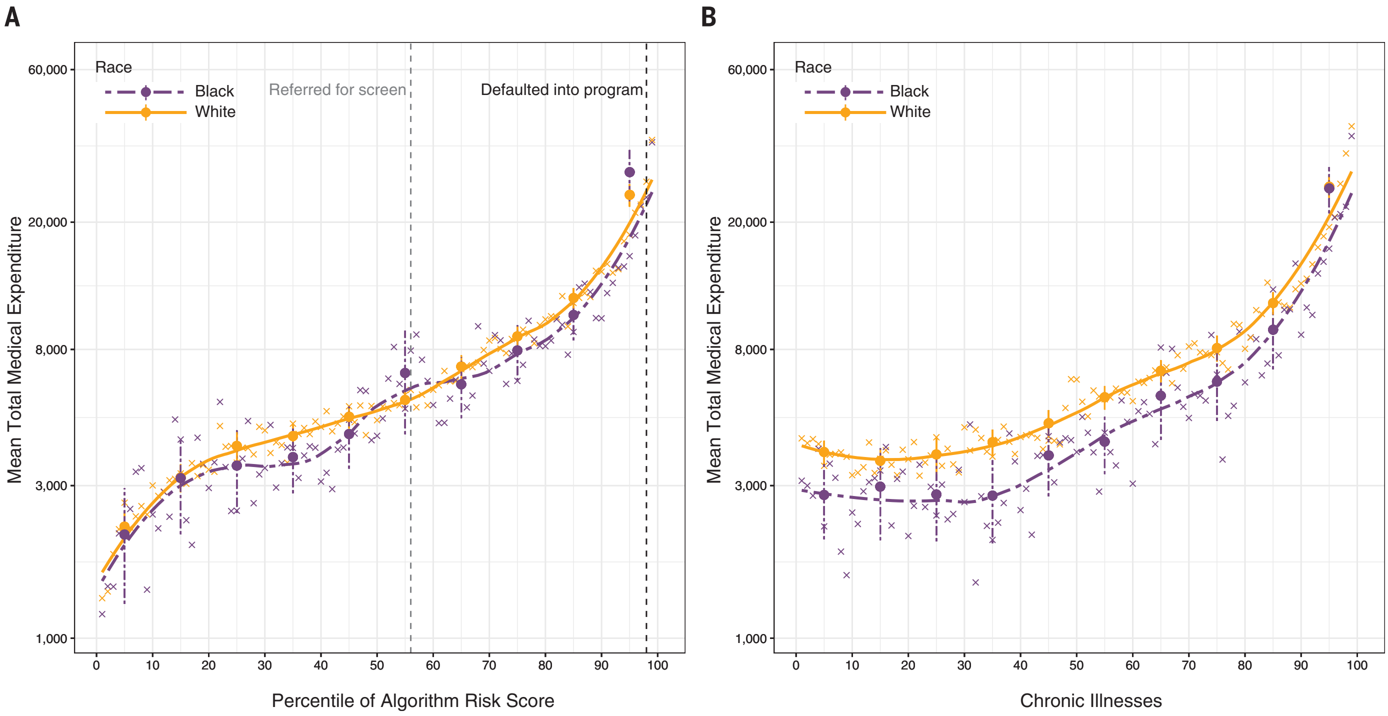

Fig. 3. Costs versus algorithm-predicted risk, and costs versus health, by race. (**A**) Total medical expenditures by race, conditional on algorithm risk score. The dashed vertical lines show the auto-identification threshold (black line: 97th percentile) and the screening threshold (gray line: 55th percentile). (**B**) Total medical expenditures by race, conditional on number of chronic conditions. The × symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The *y* axis uses a log scale.

of chronic conditions. We then test for label choice bias, defined analogously to calibration bias above: For two algorithms trained to predict Y and Y', and using a threshold τ indexing a (similarly sized) high-risk group, we would test $p[B|R > \tau] = p[B|R' > \tau]$ (here, p denotes probability and B represents Black patients).

We find that the racial composition of this highest-risk group varies far more across algorithms: The fraction of Black patients at or above these risk levels ranges from 14.1% for the cost predictor to 26.7% for the predictor of chronic conditions. Thus, although there could be many reasonable choices of label—all predictions are highly correlated, and any could be justified as a measure of patients' likely benefit from the program—they have markedly different implications in terms of bias, with nearly twofold variation in composition of Black patients in the highest-risk groups.

Relation to human judgment

As noted above, the algorithm is not used for program enrollment decisions in isolation. Rather, it is used as a screening tool, in part to alert primary care doctors to high-risk patients. Specifically, for patients at or above a certain level of predicted risk (the 55th percentile), doctors are presented with contextual information from patients' electronic health records and insurance claims and are prompted to consider enrolling them in the program. Thus, realized enrollment decisions largely reflect how doctors respond to algorithmic predictions, along with other administrative factors related to eligibility (for instance, primary care practice site, residence outside of a nursing home, and continual enrollment in an insurance plan).

Table 3 shows statistics on those enrolled in the program, accounting for 1.3% of observations in our sample: The enrolled individuals are 19.2% Black (versus 11.9% Black in our entire sample) and account for 2.9% of all costs and 3.3% of all active chronic conditions in the population as a whole. We then perform four counterfactual simulations to put these numbers in context; naturally, these simulations use only observable factors, not the many unobserved administrative and human factors that also affect enrollment. First, we calculate the realized program enrollment rate within each percentile of the original algorithm's predicted risk bins and randomly sample patients in each bin for enrollment. This simulation, which mimics "race-blind" enrollment conditional on algorithm score, would yield an enrolled population that is 18.3% Black (versus 19.2% observed; P = 0.8348). Second, rather than randomly sampling, we sample those with the highest predicted number of active chronic conditions within a risk bin (using our experimental algorithm described above); this would yield a population that is 26.9% Black. Finally, we compare this to simply assigning those with the highest predicted costs, or the highest number of active chronic conditions, to the program (also using our own algorithms detailed above), which would yield 17.2 and 29.2% Black patients, respectively. Thus, although doctors do redress a small part of the algorithm's bias, they do so far less than an algorithm trained on a different label.

Discussion

Bias attributable to label choice—the difference between some unobserved optimal prediction and the prediction of an algorithm trained on an observed label—is a useful framework through which to understand bias in algorithms, both

Table 2. Performance of predictors trained on alternative labels. For each new algorithm, we show the label on which it was trained (rows) and the concentration of a given outcome of interest (columns) at or above the 97th percentile of predicted risk. We also show the fraction of Black patients in each group.

Algorithm training label	Concentration in highest-risk patients (SE)						Fraction of Black patients in group with highest risk (SE)	
	Total costs		Avoidable costs		Active chronic conditions			
Total costs	0.165	(0.003)	0.187	(0.003)	0.105	(0.002)	0.141	(0.003)
Avoidable costs	0.142	(0.003)	0.215	(0.003)	0.130	(0.003)	0.210	(0.003)
Active chronic conditions	0.121	(0.003)	0.182	(0.003)	0.148	(0.003)	0.267	(0.003)
Best-to-worst difference	0.044		0.033		0.043		0.126	

Table 3. Doctors' decisions versus algorithmic predictions. For those enrolled in the high-risk care management program (1.3% of our sample), we first show the fraction of the population that is Black, as well as the fraction of all costs and chronic conditions accounted for by these observations. We also show these quantities for four alternative program enrollment rules, which we simulate in our dataset (using the holdout set when we use our experimental predictors). We first calculate the program enrollment rate within each percentile bin of predicted risk from the original algorithm and either (i) randomly sample patients or (ii) sample those with the highest predicted number of active chronic conditions within a bin and assign them to the program. The resultant values are then compared with values obtained by simply assigning the aforementioned 1.3% of our sample with (iii) the highest predicted cost or (iv) the highest number of active chronic conditions to the program.

Population	Fraction Black (SE)		Fraction of all costs (SE)		Fraction of all active chronic conditions (SE)	
Observed program enrollment (1.3%)	0.192	(0.003)	0.029	(0.001)	0.033	(0.001)
		Simulated alternative enrollment rules				
Random, in predicted-cost bin	0.183	(0.003)	0.044	(0.002)	0.034	(0.001)
Predicted health, in predicted-cost bin	0.269	(0.003)	0.044	(0.002)	0.064	(0.002)
Highest predicted cost	0.172	(0.003)	0.100	(0.002)	0.047	(0.002)
Worst predicted health	0.292	(0.004)	0.067	(0.002)	0.076	(0.002)

in the health sector and further afield. This is because labels are often measured with errors that reflect structural inequalities (*39*). Within the health sector, using mortality or readmission rates to measure hospital performance penalizes those serving poor or non-White populations (*40*, *41*). Outside of the health arena, credit-scoring algorithms predict outcomes related to income, thus incorporating disparities in employment and salary (*2*). Policing algorithms predict measured crime, which also reflects increased scrutiny of some groups (*42*). Hiring algorithms predict employment decisions or supervisory ratings, which are affected by race and gender biases (*43*). Even retail algorithms, which set pricing for goods at the national level, penalize poorer households, which are subjected to increased prices as a result (*44*).

This mechanism of bias is particularly pernicious because it can arise from reasonable choices: Using traditional metrics of overall prediction quality, cost seemed to be an effective proxy for health yet still produced large biases. After completing the analyses described above, we contacted the algorithm manufacturer for an initial discussion of our results. In response, the manufacturer independently replicated our analyses on its national dataset of 3,695,943 commercially insured patients. This effort confirmed our results—by one measure of predictive bias calculated in their dataset, Black patients had 48,772 more active chronic conditions than White patients, conditional on risk score—illustrating how biases can indeed arise inadvertently.

To resolve the issue, we began to experiment with solutions together. As a first step, we suggested using the existing model infrastructure—sample, predictors (excluding race, as before), training process, and so forth—but changing the label: Rather than future cost, we created an index variable that combined health prediction with cost prediction. This approach reduced the number of excess active chronic conditions in Blacks, conditional on risk score, to 7758, an 84% reduction in bias. Building on these results, we are establishing an ongoing (unpaid) collaboration to convert the results of Table 3 into a better, scaled predictor of multidimensional health measures, with the goal of rolling these improvements out in a future round of algorithm development. Of course, our experience may not be typical of all algorithm developers in this sector. But because the manufacturer of the algorithm we study is widely viewed as an industry leader in data and analytics, we are hopeful that this endeavor will prompt other manufacturers to implement similar fixes.

These results suggest that label biases are fixable. Changing the procedures by which we fit algorithms (for instance, by using a new statistical technique for decorrelating predictors with race or other similar solutions) is not required. Rather, we must change the data we feed the algorithm—specifically, the labels we give it. Producing new labels requires deep understanding of the domain, the ability to identify and extract relevant data elements, and the capacity to iterate and experiment. But there is precedent for all of these functions in the literature and, more concretely, in the private companies that invest heavily in developing new and improved labels to predict factors such as consumer behavior (*45*). In addition, although health—as well as criminal justice, employment, and other socially important areas—presents substantial challenges to measurement, the importance of these sectors emphasizes the value of investing in such research. Because labels are the key determinant of both predictive quality and predictive bias, careful choice can allow us to enjoy the benefits of algorithmic predictions while minimizing their risks.

REFERENCES AND NOTES

1. J. Angwin, J. Larson, S. Mattu, L. Kirchner, "Machine Bias," *ProPublica* (23 May 2016); www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.
2. S. Barocas, A. D. Selbst, *Calif. Law Rev.* **104**, 671 (2016).
3. A. Chouldechova, A. Roth, arXiv:1810.08810 [cs.LG] (20 October 2018).
4. A. Datta, M. C. Tschantz, A. Datta, *Proc. Privacy Enhancing Technol.* **2015**, 92–112 (2015).
5. L. Sweeney, *Queue* **11**, 1–19 (2013).
6. M. Kay, C. Matuszek, S. A. Munson, in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (ACM, 2015), pp. 3819–3828.
7. B. F. Klare, M. J. Burge, J. C. Klontz, R. W. Vorder Bruegge, A. K. Jain, *IEEE Trans. Inf. Forensics Security* **7**, 1789–1801 (2012).
8. J. Buolamwini, T. Gebru, in *Proceedings of the Conference on Fairness, Accountability and Transparency* (PMLR, 2018), pp. 77–91.
9. A. Caliskan, J. J. Bryson, A. Narayanan, *Science* **356**, 183–186 (2017).
10. S. Corbett-Davies, S. Goel, arXiv:1808.00023 [cs.CY] (31 July 2018).
11. M. De-Arteaga *et al.*, arXiv:1901.09451 [cs.IR] (27 January 2019).
12. M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, S. Venkatasubramanian, in *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM, 2015), pp. 259–268.
13. J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, S. Mullainathan, *Q. J. Econ.* **133**, 237–293 (2018).
14. C. S. Hong, A. L. Siegel, T. G. Ferris, *Issue Brief (Commonwealth Fund)* **19**, 1–19 (2014).
15. N. McCall, J. Cromwell, C. Urato, "Evaluation of Medicare Care Management for High Cost Beneficiaries (CMHCB) Demonstration: Massachusetts General Hospital and Massachusetts General Physicians Organization (MGH)" (RTI International, 2010).
16. J. Hsu *et al.*, *Health Aff.* **36**, 876–884 (2017).
17. L. Nelson, "Lessons from Medicare's demonstration projects on disease management and care coordination" (Working Paper 2012-01, Congressional Budget Office, 2012).
18. C. Vogeli *et al.*, *J. Gen. Intern. Med.* **22** (suppl. 3), 391–395 (2007).
19. D. W. Bates, S. Saria, L. Ohno-Machado, A. Shah, G. Escobar, *Health Aff.* **33**, 1123–1131 (2014).
20. J. Kleinberg, J. Ludwig, S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **105**, 491–495 (2015).
21. G. Hileman, S. Steele, "Accuracy of claims-based risk scoring models" (Society of Actuaries, 2016).
22. J. Kleinberg, S. Mullainathan, M. Raghavan, arXiv:1609.05807 [cs.LG] (19 September 2016).
23. A. Chouldechova, *Big Data* **5**, 153–163 (2017).
24. V. de Groot, H. Beckerman, G. J. Lankhorst, L. M. Bouter, *J. Clin. Epidemiol.* **56**, 221–229 (2003).
25. J. J. Gagne, R. J. Glynn, J. Avorn, R. Levin, S. Schneeweiss, *J. Clin. Epidemiol.* **64**, 749–759 (2011).
26. A. K. Parekh, M. B. Barton, *JAMA* **303**, 1303–1304 (2010).
27. D. Ettehad *et al.*, *Lancet* **387**, 957–967 (2016).
28. K.-T. Khaw *et al.*, *BMJ* **322**, 15 (2001).
29. K. Fiscella, P. Franks, M. R. Gold, C. M. Clancy, *JAMA* **283**, 2579–2584 (2000).
30. N. E. Adler, K. Newman, *Health Aff.* **21**, 60–76 (2002).
31. N. E. Adler, W. T. Boyce, M. A. Chesney, S. Folkman, S. L. Syme, *JAMA* **269**, 3140–3145 (1993).
32. M. Alsan, O. Garrick, G. C. Graziani, "Does diversity matter for health? Experimental evidence from Oakland" (National Bureau of Economic Research, 2018).
33. K. Armstrong, K. L. Ravenell, S. McMurphy, M. Putt, *Am. J. Public Health* **97**, 1283–1289 (2007).
34. M. Alsan, M. Wanamaker, *Q. J. Econ.* **133**, 407–455 (2018).
35. M. van Ryn, J. Burke, *Soc. Sci. Med.* **50**, 813–828 (2000).
36. K. M. Hoffman, S. Trawalter, J. R. Axt, M. N. Oliver, *Proc. Natl. Acad. Sci. U.S.A.* **113**, 4296–4301 (2016).
37. J. J. Escarce, F. W. Puffer, in *Racial and Ethnic Differences in the Health of Older Americans* (National Academies Press, 1997), chap. 6; www.ncbi.nlm.nih.gov/books/NBK109841/.
38. S. Passi, S. Barocas, arXiv:1901.02547 [cs.CY] (8 January 2019).
39. S. Mullainathan, Z. Obermeyer, *Am. Econ. Rev.* **107**, 476–480 (2017).
40. K. E. Joynt Maddox *et al.*, *Health Serv. Res.* **54**, 327–336 (2019).
41. K. E. Joynt Maddox, M. Reidhead, A. C. Qi, D. R. Nerenz, *JAMA Intern. Med.* **179**, 769–776 (2019).
42. K. Lum, W. Isaac, *Significance* **13**, 14–19 (2016).
43. I. Ajunwa, "The Paradox of Automation as Anti-Bias Intervention," available at SSRN (2016); https://ssrn.com/abstract=2746078.
44. S. DellaVigna, M. Gentzkow, "Uniform pricing in US retail chains" (National Bureau of Economic Research, 2017).
45. C. A. Gomez-Uribe, N. Hunt, *ACM Trans. Manag. Inf. Syst.* **6**, 13 (2016).

ACKNOWLEDGMENTS

We thank S. Lakhtakia, Z. Li, K. Lin, and R. Mahadeshwar for research assistance and D. Buefort and E. Maher for data science expertise. **Funding:** This work was supported by a grant from the National Institute for Health Care Management Foundation. **Author contributions:** Z.O. and S.M. designed the study, obtained funding, and conducted the analyses. All authors contributed to reviewing findings and writing the manuscript. **Competing interests:** The analysis was completely independent: None of the authors had any contact with the algorithm's manufacturer until after it was complete. No authors received compensation, in any form, from the manufacturer or have any commercial interests in the manufacturer or competing entities or products. There were no confidentiality agreements that limited reporting of the work or its results, no material transfer agreements, no oversight in the preparation of this article (besides ethical oversight from the approving IRB, which was based at a non-profit academic health system), and no formal relationship of any kind between any of the authors and the manufacturer. **Data and materials availability:** Because the data used in this analysis are protected health information, they cannot be made publicly available. We provide instead a synthetic dataset (using the R package synthpop) and all code necessary to reproduce our analyses at https://gitlab.com/labsysmed/dissecting-bias.

SUPPLEMENTARY MATERIALS

science.sciencemag.org/content/366/6464/447/suppl/DC1
Materials and Methods
Figs. S1 to S5
Tables S1 to S4
References (*46*–*51*)

8 March 2019; accepted 4 October 2019
10.1126/science.aax2342