

☐

I'm not robot


reCAPTCHA

Continue

Brigham financial management theory and practice pdf

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 illness. 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 work-arounds 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 address their elevated risks. This algorithm also informs decisions about which patients to enroll in programs designed to reduce the financial burden of care on the insurer. The program also identifies patients who are at greatest need for social services, 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. 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 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). Table 1 Descriptive statistics on our sample, by race, BP, blood pressure; LDL, low-density lipoprotein. View this table: 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 (22, 23)], others on statistical parity of some decision D computed by the algorithm (40), and still others on balance of average predictions, conditional on the realized outcome (22). Given this multiplicity and the recognition that not all conditions can be met simultaneously satisfied (3, 10, 22), we focus on metrics most relevant to the real-world use of this algorithm. Specifically, we ask whether the algorithm is able to predict accurately, for its intended purpose, while being fair to patients. Fairness here means that the algorithm does not discriminate against groups of patients based on characteristics unrelated to the goal of the prediction. To measure fairness, we compare the algorithmic risk scores for patients by race (R_{race}), formed on the basis of claims data $X_i(t-1)$ from the prior year $t-1$, to the patients’ realized health H_i . We call this comparison the algorithmic health index (AHI). We also ask whether the algorithm is calibrated for its intended purpose. To focus on this question, we compare the distribution of outcomes in health care 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. 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; $P < 0.001$). Fig. 1B Number of chronic illnesses versus algorithm-predicted risk, by race (A) Mean number of chronic conditions by race, plotted against algorithmic 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 x 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). 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 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 (j) with $R_j \geq \alpha$ and compare this patient’s health to that of the inframarginal Black patient (i) with $R_i < \alpha$. 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 $\alpha = 97$ th 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 algorithmic 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 algorithmic 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. Fig. 2 Biomarkers of health versus algorithm-predicted risk, by race (A) E Race differences in a range of biological measures of disease severity, conditional on algorithmic risk score, for the most common chronic illnesses in our sample. (B) Biomarker distributions by race, conditional on algorithmic risk score. The y-axis shows the standardized residuals of the biomarker values, conditional on algorithmic risk score. The x-axis shows the algorithmic risk score, with the screening threshold (gray line: 55th percentile) and the auto-identification threshold (black line: 97th percentile). The shaded regions represent 95% confidence intervals clustered by patient. 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 $Y_i(t)$ (e.g., the label). In this instance, the algorithm takes total medical expenditures (for simplicity, we denote “costs” C) 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 C. 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, Blacks 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. Fig. 3 Costs versus algorithm-predicted risk, and costs versus health, by race (A) Total medical expenditures by race, conditional on algorithmic 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 x symbols show risk percentiles; circles show risk deciles with 95% confidence intervals clustered by patient. The y axis uses a log scale. 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, \$180.1 less per year, holding constant the number of chronic illnesses (or \$114.4 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

[illegible]