ANNUAL REVIEWS

Annual Review of Public Health

Racial Bias in Clinical and Population Health Algorithms: A Critical Review of Current Debates

Madison Coots,[1] Kristin A. Linn,[2] Sharad Goel,[1] Amol S. Navathe,[2] and Ravi B. Parikh[3]

[1]Harvard Kennedy School, Harvard University, Cambridge, Massachusetts, USA

[2]The Parity Center, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA

[3]School of Medicine, Emory University, Atlanta, Georgia, USA; email: ravi.bharat.parikh@emory.edu

Annu. Rev. Public Health 2025. 46:507–23

First published as a Review in Advance on December 3, 2024

The *Annual Review of Public Health* is online at publhealth.annualreviews.org

https://doi.org/10.1146/annurev-publhealth-071823-112058

Keywords

race, bias, algorithms, fairness, equity, health care

Abstract

Among health care researchers, there is increasing debate over how best to assess and ensure the fairness of algorithms used for clinical decision support and population health, particularly concerning potential racial bias. Here we first distill concerns over the fairness of health care algorithms into four broad categories: (*a*) the explicit inclusion (or, conversely, the exclusion) of race and ethnicity in algorithms, (*b*) unequal algorithm decision rates across groups, (*c*) unequal error rates across groups, and (*d*) potential bias in the target variable used in prediction. With this taxonomy, we critically examine seven prominent and controversial health care algorithms. We show that popular approaches that aim to improve the fairness of health care algorithms can in fact worsen outcomes for individuals across all racial and ethnic groups. We conclude by offering an alternative, consequentialist framework for algorithm design that mitigates these harms by instead foregrounding outcomes and clarifying trade-offs in the pursuit of equitable decision-making.

1. INTRODUCTION

Over the past two decades, organizations across sectors have developed and deployed algorithms to enhance decision-making. In health care, algorithms are increasingly used to guide high-stakes decisions, including disease screening and treatment, as well as to allocate limited health care resources. Algorithmic decision-making promises to increase efficiency and reduce subjectivity, but researchers and clinicians have raised concerns that health care algorithms exacerbate inequities, particularly relating to race and ethnicity.

In this article, we survey a wide range of prominent health care algorithms used for population health and individual decision support, all of which have been the subject of extensive debates over their fairness. While the term fairness itself is contested, we use it here to describe an algorithm's tendency to enhance rather than diminish the equity of decisions. The clinical algorithms we consider span several medical fields, including oncology, obstetrics, cardiology, and nephrology, and concern both shared decision-making and the allocation of limited resources. Using these algorithms as case studies, we distill into four categories the myriad fairness concerns that have been raised for health care algorithms: (*a*) the inclusion (or, conversely, the exclusion) of race and ethnicity in algorithm inputs, (*b*) unequal algorithm decision rates across groups, (*c*) unequal error rates across groups, and (*d*) potential bias in the target variable used in prediction. In **Table 1**, we situate into this structure each of the algorithms we consider.

We briefly summarize these four broad concerns before discussing them in depth in subsequent sections. The use of race and ethnicity in health algorithms is heavily debated (18–20, 56, 61–63, 67). Advocates for race-aware algorithms argue that explicitly considering race improves the accuracy of decisions for all groups (1, 38, 39, 42, 67). Conversely, proponents of race-unaware algorithms—commonly called "race-blind" algorithms—argue that using race perpetuates pernicious racial attitudes and exacerbates racial and ethnic inequities (23, 33, 56, 61–63). In addition to scrutinizing an algorithm's inputs, researchers have sought to assess the fairness of algorithms by examining differences in decision and error rates across racial and ethnic groups. For example, in

Table 1 Taxonomy of health care algorithms and their fairness concerns

Algorithm		Use of race	Unequal decision rates	Unequal error rates	Label bias
Not resource constrained	Lung cancer incidence risk model		✓	✓	
	Lung cancer LYFS model	✓	✓	✓	✓
	VBAC success calculator	✓	✓		
	CVD incidence risk model	✓	✓		✓
Resource constrained	CVD hospital mortality risk model		✓		
	Kidney function (eGFR) equation	✓	✓		
	Health care need prediction models				✓

Abbreviations: CVD, cardiovascular disease; eGFR, estimated glomerular filtration rate; LYFS, life-years from screening; VBAC, vaginal birth after cesarean.
We situate each of the seven algorithms considered in this review within our taxonomy of the four main fairness concerns for clinical and population health algorithms.

the context of lung cancer, researchers have noted that common algorithms recommend different screening rates across racial groups and also exhibit racial gaps in missed referrals for screening (i.e., unequal error rates) (2, 50). As a result, many proponents advocate for designing algorithms to equalize decision and error rates across groups (37). That strategy seeks to ensure that the benefits and burdens of algorithmic decision-making are shared equally across groups, although it also tacitly ignores differences in need and individual preferences. Finally, there is concern among researchers over the mismeasurement of target variables used for prediction in health care algorithms, a problem also known as "label bias" (66). For example, algorithms trained to predict health care expenditure as a proxy of health care need may underallocate health care resources to disadvantaged groups due to racial disparities in health care access and expenditure (49).

Many of the fairness concerns raised for health care algorithms focus on how an algorithm makes decisions: For example, does it use race, or does it equalize decisions across groups? That focus contrasts with a perspective that foregrounds outcomes, an approach to ethical decision-making often called consequentialism. From a consequentialist perspective, what renders an algorithm fair is its impact on individuals and society, not its set of inputs or some particular statistical summary. Consequentialism suggests an algorithm be designed to maximize aggregate utility, the overall desirability or benefit that comes from a specific decision or policy. In the health care context, utility might encapsulate quantitative measures, such as life-years gained, or improvement in quality of life, as well as the monetary and nonmonetary costs associated with unnecessary testing.

In this article, we argue that the dominant approach to designing fair health care algorithms—one that, for example, seeks to equalize decision or error rates—can often harm the groups it seeks to protect. Instead, we advocate for adopting a consequentialist perspective to algorithm design. The remainder of our article is structured as follows. Beginning with an extended case study of risk algorithms for lung cancer, we unpack the four broad categories of fairness concerns for health care algorithms outlined above. We then expand our discussion by reviewing several more prominent health care algorithms that have been the subject of recent fairness debates. Finally, we conclude with an example that illustrates the value of a consequentialist framework, offering a path forward for designing equitable algorithms in health care and beyond.

2. FAIRNESS CONCERNS IN LUNG CANCER SCREENING

In the United States, lung cancer is the third most commonly diagnosed cancer and is the leading cause of cancer-related death (8). Black men in the United States have higher rates of lung cancer incidence and mortality than do men in any other racial or ethnic group (45, 46). Researchers have attributed these disparities to differences in social determinants of health—such as access to health care and exposure to carcinogens—that are correlated with race and ethnicity, with race and ethnicity acting as surrogates for these factors in clinical models for lung cancer (58).

Screening in the form of low-dose computed tomography (CT) scans remains the most effective method for diagnosing and informing treatment for lung cancer. However, screening comes with both monetary and nonmonetary costs, such as the direct cost of the scan, taking time away from work, and the psychological stress associated with screening and false positives. Given these trade-offs, the United States Preventive Services Task Force (USPSTF) recommends annual screening only for high-risk individuals: adults aged 50–80 years who have a 20 pack-year smoking history and currently smoke or have quit within the past 15 years (35). Researchers have also developed risk models to produce more personalized estimates of risk (10, 31, 60). To identify high-risk individuals, clinicians use thresholds on these risk scores to inform recommendations for follow-up screening.

The algorithms used to produce these risk scores are the subject of extensive debate, particularly with respect to their inclusion of race (37) and how the use of race interacts with the choice of target variable, tying into discussions on label bias (58). Furthermore, the fairness of these risk models has been characterized in terms of differences in decision and error rates across demographic groups (2, 50). Using lung cancer as an extended case study, we discuss in more depth the four categories of fairness concerns listed above and offer empirical evidence on the consequences of employing different approaches to fairness popular in the literature.

2.1. The Use of Race

To develop the most statistically accurate models, some researchers have recommended using race-aware algorithms that account for racial and ethnic disparities in lung cancer incidence and mortality (10, 31, 37, 60). Medical societies, such as the American Thoracic Society, have also recommended augmenting the USPSTF guidelines with race-aware predictive models to guide decisions, in part to identify more high-risk racial minorities for screening (53). However, adjustments for race in clinical algorithms are widely debated across numerous disease contexts, as well as in medicine more generally (23, 30, 56, 61–63, 67). Many researchers have advocated for eliminating race and ethnicity in similar risk estimation tools due to concerns that including race and ethnicity reifies race as biologically meaningful and may in turn result in inappropriately racialized medical treatment (56, 61–63).

2.2. Unequal Decision Rates Across Groups

Researchers have sought to evaluate the fairness of lung cancer risk models by comparing screening recommendation rates across groups. For example, Landy et al. (37) propose a (necessarily) race-aware model that ensures that screening rates for racial and ethnic minorities are equal to or greater than that of White individuals. While perhaps intuitively appealing, equalizing such decision rates can harm members of all groups. To see this potential harm, we note that in shared decision-making contexts such as with lung cancer screening, the decision threshold—in this case, the screening threshold—is set at the point of indifference, where the costs of the decision are expected to equal the benefits (51). As a result, individuals with true risk above the decision threshold are expected to have positive utility from being screened. Conversely, individuals with true risk below the threshold are expected to have negative utility from being screened (e.g., from incurring the costs of screening while expecting little benefit). However, equalizing decision rates does not consider an individual's utility for being screened and may thus impose utility losses from over- or underscreening individuals.

To see how equalizing decision rates can produce these undesired consequences, consider **Figure 1*a***, which shows distributions of lung cancer risk as estimated by the Lung Cancer Risk Assessment Tool (LCRAT), disaggregated by race and ethnicity, using data from the National Lung Screening Trial (NLST) (47). [**Figure 1** mirrors analyses performed by Chohlas-Wood et al. (11).] In this population, 62% of White individuals have a risk score above the recommended screening threshold of 2% (31, 32); 35% of Asian individuals, 36% of Hispanic individuals, and 74% of Black patients are also above the threshold. As evident in the plot, these differences in decision rates arise primarily due to differences in the group-level distributions of risk.

Now consider a policy that sets group-specific thresholds on risk so that Asian, Hispanic, and Black individuals are recommended for screening at the same rate as White individuals (62%). Under this approach, Asian and Hispanic individuals with risk above 1.2% would be screened, which includes many relatively low-risk individuals for whom screening would be expected to yield negative utility. Examples of sources of negative utility from screening include pain and

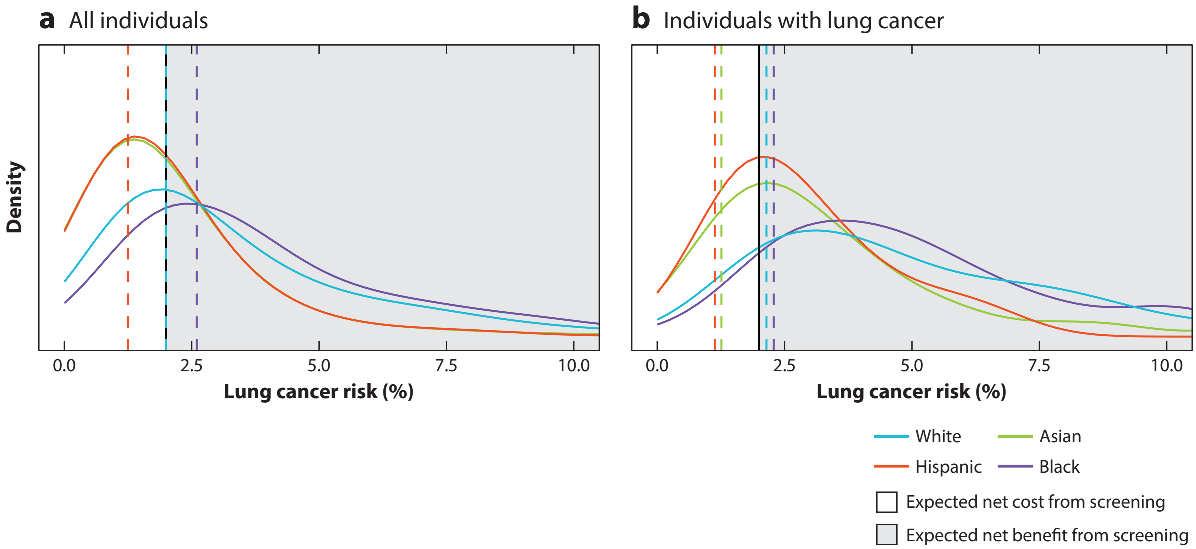

Figure 1

The distribution of lung cancer risk for all individuals and individuals with lung cancer. Estimates of risk were generated using the lung cancer risk assessment tool (LCRAT). The dashed vertical lines correspond to screening thresholds that equalize (*a*) decision rates across groups and (*b*) false negative rates across groups. The vertical black line indicates the recommended screening threshold of 2%.

complications from invasive biopsies or surgeries. Yet only Black individuals who have relatively high risk of lung cancer—above ~2.6%—would be screened, leaving out many Black individuals for whom screening would be expected to have net benefits, also resulting in lost utility. Thus, equalizing screening rates across groups would in fact harm Asian, Hispanic, and Black individuals by failing to screen some individuals who are expected to benefit from screening and screening others for whom the costs are expected to outweigh the benefits.

2.3. Unequal Error Rates Across Groups

Other work has sought to assess the fairness of the USPSTF screening criteria in terms of group-level sensitivity[1] (2, 50). This analysis amounts to a comparison of false negative rates (FNR), given that FNR is 1− sensitivity. These studies (2, 50) have shown that the USPSTF criteria fail to recommend Black individuals with lung cancer for screening at higher rates than White individuals. Consequently, some researchers recommend lowering the USPSTF criteria on smoking history to increase screening among Black individuals with lung cancer (2, 50). This argument, however, similarly fails to account for differences in risk distributions across groups.

To see how equalizing error rates—much like equalizing decision rates—can produce undesired consequences, consider **Figure 1*b***. We depict distributions of estimated lung cancer risk (as estimated by the LCRAT) among people who have lung cancer. Under a policy of screening patients above the recommended 2% threshold, Hispanic individuals have the highest false negative rate at 44%, Asian individuals have an FNR of 28%, White individuals have an FNR of 13%, and Black individuals have the lowest FNR at 10%. Equalizing false negative rates across groups (while leaving risk estimates unchanged) requires the use of group-specific screening thresholds.

[1]The sensitivity equals $TP/(TP + FN)$, where TP is the number of true positives, and FN is the number of false negatives.

For example, screening Hispanic individuals above a 1.1% threshold, screening Asian individuals above a 1.3% threshold, screening White individuals above a 2.1% threshold, and screening Black individuals above a 2.3% threshold would result in all groups having a false negative rate of ∼15%. However, such a policy would recommend screening for some relatively low-risk Hispanic and Asian individuals, resulting in expected utility losses for some members of these groups. In general, equalizing error rates across groups may harm members of all racial and ethnic groups.

2.4. The Risk of Label Bias

Not all lung cancer risk models are designed to predict the same outcome. The LCRAT and PLCOm2012 models predict lung cancer incidence, whereas the life-years gained from screening–CT (LYFS-CT) model predicts expected benefit from screening (10, 31, 60). However, researchers have expressed concern that models—particularly race-aware models—that predict life-years gained from screening are susceptible to label bias due to differences in life expectancy across racial and ethnic groups (58). Black individuals have lower average life expectancy than do White individuals, with researchers attributing these gaps to differences in social determinants of health, such as income and geography (55). A model that predicts life-years gained from screening may therefore risk predicting patterns in social determinants of health as opposed to absolute benefits of screening.

As we have seen in the case of lung cancer, fairness concerns in algorithmic decision-making are nuanced, and divorcing measures of fairness from outcomes may produce unintended and undesired consequences. These issues are not isolated to lung cancer screening but are pervasive across algorithms used in health care, which we explore further in the following sections.

3. CASE STUDIES ON THE FAIRNESS OF HEALTH CARE ALGORITHMS

We next examine several prominent—and, in some cases, controversial—health care algorithms used for risk estimation and resource allocation. For each algorithm, we briefly review its genesis and intended applications and discuss the relevant fairness concerns they invoke.

3.1. Risk Estimation

We first consider three risk scores used in obstetrics and cardiology. As shown in **Table 1**, two of these risk scores are used in a shared decision-making context, and one is used in a resource-constrained setting.

3.1.1. Vaginal birth after cesarean. Among pregnant women who previously have had cesarean sections (C-sections), there are trade-offs to a trial of labor (TOLAC, or trial of labor after cesarean). Vaginal births have well-established advantages over repeat C-sections, including shorter recovery times, lower risks of infection and hemorrhage, and better outcomes in future pregnancies (15). However, for some women, these benefits may not outweigh the risks. Vaginal birth after cesarean (VBAC) success calculators help health care providers and pregnant women weigh this decision. The first VBAC calculator—called the Grobman calculator—was developed to predict the likelihood of a successful VBAC by considering factors such as age, body mass index (BMI), any prior vaginal delivery, previous VBAC, reason for prior cesarean, and race and ethnicity (26, 27, 36).

VBAC calculators have been criticized for their use of race and the resulting disparities in TOLAC recommendation rates across racial and ethnic groups (61, 63). The Grobman calculator produces systematically higher VBAC success probabilities for White women than for Black or Hispanic women, consistent with studies showing that Black and Hispanic women are less likely to achieve successful VBAC than are White women (43). These racial disparities in estimated VBAC success rates likely exist because, conditional on other covariates, race and ethnicity capture a mix of unobserved clinical and nonclinical factors, such as income, education, or health care access.

Researchers have hypothesized that using race-aware VBAC calculators, such as the Grobman calculator, led to differences in how women of different racial and ethnic groups were counseled by their doctors to attempt TOLAC (61, 63). Fairness critiques of VBAC calculators have centered predominantly on unequal recommendation rates to attempt TOLAC across groups. Akin to the lung cancer example above, equalizing VBAC recommendation rates may similarly require using race-specific thresholds on VBAC success probability due to possible differences in risk distributions across groups. This approach, however, would likely result in doctors not recommending TOLAC for some women above the proposed threshold and recommending TOLAC for some women below that threshold—a policy that may impose utility losses for women across all groups. These utility losses would stem from both women who would have benefited from attempting TOLAC but were not counseled to do so as well as from women who were inappropriately recommended TOLAC and suffered avoidable complications during delivery.

Grobman et al. (28, 29) have since developed a race-unaware VBAC calculator, which exhibits comparable overall accuracy to the previous race-aware version, though the authors did not report accuracy or recommendation rates across racial and ethnic groups (6, 28, 29). As researchers continue to evaluate the fairness of VBAC calculators and other decision aids, we believe they should move away from scrutinizing differential recommendation rates across groups, an approach that fundamentally ignores potential differences in underlying risk distributions and which may result in utility losses for members of all groups.

3.1.2. Cardiovascular disease. Clinical algorithms are commonly used to estimate cardiovascular risks, but they have been criticized for explicitly considering race, exhibiting unequal decision rates across groups, and being impacted by label bias. Cardiovascular disease encompasses a broad range of conditions and events such as coronary heart disease (CHD), coronary heart failure (CHF), heart attack, and stroke. To reduce cardiovascular disease (CVD)-related morbidity and mortality, clinicians have long prescribed statins, a class of cholesterol-lowering drugs, to prevent CVD events. Risk calculators for CVD were developed to help physicians determine when to recommend statins as a prophylactic.

The Framingham Risk Score (FRS), a race-unaware model trained to predict incidence of CHD, was among the first such risk estimation algorithms (65). Studies found that the FRS performed reasonably well at predicting CHD events in Black individuals, but concerns emerged over the limited scope of the FRS in predicting only CHD and not other CVD events, such as CHF (16), especially because Black individuals exhibited higher rates of CHF than did other racial and ethnic groups (3, 5). Racial and ethnic differences in CHF incidence rates may lead to label bias when CHD events are used as a proxy for general cardiovascular risks. To address this concern, researchers subsequently developed an expanded (but still race-unaware) CVD risk model that additionally predicted the risk of stroke, peripheral artery disease, and heart failure (17).

Persistent racial and ethnic disparities in incidence rates across these events motivated the development of a race-aware CVD model: the pooled cohort equations (PCE), also known as the American College of Cardiology/American Heart Association (ACC/AHA) risk calculator (25). The PCE outperformed other risk scores in predicting initial atherosclerotic CVD events among Black individuals as well as others (24, 25). However, like earlier risk models, the PCE failed to include CHF in its set of considered events, again raising concerns of label bias (7). In response, the AHA released the Predicting Risk of cardiovascular disease EVENTs (PREVENT) equations for predicting the risk of CVD and CVD subtypes, including CHF (34). Notably, the PREVENT equations are also race-unaware, due to a previous decision to exclude race as a predictor (33). The model instead includes a zip code–level social deprivation index, which helps account for CVD risk factors that are likely correlated with race and ethnicity (33).

The authors of PREVENT report that the equations are suitably calibrated for Black individuals, even without including race or ethnicity, although some have questioned the decision to exclude race and ethnicity from the model. Diao et al. (19) characterized the expected changes in statin and antihypertensive therapy eligibility from the switch to the race-unaware PREVENT equations and found that their use would decrease eligibility for statin and antihypertensive therapy for ~17 million US adults and that these changes would affect a greater proportion of Black adults than White adults. However, the PREVENT equations may yield more clinically appropriate decisions, as the PCE overestimate CVD risks for members of all racial and ethnic groups—a pattern that researchers have attributed to changes in the prevalence of risk factors over time (such as smoking) and to advances in care and prevention (33, 44). Furthermore, those whose treatment recommendations change would likely have cardiovascular risks near the decision threshold, meaning they might not have benefited substantially from treatment (14). Without explicitly considering utility (e.g., in terms of quality-adjusted life-years), fully assessing the impact of excluding race and ethnicity from the model is a challenge.

While PREVENT excludes race and ethnicity, other cardiovascular risk models are still race-aware, such as the AHA Get with the Guidelines–Heart Failure Risk Score, drawing debate. That score informs triage decisions on whether to admit patients to intensive and specialty care, which are often limited resources in many hospitals (57). The model was designed to predict the risk of in-hospital mortality using data from a cohort of patients hospitalized with heart failure (52). In the training data, in-hospital mortality was lower for Black patients, counter to expectations (52). One possible explanation for this pattern relates to racial disparities in access to cardiology care. Past work has found that one of the strongest predictors of admission to the cardiology service is whether a patient was previously seen by an outpatient cardiologist at the hospital—and there were significant racial and ethnic disparities in the proportion of patients who had seen a cardiologist within the past year (21, 22). Because of those disparities, Black patients admitted to intensive care might have had better access to health care and lower risk of mortality as a result.

Due to the lower estimated risks for Black patients, researchers have raised concerns that the AHA Get with the Guidelines–Heart Failure Risk Score may misdirect intensive care away from Black patients (61). But, as discussed previously, decision rate–based criticisms of model fairness ignore potential differences in risk across groups. It may be the case that non-Black patients face higher risks in this setting and would be better served by the additional care. We would caution, however, against using this risk score in settings where the patient population differs substantially from the population used to train the score, especially if those differences are correlated with race or ethnicity. For example, in safety net hospitals, the score might underpredict mortality risk for Black individuals because those patients might have higher risks than the relatively healthy Black individuals in the training data.

3.2. Resource Allocation

We now turn to algorithms used for allocating healthcare resources, a context where the use of algorithms has raised concerns and ethical questions about how these resources are distributed, particularly across different racial and ethnic groups. Specifically, we discuss two resource allocation algorithms used in nephrology and care management.

3.2.1. Kidney transplants. Like algorithms for cardiovascular disease, those for kidney disease are both widespread and controversial, having been criticized for using race and for exhibiting unequal decision rates across groups (61). In the United States, resources for kidney transplantation are highly constrained. In 2021, the average wait time for a deceased donor kidney was five years, and more than half of listed transplant candidates were expected to die or be removed from the list before receiving a transplant (64). To make the best use of the constrained supply of donor kidneys, researchers and clinicians have turned to algorithms to estimate a patient's need for a transplant. Patients are typically recommended for transplants based on estimates of their kidney function, as measured by their glomerular filtration rate (GFR). Given challenges with measuring GFR directly, GFR has traditionally been estimated (eGFR) using an algorithm based on factors such as age, sex, race, body size (usually weight or surface area), and serum creatinine (13, 59). In addition to resource-constrained kidney allocation, eGFR equations have also been used to make non-resource-constrained chronic kidney disease diagnoses and recommendations for drug treatments or other therapies (20).

Early eGFR equations—such as the Modification of Diet in Renal Disease (MDRD) study equation and the Chronic Kidney Disease Epidemiology Collaboration (CKD-EPI) equation—used race to estimate GFR (39–41). All else being equal, these equations estimated higher GFR for Black individuals, meaning better kidney function. Race-aware eGFR equations drew concerns that they made Black patients appear healthier than patients were in reality (61). As a result, many hospitals began using ad hoc race-unaware estimates of kidney function by reporting the White/other eGFR prediction for all patients (20). This strategy was found to slightly increase chronic kidney disease diagnoses among Black adults (20).

This critique of race-aware eGFR equations and the corresponding policy response mirror calls to equalize decision rates across groups. However, as discussed above, equalizing decision rates fails to allow for group differences, and it risks harming members of all groups. Indeed, race-unaware eGFR values typically underestimate GFR for Black individuals (30); the race adjustment in the original equations was included precisely to ensure that estimates were calibrated across groups. Consequently, the race-unaware estimates make Black patients appear less healthy than they likely are in reality. That pattern may have led to inappropriate diagnoses, potentially resulting in net harm to Black patients. Researchers have since revised the CKD-EPI equation to replace race with cystatin C as a predictor of eGFR, leading to race-unaware estimates that are approximately calibrated (30).

Race-aware eGFR equations have been similarly criticized for deprioritizing Black patients for kidney transplantation and specialist care (61) because they estimate higher kidney function for Black individuals compared with otherwise similar White individuals. However, as discussed above, the race-aware equations produce largely accurate estimates of GFR for both Black and non-Black patients. Policy makers may well prefer to enforce some degree of parity in kidney allocation rates, even if that means prioritizing a healthier Black patient over a less healthy White patient. But we believe that these difficult trade-offs should be confronted directly (12). Seeking to increase Black individuals' eligibility for kidney transplants by using an eGFR equation that underestimates these patients' kidney function risks other adverse consequences in the form of overtreatment.

3.2.2. Health care costs as a proxy of need. Health care providers in the United States offer specialty care management programs to improve the care of high-risk patients with complex health needs. These programs aim to help individuals better manage their health by offering additional support from teams of dedicated nurses, social workers, and community health workers. However, these programs are expensive, and health care systems consequently use algorithms to identify patients for whom the benefits justify the additional costs (4). A common strategy is to predict patients' future medical expenses and then direct specialty care management to those expected to incur the largest costs (49).

However, past work has demonstrated that algorithms trained to predict health care costs can fail to allocate resources to high-need racial minorities. Obermeyer et al. (49) evaluated the fairness of a commercial algorithm widely used by health care systems to guide patient referrals to specialty care programs. The algorithm was trained to predict future costs—as a proxy for complex health care needs—based on insurance claims (e.g., diagnoses, procedures, medications) made by an individual in the prior year. The researchers found that the algorithm's generated risk scores were well-calibrated across race groups for predicting health care costs. Conditional on risk score, both Black and White individuals had approximately the same costs in the following year. However, the researchers found that the algorithm was poorly calibrated for predicting realized health. Conditional on risk score, Black patients had significantly more illness burden than did White patients. For example, at the 97th percentile of risk—the threshold for allocating resources—Black patients had 26% more chronic illnesses than did White patients. Due to this miscalibration, resources were diverted away from high-need Black patients to healthier White patients.

This example illustrates the problem of label bias. Health care costs are a poor proxy of health care needs, given the disparities in health care access and Medicaid enrollment that are correlated with race. Past work has shown that, conditional on need, health care spending is lower for Black individuals than for White individuals (48). Consequently, accurate prediction of health care costs necessarily leads to racially biased allocation of health care resources. Obermeyer et al. (49) estimated that changing the target of prediction to an index variable that incorporates health alongside cost prediction would lead to more resources being allocated to Black patients. This result suggests that algorithmic label bias, at least in some circumstances, is both fixable and preventable by thoughtfully selecting prediction targets.

4. TOWARD CONSEQUENTIALIST ALGORITHM DESIGN

The algorithm case studies in the previous sections reveal problems with popular approaches to fairness, which often fail to consider the impact of decisions. These issues highlight the need for a design approach that foregrounds the consequences of an algorithm's use—a challenge we take on here. To guide our discussion, we consider risk models for type 2 diabetes. Researchers have proposed using race-based models for estimating diabetes risk to address known racial disparities in diabetes diagnoses (1). Current guidelines advise using a 1.5% threshold on estimated risk for recommending follow-up screening in the form of a blood test (1). Diabetes screening is not a resource-constrained practice, so, in line with our discussion above, we do not evaluate potential differences in decision rates (or error rates) across groups, as they tell us little about the fairness of an algorithm. Rather, we address the other two fairness concerns considered in this article: label bias and the inclusion (or exclusion) of race and ethnicity. Using a consequentialist framework, we show how to arbitrate between race-aware and race-unaware risk models, following Coots et al. (14).

For our analysis, we use data from the National Health and Nutrition Examination Survey (NHANES) (four cycles from 2011 to 2018) (9). We restricted our sample to ~18,000 patients who were not pregnant, were 18–70 years old, and had a BMI between 18.5 kg/m^2 and 50.0 kg/m^2. Using this data, we trained two linear diabetes risk models based on age and BMI; these models differed only in their inclusion of race and ethnicity as a predictive variable. This example is for illustrative purposes only, and we caution against using these models to guide clinical decisions.

4.1. Selecting Appropriate Prediction Targets

When designing or evaluating a model, investigators must scrutinize the prediction target for label bias. Researchers should be careful to consider the ways in which the proposed label may mismeasure the true outcome through systemic mechanisms of inequality, such as inequitable health care access. In our diabetes example, the label used to train the risk models was constructed by combining the results of a blood test administered by the NHANES with the response to a question on whether the respondent had ever been diagnosed with diabetes by a doctor; therefore, the label likely accurately captures disease status, the true outcome of interest. If, however, our label were constructed only using the individual's response to the diagnosis question, our diabetes risk models would more closely predict diabetes diagnoses, as opposed to diabetes incidence. That misalignment between label and outcome could result in underestimating diabetes risk in groups that have less access to health care.

4.2. Foregrounding Utility

One of the major shortcomings of common algorithmic fairness arguments is that they do not consider individual utilities. In particular, many have advocated for using specific risk algorithms on the grounds that they increase decision rates (often for screening or treatment) for racial and ethnic minority groups (37). Yet these arguments often do not consider expected changes in utility from these changed decisions, and they generally overlook the possibility that increasing decision rates can lead to utility losses for members of all groups, as we demonstrated with our lung cancer example. A consequentialist approach to algorithm design instead foregrounds individual utilities in fairness evaluations.

We first examine the accuracy of race-unaware risk models. In **Figure 2*a***, we visualize the estimated risk from a race-unaware model for diabetes against the observed rates of diabetes across racial and ethnic groups. This plot reveals discrepancies between the model predictions and true diabetes incidence for all groups. This miscalibration would consequently lead the model to fail to recommend screening for some high-risk Asian individuals—whose risk is underestimated by the model—and inappropriately recommend screening for some low-risk White individuals—whose risk is overestimated by the model. In both cases, the miscalibration would impose net costs on members of both groups from suboptimal screening recommendations. Past work has suggested using race-aware risk models to correct miscalibration across groups and obtain more accurate predictions (1). However, that work has stopped short of considering not just differences in accuracy, but differences in utility between race-aware and race-unaware models.

To compare the utility of race-aware versus race-unaware models, we follow Coots et al. (14) and first construct a simplified utility function to aid our comparison of risk models. Our utility function assumes a constant cost of screening and a constant benefit of correctly detecting diabetes. [For further detail on the utility function, see Coots et al. (14).] Applying this utility function to the decisions produced under race-aware and race-unaware models, we find that the relative gain in utility from using a race-aware model is smaller than expected in light of the substantial improvements in accuracy offered by the race-aware model over the race-unaware model. Relative to a baseline policy of no screening, we estimate that the race-aware model would improve overall utility by 0.2% over the race-unaware model. **Figure 2*c*** shows that gains are similarly small across race groups. Our simplified utility function is for illustrative purposes only, to estimate the magnitude of the expected benefits, and is not intended to capture all the complex clinical considerations.

The modest utility gains stem from two factors. First, the vast majority of individuals (94%) would receive the same screening recommendation under both the race-aware and race-unaware models. In **Figure 2*b***, we plot the race-aware risk estimate for an individual against their race-unaware risk estimate. The dots in the shaded regions of the plot correspond to individuals for whom both models produce the same recommendation. In **Figure 2*d***, we show the fraction of

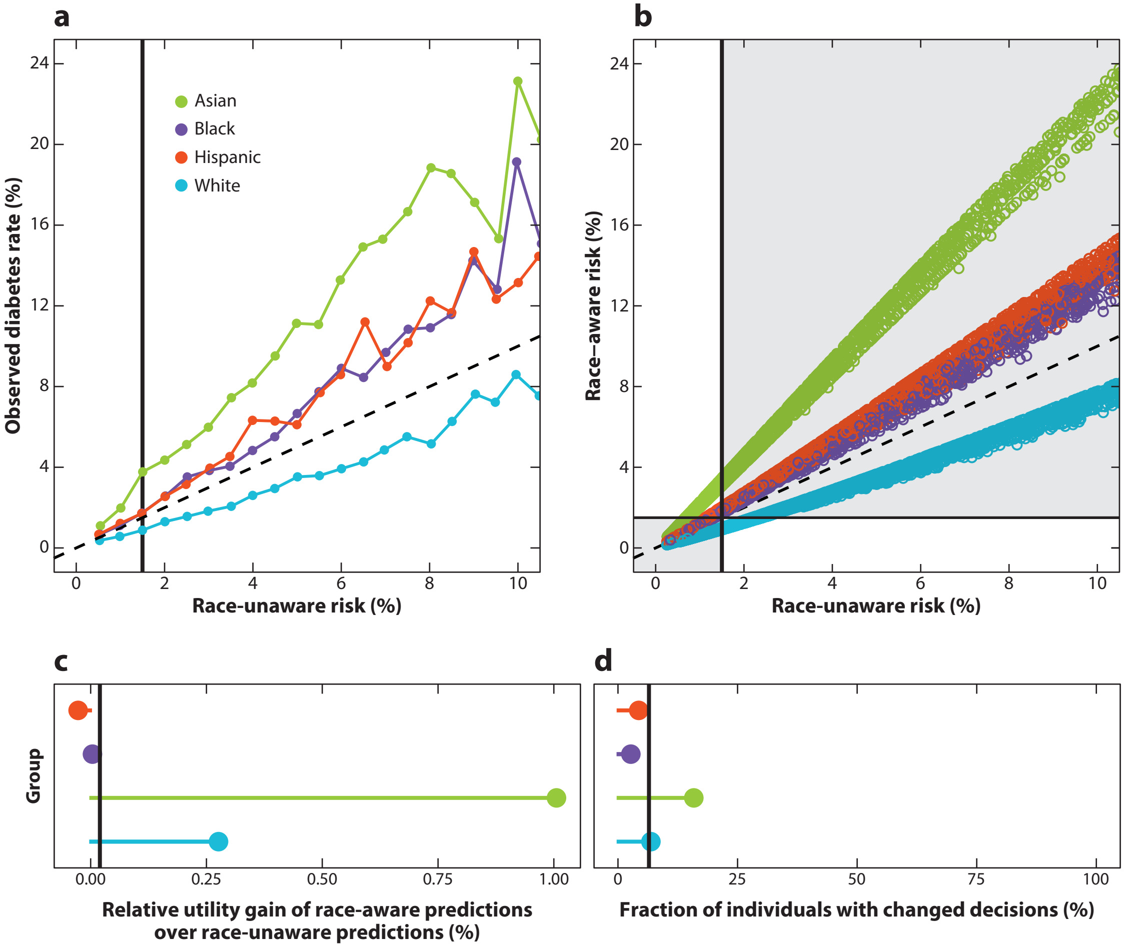

Figure 2

A consequentialist approach to the design of a diabetes risk prediction model. In line with analyses in Coots et al. (14), (*a*) estimated risk from a race-unaware model for diabetes against the observed rates of diabetes across racial and ethnic groups. Panel adapted from Chohlas-Wood et al. (11). (*b*) Scatter plot showing race-unaware risk plotted against race-aware risk for each individual in the data. Individuals in the shaded regions receive the same recommendation under both models. In panels *a* and *b*, the vertical lines correspond to the recommended diabetes risk threshold of 1.5%, above which the typical patient can expect to benefit from screening. The diagonal dashed lines represent hypothetical risk scores that are perfectly calibrated to empirical diabetes rates. (*c*) The relative gain in utility from the use of race-aware predictions to make screening recommendations across racial and ethnic groups. The vertical line denotes the average relative gain experienced across the entire population. (*d*) The fraction of individuals with different recommendations under race-unaware and race-aware models. The vertical line denotes the fraction of individuals with changed decisions across the entire population.

individuals with different decisions under each model by race and ethnicity. The second factor driving this result is that those individuals whose decisions do change are typically close to the decision threshold and therefore accrue relatively small utility gains from using a race-aware model. In short, the small subset of individuals with changed decisions should be largely ambivalent to being screened.

The race-unaware model is starkly miscalibrated but results in smaller-than-expected utility losses relative to the race-aware model. By foregrounding utility, our analysis helps clarify the expected benefits from using a race-aware model. However, the costs of using race—from risk of stigmatization or reinforcing pernicious attitudes on biological determinism, for example—remain an open concern. Ultimately, these costs must be weighed against the estimated benefits in selecting the most appropriate model for decision-making.

5. CONCLUSION

As algorithms are increasingly used to guide health care decisions, discussions around algorithmic fairness have come to the forefront, with racial equity being a particular focus of attention. By critically examining contemporary debates, we have argued for reframing what it means for an algorithm to be fair. Past fairness approaches—grounded in narrow summary statistics such as decision rates and error rates—fail to anticipate the outcomes produced by algorithms, thereby risking unintended harm, including to those in racial and ethnic minority groups. With a consequentialist approach to algorithm design, we advocate for explicitly considering the utility of decisions produced by candidate algorithms to better understand the impact of design choices. This is no easy task. A consequentialist approach requires defining an appropriate utility function, a complex assignment that may also require aligning differing values across stakeholders. This challenge has led some scholars to critique consequentialist approaches to policy (54). But we believe that in many cases of practical importance, it is both feasible and useful to articulate one's values, estimate the impacts of different algorithms, and confront the resulting trade-offs. We hope that our discussion helps researchers, clinicians, and policy makers better understand the common threads underlying ongoing debates and illuminates a path forward for designing more equitable health care algorithms.

DISCLOSURE STATEMENT

Dr. Parikh reports grants from the Prostate Cancer Foundation, National Palliative Care Research Center, NCCN Foundation, Conquer Cancer Foundation, Humana, Emerson Collective, Schmidt Futures, Arnold Ventures, Mendel.ai, and Veterans Health Administration; discloses personal fees and equity from GNS Healthcare, Thyme Care, and Onc.AI, as well as personal fees from the ConcertAI, Cancer Study Group, Mendel.ai, Biofourmis, Archetype Therapeutics, CreditSuisse, G1 Therapeutics, Humana, and Nanology; has received honoraria from Flatiron and Medscape; is an unpaid board member at the Coalition to Transform Advanced Care and the American Cancer Society and editor of the *Journal of Clinical Oncology*; and serves unpaid on a leadership consortium at the National Quality Forum, all outside the submitted work. Dr. Navathe reports grants from the Hawaii Medical Service Association, Commonwealth Fund, Robert Wood Johnson Foundation, Donaghue Foundation, Veterans Affairs Administration, Arnold Ventures, United Healthcare, Blue Cross Blue Shield of North Carolina, and Humana; personal fees from Navvis Healthcare, Elsevier Press, Medicare Payment Advisory Commission, Analysis Group, Advocate Physician Partners, Federal Trade Commission, and Catholic Health Services Long Island; equity from Clarify Health; personal fees from and board membership with the Scan Group; and unpaid board membership with Integrated Services, Inc. outside the submitted work during the past three years. Dr. Linn reports grants from the Commonwealth Fund and Robert Wood Johnson Foundation, as well as personal fees from *JAMA Network Open* outside the submitted work during the past 3 years. This article does not necessarily represent the views of the US government or the Department of Veterans Affairs. The statements contained herein are solely those of the authors and do not represent or imply concurrence with or endorsement by the National Cancer Institute (NCI).

REPRODUCIBILITY

Code to reproduce our analysis is available at **https://github.com/madisoncoots/racial-bias-in-healthcare-algs**.

ACKNOWLEDGMENTS

The authors thank David Kent, Johann Gaebler, Michael Zanger-Tishler, Shira Gur-Arieh, Jacob Jameson, and Julian Nyarko for helpful discussions. The authors additionally thank the NCI for access to data collected by the National Lung Screening Trial (NLST) (CDAS project NLST-1121). This work was supported in part by the Commonwealth Fund and Robert Wood Johnson Foundation (principal investigators, Dr. Linn, Dr. Navathe, Dr. Parikh) and the Veterans Affairs grant SDR 20-361 (principal investigator, Dr. Navathe).

LITERATURE CITED

1. Aggarwal R, Bibbins-Domingo K, Yeh RW, Song Y, Chiu N, et al. 2022. Diabetes screening by race and ethnicity in the United States: equivalent body mass index and age thresholds. *Ann. Intern. Med.* 175(6):765–73
2. Aldrich MC, Mercaldo SF, Sandler KL, Blot WJ, Grogan EL, Blume JD. 2019. Evaluation of USPSTF lung cancer screening guidelines among African American adult smokers. *JAMA Oncol.* 5(9):1318–24
3. Bahrami H, Kronmal R, Bluemke DA, Olson J, Shea S, et al. 2008. Differences in the incidence of congestive heart failure by ethnicity: the Multi-Ethnic Study of Atherosclerosis. *Arch. Intern. Med.* 168(19):2138–45
4. Bates DW, Saria S, Ohno-Machado L, Shah A, Escobar G. 2014. Big data in health care: using analytics to identify and manage high-risk and high-cost patients. *Health Aff.* 33(7):1123–31
5. Bibbins-Domingo K, Pletcher MJ, Lin F, Vittinghoff E, Gardin JM, et al. 2009. Racial differences in incident heart failure among young adults. *N. Engl. J. Med.* 360(12):1179–90
6. Buckley A, Sestito S, Ogundipe T, Roig J, Rosenberg HM, et al. 2022. Racial and ethnic disparities among women undergoing a trial of labor after cesarean delivery: performance of the VBAC calculator with and without patients' race/ethnicity. *Reprod. Sci.* 29(7):2030–38
7. Carnethon MR, Pu J, Howard G, Albert MA, Anderson CAM, et al. 2017. Cardiovascular health in African Americans: a scientific statement from the American Heart Association. *Circulation* 136(21):e393–423
8. CDC (Cent. Dis. Control Prev.). 2024. Lung cancer statistics. *CDC Lung Cancer*. **https://www.cdc.gov/lung-cancer/statistics/**
9. CDC (Cent. Dis. Control Prev.), NCHS (Natl. Cent. Health Stat.). 2024. National Health and Nutrition Examination Survey Data. *National Center for Health Statistics*. **https://wwwn.cdc.gov/nchs/nhanes/**
10. Cheung LC, Berg CD, Castle PE, Katki HA, Chaturvedi AK. 2019. Life-gained-based versus risk-based selection of smokers for lung cancer screening. *Ann. Intern. Med.* 171(9):623–32
11. Chohlas-Wood A, Coots M, Goel S, Nyarko J. 2023. Designing equitable algorithms. *Nat. Comput. Sci.* 3:601–10
12. Chohlas-Wood A, Coots M, Zhu H, Brunskill E, Goel S. 2024. Learning to be fair: a consequentialist approach to equitable decision-making. arXiv:2109.08792v4 [cs.LG]
13. Cockcroft DW, Gault H. 1976. Prediction of creatinine clearance from serum creatinine. *Nephron* 16(1):31–41
14. Coots M, Saghafian S, Kent D, Goel S. 2024. A framework for considering the role of race and ethnicity in estimating disease risk. *Ann. Intern. Med.* In press
15. Curtin SC, Gregory KD, Korst LM, Uddin SF. 2015. Maternal morbidity for vaginal and cesarean deliveries, according to previous cesarean history: new data from the birth certificate, 2013. *Natl. Vital. Stat. Rep.* 64(4):1–13
16. D'Agostino RB, Grundy S, Sullivan LM, Wilson P. 2001. Validation of the Framingham coronary heart disease prediction scores: results of a multiple ethnic groups investigation. *JAMA* 286(2):180–87

17. D'Agostino RB, Vasan RS, Pencina MJ, Wolf PA, Cobain M, et al. 2008. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation* 117(6):743–53
18. Diao JA, He Y, Khazanchi R, Nguemeni Tiako MJ, Witonsky JI, et al. 2024. Implications of race adjustment in lung-function equations. *N. Engl. J. Med.* 390(22):2083–97
19. Diao JA, Shi I, Murthy VL, Buckley TA, Patel CJ, et al. 2024. Projected changes in statin and antihypertensive therapy eligibility with the AHA PREVENT cardiovascular risk equations. *JAMA* 332:989–1000
20. Diao JA, Wu GJ, Taylor HA, Tucker JK, Powe NR, et al. 2021. Clinical implications of removing race from estimates of kidney function. *JAMA* 325(2):184–86
21. Eberly LA, Richterman A, Beckett AG, Wispelwey B, Marsh RH, et al. 2019. Identification of racial inequities in access to specialized inpatient heart failure care at an academic medical center. *Circ. Heart Fail.* 12(11):e006214
22. Eberly LA, Wispelwey B, Richterman A, Beckett AG, Manchanda ECC, et al. 2020. Response by Eberly et al to letter regarding article, "Identification of Racial Inequities in Access to Specialized Inpatient Heart Failure Care at an Academic Medical Center." *Circ. Heart Fail.* 13(6):e007193
23. Eneanya ND, Yang W, Reese PP. 2019. Reconsidering the consequences of using race to estimate kidney function. *JAMA* 322(2):113–14
24. Goff DC Jr., Lloyd-Jones DM. 2016. The pooled cohort risk equations—Black risk matters. *JAMA Cardiol.* 1(1):12–14
25. Goff DC Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, et al. 2014. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J. Am. Coll. Cardiol.* 63(25 Pt. B):2935–59
26. Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, et al. 2007. Development of a nomogram for prediction of vaginal birth after cesarean delivery. *Obstet. Gynecol.* 109(4):806–12
27. Grobman WA, Lai Y, Landon MB, Spong CY, Leveno KJ, et al. 2009. Does information available at admission for delivery improve prediction of vaginal birth after cesarean? *Am. J. Perinatol.* 26:693–701
28. Grobman WA, Sandoval G, Rice MM, Bailit JL, Chauhan SP, et al. 2021. Prediction of vaginal birth after cesarean delivery in term gestations: a calculator without race and ethnicity. *Am. J. Obstet. Gynecol.* 225(6):664.e1–7
29. Grobman WA, Sandoval GJ, Rice MM, Chauhan SP, Clifton RG, et al. 2024. Prediction of vaginal birth after cesarean using information at admission for delivery: a calculator without race or ethnicity. *Am. J. Obstet. Gynecol.* 230(3):S804–6
30. Inker LA, Eneanya ND, Coresh J, Tighiouart H, Wang D, et al. 2021. New creatinine- and cystatin C–based equations to estimate GFR without race. *N. Engl. J. Med.* 385(19):1737–49
31. Katki HA, Kovalchik SA, Berg CD, Cheung LC, Chaturvedi AK. 2016. Development and validation of risk models to select ever-smokers for CT lung cancer screening. *JAMA* 315(21):2300–11
32. Katki HA, Kovalchik SA, Petito LC, Cheung LC, Jacobs E, et al. 2018. Implications of nine risk prediction models for selecting ever-smokers for computed tomography lung cancer screening. *Ann. Intern. Med.* 169(1):10–19
33. Khan SS, Coresh J, Pencina MJ, Ndumele CE, Rangaswami J, et al. 2023. Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: a scientific statement from the American Heart Association. *Circulation* 148(24):1982–2004
34. Khan SS, Matsushita K, Sang Y, Ballew SH, Grams ME, et al. 2024. Development and validation of the American Heart Association's PREVENT equations. *Circulation* 149(6):430–49
35. Krist AH, Davidson KW, Mangione CM, Barry JM, Cabana M, et al.; US Prev. Serv. Task Force. 2021. Screening for lung cancer: US Preventive Services Task Force recommendation statement. *JAMA* 325(10):962–70
36. Landon MB, Leindecker S, Spong CY, Hauth JC, Bloom S, et al. 2005. The MFMU cesarean registry: factors affecting the success of trial of labor after previous cesarean delivery. *Am. J. Obstet. Gynecol.* 193(3):1016–23
37. Landy R, Gomez I, Caverly TJ, Kawamoto K, Rivera MP, et al. 2023. Methods for using race and ethnicity in prediction models for lung cancer screening eligibility. *JAMA Netw. Open.* 6(9):e2331155

38. Lett E, Asabor E, Beltrán S, Cannon AM, Arah OA. 2022. Conceptualizing, contextualizing, and operationalizing race in quantitative health sciences research. *Ann. Fam. Med.* 20(2):157–63
39. Levey AS, Bosch JP, Lewis JB, Greene T, Rogers N, Roth D. 1999. A more accurate method to estimate glomerular filtration rate from serum creatinine: a new prediction equation. *Ann. Intern. Med.* 130(6):461–70
40. Levey AS, Coresh J, Greene T, Stevens LA, Zhang Y, et al. 2006. Using standardized serum creatinine values in the modification of diet in renal disease study equation for estimating glomerular filtration rate. *Ann. Intern. Med.* 145(4):247–54
41. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AF III, et al. 2009. A new equation to estimate glomerular filtration rate. *Ann. Intern. Med.* 150(9):604–12
42. Manski CF, Mullahy J, Venkataramani AS. 2023. Using measures of race to make clinical predictions: decision making, patient health, and fairness. *PNAS* 120(35):e2303370120
43. Martin JA, Hamilton BE, Osterman MJK, Driscoll AK, Drake P. 2018. Births: final data for 2017. *Natl. Vital Stat. Rep.* 67(8):34
44. Muntner P, Colantonio LD, Cushman M, Goff DC Jr., Howard G, et al. 2014. Validation of the atherosclerotic cardiovascular disease pooled cohort risk equations. *JAMA* 311(14):1406–15
45. Natl. Cancer Inst. 2024. *SEER incidence data, November 2023 submission (1975–2021), SEER 22 registries*. SEER*Explorer, Surveill. Res. Prog., Bethesda, MD, retrieved May 8. **https://seer.cancer.gov/statistics-network/explorer/**
46. Natl. Cancer Inst. 2024. *U.S. mortality data (1969–2022)*. SEER*Explorer, Surveill. Res. Prog., Bethesda, MD, retrieved May 8. **https://seer.cancer.gov/statistics-network/explorer/**
47. Natl. Lung Screen. Trial Res. Team, Aberle DR, Berg CD, Black WC, Church TR, et al. 2011. The National Lung Screening Trial: overview and study design. *Radiology* 258(1):243–53
48. Natl. Res. Counc. (US) Comm. Popul. 1997. *Racial and Ethnic Differences in the Health of Older Americans*, ed. LG Martin, BJ Soldo. Washington, DC: Natl. Acad. Press
49. Obermeyer Z, Powers B, Vogeli C, Mullainathan S. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366:447–53
50. Pasquinelli MM, Tammemägi MC, Kovitz KL, Durham ML, Deliu Z, et al. 2021. Brief report: risk prediction model versus United States Preventive Services Task Force 2020 draft lung cancer screening eligibility criteria—reducing race disparities. *JTO Clin. Res. Rep.* 2(3):100137
51. Pauker SG, Kassirer JP. 1980. The threshold approach to clinical decision making. *N. Engl. J. Med.* 302(20):1109–17
52. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, et al. 2010. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association Get with the Guidelines program. *Circ. Cardiovasc. Qual. Outcomes* 3:25–32
53. Rivera MP, Katki HA, Tanner NT, Triplette M, Sakoda LC, et al. 2020. Addressing disparities in lung cancer screening eligibility and healthcare access. An official American Thoracic Society statement. *Am. J. Respir. Crit. Care Med.* 202(7):e95–112
54. Sandel MJ. 2009. *Justice: What's the Right Thing to Do?* New York: Farrar, Straus and Giroux
55. Schwandt H, Currie J, Bär M, Banks J, Bertoli P, et al. 2021. Inequality in mortality between Black and White Americans by age, place, and cause and in comparison to Europe, 1990 to 2018. *PNAS* 118(40):2104684118
56. Shaikh N, Lee MC, Stokes LR, Miller E, Kurs-Lasky M, et al. 2022. Reassessment of the role of race in calculating the risk for urinary tract infection: a systematic review and meta-analysis. *JAMA Pediatr.* 176(6):569–75
57. Smith WR, Poses RM, McClish DK, Huber EC, Clemo FLW, et al. 2002. Prognostic judgments and triage decisions for patients with acute congestive heart failure. *Chest* 121(5):1610–17
58. Stevens ER, Caverly T, Butler JM, Kukhareva P, Richardson S, et al. 2023. Considerations for using predictive models that include race as an input variable: the case study of lung cancer screening. *J. Biomed. Inform.* 147:104525
59. Stevens LA, Coresh J, Greene T, Levey AS. 2006. Assessing kidney function—measured and estimated glomerular filtration rate. *N. Engl. J. Med.* 354(23):2473–83

60. Tammemägi MC, Katki HA, Hocking WG, Church TR, Caporaso N, et al. 2013. Selection criteria for lung-cancer screening. *N. Engl. J. Med.* 368(8):728–36. Erratum. 2013. *N. Engl. J. Med.* 369(4):394
61. Vyas DA, Eisenstein LG, Jones DS. 2020. Hidden in plain sight—reconsidering the use of race correction in clinical algorithms. *N. Engl. J. Med.* 383(9):874–82
62. Vyas DA, James A, Kormos W, Essien UR. 2022. Revising the atherosclerotic cardiovascular disease calculator without race. *Lancet Digit. Health* 4(1):e4–5
63. Vyas DA, Jones DS, Meadows AR, Diouf K, Nour NM, Schantz-Dunn J. 2019. Challenging the use of race in the vaginal birth after cesarean section calculator. *Women's Health Issues* 29(3):201–4
64. Wang JH, Hart A. 2021. Global perspective on kidney transplantation: United States. *Kidney360* 2(11):1836–39
65. Wilson PWF, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. 1998. Prediction of coronary heart disease using risk factor categories. *Circulation* 97(18):1837–47
66. Zanger-Tishler, Nyarko J, Goal S. 2024. Risk scores, label bias, and everything but the kitchen sink. *Sci. Adv.* 10:eadi8411
67. Zink A, Obermeyer Z, Pierson E. 2024. Race adjustments in clinical algorithms can help correct for racial disparities in data quality. *PNAS* 121(34):e2402267121