Northwestern | INSTITUTE FOR POLICY RESEARCH

IPR Working Paper Series

WP-22-47

Using Measures of Race to Make Clinical Predictions: Decision Making, Patient Health, and Fairness

Charles F. Manski
Northwestern University and IPR

John Mullahy
University of Wisconsin-Madison

Atheendar Venkataramani
University of Pennsylvania

Version: December 5, 2022

DRAFT

Abstract

The use of race measures in clinical prediction models and algorithms has become a highly contentious issue, driven by concerns that inclusion of race as a covariate exacerbates and perpetuates long-standing disparities in quality of health care provided to racial and ethnic minority patients. The authors seek to inform and ground this debate by evaluating the inclusion of race—even if imperfectly measured—in probabilistic predictions of illness that aim to inform clinical decision making. First, adopting a utilitarian framework to formalize social welfare, their analysis reveals that patients of all races are better off when clinical decisions are jointly guided by patient race and other observable covariates. In this sense, race is not a particularly special covariate: Any covariate with predictive power (i.e., one that changes conditional probabilities of illness) should be used to optimize clinical decisions. The researchers then extend the model to a two-period setting where prevention activities that address systemic drivers of disease are relevant and find that the same basic conclusions emerge. Finally, they discuss formal non-utilitarian concepts of fairness and disparity-aversion that have been proposed to guide societal allocation of health care resources.

The authors are grateful for insightful comments and suggestions offered by several colleagues.

1. Introduction

Recent years have seen considerable debate around two related questions that are sometimes conflated:

Should the manner in which a clinician treats a patient depend on that patient's race?

Should measures of race be included as covariates in clinical prediction models/algorithms?

Numerous arguments for and against the use of race in clinical settings have been advanced. For a concise overview of this debate see the recent contributions by Manski (2022) and Briggs (2022), who summarize the main opposing points of view.

Clear articulation of the goals we wish to achieve with patient care can help clarify the underlying disagreement and perhaps even resolve some of it. Two goals appear prominent. The first concerns the clinician's role in improving the health of the individual patient in front of them. The second concerns a societal objective of eliminating racial disparities in health care and health. Reasonable people may differ in how much weight they place on each goal. This matters because efforts to advance achievement of one goal may not advance, indeed may hinder, achievement of the other. Regardless of one’s priorities, conceptual clarity is required to achieve either or both goals.

This study aims to lay a clear and rigorous foundation to inform ongoing debates about the use of measures of race in clinical settings. We develop a simple model of the effects of considering race on clinical prediction and decisions and then generalize it. We begin with a canonical model that uses a single-period utilitarian framework wherein a clinician’s sole objective in any clinical encounter is to make the treatment recommendation that maximizes the patient’s expected health, conditional on all the information available at the time the treatment recommendation is made. The main result is that failure to use all available information in clinical prediction models or in a particular clinical encounter results in sub-optimal expected health for patients. In particular, statisticians' failure to use observed measures of race in developing clinical prediction algorithms or clinicians' failure to use available measures of race in treatment decisions will generally result in sub-optimal expected health outcomes for patients of *all* races. This result holds regardless of the extent to which available measures of race correlate with ancestry, socioeconomic status, or other drivers of health. Algorithms that use better measures of underlying drivers (e.g., direct measures of genotype, social deprivation, or biomarkers that capture these processes) may obviate the use of race in specific clinical settings (Powe, 2020; Hsu et al., 2021; Inker et al., 2021). However, until such alternate measures are available, algorithms including race would (weakly) outperform those that do not, by virtue of capturing a range of important correlates of health, no matter how imperfectly.

We then extend the canonical model by embedding the clinician's decision in a two-period framework wherein the period-two clinical decision is made in a context where social and environmental factors differentially affect patients in the first period. This model takes into account theoretical models of structural or systemic racism, wherein socially-produced disadvantages over the patient's life course adversely affect their health and well-being (Bailey et al., 2017; O'Brien et al., 2020; Bohren et al., 2022; Darity, 2022). As a result, the circumstances of patients seeking care in period two will differ, with some patients being more advantaged than others. Specifically, when period two arises, there will be disparities across the patient population in economic opportunities, healthcare access, and other determinants of health. Despite these disparities, our analysis shows that the clinician's role in period two *as a clinician* should still be to provide optimal care to each patient in the same manner as recommended by our canonical model. Increasingly, clinicians are advocating and striving for reductions in health disparities. But this is a separate matter than the activities they pursue and decisions they make in a clinical encounter. In our two-period utilitarian model, these activities are best pursued in period one, when prevention activities addressing social and structural drivers of health can reduce disadvantages in period two.

We recognize that the utilitarian framing of our analysis may not appeal to practitioners and policymakers who prioritize non-utilitarian notions of justice and fairness. For example, they may prioritize ensuring groups receiving similar (access to) health care resources as a first order goal. To situate these perspectives, we formally discuss alternate notions of fairness and disparity-aversion proposed to guide societal allocation of resources.

We also offer an Appendix to highlight two additional, related literatures. First, we note the rapidly expanding literature on *algorithmic fairness* in economics, computer science, and related fields. With some exceptions, this recent literature poses criteria for fairness and seeks to empirically measure adherence to them, without embedding them in a problem of welfare maximization as we do here. While our formal analysis does not link explicitly to this literature, we outline some of its noteworthy features (section A.1 of the Appendix). Second, we consider the practice of "race-norming," and illustrate that it is conceptually distinct from the main issue we consider here, which is the use of race measures to inform clinical predictions (section A.2 of the Appendix).

We proceed as follows. Section 2 provides background on the recent concern with and several examples of the use of race in medical decision making. Section 3 presents the one-period model of utilitarian treatment choice. Section 4 extends the model to a two-period setting in which preventive care precedes treatment choice. Section 5 discusses non-utilitarian views on fairness and justice. Section 6 concludes.

2. Background

2.1 Decision Making Contexts

Much of clinical decision-making involves the quantitative prediction of disease risk, treatment effectiveness, or other outcomes based on various sources of data. The canonical empirical prediction model, sometimes called a clinical algorithm, is an estimate $\hat{P}(y|x,z)$ of a probability $P(y|x,z)$ where y is the health outcome of interest and x and z denote vectors of covariates on which the prediction is based (Manski, 2019). For the present discussion, we take x to be covariates that will certainly be included in the model (e. g., patient age) and z to be variables that may be included at the analyst's discretion.[1]

The consideration that occupies our attention in this paper is the selection of the z variables. We do not endow $P(y|x,z)$ with any causal interpretation. The task at hand is to choose z to generate conditionally optimal predictions. It will be demonstrated formally in sections 3 and 4 that richer specifications of z generally yield superior clinical predictions than do sparser ones. So long as a candidate covariate z_k has some predictive power, its inclusion in the z vector will result in superior predictions.

2.2 Inclusion of Race in Prediction Models

There is presently considerable debate around the inclusion in prediction models of a particular z_k: patient race. In the clinical literature, Vyas *et al.* (2020) and Cerdeña *et al.* (2020) represent notable examples of calls to remove the consideration of race in prediction models, while Powe (2020) summarizes a range of arguments. In economics, Manski (2022) and Briggs (2022) summarize the key arguments supporting (Manski) and criticizing (Briggs) the inclusion of measures of race in clinical prediction models.

Manski (2022) describes and then questions four assertions that have been advanced as arguments against the inclusion of race in prediction models. These assertions are: (1) race is a social, not biological, concept; (2) race should not be considered if there is no established causal link between race and the illness; (3) using race may perpetuate or worsen racial health inequities; and (4) many persons are offended by the use of race in risk assessment. With the stated goal of making clinical decisions that would be expected to yield the best outcome for each patient, Manski concludes his paper with this observation:

[1] Section A.2 of the Appendix discusses the relationship between such clinical prediction models and the so-called race-norming of outcomes.

> If an alternative perspective is to have a compelling foundation, it should explain why society should find it acceptable to make risk assessments using other patient characteristics that clinicians observe, but not race. It should explain why the social benefit of omitting race from risk assessment is sufficiently large that it exceeds the harm to the quality of patient care.

In a somewhat similar vein, Powe (2020) states:

> There is no time more important than now to understand how race and social and biological factors interact to affect health. Estimation of essential physiologic processes, such as kidney function, with variables that do not incorporate race and are more accurate than race is a worthy aspiration. Those estimating tools should have equal or greater precision, be soundly grounded in evidence on health outcomes, and be acceptable to patients.

A common concern voiced by those arguing against the inclusion of race in prediction is that its very definition is complex and elusive. Many writers note that measured race historically has reflected a social rather than biological concept. When they argue that this precludes the use of race in making clinical predictions, they may have in mind that there is not a one-to-one mapping between a specified racial categorization and elements of ancestry or epigenetic modifications – e.g., a particular set of genotypes or stress-led differences in patterns of gene expression – that predict disease risk (Vyas et al., 2020). Another argument against the inclusion of race stems from a belief that, no matter how race is defined, including race as a predictor of outcomes will result in inferior health care for racial and ethnic minority populations relative to the care received by others. Health care that is sensitive to a patient's race, it is asserted, risks exacerbating systemic racial biases that are argued to prevail (Jones, 2021). Briggs captures this sentiment concisely:

> …while it is difficult to refute the central contention that optimal decision making requires the use of all covariates that are associated with outcome, the assumption that racial covariates, and their application within the medical arena, are sufficiently free from bias (structural, institutional or personal) misses the point of the underlying argument: that race is not the same as every other covariate in our arsenal. It is a covariate that is acting as a proxy for a wide range of other explanatory variables that could be genetic/biological but in many circumstances are more likely to be sociological/socioeconomic.

As we consider the controversy regarding inclusion of race in clinical prediction and decision-making, we think it important to recognize that the two sides of the debate do not simply differ in advocating different strategies for attaining the same goal. Rather, the core differences also concern the attainment of different goals, though these goals are often not explicitly or clearly articulated. Advancing the debate will thus require clarity around the desired goals and the measurements necessary to assess whether or not these goals are being achieved. Improved patient health is a straightforward goal to define and measure. It is less obvious how to conceptualize clearly and measure other objectives that have often been stated, which include the achievement of equity and elimination of disparities in health care (e.g., Sen, 2002). It is understandable that different parties may have different goals, reflecting different values.

This is all the more reason that a clear articulation of such goals is a necessary antecedent of productive discussion and policymaking.

One notable example of the lack of clarity in the literature concerns the common use of the term "bias." This term can mean different things to different people in different contexts. Bias in the sense typically encountered in discussions or race has little or nothing to do with how the expectation of a statistical estimator compares to its true value. Bias has a relatively clear meaning in the study of “algorithmic bias,” an area of inquiry kindred to but distinct from what we pursue in this paper (Obermeyer et al., 2019).[2] Beyond this literature, we find that bias is often weakly conceptualized. Careful scrutiny of statements about the presence of racial bias is essential for advancing understanding.

2.3 Examples

The issues raised above are of more than just academic methodological interest. Across an increasingly broad spectrum of clinical decision-making contexts, debates about whether to include race measures—however defined—in clinical decision making are shaping clinical practice. Examples include the treatment and management of kidney disease (Delgado *et al.* 2021), the assessment of osteoporosis and fracture risk associated with osteoporosis (Kanis et al., 2020), the use of spirometry to assess lung function (Bonner and Wakeam, 2022), and the determination of appropriate X-ray radiation doses (Bavli and Jones, 2022).[3] Vyas et al. (2020) and Cerdeña et al. (2020) review clinical contexts wherein such considerations have arisen. Decisions to include or exclude race in clinical prediction are already affecting the health care being delivered to patients of all races.

In some fields, leading institutions have formally recommended race-free risk assessment. A notable case is Delgado *et al*. (2021), which presents the recommendations of the National Kidney Foundation-American Society of Nephrology (NKF-ASN) Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. The Task Force considered the prevailing use of race in computation of estimated glomerular filtration rate(eGFR), a measure of kidney function. It recommended removal of race as a determinant of eGFR, writing (pp.5-6):

[2] Briefly, a prominent version of algorithmic bias arises when proxy outcome measures are used instead of true measures in prediction models and when the gap between the proxy and the true measures depends on some variable of interest, e.g. race. See section A.1 of the Appendix for further discussion.

[3] In some clinical contexts attention has been paid to whether diagnostic accuracy using traditional diagnostic standards may depend not per se on race but rather on skin tone. Examples include pulse oximetry and assessment of pressure injuries or ulcers. Many of the issues raised herein with respect to race inclusion would appear equally applicable to skin tone, with a key distinction being that there are accepted objective measures of skin tone (e.g. McCreath et al., 2016).

> For U.S. adults (>85% of whom have normal kidney function), we recommend immediate implementation of the CKD-EPI creatinine equation refit without the race variable in all laboratories in the U.S. because it does not include race in the calculation and reporting, includes diversity in its development, is immediately available to all labs in the U.S., and has acceptable performance characteristics and potential consequences that do not disproportionately affect any one group of individuals.

Research related to this recommendation is documented by Williams, Hogan, and Ingelfinger (2021). the underlying concern being that the use of race predictions may increase eGFR for Black patients, potentially reducing the likelihood of receiving therapy for chronic kidney disease or being listed for transplantation. The recommendation has already been implemented in multiple major medical centers. Powe (2020) urged caution, noting that calls to de-adopt race-inclusive algorithms did not consider “all of the ramifications and long-term health consequences,” some of which may introduce harm. For example, the move from using race-free measures of serum creatinine to race-adjusted eGFR measures helped reduce racial disparities in receipt of metformin, a diabetes medication with well-established short- and long-term benefits, by increasing use among patients who otherwise would have been contraindicated from receiving it under race-free measures (Shin et al., 2020). Some writers have argued that removal of race-adjustment may also increase rates of rejection of Black kidney donor candidates and lead to reductions in doses of chemotherapies (Levey et al., 2020). A recent analysis found a greater degree of misclassification in many race-free eGFR algorithms relative to algorithms including race, with the exception being a newly developed race-free algorithm that relies on an alternate biomarker, cystatin, instead of the more widely collected biomarker creatinine (Hsu et al., 2021).

Another prominent example is the use of race in models that predict fracture risks, with a particular focus on osteoporosis. For example, the FRAX algorithm (Fracture Risk Assessment Tool; https://www.sheffield.ac.uk/FRAX/index.aspx) for the U.S. population has four different versions to assess fracture risk for Black, Caucasian, Hispanic, and Asian patients. The use of these differential algorithms has been criticized for reasons that include the possibility that, ceteris paribus, lower fracture risk predictions for Black than White, Asian, and Hispanic patients may result in the former being undertreated to prevent osteoporosis or slow its progression if diagnosed (Vyas et al., 2020). In response to these criticisms, Kanis et al. (2020) note that there are indeed racial differences in osteoporosis treatment even after adjusting for fracture risk. Kanis et al. argue that racial differences in treatment gaps—the fraction of a population indicated for but not receiving treatment—are best understood and acted upon when risks are predicted accurately: "The quantification afforded by FRAX has allowed inequalities in the treatment gap to be identified." While Kanis et al. urge that FRAX not be used uncritically, they conclude that its use ultimately "helps to resolve, rather than exacerbate, racial inequalities."

2.4 The State of the Debate

Across the many clinical contexts where these issues have arisen, perspectives like those offered by Powe (2020), Kanis et al. (2020), and Manski (2022) are heard far less than calls to proceed with the removal of race from clinical decision-making (Vyas et al., 2020; Briggs, 2022). This dynamic holds even as arguments have been advanced to explicitly consider race in other (more population-level) settings. For example, there have been calls to allocate COVID-19 vaccines on the basis of existing and predicted disparities in disease risk (e.g. Schmidt et al, 2020) and for greater inclusivity in clinical trials of subjects from sub-populations that have been typically underrepresented (e.g. racial minorities) in such studies (U.S. Food and Drug Administration, 2020).[4] These tensions are borne out in statements such as those by Bhakta et al (2022):

> Research should continue to ascertain and use race, not as an explanatory variable, but rather in a manner that highlights where racial health disparities exist, that reduces enrollment bias, and that maximizes the generalizability of the results.

The lack of clarity and agreement in when and how race should be considered illustrates the need for a common, rigorous foundation to fix ideas and clarify goals. We demonstrate in the next sections that the application of one conceptually rigorous—and, we would argue, familiar and broadly applicable—foundation leads to the conclusion that the health and welfare of patients of all races is likely to suffer if the use of race in prediction models is proscribed. We offer this framework to help both sides of the debate identify incorrect, misguided, or unacceptable assumptions and to underscore that clear articulation of goals is indispensable. We thereafter describe alternative frameworks whose properties, assumptions, and implications can be similarly discussed and debated.

3. Utilitarian Patient Care

This section presents a standard medical-economics framing of utilitarian patient care as a single-period problem of treatment choice with predetermined risks of illness. Section 4 extends the framework to encompass preventive efforts that reduce risks of illness.

[4] U.S. Food and Drug Administration (2020) states: "Differences in response to medical products (e.g., pharmacokinetics, efficacy, or safety) have been observed in racially and ethnically distinct subgroups of the U.S. population. These differences may be attributable to intrinsic factors (e.g., genetics, metabolism, elimination), extrinsic factors (e.g., diet, environmental exposure, sociocultural issues), or interactions between these factors. Analyzing data on race and ethnicity may assist in identifying population-specific signals."

3.1. Optimal Care with Predetermined Illness Risk

Medical economists have commonly studied clinical decision making in a static setting of individual patient care. This setting supposes that a clinician must choose how to treat each person in a population of patients. The clinician observes certain predetermined covariates for each patient, with associated predetermined risks of illness. The objective is to maximize a utilitarian welfare function, one that sums up the benefits and costs of treatment across the population of patients. A utilitarian welfare function formalizes the idea of "patient-centered care." The assumption of individualistic care means that the care received by one patient may affect that person but does not affect other members of the population. This assumption is generally realistic when considering non-infectious diseases.

A common problem in clinical decision making is that treatments must be chosen with incomplete knowledge of their potential outcomes. A central idealization of the setting usually studied by medical economists has been to assume that the clinician knows the probability distribution of health and welfare outcomes that may potentially occur if a patient with specified observed attributes is given a specified treatment – i.e., the clinician has rational expectations. The assumption of rational expectations does not assert that the clinician can predict patient outcomes with certainty. Instead, it means that the clinician makes accurate probabilistic predictions conditional on observed patient covariates.

Analysis shows that if a clinician has rational expectations, the problem of optimizing patient care has a simple solution: patients should be divided into groups having the same observed covariates and all patients in a such a group should be given the care that yields the highest within-group mean utility. Our analysis with a utilitarian social welfare function suggests that it is optimal to differentially treat patients with different observed covariates if different treatments maximize their within-group mean utility. In our model, patients with the same observed attributes should be treated uniformly.

Analysis also shows that achievable utilitarian welfare across the population weakly increases as more patient covariates are observed. Observing more covariates enables a clinician to refine the probabilistic predictions of treatment outcomes on which decisions are based. Refining these predictions is beneficial to the extent that doing so affects optimal treatment choices.

These findings have been documented extensively. Abstract analyses, not specific to medical applications, include Good (1967), Manski (2007), and Kadane, Schervish, and Seidenfeld (2008). Analyses in the literature on medical economics include Phelps and Mushlin (1988), Meltzer (2001), Basu and Meltzer (2007), and Manski (2013). We prove the findings here in the simple instructive setting of choice between two treatments. Section 3.2 presents well-known findings. Section 3.3 develops a new finding that strengthens the conclusion drawn about the value of covariate information to treatment choice.

3.2. Optimal Choice Between Two Treatments

3.2.1. Treatments, Covariates, and Illness Probabilities

Suppose that a clinician must choose between treatments A and B. The choice is made without knowing a patient’s illness outcome, $y = 1$ or 0. We assume throughout that y measures accurately the health state on which a patient's welfare depends; that is, y is not a proxy of the sort considered by Obermeyer et al. (2019) in their investigation of algorithmic bias (see Appendix A.1). The clinician observes predetermined patient covariates (x, z). Having rational expectations, the clinician knows a patient’s probability of illness conditional on these covariates. Thus, the clinician knows $p_x = P(y = 1|x)$ and $p_{xz} = P(y = 1|x, z)$.

Consider patients who have the same value of x but who vary in their values of z. Assume that z takes values in a finite set Z. Each value of z occurs for a positive fraction of patients; thus, $P(z|x) > 0$ for all $z \in Z$. Assume that p_{xz} varies with z. To relate this setup to the concern with use of race in medical decision making, one may take z to be an observable measure of race. However, our abstract analysis does not require this specific interpretation of z. It holds when z is any observable covariate.

We show that maximum utilitarian welfare using p_{xz} to predict illness is always at least as large as using p_x. We note that this captures cases such as the prediction problem for kidney disease, where better predictors (of underlying drivers) of kidney function (eGFR) for all groups z have been identified, potentially obviating the need to include z as a covariate. We also note that maximum utilitarian welfare is strictly larger if optimal treatment choice varies with z. In the latter case, we characterize the determinants of the magnitude of the improvement.

3.2.2. Maximizing Expected Utility

Patient outcomes with each treatment depend on whether a patient has the disease. Patients are heterogeneous, so treatment response may differ across patients. Let $U_x(y, t)$ denote the expected utility that a patient with covariates x would experience with treatment t, should the illness outcome be y. We suppose for simplicity that this expected utility does not vary across patients with different values of z. However, illness probabilities may vary with z. A utilitarian clinician with rational expectations need not know the personal utility function of each individual patient, but is assumed to know expected utility $U_x(y, t)$ for each possible value of (x, t, y).

A clinician making a treatment decision does not know a patient’s illness outcome but does know the illness probabilities p_x and p_{xz}. With this knowledge, the clinician can compute expected utility in two ways, as $p_x{\cdot}U_x(1, t) + (1 – p_x){\cdot}U_x(0, t)$ or as $p_{xz}{\cdot}U_x(1, t) + (1 – p_{xz}){\cdot}U_x(0, t)$. Presuming that the objective is to maximize expected utility, the clinician may therefore use the criterion

(1a) choose treatment A if $p_x \cdot U_x(1, A) + (1 - p_x) \cdot U_x(0, A) \geq p_x \cdot U_x(1, B) + (1 - p_x) \cdot U_x(0, B)$,
(1b) choose treatment B if $p_x \cdot U_x(1, A) + (1 - p_x) \cdot U_x(0, A) \leq p_x \cdot U_x(1, B) + (1 - p_x) \cdot U_x(0, B)$,

or the criterion

(2a) choose treatment A if $p_{xz} \cdot U_x(1, A) + (1 - p_{xz}) \cdot U_x(0, A) \geq p_{xz} \cdot U_x(1, B) + (1 - p_{xz}) \cdot U_x(0, B)$,
(2b) choose treatment B if $p_{xz} \cdot U_x(1, A) + (1 - p_{xz}) \cdot U_x(0, A) \leq p_{xz} \cdot U_x(1, B) + (1 - p_{xz}) \cdot U_x(0, B)$.

With criterion (1), the maximized value of expected utility for patients with covariates x is

$$\max\,[p_x \cdot U_x(1, A) + (1 - p_x) \cdot U_x(0, A),\ p_x \cdot U_x(1, B) + (1 - p_x) \cdot U_x(0, B)]. \tag{3}$$

With criterion (2), the maximized value of expected utility for patients with covariates (x, z) is

$$\max\,[p_{xz} \cdot U_x(1, A) + (1 - p_{xz}) \cdot U_x(0, A),\ p_{xz} \cdot U_x(1, B) + (1 - p_{xz}) \cdot U_x(0, B)]. \tag{4}$$

In this case, the maximized value of expected utility for patients with covariates x is the mean of (4) with respect to the distribution of z conditional on x; that is,

$$E_{z|x}\{\max\,[p_{xz} \cdot U_x(1, A) + (1 - p_{xz}) \cdot U_x(0, A),\ p_{xz} \cdot U_x(1, B) + (1 - p_{xz}) \cdot U_x(0, B)]\}. \tag{5}$$

3.2.3 Comparison of the Criteria Using Jensen's Inequality

Jensen's inequality shows that the magnitude of (5) weakly exceeds that of (3), implying that criterion (2) performs at least as well as (1) from the utilitarian perspective. In particular,

$$\begin{aligned} &E_{z|x}\{\max\,[p_{xz} \cdot U_x(1, A) + (1 - p_{xz}) \cdot U_x(0, A),\ p_{xz} \cdot U_x(1, B) + (1 - p_{xz}) \cdot U_x(0, B)]\} \\ &\quad \geq \max\,\{E_{z|x}(p_{xz}) \cdot U_x(1, A) + [1 - E_{z|x}(p_{xz})] \cdot U_x(0, A),\ E_{z|x}(p_{xz}) \cdot U_x(1, B) + [1 - E_{z|x}(p_{xz})] \cdot U_x(0, B)]\} \\ &\quad = \max\,[p_x \cdot U_x(1, A) + (1 - p_x) \cdot U_x(0, A),\ p_x \cdot U_x(1, B) + (1 - p_x) \cdot U_x(0, B)]. \end{aligned} \tag{6}$$

The first inequality follows from Jensen's inequality because $\max(\cdot, \cdot)$ is a convex function. The second equality holds because $E_{z|x}(p_{xz}) = p_x$. The first inequality in (6) is strict if there exist some values of z for which criterion (2) yields a different treatment than criterion (1). The inequality is an equality if criteria (2) and (1) have the same solution for all values of z.

3.3. Direct Comparison of the Criteria

Jensen's inequality provides a simple proof of the qualitative result that a utilitarian clinician should use all observed covariates to predict illness. However, it does not reveal quantitatively the extent to which criterion (2) outperforms criterion (1). We can do this through direct comparison of the criteria. As far as we are aware, the derivation presented here is new.

Without loss of generality, let treatment A be optimal in (1). Let A be optimal in (2) for all $z \in Z_A$ and let Z_B be the complement of Z_A. Thus, inequality (2a) holds for $z \in Z_A$, some non-empty proper subset of Z, and does not hold for $z \in Z_B$, also a non-empty proper subset of Z. Criterion (2) yields better outcomes than (1) for persons with $z \in Z_B$ and the same outcomes as (1) for persons with $z \in Z_A$.

Now use the decomposition of Z into (Z_A, Z_B) to rewrite (3) and (5) as

$$
\begin{aligned}
(7)\quad & \max\,[p_x \cdot U_x(1, A) + (1 - p_x)\cdot U_x(0, A),\ p_x \cdot U_x(1, B) + (1 - p_x)\cdot U_x(0, B)] \\
& = p_x \cdot U_x(1, A) + (1 - p_x)\cdot U_x(0, A) \\
& = P(z \in Z_A|x)\cdot E[p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A)|\ x, z \in Z_A] \\
& \quad + P(z \in Z_B|x)\cdot E[p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A)|x, z \in Z_B]
\end{aligned}
$$

and

$$
\begin{aligned}
(8)\quad & E_{z|x}\{\max\,[p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A),\ p_{xz}\cdot U_x(1, B) + (1 - p_{xz})\cdot U_x(0, B)]\} \\
& = P(z \in Z_A|x)\cdot E[p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A)|\ x, z \in Z_A] \\
& \quad + P(z \in Z_B|x)\cdot E[p_{xz}\cdot U_x(1, B) + (1 - p_{xz})\cdot U_x(0, B)|x, z \in Z_B],
\end{aligned}
$$

Subtracting (7) from (8) yields

$$
(9)\quad P(z \in Z_B|x)\cdot E\{[p_{xz}\cdot U_x(1, B) + (1 - p_{xz})\cdot U_x(0, B)] - [p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A)]|x, z \in Z_B\}.
$$

The inequality $p_{xz}\cdot U_x(1, B) + (1 - p_{xz})\cdot U_x(0, B) > p_{xz}\cdot U_x(1, A) + (1 - p_{xz})\cdot U_x(0, A)$ holds for all $z \in Z_B$. Hence, (9) is positive. This qualitative finding repeats the earlier one using Jensen's inequality. What is new here is that (9) quantifies the extent to which criterion (2) outperforms criterion (1). The magnitude of (9) is the product of two factors. One is the fraction $P(z \in Z_B|x)$ of patients for whom treatment B yields strictly larger expected utility than treatment A. The other is the mean gain in expected utility that criterion (2) yields for the subset Z_B of patients.

3.4. Utilitarian Care and Disparities in Treatment and Health

Utilitarian treatment choice aims to maximize patient well-being, optimizing care within groups of patients who share common observed covariates. In this sense, it embeds a specific, clear idea that clinical decision making should be fair and just, expressing the idea that the primary concern of a clinician is to do best by each individual patient. Nevertheless, it does not imply that patients with different observed covariates receive the same treatments or experience the same health. In this sense it yields treatment, although not necessarily health, disparities across groups of patients (see Box 1).

Consider persons with covariates (x, z). The optimal utilitarian treatment of these persons is A if (2A) holds and is B if (2B) holds. The value of the maximized expected utility of these persons is given by (4), which varies with the illness probability p_{xz} and expected utility function $U_x(\cdot, \cdot)$. Cross-covariate disparities in treatment and health are well-motivated from the utilitarian perspective if clinicians have rational expectations and act accordingly. Of course, if $P(y|x,z)=P(y|x)$, then there is no compelling reason to include z in a prediction model. For example, use of race measures does not additionally improve predictions of kidney function in newly developed algorithms that use the biomarker cystatin (Hsu et al., 2021).

Cross-covariate disparities may not be well-motivated if, not having rational expectations, clinicians act based on imperfect knowledge of illness probabilities and expected utilities. Then clinical decisions may be sub-optimal. Moreover, the degree of sub-optimality may vary across patients with different observed covariates. For example, research suggests that clinicians' assessments of physical pain or cardiac risk among Black patients may be noisier or more prone to bias than those among White patients (Schulman et al., 1999; Hoffman et al., 2016; Sun et al., 2022) or that take-up of preventive services among Black patients may increase markedly if their doctor is also Black (Alsan et al., 2019). Sub-optimality of clinical decision making is always undesirable from the utilitarian perspective, regardless of which specific groups of patients most suffer the consequences. The utilitarian prescription to improve decision making is to improve knowledge of patient illness probabilities and expected utilities. One could do so by addressing biases in medical education (Burgess et al., 2007) or developing enhanced decision-support tools (Pierson et al, 2021). The utilitarian perspective implies that these solutions would dominate solutions that discard information from clinical prediction and decision-making on race entirely.

Some have deemed disparities undesirable from non-utilitarian perspectives on fairness and justice, particularly when z measures race or ethnicity. We discuss these alternate perspectives in Section 5.

4. Optimal Prevention and Treatment in a Two-Period Model

We now consider a two-period problem of utilitarian patient care. Period 2 remains as in Section 3, with a clinician choosing treatments for patients having predetermined risks of illness. The change is that we introduce an earlier period 1, in which society may undertake preventive efforts that reduce illness risk in period 2. If preventive care were costless, it would be optimal to provide it to all patients. However, such care is costly. Hence, the decision to provide it in period 1 is non-trivial.

To formalize the preventive-care decision in a simple manner, let s = 1 if a patient receives such care in period 1 and s = 0 otherwise. Suppose that society can personalize such care, choosing s separately for patients with different observed covariates. Let $C_{xz} > 0$ be the social cost of providing preventive care to patients having covariates (x, z), with cost expressed in units commensurate with patient utility. Let p_{sxz} denote the probability of illness in period 2, which now varies with s as well as with (x, z). We assume that preventive care reduces the risk of illness; thus, $p_{1xz} < p_{0xz}$.

We have found that, in period 2, it is optimal to condition illness risk on (x, z) when choosing a treatment, yielding (4) as the maximized value of expected utility. From the perspective of period 1, two versions of (4) are feasible, one setting s = 1 and the other setting s = 0. The resulting feasible maximized values of expected utility respectively are:

$$\text{(10a)}\quad \max\,[p_{1xz}\cdot U_x(1, A) + (1 - p_{1xz})\cdot U_x(0, A),\ p_{1xz}\cdot U_x(1, B) + (1 - p_{1xz})\cdot U_x(0, B)]\ -\ C_{xz},$$

$$\text{(10b)}\quad \max\,[p_{0xz}\cdot U_x(1, A) + (1 - p_{0xz})\cdot U_x(0, A),\ p_{0xz}\cdot U_x(1, B) + (1 - p_{0xz})\cdot U_x(0, B)].$$

Providing preventive care is optimal if and only if (10a) exceeds (10b). This inequality holds if preventive care reduces illness risk sufficiently and if such care is not too costly.

We present this model to connect our analysis to the emerging literature on systemic racism and its onward consequences on health and health disparities (Bailey et al 2017; O’Brien et al., 2020; Bohren et al., 2022; Darity, 2022). In that literature, long-standing discriminatory processes harm health over the life-course by increasing stress, reducing economic opportunities and financial security, and reducing access to health care. With this in mind, the term “preventive care” as used here can be interpreted either as direct preventive health care or more broadly to include any number of social activities or policies that reduce disease risk in period 2. Such policies may include broad economic or social interventions (e.g., minimum wage policy, social safety net policies, environmental policies, reparations) or interventions specifically targeted to address discrimination in period 1 (e.g., legal standards, affirmative action).

The key point is that while assessing the costs and benefits of such policies may be valuable, the clinician's optimal period 2 treatment decision remains as above, given whatever hand period 1 has dealt. Engaging clinicians to work on ameliorating problematic systemic drivers of health may be laudable. But our analysis shows that such work is generally separate from the care a clinician provides to a specific patient in period 2.

Consideration of a specific clinical setting may further elucidate these points. A clinician may be interested in prescribing insulin for a hospitalized patient with Type I diabetes, a condition for which failure to take insulin will result in death. That patient is experiencing economic precarity, thereby leading him to hold less generous health insurance. The clinician may want to prescribe a contemporary insulin regimen that most closely mimics physiology and thus now represents standard of care. However, despite the clinician's best efforts, coverage for the contemporary regimen is denied, while coverage for an older type of insulin regimen is provided. The clinician correctly discharges the patient with the older regimen, because this particular patient can afford it and thereby is more likely to adhere to it. However, outside this immediate encounter, the same clinician may advocate for policies that broaden population access to the latest medical technology (e.g., Essien et al, 2021) or, in a longitudinal relationship with the same patient, lay the groundwork for future care by address the patient's social needs in period 1.

5. Non-Utilitarian Concepts of Fairness

We showed in Section 3 that, if clinicians have rational expectations, utilitarian treatment choice optimizes care within groups of patients who share common observed covariates. This expresses a specific version of the idea that clinical decision making should be fair and just, without implying that patients with different observed covariates should receive the same treatments or will experience the same health. We observed that cross-covariate disparities may not be well-motivated from the utilitarian perspective if clinicians make sub-optimal decisions based on imperfect knowledge. The utilitarian prescription to cure sub-optimality is to improve clinical knowledge, thereby improving decisions for all groups of patients.

Writers arguing for removal of race from medical risk assessment appear to have something other than a utilitarian perspective in mind. Vyas *et al.* (2020) call for a general reconsideration of the use of race in risk assessments, stating: "Many of these race-adjusted algorithms guide decisions in ways that may direct more attention or resources to white patients than to members of racial and ethnic minorities." Cerdeña *et al.* (2020) states: "race-based medicine…perpetuates health-care disparities." Briggs (2022) states: "race is not the same as every other covariate in our arsenal." Unfortunately, such statements are typically too nebulous to enable readers to either reconcile or contrast them with the principles of utilitarian decision making.

However, here we do take note of two bodies of work that have sought to generate well-defined non-utilitarian concepts and measures of fairness, with particular application to racial equity. We discuss ideas developed in the economics literature presently. We describe computer-science work on algorithmic fairness in section A.1 of the Appendix.

5.1. Fairness in the Economics Literature

A simple framework illustrates key points, made originally by Foley (1968) and extended by Varian (1974). In the present health context, an *allocation* of resources in an N-member population is an allocation of treatments, $\mathbf{T}$, where, as above, each individual's treatment t_n is either A or B. Consider a population of N = 2, comprising of one Black patient (denoted J) and one White patient (denoted K), so that $\mathbf{T} = (t_J, t_K)$ where **T** is one of (A,A), (A,B), (B,A), or (B,B). Treatments result in scalar health outcomes, or expected outcomes if treatment response is uncertain, via individual-specific health production functions $h = h_n(t_n)$. We initially consider fairness of treatments rather than outcomes since, in general, the clinician—whose fairness is in question—is likely to have greater control over the former.

The treatment that results in the greatest health for J may or may not be the same as the treatment that is best for K; that is, $\text{sign}(h_J(A) - h_J(B)) \underset{<}{\geq} \text{sign}(h_K(A) - h_K(B))$. An algorithm that recommends uniform treatment for J and K when the signs are opposite implies that either J or K is not maximally healthy.

When there are no consumption externalities, **T** is said to be *envy-free* if $t_J \succeq_J t_K$ and $t_K \succeq_K t_J$; that is J and K each (weakly) prefers the treatment they receive under **T** to the treatment received by the other individual. It is natural to assume that the preference relationship corresponds directly to each individual's own h, although this is reconsidered below. If **T** is also Pareto efficient, then **T** is said to be *fair*. An algorithm that recommends a fair allocation to a clinician might be considered to be a *fair algorithm*. (Note however that the term “algorithmic fairness” is often used in a different sense - see section A.1 of the Appendix).

Suppose now that there are externalities in consumption (Thomson, 2011) or interdependent preferences (Pollack, 1976), wherein J and K care not just about the treatment they receive but also care about the treatment received by the other; that is, they have preferences over the allocations **T**. Such externalities might take many different forms, but one that is perhaps particularly relevant to recent discussions of racial equity is where individuals are *disparity-averse*.

Disparities of concern might imaginably be with respect to treatment allocations themselves or to the health outcomes arising from different treatment allocations. (See Box 1 for an example that underscores the distinction between treatment and outcome disparities.) Suppose for instance that $h_J(A) > h_J(B)$. If J is disparity-averse with respect to outcomes, it may be that J's preferences over the possible **T** outcomes nonetheless are:

$$(\mathrm{A},\mathrm{A}) \succeq_{\mathrm{J}} (\mathrm{B},\mathrm{B}) \succeq_{\mathrm{J}} (\mathrm{A},\mathrm{B}) \succeq_{\mathrm{J}} (\mathrm{B},\mathrm{A}) \qquad \text{(strong disparity aversion, SDA)}$$

or

$$(\mathrm{A},\mathrm{A}) \succeq_{\mathrm{J}} (\mathrm{A},\mathrm{B}) \succeq_{\mathrm{J}} (\mathrm{B},\mathrm{B}) \succeq_{\mathrm{J}} (\mathrm{B},\mathrm{A}) \qquad \text{(weak disparity aversion, WDA)}$$

Now J's preference ordering over treatment allocations depend on more than just J's own health; for example, $(\mathrm{B},\mathrm{B}) \succeq_{\mathrm{J}} (\mathrm{A},\mathrm{B})$ under SDA.

An alternative characterization of disparity aversion would take a societal perspective rather than the perspective of a given member of society. Without privileging any individual's treatment preferences, one might arrive at a societal ("S") partial ordering of the treatment allocations if social disparity aversion is with respect to treatments

$$\left\{ \begin{array}{c} (\mathrm{A},\mathrm{A}) \\ (\mathrm{B},\mathrm{B}) \end{array} \right\} \succeq_{\mathrm{S}} \left\{ \begin{array}{c} (\mathrm{A},\mathrm{B}) \\ (\mathrm{B},\mathrm{A}) \end{array} \right\}$$

Corresponding health outcomes might be reasonable partial tie-breakers if both individuals are, say, better off under A than B:

$$(\mathrm{A},\mathrm{A}) \succeq_{\mathrm{S}} (\mathrm{B},\mathrm{B}) \succeq_{\mathrm{S}} \left\{ \begin{array}{c} (\mathrm{A},\mathrm{B}) \\ (\mathrm{B},\mathrm{A}) \end{array} \right\}.$$

While the Foley-Varian notions of envy and fairness are not obviously applicable in this externality context, recognition that real-world individual and/or social preferences may sometimes have elements of disparity aversion could help us understand why some may view "fair" allocations in clinical contexts to differ from ones that are, by the above definitions, envy-free and Pareto efficient. Indeed, some recent calls for the removal of race measures in clinical predictions hint strongly at authors' disparity aversion. For instance, Vyas *et al.* (2020) write: "If doctors and clinical educators rigorously analyze algorithms that include race correction, they can judge, with fresh eyes, whether the use of race or ethnicity is appropriate.…They can discern whether the correction is likely to relieve or exacerbate inequities. If the latter, then clinicians should examine whether the correction is warranted."

6. Conclusions

One characterization of patient-centered care and personalized medicine (NEJM Catalyst, 2017) holds that:

> Not only are care plans customized, but medications are often customized as well. A patient's individual genetics, metabolism, biomarkers, immune system, and other "signatures" can now be harnessed in many disease states—especially cancer—to create personalized medications and therapies, as well as companion diagnostics that help clinicians better predict the best drug for each patient.

Marshaling data that provide informative measures of these "signatures" is an important empirical task. These signatures' measures may not be perfect. Nevertheless, as we show formally in sections 3 and 4, they have the potential to improve clinical predictions and, therefore, patients' health and well-being.[5]

A familiar mantra conceives healthcare quality as the delivery of the right care at the right time in the right place (e.g. Nowak et al., 2012). Strongly implied in this statement is that "the right care" means "the particular care that's best for each patient." Consider a thought experiment wherein two otherwise-identical healthcare delivery systems (say R and S) differ only in how clinical prediction models are used to inform clinicians' treatment decisions. In R treatment decisions are guided by race-inclusive models P(y|x,z), while in S treatment decisions are guided by race-exclusive models P(y|x). In which system would fully informed patients choose to enroll? In which system would fully informed clinicians choose to practice? While we cannot pretend to understand patients' and clinicians' motives for the choices they would make, we submit that questions of this nature are worth contemplating.

If implemented in practice, the arguments we have offered here ultimately support better quality care being delivered to patients of all races. While the model we outline meets a well-known mantra of healthcare quality, in practice, the way forward when making decisions around whether and how to consider race in clinical-decision making will require a much clearer sense of policymaker and public goals than has been achieved through the public debate to date. The present contribution is a step towards putting these debates on firmer and more transparent footing.

Our results have immediate and potentially widespread policy implications. For example, the U.S. Dept. of Health and Human Services (DHHS) has recently undertaken efforts to revise Section 1557 of the Affordable Care Act ("Nondiscrimination"). Among other things, these efforts include proposing a new § 92.210 that focuses on discrimination and clinical algorithms:

[5] Kanis et al. (2020) write: "...risk factors should be chosen according to established criteria irrespective of our understanding of their basis or their accuracy. A good example is consumption of alcohol, which is notorious for being inaccurately reported. In general, people who drink alcohol tend to neglect or underestimate their alcohol consumption.... It matters not whether the return is accurate—only that it provides a consistent indication of risk, which it does. Thus, we are more interested in association than causality. The same goes for race, location, and ethnicity."

> Proposed § 92.210 states that a covered entity must not discriminate against any individual on the basis of race, color, national origin, sex, age, or disability through the use of clinical algorithms in its decision-making.…The intent of proposed § 92.210 is not to prohibit or hinder the use of clinical algorithms but rather to make clear that discrimination that occurs through their use is prohibited.…The Department notes that the use of algorithms that rely upon race and ethnicity-conscious variables may be appropriate and justified under certain circumstances, such as when used as a means to identify, evaluate, and address health disparities. (U.S. DHHS, 2022)

It is for lawyers and regulators to offer suitable definitions of "discriminate against" and "identify, evaluate, and address health disparities." Our analysis suggests that if the goal of the health care system is to deliver optimal care by the standards developed in section 3, clinical appropriateness—what is best for the patient—must figure centrally in the deliberations. To this end clear distinctions must be drawn between differences in health care and differences in health outcomes.

We note in closing that there has been and remains considerable controversy regarding the definition of race. Race has been defined by ancestry, biology, social context, or a combination thereof. Race has also been defined based on self-identification or external perception. Moreover, these definitions may be fluid over time, even at the individual level. The challenges that definitional considerations like these pose for the preceding analysis are subtle but potentially important. Specifically, suppose a clinical trial is used as the basis of the evidence on which a clinician's decisions will be made, and suppose further that in the trial's data each participant's race was coded in some manner (the particular manner is unimportant). The evidence at hand for the clinician will then be a set of estimated prediction models $\hat{P}(y|x, z = z_t)$, one for each race category z_t and each x represented in the trial. The clinician now encounters a patient seeking treatment, observing their x characteristics and conceiving their race in some manner, say z_c. The question is whether the race category that serves as the basis of the clinician's treatment decision, z_c, is the same race category that would have been coded for this patient had they been a participant in the clinical trial. If so, the preceding analysis goes through without modification. If not, a more complex analysis must be pursued that is beyond this paper's scope.

References

Alsan, M., Garrick, O., and G. Graziani (2019), “Does Diversity Matter for Health? Experimental Evidence from Oakland,” *American Economic Review*, 109, 4071-4111.

Bailey Z., Krieger, N., Agenor, M., Graves, J., Linos, N., and M.T. Bassett. (2017), “Structural Racism and Health Inequities in the USA: Evidence and Interventions,” *The Lancet*, 389, 1453-1463.

Barocas, S., M. Hardt, and A. Narayanan (2021), *Fairness and Machine Learning—Limitations and Opportunities*. https://fairmlbook.org/ (accessed June 10, 2022)

Basu, A. and D. Meltzer (2007), "Value of information on preference heterogeneity and individualized care," *Medical Decision Making*, 27, 112-27.

Bavli, I. and D.S. Jones (2022), "Race Correction and the X-Ray Machine—The Controversy over Increased Radiation Doses for Black Americans in 1968," *NEJM*, 387, 947-952.

Bhakta, N.R., Kaminsky, D.A., Bime, C., Thakur, N., Hall, G.L., McCormack, M.C., and S. Stanojevic (2022), “Addressing Race in Pulmonary Function Testing by Aligning Intent and Evidence with Practice and Perception,” *Chest*, 161, 288-297.

Bohren, J.A., Hull, P., and A. Imas (2022), “Systemic Discrimination: Theory and Measurement,” *NBER Working Paper No 29820.*

Bonner, S.N. and E. Wakeam (2022), "The End of Race Correction in Spirometry for Pulmonary Function Testing and Surgical Implications," *Annals of Surgery*, 276, e3-e5. doi: 10.1097/SLA.0000000000005431

Briggs, A.H. (2022), "Healing the Past, Reimagining the Present, Investing in the Future: What Should Be the Role of Race as a Proxy Covariate in Health Economics Informed Health Care Policy?" *Health Economics*, 31, 2115-2119.

Burgess, D., van Ryn, M., Dovidio, J., and S. Saha (2007), “Reducing Racial Bias Among Health Care Providers: Lessons from Social-Cognitive Psychology,” *Journal of General Internal Medicine*, 22, 882-887.

Cerdeña, J.P., M.V. Plaisime, and J. Tsai. (2020), "From Race-Based to Race-Conscious Medicine: How Anti-Racist Uprisings Call Us to Act," *Lancet*, 396, 1125-1128.

Chen, V. and J.N. Hooker (2021), "Welfare-based Fairness through Optimization," Working Paper, Carnegie-Mellon University.

Christensen, D.M., J. Manley, J. Resendez (2021), "Medical Algorithms Are Failing Communities of Color," *Health Affairs Forefront*, DOI: 10.1377/forefront.20210903.976632

Claxton, K. (1999), "The Irrelevance of Inference: a Decision-making Approach to the Stochastic Evaluation of Health Care Technologies," *Journal of Health Economics*, 18, 341-364.

Darity, W.A. (2022), “Positions and Possessions: Stratification Economics and Intergroup Inequality,” *Journal of Economic Literature*, 60, 400-426.

Delgado, C., Baweja, M. Crews, D. *et al.* (2021). A Unifying approach for GFR Estimation: Recommendations of the NKF-ASN Task Force on Reassessing the Inclusion of Race in Diagnosing Kidney Disease. *Journal of the American Society of Nephrology*.

doi.org/10.1681/ASN.2021070988 .ones, D.S. (2021), "Moving Beyond Race-Based Medicine," *Annals of Internal Medicine*, 174, 1745-1746.

Essien, U.R., Dusetzina, S.B., and W. Gellad (2021), “A Policy Prescription for Reducing Health Disparities-Achieving Pharmacoequity,” *Journal of the American Medical Association*, 328, 1793-1794.

Foley, D. (1967), "Resource Allocation and the Public Sector," Yale Economic Essays, 7, 45-98.

Good, I. (1967), "On the Principle of Total Evidence," *The British Journal for the Philosophy of Science*, 17, 319-321.

Hobson, W. (2021), "How ‘race-norming’ was built into the NFL concussion settlement," *Washington Post*, Published August 2, 2021.

Hoffman, K.M., Trawalter, S., Axt, J.R., and M.N. Oliver (2016), “Racial Bias in Pain Assessment and Treatment Recommendations, and False Beliefs About Biological Differences Between Blacks and Whites,” *Proceedings of the National Academies of Science*, 113, 4296-4301.

Hsu, C., Yang, W., Parikh, R.V., Henderson, A.H., Chen, T.K., Cohen, D.L., He, J., Mohanty, M.J., Lash, J.P., Mills, K.T., Muiru, A.N., Parsa, A., et al (2021), “Race, Genetic Ancestry, and Estimating Kidney Function in CKD,” *New England Journal of Medicine*, 385, 1750-1760.

Inker, L.A., Eneanya, N.D., Coresh, J., et al (2021), “New Creatinine- and Cystatin C-Based Equations to Estimate GFR without Race,” *New England Journal of Medicine*, 385, 1737-1749.

Kadane, J., M. Shervish, and T. Seidenfeld (2008), "Is Ignorance Bliss?" *Journal of Philosophy*, 105, 5-36.

Kanis, J.A. et al. on behalf of the International Osteoporosis Foundation (2020), "FRAX and Ethnicity," *Osteoporosis International*, 31, 2063-2067.

Levey, A.S., Titan, S.M., Powe, N.R., Coresh, J., and Inker, L.A. (2020), “Kidney Disease, Race, and GFR Estimation,” *Clinical Journal of the American Society of Nephrology*, 15, 1203-1212.

Liang, A., J. Lu, and X. Mu (2022), "Algorithmic Design: Fairness versus Accuracy." Working Paper, Northwestern University.

Manski, C.F. (2007), *Identification for Prediction and Decision*, Cambridge, MA: Harvard University Press.

Manski, C.F. (2013), “Diagnostic Testing and Treatment under Ambiguity: Using Decision Analysis to Inform Clinical Practice,” *Proceedings of the National Academy of Sciences*, 110, 2064-2069.

Manski, C. F. (2019), *Patient Care under Uncertainty*, Princeton: Princeton University Press.

Manski, C.F. (2022), "Patient-Centered Appraisal of Race-Free Clinical Risk Assessment," *Health Economics*, 31, 2109-2114.

McCreath, H.E. et al. (2016), "Use of Munsell Color Charts to Measure Skin Tone Objectively in Nursing Home Residents at Risk for Pressure Ulcer Development," *Journal of Advanced Nursing*, 72, 2077-2085.

NEJM Catalyst (2017), "What is Patient-Centered Care?" January 1, 2017.

Nowak, N.A. et al. (2012), "Right Care, Right Time, Right Place, Every Time," *Healthcare Financial Management*, 66, 82-88.

Obermeyer, Z. et al. (2019), "Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations," *Science*, 366, 447-453.

Obermeyer, Z. et al. (2021), *Algorithmic Bias Playbook*. University of Chicago, Chicago Booth Center for Applied Artificial Intelligence.

O'Brien, R., Neman, T., Seltzer, N., Evans, L., and A.S. Venkataramani (2020), "Structural Racism, Economic Opportunity, and Racial Health Disparities: Evidence from U.S. Counties," *SSM-Population Health*, 11, e100564.

Pierson, E., Cutler, D., Leskovic, J., Mullainathan, S., and Z. Obermeyer (2021), "An Algorithmic Approach to Reducing Unexplained Pain Disparities in Underserved Populations," *Nature Medicine*, 27, 136-140.

Phelps, C. and A. Mushlin (1988), "Focusing Technology Assessment Using Medical Decision Theory," *Medical Decision Making*, 8, 279-289.

Pollak, R.A. (1976), "Interdependent Preferences," *American Economic Review*, 8, 309-320.

Powe, N. (2020). "Black Kidney Function Matters – Use or Misuse of Race?" *Journal of the American Medical Association*, 324, 737-738.

Rambachan, A. et al. (2020), "An Economic Perspective on Algorithmic Fairness," *AEA Papers and Proceedings*, 110, 91–95.

Sen, A.K. (2002), "Why Health Equity?" *Health Economics*, 11, 659-666.

Schmidt, H., Gostin, L.O., and M.A. Williams (2020), "Is it Lawful and Ethical to Prioritize Racial Minorities for COVID-19 Vaccines?" *JAMA*, 324, 2023-2024.

Schulman, K.A., Berlin, J.A., Harless, W., et al (1999), "The Effect of Race and Sex on Physicians Recommendations for Cardiac Catheterization," *New England Journal of Medicine*, 340, 618-626.

Shin, J., Sang, Y., Chang, A.R., Dunning, S.C., Coresh, J., Inker, L.I., Selvin, E., Ballew, S.H., and M.E. Grams (2020), "The FDA Metformin Label Change and Racial and Sex Disparities in Metformin Prescription Among Patients with CKD," *Journal of the American Society of Nephrology,* 31, 1847-1858.

Sun, M., Oliwa, T., Peek, M.E., and E.L. Tung (2022), “Negative Patient Descriptors: Documenting Racial Bias in the Electronic Health Record,” *Health Affairs*, 41, 203-211.

Thomson, W. (2011), "Fair Allocation Rules," Chapter 21 in *Handbook of Social Choice and Welfare*, Volume II. K. Arrow, A. Sen, and K. Suzumura, Eds. Amsterdam: Elsevier.

Tong, M. and S. Artiga. (2021), "Use of Race in Clinical Diagnosis and Decision Making: Overview and Implications," *Kaiser Family Foundation Issue Brief*, December 9, 2021.

U.S. Food and Drug Administration (2020), *Enhancing the Diversity of Clinical Trial Populations — Eligibility Criteria, Enrollment Practices, and Trial Designs: Guidance for Industry*, Center for Drug Evaluation and Research (CDER), Center for Biologics Evaluation and Research (CBER), U.S. Food and Drug Administration, Silver Spring, MD, November 2020.

U.S. Dept. of Health and Human Services (2022), "Nondiscrimination in Health Programs and Activities," *Federal Register*, August 4, 2022, 47824-47920, Docket ID: HHS-OS-2022-0012.

Varian, H.R. (1974), "Equity, Envy, and Efficiency," *Journal of Economic Theory*, 9, 63-91.

Vyas, D., L. Eisenstein, and D. Jones. (2020), "Hidden in Plain Sight—Reconsidering the Use of Race Correction in Clinical Algorithms," *New England Journal of Medicine* 383, 874-882.

Williams, W., Hogan, J., & Ingelfinger, J. (2021). Time to eliminate health care disparities in the estimation of kidney function. *New England Journal of Medicine*. doi/full/10.1056/NEJMe2114918.

Appendix

A.1. Algorithmic Bias and Algorithmic Fairness

In recent years considerable attention has been devoted to notions of algorithmic bias. Obermeyer et al. (2021) have authored a plain-language "playbook" that decision makers might consult to determine whether algorithms on which they rely are susceptible to bias. For these authors, biased algorithms are defined as follows:

> More generally, in many important social sectors, algorithms guide decisions about who gets what. In these situations, we believe that if an algorithm scores two people the same, those two people should have the same basic needs—no matter the color of their skin, or other sensitive attributes…We consider algorithms that fail this test to be biased. [page 1]

Algorithmic fairness is not defined merely by the absence of bias in an algorithm. Barocas et al. (2021) outlines challenges in defining algorithmic fairness:

> We should not expect work on fairness in machine learning to deliver easy answers. And we should be suspicious of efforts that treat fairness as something that can be reduced to an algorithmic stamp of approval. At its best, this work will make it far more difficult to avoid the hard questions when it comes to debating and defining fairness, not easier. It may even force us to confront the meaningfulness and enforceability of existing approaches to discrimination in law and policy, expanding the tools at our disposal to reason about fairness and seek out justice. [page 34]

These authors underscore such definitional challenges by offering *nineteen* distinct algorithmic fairness criteria (table 6 on page 75).

To investigate *how* algorithms may be unfair, Rambachan et al. (2020) suggest a decomposition of the difference in an algorithm's estimated predictions between two groups, $\hat{E}[\tilde{Y}|G_1] - \hat{E}[\tilde{Y}|G_2]$, where $\tilde{Y}$ is the measured outcome predicted by the algorithm. They write:

> A raw difference in predictions across groups may arise for three reasons: differences in base rates, differences in measurement error, or differences in estimation error across groups. Intervening at the level of the algorithm may address only the last two components of the disparity by investing in better training data and collecting a better proxy for the outcome of interest. In contrast, the base rate difference is a product of the underlying socioeconomic context itself, not the algorithm. [page 92]

While the growing literature on algorithmic fairness is challenging to summarize succinctly, we mention here a few recent contributions. One is the e-book by Barocas et al. (2021) that explores a large set of criteria for algorithmic fairness. Another is Chen and Hooker (2021) who approach algorithmic fairness (which they call fairness in AI systems) via welfare optimization approaches that involve definition of relevant social welfare functions and corresponding considerations of computational issues involved in their optimization. Liang et al. (2022) focus on algorithmic fairness and algorithmic predictive accuracy, exploring in particular when there will and will not tend to be tradeoffs (e.g., when greater fairness can be attained only at the expense of reduced predictive accuracy).

A.2. Race-Normed Outcomes

Distinct from this paper's focus is the issue of race norming of outcomes. While widely used, the term race norming does not to our knowledge have a commonly accepted technical definition. It does, however, claim a Wikipedia page (https://en.wikipedia.org/wiki/Race-norming) that offers this definition: "Race-norming, more formally called within-group score conversion and score adjustment strategy, is the practice of adjusting test scores to account for the race or ethnicity of the test-taker." This definition suggests transformation of an observed health outcome y that is measured in a uniform manner across races into a race-specific adjusted outcome, say $w(y, z)$ for a specified function $w(\cdot, \cdot)$. Then the conditional probability distribution of interest becomes $P[w(y, z)|x, z]$ rather than the $P(y|x, z)$ that has been our concern.

Race norming of clinical outcomes has been controversial. A prominent recent example is the National Football League's financial settlement with former players for concussion-related brain injuries (Hobson, 2021). The original settlement that denied compensation to some Black players was based on race-normed cognitive test outcomes that indicated Blacks to have different baseline cognitive capabilities than non-Blacks. Original settlements have been reconsidered given the questionable credibility of these baselines.

An important yet distinct issue in the NFL case is whether cognitive test scores are appropriate outcome measures on which to base compensation. It may be that what these tests measure is at best a proxy for true health status with the gaps between proxy and truth being race dependent. The use of such problematic proxies is one of the main concerns in the algorithmic bias work discussed earlier (e.g. Obermeyer et al., 2019).

Box 1: Optimal Treatments and Health Outcome Disparities

Suppose there is credible evidence that the optimal treatment for Black patients (who are denoted J) is treatment B and optimal treatment for White patients (who are denoted K) is treatment A. Assume that treatments A and B are different dosages of the same drug or intensities of the same intervention, with treatment intensity denoted t. Suppose the respective health production functions are:

$$h_J(t_J) = 8 - (t_J - 4)^2 \text{ and } h_K(t_K) = 10 - (t_K - 2)^2$$

Then A is $t_K = 2$ and B is $t_J = 4$. Thus, Black patients must receive higher-intensity treatment than White patients to attain optimal health.

Define health disparity as $D(t_J, t_K) = h_J(t_J) - h_K(t_K)$. When no patient receives treatment ($t_J = t_K = 0$), health is worse for Black patients than for White patients, with $D(0,0) = -14$.

The treatment allocation that is optimal by the utilitarian criteria developed in section 3 has Black patients receive B ($t_J = 4$) and White patients receive A ($t_K = 2$). The disparity is thus less negative than when no treatment is received, with $D(4,2) = -2$, although White patients' health is still better than Black patients'.

From this baseline, consider the implications for health of four alternative treatment allocations, each of which treats Black and White patients with a common intensity:

Alternative 1. Everyone is treated with what evidence indicates is best for White patients. Compared with the baseline, the disparity is more negative, with $D(2,2) = -6$.

Alternative 2. Everyone is treated with what evidence indicates is best for Black patients. The disparity is now positive, with $D(4,4) = 2$.

Alternative 3. Everyone is treated with a population weighted average of the optimal treatments, $C = pA + (1-p)B$, where p is Whites' population share). Assume p=.8. Then the common intensity is 2.4. Now disparity is more negative than at baseline, with $D(2.4,2.4) = -4.4$.

Alternative 4. Treatments are allocated to eliminate disparity, which occurs if the common intensity is 3.5. Then $D(3.5,3.5) = 0$. While disparity is eliminated, the health levels of both Black and White patients suffer relative to the baseline.

The implications of these treatment allocations for health levels and disparity are summarized in this table:

Scenario	t		Health		$D(t_J,t_K)$
	Black (J)	White (K)	Black (J)	White (K)	
No Treatment	0	0	-8	6	-14
Baseline: Optimal	4	2	8	10	-2
Alternative 1	2	2	4	10	-6
Alternative 2	4	4	8	6	2
Alternative 3	2.4	2.4	5.44	9.84	-4.4
Alternative 4	3.5	3.5	7.75	7.75	0