NBER WORKING PAPER SERIES

UNDERSTANDING ALGORITHMIC DISCRIMINATION IN HEALTH ECONOMICS THROUGH THE LENS OF MEASUREMENT ERRORS

Anirban Basu
Noah Hammarlund
Sara Khor
Aasthaa Bansal

Working Paper 29413
http://www.nber.org/papers/w29413

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2021

We would like to thank Peter Hull for his very useful comments. We also acknowledge support from a consortium of ten biomedical companies to the University of Washington through an unrestricted gift. All errors are ours. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Understanding Algorithmic Discrimination in Health Economics Through the Lens of Measurement Errors

Anirban Basu, Noah Hammarlund, Sara Khor, and Aasthaa Bansal

NBER Working Paper No. 29413
October 2021
JEL No. C53,I10,I14

ABSTRACT

There is growing concern that the increasing use of machine learning and artificial intelligence-based systems may exacerbate health disparities through discrimination. We provide a hierarchical definition of discrimination consisting of algorithmic discrimination arising from predictive scores used for allocating resources and human discrimination arising from allocating resources by human decision-makers conditional on these predictive scores. We then offer an overarching statistical framework of algorithmic discrimination through the lens of measurement errors, which is familiar to the health economics audience. Specifically, we show that algorithmic discrimination exists when measurement errors exist in either the outcome or the predictors, and there is endogenous selection for participation in the observed data. The absence of any of these phenomena would eliminate algorithmic discrimination. We show that although equalized odds constraints can be employed as bias-mitigating strategies, such constraints may increase algorithmic discrimination when there is measurement error in the dependent variable.

Anirban Basu
The CHOICE Institute
Departments of Pharmacy,
Health Services, and Economics
University of Washington, Seattle
1959 NE Pacific St., Box - 357660
Seattle, WA 98195
and NBER
basua@uw.edu

Noah Hammarlund
The CHOICE Institute
School of Pharmacy
University of Washington, Seattle
1959 NE Pacific St., Box - 357660
Seattle, WA 98195
noahh@uw.edu

Sara Khor
The CHOICE Institute
School of Pharmacy
University of Washington, Seattle
khors@uw.edu

Aasthaa Bansal
The CHOICE Institute
School of Pharmacy
University of Washington, Seattle
abansal@uw.edu

INTRODUCTION

Health equity among protected or disadvantaged groups is a core aim of federal, local, and private health policy initiatives (National Academies of Sciences, Engineering, and Medicine 2016, 2017). Disadvantaged groups may be defined by features such as race, gender, sexual orientation, disability, and other economic and social statuses. In 2003, the Institute of Medicine identified equitable care as an aim for high-value healthcare (Institute of Medicine 2003). In 2003, Congress mandated the production of the annual National Healthcare Disparities Report in an effort to monitor national progress (Ayanian, 2015). Still, differences in access to quality care, mortality from disease, and treatment rates across protected populations provide evidence of systematic health disparities in the United States (National Academies of Sciences, Engineering, and Medicine 2017). Racial treatment disparities are well documented (Institute of Medicine 2003; 2019 National Healthcare Quality and Disparities Report; Hammarlund 2021b). More recently, attention has been drawn to the phenomenon where decision-making following algorithmic predictions using machine learning-based approaches may exacerbate such disparities (Char et al.., 2018; Rajkomar 2018; Parikh 2019).

Predictive algorithms, which can induce and/or perpetuate inequity, are sometimes referred to as "unfair regressions". This phenomenon, known as "algorithmic bias" or "algorithmic discrimination", is not unique to healthcare. Analyses recently uncovered biased outcomes in algorithms across domains such as judicial practices (Angwin et al. 2016; Chouldechova 2917; Nabi and Shpitser, 2018), hiring, and promotion decisions (Dastin, 2018), and customs and border protection (Grother, 2019). Health care systems in the U.S. are increasingly incorporating machine learning into clinical medicine with the hope that more accurate predictions can improve health care delivery. Machine learning-based tools built by private companies and hospitals currently support physicians and sometimes function independently of them (Char et al, 2018). Decision support systems based on EMR analyses may now be able to decrease adverse events and complications in certain settings for the average patient and outperform clinicians on certain metrics (Freedman et al, 2018; Davenport, *T*. and Kalakota, *R*. 2019). However, in 2014 the President's Big Data Working Group argued that discrimination could be the "inadvertent outcome of the way big data technologies are structured and used" and pointed toward "the potential of encoding discrimination in automated decisions" (Podesta 2014; Hardt et al, 2016). Artificial intelligence applications can transform healthcare, but if concerns around algorithmic bias are not sufficiently addressed, historical health inequities will be perpetuated and potentially worsen disparities.

This paper brings together the literature from statistics, economics, and computer science on algorithmic discrimination to provide an overarching generic statistical description of how algorithm-based decisions can lead to inequities. In the next section, we start with a brief

overview of the literature on the use of machine learning methods in health economics, which have been shown to be subjected to algorithmic discrimination or vulnerable to it. In Section 3, we use a wide lens of measurement errors to describe this framework for the first time to our knowledge. We show that unlike the traditional literature on measurement errors, where the measurement errors in the independent variables (errors-in variable) generate bias in coefficient estimates while (classical) measurement error in the dependent variable affects efficiency but not bias, both types of measurement errors will contribute towards algorithmic bias. In Section 4, we discuss some of the proposed solutions in computer sciences and statistics literature and highlight their limitations in the context of measurement errors. Specifically, we focus on the most commonly used equalized odds constraint and show that such an approach can increase bias when there is a measurement error in the dependent variable. We hope that this paper can serve as guidance for applied researchers planning to develop algorithms to inform health and healthcare decisions and understand the limitations and solutions to some of these issues.

USE OF ALGORITHMS IN HEALTH ECONOMICS AT RISK OF ALGORITHMIC DISCRIMINATION

Machine learning (ML) has been increasingly used in health economics and outcomes research applications, hoping to unlock the potential of using large volumes of structured and unstructured data to transform healthcare and payment models (Rueda et al, 2019). Some of the most common types of health economics and outcomes research activities where ML has been applied are clinical decision support and predictive analytics. The goal is to use ML algorithms to predict individual risk of developing a disease or experiencing a health event. These algorithms can be embedded in tools to help clinicians allocate appropriate and timely care such as preventative care, intensive interventions, or disease surveillance. ML may be especially useful in providing early alerts to highly complex diseases, highlighting efficient care pathways, and improving the quality of care by detecting patterns that may be previously time-consuming or highly dependent on providers' skills and experiences. Existing applications have included predicting cardiovascular disease (Lloyd-Jones, 2010; Krittanawong et al, 2020), peripheral artery disease (Ross et al, 2016), heart failure (Desai, 2020), hypertension (Ye, 2018), aneurysm ruptures (Silva et al, 2019), diabetes (Lai et al, 2019), liver disease (Spann et al, 2020), postpartum depression (Zhang et al, 2021), COVID-19 (Gao et al, 2020), multimodal disease (Chen et al, 2017), cancer diagnosis and recurrence (Johri et al, 2021, Zafar et al, 2020) and all-cause mortality (Bergquist et al, 2020). Often these models rely on the accurate capture of the predictors and health outcomes in administrative databases and the electronic health records (EHR) based on ICD-9 diagnosis or procedure codes. However, these predictors and outcomes could be biased due to systemic inequality in the health care system. For

example, patients of lower SES may be less likely to receive appropriate and timely care due to healthcare access barriers and implicit biases by healthcare providers (e.g, undiagnosed hypertension or delayed treatment for cancer recurrence). These patients may also be more likely to be seen in different care settings (e.g, safety net clinics or emergency room) than the higher SES groups. Their clinical history and reasoning documentation may be incomplete or systematically different. Consequently, algorithms that are based on these biased data may "learn" to replicate the systemic biases in the data and recommend less testing and treatment to historically underserved patients, further perpetuating the systemic disparities in health care and health outcomes.

Another type of health economics application of ML is predicting healthcare spending. There has been increasing interest in using algorithms to identify high-cost, high-needs individuals and to augment health plan payment risk-adjustment formulas (Rose et al, 2017, Yang et al, 2018, Park et al, 2018, Rose 2016). When used appropriately, healthcare cost prediction can facilitate allocating scarce care management resources (Morid 2017, Obermeyer 2019). However, when observed health expenditure and healthcare utilization are used as prediction outcomes, bias may arise due to unequal access to health care across groups. Groups with worse healthcare access may use fewer health services and may appear to have lower observed healthcare expenditure despite poor health. Algorithms that assume that health care expenditure or utilization are accurate proxies for true health across all groups may make erroneous inferences about the risk for individuals with incomplete data, unintentionally directing resources away from patients who need them the most (Obermeyer et al. 2019). Algorithms used for predicting health plan payments have also been shown to differentially underpredict spending for some subgroups. This can lead to under-compensating health plans for selected individuals, which disincentives insurers from enrolling or providing full insurance to these individuals. Consequently, differential access to insurance systematically generates differential patterns of heath care across different groups, further propagating biases in any algorithms developed with such data. Fair regression methods have been proposed as a way to address some of these issues (Zink et al, 2020).

The third type of health economics and health services application of ML is for monitoring health through wearables and personal devices. Smartphones, watches, and other portable devices are now available to track patients' health and behavior around the clock in the comfort of their own home, providing a unique and large treasure trove of data that could offer valuable insights about their health, such as low heart rate alert, ECG monitoring, or sleep tracking. While these portable devices may be powerful remote diagnostic and health monitoring tools, their adoption has been uneven across groups. National surveys showed that the use of wearable devices differed by age, wealth, education level, and health status (Chandrasekaran et al, 2020). This

"digital divide" has implications on the differential benefits from the health technologies and the representativeness of the data collected. There are also concerns that these devices are not as accurate for individuals with darker skin tones (Colvonen et al, 2020). The lack of diversity in the training and validation data used to develop the ML algorithms may have contributed to the biased performance across groups with different skin tones. The differential inaccuracy of the information provided by these wearables, if not addressed, can reinforce existing healthcare disparities for those with darker skin tones.

A GENERIC STATISTICAL FRAMEWORK FOR UNFAIR REGRESSION

We present a generic statistical framework for unfair regression, which implies algorithmic discrimination for resource allocation. This framework is inspired by the works by Arnold et al. (2021) and Rambachan et al. (2020) and presents this problem through the lens of measurement errors.

Definitions of Discrimination and Its Components

Let us consider two groups of individuals in a population, differentiated by an observable characteristic (e.g., race) that is denoted by $R_i \in$ {a, b}. Let us also consider an allocation problem for each individual, where "qualification" to receive a specific level of a resource T_i is based on some unobserved (latent) variable Y_i^*. We denote Y_i^* as the "qualifying variable" (Arnold et al., 2021). Without loss of generality, let Y_i^* be absolutely continuous with respect to Lebesgue measure and indicate the "true" qualification for the allocation level of that resource for all individuals. In medicine, one can think of this qualifying variable as the "true" health status of an individual, and treatment is efficacious below a certain threshold of health. Therefore, they qualify to receive treatment.

In an ideal world, the absence of discrimination implies that one can allocate resources fairly to those who qualify for it, and this allocation rule is the same in each group. Suppose the allocation rule is, for some reason, different between the two groups. In that case, the difference in the mean allocation between the two groups, conditional on the latent variable Y^*, should be viewed as the extent of "discrimination" between these groups. That is, total discrimination between the two groups is given by (Arnold et al, 2021):

$$\Delta = E_{Y^*}\left[E[T_i|R_i = a, Y_i^*] - E[T_i|R_i = b, Y_i^*]\right] \quad (1)$$

The inner difference in Δ represents the difference in the average allocation of a resource between the two groups holding their true qualification Y^* constant. The outer expectation

averages this comparison over the marginal distribution of the qualification. If $\Delta \neq 0$, discrimination exists. $\Delta < 0$ indicates discrimination exists against group a, and vice versa.

It is essential to distinguish the concepts of discrimination versus disparity at this point. Our definition of *T* represents what is allocated or prescribed and not what is received. There are many reasons why observed receipt of resources may differ from that is being allocated. For example, physicians may prescribe a particular treatment, but a patient may or may not receive the treatment due to their own preferences, prejudices, or other social determinants such as structural racism (Alsan et al. 2021). All of these would lead to differences in observed receipt of treatment between two groups, representing disparity.[a] In our framework, we are mainly concerned about what is being prescribed or allocated, and therefore, focus on discrimination based on such allocation.

Because Y_i^* often remains unobserved or only partially observed, we typically develop an algorithm using data on an observed vector of characteristics (X_i) and a set of observed outcomes (Y_i) from other individuals to predict individual *i*'s qualification for treatment. An example of such an approach would be to use observed clinical characteristics to predict the long-term risk of developing a disease and then using these predictions to determine who should receive screening. Typically, *Y* could be the same variable as *Y** (e.g., death data), but *Y* is only observed for a selected group of individuals (e.g., poor surviellance in underserved population). Alternatively, *Y* could be a completely different variable from *Y** and acts as a proxy for *Y** (e.g, health care expenditures as a proxy for underlying health risk).

Because of the need for the intermediate step to develop an algorithm, let an algorithm predict $\hat{Y}_i$ based on a set of covariates or features (*X*) and allocation is made based on this algorithmic prediction. $\hat{Y}_i$ is also assumed to be absolutely continuous with respect to Lebesgue measure. A naïve option for *Y* could be the observed allocations (T_{obs}) in practice. This approach would be problematic for obvious reasons as an algorithm will only reproduce the existing discrimination in practice. This is similar to concepts of "selective labels" (Lakkaraju et al, 2017) and "bias - in/bias out" (Rambhachan and Roth, 2019) in the literature and highlights issues of biased human decision-makers generating the data. Ideally, *Y* would be some form of risk that mimics the appropriateness of receiving allocations.

Total discrimination, as defined in (1), can be broken down into two components – *human discrimination* and *algorithmic discrimination*:

[a] It should be noted that differences in treatment receipt driven by patients' preferences were not considered to be part of disparity by the IOM Panel on "Unequal treatment" (IOM, 2003).

$$\Delta = \mathrm{E}_{\mathrm{Y}^*}\left[\mathrm{E}\left[\mathrm{E}\left[\mathrm{T_i}|\mathrm{R_i}=\mathrm{a},\hat{Y}_i\right] \mid \mathrm{R_i}=\mathrm{a},\mathrm{Y_i^*}\right] - \mathrm{E}\left[\mathrm{E}\left[\mathrm{T_i}|\mathrm{R_i}=\mathrm{b},\hat{Y}_i\right] \mid \mathrm{R_i}=\mathrm{b},\mathrm{Y_i^*}\right]\right] \quad (2)$$

For notational simplicity, we drop subscript *i* henceforth. *Human discrimination* is driven by prejudice, preferences, and also statistical discrimination in decision making (Balsa and McGuire 2001; Alsan et al. 2019), and arises when allocation is different between the two groups conditional on a predicted score of $\hat{Y}$, i.e.,

$$\mathrm{E}\left[T|R=\mathrm{a},\hat{Y}\right] \neq \mathrm{E}\left[T|R=\mathrm{b},\hat{Y}\right] \quad (3)$$

Since the focus of this paper is on *algorithmic discrimination*, we will assume that there is no human discrimination in our framework. This is merely to highlight the role of algorithmic discrimination, and not to say that human discrimination does not exist in practice, because it certianly does. This assumption implies that allocation decisions are *fair* and not inherently disparate conditional on the predicted index.[b] For example, without loss of generality, let the allocation of treatment follow:

$$\mathrm{E}\left[T|R=\mathrm{a},\hat{Y}\right] = \mathrm{E}\left[T|R=\mathrm{b},\hat{Y}\right] = g(\hat{Y}) \quad (4)$$

Therefore, without human discrimination, total discrimination in (2), now denoted as Δ^*, can be written as:

$$\Delta^* = \mathrm{E}_{\mathrm{Y}^*}\left[\mathrm{E}\left[\, g(\hat{Y})|\, R=\mathrm{a},\mathrm{Y}^*\,\right] - \mathrm{E}\left[\, g(\hat{Y}) \mid R=\mathrm{b},\mathrm{Y}^*\,\right]\right] \quad (5)$$

$g(\hat{Y})$ can be either a linear or a non-linear function depending on the nature of allocation. In this work, we focus on a linear function to illustrate the role of measurement errors more clearly. We believe that our results can be extended to a non-linear $g(\hat{Y})$, but we delegate this work to the future.[c]

Under the assumption of linearity in $g(\hat{Y})$, and the absence of human discrimination, total discrimination can be written as:

$$\Delta^* = \mathrm{E}_{\mathrm{Y}^*}\left[\mathrm{E}\left[\, \hat{Y}|\, R=\mathrm{a},\mathrm{Y}^*\,\right] - \mathrm{E}\left[\, \hat{Y}|\, R=\mathrm{b},\mathrm{Y}^*\,\right]\right] \quad (6)$$

Δ^* represents *algorithmic discrimination* (AD) (Maxwell and Tomlinson, 2020; Kleinberg et al. 2018a; Arnold et al, 2020) or *algorithmic statistical discrimination* (Jelveh and Luca, 2015). It is

[b] We later discuss how algorithmic bias may interact with such human discrimination.

[c] In the case of a binary treatment allocation, as long as conditional distribution of $\hat{Y}$ is symmetric and identical in two groups, (6) implies (5).

also sometimes noted as *algorithmic bias* (Friedman and Nissenbaum, 1996), although we did not find any clear delineation between the use of these terms.

Measurement Errors and Endogenous Selection

One can view the relationship between $\hat{Y}$, Y, and Y^* through the lenses of classical measurement errors:

$Y = Y^* + \varepsilon_Y$, and (7)

$\hat{Y} = Y + \varepsilon_{\hat{Y}}$,[d] (8)

where $E[\varepsilon_Y] = E[\varepsilon_{\hat{Y}}] = 0$. Here, in the first equation, ε_Y reflects the measurement error in Y representing Y^*. For example, Y^* maybe true health, while Y is the observed health expenditure (Obermeyer et al., 2019). In the second equation, $\varepsilon_{\hat{Y}}$ reflects the estimation error in predicting $\hat{Y}$ using an algorithm. For example, $\hat{Y}$ could be Y predicted as a function of X's. In that case, estimation error can arise both due to model misspecification and measurement errors in the X's.

Following (6), (7), and (8), algorithmic discrimination is then given as

$$\Delta^* = \mathrm{E}_{\mathrm{Y}^*}[\,\mathrm{Y}^* \mid R = \mathrm{a}, \mathrm{Y}^*\,] - \mathrm{E}_{\mathrm{Y}^*}[\,\mathrm{Y}^* \mid R = \mathrm{b}, \mathrm{Y}^*\,] +$$

$$\mathrm{E}_{\mathrm{Y}^*}[\varepsilon_Y \mid R = \mathrm{a}, \mathrm{Y}^*\,] - \mathrm{E}_{\mathrm{Y}^*}[\varepsilon_Y \mid R = \mathrm{b}, \mathrm{Y}^*\,] +$$

$$E_{Y^*}[\,\varepsilon_{\hat{Y}} \mid R = a, Y^*\,] - E_{Y^*}[\,\varepsilon_{\hat{Y}} \mid R = \mathrm{b}, Y^*\,] \qquad (9)$$

In Eq (9), the first difference drops out by construction. The second difference in (9) represents the difference in measurement error for the qualifying variable between the two groups. The third difference in (9) reflects the difference in estimation error between the two groups. Together, these last two components provide the primary source of algorithm-induced bias or discrimination even when fair treatment allocation decisions are based on these predicted scores.

The fundamental question is why these measurement errors would be different across the groups. Availability of outcomes data and their underlying data generating processes contribute to the first component of AD. Choice of statistical methods, estimation heuristics, and data

[d] In certain binary setting, e.g. in the criminal justice lietarature (Arnold et al. 2020), Y = D·Y*, where D is a treatment choice made by an agent, which are informed by algorithmic predictions. Therefore, in such settings, both measurement errors in (7) and (8) are present.

generating processes for the predictors contribute to the second component of AD. The literature on unfair regression and algorithmic bias has discussed many situations that may give rise to such differential measurement errors. However, in many instances, the framework of measurement errors may not have been mentioned. These include biases caused by missing data, faulty device measurements (Bent et al., 2020), historically biased human decisions (Lakkaraju et al., 2017; Rambhachan and Roth, 2019), and using unfair algorithmic objectives (Corbett-Davies and Goel S 2018). Although our model can also readily extend to any of these situations, we specifically focus on the general criteria of endogenous selection to be the generator of these differential measurement errors. Endogenous selection arises when the available/observed data are generated under different processes for two groups, and these processes are also correlated with Y^*. A classic example of such processes includes differential healthcare access barriers by race, which leads to a differential rate of engagement with the health care system. We illustrate this case in the next section.

A STYLIZED ILLUSTRATION OF THE RISE OF ALGORITHMIC DISCRIMINATION WHEN DATA ARE GENERATED UNDER DIFFERENTIAL ACCESS

Consider a scenario where the goal is to develop a predictive model that will predict the appropriateness of heart surgery for individuals (T) or predict the specific dose of a drug to be prescribed.[e] To develop the predictive model, we rely on the EHR of a cohort of patients and predict a severity score ($\hat{Y}$) derived from electrocardiograms measurements (Y) as a function of patient demographics and other diagnosed clinical conditions (X).[f] There are many ways the data generating processes for Y and X can be different for different groups of individuals with the same underlying severity of heart health. We will focus on one mechanism: differential access.

Let there be two groups of individuals ($R_i \in \{a, b\}$) who have differential levels of access to healthcare due to differences in insurance statuses, systemic barriers, and also other social determinants of health. We assume that individuals with more severe health states will more urgently seek care, but access barriers independently reduce the ability for patients to seek care and enter the healthcare system. A formal representation of these mechanisms using a Roy model (1951) is given in the Appendix. These relationships also imply that populations with

[e] Based on our discussion in last section, we refrain from making this a binary decision, e.g. there is a certain threshold for the true underlying severity of heart health (Y*), above which individuals should qualify for heart surgery. Our framework readily extends to such binary decision under conditions described in footnote b.

[f] We defer from using a Y that directly represent any human choices due to the well-establishes "selective labels" (Lakkaraju et al, 2017) and "bias -in/bias out" (Rambhachan and Roth, 2019) issues.

greater access barriers that appear in the EHR data are more severe on average, which in turn generates two independent sources of differential measurement errors across the groups in the observed data:

First Source: The endogenous selection across the two groups described above implies that the true severity of heart health may be different for two groups of individuals even with the same severity score from electrocardiograms. This logic is illustrated in stylized Figure 1, which shows the association between Y^* and Y across all individuals in the population, irrespective of the R status. In general, they are positively associated. However, since Y is not a perfect proxy from Y^*, any given value of Y could represent a range of true underlying health severity Y^*. This measurement error is representative of ε_Y, which again, unconditionally, does not vary by R. Now, under the discussed access barriers, one group has their electrocardiograms measured only when their true severity of the disease is higher.

Under this data generating process, *conditional on* observed data, i.e, among those who have their electrocardiogram measured, the same score would reflect different average true severity for the two groups (See Appendix for a derivation).

$$E(Y^* \mid Y, R = \text{a}) \neq E(Y^* \mid Y, R = \text{b}).^{g} \qquad (10)$$

Following the relationship between Y and Y^* from (1), and by Bayes rule,[h] it can be said that the expected measurement error ε_Y is different for the two groups:

$$E[\varepsilon_Y \mid R = \text{a}, Y^*] \neq E[\varepsilon_Y \mid R = \text{b}, Y^*] \neq 0 \qquad (11)$$

Without loss of generality, if group *a* is the one that is facing more stringent access, then $E(Y^* \mid Y, R{=}\text{a}) > E(Y^* \mid Y, R{=}\text{b})$. That is, true severity will be higher for group a, given the level of observed severity score, compared to group b. This implies[h]

$$E[\varepsilon_Y \mid R = \text{a}, Y^*] < 0 < E[\varepsilon_Y \mid R = \text{b}, Y^*].^{i} \qquad (12)$$

Second Source: Since some of the factors in X that we use for prediction may also be affected by the same endogenous selection as in Y, we would expect a similar form of measurement errors in them. For example, let X represent diagnosed conditions as in an Elixhauser index, Charlson comorbidity index, or the Hierarchical Condition Categories. These diagnoses are recorded conditional on individuals engaging with the health care system. For example, a binary

[g] Hereon, Y represents observed Y in the data.

[h] $E_Y[Y^* \mid Y, R] = E_{Y^*}[E_Y[Y^* \mid Y, Y^*, R]] = E_{Y^*}[E_Y(Y - \varepsilon_Y \mid Y, Y^*, R]] = Y - E_{Y^*}[\varepsilon_Y \mid Y^*, R]$

[i] Since $E[\varepsilon_Y]=0$ and $E[\varepsilon_Y \mid Y^*] = 0$

diagnosed diabetes status of one may reflect a more severe form for the disease among the group with stringent access barriers than those with fewer barriers. Let the measurement error in X be denoted as:

$$X = X^* + \varepsilon_X, \quad (13)$$

where X^* is the true severity of the condition, and ε_X is the measurement error, $E[\varepsilon_X] = 0$ and $E[\varepsilon_X|X^*] = 0$. This is a typical error-in-variables model. From the above discussions and mirroring Eqs (8) and (9), we can say

$$E[\varepsilon_X \mid R = a, X^*] \neq E[\varepsilon_X \mid R = b, X^*] \neq 0. \quad (14)$$

Consequently, this implies that the true severity of diabetes may be higher among individuals with diagnosed diabetes in group a (with access barriers) than in the group b :

$$E[\varepsilon_X \mid R_i = a, X^*] < 0 < E[\varepsilon_X \mid R_i = b, X^*]. \quad (15)$$

Deviation from typical measurement error formulations: In measurement error formulations, when the goal is to regress X on Y, it is well known that ε_Y, the classical measurement errors in Y (first source) do not bias the coefficient estimate on X, whereas ε_X, the error-in variable for X (second source) makes X endogenous and hence biases the coefficient estimate on X. However, because we do not use the information on R to build these algorithms, the deviation of $\hat{Y}$ from Y^* will be differential for the two groups, and these differences are affected by the measurement errors in both Y and X.

To see this intuitively, if one is able to run a regression like: $\hat{Y} = \beta_0 + \beta_1 R + \beta_2 Y^* + w$, algorithmic discrimination would be captured by the coefficient β_1. This, in spirit, is similar to the outcomes-based benchmark test proposed by Arnold et al. (2020),[j] where they rely on quasi-experimental data that ensures equivalent distribution of Y^* across R and therefore run this regression without observing Y^*, which they extrapolate back to the whole data. In the absence of quasi-experimental approach, one must rely on some external data where $Y^*$is available to implementthe above regression. If AD is present, the estimates of β_1 will deviate from zero because the implied regression here is: $Y^* = \beta_0/(1-\beta_2) + \beta_1 R/(1-\beta_2) + h(w, \varepsilon_Y, \varepsilon_X)$, and R is correlated with both ε_Y and ε_X through endogenous selection (based on (12) and (14)). A more structured derivation of this intuition is presented next.

[j] This test is distinct from marginal outcomes tests used to detect total discrimination, as in Hull (2021), which entails estimating and comparing race-specific marginal treatment effects of certain decisions.

The implication of Measurement Errors for Predictions: Let a true stylized relationship between the X^* and Y^* be represented in the form of a linear model:

$Y^* = \alpha_0 + \alpha_1 X^* + u$, where u represents the error with $E(u|X^*) = E(u|X^*, R) = 0$ (16)

Assume that $E(u \cdot \varepsilon_Y) = E(u \cdot \varepsilon_X) = 0$, that is u is independent of the measurement errors in Y and X. Since we neither observe Y^* or X^* in the data, introducing these measurement errors gives:

$(Y - \varepsilon_Y) = \alpha_0 + \alpha_1(X - \varepsilon_X) + u$

$\Rightarrow Y = \alpha_0 + \alpha_1 X + (u + \varepsilon_Y - \alpha_1 \varepsilon_X)$ (17)

We train our algorithm using Y and X, and even if we make sure that we have found an unbiased estimator, signifying $E((u - \varepsilon_Y - \alpha_1 \varepsilon_X)) = 0$, $\hat{Y} = \hat{E}[Y|X]$ will be a biased estimator of Y^*. This is because $\hat{Y} = \hat{\alpha}_0 + \hat{\alpha}_1 X$, and $E(\hat{\alpha}_1) \neq \alpha_1$ because of the correlation of X with the error term.

Specifically,

$E(\hat{\alpha}_1) = \alpha_1 + \gamma E[X \cdot (u + \varepsilon_Y - \alpha_1 \varepsilon_X)]$, where $\gamma > 0$, and

$E(\hat{\alpha}_0) = \alpha_0 - \gamma E(X^*) \cdot E[X(u + \varepsilon_Y - \alpha_1 \varepsilon_X)] = \alpha_0$, assuming X^* is normalized to have mean zero. γ is a positive constant depending on the variances of X and ε_X.

Typically, we will only worry about the bias in estimating α_1 due to the correlation between X and ε_X, as $E[X(u + \varepsilon_Y)] = 0$, by definition. However, when it comes to resource allocation based on predictions, disparate allocations because of AD would involve the measurement errors in Y, too.

$\Delta^* = E[\, E[\hat{Y} \mid R = a, Y^*] - E[\hat{Y} \mid R = b, Y^*] \,]$

$= E[\, E\hat{\alpha}_0 + \hat{\alpha}_1 X \mid R = a, Y^*] - E\,[\hat{\alpha}_0 + \hat{\alpha}_1 X \mid R = b, Y^*] \,]$

$= E[\, \alpha_0 + \alpha_1 X + \gamma E[X \cdot (u + \varepsilon_Y - \alpha_1 \varepsilon_X)| \, R = a, Y^*] \,]$

$- E\,[\alpha_0 + \alpha_1 X + \gamma E[X \cdot (u + \varepsilon_Y - \alpha_1 \varepsilon_X)| \mid R = b, Y^*] \,]$

$= E[\, E[Y^*| \, R = a, Y^*] - E[Y^* \mid R = b, Y^*] \,] +$

$E[\, \gamma E[X \cdot \varepsilon_Y| \, R = a, Y^*] - \gamma E[X \cdot \varepsilon_Y| \, R = b, Y^*] \,] +$

$$E[\gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{b}, Y^*] - \gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{a}, Y^*]]$$

(18)

Eq (18) corresponds to the model we started off in (9) and indicates that the AD arises due to the two sources of measurement errors. Specifically, even though we have assumed that the measurements of *X* or *Y* are independent of the errors of the other, i.e, $E[X \cdot \varepsilon_Y] = E[Y \cdot \varepsilon_X] = 0$, conditional on *R*, these relationships are no longer independent. For example,

$E[X \cdot \varepsilon_Y|R] = E[X|R] \cdot E[\varepsilon_Y|R] \neq 0$,

where the last inequality follows from (14).[k] Also, from (15), we can say that

$E[X \cdot \varepsilon_Y|R = \text{a}, Y^*] < 0 < E[X \cdot \varepsilon_Y|R = \text{b}, Y^*]$ (19)

The real challenge of AD is that it is difficult to put a sign on this bias. The second difference in (18) is different from zero and negative by construction, following the logic in (19). The third difference, driven by the difference measurement error in *Xs* between the two groups, is positive (see Appendix for a derivation). Consequently, AD can exacerbate existing disparities and, in very specific situations, can sometimes overcome them.

In sum, AD arises when there are measurement errors in either Y or X and endogenous selection for participation in observed data. Note that both of these phenomena are required to generate algorithmic bias.

A very specific extension of this challenge with AD is the use of data generated under the presence of human discrimination between the two groups (Rambachan and Roth, 2019). We had assumed that there was no human discrimination in our framework. However, for example, if individuals in the group *a* are historically discriminated against (conditional of any algorithmic predictions), AD can sometimes help overcome this human discrimination-induced disparity (also called arbitraging discrimination, Rambachan et al. 2020) or even exacerbate it depending on the magnitudes in (18).

MITIGATING ALGORITHMIC DISCRIMINATION - CHALLENGES AND SOLUTION

[k] $E[\varepsilon_Y \mid R] = \text{E}_{Y^*}[E[\varepsilon_Y \mid R, Y^*]] \neq 0$.

Our prior discussion on algorithmic discrimination through the lens of measurement errors can also help us understand when some of the proposed solutions to AD may or may not work. The suite of methods to deal with AD can be broadly denoted as observational algorithmic fairness methods that define specific formalizations of fairness objectives and then provide an overview of approaches to meet the specific forms of fairness. Reviews of such objectives can be found elsewhere (Mitchell et al. 2021). It is important to note that there is no consensus definition of universal fairness. Once one selects a specific fairness objective and the methods to achieve that objective, the method still only meets that specific goal rather than any overall form of fairness that invokes multiple fairness objectives. Therefore, analysts should first decide which general notion of fairness they wish to meet and how to operationalize that objective. It is often impossible to meet different forms of fairness simultaneously as they may inherently be in opposition to each other (Kleinberg et al, 2016; Pfohl et al, 2021). While measures of fairness may not be able to achieve the desired level of universal fairness, there is still value in considering fairness definitions and their resulting solutions. They can still play an important role by increasing the transparency of the implementation process by forcing explicit decisions concerning these tradeoffs and transparency of the task-specific scenario and fairness goals. Therefore, algorithm developers or implementers should closely examine the desired notions of fairness, its specific implementation, and the tradeoffs that best meet the given scenario's fairness objective. These decisions should consider the broader social context and be informed by stakeholders who will be affected by these decisions. In this work, our fairness objective was determined by (6).

Given any fairness objective criterion, approaches that achieve it fall into three categories based on the point in the learning process at which fairness is addressed: the preprocessing, model fitting, or post-processing phase (Zink and Rose, 2020). Preprocessing techniques attempt to fix biases in the data before the model is trained. These methods then transform or change the data to remove discrimination before a model is fit to that data (Kamiran and Calders, 2009; Zemel et al, 2013, Luong). Further details on these approaches can be found in D'Alessandro et al. (2017).

Similarly, post-processing techniques attempt to meet fairness objectives by modifying the resulting classifications from a given model to make an explicit tradeoff in outcomes between protected and non-protected groups (D'Alessandro et al, 2017). A common approach is to find a classification threshold using the resulting prediction function that optimizes the desired fairness objective (Dwork et al, 2012; Hardt et al, 2016; Kleinberg et al, 2018). Here, the approaches separate the assessment of fairness from the optimization of performance during model fitting. Therefore, the fairness objective can be applied to any predictive model to achieve the desired form of fairness (D'Alessandro et al, 2017). However, the tradeoff between model performance

and fairness is then not directly controlled. Also, post-processing approaches naturally imply that a different standard is used to allocate treatment across two groups even when using the same prediction model. Such differential allocation standard, in itself, can be viewed as unfair. Further details about post-processing approaches can be found in Lohia et al. (2018).

Here, we will focus mainly on the model-fitting approach using a specific fairness criterion.

Equalized Odds (EO) Criterion

We focus on a few objectives that are considered to be a better representative of fairness and highlight how certain implementation of these goals in algorithms may or may not fail to meet these goals under the different measurement errors described above. The main fairness objectives that we will focus on is the *Equalized Odds (EO)* criterion and its several sub-criteria. The EO criteria are usually invoked when Y^* is binary, and consists of balancing True Positive Ratios (TPRs) and False Positive Rations (FPRs) across groups. Subset criteria for Equalized Odds are also used. For example, the notion of Equal Opportunity only requires the TPRs to be similar across groups (Hardt et al, 2016). Similarly, the notion of Predictive Equality requires similarity of FPRs only (Verma & Rubin, 2018). However, following our framework, we will use a version of the EO criterion that can be applied to continuous Y^*. Specifically, this criterion mirrors the AD criterion set in (6) and demands that this should be less than certain thresholds for every level of Y^*.

$$E[\hat{Y} \mid R = a, Y^*] - E[\hat{Y} \mid R = b, Y^*] \leq \varepsilon \qquad (20)$$

An accurate classifier, which does not generate AD, should meet this condition. Typically, these criteria are implemented in a prediction model through a constrained optimization approach or a penalized regression (Calders et al. 2013; Zink and Rose 2020). For example, a penalized regression likelihood function can be written as:

$$\text{Max}_{\alpha}\, F(Y^* \mid X, \alpha) + \pi \cdot \left(E_{\in R=a}\left[Y^* - \hat{Y}\right] - E_{\in R=b}\left[Y^* - \hat{Y}\right]\right)^2, \qquad (21)$$

where π is a hyperparameter π that can be user-specified or chosen via cross-validation.

If there is no measurement error in the outcome, then such a criterion helps to eliminate the AD directly. The real challenge lies in implementing such a criterion to develop algorithms when Y^* is not directly observed in the data. Consequently, implementation of EO is based on Y rather than Y^* (Calders et al., 2013):

$$E[\hat{Y} \mid R = a, Y] - E[\hat{Y} \mid R = b, Y] \leq \varepsilon \qquad (22)$$

Enforcing (22) in algorithms will not necessarily eliminate AD, since AD will still persist in the form:

$$\Delta^* = E[\, E[\hat{Y} \mid R = a, \varepsilon_Y]] - E[\hat{Y} \mid R = b, \varepsilon_Y]] \,], \quad (23)$$

which may or may not exacerbate the AD problem (see Appendix for derivations). Therefore, although the measurement error literature on endogeneity has mostly focused on paying close attention to measurement errors in the independent variables, to address AD, one must also pay close attention to measurement errors in the outcomes.

DISCUSSIONS

The implementation of predictive algorithms has the potential to transform health care. However, methodological and broader fairness concerns point towards the potential for these algorithms to directly or indirectly perpetuate or introduce health disparities. Additional difficulties arise with the complexities of interactions between membership towards protected groups. Problems can arise due to historical inequities in healthcare but also through healthcare's intersection with broader discrimination against protected classes. For instance, structural racism in domains such as housing, education, employment, and criminal justice create complicated relationships that will impact health scenarios where developers attempt to achieve fairness with respect to race (Pfohl et al, 2021; Bailey et al, 2017; Bailey et al, 2018). Additionally, classification by race, in particular, entangles historical and ongoing structural racism while also reinforcing the idea of race as a valid way to categorize people rather than a socially constructed classification (Hanna et al, 2020; Pfohl et al, 2021).

This paper provides a brief overview of the range of health economics studies vulnerable to AD. We then offer an overarching statistical framework of AD through the lens of measurement errors, which is familiar to the health economics audience. Specifically, we show that AD only exists when measurement errors exist in either the outcome or the predictors and when there is endogenous selection for participation in the observed data. The absence of any of these phenomena would eliminate algorithmic bias. We then discuss under what conditions some of the bias-mitigating strategies proposed in the computer sciences literature may or may not work. Specifically, we focus on the equalized odds constraints and show that such constraints may worsen the problem in the presence of measurement error in the dependent variable.

The general recommendation to applied researchers is to pay special attention to selecting the depending variable for their algorithms. Measurement errors in the dependent variable, in the presence of endogenous selection, will give rise to AD that is not easily addressable using

model fitting approaches. Another recommendation is to explore penalized approaches to explicitly enforce fairness criteria in their regressions.

Difficulties over complex empirical problems and difficulties in fairness objectives do not mean that algorithms cannot add value to health outcomes research and clinical decision-making. However, their use must be informed by a broader view of the social contexts surrounding protected classes' membership. Consideration of fairness notions can increase transparency and force developers to explicitly and transparently consider context-specific fairness and its tradeoffs. Alternative fairness definitions can also be implemented to monitor progress towards fairness. Rigorous focus, though, should be given to understanding the scenario-specific importance of different forms of fairness, the data generating process, the causal structure of involved data features, and the inherent tradeoffs between important measures of fairness when training and applying models. Otherwise, discrimination may be an unintended consequence of implementing care based on even generally accurate predictive models. These difficulties also require engagement from the stakeholders that will be affected by these decisions as well as constant and iterative monitoring. Only with these considerations can we hope that all patients can benefit from these technologies.

APPENDIX

A Roy's Model representing endogenous selection

A standard binary choice threshold crossing model for seeking care (*D*) is written as

$$D = \mathbf{1}[D^* > 0],$$

where **1**[.] is an indicator. A typical random utility model representing *D** is

$$D^* = g(Y^*, C) - V.$$

Here, *C* represents the cost of accessing health care. $V \perp \varepsilon_Y$ (V is independent of ε_Y). The propensity score or choice probability is

$$P(y^*, c) = \Pr(D = 1 | Y^* = y^*, C = c) = \Pr(g(y^*, c) > V) = F_V(g(y^*, c))$$

Where F_V is the distribution of V which is assumed to be continuous. Without loss of generality, it is asserted that $\partial P(Y^*, C)/\partial Y^* > 0$ (increase in severity of health condition increases the likelihood to seek care), and $\partial P(Y^*, C)/\partial C < 0$ (increases in the costs of accessing health care decreases the liklehood of seeking care).

Let the cost of accessing health care be greater for group *R* = a compared to group *R* = b. This implies that $C_a > C_b \Rightarrow \Pr(D|Y^*, R = a) < \Pr(D|Y^*, R = b)$. Therefore, even when the marginal distribution of Y^* is the same across the two groups, (i.e, $\Pr(Y^*|R = a) = \Pr(Y^*|R = b)$, by Bayes Rule, $E(Y^*|D, R = a) > \mathrm{E}(Y^*|D, R = b)$. This inequality can also be extended conditional on every value of *Y* among those who seek care. Ie.e

$$E(Y^*|D, Y\ R = a) > \mathrm{E}(Y^*|D, Y, R = b).$$

Let observed data on measured outcomes, e.g, electrocardiogram scores, be denoted as

$$Y_{OBS} = D \cdot Y$$

Therefore,

$$E(Y^*|Y_{OBS} = D \cdot Y, R = a) > \mathrm{E}(Y^*|Y_{OBS} = D \cdot Y, R = b).$$

Derivations for Eq (16)

$E[X \cdot \varepsilon_X \mid R, Y^*]$

= $E_{X^*}[E[(X^* + \varepsilon_X) \cdot \varepsilon_X \mid R, Y^*, X^*]]$

= $E_{X^*}[X^* \cdot E[\varepsilon_X \mid R, Y^*, X^*] + E[\varepsilon_X^2 \mid R, Y^*, X^*]]$

Replacing this in the last term of (16):

$E[\gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{b}, Y^*] - \gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{a}, Y^*]]$

= $E[\ E_{X^*}[X^* \cdot (E[\varepsilon_X \mid R\text{=b}, Y^*, X^*] - E[\varepsilon_X \mid R\text{=a}, Y^*, X^*])$

$+ (E[\varepsilon_X^2 \mid R\text{=b}, Y^*, X^*]] - E[\varepsilon_X^2 \mid R\text{=a}, Y^*, X^*])]] > 0,$

where the last inequality arises from the assumption that ε_X is homoscedastic (which implies that the last difference in the above equation drops out), and the established relationship from (12).

Derivations for Eq (20)

$E[\hat{Y} \mid R, Y^*] = E[\hat{Y} \mid R, Y, \varepsilon_Y] = h_R(Y, \varepsilon_Y)$, which is a function of Y and ε_Y. Assuming this is a linear function,

$h_R(Y, \varepsilon_Y) = g_{1R}(Y) + g_{2R}(\varepsilon_Y) = E[\hat{Y} \mid R, Y] + E[\hat{Y} \mid R, \varepsilon_Y]$

Therefore,

$\Delta^* = E[\ E[\hat{Y} \mid R = \text{a}, Y^*] - E[\hat{Y} \mid R = \text{b}, Y^*]\ ]$

~ $E[\ E[\hat{Y} \mid R = \text{a}, Y] - E[\hat{Y} \mid R = \text{b}, Y]\ ] + E[\ E[\hat{Y} \mid R = \text{a}, \varepsilon_Y]] - E[\hat{Y} \mid R = \text{b}, \varepsilon_Y]]\ ]$

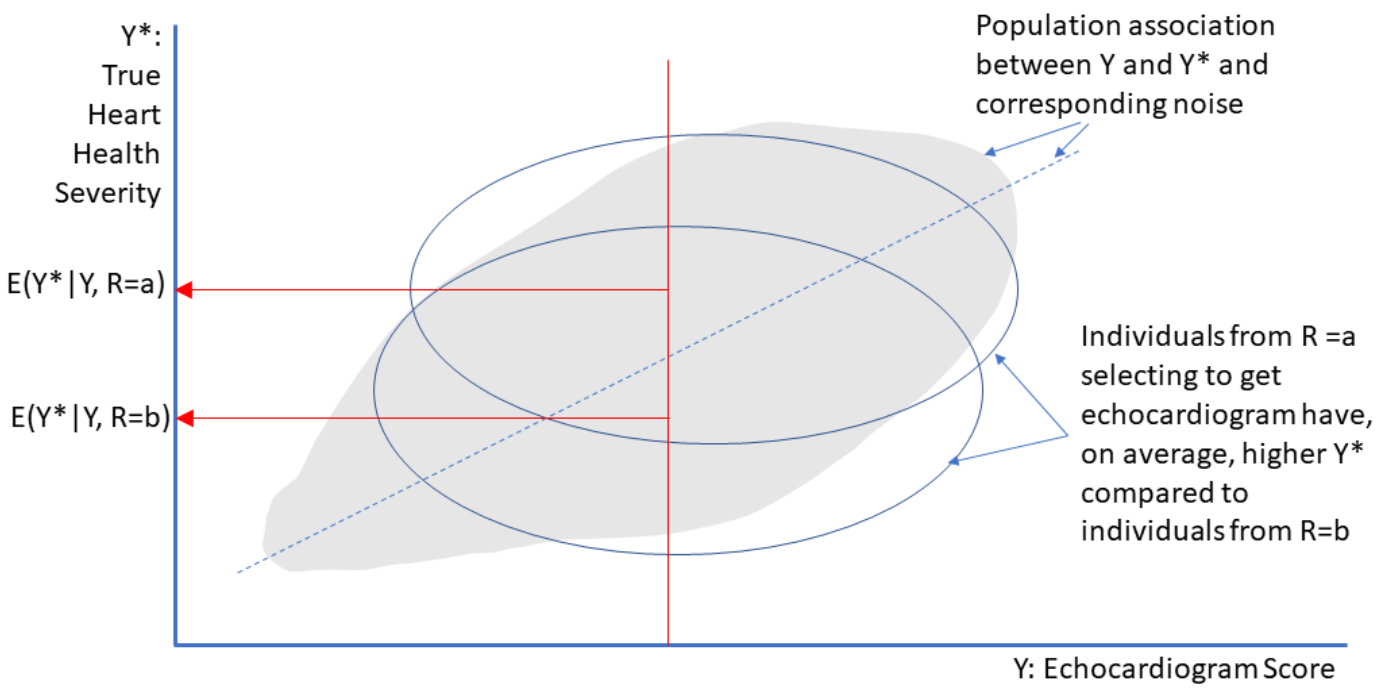
Even if we enforce the first difference to be zero through the equalized odds assumption, AD. persists as:

$\Delta^* = E[\ E[\hat{Y} \mid R = \text{a}, \varepsilon_Y]] - E[\hat{Y} \mid R = \text{b}, \varepsilon_Y]]\ ]$

= $E[\ E[Y^* \mid R = \text{a}, \varepsilon_Y] - E[Y^* \mid R = \text{b}, \varepsilon_Y]\ ] +$

$E[\gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{b}, \varepsilon_Y] - \gamma\alpha_1 E[X \cdot \varepsilon_X \mid R = \text{a}, \varepsilon_Y]]$

Figure 1: Measurement error due to differential access barriers



References

Alsan M, Owen G, Graziani G. Does Diversity Matter for Health? Experimental Evidence from Oakland. American Economic Review 2019; 109 (12): 4071-4111.

Alsan M, Stanford FC, Banerjee A, Breza E, Chandrasekhar AG, Eichmeyer S, Goldsmith-Pinkham P, Ogbu-Nwobodo L, Olken BA, Torres C, Sankar A, Vautrey PL, Duflo E. Comparison of Knowledge and Information-Seeking Behavior After General COVID-19 Public Health Messages and Messages Tailored for Black and Latinx Communities : A Randomized Controlled Trial. Annals of Internal Medicine 2021;174(4):484-492.

Angwin J, Larson J, Mattu S, Kirchner L. Machine Bias. Propublica 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, accessed September 30, 2021.

Arnold D, Dobbie WS, Hull P. Measuring Racial Discrimination in Algorithms. AEA Papers and Proceedings 2021, 111: 49–54

Ayanian JZ. The Costs of Racial Disparities in Health Care. Harvard Business Review 2015; https://hbr.org/2015/10/the-costs-of-racial-disparities-in-health-care, accessed September 30, 2021.

Balsa AI, McGuire TG. Statistical discrimination in health care. Journal of Health Economics 2001; 20(6): 881-907.

Bent, B., Goldstein, B.A., Kibbe, W.A. et al. Investigating sources of inaccuracy in wearable optical heart rate sensors. npj Digit. Med. 2020, 3(18). https://doi.org/10.1038/s41746-020-0226-6.

Bergquist T, Yan Y, Schaffter T, Yu T, Pejaver V, Hammarlund N, Prosser J, Guinney J, Mooney S. Piloting a model-to-data approach to enable predictive analytics in health care through patient mortality prediction. Journal of the American Medical Informatics Association 2020; 27(9): 1393–1400.

Calders T, Karim A, Kamiran F, Ali W, Zhang X. Controlling Attribute Effect in Linear Regression. 2013 IEEE 13th International Conference on Data Mining 2013; pp. 71-80, https://doi.org/10.1109/ICDM.2013.114.

Chandrasekaran R, Katthula V, Moustakas E. Patterns of Use and Key Predictors for the Use of Wearable Health Care Devices by US Adults: Insights from a National Survey. Journal of Medical Internet Research 2020; 22(10):e22443.

Char DS, Shah NH, Magnus D. Implementing Machine Learning in Health Care — Addressing Ethical Challenges. New England Journal of Medicine 2018; 378(11), 981–983.

Chouldechova A. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. Big Data 2017; 5(2):153–163.

Chouldechova A, Roth A. The Frontiers of Fairness in Machine Learning. ArXiv 2018. 1810.08810.pdf (arxiv.org), accessed September 30, 2021.

Colvonen PJ, DeYoung PN, Bosompra NA, Owens RL. Limiting racial disparities and bias for wearable devices in health science research. Sleep 2020; 13:43(10):zsaa159.

Corbett-Davies S, Goel S. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. ArXiv 2018. arXiv:1808.00023, accessed September 30, 2021.

d'Alessandro B, O'Neil C, LaGatta C. Big Data 2017;120-134.

Dastin, J. Amazon scraps secret A.I. recruiting tool that showed bias against women. Reuters, 2018, October 18. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G, accessed September 30, 2021.

Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future Healthcare Journal 2019; 6(2):94–98.

Desai RJ, Wang SV, Vaduganathan M, Evers T, Schneeweiss S. Comparison of Machine Learning Methods with Traditional Models for Use of Administrative Claims with Electronic Medical Records to Predict Heart Failure Outcomes. JAMA Network Open 2020; 3(1), 1918-962.

Dwork C, Hardt M, Pitassi T, Reingold O, Zemel R. Fairness through awareness. ITCS 2012 - Innovations in Theoretical Computer Science Conference 2012; 214–226.

Freedman S, Lin H, Prince J. Information Technology and Patient Health: Analyzing Outcomes, Populations, and Mechanisms. American Journal of Health Economics 2018; 4(1), 51–79.

Friedman B, Nissenbaum H. Bias in computer systems. ACM Transactions on Information Systems1996; 14(3): 330–347.

Grother P, Ngan M, Hanaoka K. Face Recognition Vendor Test (FRVT) Part 3: Demographic Effects. Nistir 8280, December 2019. Face Recognition Vendor Test (FRVT), Part 3: Demographic Effects (nist.gov), accessed September 30, 2021.

Haider AH, Scott VK, Rehman KA, Velopulos C, Bentley JM, Cornwell EE, Al-Refaie W. Racial disparities in surgical care and outcomes in the United States: A comprehensive review of patient, provider, and systemic factors. Journal of the American College of Surgeons 2013; 482-492.e12.

Hammarlund, N. Racial Treatment Disparities after Machine Learning Surgical Risk-Adjustment. Health Services & Outcomes Research Methodology 2021; 21, pages248–286.

Hanna A, Denton E, Smart A, Smith-Loud J. Towards a critical race methodology in algorithmic fairness. FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency 2020;501–512.

Hardt M, Price E, Srebro N. Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems 2016; 3323–3331.

Institute of Medicine. Unequal Treatment: Confronting Racial and Ethnic Disparities in Health Care. Washington, DC: The National Academies Press. 2003.

Jelveh Z, Luca M. Algorithmic statistical discrimination in criminal risk prediction. Paper presented at the II Fairness, Accountability and Transparency in Machine Learning Conference. Lille, France (July 11), 2015.

Podesta J, Pritzker P, Moniz EJ, Holdren J, Zients J. Big data: Seizing opportunities and preserving values. Executive Office of the President, 2014. https://obamawhitehouse.archives.gov/sites/default/files/docs/big_data_privacy_report_may_1_2014.pdf, accessed September 30, 2021.

Johri P, sen Saxena V, Kumar, A. (2021). Rummage of Machine Learning Algorithms in Cancer Diagnosis. International Journal of E-Health and Medical Communications 2021; 12(1): 1-15.

Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 2012; 33(1):1–33.

Kleinberg J, Mullainathan S, Raghavan M. (2016). Inherent Tradeoffs in the Fair Determination of Risk Scores. ArXiv 2016. https://arxiv.org/pdf/1609.05807.pdf, accessed September 30, 2021.

Kleinberg J, Ludwig J, Mullainathan S, Sunstein CR. Discrimination in the Age of Algorithms. Journal of Legal Analysis, 2018a; 10:113–174.

Kleinberg, J, Lakkaraju, H, Leskovec, J, Ludwig, J, & Mullainathan, S. Human Decisions and Machine Predictions. Quarterly Journal of Economics 2018b; 133(1):237–293.

Krittanawong, C, Virk, H. U. H, Bangalore, S, Wang, Z, Johnson, K. W, Pinotti, R, Zhang H, Kaplin S, Narasimhan B, Kitai T, Baber U, Halperin JL, Tang, WHW. (2020). Machine learning prediction in cardiovascular diseases: a meta-analysis. Scientific Reports 2020; 10(1):16057.

Lai H, Huang H, Keshavjee K, Guergachi A, Gao X. Predictive models for diabetes mellitus using machine learning techniques. BMC Endocrine Disorders 2019; 19(1):101.

Lakkaraju H, Kleinberg J, Leskovec J, Ludwig J, Mullainathan S. The Selective Labels Problem: Evaluating Algorithmic Predictions in the Presence of Unobservables. KDD. 2017:275-284.

Lloyd-Jones DM. Cardiovascular risk prediction: Basic concepts, current status, and future directions. Circulation 2010; 121: 1768–1777.

Lohia PK, Ramamurthy KN, Bhide M, Saha D, Varshney KR, Puri R. Bias Mitigation Post-processing for Individual and Group Fairness, ArXiv 2018. https://arxiv.org/abs/1812.06135, accessed September 30, 2021.

Maxwell J, Tomlinson J. Proving algorithmic discrimination in government decision-making. Oxford University Commonwealth Law Journal 2020; 20(2): 352-360.

Mitchell S, Potash E, Barocas S, D'Amour A, Lum K. Algorithmic Fairness: Choices, Assumptions, and Definitions. Annual Review of Statistics and Its Application 2021; 8(1):141-163.

Nabi R, Shpitser I. Fair Inference on Outcomes. Proceedings of the AAAI Conference on Artificial Intelligence 2018; 1931-1940.

National Academies of Sciences, Engineering, and Medicine. The Private Sector as a Catalyst for Health Equity and a Vibrant Economy: Proceedings of a Workshop. Washington, DC: The National Academies Press. 2016.

National Academies of Sciences, Engineering, and Medicine. Communities in Action: Pathways to Health Equity. Washington, DC: The National Academies Press. 2017.

National Healthcare Quality and Disparities Report. Content last reviewed June 2021. Agency for Healthcare Research and Quality, Rockville, MD. 2019. https://www.ahrq.gov/research/findings/nhqrdr/nhqdr19/index.html, accessed September 30, 2021.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S. Dissecting racial bias in an algorithm used to manage the health of populations. Science 2019; 366(6464): 447–453.

Parikh RB, Teeple S, Navathe AS. Addressing Bias in Artificial Intelligence in Health Care. JAMA - Journal of the American Medical Association 2019; 322: 2377–2378.

Park S, Basu A. Alternative evaluation metrics for risk adjustment methods. Health Economics 2018;27(6):984-1010.

Pfohl SR, Foryciarz A, Shah NH. An empirical characterization of fair machine learning for clinical risk prediction. Journal of Biomedical Informatics 2021; 113:103621.

Rajkomar A, Hardt M, Howell MD, Corrado G, Chin MH. Ensuring fairness in machine learning to advance health equity. Annals of Internal Medicine 2018: 169(12): 866–872.

Rambachan A, Kleinberg J, Mullainathan S, Ludwig J. An economic approach to regulating algorithms. National Bureau of Economic Research w27111, 2020. https://www.nber.org/papers/w27111, accessed September 30, 2021.

Rambhachan A, Roth J. Bias in, bias out? Evaluating the folk wisdom. arXiv 2019; arXiv:1909.08518

Rose S. A Machine Learning Framework for Plan Payment Risk Adjustment. Health Services Research 2016; 51(6):2358-2374.

Rose S, Bergquist SL, Layton TJ. Computational health economics for identification of unprofitable health care enrollees. Biostatistics 2017; 18:682–94.

Ross EG, Shah NH, Dalman RL, Nead KT, Cooke JP, Leeper NJ. (2016). The use of machine learning for the identification of peripheral artery disease and future mortality risk. Journal of Vascular Surgery 2016; 64(5): 1515-1522.e3.

Roy A. Some Thoughts on the Distribution of Earnings. Oxford Economic Papers 1951; 3 (2): 135–146.

Rueda J.-D, Cristancho RA, Slejko JF. Is Artificial Intelligence the Next Big Thing in Health Economics and Outcomes Research? ISPOR Value in Health Spotlight. 2019. https://www.ispor.org/docs/default-source/publications/value-outcomes-spotlight/march-april-2019/vos-heor-articles---rueda.pdf?sfvrsn=18cb16f5_0, accessed September 30, 2021.

Silva MA, Patel J, Kavouridis V, Gallerani T, Beers A, Chang K, Hoebel KV, Brown J, See AP, Gormley WB, Aziz-Sultan MA, Kalpathy-Cramer J, Arnaout O, Patel NJ. Machine Learning Models Can Detect Aneurysm Rupture and Identify Clinical Features Associated with Rupture. World Neurosurgery 2019; 131: e46–e51.

Spann A, Yasodhara A, Kang J, Watt K, Wang B, Goldenberg A, Bhat M. Applying Machine Learning in Liver Disease and Transplantation: A Comprehensive Review. Hepatology 2020; 71: 1093–1105.

Verma S, Rubin J. Fairness Definitions Explained. IEEE/ACM International Workshop on Software Fairness 2018:18.

Yang C, Delcher C, Shenkman E, Ranka S. Machine learning approaches for predicting high cost high need patient expenditures in health care. Biomedical Engineering OnLine. 2018 Nov 20;17(Suppl 1):131.

Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, Xia M, Liu M, Zhou X, Wu Q, Guo Y, Zhu C, Li Y, Culver DS, Alfreds ST, Stearns F, Sylvester KG, Widen E, McElhinney D, Ling X. Prediction of incident hypertension within the next year: Prospective study using statewide electronic health records and machine learning. Journal of Medical Internet Research 2018; 20(1): e22.

Zemel R, Wu Y, Swersky K, Pitassi T, Dwork C. Learning fair representations. 30th International Conference on Machine Learning, ICML 2013; PART 2:1362–1370.

Zhang Y, Wang S, Hermann A, Joly R, Pathak J. Development and validation of a machine learning algorithm for predicting the risk of postpartum depression among pregnant women. Journal of Affective Disorders 2021; 279, 1–8.

Zink A, Rose S. Fair regression for health care spending. Biometrics 2020; 76(3), 973–982.