Fair prediction with disparate impact:
A study of bias in recidivism prediction instruments

Alexandra Chouldechova *

Last revised: February 8, 2017

Abstract

Recidivism prediction instruments (RPI's) provide decision makers with an assessment of the likelihood that a criminal defendant will reoffend at a future point in time. While such instruments are gaining increasing popularity across the country, their use is attracting tremendous controversy. Much of the controversy concerns potential discriminatory bias in the risk assessments that are produced. This paper discusses several fairness criteria that have recently been applied to assess the fairness of recidivism prediction instruments. We demonstrate that the criteria cannot all be simultaneously satisfied when recidivism prevalence differs across groups. We then show how disparate impact can arise when a recidivism prediction instrument fails to satisfy the criterion of error rate balance.

Keywords: disparate impact; bias; recidivism prediction; risk assessment; fair machine learning

1 Introduction

Risk assessment instruments are gaining increasing popularity within the criminal justice system, with versions of such instruments being used or considered for use in pre-trial decision-making, parole decisions, and in some states even sentencing[1,2,3]. In each of these cases, a high-risk classification—particularly a high-risk misclassification—may have a direct adverse impact on a criminal defendant's outcome. If the use of RPI's is to become commonplace, it is especially important to ensure that the instruments are free from discriminatory biases that could result in unethical practices and inequitable outcomes for different groups.

In a recent widely popularized investigation conducted by a team at ProPublica, Angwin et al.[4] studied an RPI called COMPAS[a], concluding that it is biased against black defendants. The authors

*Heinz College, Carnegie Mellon University

[a]COMPAS[5] is a risk assessment instrument developed by Northpointe Inc.. Of the 22 scales that COMPAS provides, the Recidivism risk and Violent Recidivism risk scales are the most widely used. The empirical results in this paper are based on decile scores coming from the COMPAS Recidivism risk scale.

found that the likelihood of a non-recidivating black defendant being assessed as high risk is nearly twice that of white defendants. Similarly, the likelihood of a recidivating black defendant being assessed as low risk is nearly half that of white defendants. In technical terms, these findings indicate that the COMPAS instrument has considerably higher false positive rates and lower false negative rates for black defendants than for white defendants.

ProPublica's analysis has met with much criticism from both the academic community and from the Northpointe corporation. Much of the criticism has focussed on the particular choice of fairness criteria selected for the investigation. Flores et al.[6] argue that the correct approach for assessing RPI bias is instead to check for *calibration*, a fairness criterion that they show COMPAS satisfies. Northpointe in their response[7] argue for a still different approach that checks for a fairness criterion termed *predictive parity*, which they demonstrate COMPAS also satisfies. We provide precise definitions and a more in-depth discussion of these and other fairness criteria in Section 2.1.

In this paper we show that the differences in false positive and false negative rates cited as evidence of racial bias by Angwin et al.[4] are a direct consequence of applying an RPI that that satisfies predictive parity to a population in which recidivism prevalence[a] differs across groups. Our main contribution is twofold. (1) First, we make precise the connection between the predictive parity criterion and error rates in classification. (2) Next, we demonstrate how using an RPI that has different false postive and false negative rates between groups can lead to disparate impact when individuals assessed as high risk receive stricter penalties. Throughout our discussion we use the term *disparate impact* to refer to settings where a penalty policy has unintended disproportionate adverse impact on a particular group.

It is important to bear in mind that fairness itself—along with the notion of disparate impact—is a social and ethical concept, not a statistical one. A risk prediction instrument that is fair with respect to particular fairness criteria may nevertheless result in disparate impact depending on how and where it is used. In this paper we consider hypothetical use cases in which we are able to directly connect particular fairness properties of an RPI to a measure of disparate impact. We present both theoretical and empirical results to illustrate how disparate impact can arise.

1.1 Outline of paper

We begin in Section 2 by providing some background on several of the different fairness criteria that have appeared in recent literature. We then proceed to demonstrate that an instrument that satisfies predictive parity cannot have equal false positive and negative rates across groups when the recidivism prevalence differs across those groups. In Section 3 we analyse a simple risk assessment-based sentencing policy and show how differences in false positive and false negative rates can result in disparate impact under this policy. In Section 3.3 we back up our theoretical analysis by presenting some empirical results based on the data made available by the ProPublica investigators. We conclude with a discussion of the issues that biased data presents for the arguments put forth in this paper.

[a] *Prevalence*, also termed the *base rate*, is the proportion of individuals who recidivate in a given population.

1.2 Data description and setup

The empirical results in this paper are based on the Broward County data made publicly available by ProPublica[8]. This data set contains COMPAS recidivism risk decile scores, 2-year recidivism outcomes, and a number of demographic and crime-related variables on individuals who were scored in 2013 and 2014. We restrict our attention to the subset of defendants whose race is recorded as African-American (b) or Caucasian (w).[a] After applying the same data pre-processing and filtering as reported in the ProPublica analysis, we are left with a data set on $n = 6150$ individuals, of whom $n_b = 3696$ are African-American and $n_c = 2454$ are Caucasian.

2 Assessing fairness

2.1 Background

We begin by with some notation. Let $S = S(x)$ denote the risk score based on covariates $X = x \in \mathbb{R}^p$, with higher values of S corresponding to higher levels of assessed risk. We will interchangeably refer to S as a *score* or an *instrument*. For simplicity, our discussion of fairness criteria will focus on a setting where there exist just two groups. We let $R \in \{b, w\}$ denote the group to which an individual belongs, and do not preclude R from being one of the elements of X. We denote the outcome indicator by $Y \in \{0, 1\}$, with $Y = 1$ indicating that the given individual goes on to recidivate. Lastly, we introduce the quantity s_{HR}, which denotes the high-risk score threshold. Defendants whose score S exceeds s_{HR} will be referred to as *high-risk*, while the remaining defendants will be referred to as *low-risk*.

With this notation in hand, we now proceed to define and discuss several fairness criteria that commonly appear in the literature, beginning with those mentioned in the introduction. We indicate cases where a given criterion is known to us to also commonly appear under some other name. All of the criteria presented below can also be assessed *conditionally* by further conditioning on some covariates in X. We discuss this point in greater detail in Section 3.1.

Definition 1 (Calibration)**.** A score $S = S(x)$ is said to be *well-calibrated* if it reflects the same likelihood of recidivism irrespective of the individuals' group membership. That is, if for all values of s,

$$\mathbb{P}(Y = 1 \mid S = s, R = b) = \mathbb{P}(Y = 1 \mid S = s, R = w). \tag{2.1}$$

Within the educational and psychological testing and assessment literature, the notion of *calibration* features among the widely accepted and adopted standards for empirical fairness assessment. In this literature, an instrument that is *well-calibrated* is referred to as being *free from predictive bias*. This criterion has recently been applied to the PCRA[b] instrument, with initial findings suggesting that calibration is satisfied with respect race[10,11], but not with respect to gender[12]. In

[a] There are 6 racial groups represented in the data. 85% of individuals are either African-American or Caucasian.

[b] The Post Conviction Risk Assessment (PCRA) tool was developed by the Administrative Office of the United States Courts for the purpose of improving "the effectiveness and efficiency of post-conviction supervision" [9]

their response to the ProPublica investigation, Flores et al.[6] verify that COMPAS is well-calibrated using logistic regression modeling.

Definition 2 (Predictive parity)**.** A score $S = S(x)$ satisfies *predictive parity* at a threshold s_{HR} if the likelihood of recidivism among high-risk offenders is the same regardless of group membership. That is, if,

$$\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = b) = \mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = w). \tag{2.2}$$

Predictive parity at a given threshold s_{HR} amounts to requiring that the *positive predictive value* (PPV) of the classifier $\hat{Y} = \mathbb{1}_{S > s_{\text{HR}}}$ be the same across groups. While predictive parity and calibration look like very similar criteria, well-calibrated scores can fail to satisfy predictive parity at a given threshold. This is because the relationship between (2.2) and (2.1) depends on the conditional distribution of $S \mid R = r$, which can differ across groups in ways that result in PPV imbalance. In the simple case where S itself is binary, a score that is well-calibrated will also satisfy predictive parity. Northpointe's refutation[7] of the ProPublica analysis shows that COMPAS satisfies predictive parity for threshold choices of interest.

Definition 3 (Error rate balance)**.** A score $S = S(x)$ satisfies *error rate balance* at a threshold s_{HR} if the false positive and false negative error rates are equal across groups. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = b) = \mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = w)\,, \quad \text{and} \tag{2.3}$$
$$\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = b) = \mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = w), \tag{2.4}$$

where the expressions in the first line are the group-specific false positive rates, and those in the second line are the group-specific false negative rates.

ProPublica's analysis considered a threshold of $s_{\text{HR}} = 4$, which they showed leads to considerable imbalance in both false positive and false negative rates. While this choice of cutoff met with some criticism, we will see later in this section that error rate imbalance persists—indeed, must persist—for any choice of cutoff at which the score satisfies the predictive parity criterion. Error rate balance is also closely connected to the notions of *equalized odds* and *equal opportunity* as introduced in the recent work of Hardt et al.[13].

Definition 4 (Statistical parity)**.** A score $S = S(x)$ satisfies *statistical parity* at a threshold s_{HR} if the proportion of individuals classified as high-risk is the same for each group. That is, if,

$$\mathbb{P}(S > s_{\text{HR}} \mid R = b) = \mathbb{P}(S > s_{\text{HR}} \mid R = w) \tag{2.5}$$

Statistical parity also goes by the name of *equal acceptance rates*[14] or *group fairness*[15], though it should be noted that these terms are in many cases not used synonymously. While our discussion focusses primarily on first three fairness criteria, statistical parity is widely used within the machine learning community and may be the criterion with which many readers are most familiar[16,17]. Statistical parity is well-suited to contexts such as employment or admissions, where it may be desirable or required by law or regulation to employ or admit individuals in equal proportion across racial, gender, or geographical groups. It is, however, a difficult criterion to motivate in the recidivism prediction setting, and thus will not be further considered in this work.

2.2 Further related work

Though the study of discrimination in decision making and predictive modeling is rapidly evolving, it also has a long and rich multidisciplinary history. Romei and Ruggieri [18] provide an excellent overview of some of the work in this broad subject area. The recent work of Barocas and Selbst [19] offers a broad examination of algorithmic fairness framed within the context of anti-discrimination laws governing employment practices. Hannah-Moffat [20], Skeem [21], and Monahan and Skeem [22] examine legal and ethical issues relating specifically to the use of risk assessment instruments in sentencing, citing the potential for race and gender discrimination as a major concern.

In work concurrent with our own, several other researchers have also investigated the compatibility of different notions of fairness. Kleinberg et al. [23] show that calibration cannot be satisfied simultaneously with the fairness criteria of *balance for the negative class* and *balance for the positive class*. Translated into the present context, the latter criteria require that the average score assigned to non-recidivists (the negative class) should be the same for both groups, and that the same should hold among recidivists (the positive class). The work of Corbett-Davies et al. [24] closely parallels the results that we present in Section 2.3, reaching the same conclusion regarding the incompatibility of predictive parity and error rate balance in the setting of unequal prevalence.

2.3 Predictive parity, false positive rates, and false negative rates

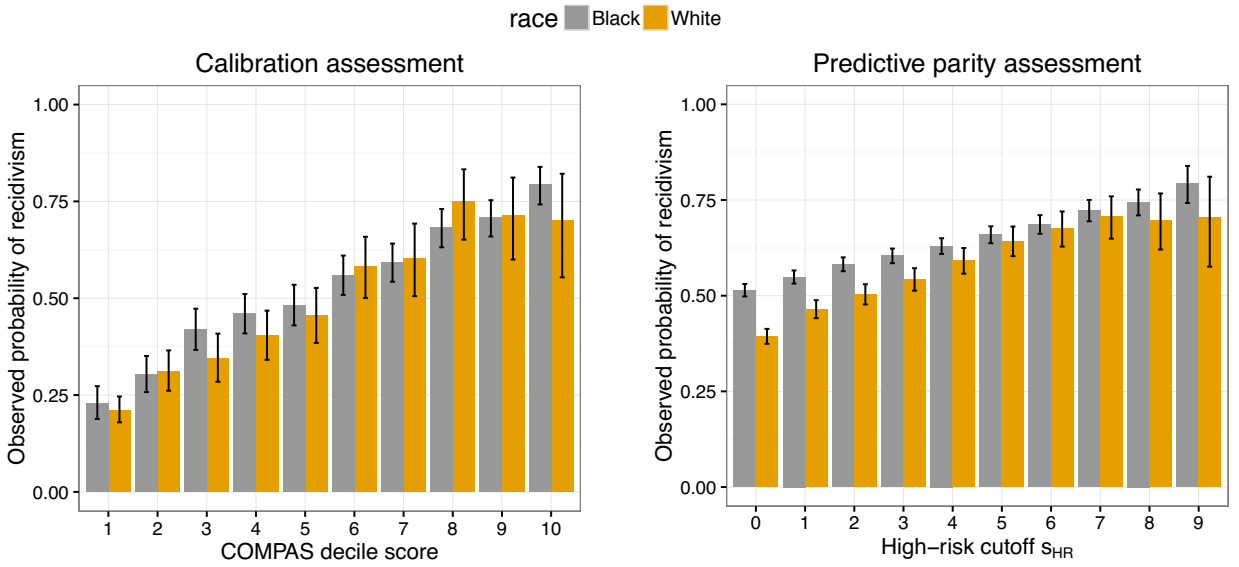
In this section we present our first main result, which establishes that predictive parity is incompatible with error rate balance when prevalence differs across groups. To better motivate the discussion, we begin by presenting an empirical fairness assessment of the COMPAS RPI. Figure 1 shows plots of the observed recidivism rates and error rates corresponding to the fairness notions of calibration, predictive parity, and error rate balance. We see that the COMPAS RPI is (approximately) well-calibrated, and also satisfies predictive parity provided that the high-risk cutoff s_{HR} is 4 or greater. However, COMPAS fails on both false positive and false negative error rate balance across the range of high-risk cutoffs.

Angwin et al. [4] focussed on a high-risk cutoff of $s_{\text{HR}} = 4$ for their analysis, which some critics have argued is too low, suggesting that $s_{\text{HR}} = 7$ is more suitable. As can be seen from Figures 1c and 1d, significant error rate imbalance persists at this cut-off as well. Moreover, the error rates achieved at so high a cutoff are at odds with evidence suggesting that the use of RPI's is of interest in settings where false negatives have a higher cost than false positives, with relative cost estimates ranging from 2.6 to upwards of 15. [25,26]

As we now proceed to show, the error rate imbalance exhibited by COMPAS is not a coincidence, nor can it be remedied in the present context. When the recidivism prevalence–i.e., the base rate $\mathbb{P}(Y = 1 \mid R = r)$—differs across groups, any instrument that satisfies predictive parity at a given threshold s_{HR} *must* have imbalanced false positive or false negative errors rates at that threshold. To understand why predictive parity and error rate balance are mutually exclusive in the setting of unequal recidivism prevalence, it is instructive to think of how these quantities are all related.

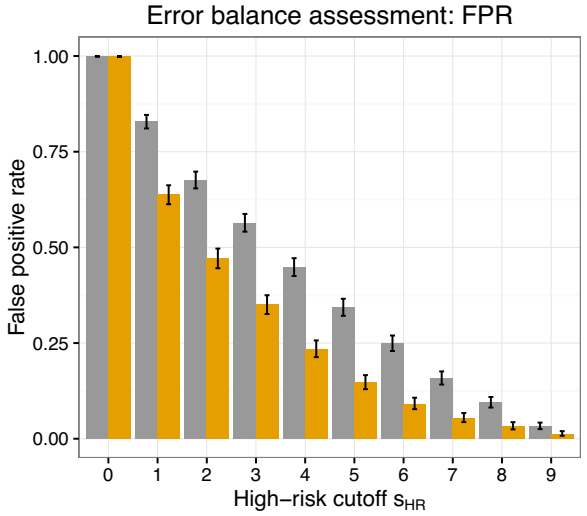
Given a particular choice of s_{HR}, we can summarize an instrument's performance in terms of a confusion matrix, as shown in Table 1 below.
All of the fairness metrics presented in Section 2.1 can be thought of as imposing constraints on

the values (or the distribution of values) in this table. Another constraint—one that we have no direct control over—is imposed by the recidivism prevalence within groups. It is not difficult to

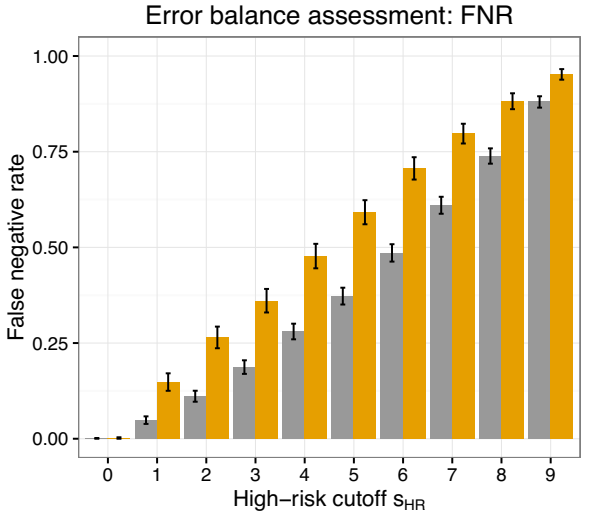


(a) Bars represent empirical estimates of the expressions in (2.1): $\mathbb{P}(Y = 1 \mid S = s, R = r)$ for decile scores $s \in \{1, \ldots, 10\}$.

(b) Bars represent empirical estimates of the expressions in (2.2): $\mathbb{P}(Y = 1 \mid S > s_{\text{HR}}, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$



(c) Bars represent observed false positive rates, which are empirical estimates of the expressions in (2.3): $\mathbb{P}(S > s_{\text{HR}} \mid Y = 0, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$



(d) Bars represent observed false negative rates, which are empirical estimates of the expressions in (2.4): $\mathbb{P}(S \leq s_{\text{HR}} \mid Y = 1, R = r)$ for values of the high-risk cutoff $s_{\text{HR}} \in \{0, \ldots, 9\}$

Figure 1: Empirical assessment of the COMPAS RPI according to three of the fairness criteria presented in Section 2.1. Error bars represent 95% confidence intervals. These Figures confirm that COMPAS is (approximately) well-calibrated, satisfies predictive parity for high-risk cutoff values of 4 or higher, but fails to have error rate balance.

	Low-Risk	High-Risk
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Table 1: T/F denote True/False and N/P denote Negative/Positive. For instance, FP is the number of false positives: individuals who are classified as high-risk but who do not reoffend.

show that the prevalence (p), positive predictive value (PPV), and false positive and negative error rates (FPR, FNR) are related via the equation

$$\text{FPR} = \frac{p}{1-p}\frac{1-\text{PPV}}{\text{PPV}}(1-\text{FNR}). \tag{2.6}$$

From this simple expression we can see that if an instrument satisfies predictive parity—that is, if the PPV is the same across groups—but the prevalence differs between groups, the instrument cannot achieve equal false positive and false negative rates across those groups.

This observation enables us to better understand why we observe such large discrepancies in FPR and FNR between black and white defendants in Figure 1. The recidivism rate among black defendants in the data is 51%, compared to 39% for White defendants. Thus at any threshold s_{HR} where the COMPAS RPI satisfies predictive parity, equation (2.6) tells us that some level of imbalance in the error rates must exist. Since not all of the fairness criteria can be satisfied at the same time, it becomes important to understand the potential impact of failing to satisfy particular criteria. This question is explored in the context of a hypothetical risk-based sentencing framework in the next section.

3 Assessing impact

In this section we show how differences in false positive and false negative rates can result in disparate impact under policies where a high-risk assessment results in a stricter penalty for the defendant. Such situations may arise when risk assessments are used to inform bail, parole, or sentencing decisions. In Pennsylvania and Virginia, for instance, statutes permit the use of RPI's in sentencing, provided that the sentence ultimately falls within accepted guidelines[1]. We use the term "penalty" somewhat loosely in this discussion to refer to outcomes both in the pre-trial and post-conviction phase of legal proceedings. For instance, even though pre-trial outcomes such as the amount at which bail is set are not punitive in a legal sense, we nevertheless refer to bail amount as a "penalty" for the purpose of our discussion.

There are notable cases where RPI's are used for the express purpose of informing risk reduction efforts. In such settings, individuals assessed as high risk receive what may be viewed as a benefit rather than a penalty. The PCRA score, for instance, is intended to support precisely this type of decision-making at the federal courts level[11]. Our analysis in this section specifically addresses use cases where high-risk individuals receive stricter penalties.

To begin, consider a setting in which guidelines indicate that a defendant is to receive a penalty

$t_{\min} \leq T \leq t_{\max}$. A very simple risk-based approach, which we will refer to as the MinMax[a] policy, would be to assign penalties as follows:

$$T_{\text{MinMax}}(s) = \begin{cases} t_{\min} & \text{if } s > s_{\text{HR}} \\ t_{\max} & \text{if } s < s_{\text{HR}} \end{cases}. \tag{3.1}$$

In this simple setting, we can precisely characterize the extent of disparate impact in terms of recognizable quantities. Our analysis will focus on the quantity

$$\Delta = \Delta(y_1, y_2) \equiv \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2),$$

which is the expected difference in sentence duration between defendants in different groups, with potentially different outcomes $y_1, y_2 \in \{0, 1\}$. Δ is taken to serve as our the measure of disparate impact.

Proposition 3.1. *The expected difference in penalty under the MinMax policy is given by*

$$\begin{aligned} \Delta &\equiv \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2) \\ &= (t_{\max} - t_{\min})\big(\mathbb{P}(S > s_{\text{HR}} \mid R = b, Y = y_1) - \mathbb{P}(S > s_{\text{HR}} \mid R = w, Y = y_2)\big) \end{aligned}$$

A proof can be found in Appendix A. We will discuss two immediate Corollaries of this result.

Corollary 3.1 (Non-Recidivists)**.** *Among individuals who* do not recidivate*, the difference in average penalty under the MinMax policy is*

$$\Delta = (t_{\max} - t_{\min})(\text{FPR}_b - \text{FPR}_w), \tag{3.2}$$

where FPR_r *denotes the false positive rate among individuals in group* $R = r$*.*

Corollary 3.2 (Recidivists)**.** *Among individuals who* recidivate*, the difference in average penalty under the MinMax policy is*

$$\Delta = (t_{\max} - t_{\min})(\text{FNR}_w - \text{FNR}_b), \tag{3.3}$$

where FNR_r *denotes the false negative rate among individuals in group* $R = r$*.*

When using an RPI that satisfies predictive parity in populations where recidivism prevalence differs across groups, it will generally be the case that the higher recidivism prevalence group will have a higher FPR and lower FNR. From equations (3.2) and (3.3), we can see that this would on average result in greater penalties for defendants in the higher prevalence group, both among recidivists and non-recidivists.

An interesting special case to consider is one where $t_{\min} = 0$. This could arise in sentencing decisions for offenders convicted of low-severity crimes who have good prior records. In such cases, so-called restorative sanctions may be imposed as an alternative to a period of incarceration. If

[a]The term MinMax as used throughout this paper has no intended connection the decision-theoretic notion of minimax decision rules. Min and Max in this context refer to the minimum and maximum allowable sentences as stipulated by sentencing guidelines.

we further take $t_{\max} = 1$, then $\mathbb{E}T = \mathbb{P}(T \neq 0)$, which can be interpreted as the probability that a defendant receives a sentence imposing some period of incarceration.

It is easy to see that in such settings a non-recidivist in group b is $\mathrm{FPR}_b/\mathrm{FPR}_w$ times more likely to be incarcerated compared to a non-recidivist in group w.[a] This naturally raises the question of whether overall differences in error rates are observed to persist across more granular subpopulations, such as the subset of individuals eligible for restorative sanctions. We explore this question in the section below.

3.1 Conditioning on other covariates

One might expect that differences in false positive rates are largely attributable to the subset of defendants who are charged with more serious offenses and who have a larger number of prior arrests/convictions. While it is true that the false positive rates within both racial groups are higher for defendants with worse criminal histories, considerable between-group differences in these error rates persist across low prior count subgroups. Figure 2 shows plots of false positive rates across different ranges of prior count for all defendants and also for the subset charged with a misdemeanor offense, which is the lowest severity criminal offense category. As one can see, differences in false positive rates between Black defendants and White defendants persist across prior record subgroups.

In general, all of the theoretical results presented in this section extend to the setting where we further condition on the covariates X. The main difference is that all classification metrics would need to be evaluated conditional on X. For instance, assuming that $t_{\min}$ and $t_{\max}$ are constant on a set $\mathcal{X}$, Corollary 3.1 would say that the difference in average penalty under the MinMax policy among non-recidivists for whom $X \in \mathcal{X}$ is given by

$$\Delta = (t_{\max} - t_{\min})\left(\mathrm{FPR}_b(\mathcal{X}) - \mathrm{FPR}_w(\mathcal{X})\right) \tag{3.4}$$

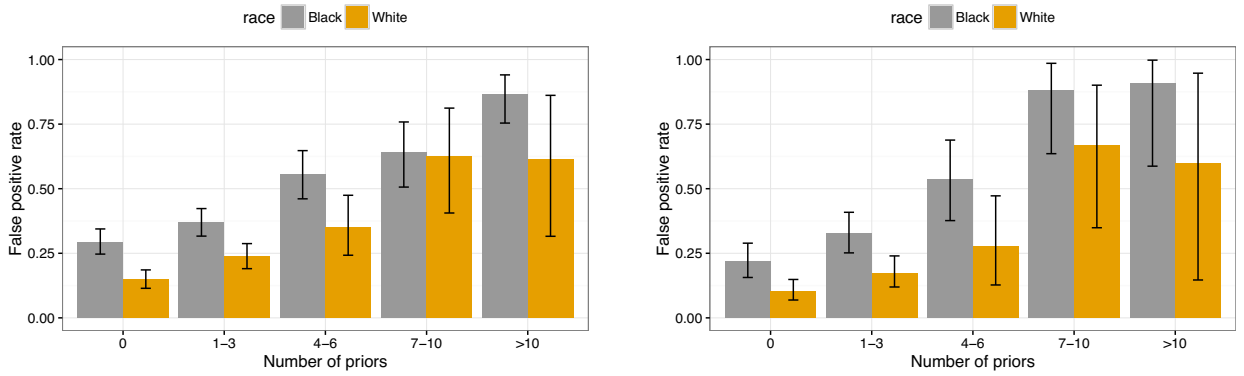
$$\equiv (t_{\max} - t_{\min})\left(\mathbb{P}(S > s_{\mathrm{HR}} \mid R = b, Y = 0, X \in \mathcal{X}) - \mathbb{P}(S > s_{\mathrm{HR}} \mid R = w, Y = 0, X \in \mathcal{X})\right). \tag{3.5}$$

The false positive rates shown in Figure 2(a) correspond precisely to the quantities $\mathrm{FPR}_r(\mathcal{X})$ for choices of $\mathcal{X}$ given by different prior record count bins. The leftmost bars correspond to taking $\mathcal{X} = \{\#\text{priors} = 0\}$. Similarly the leftmost bars in Figure 2(a) correspond to taking $\mathcal{X} = \{\#\text{priors} = 0, \text{charge degree} = M\}$. In Appendix B we present a logistic regression analysis showing that significant differences in false positive rates persist even after adjusting for a number of other recidivism-related covariates.

3.2 Connections to measures of differences in distribution

In their analysis of the PCRA instrument, Skeem and Lowenkamp [11] remark that some applications of the risk score could create disparate impact due to differences in the score distributions between black and white offenders. To summarize the distributional difference in scores between the two groups, the authors report a Cohen's d of 0.34, with a corresponding non-overlap of 13.5%. A

[a] We are overloading notation in this expression: Here, $\mathrm{FPR}_r = \mathbb{P}(\mathrm{HR} \mid R = r, t_L = 0)$, similarly for FNR_r.



(a) All defendants.

(b) Defendants charged with a Misdemeanor offense.

Figure 2: False positive rates across prior record count. Plot is based on assessing a defendant as "high-risk" if their COMPAS decile score is $> s_{\text{HR}} = 4$. Error bars represent 95% confidence intervals.

natural question to ask is whether the level of disparity in sentence duration, Δ, is in some sense closely related to such measures of distributional difference. With a small generalization of the *% non-overlap* measure, we can answer this question in the affirmative.

The % non-overlap of two distributions is generally calculated assuming both distributions are normal, and thus has a one-to-one correspondence to Cohen's d[27].[a] However, as we can see from Figure 3, the COMPAS decile score is far from being normally distributed in either group. A more reasonable way to calculate % non-overlap in such cases is to note that in the Gaussian case % non-overlap is equivalent to the total variation distance. Letting $f_{r,y}(s)$ denote the score distribution among individuals in group r with recidivism outcome y, one can establish the following sharp bound on Δ.

Proposition 3.2 (Percent overlap bound)**.** *Under the MinMax policy,*

$$\Delta(y_1, y_2) \leq (t_{\max} - t_{\min}) d_{\text{TV}}(f_{b,y_1}, f_{w,y_2}).$$

This result is simple to understand. When there is some non-overlap between the score distributions for two groups, the worst case scenario is that the non-overlap is entirely due to mass shifting from scores below s_{HR} to those above s_{HR}. In such cases, the inequality becomes an equality.

3.3 Empirical results

In this section we present some empirical results based on two hypothetical sentencing rules: the *MinMax* rule introduced in the previous section, and the *Interpolation* rule, which we will introduce below. Though the offenders in our data set come from Broward County, Florida, our empirical analysis is modelled on the sentencing guidelines of the State of Pennsylvania.

[a] $d = \frac{\bar{S}_b - \bar{S}_w}{SD}$, where SD is a pooled estimate of standard deviation.

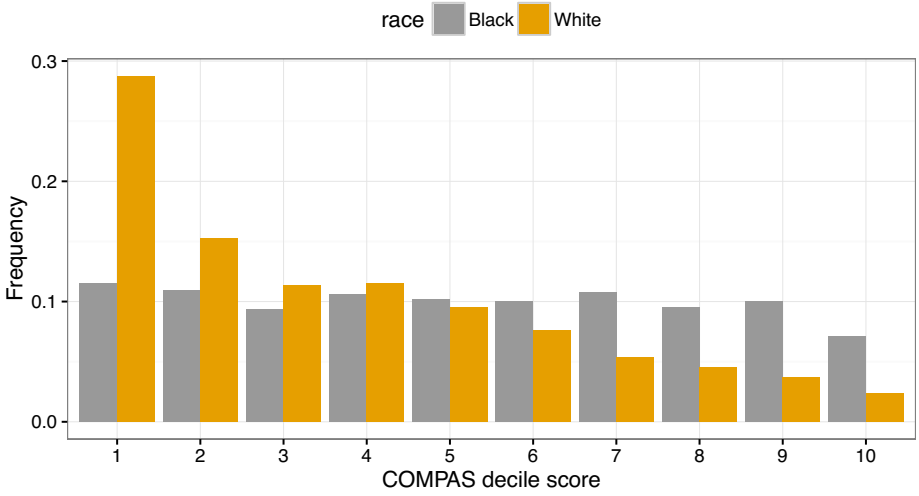


Figure 3: COMPAS decile score histograms for Black and White defendants. Cohen's $d = 0.60$, non-overlap $d_{\mathrm{TV}}(f_b, f_w) = 24.5\%$.

The penalty ranges $t_{\min}$ and $t_{\max}$ are selected by approximately matching each offender's charge degree (M2 - F1) to a sentence range in Pennsylvania's Basic Sentencing Matrix (PA Code §303.16). This matrix provides sentence ranges based on the charge degree for the current offense and the defendant's prior record score (0 - 5+). We do not have enough information in the Broward County data to reliably assign a prior record score for each individual. Our results are based on using the sentencing range corresponding to a prior record score of 1 for all defendants in the data.

Figure 4 shows the expected sentences for black and white defendants broken down by observed recidivism outcome. The x-axis in these figures is taken to be the offense gravity score, which for the purpose of this analysis is mapped to charge degree as indicated in Table 2.

Offense gravity score	2	3	5	7	8
Charge Degree	(M2)	(M1)	(F3)	(F2)	(F1)

Table 2: Mapping between offense gravity score and charge degree used in the empirical analysis.

Results are shown for both the MinMax policy introduced earlier in this section, and the Interpolation policy, which is given by

$$T_{\mathrm{Int}}(s) = t_{\min} + \frac{s-1}{9}(t_{\max} - t_{\min}). \tag{3.6}$$

Unlike the MinMax policy, which is based on the coarsened score, the Interpolation policy assigns sentences by linearly interpolating between $t_{\min}$ and $t_{\max}$ based on the assigned decile score. We see that under both policies there are consistent trends in the expected sentences. Black defendants are observed to receive higher sentences than white defendants both within the non-recidivating subgroup and the recidivating subgroup (except in the F1 charge degree category, where sample sizes are small and results are non-significant). Since white defendants have higher false negative rates and lower false positive rates than black defendants, the empirical results are consistent with the theoretical results presented earlier in this section.

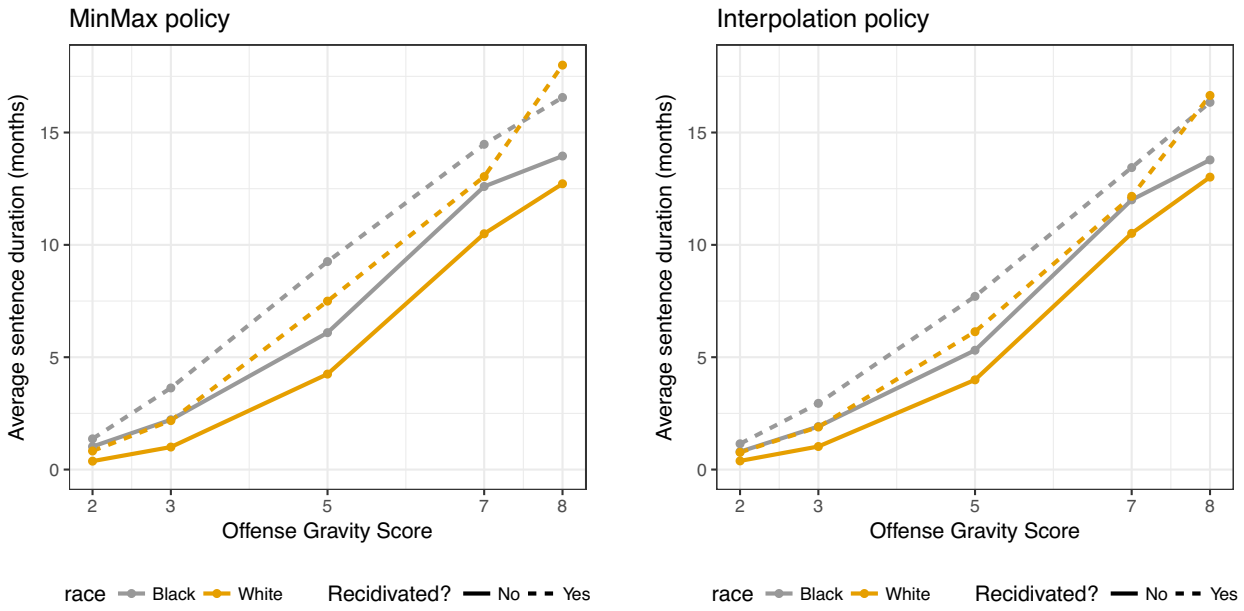


Figure 4: Average sentences under the hypothetical sentencing policies described in Section 3.3. The mapping between the x-axis variable and the offender's charge degree is given in Table 2. For all OGS levels except 8, observed differences in average sentence are statistically significant at the 0.01 level.

4 Revisiting predictive parity

In this final section we revisit the notion of predictive parity and further discuss its implications for general classifiers. We know from equation (2.6) that when the positive predictive values are constrained to be equal but the prevalences differ across groups, the false positive and false negative rates cannot both be equal across those groups. While we have no direct control over recidivism prevalence, we do have some control over the PPV and error rates of our classifiers. At least in principle, we are free to tune our classifiers in any of the following ways:

(i) Allow unequal false negative rates to retain equal PPV's and achieve equal false positive rates

(ii) Allow unequal false positive rates to retain equal PPV's and achieve equal false negative rates

(iii) Allow unequal PPV's to achieve equal false positive and false negative rates

Figure 5 helps to put these trade-offs into perspective. From (2.6), we can see that FPR is a linear function of FNR under constraints on PPV and p. This means that, if PPV is fixed at a given value, tuning strategy (i) may require a very large increase in FNR in order to balance FPR. The black line shows feasible combinations of $(\text{FNR}_b, \text{FPR}_b)$ when PPV_b is forced to equal the observed value $\text{PPV}_w = 0.591$. We can see that to get FPR_b to match FPR_w, we would need to increase FNR_b to around 0.7, which would be a substantial drop in accuracy. In view of Corollaries 3.1 and 3.2 Strategies (i) and (ii) may generally be undesirable because while they reduce disparate impact for one subgroup (e.g., among non-recidivists), they may increase it in the other.

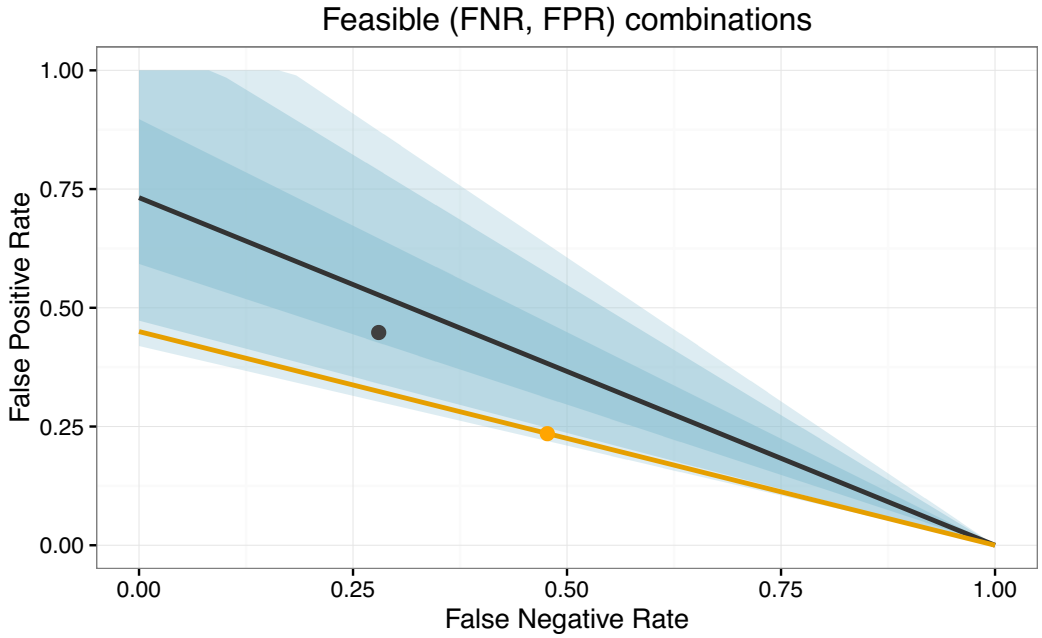


Figure 5: The two points represent the observed values of (FNR, FPR) for Black and White defendants. The orange line represents feasible values of (FNR, FPR) for White defendants when the prevalence p_w and PPV_w are both held fixed at their observed values in Table 1. The dark grey line represents feasible values of $(\text{FNR}_b, \text{FPR}_b)$ when the prevalence p_b is held fixed at the observed value and PPV_b is set equal to the observed value of $\text{PPV}_w = 0.591$. *Nested* shaded regions correspond to feasible values of $(\text{FNR}_b, \text{FPR}_b)$ if we allow PPV_b to vary under the constraint $|\text{PPV}_b - 0.591| < \delta$, with $\delta \in \{0.05, 0.1, 0.125\}$. The smaller δ, the smaller the feasible region.

The preferred approach, at least in some cases, may be to pursue strategy (iii). This amounts to using a score S that does not satisfy predictive parity in the first place, but can also be achieved by allowing the high-risk cutoff $s_{\text{HR},r}$ to differ across groups. The shaded regions in Figure 5 show feasible values of $(\text{FNR}_b, \text{FPR}_b)$ when we allow PPV_b to be within some δ of the observed value of PPV_w. We can see that even at small values of δ the feasible region is quite large.

5 Discussion

The primary contribution of this paper was to show how disparate impact can result from the use of a recidivism prediction instrument that is known to satisfy the fairness criterion of predictive parity. Our analysis focussed on the simple setting where a binary risk assessment was used to inform a binary penalty policy. While all of the formulas have natural analogs in the non-binary score and penalty setting, we find that many of the salient features are already present in the analysis of the simpler binary-binary problem.

A key limitation of our analysis stems from potential biases in the observed data that may affect our ability to draw valid inferences concerning the fairness of an RPI. Throughout this paper we have implicitly operated under the assumption that the observed recidivism outcome Y is a suitable outcome measure for the purpose of assessing the fairness properties of a recidivism

prediction instrument. However, the true outcome of interest in this context is *reoffense*, which is not what we observe. In the latest statistics released by the Federal Bureau of Investigation[28], it is reported that 46% of violent crimes and 19.4% of property crimes were successfully cleared by law enforcement agencies. Many criminal offenders are simply never identified. It is therefore possible that a non-negligible fraction of the individuals in our data for whom we observed $Y = 0$ did in truth reoffend. If this is indeed the case, and if there are group differences in the rates at which offenders are caught, the findings of empirical fairness assessments may be misleading. Understanding how such forms of data bias affect the ability to assess instruments with respect to different fairness criteria is a subject of our ongoing research efforts.

6 Conclusion

In closing, we would like to note that there is a large body of literature showing that data-driven risk assessment instruments tend to be more accurate than professional human judgements[29,30], and investigating whether human-driven decisions are themselves prone to exhibiting racial bias[31,32]. We should not abandon the data-driven approach on the basis of negative headlines. Rather, we need to work to ensure that the instruments we use are demonstrably free from the kinds of biases that could lead to disparate impact in the specific contexts in which they are to be applied.

A Proofs

Proof of Proposition 3.1. To simplify notation, we let HR denote the event $\{S > s_{\text{HR}}\}$.

$$\begin{aligned}
\mathbb{E}(\Delta(y_1, y_2)) &= \mathbb{E}(T \mid R = b, Y = y_1) - \mathbb{E}(T \mid R = w, Y = y_2) \\
&= t_{\max}\mathbb{P}(HR \mid R = b, Y = y_1) + t_{\min}(1 - \mathbb{P}(HR \mid R = b, Y = y_1)) \\
&\quad - t_{\max}\mathbb{P}(HR \mid R = w, Y = y_2) - t_{\min}(1 - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&= t_{\max}(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&\quad + t_{\min}(\mathbb{P}(HR \mid R = w, Y = y_2) - \mathbb{P}(HR \mid R = b, Y = y_1)) \\
&= (t_{\max} - t_{\min})(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2))
\end{aligned}$$

□

Proof of Proposition 3.2. By definition of total variation distance, for any event A,

$$|(\mathbb{P}(A \mid R = b, Y = y_1) - \mathbb{P}(A \mid R = w, Y = y_2))| \leq d_{\text{TV}}(f_{b,y_1}, f_{w,y_2})$$

Applying this inequality to Proposition 3.1 with $A = \{S_c = \text{HR}\}$ gives

$$\begin{aligned}
\mathbb{E}(\Delta(y_1, y_2)) &= (t_{\max} - t_{\min})(\mathbb{P}(HR \mid R = b, Y = y_1) - \mathbb{P}(HR \mid R = w, Y = y_2)) \\
&\leq (t_{\max} - t_{\min}) d_{\text{TV}}(f_{b,y_1}, f_{w,y_2})
\end{aligned}$$

□

B Covariate-adjusted false positive rates

In this section we present the results of a logistic regression analysis that we conducted in order to assess whether the observed differences in false positive rates between black and white defendants can be entirely accounted for by other covariates. We find that adjusting for covariates decreases the gap, but it nevertheless remains large and statistically significant.

For the purpose of this analysis we consider only the subset of defendants who *do not* recidivate. The outcome variable for the logistic regression is taken to be

$$y = \begin{cases} 1, & S > 4 \\ 0, & S \leq 4 \end{cases},$$

where S denotes the COMPAS decile score. In this setup, $y = 0$ denotes a True Negative and $y = 1$ denotes a False Positive. Statistically significant positive coefficient estimates correspond to variables associated with increased likelihood of false positives.

Table 3 shows the results of regressing y on race alone. The coefficient of race in this model is large, positive, and statistically significant. Without adjusting for other covariates, the odds that a non-recidivating Black defendant receives a high-risk assessment are $e^{0.976} = 2.6$ times higher than those of a White defendant.

Table 4 shows the results of regressing y on race, age, gender, number of priors, and charge degree. The coefficient of race is smaller than it was in the un-adjusted model, but it is nevertheless large and statistically significant. Even after adjusting for these other factors, the odds that a non-recidivating Black defendant receives a high-risk assessment are $e^{0.547} = 1.72$ times higher than those of a White defendant.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-1.183	0.061	-19.33	0.0000
raceBlack	0.976	0.077	12.60	0.0000

Table 3: Logistic regression with race alone.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	1.397	0.176	7.92	0.0000
raceBlack	**0.547**	0.087	6.30	0.0000
Age	-0.079	0.005	-17.48	0.0000
sexMale	-0.291	0.098	-2.97	0.0030
Number of Priors	0.283	0.016	17.78	0.0000
chargeMisdemeanor	-0.109	0.088	-1.25	0.2123

Table 4: Logistic regression with race and other covariates that may be associated with recidivism

References

[1] Model penal code: Sentencing. American Law Institute, 2016.

[2] Thomas Blomberg, William Bales, Karen Mann, Ryan Meldrum, and Joe Nedelec. Validation of the compas risk assessment classification instrument. 2010.

[3] Ben Casselman Anna Maria Barry-Jester and Dana Goldstein. Should prison sentences be based on crimes that haven't been committed yet?

[4] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

[5] Northpointe. Compas risk need assessment system: Selected questions posed by inquiring agencies.

[6] Anthony W Flores, Kristin Bechtel, and Christopher T Lowenkamp. False positives, false negatives, and false analyses: A rejoinder to "machine bias: There's software used across the country to predict future criminals. and it's biased against blacks.". *Unpublished manuscript*, 2016.

[7] William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. 2016.

[8] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. How we analyzed the compas recidivism algorithm. 2016. URL `https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm`.

[9] Administrative Office of the United States Courts. An overview of the federal post conviction risk assessment, September 2011.

[10] Jay P Singh. Predictive validity performance indicators in violence risk assessment: A methodological primer. *Behavioral Sciences & the Law*, 31(1):8–22, 2013.

[11] Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, & recidivism: Predictive bias and disparate impact. *Available at SSRN*, 2015.

[12] Jennifer L Skeem, John Monahan, and Christopher T Lowenkamp. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Available at SSRN 2718460*, 2016.

[13] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, pages 3315–3323, 2016.

[14] Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012.

[16] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

[17] Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining*, pages 144–152. SIAM, 2016.

[18] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, 29(05):582–638, 2014.

[19] Solon Barocas and Andrew D Selbst. Big data's disparate impact. 2016.

[20] Kelly Hannah-Moffat. Actuarial sentencing: An ?unsettled? proposition. *Justice Quarterly*, 30(2):270–296, 2013.

[21] Jennifer Skeem. Risk technology in sentencing: Testing the promises and perils (commentary on hannah-moffat, 2011). *Justice Quarterly*, 30(2):297–303, 2013.

[22] John Monahan and Jennifer L Skeem. Risk assessment in criminal sentencing. *Annual review of clinical psychology*, 12:489–513, 2016.

[23] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016.

[24] Sam Corbett-Davies, Emma Pierson, Avi Feller, and Sharad Goel. A computer program used for bail and sentencing decisions was labeled biased against blacks. it's actually not that clear.

[25] Nancy Ritter. Predicting recidivism risk: New tool in philadelphia shows great promise. *National Institute of Justice Journal*, 271, 2013.

[26] PCS. Validation of risk scale. Technical report, Pennsylvania Commission on Sentencing, 2013.

[27] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences (2nd Edition)*. Lawrence Erlbaum Associates, 1988.

[28] Uniform crime report: Crime in the united states, 2015 - offenses cleared. U.S. Department of Justice, 2016.

[29] Paul E Meehl. *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.* University of Minnesota Press, 1954.

[30] William M Grove, David H Zald, Boyd S Lebow, Beth E Snitz, and Chad Nelson. Clinical versus mechanical prediction: a meta-analysis. *Psychological assessment*, 12(1):19, 2000.

[31] Shamena Anwar and Hanming Fang. Testing for racial prejudice in the parole board release process: Theory and evidence. Technical report, National Bureau of Economic Research, 2012.

[32] Laura T Sweeney and Craig Haney. The influence of race on sentencing: A meta-analytic review of experimental studies. *Behavioral Sciences & the Law*, 10(2):179–195, 1992.