Human Rights Data Analysis Group hrdag.org
Using science and data to advance human rights and justice

MEASURES OF FAIRNESS FOR NEW YORK CITY'S SUPERVISED RELEASE RISK ASSESSMENT TOOL

KRISTIAN LUM AND TARAK SHAH
OCTOBER 1, 2019

1. Introduction

In 2014, the Criminal Justice Agency (CJA) was asked by the New York City Mayor's Office of Criminal Justice (MOCJ) to create a risk assessment tool to screen defendants'[1] suitability for acceptance into a pretrial supervised release program. This effort involved compiling an extensive dataset on demographic, criminal history, and arrest information by combining information contained in CJA's database[2] with data from the New York Division of Criminal Justice Services (DCJS).[3] Based on this data, CJA worked with MOCJ to create a statistical model to predict the likelihood that a defendant will be re-arrested for a felony during the pre-trial period– the prediction objective set by MOCJ. The model estimated from the data was then manually adjusted to incorporate input from MOCJ to arrive at a final 'compromise Risk Point system.'

This final model takes demographic and criminal history information about a defendant as inputs and outputs a risk category, ranging from low to high risk. Defendants classified in lower risk groups are considered better candidates for the supervised release program than those in higher risk groups. Defendants whose risk is assessed as 'high' are ineligible for release into the program [Redcross et al., 2017]. A full description of the model, its development, and our reproduction is given in Appendix A.

Following the development of this model, CJA compiled an unpublished report validating the predictive utility of the model [Healy, 2015]. In this report, CJA demonstrated that the model achieved predictive accuracy (measured by AUC, a common measure of predictive accuracy in risk assessment tools) of about 0.67, which is on par with many other available risk assessment tools [Desmarais et al., 2016]. They also demonstrated the tool's ability to predict not only felony re-arrest (the outcome the tool was designed to predict) but also any re-arrest and failure

[1]Throughout this document, we use the term 'defendant' to refer to the arrested people who appear in the dataset and to whom the model will be and has been applied. In the interest of clarity, we point out that all of these individuals are presumed innocent, and by using the term defendant we intend not to create stigma, but rather to refer to the population under discussion with adequate precision.

[2]CJA is a not-for-profit organization and information in its database is not an official source of criminal justice data. CJA is not responsible for the design, methods, or conclusions of the paper.

[3]This data is provided by the New York State Division of Criminal Justice Services (DCJS). The opinions, findings, and conclusions expressed in this publication are those of the authors and not those of DCJS. Neither New York State nor DCJS assumes liability for its contents or use thereof.

to appear. The evaluation, however, included little analysis of differences in the model's predictive performance between racial and ethnic groups.

Over the past several years, major concerns about the role similar risk assessment models play in alleviating or exacerbating racial inequalities in the criminal justice system have been at the forefront of public attention.[4] With the cooperation of CJA, the goal of this short document is to evaluate the supervised release risk assessment tool according to several mathematical notions of fairness, focusing on fairness with respect to race.[5]

This task is not straight-forward, as there are several– potentially mathematically incompatible– definitions of fairness that have been argued for in this setting. With that in mind, we present simple analyses that evaluate the risk assessment tool's adherence to definitions of fairness that are central to the current debate. This is by no means an exhaustive analysis of all notions of fairness.

2. Data

2.1. **General considerations.** Risk assessment models, including the one under consideration here, typically operate by using some information about the defendant –such as criminal history and demographic information– to predict an outcome of interest. Here, the outcome of interest is whether the defendant was re-arrested for a felony during the pre-trial. Implicitly, this outcome is meant to be a measure of the defendant's participation in serious criminal activity.

Nearly all evaluations of the fairness of a risk assessment model hinge on comparisons between the model's predictions and the outcome variable. For example, one evaluation measure might be how often the model correctly predicted the outcome. When evaluating fairness with respect to race, large discrepancies in the chosen measure between race categories are typically considered evidence of a racially biased model. Lack of a large discrepancy is considered evidence of no bias. However, if the measured outcome (felony re-arrest) is an *unfair* measure of the concept of interest (participation in serious criminal activity), then any such comparisons may be misleading.

To make this concrete, consider a hypothetical world in which exactly half of Black defendants and half of White defendants participated in serious criminal activity (re-offended) during their pre-trial period. In this world, suppose all of the Black defendants who re-offended were re-arrested whereas only half of the White defendants who re-offended were re-arrested. The differential probability of arrest could be the result of racially biased patterns of enforcement. Suppose also that we have access to a risk assessment instrument that is able to perfectly predict who will be re-arrested. Such an instrument would predict that half of all Black defendants would be re-arrested but that only one-quarter of White defendants would be re-arrested. If we use accuracy as the relevant measure of fairness, we would find that the model exhibited no racial unfairness–in this hypothetical world, the model would be correct 100% of the time for both the White defendants and Black defendants. However, such a conclusion would be misleading because Black defendants would be predicted to be future recidivists at twice the rate of White

[4]See, for example [Harcourt, 2014] or a Statement of Concern signed by over 100 civil rights organizations.

[5]The evaluations are done using a categorization scheme that includes both race and ethnicity. For brevity, references to 'race' refer to the combination of race and ethnicity.

defendants, despite the fact that there was no difference in the rate of re-offense between the two groups. In this case, our evaluation would fail to reveal the racial bias in the model due to unfairness in the way in which the outcome variable is measured. Similar problems with a disconnect between a "construct space" and a "measurement space" are detailed in [Friedler et al., 2016].

Most measures of predictive fairness of the type we will consider in section 3 are dependent on the assumption that felony re-arrest is a fair measure of re-offense. Thus, before we delve into the various ways in which one might assess fairness based on racial disparities in the relationship between model predictions and felony re-arrest, it is important that we first investigate the extent to which felony re-arrest might itself be a racially biased measure of involvement in serious criminal activity.

2.2. **Fair collection of NYC data.** In order to do this, it is useful to consider the historical setting during which the training data was collected. The data used to train and evaluate the supervised release risk assessment model is made up of all defendants arrested during the first half of 2009– January 1, 2009 to June 30, 2009– whose cases reached a final disposition or had a warrant ordered prior to June 30, 2011. If the defendant was re-arrested for a felony before the final disposition of his case, he is considered to have had a felony re-arrest. Though researchers have long argued (not uncontroversially) that arrests for violent crime are less racially biased than other types of crimes [Beck and Blumstein, 2018; Skeem and Lowenkamp, 2016; Hindelang, 1978], felony arrests encompass many types of crime beyond just the violent. For the purposes of assessing fairness with respect to race, felony re-arrest will only be a fair measure of future involvement in serious criminal activity if the likelihood that a re-offense results in a re-arrest is equal regardless of an individual's race during the time period considered in the data.

The data used to train and evaluate the model at hand was collected during the peak years of Stop, Question, and Frisk (SQF), a practice that was ultimately ruled to have been applied in a racially discriminatory manner in Floyd vs. City of New York (959 F. Supp. 2d 540). According to the New York Civil Liberties Union, from 2009-2011, 87 percent of SQFs were of Blacks and Latinos.[6] To understand the impact that SQF had on this particular data set, one would ideally tabulate the proportion of the arrests in the data that were the result of SQF. However, the provided data does not include this information, so it is not possible to evaluate this directly. Furthermore, we could find no variable in the provided data that describes the charges associated with the *re-arrest* in detail, making it impossible to identify which re-arrests are similar to those generated by SQF.

According to a report by the New York State Office of the Attorney General [Bureau and Schneiderman, 2013], SQF arrests most commonly resulted in the following charge types: marijuana possession (14.9%), trespass (13.8%), violence (12.9%), weapons offenses (12.3%), and minor property crimes (11.6%). Drug charges[7] made up 23.66%[8] of SQF arrests. As an alternative to directly calculating how many arrests were due to SQF, we compare the composition of the arrests in the training data to the arrest types identified as commonly generated under SQF. In particular, we focus on drug crimes and weapons possession, which are both identified as

[6]NYCLU's Stop-and-Frisk Data

[7]including marijuana charges, per Appendix D of [Bureau and Schneiderman, 2013]

[8]sum across marijuana and non-marijuana crimes in Appendix G of [Bureau and Schneiderman, 2013]

among the most likely types of charges to result from a SQF stop. Of course, it is possible that many of the arrests of these types were generated by some process entirely unrelated to SQF, so the extent of the impact of SQF (or any other type of discriminatory policing) on the training data cannot be definitively determined from this dataset.

A variable representing "the most severe arrest charge" is present in the provided data set. In some cases, this variable is reported using a charge code (e.g. '220.18'), and in some cases an abbreviation is given (e.g. 'cpcs'– assumed to be 'criminal possession of a controlled substance'). In all cases, DCJS's manual is used to interpret charge codes.[9] One of the most common designations used is 'cp marij 5', and we assume that other 'cpcs' designations refer to drug possession other than marijuana. Table 1 gives the charge codes and text definitions used to classify each arrest in the dataset. For example, an arrest was classified as having had a top charge that was a drug crime if the charge code began with 220 or 221 or if the text in the field contained either 'cpcs', 'cscs', or 'marij'. The right-most columns in the table give the percent of all arrests for which the most serious charge fell into each category and the percent of felony arrests for which the most serious charge fell into each category.

Category	Codes	Abbreviations	% Arrests	% Fel. Arrests
Marij. Poss.	221.05-30	cp marij	11	1
Drug Crime	220.*, 221.*	cpcs, cscs, marij	29	30
Weapon Poss.	265.01-04	cpw	5	9
Trespassing	140.05-17	tresp	6	0

TABLE 1. Definitions used to classify arrests as drug or weapons related, along with each type's percentage of total and felony arrests.

Nearly 30% of the arrests that appear in the dataset are drug-related. This figure is similar when considering only felony arrests. If we assume that re-arrests follow a similar distribution to initial arrests, then it stands to reason that the outcome variable– felony re-arrest– is in large part made up of drug-related arrests. Additionally, 9% of all felony arrests in the dataset were for weapons possession. Again, if re-arrests follow a similar pattern as initial arrests, then it also stands to reason that some, likely non-negligible, proportion of the re-arrests were for a crime that was a stated target of the SQF policy. Restricting the definition of the outcome variable to be strictly *felony* re-arrest does, however, remove the effect of arrests due to trespassing, another crime associated with SQF.

	race	Brooklyn	Manhattan	Queens	Staten Island	Bronx
1	WHITE	0.30	0.29	0.38	0.63	0.98
2	BLACK	1.62	4.51	1.12	2.26	2.67
3	HISPANIC	1.21	2.26	0.54	0.65	2.11

TABLE 2. Drug-related felony arrests per 1000 people

[9]Link to DCJS's codebook

	race	Brooklyn	Manhattan	Queens	Staten Island	Bronx
1	WHITE	0.06	0.09	0.08	0.08	0.08
2	BLACK	0.69	1.43	0.85	0.83	0.64
3	HISPANIC	0.32	0.64	0.23	0.28	0.28

Table 3. Weapons possession felony arrests per 1000 people

To address whether in this particular dataset, these types of arrests were more commonly made for racial/ethnic minorities, we disaggregate the analysis by the provided race/ethnicity variable.[10] The number of felony drug-related and felony weapons possession-related arrests that appear in the dataset for each borough and racial/ethnic group is divided by the total number of individuals with that same racial/ethnic designation and borough of residence.[11] To be consistent with the way race and ethnicity were combined in the original report, all people of Hispanic origin are considered Hispanic. White and Black non-Hispanic people are defined as White and Black, respectively. Due to insufficient sample sizes for other groups, these are the only three groups considered in this evaluation. This convention is used throughout.

This gives us race/ethnicity normalized felony arrest rates. Results are shown in Tables 2 and 3. We see that Black and Hispanic people in this dataset experience arrest for felony drug crimes and felony weapons possession at a higher rate relative to their rate of residence in the borough. This result has many possible explanations, including differential participation in drug activity and illicit weapons possession. However, given the historical context in which this data was collected, it is also possible that these observed disparities are partially the result of racially biased enforcement and have a non-negligible impact on both the model development and the evaluation itself. With this caveat in mind, we now evaluate the New York City supervised release risk assessment tool according to various notions of algorithmic fairness that have been proposed in the literature.

3. Analysis of common measures of fairness

Many different measures of the fairness have been proposed in the context of risk assessment [Berk et al., 2018]. In this section, we evaluate the supervised release tool for fairness following several published evaluations of similar risk assessment tools. We begin by reporting compliance of the supervised release tool with several parity-based measures of fairness, which are among the most commonly considered in the current debate around risk assessment. We continue by exploring whether the supervised release tool exhibits racial bias according to several other notions of fairness, including those that consider a predictive model fair based on the procedure used to create the model.

[10]Because the burden of SQF fell primarily on young people of color, further analysis exploring the impact disaggregated by age and race/ethnicity may be revealing. Ultimately, because the mandate of this report is to assess the model for racial bias, we do not consider this here.

[11]Population totals are obtained from page 14 of NYC 2010: Results from the 2010 Census, using data from 2010. The arrest data does not include a person's home address, only the location of the arrest – we are assuming here that most drug- and weapons-related arrests occurred close to home.

Because the outcome variable is felony re-arrest before case disposition, for many of the considered notions of fairness, metrics can only be calculated for defendants who were released before their case was disposed. For those who were not released pre-trial, we do not know whether they would have been re-arrested pre-trial and the outcome measure is not available. This introduces a potential source of bias, as it could be the case that there is race-specific variation–even for a given likelihood of felony re-arrest– in the rate at which defendants were granted pre-trial release. Nonetheless, in keeping with the initial analysis of this data, for those measures that require a comparison between the prediction and the outcome, we present fairness measures using the set of all defendants who were released pre-trial (n = 47,370). For the measure that does not require a comparison of the prediction with the outcome, we use the full dataset (n = 92,004). [12]

One other consideration is that the data used to train the model includes cases that, by virtue of being classified as Violent Felony Offenses (VFO), would not be eligible for release regardless of risk score based on New York rules. The results presented here are based on analysis of the same data used to train the model. We re-ran the analysis using only the subset of cases whose arraignment charges did not include a VFO, and the results did not change substantively. However, the rate of VFO charges varies by race: 7.3% of white defendants had a VFO charge at arraignment, compared to 12.7% of black defendants and 10.6% of Hispanic defendants. In this way, supervised release is more available to White defendants than Black or Hispanic defendants, even before considering risk scores.[13]

3.1. **Parity-based measures.** At least partially sparked by a news article that revealed that a popular risk assessment model, COMPAS, had nearly twice the false positive rate for Black defendants as for White defendants [Angwin et al., 2016], a robust discussion about how to appropriately characterize a predictive model's fairness with respect to a sensitive variable, such as race, has emerged in the context of risk assessment in the machine learning and statistics literature. In response to the criticism that variation in false positive rates by race constituted evidence of a racially biased tool, the makers of the COMPAS tool published an article demonstrating that according to two different parity-based notions of fairness– 'predictive parity' and 'accuracy equity'– their tool was, in fact, fair with respect to race [Dieterich et al., 2016]. [Kleinberg et al., 2016; Chouldechova, 2017] then showed mathematically that the notions of fairness championed by both sides were mutually incompatible under realistic conditions– specifically, variation by race in the rate at which defendants were re-arrested and a model that is not correct 100% of the time. Given that in this dataset Black and Hispanic defendants were re-arrested for felony offenses at a higher rate than White defendants and no models in this domain even approach perfect prediction, it will not be possible to simultaneously satisfy all group-wise parity-based notions of fairness. It is also worth noting that several of these definitions of fairness are referred to using different names in different papers and literature.

[12]Due to the large sample sizes used here, all reported differences are statistically significant at at least the 0.05 level unless otherwise noted.

[13]Overall, of 5,198 cases where the arrest charges included a VFO and the charged person was released pre-trial, 8.8% had a felony re-arrest pre-trial, compared to 6.6% of the 42,172 non-VFO cases.

3.1.1. *Predictive parity and accuracy equity.* We first consider the definitions of fairness supported by [Dieterich et al., 2016] in their defense of the COMPAS model. Definitions similar to these are among the most common used in this context. We calculate whether the supervised release tool exhibits 'predictive parity'– equal re-arrest rates by race within each risk category– and 'accuracy equity'– equal predictive accuracy for all race groups as measured by the area under the ROC curve (AUC). Results for predictive parity are shown in Figure 1, which shows the rate of felony re-arrest per one hundred defendants by race and risk score. A model is considered to have achieved predictive parity if within each risk score category, the rate of felony re-arrest is the same across all race groups. We find that across all but the lowest risk category, White defendants experienced felony re-arrest at a lower rate than their Black and Hispanic counterparts. This indicates, for example, a White defendant with a risk category of five is less at risk of felony re-arrest than a Black or Hispanic defendant assigned the same risk category. Thus, *this model does not meet the 'predictive parity' criterion.*

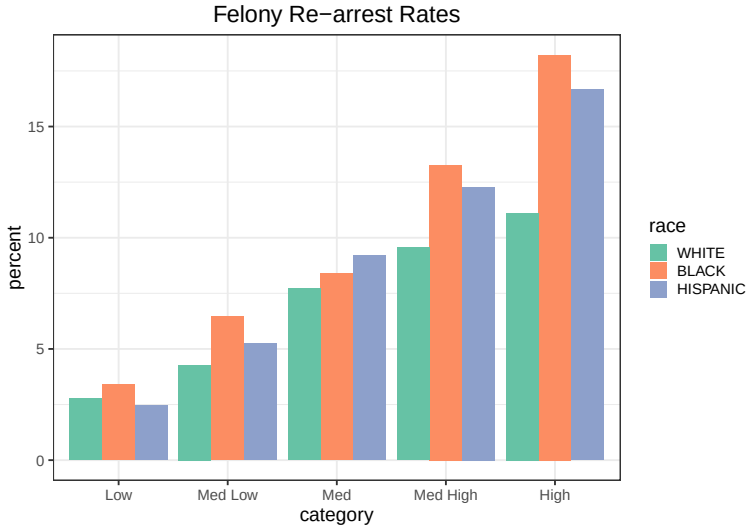

FIGURE 1. Percent of defendants with a pre-trial felony re-arrest for each risk category and race group

A related notion of fairness supported by [Dieterich et al., 2016] is 'accuracy equity'– equal predictive accuracy for all race groups. In this case, predictive accuracy is measured by the AUC. Table 4 gives the AUC for each race-group separately. A larger value of the AUC indicates greater predictive accuracy. For this population, the model is most predictive for Hispanic defendants and least for White defendants, though qualitatively the predictive accuracy is similar across all considered groups (the difference in AUC between the White and Black groups is also not statistically significant at the 0.05 level).

3.1.2. *Equal false positive rates.* In order to conduct a false-positive-based analysis similar to that of the ProPublica analysis of COMPAS [14], it is necessary to define

[14][Hardt et al., 2016] also argues for the appropriateness of this and related definitions of fairness in the context of lending.

race	AUC
WHITE	0.64
BLACK	0.66
HISPANIC	0.68

TABLE 4. Area under ROC curve by race

what the model considers a 'positive' prediction of felony re-arrest. This then defines a 'false positive' as any person who had a positive prediction for felony re-arrest but was not re-arrested for a felony.

We consider two definitions. In the deployment of the model, any defendant with a 'high' risk categorization is ineligible for the supervised release program [Redcross et al., 2017]. Thus, one definition we consider is any defendant with a 'high' risk score (the highest category). The second definition we consider is any individual classified 'medium high' or 'high'. This is arguably also a reasonable definition of a prediction of felony re-arrest, as the language for both categories indicates a high risk.

Recent changes to the threshold for determining what constitutes 'high' risk have resulted in a large increase in the number of people who are eligible for the supervised release program under this risk assessment model.[15] The findings in this report pertain to the thresholds that were in use from the deployment of the model through mid-2019 when the changes were enacted.

We first calculate race-specific false positive rates under the first definition– the model considers a future recidivist to be any individual classified as 'high' risk. That is, for each race group, we calculate the percent *of those defendants who would not go on to be re-arrested* that were identified by the model as at 'high' risk of recidivism. The results of this calculation are given in column FPR (high) of Table 5. We find that the false positive rate– the rate at which non-recidivists are classified as likely future recidivists by the model and thus made ineligible for the supervised release program– is largest for Black defendants, intermediate for Hispanic defendants, and lowest for White defendants. While the rates differ by group, they are all quite small due to the restrictive definition of positive in this case. The equivalent analysis for the second definition is shown in the FPR (medium-high+) column of Table 5. Though the rates are higher under this definition, qualitatively the results remain the same– Black and Hispanic individuals who do not experience a felony re-arrest have a higher likelihood of being classified in the higher risk groups. *Under both definitions, the supervised release model also does not meet the equal false positive rate standard of fairness championed in the ProPublica analysis.*

race	FPR (high)	FPR (medium-high +)
WHITE	0.03	0.11
BLACK	0.07	0.22
HISPANIC	0.05	0.17

TABLE 5. False positive rates by race.

[15]See, for example, this NY Post Article about the changes.

3.1.3. *Demographic parity.* Finally, we turn to a notion of fairness that considers the rate at which defendants are classified into different risk categories, regardless of their status as felony re-arrestees or not.[16] This notion of fairness can be interpreted as a measure of disparate impact. If there are large differences by race in the rate at which defendants are assigned to high risk score categories, then (if the tool's recommendations are followed), this will result in large racial disparity in eligibility for the supervised release program.

Because this notion of fairness does not account for re-arrest, we calculate this using the full dataset of defendants charged with a felony or misdemeanor at the time of arrest. This includes defendants that were not released pre-trial whose pre-trial re-arrest outcomes are by definition undefined, as they had no opportunity to be re-arrested. Figure 2 shows the percent of defendants of each race group that are assigned to each category. Here, we see that a bit over 40% of White defendants were assigned the lowest risk score as compared to about 22% of Black defendants. At the higher end of the spectrum, 11% of Black defendants were assigned the highest risk score, whereas about 5% of White defendants received that score. This shows the potential for this model (if adhered to by judges) to result in a disproportionate availability of the supervised release program for White defendants as opposed to Black defendants. These numbers are in exact agreement with Exhibit 20 of the original methodology report, and thus we only reproduce the results here for completeness as this was already part of the original analysis of the method conducted by CJA. *Based on these results, the risk assessment also does not achieve statistical parity by race.*

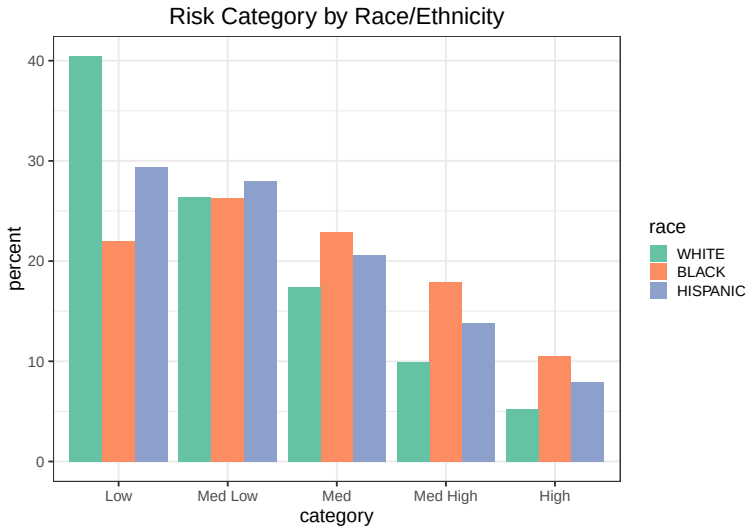

FIGURE 2. Percent of defendants, by race, assigned to each risk group.

[16]This definition of fairness has been argued for in the algorithmic fairness literature generally [Kamiran and Calders, 2009; Calders and Verwer, 2010; Feldman et al., 2015], including in the context of risk assessment [Johndrow and Lum, 2019].

3.2. **Procedural definitions of fairness.** Several additional approaches to fairness have been proposed that indicate that fairness is achieved so long as a particular procedure is followed. The simplest of these, which has come to be known as "fairness through unawareness" [Kusner et al., 2017; Grgic-Hlaca et al., 2016], is the requirement that the model not explicitly have access to information on the race variable. Though this definition is largely used as a strawman in the modern algorithmic fairness literature, as it has been shown to have some shortcomings [Pedreshi et al., 2008], the tool analyzed here meets this definition.

[Corbett-Davies et al., 2017] argue that a risk assessment model maximizes social welfare so long as the same thresholding rules are applied to all, regardless of race. That is, a model in this framework would meet this definition if the cut points that translate predicted probabilities to recommendations are the same for everyone regardless of race. In the model considered here, it is clear that this standard is met. However, this argument is made in the context where the cut points are determined based on a cost-benefit analysis, where those individuals whose expected cost of detention (presumably in terms of the cost of the facilities to house them during the pre-trial period) outweigh the expected cost of their future crimes committed if released. In developing the cut points, the methodology report states that they were determined by selecting the set of cut points with the best AUC. Thus, although the same cut points are applied to all, the manner in which those cut points was determined is not in line with the notion of fairness implied under this framework.

Finally, one procedure that has been suggested as a measure for "anti-discrimination" is to include the sensitive variable (race) in the regression modeling to avoid omitted variable bias. Then, to remove the effect of race, predictions are made by averaging across the race distribution [Pope and Sydnor, 2011]. This procedure is very similar to that used in the development of the supervised release model to account for variation in the length of the pre-trial period. To integrate this approach into the existing framework, we fit the same model as CJA, now including the race variable as one of the covariates. We then compare the predictions of the original model (which does not include race) to the predictions generated by fitting the same model including race as a covariate and averaging across the race distribution as in [Pope and Sydnor, 2011]. Although the race variable is statistically significant when included in the regression model, the predictions generated under the two methods are highly correlated. For example, if we assume that the same percentage of defendants would be classified as high risk under the two different methods, the racial composition of the high risk group under this method for accounting for race effects is not substantially different from that in the original implementation. Under that same assumption, the race-marginalized model performs qualitatively similarly to the original model.

4. Conclusion

We report on several common measures of fairness as applied to the supervised release risk assessment tool. In evaluating the composition of the charges in the training data and the historical context in which the data were collected, it seems likely that the data itself encodes racial bias. As discussed, this complicates evaluation of the model because many definitions of fairness depend upon a comparison of the model's predictions to the 'truth'. When the data measure the truth in

a racially biased way, measures of fairness that rely on this comparison may inadvertently obscure or understate the extent of racial bias present in the model. One potential remedy for this problem would be to select training (and evaluation) data that is less likely to be generated by biased enforcement, perhaps by using a more recent time period or by focusing on arrest types with less discretion in enforcement. Substantive knowledge about the policing practices generating the arrests that make up a potential dataset will be necessary to determine if this is even possible.

We also find that the tool does not comply with any of the common group-wise parity-based metrics except 'accuracy equity'. In practical terms, this has several different interpretations. In this case, the lack of compliance with predictive parity means that, for example, a White individual who is classified as high risk is actually less likely to be re-arrested than a Black or Hispanic person who is classified as high risk. This could be interpreted as bias *against* White people. Alternatively, a model that exhibits lack of predictive parity in this way could be justified if the model developers believe that the felony re-arrests decisions were made in a manner that was racially discriminatory, thus artificially inflating the measured likelihood of re-arrest for people of color. That is, by categorizing White people with slightly lower empirical risks of felony re-arrests as similar to people of color with higher risk of felony re-arrest, this could be seen as a correction for the over-policing of Black and Hispanic people.

Lack of racial equality in false positives in this case indicates that a Black person who does not go on to be re-arrested for a felony is about 2 times as likely to be ineligible for supervised release (or classified in one of the higher risk categories) than a White person who does not go on to be re-arrested for a felony. Relative to White people who are not re-arrested, Hispanic people who are not re-arrested are roughly 1.5 times times more likely to experience ineligibility for supervised release or be classified in one of the higher risk categories. This could be interpreted as bias against Black and Hispanic people because 'innocent' Black and Hispanic people are more likely to be denied access to supervised release or be described as at high risk of re-arrest than White people.

Finally, differences by race in the rate at which individuals–regardless of eventual outcome– are classified into the various risk categories indicates a lack of racial statistical or demographic parity and potential issues with disparate impact. In this case, Black defendants were about twice as likely as White defendants to be made ineligible for the supervised release program based on the risk assessment. Hispanic defendants were about 1.5 times as likely to be ineligible as White defendants. Thus, this tool has the potential to disproportionately impact communities of color relative to White communities by denying access to a potentially beneficial program at a higher rate.

Ultimately, whether a risk assessment is considered 'fair' is dependent on the definition of fairness used, and different definitions could result in different conclusions regarding *to whom* the risk assessment model is unfair, if at all. Attempts to compromise between various notions of fairness could result in a model similar to the one analyzed here, where nearly none of the group-wise parity-based definitions are met. Put differently, the results we have seen could also arise if the authors chose to 'trade off' some parity in predictive accuracy in order to get closer to achieving other competing fairness measures or to incorporate other legal or moral

considerations. However, no mention of explicitly accounting for these sorts of considerations was included in the methodology report, so it is not clear whether this was the intent of the authors.

Finally, this report and the model developed do not directly address whether there is any correspondence between the risk categories and the extent to which individuals might benefit from the supervised release program. It is implicit in the implementation of the tool that those in the highest risk group are unsuitable for the supervised release program, but no evidence within the original report or here addresses whether this is true. People who are deemed high risk of felony re-arrest are excluded from the program, yet it is possible that this group of people would benefit from the program as much if not more than those who were already at low risk of re-arrest. If it is true that the program could have the greatest effect on the highest risk group, withholding access to the program based on risk alone would be suboptimal from the point of view of reducing felony re-arrests. These are questions that can't be answered with the existing data and are not addressed by this report, the existing risk assessment tool, or any tool that focuses solely on predictions of risk without an accompanying analysis of the efficacy of the program for all individuals who might be evaluated by the tool.

Acknowledgement

We are grateful to CJA for providing data, code, and the methodology report that allowed us to do this work. We also thank Logan Koepke for helpful comments. This work was supported by the The Ethics and Governance of AI Initiative, the MacArthur Foundation, and the Open Society Foundation.

References

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. *ProPublica*, 2016. URL `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Allen J Beck and Alfred Blumstein. Racial disproportionality in us state prisons: Accounting for the effects of racial and ethnic differences in criminal involvement, arrests, sentencing, and time served. *Journal of Quantitative Criminology*, 34(3): 853–883, 2018.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 2018.

New York (State). Civil Rights Bureau and Eric T Schneiderman. *A Report on Arrests Arising from the New York City Police Department's Stop-and-frisk Practices*. Office of the NYS Attorney General, Civil Rights Bureau, 2013.

Shawn Bushway, Brian D Johnson, and Lee Ann Slocum. Is the magic still there? the use of the heckman two-step correction for selection bias in criminology. *Journal of quantitative criminology*, 23(2):151–178, 2007.

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

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

Sarah L Desmarais, Kiersten L Johnson, and Jay P Singh. Performance of recidivism risk assessment instruments in us correctional settings. *Psychological Services*, 13(3):206, 2016.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe, 2016.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Sorelle A Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.

Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law, Barcelona, Spain*, volume 8, 2016.

Bernard E Harcourt. Risk as a proxy for race: The dangers of risk assessment. *Fed. Sent'g Rep.*, 27:237, 2014.

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.

Eoin Healy. Research report for MOCJ's pretrial felony re-arrest risk assessment tool. Technical report, New York City Criminal Justice Agency, October 2015.

James J Heckman. Sample selection bias as a specification error (with an application to the estimation of labor supply functions), 1977.

Michael J Hindelang. Race and involvement in common law personal crimes. *American sociological review*, pages 93–109, 1978.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *The Annals of Applied Statistics*, 13(1):189–220, 03 2019.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6. IEEE, 2009.

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

Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pages 4066–4076, 2017.

Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

Devin G Pope and Justin R Sydnor. Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, 3 (3):206–31, 2011.

Cindy Redcross, Melanie Skemer, Dannia Guzman, Insha Rahman, and Jessi Lachance. New york city's preterial supervised release program: An alternative to bail. Technical report, MDRC and Vera Institute of Justice, 2017.

Jennifer L Skeem and Christopher T Lowenkamp. Risk, race, and recidivism: predictive bias and disparate impact. *Criminology*, 54(4):680–712, 2016.

Appendix A. Description of the model

This section gives an overview of the methods implemented by CJA in building the discussed risk assessment model. All information is summarized from [Healy, 2015], which was provided by CJA to help this project. SPSS code as well as anonymized training data was also provided by the original authors for replication purposes.

The population included in the dataset used to create the model is composed of individuals with prosecuted Summary (custodial) arrests in first half of 2009 with an adjudicated outcome as of June 30, 2011, or in which the case had any pretrial FTA. The unit of analysis is an individual. To avoid including the same individual multiple times in the dataset, each person's first custodial arrest is included in the dataset and subsequent arrests during the time period are dropped. People who only had arrests related to Desk Appearance Tickets in the first half of 2009 were also excluded from the data. This data was then matched to data held by DCJS to include information about each individual's re-arrests. Individuals whose cases had not reached a final disposition and had not had a warrant ordered by the final date of the study (June 30, 2011) were dropped from the training data. An individual was considered to have had a felony re-arrest for the purposes of this dataset if he was arrested for a felony offense during the pre-trial period, defined to be the period from arraignment to plea or disposition. The dataset also contains information about the individual's criminal history, demographic information, etc.

The data was randomly split into two parts: one to be used for model fitting (the 'analytic sample') and one to be used for model validation (the 'validation sample').

Logistic regression models were fit to the analytic sample in which the dependent variable was felony re-arrest and various combinations of criminal history and demographic variables were included as independent variables.[17] Also included as covariates in the regression were three 'control variables': the number of days at risk (daysatrisk2) and an estimated probability of appearing in the dataset (PROBRELEASEDPRE + PROBCONTINUED). The former variable is meant to control for variable length of time during which an individual could be re-arrested such that defendants whose cases required more time to reach a final disposition would not be at an artificial disadvantage due to the additional time during which an arrest could have been made relative to a defendant whose case was disposed quickly. The latter variables are an attempt at using a Heckman correction [Heckman, 1977], which is a technique for accounting for selection bias– the variability across defendants in likelihood of appearing in the training data (recall that those who were not released pre-trial cannot be used to train the model as we do not know if they would have been re-arrested pre-trial, causing systematic under-representation in the data of the types of defendants who are less likely to gain pre-trial release).

Unfortunately, despite the prevalence of its use in the criminology literature in exactly this setting, the Heckman correction is not a valid method for accounting for selection bias in logistic regressions [Bushway et al., 2007]. Furthermore, this model fails to implement the Heckman correction correctly in two ways. In this case, the probability of selection into the training data decomposed into two separate probabilities– an estimated probability that the defendant was released pre-trial

[17]When variables are first introduced, their names in the provided dataset are given in parentheses.

unconditional on whether the case was continued and an estimated probability that the case was continued beyond arraignment. Each of these probabilities range approximately from zero to one. We interpret the documentation to mean that the sum of these two probabilities is meant to represent the overall probability of inclusion in the dataset. However, the sum of these two probabilities range from approximately 0.3 to 2, indicating that the sum of the two variables is not a valid estimate of the probability of selection into the dataset, as estimated probabilities of selection should not be greater than one.[18] The Heckman correction requires that the inverse Mills ratio of the probability of inclusion in the dataset be included as an independent variable in the regression model. Even if the probabilities were estimated reasonably, the estimated probabilities are included in the regression model in their native scale–that is, the model includes the estimated probabilities themselves, not the inverse Mills ratio of the probabilities, as would be required to perform the Heckman correction. Thus, despite the authors' accurate observation that it is important to account for selection bias, there is no theoretical or empirical justification for the method used to correct for it.

Regardless, the authors arrive at a final logistic regression model that includes, in addition to the above mentioned control variables, the following variables: a categorical age variable consisting of age groups 16-19, 20-29, 30-39, 40+ (AGE_EH); a binary indicator of first arrest (FIRSTARREST); a binary indicator of whether the defendant had open cases at the time of arrest (open1_eh); a binary indicator of whether the defendant was engaged in full-time caregiving, employment, or training activity(FTactivity); a binary indicator of whether the defendant had a previous warrant in the last four years (DCJSPWorOW4YR); a binary indicator of whether the defendant had a misdemeanor conviction in the past nine years (MCONV1_YN1); a binary indicator of whether the defendant had any felony convictions in the previous 9 years (FCONV9_YN); and, a binary indicator of whether the defendant had a drug conviction in the past nine years (DRCONV9_YN). All included variables except identified control variables are categorical, and sum contrasts are used for each as a way of parameterizing the model such that the coefficients for each categorical variable are constrained to sum to zero.

The result of fitting this logistic regression model is a set of regression coefficients, each one representing a 'weight' to be applied to each of the levels of the factors included in the model. These coefficients are, in general, not integers. In order to transform the model's coefficients into integers, each coefficient is first divided by the smallest (in magnitude) statistically significant coefficient (at the $\alpha = 0.05$ level) and then rounded to the nearest integer. Coefficients that are not statistically significant are set to zero. The result is an integer point value associated with each level of each of the factors included in the model.

These point values make up the scoring model, and a defendant's score is calculated by tallying the number of points they receive.

[18] An alternative interpretation is that each of these probabilities are components of the probability of selection into the dataset, and a sort of double Heckman correction is applied by including both in the regression equation. This interpretation is supported based on the fact that each term has its own coefficient in the regression. The following critiques remain under this interpretation as well.

Appendix B. Reproduction of CJA's findings

We have translated the provided SPSS code to R. We first filter the data according to the rules provided for inclusion in the sample: the arrest must be for a felony or misdemeanor charge, the defendant must have been released pre-trial, and the case must have concluded by the end of the study period or a warrant had been issued.

We then fit the following model, where preFEL is the indicator variable that takes value one if the defendant was re-arrested for a felony offense prior to the case disposition and zero otherwise.

```
glm.mod <- glm(preFEL ~ AGE_EH + open1_eh + FIRSTARREST +
                    FTactivity + DCJSPWorROW4YR + MCONV1_YN +
                    FCONV9_YN + DRCONV9_YN + daysatrisk2 +
                    PROBRELEASEDPRE + PROBCONTINUED,
               contrasts = list(AGE_EH = "contr.sum",
                                open1_eh = "contr.sum",
                                FIRSTARREST = "contr.sum",
                                FTactivity = "contr.sum",
                                DCJSPWorROW4YR = "contr.sum",
                                MCONV1_YN = "contr.sum",
                                FCONV9_YN = "contr.sum",
                                DRCONV9_YN = "contr.sum"),
               data = data, family = binomial(link = "logit"))
```

The output of this code is a set of coefficients or 'weights', one for each level of the categorical variables except one. For example, this outputs one weight for each age category except the final category, 40+. The weight for the final level is set such that the sum of the weights across all levels of a given variable sum to zero. The results of applying this procedure to each of the analytic sample, validation sample, and full dataset are shown in Table 6. The coefficient estimates exactly match those given in Exhibit 2 of [Healy, 2015].

We then derive the point values associated with each level of the variables by applying the described rule: divide all coefficients by the smallest (in magnitude) coefficient that has an associated p-value less than 0.05. As in the original methodology, the intercept and the control variables are excluded from receiving point values. This results in the following point values associated with each model, shown in Table 7.

The point values for the final model shown in Exhibit 2 of [Healy, 2015] *do not* correspond to the point values derived from any one of the datasets. As noted in [Healy, 2015], the point values used in the final, deployed model are a 'compromise Risk Point system,' which appears to be a 'compromise' across the point values derived from each of the three different samples. The report describing the final point values used in the supervised release tool did not fully explain how a decision was made regarding which of the three possible point values for each factor ought to be used. Specifically, while most of the final point values are those derived by applying the methodology to the validation set, we believe the point value for the second age category was taken from the 'total' model, and the point value for the oldest age category is a compromise of averaging the point values from the analysis set and the validation set. Another possible interpretation is that for all factors

	Analysis		Validation		Total	
	Estimate	Sig	Estimate	Sig	Estimate	Sig
Age: 16 to 19	0.588	***	0.539	***	0.565	***
Age: 20 to 29	0.052		0.078	.	0.064	*
Age: 30 to 39	-0.341	***	-0.221	***	-0.281	***
Age: 40 and older	-0.300	***	-0.395	***	-0.348	***
Open Cases: No	-0.152	***	-0.099	**	-0.126	***
Open Cases: Yes	0.152	***	0.099	**	0.126	***
First Arrest: No	0.286	***	0.268	***	0.277	***
First Arrest: Yes	-0.286	***	-0.268	***	-0.277	***
Fulltime Activity: No	0.147	***	0.156	***	0.152	***
Fulltime Activity: Yes	-0.147	***	-0.156	***	-0.152	***
Warrant (4yr): No	-0.140	***	-0.106	**	-0.123	***
Warrant (4yr): Yes	0.140	***	0.106	**	0.123	***
Misd. conv. (1yr): No	-0.176	***	-0.197	***	-0.186	***
Misd. conv. (1yr): Yes	0.176	***	0.197	***	0.186	***
Felony conv. past (9yr): No	-0.102	*	-0.085	*	-0.093	**
Felony conv. past (9yr): Yes	0.102	*	0.085	*	0.093	**
Drug conv. past (9yr): No	-0.127	**	-0.207	***	-0.167	***
Drug conv. past (9yr): Yes	0.127	**	0.207	***	0.167	***

TABLE 6. Coefficient estimates and associated significance codes for model as applied to each subdivision of the data. Only the coefficients used to calculate risk points are reported. P-value significance codes are as follows: $p < 0.001$ ***, $p < 0.01$ **, $p < 0.05$ *, $p < 1$.

except age, the final point values were chosen to be the most common value or the median value across the three candidate point values. Ultimately, because this more subjective part of the model development was not documented, it is not possible to know how or why the specific point values that make up the final model were chosen.

This is a non-standard way of arriving at a final model. An example of a more standard methodology would be to first derive point values using only the analysis set. Then, one would verify that the point values derived from the analysis set perform satisfactorily when applied to the validation set. For example, one might check that the point values are nearly as predictive in the validation set as they are in the analysis set. If so, one would then apply the same procedure used to initially derive the point values to the total dataset and use the point values derived from the total dataset as the final model. Picking and choosing point values across the different 'folds' of the data risks creating a model that assigns too-high scores to some defendants relative to others that is not justified by differences in risk of re-arrest as estimated from the data or by explicit fairness-based constraints.

It is difficult to assess whether points based on a single model fit on the full dataset, as opposed to the hand-selected compromise points taken from models fit on different partitions of the data, would have been substantially different in terms of the qualitative results of the fairness evaluation. This is because the fairness analyses rely on mapping the point scale to risk categories. This is done by

	Analysis	Validation	Total	CJA Reported
Age: 16 to 19	6	6	9	6
Age: 20 to 29	0	0	1	1
Age: 30 to 39	-3	-3	-4	-3
Age: 40 and older	-3	-5	-5	-4
Warrants (4yr): No	-1	-1	-2	-1
Warrants (4yr): Yes	1	1	2	1
Drug Conv. (9yr): No	-1	-2	-3	-2
Drug Conv. (9yr): Yes	1	2	3	2
Felony Conv. (9yr): No	-1	-1	-1	-1
Felony Conv. (9yr): Yes	1	1	1	1
First Arrst: No	3	3	4	3
First Arrst: Yes	-3	-3	-4	-3
Fulltime Activity: No	1	2	2	2
Fulltime Activity: Yes	-1	-2	-2	-2
Misd. Conv. (1yr): No	-2	-2	-3	-2
Misd. Conv. (1yr): Yes	2	2	3	2
Open Cases: No	-1	-1	-2	-1
Open Cases: Yes	1	1	2	1

Table 7. Derived point values for each variable. Dark gray cells indicate a match with the final score used in the supervised release risk assessment model. Light gray indicates values that were likely averaged to arrive at the final point value.

defining 'cut points' such that, for example, all individuals with less than negative nine points are considered 'low risk'. How these cut points were determined isn't fully detailed, and so it is difficult to assess whether using a model-derived point scale and associated cut points would have resulted in different risk categorizations and, ultimately, different conclusions regarding which notions of fairness the model does and does not meet.

We can, however, look at how the raw point scores changed as a result of hand-selected deviations from a baseline model, where we take the baseline model to be that developed using the more standard procedure described above. Figure 3 shows the distribution of the difference in raw point values between a model developed by applying the statistical models directly to the 'total' dataset and the model developed by hand-picking the desired point values, disaggregated by various demographic variables. Positive values here indicate that the compromise point values are higher than they would have been under the baseline model. Negative values here indicate that they are lower. The blue line shows the median difference between the two models. The center bar within each box shows the median difference for each group. Relatively speaking, at the median the adjustments to the model most benefit young, Black, and male defendants.

Appendix C. Potential issues with current model implementation

Based on reproducing the New York City supervised release model, we have identified several potential issues with the methodology used to fit the model. As already noted in previous sections, the final point values were not derived from a

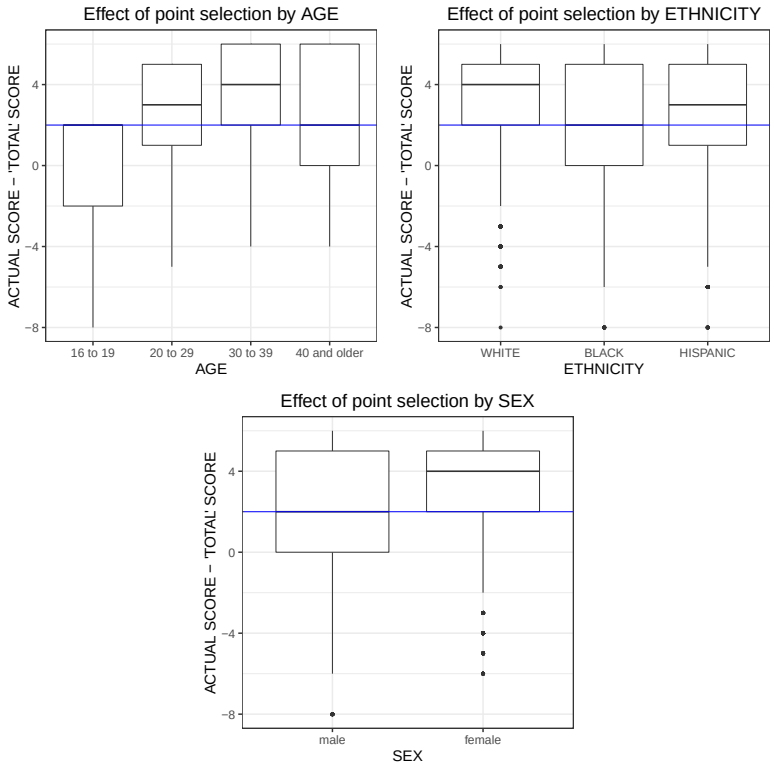

FIGURE 3. Boxplot of differences between raw scores of the 'compromise risk score model' and the statistical model fit to the 'total' dataset, disaggregated by defendant age, race/ethnicity, and sex.

single model but rather were the result of combining point values from different models in an ad hoc way. Further, the Heckman correction is not theoretically motivated to use in this setting and was not implemented correctly even if it were. Finally, we interpret the inclusion of the days at risk variable as being meant to control for variation in the time during which each defendant could have been arrested. We believe a survival analysis would be more methodologically justifiable in this setting rather than simply including it in the regression if one wants to control for that effect.

It is also worth reiterating the amount of discretion that MOCJ exercised in developing the model and implementing the supervised release tool. MOCJ defined felony re-arrest as the outcome to be predicted, and as we've shown, there is evidence that this is a racially biased measure of someone's impact on public safety. Additionally, CJA tested a number of models for MOCJ using different variables, but CJA did not ultimately choose the model on which the tool was based. MOCJ also made the decision to group scores into five risk categories. Furthermore, MOCJ may have provided input on the final risk point values through an undocumented process, rather than deriving them directly from fitted models [Healy, 2015]. Each of these decisions impact the ability of individuals to participate in the supervised release program, and thus merit close scrutiny. Though risk assessment models are

often promoted as 'evidence-based', the reality of how this model was developed leaves us with a model that is at best 'evidence-informed.'