The Effect of Ethnicity on Recidivism Risk Scores in US Courts

Fabian Beigang

May 2020

1 Introduction

Over the last years, a growing interest emerged around the topic of unfair bias in algorithmic decision support systems. It gained particular popularity when COMPAS, an algorithm used in US courts to determine defendants' risks of reoffending, came under scrutiny by a group of journalists. This case engendered a general debate about what it exactly means that an algorithm is unfairly biased.

A number of different formal definitions of fairness have been put forward and discussed. However, none of the as of yet presented notions seems to appropriately capture all the intuitions about which algorithms should count as fair and which not. In this paper, I want to introduce a new notion of fairness which is based on causal inference methods. It formalizes the idea that discriminatory bias means that an attribute such as ethnicity or gender makes an unjustified difference to the algorithmic classification. After introducing the notion, I will conduct a case study, analyzing the famous COMPAS data set according to this definition of algorithmic fairness. To this end, I will estimate the causal effect of ethnicity on the algorithmic risk assessment, and in parallel, the causal effect of ethnicity on actual recidivism rates, using distance based matching.

2 Background

2.1 The COMPAS case

In 2016, ProPublica, an investigative journalism organization, published an article (Angwin et al., 2016) in which the authors aimed to show that the algorithmic decision support systems used in criminal sentencing in some US states exhibit significant racial bias. In their analysis, they focussed on a software solution called COMPAS, developed and distributed by the company Northpointe, which calculates the risk that a given defendant will recidivate. In other words, it predicts the probability that the defendant will commit another crime within the next two years after trial. The score is based on a detailed questionnaire that the defendants have to complete.

The analysis of the COMPAS algorithm found that while the overall accuracy of the predictions were roughly the same for black (67%) and for white (69%) defendants, the two groups differed significantly in their respective error rates: The false positive rate for black defendants was 46%, while only 23% for white defendants - indicating that black defendants were twice as often incorrectly classified as future recidivists; at the same time, the false negative rate for black defendants was 28%, while 48% for whites - which means that white defendants that would actually go on to commit another crime after trial were much more likely to nevertheless receive a low risk score. This was taken to show that COMPAS' predictions were biased against blacks, and that they could potentially result in discriminatory sentencing decisions.

Northpointe contested the claims (see Flores et al., 2016), arguing that the algorithm is not unfairly biased since it can be shown to be well-calibrated by group. This means that for both groups, blacks and whites, of those defendants that were assigned x% risk of recidivating by the algorithm, a proportion of roughly x% did indeed turn out to recidivate. Calibration by group has commonly been taken to be a desired property of risk assessment instruments (Pleiss et al., 2017). The idea behind it is that if a predictor would not yield results that are calibrated by group, the predicted probability estimate would not have a consistent meaning across different demographic groups. But does calibration alone suffice for an algorithm to be fair?

A large debate ensued. It was argued that calibration by group may be a necessary, but certainly not sufficient condition for fairness. On the other hand, ProPublica's implicit definition of fairness as similar error rates did not go uncontested either. It was, for instance, shown that lower false positive rates (of one demographic group) can be achieved by lowering the quality of the predictor for that group, which seems undesirable from the standpoint of that group (Corbett-Davies et al., 2017). In addition to arguments from the potential manipulability of the equal error rates requirement, it was maintained that the definition is conceptually inadequate - the mathematical definition does not appropriately capture the more commonsensical meaning of the notion of algorithmic fairness.

2.2 Fairness as causal adequacy

In this paper I want to analyze the COMPAS algorithm according to a different notion of fairness. The underlying idea of this notion is that discriminatory bias is present when a sensitive attribute (such as ethnicity, gender, or disability) makes an unjustified difference to the algorithmic prediction. In other words, if a sensitive attribute is a cause of the algorithmic prediction, while it is not a cause of the actual state of the world. To illustrate this with two examples, think of an algorithm that makes a prediction about whether someone will likely have a car accident over the course of their life - for instance as part of the process of obtaining a driving license. If the applicant's ethnicity makes a difference to the prediction, it would count as unfair according to this notion, since ethnicity does not have a causal effect on the actual rate of accidents. On the other hand, if the applicant's disability (say, vision impairment) makes a difference to the prediction, it would not count as unfair, as a severe vision impairment will arguably have a causal effect on the person's rate of accidents.

To express the foregoing more formally in the language of potential outcomes, we need to introduce a number of definitions. Let Y denote the actual state of the world that an algorithm is supposed to predict, and $\hat{Y}$ the algorithm's prediction of Y. Let D denote the sensitive attribute relative to which we want to evaluate the algorithm's fairness. We can then define the causal effect of the sensitive attribute on the algorithmic prediction as:

$$\alpha_{\hat{Y}|D} = Pr(\hat{Y}_{D=1} = 1) - Pr(\hat{Y}_{D=0} = 1)$$

And the causal effect of the sensitive attribute on the actual state of the world as:

$$\alpha_{Y|D} = Pr(Y_{D=1} = 1) - Pr(Y_{D=0} = 1)$$

With these two definitions at hand, we can express our criterion of algorithmic fairness as causal adequacy more precisely: the causal effect of the sensitive attribute D on the algorithmic prediction $\hat{Y}$ cannot be greater than the actual causal effect of D on the corresponding actual state of the world Y. Formally:

$$\alpha_{\hat{Y}|D} \leq \alpha_{Y|D}$$

This formalization of algorithmic fairness is in line with arguments about the ethical evaluation of discrimination. A common line of argument is that discrimination is wrong because it fails to treat people as individuals, but rather treats them as members of a specific group. Properties or past behaviors of other members of the group are projected onto an individual merely due to their membership in that group. A decision is then discriminatory if it is based on the property of being a member of a social group, while the property of belonging to that group is actually irrelevant for the decision in question. In the driving license case mentioned above, for instance, ethnicity is irrelevant while a disability might not, and consequently the former counts as discriminatory while the latter does not. Irrelevance can here be interpreted as a causal notion:

irrelevant properties are those that do not possibly have a causal effect on the actual outcome.

Let us return to the COMPAS case. The sensitive attribute D here is clearly *ethnicity*, the relevant state of the world Y is *future criminal behavior (recidivism)*, and the algorithmic prediction $\hat{Y}$ is the *recidivism risk score*. There is a broad literature on racial disparities in criminal behavior. Generally, the evidence supports the thesis that there is no causal link between ethnicity and crime - when other factors such as exposure to violence and socioeconomic status are controlled for, the correlation between ethnicity and crime vanishes (see, e.g., Aliprantis, 2017, Ulmer et al., 2012). Hence, it seems there is no scientific basis on which we could assume that being African-American would influence the likelihood of recidivism. We can thus make the following assumption for our analysis:

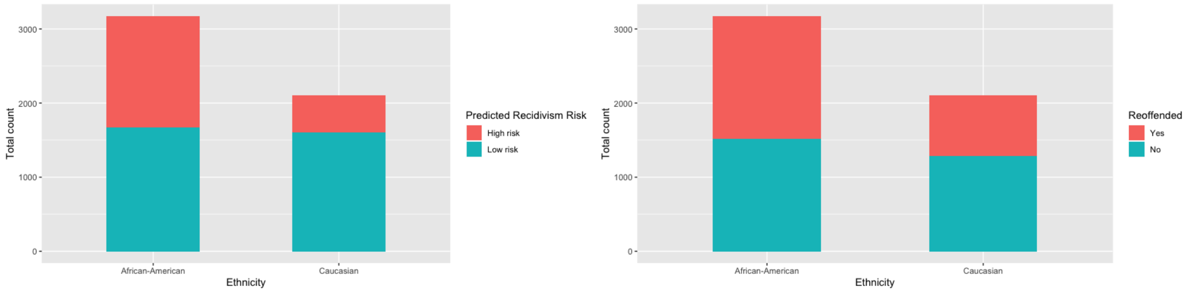
$$\alpha_{Y|D} = 0$$

In order to determine whether the COMPAS algorithm is unfair relative to our notion of *algorithmic fairness as causal adequacy*, we consequently only need to determine $\alpha_{\hat{Y}|D}$. This is what we will attempt in the following sections.

3 The data set

We will base our analysis on the data set that ProPublica has made available[1]. To create the data set, ProPublica merged COMPAS scores they received from the Broward County Sheriff's Office in Florida with public criminal records from the Broward County Clerk's Office website. The resulting data set contains 7214 entries, each representing one defendant. The features of interest to our analysis are *age*, the *charge degree* (which takes values "M" for misdemeanor, or "F" for felony), *ethnicity*, the *number of prior convictions*, the *risk score* assigned by the COMPAS algorithm, and whether they *actually recidivated* within two years

[1] `https://github.com/propublica/compas-analysis`



(a) Predicted recidivism risk

(b) Actual recidivism

Figure 1: Proportion of outcomes by ethnicity

after the trial.

In line with ProPublica's analysis, we removed a number of rows from the data set. First, those in which the date of the COMPAS evaluation was more than 30 days after the arrest, as this could indicate that the recorded COMPAS score is not for the recorded crime the defendant was arrested for. Second, those for which there was no COMPAS risk score. Third, those for which the charge was an ordinary traffic offense. Since in this analysis we want to focus on the difference between being black and white, we removed all those cases of defendants whose ethnicity entry was neither "Caucasian" nor "African-American". This leaves us with 5278 entries.

These 5278 defendants can be divided into 3175 African-American and 2103 Caucasian defendants, 4247 male and 1031 female defendants. Of the defendants, 2002 received a high risk score (above five on the one to ten scale); 2647 did in fact recidivate. We will interpret a risk score of above five as the prediction that a defendant will recidivate within two years.

Of the African-American defendants, 1506 were categorized as high risk (i.e. above 5) by the COMPAS algorithm, and 1661 did actually recidivate. Of the Caucasian defendants, 496 were categorized as high risk, and 822 did recidivate.

If we look at how high risk scores are distributed among different age categories, we notice that for those below 25, more than half of the defendants

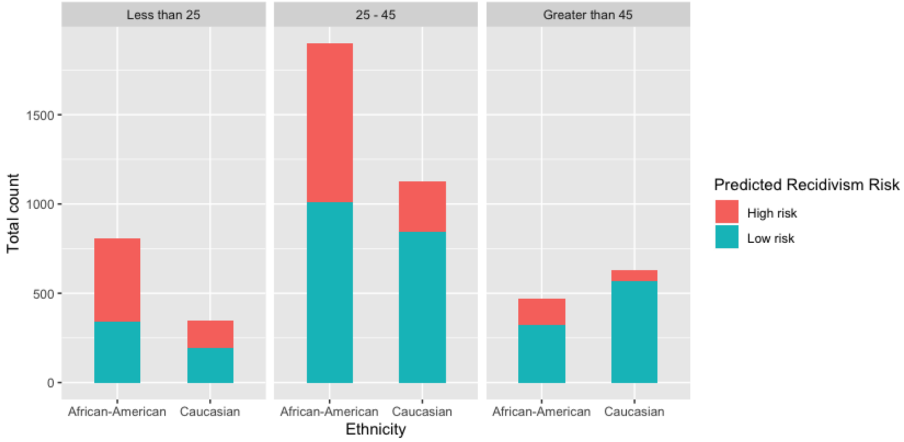


Figure 2: Proportion of predicted recidivism scores by ethnicity and age group

received a high risk score by the COMPAS algorithm, while for those between 25 and 45 it was still a significant portion - about one third - while for those above 45 it is only a relatively small fraction. This distribution is similar for actual recidivism.

If we look at risk scores and actual recidivism by ethnicity while controlling for age, we find that young and middle-aged African-Americans are the group that had the greatest proportion of high risk scores (see Figure 2). This is also the case for actual recidivism (see Figure 3). But, comparing the two box plots, it is striking that especially in the middle-aged group (25-45 years) the proportion of Caucasians that received a high risk score is lower than the proportion that actually recidivated, while the reverse is true for African-Americans, which have a higher proportion of high risk scores than actual recidivists in the middle-aged group.

To get a more precise sense of the disparities between African-American and Caucasian defendants in terms of risk scores and actual recidivism, we will perform two t-tests in order to assess whether the seeming disparities are due to chance or whether we can assume that there are underlying systematic differences which drive the observed results. The t-test for a high risk evaluation

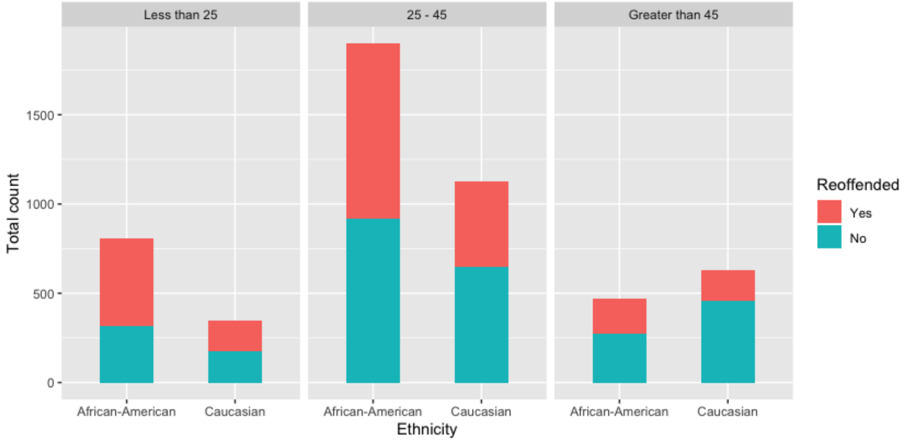


Figure 3: Proportion of actual recidivism by ethnicity and age group

yields a difference in proportion of 0.24. The 95% confidence interval for the difference in proportion of high risk evaluations is from 0.21 to 0.26. On the other hand, the t-test for actual recidivism yields a difference in proportion of 0.14 (with a 95% confidence interval from 0.12 to 0.17), which is significantly lower than the difference in proportion of a high risk evaluation. In words, this means that the proportion of defendants that are categorized as high risk is 24% higher in the group of African-American than it is in the group of Caucasian defendants. However, the proportion of African-Americans in our sample that actually recidivate is only 14% higher than the proportion of Caucasians.

Note that the above differences result from comparisons in which we did not control for any other variables. Running a multivariate logistic regression for both, high risk evaluation and actual recidivism, while controlling for age, sex, number of priors, and charge degree, yields that for an average African-American defendant, the probability of a high risk evaluation is 11% higher than for a Caucasian defendant, but the probability of actually recidivating is only about 1% higher.

4 Analysis

We begin the analysis by graphically summarizing our intuitive or evidence backed causal judgments about the relevant variables in the data set (Figure 4). We are interested in the potential causal link from *ethnicity* to the *risk score*. To contrast the causal effect of ethnicity on the risk score, we will also analyse the effect of *ethnicity* on *actual recidivism*, which, however, is not represented in the causal graph. We assume there is an unobserved variable (in the graph in light grey), which is a common cause of ethnicity, count of prior crimes, and the severity of charges. Without specifying exactly what the variable "Background circumstances" exactly refers to, it could describe something like the family an individual is born into, which might determine or at least partially influence genetic factors, socioeconomic level, exposure to violence, et cetera. Further, we draw links from *age* and *sex* to *charge degree* and *number of prior convictions*. This seems reasonable, as there is evidence that both sex (e.g. Mawby, 1980) and age (e.g. Ulmer and Steffensmeier, 2014) have an effect on criminal behaviour. We do not draw a link from *ethnicity* to either *number of prior convictions* nor *charge degree*, as scientific studies indicate that if one controls for the right background variables, the correlation between ethnicity and crime rates vanish (e.g. Ulmer et al, 2012).

Based on the COMPAS handbook[2], we know that sex, age, and criminal behavior are taken into account by the COMPAS algorithm, and hence we can assume they potentially have a causal effect on the risk prediction. While ethnicity is not explicitly recorded in the data on which the prediction is based, there may be many proxies for ethnicity in the data. This makes it possible that ethnicity makes a causal difference to the risk prediction without being explicitly recorded in the data set.

Further potential confounding factors for the risk prediction could include socioeconomic status and previous exposure to crime, which are not included in the data set. However, assuming that the number of priors and charge degree

[2] `http://northpointeinc.com/files/technical_documents/FieldGuide2_081412.pdf`

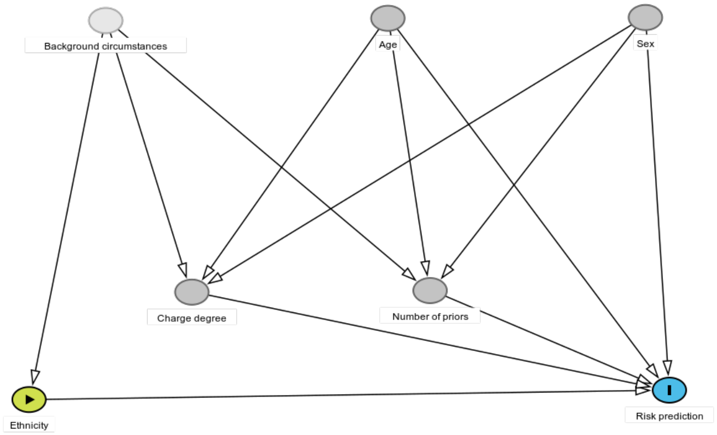


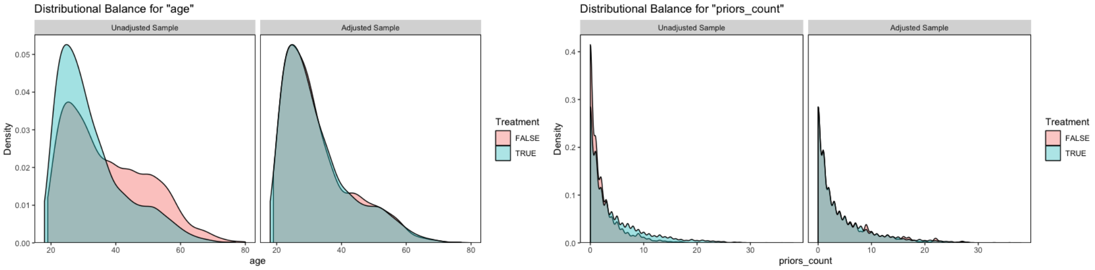
Figure 4: Causal graph

are sufficiently correlated with these potential confounders, and that they can be hence seen as proxies, justifies the assumption that we have sufficiently precise observations of all the factors that influence the risk prediction. With this assumption, our problem lends itself to an analysis via matching.

Given the causal graph above, it seems plausible to include the four variables *age*, *sex*, *number of prior convictions*, and *charge degree* in the set of covariates in order to create a matched control group. Since our set of covariates is not highly dimensional, we will use a distance based matching method, using the Mahalanobis distance.

We define being of "African-American" ethnicity as the treatment, and being of "Caucasian" ethnicity as the control. Since the number of observations in the treatment group is significantly greater than the number of observations in the control group, we will use a matching procedure with replacement. This means, some of the Caucasian defendants will be used more than once in estimating the causal effect.

We will try 1:1 and 1:2 matching. Depending on which of the two methods yields the better balance between samples, we will make our choice of matching



(a) Age

(b) Count of prior convictions

Figure 5: Balance of covariates after 1:1 matching

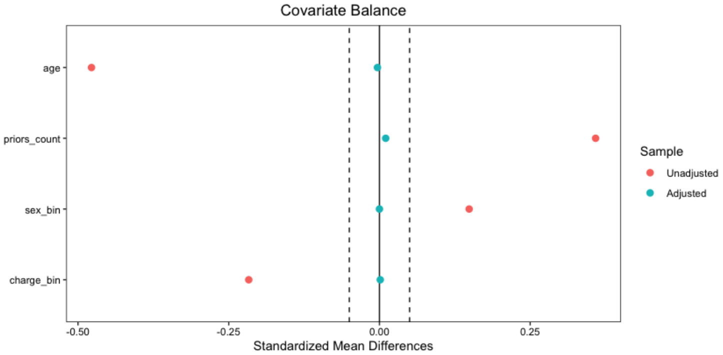
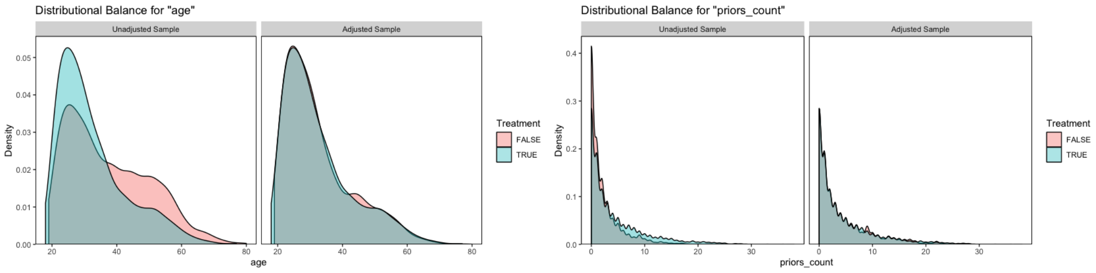


Figure 6: Summary covariate balance after 1:1 matching

method for estimating the treatment effect. We will first analyze the adjusted sample using 1:1 matching.

For *sex* and *charge degree*, the plot indicates that we achieved a (close to) perfect balance, and also for *age* (see Figure 5a) and *number of prior convictions* (see Figure 5b) it looks like the balance the 1:1 matching achieves is sufficient. If we look at the standardized difference in means for the adjusted sample, we see that with *sex* we indeed have a perfect match, with *charge degree* we have a negligible difference of 0.0006, for *age* 0.0033, and for *priors count* 0.0104. This is by all standards a very close match. The numbers are summarized in Figure 6.

Next, we will perform 2:1 matching and see whether we can improve the bal-



(a) Age

(b) Count of prior convictions

Figure 7: Balance of covariates after 1:2 matching

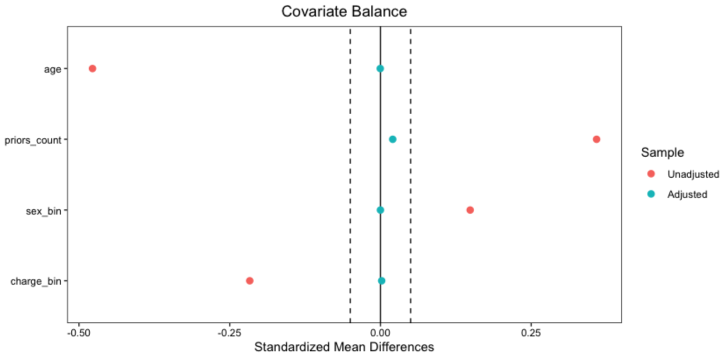


Figure 8: Summary covariate balance after 1:2 matching

ance achieved by 1:1 matching. Again, the diagrams indicate that our adjusted sample matches the treatment group relatively closely.

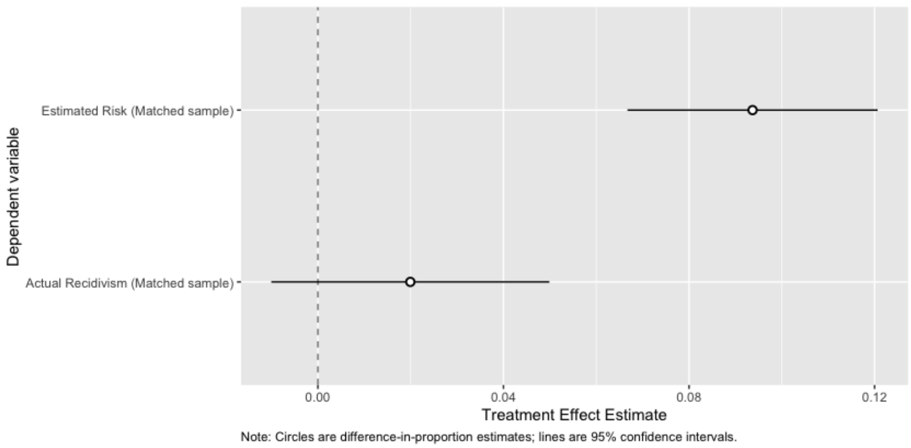
More precisely, there is no difference in means for *sex*, and a negligible one for *charge degree* (0.0009), the stadardized difference in means for *age* has gone down a tiny bit (to 0.0003) but the difference in means for the *number of prior convictions* has doubled to 0.0203. Since a good balance on the number of prior counts seems desirable, we are consequently going to use 1:1 matching for our estimation of the causal effect.

Before we begin the estimation of the causal effect, we have to address the questions whether we should estimate the average treatment effect (ATE) or the average treatment effect for the treated (ATT). The latter would be a good choice in those cases, in which there naturally is more interest in those cases in which the treatment condition is present. An example of this sort would be smoking: intuitively, what a study on effects of smoking intends to assess is what would have been the case if the smoker hadn't smoked. It is less interesting to ask what would have happened if a given non-smoker would have smoked. In our case, however, we are interested in both counterfactuals. How would the risk evaluation have differed, if a given person would have been of Caucasian rather than African-American ethnicity? And, equally relevant, how would the risk evaluation have differed, if a given person would have been of African-American rather than Caucasian ethnicity? Estimating the ATE can be problematic if there is not sufficient overlap between the treatment and the control group. This, however, is not the case in our sample, as the foregoing analysis has shown.

The estimated difference in proportion we obtain as a result from our treatment and matched control group for a *high risk evaluation* is 0.094. In other words, being African American makes it almost 10% more likely to be assessed as high risk as compared to being Caucasian by the COMPAS algorithm. The standard error of our estimate is 0.014. The p-value is well below the 0.01 level of significance. Given that our identification assumptions hold, we can conclude that the ethnicity of a defendant does indeed have a significant effect on the COMPAS risk of recidivism evaluation.

To check how robust our results are with regards to unobserved confounders, we conduct a sensitivity analysis using Rosenbaum bounds. The upper bound of the p-value remains below the 0.01 level of significance up to a gamma value of 1.7 - that is, we would only change our conclusion if there were an unobserved characteristic that is associated with high risk scores and that is 1.7 times more common among African-Americans rather than Caucasian defendants. While this shows that our causal estimate is somewhat sensitive to the presence of unobserved confounders, the result that conclusions are only valid up to such a confounding level is not uncommon in the social sciences.

To contrast, the estimated difference in proportion from the adjusted sam-



Note: Circles are difference-in-proportion estimates; lines are 95% confidence intervals.

Figure 9: Comparison causal estimates

ple for actual recidivism is only 0.020. The standard error of this estimate is 0.015, and the p-value is 0.19. This means the result is not significant at any conventional level of significance. We have to conclude that our analysis does not establish a causal link between ethnicity and recidivism. In other words, being African-American rather than Caucasian does not make one more likely to reoffend, and vice versa. This is in line with scientific evidence on racial disparities in crime behavior.

5 Discussion

The above analysis confirms the hypothesis we set out to investigate, namely the claim that the COMPAS algorithm is racially biased. It shows that higher risk scores on average for African-American defendants cannot be explained away as mere correlations that come about through differences in other factors (as for instance different age distributions). Being African-American makes a difference to the risk assessment provided by the algorithm. The evidence, however, does not lend credence to the claim that being African-American makes any difference whatsoever to whether someone actually ends up reoffending.

Our investigation supports ProPublica's hypothesis that the COMPAS recidivism risk predictions are not fair, yet it does so from a different perspective than ProPublica's own analysis. Applying *causal adequacy* as the fairness definition produces, at least in principle, a more robust assessment of racial bias than merely comparing different error rates between ethnic groups. The reason for this is that we aim at identifying the cause of the disparity in outcome, rather than just observing a correlation of a specific ethnicity with higher or lower error rates. Hence, on our account unfairness means that ethnicity explains[3] the disparity in predictive outcomes, while it does not explain actual outcomes. Comparing error rates, on the other hand, allows for the scenario that there is an unobserved factor which incorrectly drives the predictions in one direction, and which happens to be correlated with being of a specific demographic group.

In the COMPAS case, this could be the number of prior crimes committed by a defendant. Suppose the COMPAS algorithm would make its prediction only on the basis of prior crimes committed. A higher number of prior crimes certainly justifies the prediction that a defendant is more likely to reoffend in the future. Now imagine that the African-American group of our sample happened (for whatever reason) to on average have a higher count of prior crimes on their record. But imagine further that (for whatever reason) none of the defendants reoffended. Intuitively, the algorithm's predictions do not seem to be racially biased, despite different error rates for different demographic groups. On ProPublica's definition, this case would count as unfair, whereas on the definition outlined in this paper, it wouldn't, since the disparity would not be explained by the defendant's ethnicity but by the number of prior convictions.

Yet, this robustness comes at a price. The assumptions we have to make in order to conduct the analysis of whether an algorithm satisfies causal adequacy are stronger than to check for equal error rates. This is particularly problematic when the full data set on which predictions were based is not available, because then there is no guarantee that the crucial *selection on observables* assumption holds. A potential confounder we did not control for is socioeconomic

[3] We are taking explanation to mean a causal explanation along interventionist lines, see Woodward (2005)

status. Our analysis implicitly relied on the assumption that charge degree and number of prior crimes is sufficiently correlated with socioeconomic status, such that matching on these variables also yields balance on socioeconomic status. Hence the validity of our fairness evaluation depends on the plausibility of this assumption.

6 Conclusion

In this paper we have shown that the COMPAS algorithm, which is used in US courts in order to assess a defendant's risk of recidivism, is racially biased against African-Americans as compared to Caucasian defendants. We first outlined and argued for the definition of unfair bias as causal inadequacy, meaning that a sensitive attribute such as ethnicity or gender has a stronger causal effect on the predictor of an event, than on the event itself. In order to check whether the algorithm was unfair according to causal adequacy, we applied Mahalanobis distance-based 1:1 matching to determine the causal effect of ethnicity on COMPAS risk scores, as well as actual recidivism rates. The analysis yielded that there was a significant causal effect of ethnicity on risk scores, but no significant effect on actual recidivism rates.

References

[Aliprantis, 2017] Aliprantis, D. (2017). Human capital in the inner city. Empirical Economics, 53(3), 1125-1169.

[Angwin et al., 2016] Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias. ProPublica. Retrieved from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[Corbett-Davies et al., 2017] Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017, August). Algorithmic decision making and the cost of fairness. In Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 797-806).

[Flores et al., 2016] Flores, A. W., Bechtel, K., and Lowenkamp, C. T. (2016). 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. Fed. Probation, 80, 38.

[Mawby, 1980] Mawby, R. (1980). Sex and crime: The results of a self-report study. The British journal of sociology, 31(4), 525-543.

[Pleiss et al., 2017] Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. (2017). On fairness and calibration. In Advances in Neural Information Processing Systems (pp. 5680-5689).

[Ulmer et al., 2012] Ulmer, J. T., Harris, C. T., and Steffensmeier, D. (2012). Racial and ethnic disparities in structural disadvantage and crime: White, Black, and Hispanic comparisons. Social Science Quarterly, 93(3), 799-819.

[Ulmer and Steffensmeier, 2014] Ulmer, J. T., and Steffensmeier, D. J. (2014). The age and crime relationship: Social variation, social explanations. In The nurture versus biosocial debate in criminology: On the origins of criminal behavior and criminality (pp. 377-396). SAGE Publications Inc.

[Woodward, 2005] Woodward, J. (2005). Making things happen: A theory of causal explanation. Oxford university press.