# Methods & Results

Yijia Lyu

May 2020

## Abstract

The algorithm COMPAS, which is used by many courts in USA to assess the recidivism likelihood of defendants, has been accused of racial bias. This paper validates COMPAS's classification and threshold setting in defining 'low', 'medium', 'high' risks and tries to select proper threshold for each racial group so that the tool can improve its algorithm fairness whilst having good predictive performance. By examining the accuracy, FT, FP, the result shows that this tool is indeed relatively biased towards the black defendants and selecting different thresholds can improve accuracy fairness but the false predictions are still unavoidable.

**Keywords:** Recidivism, Race, Bias, Prediction

# 1. Methods and Data

## 1.1 Key Definitions

• Recidivism

Following the result of a recent recidivism study by the U.S. Sentencing Commission (2019) that most recidivists commit a new crime within the first two years after their release, we define recidivism in this study as a new arrest in two-year period. With this definition, we validate whether the COMPAS scores of criminal suspects in our dataset correspond with their criminal behaviors in the next two years.

• High, Medium, Low Decile Score

According to COMPAS, all scores range from 1 to 10. Decile scores between 1 to 4 are classified as low scores, scores between 5 to 7 are labelled as medium scores, and scores between 8 to 10 are regarded as high scores. We use this classification to verify the algorithm bias when we look at the raw data. But when in the threshold selection section, we define high risk as score greater than the threshold value.

• Algorithm Fairness

Statistical parity and calibration are two key definitions when assessing algorithm fairness, but they are not often compatible. Statistical parity means that statistical features such as classification error, false positive rate, false negative rate, precision, etc., should be equal across groups when classified with protected attributes. This may not be satisfied at the same time, in this paper, we first focus on the equal error rate. Calibration means that predictive outcomes should be independent of protected attributes like demographic features, for example, race or gender. In the context of pretrial, we refer calibration to be same probability of true recidivism across different racial groups among defendants who are given with a reoffend risk score after release. A score s = s(x) is calibrated if it satisfies the following equation:

$$\Pr(Y = 1 \mid s(X), X_p) = \Pr(Y = 1 \mid s(X)).$$

## 1.2 Data Sources

As the goal of this study is to assess recidivism algorithm fairness of COMPAS and improve model fairness towards different racial groups, we select data from Broward County in Florida whose jurisdiction is known to use COMPAS risk evaluation tool to decide pretrial release. Florida is also a state that has strong open-records laws. To start with, we use records for 11757 criminal suspects who were assessed at the pretrial stage and scored for recidivism risk using the COMPAS in Broward County during 2013 and 2014. The dataset is publicly available, including information of person id, assessment id, case id, agency text, last name, first name, middle name, sex code, ethnic code, date of birth and other nineteen variables. We use jail data and public incarceration data from 2013 to 2016, provided by Broward County Clerk's Office website, to match with previous records using identifiable variables like name, case id, etc.

## 1.3Data Processing

With datasets above, we are able to build a criminal history dataset. However, some records are not correctly matched for a number of reasons: (1) The basic information of defendants were entered with spelling or numeric mistakes; (2) There were no corresponding crime charge cases to COMPAS scores. (3) Recidivism information is less than two years. The error rate is close to 3.8% with confidence interval to positive and negative 1.75%, we remove mismatching records and merge the matched ones into a final dataset with population of 7214 defendants.

## 1.4 Methods

The COMPAS tool assess criminals in three aspects: Risk of Recidivism, Risk of Violent Recidivism and Risk of Failure to Appear. In this paper, we focus on the racial disparity and accuracy for the overall risk score of recidivism only.

We first validate the predictive scores of COMPAS with actual recidivism of different racial groups to see if the scores are classified correctly and if the scores are accurate in assessing one's likelihood of recommitting crimes. Based on the initial result, we will use threshold bias correction model that improves algorithm fairness (calibration and statistical parity). This on the other hand verifies if COMPAS's uniform classification on low, medium, high risks is reasonable. Many risk assessment tools compare true risks $\hat{U}(x)$ with the risk score $s(x)$ and set the corresponding $d(x) = 1$ when it satisfies the condition that $\hat{U}(x) >= t$, 't' here represents an appropriate threshold which we adjust in different context. In this study, 't' is a threshold that

differentiate high-risk recidivists. But note that the error rates differ across different groups with certain attributes, which needs to be reduced. We will then perform calculations to compare the confusion matrix of each threshold and find the proper threshold for each racial group to improve algorithm fairness, as concepts formally defined by Hardt et al. (2016) and Kleinberg et al. (2016). This study does not focus on reducing one particular false predictions (FN/FT), for the trade-off between false positives (defendants who do not recommit crimes but classified as high risks) and false negatives (defendants who are classified as low risks but recommit crimes) is another problem that should be carefully discussed in terms of jail resources and social threat.

# 2. Results and Discussions

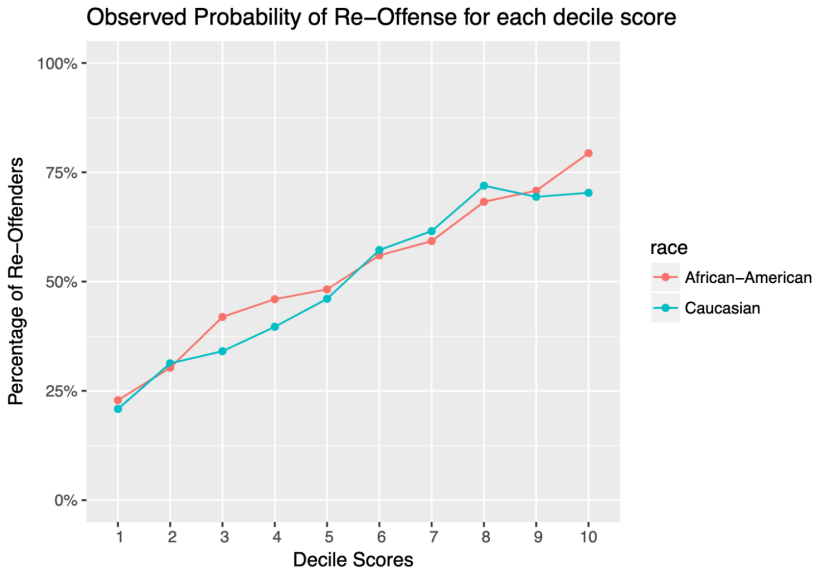
## 2.1 Results

2.1.1 Raw data overview

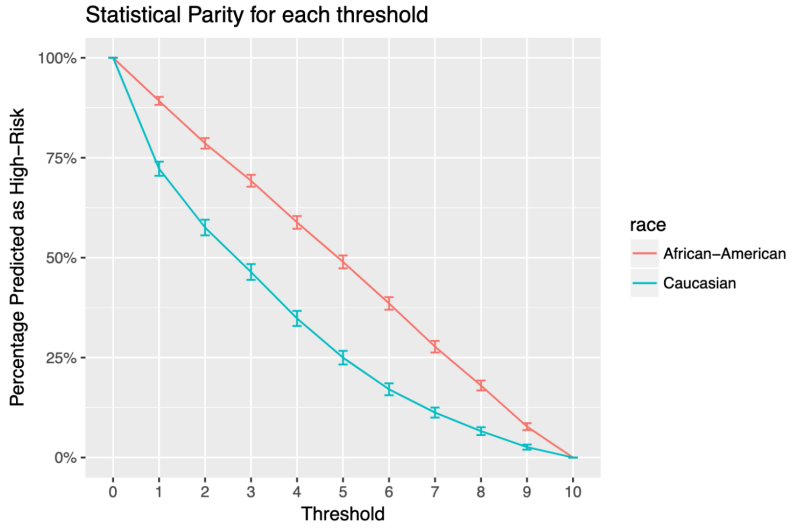
| decile_score | African-American | Asian | Caucasian | Hispanic | Native American | Other |
|---|---|---|---|---|---|---|
| 1 | 398 | 15 | 681 | 196 | NA | 150 |
| 2 | 393 | 4 | 361 | 113 | 4 | 66 |
| 3 | 346 | 5 | 273 | 86 | 1 | 36 |
| 4 | 385 | NA | 285 | 52 | 1 | 46 |
| 5 | 365 | 1 | 241 | 52 | NA | 22 |
| 6 | 384 | 3 | 194 | 37 | 2 | 21 |
| 7 | 400 | 1 | 143 | 34 | 4 | 10 |
| 8 | 359 | 2 | 114 | 26 | 1 | 10 |
| 9 | 380 | NA | 98 | 20 | 2 | 8 |
| 10 | 286 | 1 | 64 | 21 | 3 | 8 |

(Figure1 COMPAS score distribution: predicted results)

The graph above shows that in general, COMPAS gave decile scores from 1 to 10 to the black defendants fairly averagely while it gave scores to the white defendants with high discrepancy. Among the 2454 white defendants, 35.4% of them were classified as ‘high/medium risk’. In comparison, 58.8% of the 3696 black defendants were labelled as high/medium, suggesting a higher likelihood of recidivism than other races.




(Figure 2 Calibration check - actual recidivists proportion on each decile scores)

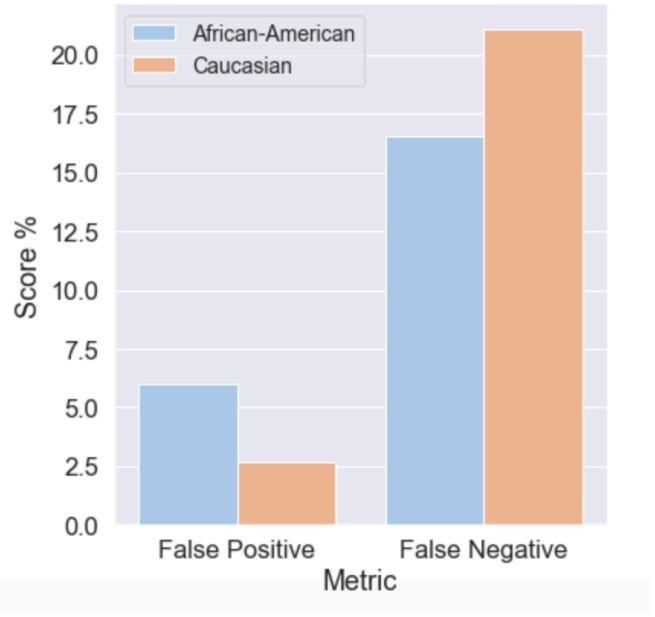



(Figure 3 Statistical parity check)

Then we plot the calibration graph and statistical parity prediction graph for each racial group. Figure2 examines the calibration and demonstrates that current COMPAS system does not well calibrate. On each decile score level, prediction accuracy on black and white defendants differs from each other. The differences at score=3, 4, 10 is huge. Figure3 checks statistical parity by comparing the probability Pr(S) that the algorithm classifies a defendant with a score (S) above certain decile scores. It must be the same with the probability for African-Americans (X = a), as it is for Caucasians (X = c). But the result is as follows:

$$Pr(S > 4 \mid X = a) = 0.59 > Pr(S > 4 \mid X = c) = 0.35$$
$$Pr(S > 7 \mid X = a) = 0.28 > Pr(S > 7 \mid X = c) = 0.11$$

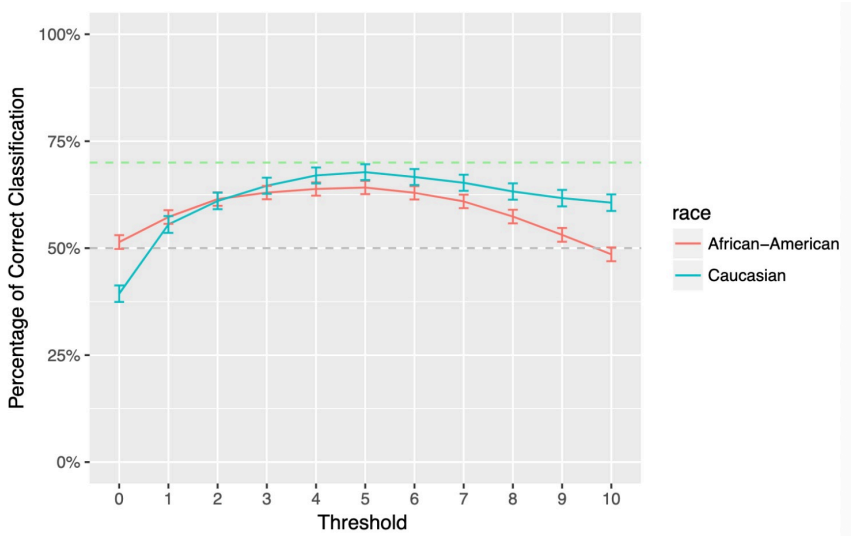
This means that COMPAS does not satisfy the algorithm fairness.




(Figure 4 FP and FN on different racial groups)

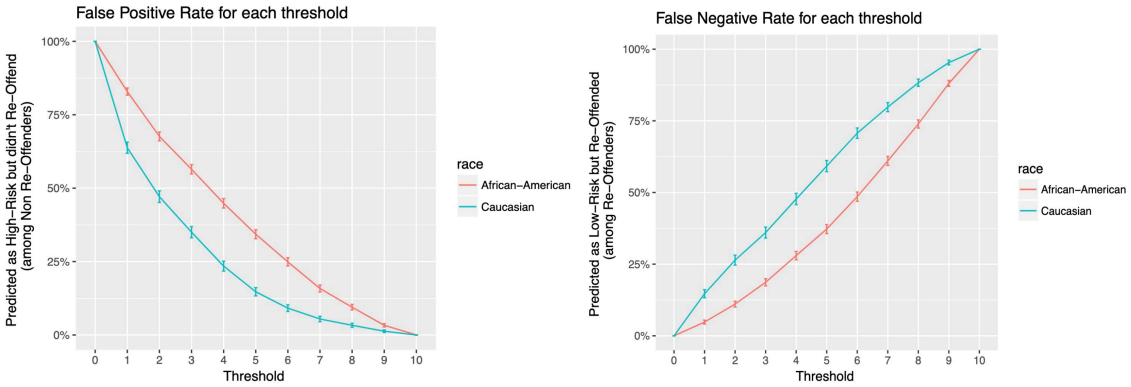
Figure 4 further shows that the black population experiences both a higher false positive rate and lower false negative rate. Its FP is almost twice the rate of the white and its FN is 25% less than Caucasians. This implies that COMPAS underestimates the recidivism likelihood of white re-offenders but overestimate the recidivism probability of the black population. This metric is a clear indicator that the scoring system is biased to different race groups.

2.1.2 Threshold Selection




(Figure 5 Accuracy and threshold for two race groups)

The x-axis of the plot represents different threshold from 1 to 10, and the y-axis represents the accuracy. The graph shows that the accuracy for African-American is better than Caucasian when threshold is under 3, whereas from threshold 4 onwards, the accuracy for Caucasian increases and performs better than African American. There is also substantiate difference in accuracy at threshold above 8. The result shows that for both the two groups, threshold=5 has the highest accuracy. The predictive accuracy when threshold = 5 for the black population is closest to the accuracy when threshold = 3 for the white.




(Figure 6 FP and FT for each threshold for the two race groups)

However, the results of false positive rate and false negative rate for each threshold shows that even if we re-evaluate the threshold to classify risk categories for different race groups and find new threshold values that maximize the predictive accuracy to the two groups on the same level,

the COMPAS tool systematically imposes high FN and low FT on the black population whilst giving high tolerance on the white defendants.

## 2.2 Discussion

This study verifies the racial bias in the COMPAS algorithm and uses threshold rule for bias correction by looking for proper threshold for each racial group that minimizes bias. However, algorithm bias in COMPAS is reflected in multiple ways. First, it sets the same threshold classification for each racial group, leading to unequal accuracy. Second, the two types of predictive errors, false positives and false negatives, are particularly concentrated on the black population. Following the idea of calibration and statistical parity when examining algorithm fairness, we test the predictive accuracy, FP, FT for each threshold of the African American and the Caucasian. Although we find that threshold = 3 for the white and threshold = 5 for the black might be almost equal in satisfying the predictive accuracy to be highest and close, our goal to achieve similar FP and FT cannot be realized at the same time. Threshold re-selection individually may not fully meet our expectation on achieving fairness. It needs to be supplemented with other methods that correct the bias in COMPAS. For example, we also try to pre-classify different race groups and test several machine learning methods like Lasso, Naïve Bayes, and Random Forest on each group. However, since this paper focuses on correcting bias in COMPAS based on the discrepancies between its predictions and facts, discussing machine learning methods would be another story and we would like to conduct subsequent related research in another paper.

## Reference:

Broward County Clerk's Office. Available at: https://www.clerk-17th-flcourts.org/

Hardt, M., Price, E., and Srebro, N (2016). *Equality of opportunity in supervised learning.* CoRR, abs/1610.02413.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. Inherent trade-offs in the fair determination of risk scores. CoRR, abs/1609.05807, 2016.

United States Sentencing Commission (2019). *Recidivism Among Federal Offenders: A Comprehensive Overview*. Available at: https://www.ussc.gov/research/research-reports/recidivism-among-federal-offenders-comprehensive-overview