Université Paris-Dauphine
Université Paris Sciences et Lettres
Cycle Pluridisciplinaire d'Etudes Supérieures

June 26^{th} 2020

Mémoire de recherche

What is a fair algorithm ?

TABET GONZALEZ Salwa

Acknowledgements

First, I would like to acknowledge for the help and kindness of my supervisor Mr. Alexis Tsoukiàs. He gave me my first insight of the research world in Computer Science and Algorithmic Decision Theory, and more generally allowed me to make my first steps in research by giving me the time and patience that I needed.

I would also like to thank the CPES Computer Science dean Mrs. Virginie Gabrel-Willemin for her support all throughout the academic year, and specially during the second semester. Thanks to her, I am now convinced of wanting to pursue a career in Computer Science, despite all the doubts that I could have this year.

Contents

Introduction

Many structures in our surroundings today use algorithms to make decisions: college admissions, credit allowances, access to facilities are typical cases of such applications. Less typical cases exist, such as assessing the risk of a defendant to recidivate in a two-year window after being released. The question of "fairness" of these decision-making algorithms arise.

This memoir aims to explore the existing Computer Science and Law literature with respect to studies conducted about fairness in a specific case: the COMPAS algorithm.

1 Algorithmic fairness

1.1 Definitions of fairness

The concept of "algorithmic fairness" is not clearly established, despite numerous attempts in the literature. There exist different definitions of fairness, corresponding to different legal approaches. The use of any of these mechanisms depends on the context: this means that it is impossible to achieve a universal definition of fairness and that one has to assess whether a certain mechanism is suitable for a specific situation. Therefore, the goal of being aware of these different notions of fairness is to be able to choose the right tool that matches best the situation.[1]

Notion	Sub-notion	Corresponding Legal Mechanism
Individual fairness	The unaware approach	Equal opportunity as colorblindness
	Fairness through awareness	Equal opportunity based on similarities, and levels of scrutiny
Group fairness	Decoupling	Affirmative action (as separate but equal)
	Statistical or conditional parity	Affirmative action (preferably through critical diversity)
	Equal opportunity	Affirmative action (as equal opportunity)
	Equalized odds	Achieving equlity by equalizing the false positive and false negative errors
	Calibration	Achieving equality by statistical significance
	Multicalibration	Achieving equality by statistical significance, and accounting for intersectionality
Causal Reasoning	Counterfactual fairness	Disparate treatment and disparate impact analysis

Table 1: Notions of fairness and summary of their corresponding legal mechanisms [1, p.9,46]

We will give a definition and an example for each sub-notion shown in the table. All subnotions aim to be fair towards the individual, but they give a different weight to each characteristic and therefore may have different outcomes.

1.1.1 Individual fairness

The aim of this notion of fairness is to address fairness towards the individual regardless of the group he/she belongs to. This corresponds with the principle of equal opportunity and it is the easiest notion to understand, since it is quite intuitive. It works by finding similarities and disparities between individuals, without focusing on any particular characteristics.

Its goal is to consider the individual as the main object without considering its group affiliations. Therefore, individual fairness answers to an individualized justice notion and is the key to equality before the law. This notion is well-established in many legal systems, such as the American legal system. Ideally, by focusing on the individual and blurring some of its characteristics, we would be able to limit bias against the members of a certain group (minorities for race and gender, for example). Still, it is very difficult to separate an individual from his group as certain characteristics may act as a proxy: for example, a Black individual would be more likely to live in the poorer areas of the city, while a White individual would be more likely to live in the suburbs.

The unaware approach

This approach refers to giving zero constraints for fairness. According to this approach, the algorithm should be blinded (or unaware) of any identifiable factors and protected attributes by law such as gender, race and sexual orientation[2]. It is built on the principle of meritocracy: according to it, everyone can succeed with equal amounts of work and talent. Therefore, meritocracy rejects any form of group affiliation, which is seen as a denial of individuality and merit. Yet, Colorblindness allows to ignore racial and cultural issues and gives a false illusion of fairness. In fact, the unaware approach ignores differences and characteristics that lead to different outcomes, so it is not able to rectify bias already present in the data[3, p.45]. In addition, there are empirical studies that show that this approach does not give a suitable form of fairness since it ultimately helps maximizing the profit of the dominant group and further marginalizes already disadvantaged groups [4].

Fairness through awareness

This approach seeks to treat similar individuals similarly [5]. In order to define whether two individuals are similar or not, Dwork assumes a distance metric between them and other individuals: that is why it is called ”fairness through awareness”. Its purpose is to make sure that if a pair of individuals are considered similar by a human, then the distance between them and each one of the compared individuals is similar. This implies that there is a need to assess whether two individuals are similar or not and this depends on human perceptions and beliefs. The metric is hence developed with the help of experts in the specific domain that the algorithm will evolve in. Because its goal is to compare every pair of individuals in the dataset, fairness through awareness falls into the individual fairness category and is not to be confused with group fairness. Nonetheless, even if this form of fairness seems to be a suitable answer for fighting against discrimination, it is very difficult to identify which characteristics are relevant for the metric and which categories are needed to classify similar individuals.

1.1.2 Group fairness

The aim of this notion of fairness is to achieve fairness by being aware of the individual's group affiliations, because historical discrimination has shown that different groups have different outcomes for the same situation. This corresponds with the notion of affirmative action. This notion is quite difficult to understand since it aims to achieve fairness by discriminating, generally against the majority (i.e., affirmative action). Unlike individual fairness, group fairness does explicitly highlight some characteristics such as gender, race or age group in order to choose the mechanism that would give the fairest outcome.

Definition 1. *(Affirmative action)[6] Legal mechanism that aims to improve the position of historically disadvantaged minorities in the society by prioritizing them in resource allocation.*

Definition 2. *(Affirmative action)[1, p.21] Treating groups that face differently discriminatory conditions outside the context differently, so as to achieve outcomes within the context that are less tainted by the discriminatory treatment those groups face outside of the context.*

To successfully address affirmative action, the decision-maker needs to fully understand the situation and the discrimination at play. This leads to an increased mistrust of affirmative action because it is seen as ineffective and questionable. For instance, legal scholars argue about the importance of such mechanisms in reinforcing

diversity [7] and their violation of traditional values such as merit, as stated before.

Decoupling

This approach consists in multiplying the decision-making processes to cater to different groups. Its goal is to rule out the possibility of making the algorithm more suitable for the dominant party, thus increasing fairness. It works by acknowledging the fact that the features and their weight might differ between groups to have the same outcome. However, this approach has its issues: for instance, it is difficult to decide which groups need to have a specialized tool and which do not. Also, for some protected attributes (such as race), this approach might be dangerous: it would bring back to the Segregation Era's "separate but equal" notion. In this aspect, segregation had a negative impact on the quality and level of services and goods provided to blacks in the separated premises [8]. The notion of intersectionality is overlooked here too: an individual could belong to different minorities thus be discriminated against on several grounds. Finally, this approach could constitute a disparate treatment on the basis of race since the decision-making process would use features whose predictive power is different among different groups [9, p.38-39].

Statistical or conditional parity

This approach consists in equalizing the fraction of people having a certain outcome and the fraction these people represent in the general population. For example, since women represent 50% of the general population, then 50% of college admissions should be for women. Conditional parity is a special form of statistical parity, based on equalizing the odds of having a certain outcome with respect to different features such as age, race and gender. However, this approach is highly questioned because the common perception is that this notion of fairness allows the creation of quotas and the positive discrimination of minorities on the expense of the dominant party (and its more qualified members). Also, it is not clear what does the general population mean: it could be the national population, the applicant pool, the general population in a certain field for job applications, and so on. Choosing a definition over another could perpetuate bias instead of fighting against it: for example, women are underrepresented in male-dominated fields and therefore are less likely to apply for a job. Choosing to build quotas based on the application pool would not be of any help in fighting for more women representation in those fields.

Equal opportunity

This approach aims to favor the group of individuals who really belong to the true positive class. This means that it equalizes the opportunity to be classified as the positive class for those who truly belong to this group. It is a weaker form of equalized odds: for example, the decision-maker would make sure that the people who pay back their loan have an equal opportunity of getting the loan in the first place, without specifying any special treatment for those who would ultimately fail to pay it back [4].

Equalized odds

Equalized odds aims to achieve equal opportunity for the positive and negative classes simultaneously. In other words, its goal is to equalize false negative and false positive rates, therefore ensuring that the decision-maker erroneously categorizes individuals in a similar manner across groups (such as races or genders) and that the percentage of erroneous predictions is also equal. Nevertheless it is difficult to achieve such fairness because of several issues. First, the actual cost and impact of false negatives and false positives is different: actual recidivists categorized as low-risk have a heavier impact on society than false high-risk defendants, but the impact on the individual's life is heavier for the latter [9]. Second, it is difficult to predict whether an individual will be a false negative or a false positive. Therefore, it is almost impossible to predict what would have happened if we had assigned said individual the positive outcome instead of the negative, and vice-versa. This is called a counterfactual outcome and it is very difficult to estimate [10]. Third, by equalizing error rates the decision-maker is changing the predictive values across group, thus reinforcing mistrust on the decision-making process.

Calibration

Theoretically, a calibrated decision-making process is fair within any given score category that it creates. In other words, its probabilities carry semantic meaning [1, p.35]. Calibration answers to a need of equal opportunity at least, and also that among the positive class the odds of actually belonging to it are equalized across all groups [11]. Therefore, calibration ensures that we can treat similarly two individuals with the same score with respect

to the actual outcome [12]. This method is well-perceived as significant in the public eye [13], but calibration and other notions of fairness may collide together and be mutually exclusive: for example, it is mathematically impossible to achieve calibration, equal false positive rates and equal false negative rates between two groups [11].

Multicalibration

Multicalibration corresponds with calibration plus accounting for intersectionality. It aims to reach a better balance between group fairness (calibration) and individual fairness (intersectionality, making the individual unique). It is achieved by defining a metric for calibration and giving up on equalizing false positive and false negative rates [14]. Multicalibration operates by computing different thresholds without any prior information of hypothetically discriminated groups: therefore, it will try to adapt its calibration on groups that it is able to detect, even if they were not previously detected by the user. However, the choice of features considered by the decision-maker directly influences the outcome of the decision-making process, thus it is easy to include bias in it. Also, if the subset of a certain group is too small, it might be hard for decision-maker to successfully calibrate this group.

1.1.3 Causal reasoning

The aim of this notion of fairness is to focus on the causal relationship between the factors (or features) and the outcome (or decision). This corresponds with the mechanism of due process that will be defined later. It works in a similar manner of group fairness by highlighting some characteristics of the individual. However, it assesses which factors have a higher correlation with the outcome in order to focus on these and include them in the decision-making process.

Counterfactual fairness

In general, causal reasoning-based approaches focus on including only the features that have been proven to cause a certain outcome in the decision-making process [15]. It eliminates superfluous and potentially error-inducing factors with a high correlation. Mathematically, this approach aims to achieve counterfactual fairness, which identifies the factors that can actually cause discrimination and sort out the effects of said factors. This process is executed by creating a hypothetical world in which a minority individual would belong to the dominant party, and then assessing which outcome it would have in this world. Since it is a very complex task to rule out all the proxies that can cause bias, this approach addresses all the relationships between the attributes and intervenes by assigning different values to each attribute in order to create the counterfactual world [16]. However, this approach also bears some drawbacks: on one hand, it is very difficult to identify the causal variables and only experts in the specific domain of the decision-making could do so. On the other hand, the conclusions cannot be immediate based on the counterfactual world. In fact, the counterfactual world does not directly and explicitly inform the user of the individual's behavior in the real world.

1.2 Formal definitions of algorithmic fairness

As an increasing number of decisions are being controlled by artificial intelligence, the quest of a fair algorithm is a key problem nowadays. The popularity of such algorithms is based on the belief that because they can take into account much more features than humans and suppress perception bias, they make fairer decisions. However, this is not always the case since even the data used to fit and train the models can be biased, therefore perpetuating historical forms of bias and discrimination.

All the previously shown forms of fairness were mathematically defined in order to use them for algorithmic purposes [17]. We will only include the easiest definitions, since very-specific fairness definitions are better introduced in other papers.
Let S the protected attribute (e.g., race or gender), $S = 1$ the dominant party, and $S \neq 1$ the unprivileged group. $\hat{Y} = 1$ means that the prediction is positive.

The unaware approach

The unaware approach requires that the protected attribute is not used in the algorithm. Let i and j be two

individuals and X_i, X_j their attributes.

$$X_i = X_j \longrightarrow \hat{Y}_i = \hat{Y}_j \tag{1}$$

Fairness through awareness

The aware approach requires that the protected attribute is used in the algorithm in order to treat similar individuals similarly. Let i and j be two individuals, X_i, X_j their attributes and S_i, S_j their protected attributes.

$$\left.\begin{array}{l} X_i = X_j \\ S_i = S_j \end{array}\right\} \longrightarrow \hat{Y}_i = \hat{Y}_j \tag{2}$$

Decoupling

This requires different algorithms for each class.

Statistical or conditional parity

Statistical parity means that the positive prediction rate is the similar across groups.

$$|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]| \leq \epsilon \tag{3}$$

A lower value of ϵ means better fairness.

Equal opportunity

Equal opportunity seeks to equalize true positive rates across groups.

$$|P[\hat{Y} = 1|S \neq 1, Y = 1] - P[\hat{Y} = 1|S = 1, Y = 1]| \leq \epsilon \tag{4}$$

Equalized odds

To satisfy equalized odds fairness, false positive rates must be similar across groups, and false negatives rates too.

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \epsilon \tag{5}$$
$$|P[\hat{Y} = 0|S = 1, Y = 1] - P[\hat{Y} = 0|S \neq 1, Y = 1]| \leq \epsilon \tag{6}$$

A smaller difference between false positive rates between minorities and dominant party, plus a smaller difference between false negative rates between these same groups mean better fairness.

Calibration

Calibration works in a probabilistic classifying situation. It means that for any predicted probability value, all groups will have similar positive predictive values. Let V be the predicted probability value.

$$|P[Y = 1|S \neq 1, V = v] - P[Y = 1|S = 1, V = v]| \leq \epsilon \tag{7}$$

Multicalibration

Multicalibration relies on calibration, without basing on the protected attribute S. The subgroups are those that the algorithm can identify as categorization sub-classes.

Counterfactual fairness [16]

In his paper, Kusner defined a causal model as a triple (U, V, F) of sets. Here, we only need the definition of U, which is a set of latent background variables, and V the set of observable variables. The factors in U are not caused by any variable in the set V.

Let S be the protected attributes, $\bar{S}$ the remaining attributes and Y the actual outcome. Then, predictor $\hat{Y}$ is counterfactually fair if under any context $S = s$ and $\bar{S} = \bar{s}$,

$$|P[\hat{Y}_{S \longleftarrow s}(U) = y|\bar{S} = \bar{s}, S = s] - P[\hat{Y}_{S \longleftarrow s'}(U) = y|\bar{S} = \bar{s}, S = s]| \leq \epsilon \tag{8}$$

for all y and for any value s' attainable by S.

1.3 Fairness verification

Fairness can be verified formally with a probabilistic verification assistant, or data-wise with statistical studies. In the former, the fairness verifier takes the decision-making program and a population model as input for proving that the algorithm is formally fair. The population model is built from an objective standpoint, typically generated from census data [18].

In the latter, the literature has indicated several issues related to data that may lead to unfairness. For instance, some research papers [19, 20] describe types of bias that can be insidiously included in the datasets used to train and fit the models:

- Datasets can include forms of bias, such as biased device measurements, historically biased human decisions, erroneous reports and others;
- Missing data can also lead to bias, knowing that different base rates can lead to unfairness. The resulting datasets are not representative of the target population, therefore unfit for real verification;
- Proxy attributes for sensitive attributes are hard to detect. Protected attributes determine privileged and unprivileged groups and are generally not allowed for use un decision-making. Proxy attributes are non-protected attributes that can be characteristic of sensitive attributes, such as adress or zipcode for race, dependant on historically segregated housing (see section 1.1.1). If the learning dataset does contain proxy attributes, the algorithm may implicitly make decisions based on the sensitive attributes [21], which is what the user was trying to avoid in the first place.

Therefore, when assessing whether an algorithm is fair or not, the controller has to be careful with these points. We will see in section 3 that ProPublica did not always fulfill these specifications in their study, and that we do not know if Equivant's COMPAS algorithm does completely avoid proxy attributes.

2 COMPAS algorithm

2.1 Introduction

COMPAS is a proprietary actuarial risk and needs assessment tool, developed by Northpointe (now owned by Equivant), and is being used by criminal justice agencies in many jurisdictions in order to determine defendants' risk to recidivate.[22] However, due to its proprietary nature, the details how it works and the way the score is computed are not known to the public (including the judges and defendants) [23]. It was initially made to inform the authorities of decisions regarding placement, supervision and case management of defendants. It was developed empirically with a focus on characteristics known to affect recidivism, such as criminal history, relationships and education. This focus is based on several of criminological theories that explain how people become involved in criminal behavior and may provide guidance for effective interventions [22, p.5-6]

COMPAS has two principal risk models: General Recidivism Risk and Violent Recidivism Risk. In order to compute each score, the algorithm takes into account both dynamic risk (crimogenic factors) and static risk (historical factors). It is particularly helpful in helping avoid overloading criminal justice systems, by guiding the authorities in their decision to keep or not a defendant based on his/her risk to reoffend.

2.2 Scaling

The Risk Scales stated before use methods and strategies for predictive modelling in order to predict if a defendant is likely to recidivate. The raw scores are then transformed in decile scores by ranking the scale scores of a normative group in ascending order and then dividing these scores into ten equal sized groups. COMPAS scores for each defendant ranged from 1 to 10, with ten being the highest risk. Scores 1 to 4 were labeled by COMPAS as "Low"; 5 to 7 were labeled "Medium"; and 8 to 10 were labeled "High". Because the decile scores are computed

in relation to the highest an lowest raw score, they can only be interpreted in a relative sense.

The defendant has to, among other tools, answer to a risk assessment that covers current charges, criminal history, compliance, family criminality, relationships, substance abuse, stability, residence, social environment, education, work, recreation, social isolation, personality, anger issues and criminal attitude. The algorithm then takes into account over 137 features to assess if a defendant is likely to recidividate up to 2 years after an hypothetical release. These answers are injected into the COMPAS software to generate scores. Both Recidivism Risk Scales are used in different situations: General Recidivism Risk is used for arrest, felony arrest, noncompliance and return to prison outcomes, while Violent Recidivism Scale is tested only for person offense (manslaughter for example).

2.2.1 General recidivism

The recidivism risk scale was designed to predict new offenses that would be done after the assessment. The original scale construction used the existence of a new offense within two years of the assessment date as the "positive" event. The main features taken into account to compute such score are primarily prior criminal history, criminal associates, drug involvement and early indicators of juvenile delinquency problems, which are known predictors of recidivism [22, p.27].

2.2.2 Violent recidivism

For this recidivism risk scale, input includes history of violence, history of non-compliance, vocational and educational problems, current age and age of first arrest. These factors are considered to be strongly associated with future violence for people without mental disorders [24].

2.3 Computing scores

There is ongoing debate between professionals advocating the superiority of actuarial risk assessments and those pleading for the use of structured clinical judgements when assessing risk for violent recidivism.

Definition 3. *(Actuarial Risk Assessment)[25] A statistically calculated prediction of the likelihood that an individual will pose a threat to others or engage in a certain behavior (e.g., violence) within a given period. Unlike in a clinical risk assessment, someone conducting an actuarial risk assessment relies on data from specific, measurable variables (e.g., age, gender, prior criminal activity) that have been validated as predictors and uses mathematical analyses and formulas to calculate the probability of dangerousness or violent behavior.*

COMPAS is an actuarial risk assessment tool, which means that it is a method of estimating the likelihood of reoffending. The score of an individual is estimated with respect to known recidivism rates of offenders sharing the same characteristics. Each feature is multiplied by a weight ω, which is determined by the importance of the feature in predicting recidivism on the study data. However, the weighting process is not available for the public.

COMPAS risk assessment is said to be about predicting group behavior, therefore not focusing on individual predictions. Equivant states that their risk scales are "able to identify groups of high-risk offenders - not a particular high-risk individual." [22, p.31]. Their goal is to identify groups of low, medium and high risk offenders, and not to predict at an individual level. The final decision is up to the authorities, who can override the computed risk to include information that is not taken into account by the algorithm (e.g., mitigating or aggravating circumstances).

3 A review of ProPublica's analysis

In 2016, an article sparked debate in the algorithmic fairness field. The media outlet ProPublica examined the fairness of COMPAS's classification and concluded that the algorithm is biased against African-Americans. The

study was made by assessing whether defendants who were released from jail or prison actually recidivated after two years, and comparing it to COMPAS's predictions. For ProPublica, COMPAS is prone to racial bias because African-American defendants were more likely misclassified as higher risk individuals when Caucasian defendants were more likely misclassified as lower risk individuals. We will discuss more in depth the statements of ProPublica in this section.

The disagreement between Equivent and ProPublica relies on their different definitions of fairness. For Equivent, for instance, COMPAS algorithm is fair because in each category (e.g., racial group) the same percentage of defendants recidivated. For ProPublica, fairness means that the algorithm should make the same type of error (false positives and false negatives) across all groups. As seen before, satisfying different notions of fairness at the same time is nearly impossible, so their disagreement is what makes this real-life example very interesting for Computer Science studies.

3.1 Dataset

The details of the database are given by ProPublica as an introduction to their study [23]. The database contains the data of criminal defendants in Broward County, Florida. It also contains their recidivism rate predicted by COMPAS, and the rate that actually occurred over a two-year period. ProPublica chose Broward County to collect data because it is a large jurisdiction using the COMPAS tool in pretrial release decision and Florida has strong open-records laws, which facilitated data collection.

ProPublica obtained two years worth of COMPAS scores from the Broward County for all 18610 people who were scored in 2013 and 2014. However, some data was discarded : Broward County primarily uses the score to determine whether to release or not a defendant before his or her trial, so they discarded scores that were assessed at parole, probation or other stages of the criminal justice system. These previous steps left 11757 people who were assessed at the pretrial stage.

The data is arranged in 7 subdatabases :

- charge : contains the charges of each defendant
- casearrest : contains the data of arrestations
- compas : contains the scores predicted by COMPAS algorithm
- people : contains the personal information of the defendants
- jailhistory : contains the criminal history of detention in jail (minor crimes)
- prisonhistory : contains the criminal history of detention in prison (major crimes and felony)

Each pretrial defendant received at least three COMPAS scores: "Risk of Recidivism", "Risk of Violence" and "Risk of Failure to Appear", but for the study they only focused on the first two scores. Also, they found that sometimes defendants' information was incorrectly entered, which led to only valid cases.

Then, the definition of recidivism had to be discussed in order to consistently recognize recidivists. They considered recidivism as "a criminal offense (minus traffic tickets and municipal ordinance violations) that resulted in a jail booking and that took place after the crime for which the person was COMPAS scored." [23] They also ignored people who failed to appear at their court hearings and were arrested for it. For violent recidivism, they used the FBI's definition of violent crime.

Definition 4. *(Violent crime)[26] In the FBI's Uniform Crime Reporting (UCR) Program, violent crime is composed of four offenses: murder and nonnegligent manslaughter, forcible rape, robbery, and aggravated assault. Violent crimes are defined in the UCR Program as those offenses which involve force or threat of force.*

However, ProPublica did keep more cases than needed in their study [27]. The COMPAS survey is generally taken the same day or the day after the defendant is jailed, therefore the waiting time between arrest and screening is very short. ProPublica then collected data on arrest that took place until the end of March 2016 in order to label recidivism. This means that the limit date of screening for defendants should be April 1, 2014. This was

not the case: a simple look to the files used by ProPublica shows that they kept defendants with screening date between that limit date and December 31, 2014. ProPublica correctly dropped all non-recidivists with screening dates after April 1, 2014, but failed to drop recidivists with screening dates after the limit date. Thus, the initial dataset has more recidivists than expected.

We constructed a corrected dataset by eliminating all the entries for which the screening date is posterior to April 1, 2014. The demographic study of the corrected dataset is presented in the following tables.

Race	Number	Percentage
African-American	2682	50.57%
Asian	27	0.51%
Caucasian	1829	34.48%
Hispanic	448	8.45%
Native American	9	0.17%
Other	309	5.83%
Total	5304	100%

(a) General recidivism risk

Race	Number	Percentage
African-American	1813	47.11%
Asian	25	0.65%
Caucasian	1399	36.56%
Hispanic	347	9.06%
Native American	6	0.16%
Other	247	6.45%
Total	3827	100%

(b) Violent recidivism risk

Table 2: Demographic study for ethnicity

Sex	Number	Percentage
Female	1017	19.17%
Male	4287	80.83%
Total	5304	100%

(a) General recidivism risk

Sex	Number	Percentage
Female	812	21.22%
Male	3015	78.78%
Total	3827	100%

(b) Violent recidivism risk

Table 3: Demographic study for sex

Age category	Number	Percentage
Less than 25	1108	20.89%
25 - 45	3031	57.15%
Greater than 45	1165	21.96%
Total	5304	100%

(a) General recidivism risk

Age category	Number	Percentage
Less than 25	714	18.66%
25 - 45	2178	56.91%
Greater than 45	935	24.43%
Total	3827	100%

(b) Violent recidivism risk

Table 4: Demographic study for age category

Recidivism	Number	Percentage
Negative	3363	63.40%
Positive	1941	36.60%
Total	5304	100%

(a) General recidivism risk

Violent recidivism	Number	Percentage
Negative	3368	88.01%
Positive	459	11.99%
Total	3827	100%

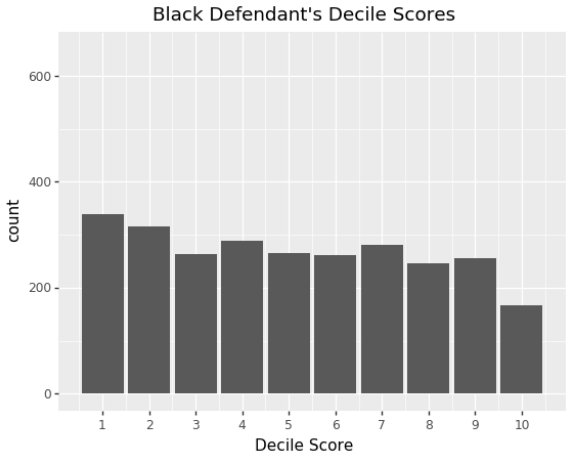
(b) Violent recidivism risk

Table 5: Recidivism rates

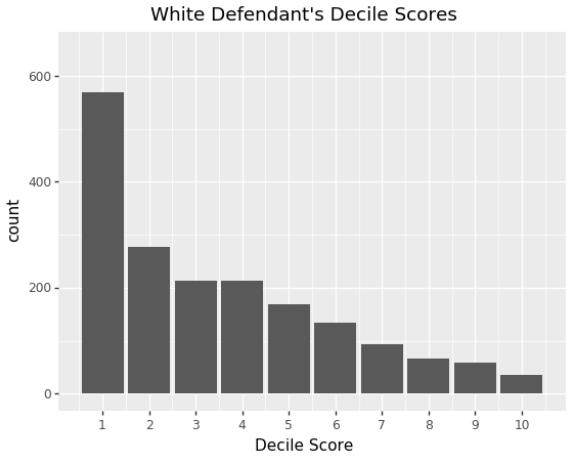
3.2 Score distribution

The analysis begins by comparing the distribution of the COMPAS decile scores among African-Americans and Caucasians, the two dominant groups in terms of population in the dataset. In accordance with ProPublica, we

found that scores for Caucasian defendants were not distributed evenly and tended to be lower. The scores of African-American defendants were more evenly distributed, which show a disparity in score distribution between these groups.



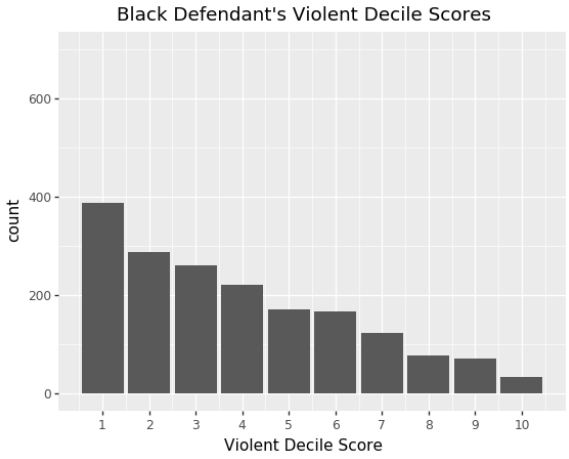
(a) African-American defendants



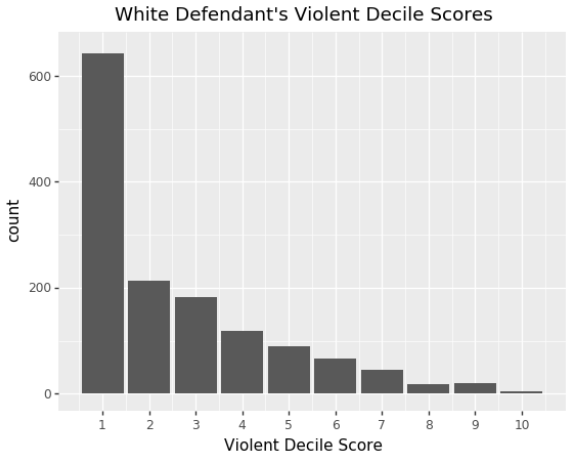
(b) Caucasian defendants

Figure 1: Distribution of decile scores

The same observation can be made for violent recidivism decile scores. However, the simple distributions of the decile scores across groups do not inform about the racial disparities in score.



(a) African-American defendants



(b) Caucasian defendants

Figure 2: Distribution of violent decile scores

Then, we create a logistic regression model that considers race, age category, gender, crimes, number of priors and actual future recidivism. ProPublica found that several factors were quite predictive of a higher score. For instance, defendants under 25 had 149.61% chances to be ranked as higher risk than the reference population (e.g.,between 25 and 45 years old). African-American defendants were 45.28% more likely to obtain a higher score than the baseline population (e.g., Caucasian) and women were 12.48% more likely to obtain such score. However, this could be due to the significantly smaller sample of women: there exists a femenine version of COMPAS that tailored for this particular subgroup.

We found different rates with our corrected dataset: young defendants had 159.53% chances to be ranked as higher risk, African-American had 46.92% chances and women 20.28% more chances. The slightly higher rates found could be caused by the smaller corrected sample, allowing less entries to train the model. The same observations were made with the violent recidivism sub-data: for ProPublica, African-American defendants and

young defendants were 77.39% and 641.42% more likely to obtain a higher score than the reference population. In our study, these rates were of 76.84% and 665.99% respectively.

For the next analyses, medium and high scores were generally mashed into a single category, named High. This was done accordingly to Equivant's user guide, which states: "scores in the medium and high range garner more interest from supervision agencies." [22, p.27]

3.3 Predictive accuracy

We ran a Cox Proportional Hazards model in order to assess if COMPAS could accurately predict if a defendant recidivated or not, based on the score it gave in the first place. This was done by ProPublica and Equivant in their own validation study [28]. For their analysis, the sample size was of 10314 defendants, of which 3569 are Caucasian and 5147 African-American. In our study, we corrected the dataset so the sample size was of 6694, of which 2259 are Caucasian and 3437 African-American.

The Cox proportional-hazards model is a regression model commonly used for investigating the association between the survival time of subjects and several predictor variables.It takes as arguments the duration until the event and the resulting event. In this case, the duration was the time elapsed between a defendant's screening date and his/her re-offense, or April 1, 2016 if there is no recidivism. The event is recidivism: 1 for reoffenders and 0 for the others.

In our corrected dataset we had 1961 high-risk, 2414 medium-risk and 4472 low-risk defendants. ProPublica had found that people with high scores had 3.50 more chances to recidivate than those with low risk scores. Equivant had found that this rate was of 5.60, and our study (table 6) found that this rate is of 2.94. In any case, Cox regression shows that there is indeed a predictive characteristic in these scores. Also, by looking at the concordance we notice that the model accurately predicted the event 61% percent of time, rather than 63% (as ProPublica found). For Equivant, this rate was even higher.

	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	z	p	log2(p)
High	1.08	2.94	0.05	0.98	1.17	2.66	3.24	21.73	< 0.005	345.49
Medium	0.66	1.94	0.05	0.57	0.76	1.76	2.13	13.64	< 0.005	138.36

Number of observations	8847
Number of events	2475
Concordance	0.61
Partial log-likelihood	-21180.76
Partial AIC	42365.51
log-likelihood ratio test	485.52 on 2 df
-log2(p) of ll-ratio test	350.23

Table 6: Cox Proportional Hazards model on scores

When doing the same test on decile scores, we found that the predictive accuracy was of 64%, rather than 66% (as ProPublica found). In all cases, this means that decile scores are slightly more accurate in predicting the defendant's behavior. Then, we added an interaction term into the data and ran another Cox regression model (table 7).

The interaction term shows that there are disparities between African-American and Caucasian defendants. However, here again our results are different from ProPublica's. For ProPublica, high risk white defendants are 3.61 more likely than low risk white defendants, while High risk black defendants are 2.99 more likely than low.[23]. In our study, these rates were of 2.86 and 2.56. The disparity is therefore less pronounced in our study, but it is still present. These results are included in table 8. Also, here the accuracy is of 60%, which is lower than the threshold for reliability described by Equivant: "A rule of thumb according to several recent articles is

that AUCs of .70 or above typically indicate satisfactory predictive accuracy, and measures between .60 and .70 suggest low to moderate predictive accuracy". [28]

Black High Hazard	2.56
White High Hazard	2.86
Black Medium Hazard	1.86
White Medium Hazard	2.08

Table 8: Summary of Cox model

Then, we built a Kaplan-Meier model for each sub-group. The Kaplan–Meier estimator is a non-parametric statistic used to estimate the survival function from lifetime data. It takes the same arguments as the Cox Proportional Hazards model stated before. The Kaplan Meier survival plot (figure 3) helps visualize the difference in recidivism rates between high, medium and low risk defendants.

According to the figures, high risk defendants are more likely to reoffend than low risk defendants: they tend to survive more. Table 9 summarizes the recidivism rates with respect to the defendant's ethnicity.

Score	Overall	Caucasian	African-American
Low	0.2704	0.2416	0.3188
Medium	0.4622	0.4211	0.4837
High	0.5973	0.5525	0.6171

Table 9: Race-specific recidivism rate

We ran race-specific Cox models (tables 10 and 11) in order to check accuracy across races. We found that, just as ProPublica's study, the regression accuracy rates across races were the same: 60%.

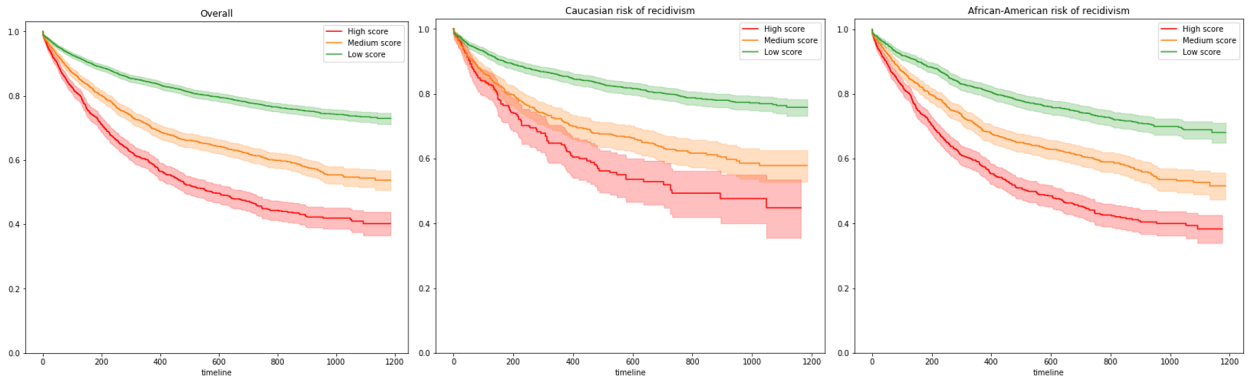


Figure 3: Race-specific Kaplan-Meier models

	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	z	p	log2(p)
High	1.02	2.78	0.11	0.82	1.23	2.26	3.42	9.70	< 0.005	71.55
Medium	0.71	2.04	0.09	0.54	0.88	1.72	2.41	8.34	< 0.005	53.35

Number of observations	2972
Number of events	703
Concordance	0.60
Partial log-likelihood	-5298.42
Partial AIC	10600.83
log-likelihood ratio test	117.98 on 2 df
-log2(p) of ll-ratio test	85.10

Table 10: Cox Proportional Hazards model on scores of Caucasian defendants

	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	z	p	log2(p)
High	0.96	2.61	0.06	0.83	1.09	2.30	2.97	14.91	< 0.005	164.64
Medium	0.53	1.70	0.07	0.40	0.66	1.50	1.94	8.02	< 0.005	49.77

Number of observations	4657
Number of events	1480
Concordance	0.60
Partial log-likelihood	-11649.87
Partial AIC	23303.74
log-likelihood ratio test	226.89 on 2 df
-log2(p) of ll-ratio test	163.66

Table 11: Cox Proportional Hazards model on scores of African-American defendants

We ran a similar analysis on COMPAS's violent recidivism score and found similar disparities to those for general recidivism (table 12). For ProPublica, the results did not give any additional information on differences across racial groups, meaning that there is no significant difference the hazards of high and low risk black defendants and high and low risk white defendants for them. This could be due to the fact that their dataset kept a larger number entries than needed: their study was based on 18178 observations for 818 events, while ours only kept 12465 observations for 598 events. The Kaplan-Meier plots show that the survival rates are very high, meaning that there is far less violent recidivism than general recidivism (figure 4).

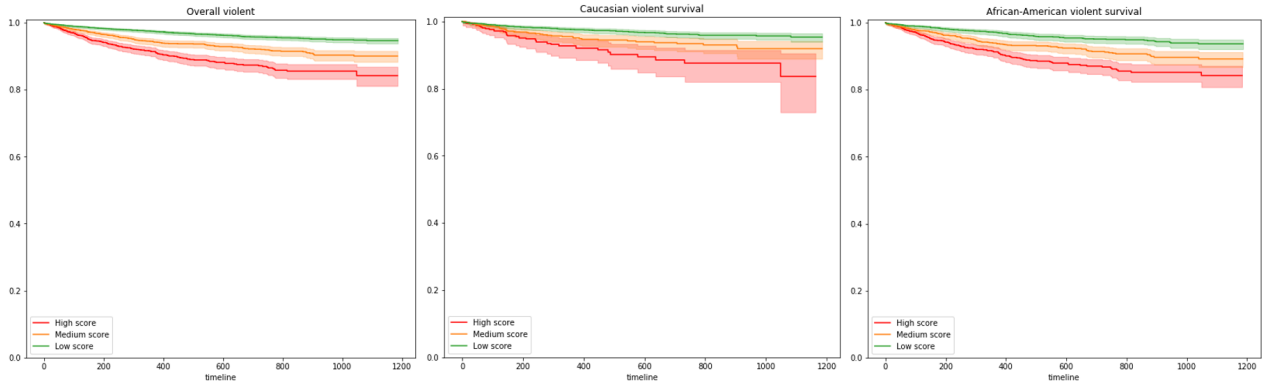


Figure 4: Race-specific Kaplan-Meier models for violent recidivism

3.4 Error rates

The previous analysis showed that COMPAS algorithm is prone to overpredicting African-American's future recidivism: their scores are generally higher but these higher scores are less significant of future recidivism than for Caucasian defendants. We also verified error rates to investigate whether false positives and false negatives were unevenly distributed among races. We used the `truth_tables` module available in ProPublica's study with some slight changes to suit our corrected dataset.

The dataset used for this study was made by removing people from the initial dataset for whom ProPublica had less than two years of recidivism information. The remaining population was of 6216, instead of ProPublica's population of 7214. It is slightly larger than the sample in the previous studies because the defendant's case information is not needed [23]. Tables 13a, 13b and 13c show disparities between false positive and false negative rates across race groups.

	Low	High
Survived	2681	1282
Recidivated	841	1412
FP	32.35%	
FN	37.33%	
PPV	0.52	
NPV	0.76	
LR+	1.94	
LR-	0.55	

(a) All defendants

	Low	High
Survived	1139	349
Recidivated	314	330
FP	23.45%	
FN	48.76%	
PPV	0.49	
NPV	0.78	
LR+	2.18	
LR-	0.64	

(b) Caucasian defendants

	Low	High
Survived	990	805
Recidivated	375	969
FP	44.85%	
FN	27.90%	
PPV	0.55	
NPV	0.73	
LR+	1.61	
LR-	0.51	

(c) African-American defendants

Table 13: Contingency tables

As shown in the previous sub-tables, the algorithm is more likely to make false positive errors on African-American defendants than on Caucasian defendants: 44.85% versus 23.45%, so nearly twice as likely compared to their counterparts. This means that COMPAS is more likely to falsely label a black defendant as high risk recidivist where ultimately this defendant would not recidivate. ProPublica found that black defendants who scored higher did recidivate slightly more often than white defendants: 63% versus 59%. We found the same tendency but with slightly lower percents: 55% versus 49%. These slight differences could be explained by the demographics of the study: we only have 2132 white defendants for 3139 black defendants, which means we have 47% more African-American entries than Caucasian entries.

COMPAS tended to make more false negative errors on Caucasian defendants than on African-American defendants: 48.76% versus 27.90%, there again nearly twice as likely compared to their counterparts. This means that the algorithm is more likely to wrongly predict that a white defendant would not reoffend if released, compared to his/her black counterparts. Then, the positive likelihood ratio for white defendants was slightly higher than for black defendants: 2.18 versus 1.61. This means that a positive result for a Caucasian defendant is more likely to label a true positive result than for his/her African-American counterparts.

We made this study with the assumption that medium-score defendants could be considered as high-score defendants. We also tested the error rates for true high-risk defendants and found the following contingency tables.

	Low	True High
Survived	3561	402
Recidivated	526	698
FP	10.14%	
FN	69.02%	
PPV	0.63	
NPV	0.70	
LR+	3.05	
LR-	0.77	

(a) All defendants

	Low	True High
Survived	1407	349
Recidivated	526	118
FP	5.44%	
FN	81.68%	
PPV	0.59	
NPV	0.73	
LR+	3.37	
LR-	0.86	

(b) Caucasian defendants

	Low	True High
Survived	1511	284
Recidivated	807	537
FP	15.82%	
FN	60.04%	
PPV	0.65	
NPV	0.65	
LR+	2.53	
LR-	0.71	

(c) African-American defendants

Table 14: Contingency tables

Here, black defendants were almost 3 times as likely as white defendants to be rated as high risk offenders, by error. The other rates are less significant because of the restrictive definition of true high risk defendants.

Finally, we made the same studies on the violent crimes dataset. Again, tables 15a, 15b and 15c show that the algorithm is prone to making more false positive errors on African-American defendants than on Caucasian defendants: 38.14% versus 18.46%, more than twice as likely compared to their white counterparts. ProPublica found these same results since they only kept more recidivists than needed, successfully eliminating superfluous non-recidivists. Also, here again black defendants who scored higher for violent recidivism did recidivate slightly more often than white defendants: 16% versus 13%, while ProPublica found 21% versus 17%. Our demographics are different once more: we only have 2198 white defendants for 3049 black defendants, which means that we have 38% more African-American entries than Caucasian entries.

COMPAS tended here again to make more false negative errors on Caucasian defendants than on African-American defendants: 38.14% versus 18.46%, so more than twice as likely as white defendants to be erroneously rated as high risk offenders. We also noticed that the positive likelihood ratios were the same as in the general recidivism studies. Then, with the true high scores as the "High" category, we could make the same observations: black defendants were 3 times as likely as white defendants to be rated as high risk offenders by error.

	Low	High
Survived	4121	1597
Recidivated	238	277
FP	27.93%	
FN	46.21%	
PPV	0.15	
NPV	0.95	
LR+	1.93	
LR-	0.64	

(a) All defendants

	Low	High
Survived	1679	380
Recidivated	83	56
FP	18.46%	
FN	59.71%	
PPV	0.13	
NPV	0.95	
LR+	2.18	
LR-	0.73	

(b) Caucasian defendants

	Low	High
Survived	1692	1043
Recidivated	120	194
FP	38.14%	
FN	38.22%	
PPV	0.16	
NPV	0.93	
LR+	1.62	
LR-	0.62	

(c) African-American defendants

Table 15: Contingency tables for violent recidivism

	Low	True High
Survived	5353	365
Recidivated	393	122
FP	6.38%	
FN	76.31%	
PPV	0.25	
NPV	0.93	
LR+	3.71	
LR-	0.82	

(a) All defendants

	Low	True High
Survived	1991	68
Recidivated	125	14
FP	3.30%	
FN	89.93%	
PPV	0.17	
NPV	0.94	
LR+	3.05	
LR-	0.93	

(b) Caucasian defendants

	Low	True High
Survived	2461	274
Recidivated	221	93
FP	10.02%	
FN	70.38%	
PPV	0.25	
NPV	0.92	
LR+	2.96	
LR-	0.78	

(c) African-American defendants

Table 16: Contingency tables for violent recidivism

3.5 Discussion on fairness

As stated before, the disagreement between Equivant and ProPublica relies on their different definitions of fairness which are mutually exclusive. Therefore, bias depends on the outlook. For Equivant, the algorithm is fair because the scores represent approximately the same risk of recidivism across races (see figure 5). This means that among defendants who were given the same score, they all have approximatively the same chances to reoffend. So, for Equivant, when the judges see a defendant's score he does not need to know his/her race before making a decision. That is Equivant's notion of fairness: a calibrated algorithm.

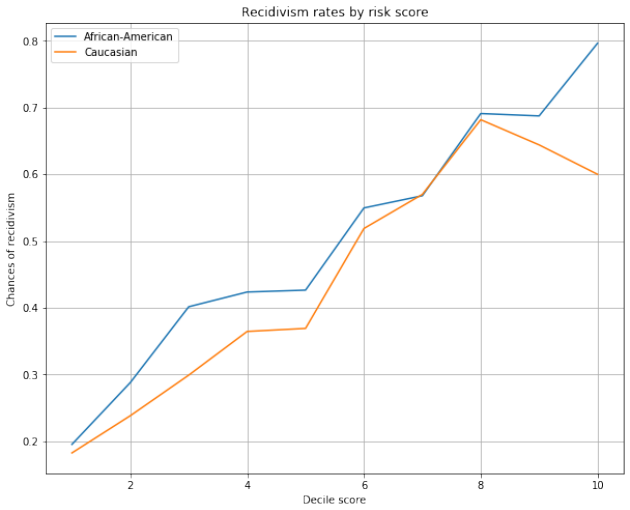


Figure 5: Recidivism rates by risk score

For ProPublica, the algorithm is not fair because the false-negative and false-positive rates are not the same across races: they aim for equalized odds fairness. For them, the algorithm should not misclassify and treat harsher a certain sub-group of defendants more frequently than others. ProPublica therefore stated that among defendants who ultimately did not reoffend, African-American defendants were nearly twice as likely to be misclassified as higher risk than their Caucasian counterparts (45% versus 24%). In this study, we found that ProPublica kept more recidivists than needed but the rate is the same.

Conclusion

Differences between ProPublica's and Equivant's studies rely on the definition of fairness. As seen in the first section, defining fairness is a very complex task: fairness is contextual, and for some forms of fairness it is impossible to meet their requirements simultaneously. For COMPAS algorithm, the stakes are even higher: falsely ranked low-risk defendants represent a public threat and misclassified high-risk defendants are heavily impacted in their life, and race is a protected attribute that quickly sparks controversy. Our study executed the same steps as ProPublica's with a different corrected dataset, but this did not change false positive rates (only false negatives) so the conclusion is still the same as ProPublica's.

Racial bias may still be included in the algorithm since it is a classifying algorithm based on machine learning, therefore relying on calibration data. Even if Equivant ensures that race is not taken into account by the algorithm, some characteristics may act as a proxy for race. For example, poor neighborhoods have higher criminality, and Black people are more likely to live in these neighborhoods, therefore their social environment, work and

education may be less stable. These characteristics are indeed taken into account by COMPAS, which could explain why African-American defendants are generally ranked as higher-risk individuals. ProPublica would have wanted to achieve equality by equalized odds, meaning that it would equalize the false positive and false negative errors for each race. But that means that the algorithm would be aware of the defendant's ethnicity, which could raise concern.

Also, there may be some biases applied to the collection of data: rearrest is considered as recidivism which is not a direct measure of reoffending. As a result, differences in the arrest rate of black and white defendants make it difficult to compare directly false-positive and false-negative rates. Black people, for example, are almost four times as likely as white people to be arrested for drug offenses [29]. This also raises a question : how to make an unbiased algorithm that bases itself on biased data? Besides, it is difficult to construct a risk score that does not include items that can be correlated with race — such as poverty, joblessness and social marginalization [23].

	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	z	p	log2(p)
High	1.05	2.86	0.10	0.85	1.26	2.33	3.51	10.05	< 0.005	76.57
Medium	0.73	2.07	0.09	0.56	0.89	1.75	2.45	8.52	< 0.005	55.76
African-American	0.29	1.34	0.07	0.15	0.43	1.16	1.54	4.04	< 0.005	14.16
African-American: High	-0.11	0.90	0.12	-0.34	0.13	0.71	1.14	-0.87	0.39	1.36
African-American: Medium	-0.20	0.82	0.11	-0.41	0.01	0.66	1.01	-1.85	0.06	3.94
Asian	-1.07	0.34	0.71	-2.46	0.32	0.09	1.37	-1.51	0.13	2.94
Asian: High	1.86	6.40	1.01	-0.11	3.83	0.89	45.91	1.85	0.06	3.95
Asian: Medium	1.55	4.71	0.92	-0.25	3.35	0.78	28.43	1.69	0.09	3.46
Hispanic	-0.05	0.96	0.12	-0.27	0.18	0.76	1.20	-0.39	0.70	0.52
Hispanic: High	-0.12	0.88	0.24	-0.59	0.35	0.55	1.41	-0.51	0.61	0.72
Hispanic: Medium	0.01	1.01	0.19	-0.37	0.40	0.69	1.49	0.08	0.94	0.09
Native	-12.09	0.00	334.26	-667.23	643.05	0.00	1.88 e+279	-0.04	0.97	0.04
Native: High	12.94	4.17 e+05	334.26	-642.20	668.08	0.00	1.40 e+290	0.04	0.97	0.05
Native: Medium	12.85	3.80 e+05	334.26	-642.29	667.99	0.00	1.27 e+290	0.04	0.97	0.04
Other	0.10	1.11	0.13	-0.15	0.35	0.86	1.42	0.79	0.43	1.21
Other: High	0.47	1.59	0.31	-0.14	1.07	0.87	2.92	1.50	0.13	2.91
Other: Medium	-0.27	0.76	0.27	-0.79	0.25	0.45	1.29	-1.01	0.31	1.67

Number of observations	8847
Number of events	2475
Concordance	0.62
Partial log-likelihood	-21159.17
Partial AIC	42352.35
log-likelihood ratio test	528.68 on 17 df
-log2(p) of ll-ratio test	334.75

Table 7: Cox Proportional Hazards model with interaction term

	coef	exp (coef)	se (coef)	coef lower 95%	coef upper 95%	exp (coef) lower 95%	exp (coef) upper 95%	z	p	log2(p)
High	1.17	3.22	0.21	0.77	1.57	2.15	4.82	5.68	< 0.005	26.17
Medium	0.62	1.86	0.19	0.26	0.99	1.29	2.69	3.34	< 0.005	10.20
African-American	0.28	1.32	0.16	-0.03	0.58	0.97	1.79	1.76	0.08	3.67
African-American: High	-0.13	0.88	0.24	-0.61	0.35	0.54	1.41	-0.54	0.59	0.77
African-American: Medium	-0.02	0.98	0.23	-0.47	0.44	0.62	1.55	-0.07	0.95	0.08
Asian	-0.09	0.91	1.01	-2.07	1.88	0.13	6.56	-0.09	0.93	0.11
Asian: High	-11.81	0.00	746.35	-1474.63	1451.02	0.00	inf	-0.02	0.99	0.02
Asian: Medium	1.64	5.15	1.24	-0.79	4.07	0.45	58.41	1.32	0.19	2.43
Hispanic	-0.03	0.97	0.25	-0.52	0.47	0.59	1.59	-0.11	0.91	0.13
Hispanic: High	0.10	1.10	0.45	-0.79	0.98	0.46	2.67	0.21	0.83	0.27
Hispanic: Medium	0.25	1.29	0.40	-0.54	1.04	0.58	2.83	0.63	0.53	0.91
Native	0.27	1.31	1.01	-1.70	2.24	0.18	9.40	0.27	0.79	0.34
Native: High	none	none	none	none	none	none	none	none	none	none
Native: Medium	0.60	1.83	1.43	-2.19	3.40	0.11	29.87	0.42	0.67	0.57
Other	0.28	1.32	0.26	-0.24	0.79	0.79	2.21	1.05	0.29	1.77
Other: High	1.0	2.77	0.47	0.09	1.94	1.10	6.99	2.16	0.03	5.02
Other: Medium	-0.42	0.66	0.58	-1.56	0.73	0.21	2.07	-0.72	0.47	1.08

Number of observations	12465
Number of events	598
Concordance	0.64
Partial log-likelihood	-5266.20
Partial AIC	10564.40
log-likelihood ratio test	157.82 on 16 df
-log2(p) of ll-ratio test	81.90

Table 12: Cox Proportional Hazards model with interaction term for violent recidivism

References

[1] D. A. Elyounes, "Contextual fairness: A legal and policy analysis of algorithmic fairness," *SSRN Electronic Journal*, 2019.

[2] D. Weinberger, "Playing with AI Fairness." https://pair-code.github.io/what-if-tool/ai-fairness.html, 2018.

[3] N. Gotanda, "A Critique of 'Our Constitution is Color-Blind'," *Stanford Law Review*, vol. 44, no. 1, pp. 1–68, 1991.

[4] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," 2016.

[5] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," 2011.

[6] J. V. White, "What is Affirmative Action?," *Scholarly Works*, no. 306, 2004. https://scholars.law.unlv.edu/facpub/306.

[7] E. C. Boddie, "The Future of Affirmative Action," *Havard Law Review*, vol. 130, 2016.

[8] A. Kull, *The Color-Blind Constitution*. Harvard University Press, 1994.

[9] D. Hellman, "Measuring algorithmic fairness." https://ssrn.com/abstract=3418528, 2019.

[10] J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, 2017.

[11] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On Fairness and Calibration," 2017.

[12] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent trade-offs in the fair determination of risk scores," 2016.

[13] N. Saxena, K. Huang, E. DeFilippis, G. Radanovic, D. Parkes, and Y. Liu, "How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness," 2018.

[14] U. Hebert-Johnson, M. P. Kim, O. Reingold, and G. N. Rothblum, "Calibration for the (computationally-identifiable) masses," 2017.

[15] J. R. Loftus, C. Russell, M. J. Kusner, and R. Silva, "Causal reasoning for algorithmic fairness," 2018.

[16] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," 2017.

[17] D. Pessach and E. Shmueli, "Algorithmic fairness," 2020.

[18] A. Albarghouthi, L. D'Antoni, S. Drews, and A. Nori, "Fairness as a Program Property," 2016.

[19] A. Chouldechova and A. Roth, "The frontiers of fairness in machine learning," 2018.

[20] F. Martínez-Plumed, C. Ferri, D. Nieves, and J. Hernández-Orallo, "Fairness and missing values," 2019.

[21] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *SSRN Electronic Journal*, 2016.

[22] Equivant, "Practitioner's Guide to COMPAS Core." https://www.equivant.com/wp-content/uploads/Practitioners-Guide-to-COMPAS-Core-040419.pdf, 2019.

[23] J. Langwin, J. Larson, S. Mattu, and L. Kirchner, "Machine Bias. There's software used across the country to predict future criminals. And it's biased against blacks.," *ProPublica*, 2019. https://www.propublica.

org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[24] P. Gendreau, T. Little, and C. Goggin, "A Meta-analysis of the Predictors of Adult Offender Recidivism: What Works!," *Criminology*, vol. 34, no. 4, 1996.

[25] American Psychological Association, "APA Dictionary of Psychology." https://dictionary.apa.org/actuarial-risk-assessment, 2020.

[26] US Department of Justice, Federal Bureau of Investigation, "Violent crime." https://ucr.fbi.gov/crime-in-the-u.s/2010/crime-in-the-u.s.-2010/violent-crime, 2020.

[27] M. Barenstein, "ProPublica's COMPAS Data Revisited," 2019.

[28] T. Brennan, W. Dieterich, and B. Ehret, "Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System," *Criminal Justice and Behavior*, vol. 36, 2008.

[29] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, vol. 4, no. 1, 2018.