arXiv:2111.09211v1 [stat.AP] 17 Nov 2021

Improving Fairness in Criminal Justice Algorithmic Risk Assessments Using Optimal Transport and Conformal Prediction Sets*

Richard A. Berk
University of Pennsylvania
Arun Kumar Kuchibhotla
Carnegie Mellon University
Eric Tchetgen Tchetgen
University of Pennsylvania

Abstract

In the United States and elsewhere, risk assessment algorithms are being used to help inform criminal justice decision-makers. A common intent is to forecast an offender's "future dangerousness." Such algorithms have been correctly criticized for potential unfairness, and there is an active cottage industry trying to make repairs. In this paper, we use counterfactual reasoning to consider the prospects for improved fairness when members of a less privileged group are treated by a risk algorithm as if they are members of a more privileged group. We combine a machine learning classifier trained in a novel manner with an optimal transport adjustment for the relevant joint probability distributions, which together provide a constructive response to claims of bias-in-bias-out. A key distinction is between fairness claims that are empirically testable and fairness claims that are not. We then use confusion tables and conformal prediction sets to evaluate achieved fairness for projected risk. Our data are a random sample of 300,000 offenders at their arraignments for a large metropolitan area in the United States during which decisions to release or detain are made. We show that substantial improvement in fairness can be achieved consistent with a Pareto improvement for protected groups.

*Cary Coglianese and Sandra Mayson provided many insightful suggestions for legal conceptions of fairness and the prospect for criminal justice reform. Discussions with Michael Kearns helped enormously to clarify the technical issues. We also received very useful feedback from a group of researchers at MIT and Harvard who work on causal inference. Special thanks go to Devavrat Shah.

Keywords
Risk Assessment; Fairness; Risk Algorithms; Machine Learning; Optimal Transport; Conformal Prediction Sets

1 Introduction

The goal of fair algorithms remains a high priority among algorithm developers and the users of those algorithms. The literature is large, scattered, and growing rapidly, but there seem to be three related conceptual clusters: definitions of fairness and the tradeoffs that necessarily follow (Kleinberg et al., 2017; Kroll et al., 2017, Corbett-Davies and Goel, 2018), claims of unbiquitous unfairness (Harcourt, 2007; Star, 2014; Tonrey, 2014; Mullainathan, 2018), and a host of proposals for technical solutions (Kamiran and Calders, 2012; Hardt et al., 2016; Feldman et al. 2015; Zafer et al., 2017; Kearns et al., 2018; Madras et al., 2018b; Lee et al., 2019; Johndrow and Lum, 2019; Romano et al., 2019; Skeem and Lowenkamp, 2020). There are also useful overviews that cut across these domains (Berk et. al. 2018; Baer et al., 2020; Mitchell et al., 2021)

In this paper, we focus on risk assessments used in criminal justice settings and propose a novel fix for algorithmic unfairness. Because the outcomes of interest are classes, we concentrate on algorithmic classifiers. Unlike most other work, the methods we offer are in part a response to a political climate in which appearances can be more important than facts, and political gridlock is a common consequence. To help break the gridlock, we seek a rigorous solution for algorithmic unfairness that is politically acceptable to stakeholders. In so doing, we take a hard look at what risk algorithms realistically can be expected to accomplish.

A recent paper by Berk and Elzarka (2020) provides a good start by proposing a novel way a fair algorithm could be trained. But their approach lacks the formal framework we offer, which, in turn, solves problems that the earlier work cannot. Building on a foundation of machine learning, optimal transport (Hütter and Rigollet, 2020; Ni et al., 2021), and conformal prediction sets (Vovk et al., 2005; 2009), we suggest a justification for risk algorithms that treats members of a less "privileged" group as if they were members of a more "privileged" group. We use Black offenders and White offenders at their arraignments to illustrate our approach with the forecasting target an arrest for a crime of violence. Many will claim that Black offenders represent a less privileged protected group and White offenders represent a more privileged protected group. Less freighted terms

are "disadvantaged" and "advantaged" respectively.

Put a little too simply, if the performance of a risk algorithm is an acceptable standard for the relevant class of White offenders, it is an acceptable standard for the relevant class of Black offenders. This helps to underscore that we propose altering how different protected groups are *treated* by a risk algorithm. We are not proposing that an algorithm change protected group membership. In causal language, our algorithmic intervention is manipulable in a manner that could be undertaken in practice (Morgan and Winship, 2015: 438 – 441).

Whatever the words to describe the protected classes, there certainly can be legitimate fairness concerns about this formulation. Treating Black offenders as if they are White may be seen as inequality of treatment. We argue later that Black offenders as a group can be, on the average, made better off while White offenders as a group can be, on the average, not made worse off. We can achieve a form of Pareto improvement.This is different from the manner in which controversial interventions such as affirmative action are designed.

We also respond constructively to a long standing ethical quandary in U.S. criminal justice (Fisher and Kadane, 1983) commonly neglected in recent overviews of fairness (Baer et al., 2020; Mitchell et al., 2021). Should adjustments towards racial fairness use the treatment of White offenders as the baseline, the treatment of Black offenders as the baseline, or some compromise between the two? Although in principle, equality may be achieved using any shared fairness baseline, those who are made better off and those who are made worse can differ dramatically. Unless an acceptable fairness baseline for all is determined, there likely will be no agreement on how fairness is to be achieved. Moreover, when the fairness baseline is not addressed along with adjustments toward fairness, one can arrive at fair results in which everyone is made equally worse off. It is difficult to imagine that stakeholders would find such a result palatable.

Finally, we sidestep a key difficulty that currently has not been adequately resolved. A common assumption is that any disparity in treatment or outcome between different protected groups is unfair and even discriminatory. Gender provides an especially clear example. Men are disproportionately over-represented in prisons compared to women. But throughout recorded history, men disproportionately have committed the vast majority of violent crimes. Is the gender disparity in imprisonment explained solely by unfairness? For criminal justice more broadly, finding comprehensive explanations for racial disparities is at least as challenging (Hudson, 1989; Yates, 1997). We do not pretend to have a resolution but offer instead what

we hope is a politically acceptable approach.

However, treating Black offenders as if they are White, complicates how the appropriate risk estimands should be defined. Proceeding as if a fair algorithm can correct fundamental and widespread racial disparities further confuses matters. Necessarily, counterfactuals in some form are being introduced because Black offenders cannot be White offenders, and a fair risk algorithm does not make fair all criminal justice decisions and actions that follow. We address these issues in the context of estimates produced by a classifier and using confusion tables and conformal prediction sets. In so doing, we introduce counterfactual estimands to help clarify distinctions between fair risk assessment *procedures* and fairness in the criminal justice *system* more generally. The two are often conflated. A risk procedure ends with the output of a risk tool. Everything that follows, including decisions as well as actions, are features of the criminal justice system beyond the risk algorithm.

In Section 2, we discuss definitions of fairness in the statistics and computer science literature commonly associated with criminal justice risk assessment and introduce two key concepts: internal fairness and external fairness. Section 3 summarizes the methods we use to improve fairness in criminal justice risk assessments: how a classifier should be trained, how to make comparable joint distributions of the data from different protected groups, and how to gauge fairness using conformal prediction sets. These methods are discussed more formally in Appendices A through D. Section 4 describes the data to be analyzed, and Section 5 discusses the results. In section 6, we return in depth to the empirical results shown as conformal prediction sets. Conclusions are offered in Section 7.

2 Defining Fairness for Protected Groups

Under the U.S. Civil Rights Act of 1964, a "protected group" is a class of people that has explicit protection against discrimination consistent with the 5th and 14th Amendments to the United States Constitution. Racial groups are perhaps the most well-known example. There are no such authoritative statements for algorithmic discrimination in part because jurisprudence is still trying to catch up (Huq, 2019). There is not even a common language to address the issues.

For criminal justice risk assessments, the analogue to discrimination is an absence of "fairness," whether intentional or unintentional. Fairness can take a variety of well-defined forms, but naming conventions vary widely.

The definitions to follow arise directly from confusion tables and are easily translated into many common fairness typologies employed in risk assessment (Kleinberg et al., 2017; Kroll et al., 2017; Berk et al., 2018, Baer et al., 2020; Mitchell et al., 2021). Other definitions briefly are considered in due time. Anticipating our later data analyses, we use White criminal justice offenders and Black criminal justice offenders at their arraignments as illustrations of the groups for which fairness is sought. Fairness centers on their algorithmic forecasts of risk. Is the *algorithmic output* fair?

- *Prediction parity* – The predictive distributions across protected groups are the same. Predictive parity can be estimated with test data by, for example, the proportions of Black offenders or White offenders forecasted to be arrested after an arraignment release. Prediction parity is sometimes called demographic parity.

  Prediction parity is judged by the risk tool output, not the decisions that follow or any subsequent actions or occurrences. An important implication is that the actual outcome class to be forecasted (e.g., an arrest) has no role in the definition of prediction parity. As a result, prediction parity has been criticized as unsatisfactory and even irrelevant (Hardt et al., 2016). Yet, an absence of prediction parity may be linked to "mass incarceration," which in practice cannot easily be disregarded. Moreover, requiring the inclusion of the actual outcome class in fairness definitions leads to challenges we address shortly.

- *Classification parity* – The false positive rates and false negative rates are the same for across protected groups. A false positive denotes that a risk algorithm incorrectly classified a case with a negative class label as a case with a positive class label. A false negative denotes that a risk algorithm incorrectly classified a case with a positive class label as a case with a negative class label.[1] The false positive rate and a false negative rate are the respective probabilities that the algorithm classifies outcomes erroneously. When there are more than two outcome classes, classification parity follows from the same reasoning, but there are no common naming conventions.

  Ideally, false positive and false negative rates are estimated with test data. Classification error for a particular outcome class, such as an

[1] Which outcome class is a positive and which is a negative is determined by the subject matter or policy being addressed. In the analysis to follow, an arrest for a violent crime is a negative, and the absence of such an arrest is a positive.

arrest, is the proportion of subjects erroneously classified as not arrested among all who actually were arrested. If an arrest is the positive class, for example, one has an estimate of the false negative rate. More formally,

$$\begin{aligned} &\text{Classification Error (for an arrest)} \\ &:= \frac{\sum_{i \in \text{test}} \mathbb{1}\{\widehat{Y}_i \neq Y_i, Y_i = \text{arrest}\}}{\sum_{i \in \text{test}} \mathbb{1}\{Y_i = \text{arrest}\}}. \end{aligned} \tag{1}$$

Y_i is the true outcome for subject i in test data, and $\widehat{Y}_i$ is the forecasted, test data outcome from the trained classifier (i.e., usually the outcome with the highest estimated probability). The classification error in (1) is an estimator of $\mathbb{P}(\widehat{Y} \neq Y | Y = \text{arrest})$. Note that one conditions on the true outcome class.

Classification error, whether through false positives or false negatives, has played a central role in fairness discussions by statisticians and computer scientists (Baer et al., 2020). However, it is often irrelevant to stakeholders, who in practice care far more about forecasting accuracy. In real forecasting settings, the actual outcome is unknown, and any subsequent decisions can be primarily informed the forecasted outcome. Further, emphasizing classification may favor interpretations akin to the prosecutor's fallacy (Thompson and Schumann, 1987); classification accuracy is used inappropriately to evaluate forecasting accuracy.

- *Forecasting accuracy parity* – Each outcome class is forecasted with equal accuracy for each protected group. A forecast is incorrect if the forecasted outcome does not correspond to the actual outcome. In contrast to classification parity, one conditions on the forecasted outcome not the actual outcome.

  Here too, estimation should be undertaken with test data. The forecasting error for a particular outcome, such as an arrest, is the proportion of subjects erroneously forecasted to not be arrested among all subjects for whom an arrest was the forecast. Formally,

$$\begin{aligned} &\text{Forecasting Error (for an arrest)} \\ &:= \frac{\sum_{i \in \text{test}} \mathbb{1}\{\widehat{Y}_i \neq Y_i, \widehat{Y}_i = \text{arrest}\}}{\sum_{i \in \text{test}} \mathbb{1}\{\widehat{Y}_i = \text{arrest}\}}. \end{aligned} \tag{2}$$

The notation Y_i and $\widehat{Y}_i$ is the same as for classification parity. Implemented with test data, forecasting error (2) is an estimator of $\mathbb{P}(\widehat{Y} \neq Y|\widehat{Y} = \text{arrest})$.

In some formulations, achieving forecasting accuracy parity requires that the forecasts are calibrated. For example, suppose a risk tool projects for certain offenders a probability of .68 for an arrest. For calibration, the actual arrest probability for all such offenders must also .68. (Baer at al., 2020). The same reasoning applies over the full set of predicted arrest probabilities. By itself, this criterion is silent on fairness, but it restricts discussion of forecasting accuracy parity to applications in which risk tools is by this yardstick performing well.

- *Cost Ratio parity* – The relative costs of false negatives to false positive (or the reciprocal), as defined above, are the same for each protected group. The cost ratio determines the way in which a risk assessment classifier trades false positives against false negatives. Commonly, some risk assessment errors are more costly than others, but the relative costs of those errors should be same for every protected group. The same reasoning applies when there are more than two outcome classes.[2]

2.1 Some Complications in Practice

Operationalizing these fairness definitions is challenging. Practice usually demands that when each kind of parity is evaluated, some form of direct and legitimate comparability is enforced. Whether individuals or groups are the observational units, they must be "similarly situated."

For much of the current fairness literature, whether observational units are similarly situated is primarily a technical problem that boils down to methods that adjust for confounders, sometimes in a causal model. Typically

[2]These costs are rarely monetized. How would one measure in dollars the "pain and suffering" a homicide victim's family or the psychological trauma of neighborhood children who witness a homicide? What matters for the risk algorithm is *relative* costs. For example, failing to accurately identify a prison inmate who after release commits a murder will be seen by many stakeholders as far more costly than failing to accurately identify a prison inmate who after release becomes a model citizen. The cost ratio might 5/1. In practice, the desired relative costs are a policy choice made by stakeholders that, in turn, is built into the risk algorithm. If no such policy choice is made, the algorithm necessarily makes one that can be very different from stakeholder preferences and even common sense. Cost ratios can affect forecasted risk, often dramatically.

overlooked is that candidate confounders possess normative as well as causal content, and both affect how confounders are selected. On the closely related notion of culpability, Horder observes (1993: 215) "... our criminal law shows itself to be the product of the shared history of cultural-moral evolution, assumptions, and conflicts that is the mark of a community of principle." As a result, controversies over fairness often begin with stark normative disagreements about what it means to be similarly situated. For example, should an offender's juvenile record matter in determining whether cases are similarly situated? The answer depends in part on how in jurisprudence psychosocial maturity is related to culpability (Loeffler and Chalfin, 2017).

Normative considerations also can create unresolved incongruities. Official sentencing guidelines, for example, often prescribe that defendants convicted of the same crimes and with the same criminal records, should receive the same sentences. Under these specific guidelines, such defendants are similarly situated (Ostrom et al., 2003: chapter 1). "Extralegal" factors such as gender, race, and income are not properly included in that determination. But if "criminal records" are significantly a product of gender, race, and income, should they not be extralegal as well? The extensive literature on fairness cited earlier makes clear that there is no satisfactory answer in sight. In the pages ahead, we offer a pragmatic way forward.

There no empirical standard for how small disparities in parity must be for the parity to be acceptable, although most stakeholders agrees that small disparities may suffice.[3] Interpretations of "small" will be contentious because harm depends on facts and judgements that are easily disputed. Moreover, there usually is no clear threshold at which some amount harm becomes too much harm. Similar issues arise across the wide variety litigation domains (Gastwirth, 2000). The fairness literature has been silent on the matter, and we do not address it here. It is peripheral to our discussion of fairness, but for fair algorithms to be used effectively in practice, a binding resolution is required.

In contrast, there is a rich and persuasive literature on provable tradeoffs between certain forms of parity and between parity, accuracy and transparency (Kleinberg et al., 2017; Barocas et al., 2018; Coglianese and Lehr, 2019; Kearns and Roth, 2020; Diana et al., 2021; Mishler and Kennedy, 2021). These tradeoffs are of more than mathematical interest because they affect virtually all proposals to make algorithms more fair. Compromises are various kinds are in practice inevitable. There does not seem to be at

[3]A tactic that has been used to preclude algorithmic risk assessment altogether is to insist on exact parity.

this point any technical resolution allowing stakeholders to have it all.

2.2 Counterfactuals: Internal and External Fairness

Despite the challenges, one has estimators for the four kind of parity that can be employed with the usual test data. Organizing the test data separately into a confusion table for each protected group, one easily can consider the degree to which each kind of parity is achieved. When each kind of parity is examined using a combination of test data and algorithmic output, one is assessing what we call *internal fairness*.

Prediction parity is a very important special case. Because outcome labels are unnecessary for its definition, one legitimately can examine this form of algorithmic fairness from test data and algorithmic output alone. If the distribution of outcome forecasts is not sufficiently same across protected groups, one properly may claim that prediction parity has been violated.[4]

Such claims may really matter. Recall that an absence of prediction parity can be a driving force for mass incarceration. Mass incarceration usually refers to the over-representation of Blacks in the jails and prisons in the United States and is seen by some as modern extension of slavery (Waquant, 2002). It has been a "hot button" issue for over a decade (Lynch, 2011), perhaps second only to police shootings in visibility and rancor.

For classification parity, forecasting accuracy parity, and cost ratio parity, one must have the labels for the actual outcomes because those labels are built into their fairness definitions. Typically, each observation in the training and test data has such a label. Yet, we have defined internal fairness such that it depends on test data outcome labels that represent a status quo, which can include racial disparities carried forward as an algorithm is trained and fairness is assessed. For our approach to fairness that treats Black offenders as if they are White, such labels may be especially misleading. We prove in Appendix D that, except for prediction parity, isolating the fairness of a risk algorithm requires untestable causal assumptions that cannot be enforced in practice. These include rank preservation and strong unconfoundedness in how the criminal justice system might treat a person

[4]There are many simple ways to achieve prediction parity. For example, at arraignment the magistrate might flip a fair coin. Heads means the offender is released. Tails means that the offender is detained. All offenders regardless of race or gender are released with a probability of .50. But the magistrate properly would be accused of procedural capriciousness (Holewinski, 2002). An approach that detained everyone would also achieve prediction parity but probably would fail by the criterion of arbitrariness. In real settings, methods that create prediction parity must also pass jurisprudential muster.

if he or she were of a certain race.[5]

In short, no matter the number of observations or how the data are collected, test data outcome labels for Black offenders, such as an arrest, cannot be assumed to accurately capture the counterfactual of policing in which Blacks are treated the same as Whites. Yet, accurate information about counterfactuals is a prerequisite for what we call *external fairness*. External fairness is a function for protected groups of the risk algorithm and unobserved fair counterfactual.

There are other fairness definitions advocated by some for which the concerns are the same. "Separation" requires that the forecasted outcome be independent of protected group membership, given the true outcome (Hardt et al., 2016). "Sufficiency" requires that the true outcome be independent of the protected group membership, given the forecasted outcome (Baer et al., 2020). "Predictive parity" (not prediction parity) requires that the probability of the true outcome conditional on the forecasted outcome be the same when one also conditions on protected group membership (Chouldechova, 2017). As before, the fundamental challenge is that the counterfactual outcome for Black offenders is not available in the test data.

Consider, for example, *counterfactual* classification parity as a form of external fairness:

$$\begin{aligned} &\text{Counterfactual Classification Error (for no arrest)} \\ &:= \frac{\sum_{i \in \text{test}} \mathbb{1}\{\widehat{Y}_i \neq Y_i, Y_i^* = \text{no arrest}\}}{\sum_{i \in \text{test}} \mathbb{1}\{Y_i^* = \text{no arrest}\}}. \end{aligned} \tag{3}$$

where, Y_i^* in our formulation denotes the underlying counterfactual outcome for a Black offender treated upon release as a similarly situated White offender, and $\widehat{Y}_i$ is the forecast from the trained algorithm. The need for similarly situated comparisons is built into the application of each form of counterfactual fairness.

Counterfactual outcomes underscore that all criminal justice risk algorithms necessarily have a circumscribed reach. They are a computational procedures that transform the input with which it is provided into information intended to help inform decisions. One can legitimately ask, therefore, if the algorithmic output by *itself* is fair; is the algorithm "intrinsically" fair? A criminal justice risk algorithm is not responsible for the deployment of police assets, the tactics that police employ, an easy access to firearms,

[5]In passing, related issues arise when DAGs and causal reasoning are used to isolate the impact of race on any criminal justice outcome (Baer et al., 2020).

gang rivalries, and myriad other factors that can affect the likelihood of a post-release arrest. Outcome labels in test data incorporate these factors over which a risk algorithm has no control. These factors can produce misleading assessments for classification parity, forecasting accuracy parity and cost ratio parity. One can have an algorithm that itself is fair despite what an analysis using test data shows. Put more strongly, stakeholders are being unrealistic to demand a fair risk algorithm fix widespread inequities in the criminal justice system and the social world more generally.

It follows that proper evaluations of classification parity, forecasting accuracy, and cost ratio parity may be at this point largely aspirational. Within our approach to fairness and later empirical application, there is no apparent path to sound estimates of the counterfactual world in which race has no role in arrests after an arraignment release such that Black offenders are treated the same as similarly situated, White offenders. One might choose instead to assume that race is unrelated to the many causes of an arrest, but that would be contrary to the overwhelming weight of evidence (Robert Wood Johnson Foundation, 2017; Rucker and Richeson, 2021; Muller, 2021).[6]

Other forms of counterfactual reasoning have been proposed for consideration of fairness. For example, Misher and Kennedy (2021; section 2.2) note the potential importance of a race counterfactual somewhat like ours. (i.e., What would happen if an individual's race were different?) Nabi and colleagues (2019) offer a DAG formulation for fair policies steeped in counterfactuals but requiring assumptions that would be difficult to defend in criminal justice settings. Kusner and colleagues (2018) provide a DAG framework for examining racial counterfactuals, but it too requires rather daunting assumptions. Imai and Jiang (2021) use counterfactual reasoning to define the concept of "principal fairness," which if achieved, subsumes many of the most common kinds of fairness, but requires conditioning on all relevant confounders. The formulation proposed by De Lara and colleagues (2021), using counterfactual thinking, optimal transport and related tools, is strongly connected to some aspects of our formulation, but makes no

[6]Why race is so strongly implicated does not require racial animus by police and other criminal justice agents. To take a tragic example, the numbers of homicides and shootings recently have increased substantially in many American cities. The vast majority of victims are Black. The vast majority of perpetrators are Black. These facts are not a product of racist "overpolicing," racially targeted "stop-and-frisk" or racially slanted data. Insofar as the causes are understood, they go to easy access to firearms and long-term structural issues ubiquitous in disadvantaged neighborhoods, perhaps exacerbated by the COVID pandemic (Sorenson et al., 2021).

distinction between disparities and unfairness such that, for example, there can be no "bona fide occupational qualifications" under Title VII of the U.S. Civil Rights Act of 1964, Section 625.[7]

These interesting papers (see also Mitchell et al., 2020) apply a rich variety of counterfactual ideas to fairness. However, they do not consider many of the foundational fairness concerns raised earlier, such as the need for a fairness baseline or the meaning of "similarly situated." As a result, they are currently are some distance from applications in real settings where risk algorithms can directly affect people's lives. They also differ substantially in method and focus from our approach, to which we turn now.

3 Achieving A Fair Criminal Justice Risk Assessment Procedure

When a criminal justice risk algorithm is trained, the data consist of two parts. There are predictors, and there are outcome class labels. The former are used to fit the latter. Both can be responsible when racial disparities are carried forward when a risk algorithm is trained.

Even when race is not included among the predictors, it is widely understood that all predictors can have racial content (Berk, 2009). For example, the number of prior arrests may be on the average larger for Blacks offenders than for White offenders. Black offenders may also tend to be younger and have been first arrested at an earlier age. Customary outcome classes used in training algorithmic risk assessments represent adverse contacts with the criminal justice system, such as arrests or convictions. Associations with race are typical here as well.

There are lively, ongoing discussions about why such racial associations exist (Alpert et al., 2007; Harcourt, 2007; Gelman et al., 2012; Grogger and Ridgeway, 2012; Starr, 2014; Tonrey, 2014; Stewart et al., 2020). Some explanations rest on charges of racial animus in the criminal justice system, some focus on criminal justice practices properly motivated but with untoward effects, and some take a step back to large inequalities in society that propagate crime.

[7]For some occupations, a person's sex, religion, or national origin may be necessary to successfully undertake a job that is a normal activity in a business or enterprise. There can also be legitimate performance requirements as long as all job applicants have an opportunity demonstrate whether they can fulfill those performance requirements. The issues can be subtle. For example, a performance test may not represent sufficiently the actual job requirements.

It is likely that some mix from each perspective is relevant, but there is no credible integration yet available. Therefore, we take no position in this paper on explanations for the role of race, although a range of associations are empirically demonstrable. Rather, we more simply build on professed differences in privilege consistent with extensive research (Rothenberg, 2008; Rocque, 2011; Van Cleve and Mayes, 2015; Leonard, 2017; Wallis, 2017; Bhopal, 2018; Edwards et al., 2019; Jackson, 2019; GBD 2019 Police Violence Subnational Collaborators, 2021). From a pragmatic point of view, we are responding as well to common suppositions and frequent stakeholder claims.

We proceed in three steps: (1) training the risk algorithm in a novel manner, (2) transporting the predictor distribution from the less privileged group to the more privileged group, and (3) constructing conformal prediction sets to serve as risk forecasts. Each step is briefly summarized next.

3.1 Training The Classifier

We train the risk algorithm *only* on White offenders; both the predictors and the outcome are taken solely from Whites. No racial distinctions between Black and White offenders can be "baked into" the risk algorithm because the algorithm has information exclusively on White offenders; race is a constant. The algorithm is necessarily blind to any racial differences.

Subsequently, when risk forecasts are sought from unlabeled data for a particular White offender, one can proceed as usual by obtaining predictions from the trained classifier. However, risk forecasts for unlabeled Black offenders obtained using the White-trained algorithm still can produce racial disparities because Black offenders can have more problematic predictor distributions, whatever their cause. Black offenders on the average will then be treated by the algorithm as particularly risky White offenders leading to less favorable forecasted outcomes.

3.2 Transporting Observations Across Protected Groups

A second adjustment is required that makes comparable the joint distributions of predictors for Black offenders and White offenders. We use a form of optimal transport (Hütter and Rigollet, 2020; Manole et al., 2021; Pooladain and Niles-Weed, 2021) to make the joint predictor distribution for Black offenders like the joint predictor distribution for White offenders. To take a toy example, an arrested Black offender 18 years of age, with 3 prior robbery arrests and a first arrest at age 14, might be given predictor values

from an arrested White offender who was 20 years age with 1 prior robbery arrest, and a first arrest at age 16. In effect, we are obtaining exact or near *quantile* matches, which operationalize "similarly situated." Empirical evidence is provided later.

Our estimation procedure builds on work by Hütter and Rigollet (2020: section 6.1), based on the Kantorovich "relaxation" (Peyré and Cuturi, 2019: section 2.3), that affords a linear programming fitting algorithm (Peyré and Cuturi, 2019: chapter 3). The transported joint predictor distribution is then smoothed with a form of nonparametric regression so that it can be used to transport predictors from new, unlabled cases for which forecasts are needed. Details and pseudocode are provided in Appendix B. Further discussion can also be found in the application.

3.3 Forecasting for Individual Cases

The practical task for a criminal justice risk assessment is to forecast one or more behavioral phenomena. By training a classifier only on White offender data and evaluating fairness separately using White test data and transported Black test data, one can compare the aggregate performance of a risk assessment tool across protected groups using conventional confusion tables. There also can be forecasts for individuals that minimize Bayes risk.

Arguably, a more defensible approach rests on conformal prediction sets (Vovl et al., 2005; 2009; Shafter and Vovk, 2008; Romano et al., 2019; Kuchibhotla and Berk, 2021). The output for a categorical response variable is a prediction set with an associated probability that the set includes the true future outcome class(es). The prediction set has valid statistical properties even in finite samples. Appendix C provides a more complete treatment and pseudocode for the form of conformal inference we employ. There is further discussion in the application.

3.4 Diagrammatic Summaries of Our Risk Algorithm

Procedural details are provided in two diagrams with brief explanations for our entire risk algorithm from the training of a classifier, to the use of optimal transport, to the output of a conformal prediction set. Figure 1 addresses how the fair algorithm was constructed. Figure 2 addresses fair forecasting. Both figures are further unpacked in the appendices. In addition, the data analysis to come provides a grounded methodological discussion.

In theory, we achieve our overall goal of constructing a fair algorithmic risk tool. But in practice, there are two challenges. First, our procedures

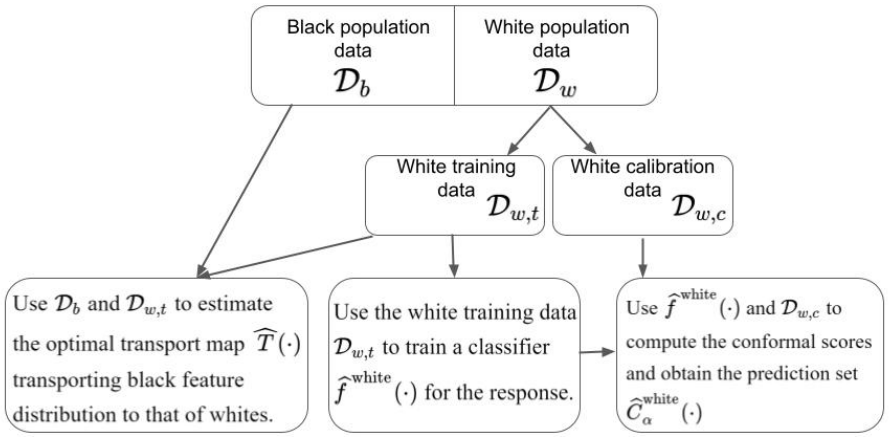

Figure 1: Flowchart for the training part of our "fair" risk algorithm. The function $\widehat{f}^{\texttt{white}}(\cdot)$ denotes the output vector of probabilities for each outcome obtained from a classifier fit on the white training data. The map $\widehat{T}(\cdot)$ denotes the estimate of the optimal transport map; see Appendix B (and Algorithm 2) for details. The set $\widehat{C}_\alpha^{\texttt{white}}(\cdot)$ is the conformal prediction set obtained using $\widehat{f}^{\texttt{white}}(\cdot)$ and the white calibration data; see Appendix C (and Algorithm 3) for details.

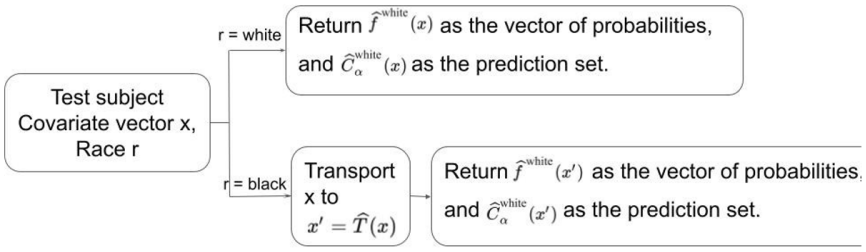

Figure 2: Flowchart for producing fair conformal prediction sets. The function $\widehat{f}^{\texttt{white}}(\cdot)$ denotes the output vector of probabilities for each outcome obtained from a classifier fit on the first split white training data. The map $\widehat{T}(\cdot)$ denotes the estimate of the optimal transport map obtained from the first split training data for whites and blacks; see Appendix B for details. The set $\widehat{C}_\alpha^{\texttt{white}}(\cdot)$ is the conformal prediction set obtained using $\widehat{f}^{\texttt{white}}(\cdot)$ and second split white training data.

must be implemented on real data. We turn to that task next. Second, as a formal matter, external fairness parities from test data will not provide valid fairness assessments without very strong assumptions. Recall that the required counterfactual outcome is unobservable in existing criminal justice datasets. The implications for policy are addressed after the application is presented.

4 The Data

We analyze a random sample of 300,000 offenders at their arraignment from a particular urban jurisdiction in the United States.[8] Because of the random sampling, the data can be treated as IID and, therefore, exchangeable. Even without random sampling, one might well be able to make an IID case because the vast majority of offenders at their arraignment are realized independently of one another.

Among those being considered for release at their arraignment, one outcome class (coded "1") to be forecasted is whether the offender will be

[8]At an arraignment, which is supposed to be held within 48 hours of an arrest, the charges are read officially to the arrested offender. The presiding magistrate then decides whether the offender can be released, sometimes with a bail bond, subject to a later return to court, or detained until that later court date.

arrested after release for a *crime of violence*. Such crimes are of special concern.[9] An absence of such an arrest (coded "0") is the alternative outcome class to be forecasted.

The follow-up time was 21 months after release. For reasons related to the ways in which competing risks were defined, 21 months was chosen as the midpoint between 18 months and 24 months. For the analysis to follow, the details are unimportant.

Candidate predictors were the usual variables routinely available in large jurisdictions. Many were extracted from adult rap sheets and analogous juvenile records. Biographical variables included race, age, gender, residential zip code, employment information, and marital status. There were overall 70 potential predictors.

In response to potential stakeholder concerns about "bias," we excluded race, zip code, marital status, employment history, juvenile record, and arrests for misdemeanors and other minor offenses. Race was excluded for obvious reasons. Zip code was excluded because, given residential patterns, it could be a close proxy for race. Employment history and marital status were eliminated because there were objections to using "life style" measures. Juvenile records was discarded because poor judgement and impulsiveness, often characteristics of young adults, are not necessarily indicators of long term criminal activity. Minor crimes and misdemeanors were dropped because many stakeholders believed that arrests for such crimes could be substantially influenced by police discretion, perhaps motivated by racial animus.

The truth underlying such concerns is not definitively known, but insofar as the discarded predictors were associated with any included predictors, potential biases remain (Berk, 2009). These decisions underscore our earlier point that there are legitimate disagreements over what features of individuals or groups should determine the when a similarly situated comparison has been properly undertaken. They also highlight the tradeoffs to be made when a suspect predictor also is an effective predictor.

In the end, the majority of the predictors were the number of prior arrests for a variety of serious crimes, and the number of counts for various charges at the arraignment. Other included predictors were whether an individual was currently on probation or parole, age, gender, the age of a first charge as an adult, and whether there were earlier arrests in the same

[9] It might seem that using a conviction rather than an arrest would convey more about the actual crime, but the vast majority of criminal trials are resolved by a guilty plea after a negotiated agreement between the defense and prosecuting attorneys. Strategic maneuvering can dominate the process. Racial factors can enter as well.

year as the current (arraignment) arrest. For the analyses to follow, there were 21 predictors.[10]

The 300,000 cases were randomly split into training data for White offenders, training data for Black offenders, test data for White offenders, and test data for Black offenders. Half the dataset was used as training data ($n = 150,000$) and half the dataset was used as test data ($n = 150,000$). Sizes of the racial splits of the training and test data were simply determined by the numbers of Black offenders and White offenders available in the data. Each racial split had at least 40,000 observations. Asymptotic performance is probably of little concern.

5 Fairness Results

We began by training a stochastic gradient boosting algorithm (Friedman, 2001) on White offenders only using the procedure *gbm* from the library *gbm* in the scripting language *R*. For illustrative purposes and consistent with many stakeholder priorities, the target cost ratio was set at 8 to 1 (Berk, 2018). Failing to correctly classify an offender who after release is arrested for a crime of violence was taken to be 8 times worse than failing to correctly classify an offender who after release is not arrested for such a crime. We were able to approximate the target cost ratio reasonably well in empirical confusion tables by weighting more heavily training cases in which there was an arrest for a crime of violence. This, in effect, changes of the prior distribution of the outcome variable.

All tuning defaults worked satisfactorily except that we chose to construct somewhat more complex fitted values than the defaults allowed.[11] The results were essentially the same when the defaults were changed by modest amounts. The number of iterations (i.e. regression trees) was determined empirically when, for a binomial loss, the reductions in the test data effectively ceased.[12]

[10]The two age-related variables and whether there were other arrests within the past year are "dynamic variables" because they can change over time. For other criminal justice decisions, such as whether to grant parole, there can be many more dynamic variables (e.g., work history in prison). At an arraignment, one is limited largely to what could be extracted from existing rap sheets and current charges.

[11]We used greater interaction depth to better approximate interpolating classifiers (Wyner et al., 2015). Even after weighting, we were trying to fit relatively rare outcomes. We needed an ensemble of regression trees each with many recursive partitions of the data.

[12]Because of the random sampling used by the *gbm* algorithm, the number of iterations in principle can vary a bit with each fit of the data. Also, the number of trees can

5.1 Algorithmic Performance Results for White Offenders

Algorithmic risk assessments can be especially challenging when the marginal distribution of the outcome is highly unbalanced. For binary outcomes, this means that if criminal justice decision-makers always forecast the most common outcome class, they will be correct the vast majority of the time. It is difficult for an algorithm to forecast more accurately. Because post-arraignment arrests for a crime of violence are well-known to be relatively rare, we were faced with the same challenge that, nevertheless, provided an instructive test bed for examining fairness.

To set the stage, Table 1 is the confusion table for White offenders using the risk algorithm trained on Whites and test data for Whites.[13] Perhaps the main message is that if arraignment releases were precluded solely by the risk algorithm when arrests for a violent crime were forecasted, more violent crime might be prevented.

Here's the reasoning. From the outcome marginal distribution of an arrest for a crime of violence, minimizing Bayes loss always counsels forecasting no such arrest after an arraignment release. That forecast would be wrong for 7.5% of the White offenders. From the left column in Table 1, when the algorithm forecasts no arrest for a violent crime after an arraignment, the forecast is wrong for 5% of the White offenders. If from the marginal distribution one always forecasted an arrest for a crime of violence, the forecast would be wrong for 92.5% of the White offenders. From right column in Table 1, the algorithm is mistaken for 85% of the white offenders. These are modest improvements in percentage units, but given the large number of White offenders, over 2000 of violent crimes might be averted if the risk algorithm determined the arraignment release decision.

Table 1: Test Data Confusion Table for White Offenders Using the White-Trained Algorithm (28% Predicted to Fail, 7.5% Actually Fail)

Actual Outcome	No Violence Predicted	Violence Predicted	Classification Error
No Violence	31630 (true positives)	11246 (false positives)	.26
Violence	1527 (false negatives)	1975 (true negatives)	.47
Forecasting Error	.05	.85	

arbitrarily vary by about 25% with very little impact.

[13]The empirical cost ratio in Table 1 is 11246/1527, which is 7.4 to 1. It is very difficult in practice to the arrive exactly at the target cost ratio, but cost ratios within about 20% of the target usually lead similar confusion tables.

At the same time, forecasting accuracy is shaped substantially by the cost ratio. Because the target cost ratio treats false negatives as 8 times more costly than false positives, predictions of violence in Table 1 are dominated by false positives. This follows directly and necessarily from the imposed tradeoffs. Releasing violent offenders is seen to be so costly that even a hint of future violence is taken seriously. But then, many mistakes are made when an arrest for a crime of violence is forecasted. In trade, when the algorithm forecasts no arrest for a violent crime, it is rarely wrong; there are relatively few false negatives. This too follows from the target cost ratio. If even a hint of violence is taken seriously, those for whom there is no such hint are likely to be very low risk releases. In short, with different target cost ratios, the balance of false positives to false negatives would change, perhaps dramatically, which means that forecast accuracy would change as well.[14]

The aversion to false negatives results in a projection that 28% of the White offenders will fail through a post-release arrest for a violent crime. In the test data, only 7.5% actually fail in this manner. The policy-determined tradeoff between false positives and false negatives produces what some call "overprediction." With different tradeoff choices, overprediction could be made better or worse. In either case, there would likely be important concerns to reconsider.

Overprediction concerns become even more prominent if test data for Black offenders are used to forecast post-arraignment crime. When the Black test data are employed with the algorithm trained on Whites, 41% of the Black offenders are forecasted to be arrested for a crime of violence, whereas 11% actually are. The base rate is a bit higher for Black offenders (i.e., 11% compared to 7.5%), but the fraction projected to arrested for a violent crime increases substantially (i.e., from 28% to 41%). As emphasized earlier, the latter disparity cannot be a product of racial differences in the algorithmic machinery. It is trained only on Whites. The likely culprit is racial disparities in the test data provided to the classifier. In any case, there is clear evidence from the test data that predictive parity is not achieved solely by training the risk algorithm on White offenders.

[14]Note that an algorithm is not a model. "An algorithm is nothing more than a very precisely specified series of instructions for performing some concrete task"(Kearns and Roth, 2020: page 4). It is not meant to explain some phenomenon, depict causal effects, or characterize how the data were generated. Consequently, the forecasted classes have nothing to say about *why* either outcome class is realized. Some arrests might be "rightous," some might be a direct or indirect product of race, and many can be a mix of the two, but the precise mechanisms are not manifest.

Table 1 also provides conventional test data performance statistics for the false positive rate, the false negative rate, forecasting accuracy for an arrest for a crime of violence, forecasting accuracy for no arrest, and the empirical cost ratio. For example, the false positive rate is .26 and when no arrest is forecasted, it is wrong 5% of the time. Comparisons could be made to the full confusion table for Black offenders, but the limitations of internal fairness would intrude. We postpone a discussion until more results are reported.

Overall, performance is roughly comparable to other criminal justice risk assessments and probably worth close scrutiny by stakeholders (Berk, 2018). No doubt some changes in the risk classifier would be requested, and the results would be reviewed for potential alternatives to implement. Our intent, however, is not to claim that the results in Table 1 are definitive. Rather, they provide a realistic context for an empirical consideration of fairness.

5.2 Optimal Transport Performance

We applied optimal transport, briefly described earlier, using the procedure *transport* in R.[15] No tuning in a conventional sense was needed. However the γ coupling matrix was $n \times n$ which meant that memory considerations came to the fore. We tried 1000, 2000, 3000, 4000 and 5000 observations in ascending order. At 5000 observations, computer memory was exceeded. We proceeded, therefore using 4000 randomly selected test data observations for Black offenders.[16]

A key diagnostic for optimal transport in practice is how well the transported joint probability distribution compares to the destination joint probability distribution. Summary fit statistics are too coarse. They can mask more than they reveal. A better option is compare the correlation matrices from the two distributions. For these results, there were no glaring inconsistencies, but it was difficult to translate differences in the correlations into implications for fairness.

Perhaps the most instructive diagnostic simply is to compare the marginal distributions for each predictor. Using histograms, we undertook such comparisons for each of the 21 predictors. The following figures show the results

[15]There are several computational options for estimating the coupling matrix. We used the default "revsimplex" (Luenberger and Ye 2008, Section 6.4) that worked very well.

[16]We were surprised by how well *transport* performed with just 4000 observations. At first, we were skeptical, and we tried several toy examples and datasets previously used by others. We found no reasons to discount our results.

for the predictors that dominated the fit when the risk algorithm was trained on the data for White offenders; these are the predictors that mattered most. There were similar optimal transport results for the other, less important, predictors.

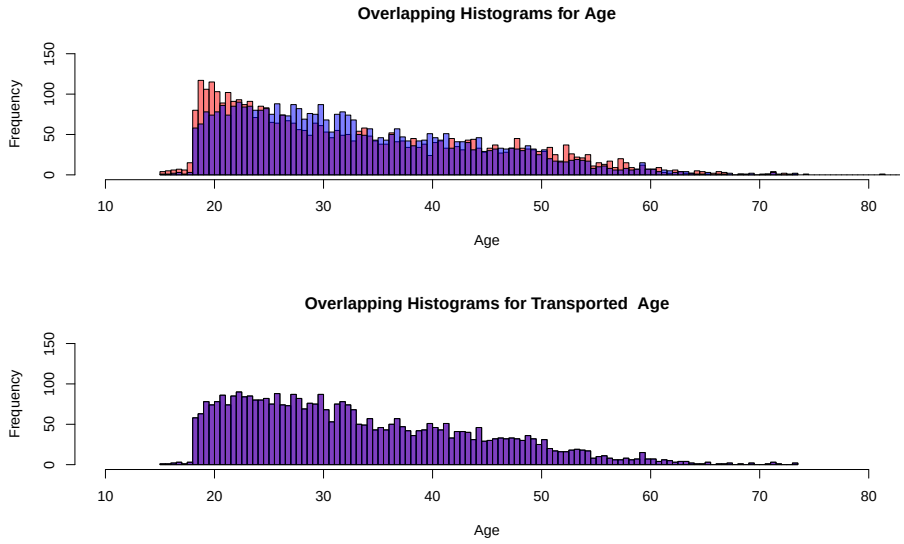

Figure 3: Histograms for an Offender's Age and Transported Age (Black offenders in orange, White offenders in blue, overlap in purple, N= 4000)

It is well-known that younger individuals have a greater affinity for violent crime than older individuals. Figure 3, constructed from the 4000 randomly selected observations provided to the procedure *transport*, shows the results for the age of the offender. The top histogram compares the test data distribution for Whites in blue to the test data distribution for Blacks in orange. The purple rectangles show where the two distributions overlap. Clearly, Black offenders at arraignment are on the average somewhat younger, especially for the youngest ages that place an offender at the greatest risk. The bottom histogram is constructed in the same manner but now, the White age distribution from test data is compared to the transported Black distribution. There are no apparent differences between the two. Clearly, Black offenders are no longer overrepresented among the youngest ages.

One must be clear that Figure 3 shows how the age *distribution* for the White test data and the Black transported test data are made virtually

indistinguishable. This does not imply exact one to one matching of Black offenders to White offenders in units of years. The matching going on as part of the linear programming algorithm is by quantiles and can one to many. More details are provided in Appendix C.

The performance of optimal transport in the bottom histogram may seem too good to be true. However, despite some distributional differences in the top histogram that may matter for risk, the overall shape of the two distributions is very similar. Both peak at low values and gradually decline in an almost linear fashion toward a long right tail. One should expect optimal transport to perform well under such circumstances.

Perhaps more surprising is that the two distributions are so similar to begin with. But arrests are a winnowing process affecting all protected groups in similar ways. The pool of individuals who are arrested is more alike than the overall populations from which they come. Regardless of race, the pool disproportionately tends to be young, male, unemployed, and unmarried, with appreciable previous police contact. It is commonly said that less than 10% of the overall population are responsible for more than 50% of the crime (e.g., Nath, 2006) This disparity is reflected in the backgrounds of individuals who are arrested, coming more likely from that 10%.

Figure 4, constructed from the same 4000 observations, shows the results for an offender's number of prior arrests for crimes of violence, which is also known to be associated with post-arraignment violent crime. It is apparent in the top histogram that Black offenders have many more priors up to about 40, at which point there are too few cases to draw any conclusions. After the application of optimal transport, the bottom histogram shows no apparent differences. As before, the two distributions were not dramatically different before optimal transport was applied.

Figure 5, using the same 4000 observations, shows the results for the earliest age at which an offender was charged as an adult. Offenders who start their criminal activities at a younger age are more crime-prone subsequently. From the top histogram, Black offenders are more common than White offenders at the younger ages. That disparity disappears in the bottom histogram after optimal transport is applied, no doubt aided by the similar shapes of the two distributions. Optimal transport seems to perform as hoped to remove racial disparities in the two joint predictor distributions.

But, the effectiveness of optimal transport in a *forecasting* setting remains to be addressed. We converted the transported joint predictor distribution constructed from the Black offender test data into conformal scores like those used in forecasting. The classifier trained on White offenders

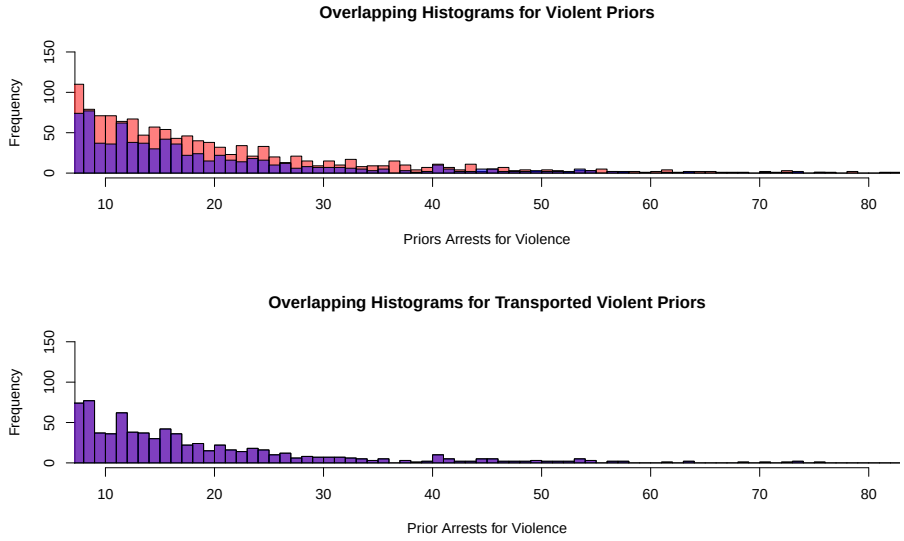

Figure 4: Histograms for the Number of Prior Arrests for a Crime of Violence and the Transported Number of Prior Arrests for a Crime of Violence (Black offenders in orange, White offenders in blue, overlap in purple, N= 4000)

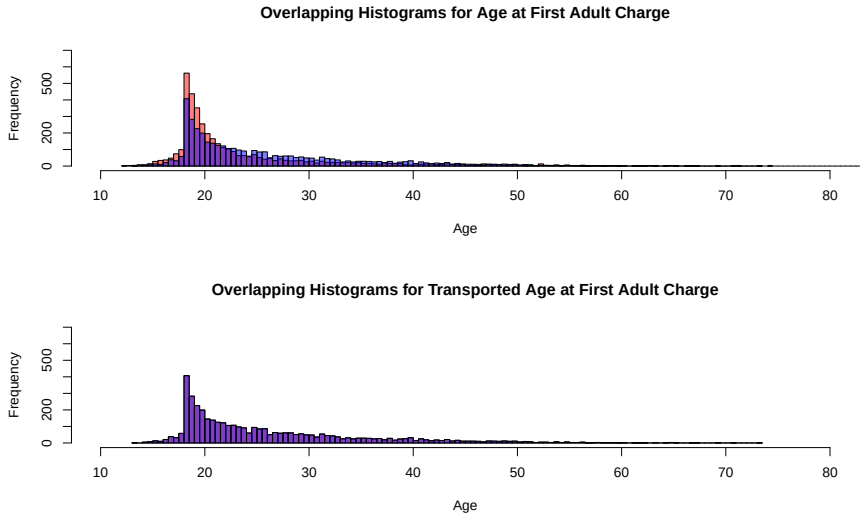

Figure 5: Histograms for the Age of the first Adult Charge and the Transported Age of the first Adult Charge (Black offenders in orange, White offenders in blue, overlap in purple, N = 4000)

was tasked with producing the probabilities of an arrest for a violent crime. These probabilities were then subtracted from "1" and from "0," yielding Black offender conformal scores for the two possible outcome classes. In other words, we were proceeding for illustrative purposes as if the Black test data were unlabeled, just as new data would be when forecasts are needed. Conformal prediction procedures for more than two outcome classes can be found in Kochibhotla and Berk (2021).

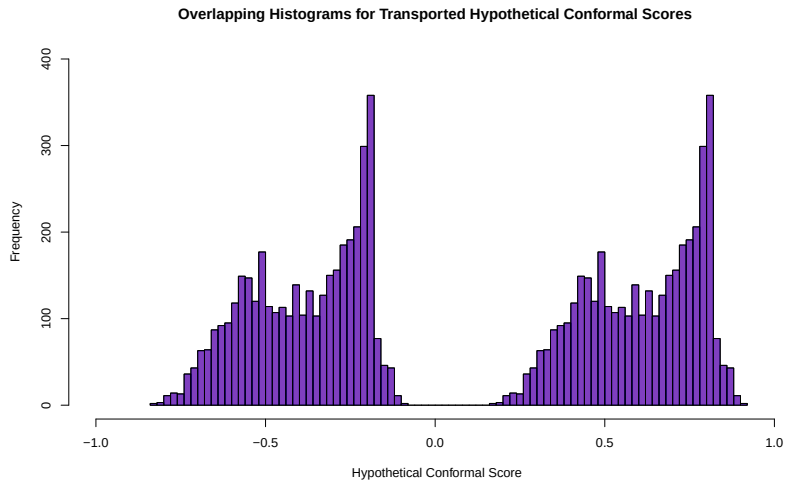

Figure 6: Histograms for White Conformal Scores and Transported Black Conformal Scores (Black offenders in orange, White offenders in blue, overlap in purple, N = 4000)

The two conformal score distributions, one for the forecasted 1s and one forecasted 0s, were then compared to the White offender conformal scores computed in the same manner from the White test data (i.e., as if the data were unlabeled). Ideally, there would be no apparent racial differences.

Figure 6 shows the results. The histogram to the left contains the conformal scores for cases in which the hypothetical outcome is no arrest for a crime of violence. The histogram to the right contains the conformal scores for cases in which the hypothetical outcome is an arrest for a crime of violence. As before, the histogram rectangles for Black offenders are in orange, the histogram rectangles for White offenders are in blue, and the overlap rectangles are in purple. Both histograms are entirely purple. The test data distribution of conformal scores for White offenders and Black offend-

ers are for all practical purposes the same.[17] The claim is strengthened that for classifiers trained on White data, conformal prediction parity might be improved by optimal transport.

When actual forecasts are required, there is another step. For Black offenders, one needs a procedure that converts the predictor values for each unlabeled case into its corresponding transported values. These new cases were not available for the earlier optimal transport exercise, and repeating optimal transport for each new unlabled case was at least impractical. Hütter and Rigollet (2020), instead suggest fitting a multivariate nonparametric smoother and using that to get good approximations of transported conformal scores. An added benefit is that the full range of predictors for the unlabeled data can have comparable transported values. We applied random forests.[18]

One begins with a conventional $n \times p$ matrix of the original joint predictor test data distribution for Black offenders denoted by X_b^{Test}. One also has an $n \times p$ matrix of the transported joint predictor distribution for Black offenders denoted by X_b^{Trans}. Each column of X_b^{Trans} is regressed in turn on X_b^{Test}. Here, that means repeating this operation 21 times. Subsequently, the predictors for any unlabeled case could be used as input for the fitted random forest to output approximations of each transported predictor. These are then collected in a matrix denoted by $\hat{X}_b^{Trans}$. When conformal scores for forecasting are needed, these approximations can be employed as usual as if they were the actual transported predictor values.

There is evidence from Figure 7 that some of the overlap in Figure 6 is lost because of the random forest approximation. For both histograms, Black offenders are somewhat overrepresented at smaller values and White offenders are somewhat overrepresented at larger values. The performance of optimal transport has been degraded. If a conventional prediction region were imposed, Black offenders might be more commonly forecasted to be arrested for a violent crime and White offenders might be more commonly forecasted not to be arrested for a violent crime.

One might do better if our random forests application were better tuned or if some superior fitting procedure were applied. But, the conformal scores that matter for the contents of conformal prediction sets are only those that

[17] The N for Whites set a 4000 because that is the number of cases used by the optimal transport procedure. This also makes comparisons between histograms easier to implement.

[18] For example, an age of 57 for an unlabled case may not exist in the transported data unless a smoother is applied. Random forests solves such problems because the inequalities responsible for recursive partitioning tree by tree leave no gaps in predictor values.

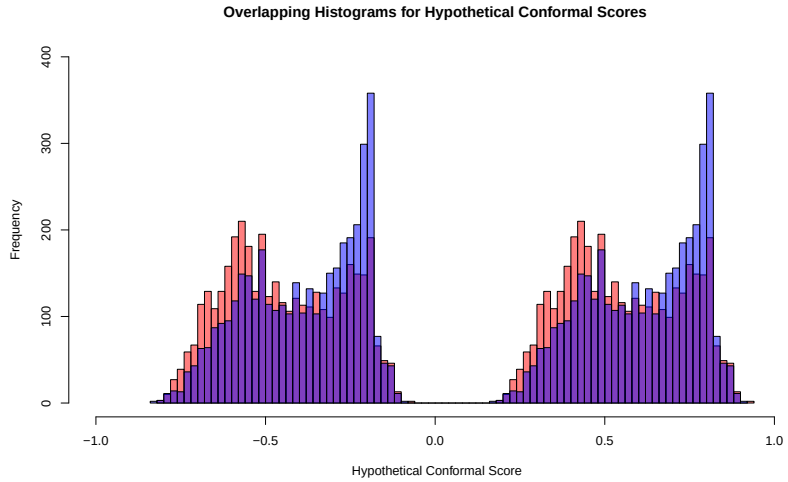

Figure 7: Histograms for White Conformal Scores and Smoothed Black Conformal Scores with the 0 Outcome Label on the Left and 1 Outcome Label on the Right (Black offenders in orange, White offenders in blue, overlap in purple, N= 4000)

fall in the near neighborhood of the prediction region's boundaries. Some may view this as a form of robustness. We need to consider whether in practice the reduction is overlap matters for fairness.

6 Evaluating Fairness in the Algorithmic Determinations of Risk

Recall that even after training the classifier only on White offenders, racial disparities remained, and these disparities were caused by differences at arraignment between the joint predictor distributions for Black offenders and White offenders. We have shown that optimal transport can remove such disparities. But they are perhaps re-introduced when forecasts need to be made.

We have argued elsewhere (Kuchibhotla and Berk, 2021) that when forecasting is the goal, accuracy is more usefully captured by conformal prediction sets than by confusion tables. One major problem with confusion tables is that forecasts are justified by minimizing Bayes loss even if there are very small differences between the estimated probabilities. For example, a dis-

tinction of .90 versus .10 for an arrest compared to no arrest, produces same forecast as a distinction of .51 versus .49. In other words, the *reliability* of the forecast is ignored. Another problem is that there are no finite sample coverage guarantees for confusion table forecasts, which can be problematic for real decisions when the number of cases is modest.

Nevertheless, standard practice and many scholarly treatments of fairness emphasize examinations of confusion tables. We take a rather different approach using conformal prediction sets. But interested readers can see in Appendix E that if one transports a joint distribution including the *response variable* as well the predictors, the confusion table from test data for White offenders is effectively the same as the confusion table from transported test data for Black offenders. For many stakeholders and some scholars, this may suffice for algorithmic fairness.

6.1 Results for Conformal Prediction Sets

We focus on predictive parity. For these data, predictive parity requires that conformal prediction sets for Black offenders and White offenders are substantially the same for a given coverage probability. Table 3 shows the results for our data on offenders at arraignment. A coverage probability was a specified somewhat arbitrarily as .95.[19]

Prediction Set	White Test Data	Black Transported data	Black Smoothed data
$\{\emptyset\}$	0.0	0.0	0.0
$\{0\}$	0.58	0.58	0.54
$\{1\}$	0.03	0.03	0.03
$\{1, 0\}$	0.39	0.39	0.43

Table 2: Estimated Proportions for Conformal Prediction Sets from White Test Data Predictors, Transported Black Predictors and Fitted Transported Black Predictors all at $1 - \alpha = .95$

In order to obtain a sufficient number of observations for instructive results, we treated the test data for White and Black offenders as if the labels were unknown. For White offenders, we obtained conformal prediction sets

[19]The coverage probability is, in effect, a tuning parameter that can be varied by the researcher. For example, with smaller coverage probabilities, the conformal prediction sets tend to contain fewer elements. There are potential gains in precision in trade for losses in certainty.

as one ordinarily would. This is a very important feature of our procedures because it guarantees that our procedures do not alter the conformal prediction sets computed for Whites. Overall, therefore, no white offenders and White offenders as a group are not made worse off by the fairness adjustments. For them, there are none.

For Black offenders we proceeded in the same manner except using two different predictor distributions: for the transported, joint predictor distribution and for it random forests smoothed, transported approximation. For Black offenders, there were 4000 observations. For White offenders, there about 10 times more.

The rows in Table 3 contain results for the four possible conformal prediction sets, where "1" denotes an arrest for a violent crime and "0" denotes no such arrest. These prediction sets are shown in the first column. In the second column are the proportions of times each prediction set materialized for the White test data. There were no empty sets implying that there were no outlier conformal scores. The most common prediction set was $\{0\}$ followed closely by $\{1, 0\}$. The prediction set $\{1\}$ surfaced very rarely. One might have expected these results because the outcome class of no arrest for a crime of violence dominated the marginal distribution of the response variable.

One important implication from the conformal prediction sets for Whites is that a substantial number of the arrest forecasts produced by the risk classifier and reported Table 1 might properly be seen as unreliable. When an arrest forecasted, the boosting classifier very often was unable produce a definitive distinction between the two possible outcome classes. Yet, unreliability forecasts are treated by confusion tables the same as reliable forecasts.

The second and third columns have identical prediction set proportions for up to two decimal places; the prediction sets from the White offenders' test data and from the Black offenders' transported test data are effectively identical. One has prediction parity in principle.

The fourth column shows that in practice there also is very close to perfect prediction parity when the approximate transported predictors are used for Black offenders. The proportion of prediction sets for which the forecast is an arrest for a violent crime remains at .03 for both Blacks and Whites. There is a slight reduction in the proportion of prediction sets that include no arrest by itself (i.e., $\{0\}$) and a slight increase in the proportion of prediction sets for which the classifier cannot make a clear choice (i.e., $\{1, 0\}$). Whether such differences matter would be for stakeholders to decide. A lot would depend on what a court magistrate would do with the $\{1, 0\}$ prediction sets. An option that might be acceptable to all would be to

withhold a decision until additional information was obtained that could improve forecasted outcome differentiation.

Finally, there is no assurance that the comparability shown in Table 3 will be achieved in other settings. The number of observations matters. So do the properties of the data. We have produced prediction parity but have not guaranteed it for new data. It remains to be seen how widely our results might generalize. The challenge comes not just from different mixes of offenders, but from different ways to define "similarly situated." Should juvenile arrests, for example, not be used?

7 Conclusions

Discussions of fairness for criminal justice risk algorithms can be puzzling. There is often a failure to appreciate that there are challenging tradeoffs between different kinds of fairness and between fairness, accuracy and transparency. You can't have it all. Yet many stakeholders argue that anything less is intolerable.

In addition, algorithmic fairness is commonly conflated with criminal justice fairness more generally. Confusion follows. A risk algorithm should not be blamed for criminal justice decisions and actions for which it is not responsible, and in any case, no algorithm can expected to fix decades of criminal justice inequities.

Another difficulty is that there is often confusion between disparate algorithmic performance and unfair algorithmic performance. If one protected group is accurately projected to be on the average a greater risk to public safety than another protected group, one should not automatically declare that the forecast is unfair. One must dig deeper to understand why the disparity is there to begin with. For example, if offenders past 50 years of age are usually forecasted to present little danger to the public, the risk algorithm simply may be exploiting the well documented fact that criminals typically "age out" of crime (Bekbolatkzy et al., 2019). Older offenders are not being given a pass just because of their seniority.

There commonly is also a failure to appreciate that a proper benchmark for risk algorithm performance is not perfection. The performance of risk algorithms should be compared to the performance of humans undertaking the same tasks. With all of the well-documented criticisms of the criminal justice system, the performance bar often can be quite low. A proper test is whether an algorithm can be more accurate, more fair, and more transparent.

In response to these complications, we have focused on the input and output of algorithmic risk assessments. Although any subsequent decisions or actions may be unfair, they are beyond a risk algorithm's reach and are best addressed by reforms tailored to those phenomena. Blaming a risk algorithm is at best a distraction and can divert remediation efforts away from fundamental change.

We have shown that prediction parity is easily achieved for a sensible fairness baseline by training on a more privileged group and then transporting the joint predictor probability distribution from a less privileged group to the joint predictor probability distribution of a more privileged group. On the average, a particular kind of Pareto improvement can follow. The more privileged group is not made worse off and the less privileged groups can be made better off. One might then argue that the risk algorithm dice are no longer loaded to favor one protected group over another. This strikes directly at concerns about mass incarceration and it many consequences.

An important distinction is made between internal and external fairness. With the exception of prediction parity, best examined with conformal prediction sets, achieving external fairness is difficult. A true label for a post-arraignment outcome is needed that accurately captures an arrest process in which members of a less privileged group are treated the same as similarly situated members of a more privileged group. In our application, for example, the post-arraignment experience of White offenders can be a reasonable counterfactual proxy for Black offenders only under extremely strong and untestable assumptions. Even arriving at effectively identical confusion tables for White and Black offenders, as we do in Appendix E, is not evidence of external fairness.

By intervening algorithmically on behalf of members of less privileged groups, we concurrently introduce a form of differential treatment. This has a long and contentious fairness history. But in most such circumstances, some groups arguably are made better off as other groups arguably are made worse off. Our approach to criminal justice risk assessment can sidestep such concerns. Still, Pareto improvement for groups must pass political and legal muster. One hurdle is whether under our approach real and consequential injuries can be avoided (Lujan v. Defenders of Wildlife, 1992); in U.S. federal court, "injury in fact" in mandatory. Another hurdle is whether there would be violations of "equal protection" under the fifth and fourteenth amendments to the U.S. Constitution (Coglianese and Lehr, 2017: 1191 - 1205), despite our intent make protection more equal.

Finally, there can be concerns about proposing risk procedures, even if rigorous, that explicitly respond to criminal justice realpolitik. But current

reform efforts are too often mired in misinformation and factional maneuvering, neither of which improve public discourse. In contrast, our foundational premises are plain. Past research is consulted. The limits of our methods are explicit. And, we have shown with real data that they can be successfully applied.

A Our Fair Risk Algorithm as Pseudocode

Algorithm 1 presents our fair risk algorithm as a pseudocode.

Algorithm 1: "Fair" Risk Algorithm

Input: Data $\mathcal{D} = \mathcal{D}_w \cup \mathcal{D}_b$ where $\mathcal{D}_w$ is the white population data and $\mathcal{D}_b$ is the black population data; Coverage probability $1 - \alpha$.

Output: A "fair" point prediction and prediction interval for the response.

1 Split the white population data $\mathcal{D}_w$ into two parts: white training data $\mathcal{D}_{w,t}$ and white calibration data $\mathcal{D}_{w,c}$.

2 Use $\mathcal{D}_{w,t}$, fit a classifier for the response given the covariates. Call this $\widehat{f}^{\texttt{white}}(\cdot)$ that takes an x from the white joint probability distribution and outputs a vector of probabilities $\widehat{f}^{\texttt{white}}(x)$.

3 Use $\mathcal{D}_{w,c}$ to obtain a conformal prediction set $\widehat{C}_\alpha^{\texttt{white}}(\cdot)$ that takes an x from the white joint probability distribution and outputs a set of outcomes that is guaranteed to contain the corresponding white outcome with a probability of at least $1 - \alpha$. See Appendix C and Algorithm 3 for more details on constructing conformal prediction sets.

4 Using the covariate observations in $\mathcal{D}_b$ and $\mathcal{D}_{w,t}$, obtain an estimate $\widehat{T}(\cdot)$ of the optimal transport map, that takes a covariate vector x from a black joint probability distribution and outputs $\widehat{T}(x)$ which resembles a white joint probability distribution. See Appendix B and Algorithm 2 for more details on estimating the optimal transport map.

5 **return** the point prediction output of our fair risk algorithm as follows. If x is the covariate vector of a white person, set $\widehat{y}^{\texttt{white}}$ as the highest probability outcome among $\widehat{f}^{\texttt{white}}(x)$. If x is the covariate vector of a black person, set $\widehat{y}^{\texttt{black}}$ as the highest probability outcome among $\widehat{f}^{\texttt{white}}(\widehat{T}(x))$.

6 **return** the prediction set output of our fair risk algorithm as follows. If x is a covariate vector of a white person, return $\widehat{C}_\alpha^{\texttt{white}}(x)$. If x is a covariate vector of a black person, return $\widehat{C}_\alpha^{\texttt{white}}(\widehat{T}(x))$.

B An Introduction to Optimal Transport

Suppose we have two distributions P and Q. In our example, think of P as the distribution of covariates for a population Black offenders and Q as that distribution for a population of White offenders. We seek to "transport" P to Q. Given a random vector X from the distribution P, we wish to create $Y = T(X)$ such that Y has the distribution Q. The function $T(\cdot)$ is called a transport map taking P to Q. There commonly exists several such maps, but (under regularity conditions) there is a unique map that minimizes the distance between X and $T(X)$ while ensuring $T(X)$ has the Q distribution. In other words, $T(X)$ moves X as little as possible to approximate a random vector from Q. Such a unique map, denoted by $T^*(\cdot)$, is called the optimal transport map.

Several kinds of distances can be used. Probably, the most common is the Euclidean distance and with this choice, the optimization problem known at the Monge formulaton, defines the optimal transport map $T^*(\cdot)$ given by

$$T^* := \underset{\substack{T:\, T(X)\sim Q, \\ \text{if } X\sim P}}{\arg\min} \; \mathbb{E}_P[\|X - T(X)\|_2^2].$$

This means that T^* is the minimizer of $\mathbb{E}_P[\|X - T(X)\|_2^2]$ over all functions T such that $T(X) \sim Q$ whenever $X \sim P$. This constraint on the functions T ensures that T^* transports P to Q and minimizes the expectation, which ensures that it is an optimal in that sense.

To illustrate, we use two very simple, univariate distributions: P and Q each supported on 5 points. Distribution P is supported at $6, 10, 15, 20, 25$, and distribution Q is supported at $10, 12, 15, 20, 30$. The probability values are given by

$$P(6) = P(10) = P(15) = P(20) = P(25) = 1/5,$$
$$Q(10) = Q(12) = Q(15) = Q(20) = Q(30) = 1/5.$$

Consider two transport maps T_1 and T_2 that convey values from $\{6, 10, 15, 20, 25\}$ into $\{10, 12, 15, 20, 30\}$.

$$T_1(6) = 10,\ T_1(10) = 12,\ T_1(15) = 15,\ T_1(20) = 20,\ T_1(25) = 30,$$
$$T_2(6) = 12,\ T_2(10) = 10,\ T_2(15) = 15,\ T_2(20) = 20,\ T_2(25) = 30.$$

In words, T_1 matches the smallest in the support of P to the smallest in the support of Q, the second smallest in the support of P to the second smallest in the support of Q, and so on. In this example, the transport

map is also the quantile-quantile map. On the other hand, T_2 does not change the same values (i.e., 10 to 10, 15 to 15, 20 to 20). It is easy to verify that if X has the distribution P, then $T_1(X)$ and $T_2(X)$ both have the distribution Q. This illustrates that a map transporting two general distributions P to Q is not unique. In this example, T_1 is the optimal transport plan minimizing $\mathbb{E}_P[|X - T(X)|^2]$, where the expectation is with respect to X from the distribution P, over all maps T such that $T(X)$ has the distribution Q.

For any transport map T, writing $Y = T(X)$, we obtain a joint distribution for the augmented vector $(X, Y) = (X, T(X))$, which is a "coupling" between the distributions P and Q. From this coupling perspective, the problem of optimal transport can be reformulated in terms of finding that coupling for a joint distribution whose marginals are fixed at P and Q. This is called the Kantorovich formulation. Estimation of the optimal transport plan is undertaken given data from P and Q, *not* the distributions P and Q themselves. We will not provide more details, and refer the reader to (Peyré and Cuturi, 2019; Deb and Sen, 2021; Deb et al., 2021; Hütter and Rigollet, 2021).

We offer pseudocode below (Algorithm 2) for estimating the optimal transport map T^* based on data, drawing on Sections 6.1.1 and 6.1.3 of Hütter and Rigollet (2021) with minor differences. The problem (4) is a linear programming task and is the most computing-intensive part of Algorithm 2. Beyond several thousand observations, solving (4) is computationally prohibitive. A simple work around is to split the data into several parts, apply Algorithm 2 on each part, and then average the estimates of optimal transport thus obtained.

Formally, suppose $\mathcal{D}_1^*$ and $\mathcal{D}_2^*$ are the initial (big) datasets available from P and Q respectively. Split $\mathcal{D}_1^*$ randomly into, say, 10 batches. Call them $\mathcal{D}_{1,1}, \mathcal{D}_{1,2}, \ldots, \mathcal{D}_{1,10}$. Similarly, split $\mathcal{D}_2^*$ randomly into, say, 10 batches. Call them $\mathcal{D}_{2,1}, \mathcal{D}_{2,2}, \ldots, \mathcal{D}_{2,10}$. Apply Algorithm 2 on $\mathcal{D}_{1,1}, \mathcal{D}_{2,1}$ to obtain an estimate $\widehat{T}^{(1)}(\cdot)$ of the optimal transport map. Similarly, apply Algorithm 2 on $\mathcal{D}_{1,2}, \mathcal{D}_{2,2}$ to obtain $\widehat{T}^{(2)}(\cdot)$, and so on to obtain $\widehat{T}^{(3)}(\cdot), \ldots, \widehat{T}^{(10)}(\cdot)$. Because the datasets $\mathcal{D}_{1,j}, \mathcal{D}_{2,j}$ are of sizes 10 times smaller than $\mathcal{D}_1*, \mathcal{D}_2^*$, problem (4) becomes more manageable computationally. Finally, set for all $x \in \mathbb{R}^d$,

$$\widehat{T}(x) := \frac{1}{10} \sum_{j=1}^{10} \widehat{T}^{(j)}(x),$$

as an estimate of the optimal transport map. For concreteness, here the data is split into 10 batches, but it can be made into a larger number of

Algorithm 2: Estimation of optimal transport map

Input: Data $\mathcal{D}_1 = \{X_1, \ldots, X_m\}$ from distribution P (in dimension d) and data $\mathcal{D}_2 = \{Y_1, \ldots, Y_n\}$ from distribution Q (in dimension d).

Output: A transport map $\widehat{T}(\cdot)$.

1 Find

$$\widehat{\Gamma} := \underset{\substack{\Gamma \in \mathbb{R}^{m \times n}, \Gamma_{ij} \geq 0 \\ \sum_{i=1}^{m} \Gamma_{ij} = 1/n, 1 \leq j \leq n, \\ \sum_{j=1}^{n} \Gamma_{ij} = 1/m, 1 \leq i \leq m}}{\arg\min} \sum_{i=1}^{m} \sum_{j=1}^{n} \|X_i - Y_j\|_2^2 \Gamma_{ij}. \tag{4}$$

This is a linear programming problem and can be solved using the R package `transport`.

2 For each $1 \leq i \leq m$, define

$$\widehat{Y}_i := \widehat{T}^{\text{emp}}(X_i) = \frac{\sum_{j=1}^{n} \widehat{\Gamma}_{ij} Y_j}{\sum_{j=1}^{n} \widehat{\Gamma}_{ij}} = m \sum_{j=1}^{n} \widehat{\Gamma}_{ij} Y_j,$$

as the transport of X_i (observations in $\mathcal{D}_1$).

3 For $1 \leq k \leq d$, perform non-parametric regression (using kernels, random forest, RKHS, etc) on the data $(X_i, \widehat{Y}_{i,k}), 1 \leq i \leq m$ with the k-th coordinate of $\widehat{Y}_i$ as the response. This yields a map $\widehat{T}_k(\cdot)$.

4 **return** $\widehat{T}(x) := (\widehat{T}_1(x), \ldots, \widehat{T}_d(x)) \in \mathbb{R}^d$ for any $x \in \mathbb{R}^d$ as the transport of x. This map $\widehat{T}(\cdot)$ serves as an estimate of the optimal transport map that transports P to Q.

batches as long as each batch contains "enough" observations.

C An Introduction to Conformal Inference

In the context of regression or classification, prediction can be a major goal, and there exist several point prediction methods that report an estimate of the true response. In practice, it is often important to also provide an uncertainty quantification along with the point prediction. Conformal inference provides such uncertainty quantification. In the context of our data, we only provide details about conformal inference for classification.

The setting of classical conformal inference is as follows. One has observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ independent and identically distributed from a distribution P. The goal is to provide a set $\widehat{C}_\alpha$ such that

$$\mathbb{P}((X_{n+1}, Y_{n+1}) \in \widehat{C}_\alpha) \geq 1 - \alpha, \tag{5}$$

when $(X_{n+1}, Y_{n+1}) \sim P$. Conformal inference provides a set $\widehat{C}_\alpha$ that satisfies (5) without any assumptions on the underlying joint probability distribution P. Furthermore, use can be made of one's favorite prediction algorithm, and the validity guarantee holds regardless of what the algorithm employed.

The basic idea of conformal inference is to assign a real valued score to each of the calibration data points, and a future point is placed in the prediction set if its score "conforms" with those of the training data points. There are several ways to construct such scores and all of them lead to a valid prediction set. In the following, we describe a simple score, and more complicated score, such as those in Kuchibhotla and Berk (2021), can also be used to obtain more precise prediction sets.

For a conformal inference procedure for classification, consider a setting in which the response/outcome Y_i takes one of two values 0, and 1. The pseudocode for the conformal inference with an absolute residual score is given in Algorithm 3. In the context of our fair risk algorithm (as shown in Figure 2), Algorithm 3 can be applied from Step 2, because the classifier $\widehat{p}(\cdot|\cdot)$ is given by $\widehat{f}^{\text{white}}(\cdot)$. In other words, $\mathcal{D}_{w,t}$ and $\mathcal{D}_{w,c}$ in Figure 2 play the roles of D_1 and D_2, respectively, in Algorithm 3. Further, $\widehat{f}^{\text{white}}(x)$ plays the role of $(\widehat{p}(0|x), \widehat{p}(1|x))$ in Algorithm 3.

Algorithm 3: Conformal prediction for classification

Input: Data splits D_1 (training data) and D_2 (calibration data), coverage probability $1-\alpha$.

Output: A prediction set $\widehat{C}_\alpha(\cdot)$ such that $\mathbb{P}(Y_\mathrm{f} \in \widehat{C}(X_\mathrm{f})) \geq 1-\alpha$, for a future observation $(X_\mathrm{f}, Y_\mathrm{f})$.

1 Train a classifier $\widehat{p}(\cdot|\cdot)$ on the training data D_1. This gives a probability distribution (estimator) for the outcomes for each x, i.e., we get for each x, probabilities $\widehat{p}(0|x)$ and $\widehat{p}(1|x)$ such that $\widehat{p}(0|x) + \widehat{p}(1|x) = 1$.

2 For each (X_i, Y_i) in the calibration data D_2, calculate the conformal scores $s(X_i, Y_i)$ as follows:

3
$$s(X_i, Y_i) := |Y_i - \widehat{p}(Y_i|X_i)|.$$

4 Compute the $(1 + 1/|D_2|)(1-\alpha)$-th quantile of $s(X_i, Y_i), i \in D_2$. Call this quantile $\widehat{\gamma}(\alpha)$.

5 **return** the prediction set

$$\widehat{C}_\alpha(x) \;:=\; \{y \in \{0,1\} :\, s(x,y) = |y - \widehat{p}(y|x)| \leq \widehat{\gamma}(\alpha)\}\,. \qquad (6)$$

D Conditions Under which Internal Fairness Implies External Fairness

Within our fairness formulation, for a risk algorithm used at arraignments to demonstrate external fairness, an appropriate estimate of the probability of a post-arraignment arrest must be available. Estimates can be obtained from the test data on hand, but they characterize criminal justice business as usual. Such estimates are appropriate for White offenders but not for the counterfactual of Black offenders who, post-release, are treated by police the same as similarly situated White offenders.[20] In this appendix, we provide sufficient conditions, under the counterfactual, allowing external fairness for classification parity, forecasting accuracy parity, and cost ratio parity to be properly inferred from internal fairness estimates. To this end, we introduce notation needed to address counterfactual outcomes.

[20] We focus on police because in practice, it is police who decide whether to make an arrest. Arrests by citizens are permitted but are extremely rare. Also, we use the term "offender" throughout to be consistent with our arraignment application.

Just in our application, there are two possible post-arraignment outcomes for offenders who are not detained: an arrest for a crime of violence or no such arrest. Let $R \in \{w, b\}$ denote the race of an offender, say White (w) or Black (b), and define $Y(r)$ as the counterfactual outcome a person would experience if, contrary to fact, the person were treated by police as a person of race $r = \{b, w\}$. Consistent with current causal inference thinking, each person, irrespective of his or her race, is associated with a pair of counterfactual outcomes $Y(w), Y(b)$, depending on how they would be treated by police were they of a given race, including a race which hypothetically differs from their actual race. That is, a person of race b is treated as a person for race w or vice versa.

Incorporating covariates, let $Y(r, x)$ likewise denote the counterfactual outcome had the person been treated by police as a person of race $r = b, w$ with covariate values x. In principle, covariates X may be multivariate and may include both continuous or discrete variables. For instance an offender's age usually is a key covariate to consider. Each offender in the data has a reported age: $X = \text{Age}$. Suppose for a 20 year old Black offender there are well-defined counterfactual outcomes $\{Y(w, x) : x = 21, 22, \ldots, 50\}$ corresponding to the person's outcomes if, contrary to fact, police treated this individual as White and of age 21, or 22, ..., or 50 years old. We will see shortly that the relationship between counterfactual variables $Y(r)$ and $Y(r, x)$ can be subtle.

Furthermore, for each Black offender, we let $Y(r, X(r^*))$ denote the person's counterfactual outcome had the offender been treated by the police as if he were of race r, with covariate values set to what they would have been had the offender been of race r^*. For example, suppose X includes age and number of prior arrests, and set $r = r^* = w$. Then, the corresponding counterfactual $Y(r, X(r^*)) = Y(w, X(w))$ for a Black offender 26 years of age with 2 prior burglary arrests such that $X = (26, 2)$ defines his or her outcome were the Black offender treated by police as if the offender were White with an age and number of prior burglaries corresponded to a similarly situated white person, such as $X(w) = (22, 1)$. For the Black offender, therefore, there is a counterfactual $Y(w, X(w)) = Y(w, X(w) = (22, 1))$. This follows from our intent to treat Black offenders as if they were White.

Throughout, we make the following consistency assumptions, which provide a necessary link between various defined counterfactuals. Mainly, we assume that $Y = Y(r) = Y(r, X(r))$ almost surely, if $R = r$. The observed outcome in the test data for an offender of race $R = r$, matches the hypothetical outcome the offender would have had were the police to treat the offender as a person of race $R = r$, which in turn matches the offender's

potential outcome if the police were to treat him or her as a person of race $R = r$ and covariates $X(r) = x$. Consistency may fail to hold, for instance, if a black offender's outcome depends on another offender's race or covariate values in addition to his own. This may arise in settings where an offender's arrest may not only depend on the offender's race but also on the race of an accomplice. Then, the potential outcome for the black offender may be ill-defined unless the race of the accomplice is introduced as a covariate.

Formally, suppose our proposed algorithm aspires to forecast the counterfactual $Y(w, X(w))$ for a Black offender (i.e., conditional on $R = b$) with observed covariates $X = X(b)$. Consider the following condition linking the optimal transport map T^* to counterfactual outcomes:

$$X(w) = T^*(X(b)) \quad w.p.1, \tag{7}$$

The assumption essentially states that for each Black offender, the joint distribution of $X(w)$ and $X(b)$ is degenerate. The assumption is far from trivial, as illustrated in the simple case where the optimal transport map is a location shift, i.e., $T^*(x) = x - \mu$ for a fixed constant μ. A violation of the assumption can arise in this setting if there were a covariate Z, say whether the offender had a family member who had been incarcerated, that although not observed in the database, interacts with race to modify the person's potential outcome as follows: $X(w) = T^*(X(b)) = X(b) - \mu_0 - \mu_1 \times Z$. Failing to account for Z would invalidate the equality assumption because the relationship between the two potential outcomes $X(w)$ and $X(b)$ cannot be made deterministic unless one also conditions on the unobserved factor Z.

Finally, consider the following strong ignorability condition, that for $r, r^* \in w, b$

$$Y(r, x) \perp\!\!\!\perp R, X(r^*), \tag{8}$$

which states that there are no common factors that determine both whether a person of race R and covariates $X(r^*)$ had he been of race r^* interacts with the criminal justice system, and how that system would treat offeneder if he or she were of race r with covariates x.

Under conditions (7) and (8), we prove that

$$Y(w, T^*(X(b)))\big|R = b, X \overset{d}{=} Y(w, X(w))\big|R = w, X.$$

This follows by noting that for any x and $x^* = T^*(x)$:

$$\begin{aligned}
&\mathbb{P}(Y(w, x^*) = y, X(w) = x^* | R = b, X = x) \\
&= \mathbb{P}(Y(w, x^*) = y | R = b, X(b) = x) Pr(X(w) = x^* | R = b, X(b) = x) \\
&= \mathbb{P}(Y(w, x^*) = y | R = b, T^* X(b) = x^*)) \times 1 \\
&= \mathbb{P}(Y(w, x^*) = y | R = b, X(w) = x^*)) \\
&= \mathbb{P}(Y = y | X = x^*, R = w)
\end{aligned}$$

establishing the result that the desired distribution $(Y^* = Y(w, x^*), X(w))|R = b$ can be obtained by first sampling $T^*(X)$ from the covariate distribution of black defenders, and subsequently sampling Y from the conditional distribution of white defenders with covariate equal to $T^*(X)$, matching the output of our proposed algorithm. Should this hold, external fairness can be said to be achieved if internal fairness can be established for the proposed algorithm. More precisely, to the extent that classification, forecasting accuracy and cost ratio parities can be demonstrated, their counterfactual analogues would in principle be implied by the stated assumptions. One would then have a firm basis for claiming external fairness.

There are several ways in which these assumptions arguably are unrealistic, at least in most jurisdictions in the United States. First, as explained above, the assumption that the relationship between counterfactual is deterministic rules out the existence of latent effect heterogeneity in the association between an offender's race and the manner in which the offender is ultimately treated by the criminal justice system. This assumption is sometimes called a rank preservation condition, which implies that for any two persons i and j, if $X_i(b) \leq X_j(b)$ for a scalar variable X, then it must be that $X_i(w) \leq X_j(w)$, thus ruling out the existence of an unmeasured factor related to race in a manner that can alter the ranking of potential covariate values.

For the location shift example previously described with say $\mu_0 = 0$ and $\mu_1 = 5$, then $X(w) = T^*(X(b)) = X(b) - 5 \times Z$. It is possible that $X_i(b) = 4 \leq X_j(b) = 6$. However $Z_i = 0$ while $Z_j = 1$ so that $X_j(w) = 1 \leq X_i(w) = 4$. The rank preservation is violated. Even in this simple example, rank preservation assumption, which is not empirically testable, may be difficult to justify because of a large number of omitted variables, particularly in practical settings where X is multivariate.

The independence condition (8) is likewise unrealistic because it rules out any common factor associated with the race of an offender and a potential apprehension outcome. It also rules out any unmeasured common cause of an offender's covariates and an apprehension outcome, conditional on race.

Given the role of race in myriad social institutions and interactions, claims that such relationships are absent would in practice strain credibility.

E Confusion Table Results for Black Test Data When the Response Variable as well as The Predictors are Transported

Stakeholders and others often focus primarily on fairness represented in confusion tables from test data. In deference to those individuals, we applied optimal transport to the Black test data in a manner that included in the joint probability distribution the response variable as well as the predictors. Table 3 is virtually the same as Table 1 within sampling error. The comparability is striking. For example, we compared the base rates from the test data for White offenders and the transported test data for Black offenders. For both, the base rate was approximately .075. 7.5% of both Black and White offenders were rearrested after an arraignment for a violent crime. We emphasize the base rate equivalence because the base rate is so central in formal fairness discussions (Kleinberg et al., 2017).

Table 3: Transported Test Data Confusion Table for Black Offenders Using White-Trained Algorithm (30% Predicted to Fail, 7.5% Actually Fail)

Actual Outcome	No Violence Predicted	Violence Predicted	Classification Error
No Violence	2658	1042 (false positive)	.28
Violence	135 (false negative)	166	.45
Forecasting Error	.05	.86	

If one is prepared, as in common practice, to evaluate fairness solely using the test data in a confusion table, optimal transport provides an effective equalizer. One may have politically acceptable risk algorithm. One is also externally fair for predictive parity because no outcome label is required. And, if one is comfortable assuming that in reality, Black offenders will be treated on the average the same as similarly situated White offenders after an arraignment release, external fairness is achieved more generally. But for most stakeholders, this last step will stretch credibility.

In short, although interpretations for fairness will vary, one has made all of the confusion table results for White offenders and Black offenders the same. All of the tradeoff concerns are bypassed as long as one is prepared to assume that if a confusion table is good enough for White offenders,

it is good enough for Black offenders. Note that such claims go only to the aggregate confusion table results by which one might evaluate a risk algorithm overall. It says nothing about the operational issues required for forecasting.

References

Alpert, G.P., Dunham, R.G., and M.R. Smith (2007) "Investigating Racial Profiling by the Maimi-Dade Police Department: A Multimethod Approach." *Criminology and Public Policy* 6(1) 24 – 55.

Baer, B.R., Gilbert, D.E., and M.T.Wells (2020) " Fairness Criteria through the Lens of Directed Acyclic Graphs: A Statistical Modeling Perspective." In Dubber, M.D., Pasquale, F., and S.Das, *The Oxford Handbook of Ethics of AI.* Oxford Press.

Barocas, S., Hardt, M., and A. Narayanan (2018) *Fairness and Machine Learning.* http://www.fairmlbook.org

Bekbolatkyzy, D.S., Yerenatovna, Yergali, D.R., Maratuly, Y.A., Makhatovna, A.G., and K.M. Beaver. (2019) "Aging Out of Adolescent Delinquency: Results from a Longitudinal Sample of Youth and Young Adults." *Journal of Criminal Justice* 60 (January - February): 108 – 116.

Berk, R.A. (2009) "The Role of Race in Forecasts of Violent Crime," *Race and Social Problems*, 1(4): 231–242.

Berk, R.A. (2017) "An Impact Assessment of Machine Learning Risk Forecasts on Parole Board Decisions and Recidivism." *Journal of Experimental Criminology* 13: 193–216.

Berk, R.A. (2018) *Machine Learning Forecasts of Risk in Criminal Justice Settings.* New York: Springer.

Berk, R.A., Heirdari, H., Jabbari, S., Kearns, M., & A. Roth (2018) "Fairness in Criminal Justice Risk Assessments: The State of the Art." *Sociological Methods and Research*, first published July 2nd, 2018, http://journals.sagepub.com/doi/10.1177/0049124118782533.

Berk, R.A., and A. A. Elzarka (2020) 11 Almost Politically Acceptable Criminal Justice Risk Assessment." *Criminology and Public Policy* 2020: 1 – 28.

Bhopal, K. (2018) *White Privilege: The Myth of a Post-Racial Society.* Policy Press.

Chouldechova, A. (2017) "Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments." *Big Data* 5(2):153 – 163.

Coglianese, C., D. Lehr (2017) "Regulating by Robot: Administrative Decision Making in the Machine-Learning Era." *Georgetown Law Journal* 105: 1147–

Coglianese, C., D. Lehr (2019) "Transparency and Algorithmic Governance." *Faculty Scholarship at Penn Law* 2123. Also *Administrative Law Review* 2019 1 (2019).

Corbett-Davies S., and S. Goel (2018) "The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning." 35th International Conference on Machine Learning (ICML 2018).

D'Amour, A., Heller, K., Adlam, B., et al., (2020) "Underspecification Presents Challenges for Credibility in Modern Machine Learning." arXiv:2011.03395v2 [cs.LG].

De Lara, L., González-Sanz, A., Asher, N., and J.-M. Loubes (2021) "Transport-Based Counterfactual Models." arXiv:2108.13025v1 [cs.AI]

Diana, E., Gill, W., Kearns, M., Kenthapadi, K., and A. Roth (2021) Minimax Group Fairness: Algorithms and Experiments) AIES' 21: Proceedings of the 2021 AAI/ACM: 66–76.

Dwork, C., Hardt, M., Patassi, T., Reingold, O., and R. Zemel (2012) "Fairness through Awareness." ITCS 2012: Proceedings of the 3rd Innovations in Theoretical Computer Science Conference: 214 – 226.

Edwards, F., Lee, H., and M. Esposito (2019) "Risk of Being Killed by Police Use of Force in the United States by Age, Race-Ethnicity, and Sex." Proceedings of the National Academy of Sciences: 116: 16793–16798.

Feldman, M., Friedler, S., Moeller, J., Scheidegger, C., and S. Venkatasubrtamanian (2015) "Certifying and Removing Disparate Impact." In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 259 – 268.

Fisher, F.M., and J.B. Kadane (1983) "Empirically Based Sentencing Guidelines and Ethical Considerations." In Alfred Blimstein et. al., ed.

Research on Sentencing: The Search for Reform, Volume II, National Criminal Justice Reference Service, Office of Justice Programs, U.S. Department of Justice, NCJ-91771.

Friedman, J.H. (2001) "Greedy Function Approximation: A Gradient Boosting Machine." *Annals of Statistics* 29 (5): 1189 – 1232.

Gastwirth, J.L. (2000) *Statistical Science in the Courtroom* Springer.

GBD 2019 Police Violence Subnational Collaborators (2021) "Fatal Police Violence by Race and State in the USA, 1980-2019: A Network Meta-Regression." *Lancet* 398: 1239-1255.

Gelman, A., Fagan, J., and A. Kiss (2012) "An Analysis of the New York City Police Department's 'Stop-and-Frisk' Policy in the Context of Claims of Racial Bias." *Journal of the American Statistical Association* 102 (2007): 813 – 823

Grogger, J., and G. Ridgeway (2012) "Testing for Racial Profiling in Traffic Step From Behind a Veil of Darkness." *Journal of the American Statistical Association* 202 (2006): 878 – 887.

Harcourt, B.W. (2007) *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* Chicago, University of Chicago Press.

Hardt, M., Price, E., N. Srebro (2016) "Equality of Opportunity in Supervised Learning." In D.D. Lee, Sugiyama, U.V. Luxburg, I. Guyon, and R. Garnett (eds.) *Equality of Opportunity in Supervised Learning.* Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain, (pp.3315 – 3323).

Holewinski, I.A. (2002) "Inherently Arbitrary and Capricious: An Empirical Analysis of Variations Among Death Penalty Statutes." *Cornell Journal of Law and Public Policy* 12: 231 – 259.

Horder, J. (1993) "Criminal Culpability: The Possibility of a General Theory." *Law and Philosophy* 12: 193– 215.

Hudson, B. (1989) "Discrimination and Disparity: The Infuence of Race on Sentencing." *Journal of Ethnic and Migration Studies* 16(1): 23 – 34.

Hütter, J., and P. Rigollet (2020) "Minimax Estimation of Smooth Optimal Transport Maps." arXiv:1905.05838v3 [math.st].

Huq, A.Z. (2019) "Racial Equality in Algorithmic Criminal Justice." *Duke Law Journal* 68 (6), 1043–1134.

Imai, K., and Z. Jaing (2021) "Principal Fairness for Human and Algorithmic Decision-Making." arXiv:2005.10400v4 cs.CY

Jackson, E.K. (2019) *Scandinavians in Chicago: The Origins of White Privilege in Modern America* University of Illinois Press.

Johndrow, J.E., and K. Lum (2019) "An Algorithm for Removing Sensitive Information: Application to Race-Independent Recidivism Prediction." *Annals of Applied Statistics* 13(1): 189 – 220.

Kamiran, F., and T. Calders (2012) "Data Preprocessing Techniques for Classification Without Discrimination." *Knowledge Information Systems* 33:1 - 33.

Kearns, M and A. Roth (2020) *The Ethical Algorithm* Oxford Press.

Kearns, M., Neel, S., Roth, A., and S. Wu (2018) "Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness." Preprint https://arxiv.org/abs/1711.05144.

Kleinberg, J., Mullainathan, S., and M. Raghavan, M. (2017) "Inherent Tradeoffs in the Fair Determination of Risk Scores." Proc. 8th Conference on Innovations in Theoretical Computer Science (ITCS).

Kroll, J.A., Huey, J., Barocas, S., Felten, E.W., Reidenberg, J.R., Robinson, D.G., and H. Yu (2017) "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633 – 705.

Kuchibhotla, A.K. and R.A. Berk (2021) "Nested Conformal Prediction Sets for Classification with Applications to Probation Data." arXiv:2104.09358

Kushner, H.J., and G.G. Yin (2003) *Stochastic Approximation and Recursive Algorithms and Appications.* Springer.

Kusner, M., Loftus, J., Russell, C. and R. Silva (2018) "Counterfactual Fairness." arXiv: 1703.06856v3. [stat.ML]

Lee, N.T., Resnick, P., and G. Barton (2019) "Algorithmic Bias Detection and Mitigation: Best Practices and Polices to Reduce Consumer Harms." Brookings institution, Washongton D.C., Bookings Report.

Leonard, D.J., (2017) *Playing While White: Privilege and Power On and Off The Field*. University of Washington Press.

Loiffler, C.E., and A. Chalfin (2017) "Estimate the Crime Effects of Raising the Age of Majority." *Criminology & Public Policy* 16(1): 45 – 71.

Luenberger, D.FG. and Y. Ye (2008) *Transportation and Network Flow Problems*. Springer.

Lujan v. Defenders of Wildlife, 504, U.S. 555 (1992).

Lynch, M. (2011) "Mass Incarceration, Legal Change, and Locale: Understanding and Remediating American Penal Overindulgence." *Criminology & Public Policy* 10(3): 673 – 698.

Madras, D., Pitassi, T., and R. Zemel (2018a) "Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer." 32nd Conference on Neural Information Processing Systems (NIPS 2018), Montréal, Canada.

Madras, D., Creager, E., Pitassi, T., and R. Zemel (2018b) "Fairness through Causal Awareness: Learning Causal Latent-Variable Models for Biased Data." arXiv: 1809.02519v3 [cs.LG]

Manole, T., Balakrishnan, S., Niles-Weed, J., and L. Wasserman (2021) "Plugin Estimation of Smooth Optimal Transport Maps." asXiv:2107.12364v1 [math.ST]

Mishler A., and E. Kennedy (2021) "FADE: Fair and Double Ensemble Learning for Observable and Counterfactual Outcomes." arXiv:2109.00173v1. [stat.ML]

Mitchel, S., Potash, E., Barocas, S., D'Amour, A, and K. Lum (2021) "Algorithmic Fairness: Choices, Assumptions, and Definitions." *Annual Review of Statistics and Its Applications* 2021: 8: 141 –163.

Morgan, S.L. and C. Winship(2015) *Counterfactuals and Causal Inference*, second edition. Cambridge University Press.

Mullainathan, S. 2018. "Biased Algorithms Are Easier to fix Than Biased People." *New York Times* December 6, 2019. https://www.nytimes.com/2019/12/06/business/algorithm-bias-fix.html. Muller, C. (2021) "Exclusion and Exploitatin: The Incarceration of Black Americans fro Slavery to the Present." *Science* 374(6565): 282 – 286.

Nabi, R., Malinsky, D., and I. Shpitser (2019) "Learning Optimal Fair Policies." arXiv:1809.02244v3 [cs.LG]

Nath, S.V. (2006) "Crime Pattern Detection Using Data Mining." *2006 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology Workshops*, 2006, pp. 41– 44.

Ostrom, C.W., Ostrom, B.J., and M. Kleinman (2003) *Judge and Discrimination: Assessing the Theory and Practice of Criminal Sentencing.* NCJRS, U.S. Department of Justice, Washington, D.C.

Peyré, G., and M. Cuturi, M. (2019) *Computational Optimal Transport With Applications to Data Science.* NOW Publishers.

Pooladian, A.-A., and J. Niles-Weed (2021) "Entropic Estimation of Transport Maps." arXiv:2109.12004v1 [math.ST]

Robert Wood Johnson Foundation (2017) "Discrimination in America: Experiences of Views of African Americans. "https://media.npr.org/assets/img/2017/10/23/discriminationpoll- african-americans.pdf

Rocque, M. (2011) "Racial Disparities in the Criminal Justice System and Perceptions of Legitimacy: A Theoretical Linkage." *Race and Justice* 1(3): 292 – 315.

Romano, Y., Barber, R.F., Sabatti, C., and E.J. Candes (2019) "With Malice Toward None: Assessing Uncertainty via Equalized Coverage." axXIiv: 1908.05428v1 [stat, ME]

Rothenberg, P.S. (2008) *White Privilege* Catherine Woods.

Rucker, J.M., and J.A. Rocheson (2021) "Toward an Understanding of Structural Racism: Implications for Criminal Justice," *Science* 374 (6565): 286 – 290.

Skeem, J., and C. Lowenkamp (2020) "Using Algorithms to Address Trade-Offs Inherent in Predicting Recidivism." *Behavioral Science and Law* 38: 259–278.

Shafer, G., and V. Vovk (2008) "A Tutorial on Conformal Prediction." *Journal of Machine Learning Research* 9: 371 – 421.

Si, N., Murthy, K., Blanchet, J., and V.A. Nguyen (2021) "Testing Group Fairness via Optimal Transport Projections." Proceedings of the 38th International Conference on Machine Learning, PMLR 139.

Sorenson, S.B., Sinko, L., and R.A. Berk (2021) "The Endemic Amid the Pandemic: Seeking Help for Violence against Women in the Initial Phases of COVID-19." *Journal of Interpersonal Violence* published online March, 2021.

Starr, S.B. (2014) "Evidence-Based Sentencing and the Scientific Rationalization of Discrimination. *Stanford Law Review* 66: 803 – 872.

Stewart, E.A., Warren, P.Y., Hughes, C., and Brunson, R.K. (2020) "Race, Ethnicity, and Criminal Justice Contact: Reflections for Future Research," *Race and Justice* 10 (2): 119 –149.

Tibshirani, R.J., Barber, R.F., Candès, E.J. and A. Ramdas (2020) "Conformation Prediction Under Covariate Shift." arXiv: 1904.06019v3 [stat.ME].

Thompson, W.C., and E.L. Schumann (1987) "Interpretation of Statistical Evidence in Criminal Trials: The Prosecutor's Fallacy and the Defense Attorney's Fallacy." *Law and Human Behavior* 11: 167 – 187.

Tonry, M. (2014) "Legal and Ethical Issues in The Prediction of Recidivism." *Federal Sentencing Reporter* 26(3): 167 – 176.

Van Cleve, N.G. and L. Mayes(2015) "Criminal Justice Through "Colorblind" Lenses: A Call to Examine the Mutual Constitution of Race and Criminal Justice." *Law & Social Inquiry* 40(2): 406 – 432.

Vovk, V., Gammerman, A., and G. Shafer (2005), *Algorithmic Learning in a Random World*, NewYork: Springer

Vovk,V., Nouretdinov, I., and A. Gammerman (2009), "On-Line Predictive Linear Regression." *The Annals of Statistics* 37: 1566 – 1590.

Wacquant, L. (2002) "From Slavery to Mass Incarceration: Rethinking the 'race question' in the US." *New Left Review* 13:41 – 73.

Wallis, J. (2017) *America's Original Sin* Brazos Press.

Wyner, A.J., Olson, M., Bleich, J, and D. Mease (2015) "Explaining the Success of AdaBoost and Random Forests as Interpolating Classifiers." *Journal of Machine Learning Research* 18(1): 1–33.

Yates, J. (1997) "Racial Incarceration Disparity Among States." *Social Science Quarterly* 78(4) 1001 – 1010.

Zafar, M.B., Martinez, I.V., Rodriguez, M.,B., and K. Gummadi. (2017) "Fairness Constraints: A Mechanism for Fair Classification." In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS). Fort Lauderdale, FL, 2017.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and C. Dwork (2013) "Learning Fair Representations." *Proceedings of Machine Learning Research* 28 (3) 325 – 333.