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

Richard Berk
University of Pennsylvania

Arun Kumar Kuchibhotla
Carnegie Mellon University

January 6, 2021

Abstract

Research Summary

Risk assessment algorithms have been correctly criticized for potential unfairness, and there is an active cottage industry trying to make repairs. In this paper, we combine statistical/machine learning, adjustments for a covariate shift, and conformal prediction sets to remove unfairness from risk algorithms themselves and the covariates used for forecasting. From a sample of 300,000 offenders at their arraignments, we construct a confusion table and its derived aggregate measures of fairness that are free of any meaningful differences between Black and White offenders. We also produce fair forecasts for *individual* offenders coupled with valid probability guarantees that the forecasted outcome is the true outcome. We believe this is a first.

Policy Implications

We see our work as a demonstration of concept for applications in a wide variety of criminal justice decisions. The procedures provided can be routinely implemented in jurisdictions with the usual criminal justice datasets used by administrators. The requisite procedures can be found in the scripting software R. However, whether stakeholders will accept our approach as a means to achieve risk assessment fairness

*The Authors have no conflict of interest. Thanks for Cary Coglianese, Eric Tchetgen Tchetgen, and two reviewers provided very helpful feedback on an earlier draft of this paper.

is unknown. There also are legal issues that would need to be resolved although we offer a Pareto improvement.

Keywords
Risk Assessment; Fairness ; Risk Algorithms ; Statistical Learning, Machine Learning, Covariate Shift, Conformal Prediction Sets

1 Introduction

The goal of fair algorithms remains a top priority among algorithm developers and the users of those algorithms (Berk, 2018; Huq, 2019; Kearns and Roth, 2020). 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 (Berk et al., 2018; 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).

In this paper, we focus on risk assessments used in criminal justice and propose novel fix for algorithmic unfairness. Because of its simplicity and apparent effectiveness, there is substantial promise for real criminal justice applications. Unlike most other work, the methods we discuss also take seriously a political climate in which appearances can be more important than facts. A recent paper by Berk and Elzarka (2020) provides a good start, but their approach lacks the formal framework that we offer, which, in turn, solves problems that the earlier work cannot. Building on a foundation of statistical/machine learning and conformal prediction sets (Vovk et al., 2005; 2009; Lei et al., 2018),[1] we suggest a statistical justification for risk algorithms that treats a less privileged group (e.g., Black offenders) as if they were a more privileged group (e.g., White offenders) and then adjusts the covariates used in risk forecasting so that there is far better balance between the groups. Instructive statistical inference can follow in either of two forms: from conventional confusion tables or from conformal prediction sets. However, their estimands differ as well as the policy questions they address. We illustrate with a sample of 300,000 offenders at arraignment.

[1]The content of statistical learning is much the same as the content of machine learning, although they have somewhat different intellectual histories. Either the term machine learning or statistical learning can be used. We will statistical learning going forward.

A didactic discussion of the statistical details is provided in Appendix A. Illustrative code in R is included in Appendix B.

2 Conceptual Framework

Despite a bit of conceptual sloppiness in the technical literature, algorithms are not models. Models are *explanations*, typically buttressed by subject-matter theory and research designs, of how the data were generated; one has a theory of how the data came to be. In conventional matrix notation, a popular scaffolding for such theory is the ubiquitous linear regression model

$$Y = \mathbf{X}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2), \tag{1}$$

where Y is the response variable, $\mathbf{X}$ is a matrix of one or more fixed regressors, $\boldsymbol{\beta}$ is the corresponding regression coefficients, and ε represents independent random perturbations, drawn from a single distribution with a mean of 0.0 and a variance σ^2. The random perturbations are the only source of uncertainty for a model conventionally assumed to be correctly specified. One has a mathematical theory of some phenomenon, although in practice the model can be right, wrong, or something in between (Buja, 2019a,b).[2]

Algorithms are a set of instructions for how to compute quantities of interest. There is no need for subject-matter theory, and no claims are made that one has characterized how a particular natural or social process generated the data. For example, the act of balancing a bank statement depends on an algorithm not a model.

Algorithms can perform poorly if they produce fitted values or forecasts that are insufficiently accurate, fair, or transparent. But there is no such thing as misspecification. Either an algorithm works satisfactorily, does not work satisfactorily or something in between (Berk, 2020a, Chapter 1). Claiming that an algorithm is right or wrong is misguided. Right or wrong compared to what? This will have important implications later in the paper.

Algorithmic uncertainty comes from the data itself. The usual requirement is that each case in the data is realized independently and at random from a joint probability distribution characterizing all of the predictor variables and the response variable. In common shorthand, the data are realized

[2]We are following the common practice of using a bold font for two-dimensional arrays. We do not use a bold font for vectors, which are one-dimensional arrays.

IID.[3] Consequently, all of the variables in a realized dataset are treated as random variables. The IID assumption cannot be satisfied by hand waving, no matter how vigorous. A case must be made from knowledge about how the data were actually generated. We provide an example later.

In this paper, algorithmic fairness is the focus. There are many kinds of fairness whose definitions and properties have been thoroughly discussed in the recent literature. For comparisons between legally protected groups, we will concentrate on five types commonly invoked, at least some of which appear in virtually every formal consideration of fair risk assessment for criminal justice decisions. There is also a very interesting literature on fairness for individuals in which *similarly situated individuals* (e.g., statistical nearest neighbors) should be treated alike (Dwork et al., 2012; Zemel et al., 2013). But so far at least, criminal justice concerns have centered on groups. It is not clear what "similarly situated" might mean for groups, although we consider that issue later.

Because actual criminal justice decisions are categorical (e.g., grant parole or not), we limit the discussion to categorical outcomes. For simplicity, we assume that the outcome to be forecasted is binary (e.g., arrested or not while on probation). There will be no important loss of generality. There are published examples of criminal justice risk assessments using more than two outcome classes (Berk and Sorenson, 2016).

A bit more formally, the binary, random response variable Y has two outcome classes, often coded as 1 or 0. The probability of an outcome class is a function of a set of covariates $\mathbf{X}$, also random variables, that may be numeric or categorical (e.g., the number of prior arrests, gender), commonly written as $\mathbb{P}(Y|\mathbf{X})$. Some algorithmic classifier, such as neural networks, is use to obtain fitted values $\hat{Y}|\mathbf{X}$. Each fitted outcome class $\hat{Y}$ can be used to characterize risk. Forecasts may be obtained for new cases that have the same set of predictor variables $\mathbf{X}$ but unknown values of Y. One uses the computed structure of $\hat{Y}|\mathbf{X}$ with the new predictor values to obtain values for $\hat{Y}$.

2.1 Defining Fairness

There is no common language for different kind of fairness, but the definitions that follow can be easily translated into most of the common typologies.

- *Prediction parity* – Is the predictive distribution for each group the

[3] IID stands for independent and identically distributed. Sometimes it is denoted by iid.

same? For example, is the proportion of Black offenders and White offenders predicted to succeed on parole the same?

- *Classification parity* – Are the false positive rates and false negative rates the same for each group? False positives and a false negatives take each binary outcome class as known and determine the proportion of times the risk algorithm incorrectly identifies it. By convention, confusion tables usually denote rows by actual outcomes and columns by forecasted outcomes. To compute the false positive and false negative rates, one conditions within rows.

- *Forecasting accuracy parity* – Is each outcome class forecasted with equal accuracy for every group? A forecast is incorrect if the forecasted outcome does not correspond to the actual outcome. Now, one conditions within columns of the confusion table to obtain for each column the proportion of cases that are incorrectly forecasted. Forecasting accuracy parity should not be confused with classification parity.

- *Cost Ratio parity* – Are the relative costs of false positives and false negatives the same for each group? The cost ratio determines the way in which a risk assessment procedure 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 of every group.[4]

There is no single concept or measure commonly used to define fairness. In practice, stakeholders examine a set of fairness measures, such as those just defined, and argue that each should be effectively equivalent across all legally protected groups. We will proceed in the same manner.[5] With IID or

[4] These costs are rarely monetized. What matters for the risk algorithm is the 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. In practice, 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 affect the forecasted risk, often dramatically.

[5] "Effectively equivalent" can be in the eye of the beholder. To require that perfect equivalence precludes algorithmic risk assessment because even if the true measures are literally identical across protected groups, the estimated measures will differ at least a bit because of random measurement error and sampling error. For example, a recent presidential executive order implicitly precludes any use of machine learning by federal

exchangeable data, discussed in Appendix A, the fairness proportions can be treated as probabilities, and statistical inference can become consequential.

2.2 Developing a Fair Risk Algorithm

Fair risk assessments depend on the performance of a risk algorithm and the data used to train it. Some argue that statistical learning algorithms should be preferred (Berk, 2018) and that all available predictors should be used except those discarded because of fairness concerns (e.g., arrests as a juvenile) or technical problems (e.g., many missing observations). We will proceed in this spirit, but the principles employed can pertain far more broadly. For concreteness, we will use Black offenders and White offenders at their arraignment hearings as illustrations throughout the paper, but the issues apply equally well to other protected groups in a variety of criminal justice settings.[6]

Some of the approaches we employ may be unfamiliar. They build on a very recent statistical literature summarized in the body of the paper and supplemented by a didactic appendix. We begin by introducing two potential corrections for possible unfairness in algorithmic risk assessments.

2.2.1 Training the Risk Algorithm on White Offenders

Suppose for now one has an IID dataset of offenders at their arraignments. The essential feature of the first correction is a risk algorithm, such as gradient boosting, trained only on Whites but through test data providing risk estimates separately for White (W) and Black (B) offenders. For a response variable Y and predictors $\mathbf{X}$, one employs White training data and a risk algorithm so that $\hat{Y}_{\rm W}^{\rm Train} = \hat{f}(\mathbf{X}_{\rm W}^{\rm Train})$. Inserting test data into the fitted $\hat{f}$, separate fitted values for Whites and Blacks respectively are $\hat{Y}_{\rm W}^{\rm Test} = \hat{f}(\mathbf{X}_{\rm W}^{\rm Test})$ and $\hat{Y}_{\rm B}^{\rm Test} = \hat{f}(\mathbf{X}_{\rm B}^{\rm Test})$. The function is *not* re-estimated with the test data; it is fixed after it is estimated with the White training data.

agencies except those association with the Department of Defense and national intelligence (https://www.whitehouse.gov/presidential-actions/executive-order-promoting-use-trustworthy-artificial-intelligence-federal-government/).

[6]Some recent concerns have been raised about the stability of machine learning results when there is a very large number of predictors that can be strongly related to one another (D'amour et al., 2020). The problems can be similar to those in linear model caused by multicolinearity and/or extrapolations into predictor regions in which there are no data. However, the statistical learning tools and data we use are not subject to such difficulties.

Just as in Berk and Elzarka (2020), the algorithm itself, trained on the data for White offenders only, cannot be responsible for any race-based unfairness because data from Black offenders play no role in the fitting enterprise. Then, all offenders are processed as if they are White. Blacks can be made better off, and no Whites can be made worse off. If Black offenders benefit, one has *a Pareto improvement.*[7]

2.2.2 Adjusting for a Covariate Shift

With the algorithm itself absolved from blame, one can exploit the test data to implement a second potential correction. Despite training the risk algorithm only on Whites, some forms of unfairness may remain because Black and White offenders can have different predictor distributions. For example, Blacks might have longer prior arrest records, which can make Black offenders appear to be more crime prone going forward. Many argue that such disparities result from police practices that can differ between Black and White citizens. Perhaps the most widely cited example is "stop-and-frisk" that has been criticized as racially motivated (Gelman et al., 2012). Stop-and frisk can be seen as a special case of racial profiling (Grogger and Ridgeway, 2012) that may include police actions after a stop is made, not just the stop itself (Alpert et al., 2007). Under these and related scenarios, Black citizens can have a larger number of prior arrests than White citizens that, in turn, can lead to algorithmic forecasts of higher risk.

In addition, there are concerns that Black individuals are at greater risk of arrest because of a greater density of police activities in their neighborhoods, even if that greater density results from legitimate law enforcement concerns. For example, as a matter of policy, more police may be assigned to neighborhoods with higher crime rates, or in practice, be dispatched disproportionately to neighborhoods with a greater concentration of 911 calls (Berk, 2020b). The claim is that disparate treatment by police, whatever

[7]We assume that consistent with common understandings and frequent stakeholder claims, White offenders get preferential treatment compared to Black offenders. If the algorithm were trained only the Black offenders, the algorithm would still not be responsible for race-based unfairness. No invidious racial distinctions could be made. But ultimately, White offenders likely would be made worse off, and there would be no systematic gains for Black offenders. The result would no longer be a Pareto improvement and would almost certainly be rejected by stakeholders. Training on Black offenders and White offenders separately would also compromise Pareto improvement and would also introduce the legally suspect practice of treating Black offenders and White offenders differently. It would also confound bias in an algorithm with bias in the data, which is precisely the problem we are trying to avoid.

the cause, is carried forward by the data used for training risk algorithms. For example, underage Black citizens may be more likely to be charged as adults. In short, unfair risk assessments can be the result even if racial differences in the data provided to a risk algorithm are not caused by racial animus.

Should the joint predictor distribution for Black offenders differ from the joint predictor distribution for White offenders, one has an instance of a "covariate shift" (Tibshirani et al., 2020). As a potential remedy when risks are forecasted, one can adjust the joint predictor distribution for Blacks to be more like the joint predictor distribution for Whites. Insofar as the two joint distributions coincide, remaining unfairness caused by "biased data" can be eliminated. One might say the groups are now "similarly situated," although we have not seen this formulation for groups before.[8]

The adjustment we implement is analogous to the methods that weight predictor distributions by propensity scores to improve causal inference in observational studies (Imbens and Rubin, 2015: section 12.4.2).[9] The weighting also is roughly similar to population-weighted adjustments common in survey research to achieve better representation (Lavrakas, 2008). However, in survey applications, the population weights typically are known. The weights needed to adjust for a covariate shift are usually unknown and are estimated. We use gradient boosting. More details are provided as part of the data analysis in section 4.2.

Propensity score adjustments can be very effective. But they are not assumption free. As discussed more formally shortly, our approach includes computing valid uncertainty estimates for individual risk forecasts: the probability that forecasted outcomes for given offenders are the true outcomes. This requires that the conditional distribution of the response $\mathbb{P}(Y|\mathbf{X})$ be the same for both protected groups, although the joint predictor distributions $\mathbb{P}(\mathbf{X})$ can differ (Tibshirani et al., 2020: Equation 6).

The implications of equivalent conditional distributions for Black and White offenders can be subtle. For our application, weighted conformal prediction under a covariate shift requires that the true probabilities of a

[8]The predictor distributions themselves are not altered. The goal is to make the two joint predictor distributions comparable *when an analysis is undertaken.* Also, comparability involves more than predictor means (cf., Oaxaca and Ransom, 1999). Characteristics of predictor distributions beyond means can be related to unfairness.

[9]Weighting differs from matching, which would not be an appropriate adjustment for a covariate shift. One fundamental problem is that there could not longer be a Pareto improvement. Another fundamental problem is that the task is not causal inference but forecasting.

post-arraignment arrest, given the available predictors, are the same for Blacks and Whites. For example, if one has only the predictors age, gender, education, and number of prior arrests available to train the risk algorithm, one must assume that a black male of age 26 with a college degree and 2 prior arrests has the same probability of being arrested after an arraignment release as that of a White male of age 26 with a college degree and 2 prior arrests.[10] For a post-arraignment arrest, the equivalence of $\mathbb{P}(Y|\mathbf{X})$ for Black and White offenders is implausible and extremely difficult to demonstrate even if it were true.

But our focus is on the algorithm. Training only on Whites and then adjusting for a racial covariate shift, racial differences in offender covariate distributions counter the impact of any *previous* differences in $\mathbb{P}(Y|\mathbf{X})$. For example, the two racial distributions for prior arrests can be made to be effectively the same. The assessments of risk *themselves* are now fair. However, one cannot expect any risk algorithm to improve fairness after a release decision is made. Going forward, therefore, Black offenders and White offenders with the same covariate values may be treated differently.

No risk algorithm, even if implemented perfectly, can reform the entire criminal justice system. Reforms must be introduced by other means. Consequently, risk forecasts free of racial content cannot correct for unjustified racial differences in the chances of a post-arraignment arrest. One possible ramification for Black offenders might be underestimates of risk when a fair risk algorithm confronts an unfair post-arraignment world. Some might argue that as a matter of justice, this is a good outcome for many offenders, and more generally provides an objectlesson on the risks of re-arrest in a fair criminal justice system.

We will carry on in the same spirit. Training a risk algorithm on White offenders necessarily means no distinctions are being made between White and Black offenders. We will see later that this alone can be sufficient for fair risk assessments. But if some unfairness remains, the data from which the algorithm learns likely are the problem. Adjusting for a covariate shift can be an effective remedy as long as one appreciates that the forecasts represent a projection into a fair, post-arraignment setting. We are proceeding as if $\mathbb{P}(Y|\mathbf{X})$ post-arraignment is the same for Black offenders and White offenders. But whether this equivalence is holds does not affect the fairness

[10] Recall that we are working with an algorithm, not a model. There are surely predictor variables that, if available, would improve forecasting accuracy. But the absence of such predictors does not invalidate the statistical inference for the forecasts. As we explain later, one can obtain valid inference *conditional on the predictor variables we have* as long as the data are IID or at least exchangeable.

of the risk *algorithm.*

2.3 Constructing Fair Conformal Prediction Sets

Two applications of propensity score weighting need to be distinguished. One is implemented with confusion tables derived from the risk algorithm. Consistent with recommended practice, these table should be constructed from test data (Berk, 2018). Propensity score weighting can be applied as needed to adjust for appearances of *aggregate* unfairness when a confusion table for White offenders is compared to a confusion table for Black offenders.

However, a given offender quite properly may want to know about fairness of his or her forecasted outcome. Comparable confusion tables for Blacks and Whites at best provide an indirect assessment. A second and complementary weighting application employs conformal prediction sets (Lei et al., 2018). This formulation may be unfamiliar to many readers. We provide some details now with further discussion in Appendix A.

One formulation begins by prescribing a statistical test for the null hypothesis that the forecasted outcome class (e.g., re-arrested) for the *given individual* corresponds to that individual's true outcome class. In practice, one test statistic is computed for each case in the test data. The test is then inverted. By inverting the test, one obtains a set of test statistics for all null hypotheses that would *not* be rejected (Rice, 1995: section 9.4).[11]

The test statistic is a conformal score, sometimes called a "nonconformity measure." Loosely speaking, it measures for a given case (e.g., an arrested individual) the degree to which a particular outcome class for Y, here 0 or 1, differs from the likely outcome class based on the predictor values for that case. For case i, this is $1 - \hat{p}_i$ or $0 - \hat{p}_i$ for $\hat{p}_i(Y = y|\mathbf{X} = \mathbf{x})$.[12] For example, If in the test data for a given case $y = 1$, and the fitted $\hat{p}_i = .8$ for $y = 1$, the conformal score is $1 - .8 = .2$. If in the test data for a given case $y = 0$, and the fitted $\hat{p}_i = .3$ for $y = 1$, the conformal score is $0 - .3 = -.3$. The larger the absolute value of the conformal score, the less conforming is the case.

[11] For a more familiar application, imagine testing the null hypothesis that, in conventional notation, $\mu = 0$. One might employ the t-statistic as the test statistic. An inverted test would include all t-statistics and their corresponding means for which the null hypothesis of $\mu = 0$ is not rejected.

[12] There are many ways to construct conformal scores (Gupta et al., 2020). The properties of these different methods are an active research area. The kind of conformal score we have used should perform well in our risk assessment setting because $\mathbb{P}(Y|\mathbf{X})$ can be reasonably well estimated by our application of gradient boosting.

For the test data, the outcome class is known. What does one do for forecasts? The outcome class for such cases is unknown. Indeed, this is precisely the setting when outcome forecasts are needed. For two possible outcome classes 1 or 0, one simple computes a conformal score for each.[13]

Given a ranked set of conformal scores from a relevant test data, it is easy determine how forecasted conformal scores compare to test data conformal scores. Consider the the ranked scores between the .025 quantile and .975 quantile. We call this the null interval. There will be four possible results for a given case.

- Class 1 falls within the null interval, but class 0 does not. The conformal procedure guarantees that for this case Class 1 is the true class with a probability of .95.

- Class 0 falls within null interval, but class 1 does not. The conformal procedure guarantees that for this case Class 0 is the true class with a probability of .95.

- Both Class 0 and class 1 fall within the null interval. There is no formal rationale for treating either outcome class by itself as the true class.[14]

- Neither Class 0 nor class 1 fall within null interval. One has an empty set. The case is treated as a highly unusual realization that some might characterize as an outlier (Guan and Tibshirani, 2019). The case's covariate values are substantially different from those of the training data cases.

A bit more formally, suppose the statistical test uses a value of .05 for α. One then has the 95% conformal prediction set. "Given a method for making a prediction $\hat{y}$, conformal prediction produces a 95% *prediction region* – a set $\Gamma^{0.05}$ that contains y with a probability at least 95%. We call $\hat{y}$ the

[13]This approach can be generalized in several very interesting ways when there are more than two outcome classes (Gupta et al., 2020). The comparative merits of the different methods are still being determined. A discussion is beyond the scope of this paper.

[14]There is some very interesting theoretical work in computer science on how decision-makers and algorithms can can improve fairness and accuracy in such situations (Madras et al., 2018a). Called "rejection learning," a risk algorithm should provide no forecast when there is too much uncertainty or a forecast is inconsistent with specified criminal justice goals. When working in tandem with a human decision-maker, the algorithm becomes adaptive rejection learning because the algorithm learns at what point to defer to the decision maker if, for instance, the decision-maker has access to information the algorithm does not. It can learn not to defer when the decision-maker is being unfair.

point prediction, and we call $\Gamma^{0.05}$ the *region prediction*" (Shafer and Vovk, 2008: 371-372, emphasis in the original). That region can be considered an interval if Y is numerical or a set if Y is categorical. For a binary outcome, if either of the results for the first two bullets occur, one forecasted class is the true class with a probability of .95. If over many offenders a claim is made that the forecasted outcome class is the true outcome class, that claim will be correct for about 95 out of 100 such forecasts. For the third bullet there is more uncertainty because the analysis does not specify which forecast is correct. For the fourth bullet, there may be reason to dig deeper into what makes such cases anomalous.

These properties are valid in finite samples. No asymptotics are required. They remain valid for virtually any estimators of $Y|\mathbf{X}$ that can produce fitted probabilities: logistic regression, neural networks, gradient boosting and more.[15]

Moreover, there is no requirement that any such risk probabilities are the true risk probabilities. In modeling parlance, the fitting procedure's mean function can be (and usually is) misspecified. That is, the conformal approach is valid in finite samples, *given* the performance of the fitting procedure. With a different estimators, different predictors, or different training data, there could be differing, but still statistically valid conclusions.[16]

The one mandatory assumption is that the original data are realized IID or at least exchangeably. When the data are generated by probability sampling implemented as part of a research design, these requirements are automatically met. Otherwise, a strong justification must be provided, typically from subject-matter knowledge and detailed information about how the data were collected. Assume-and-proceed statistics will not suffice. These issues are discussed in more depth in Appendix A.

But what about fairness? Because the risk algorithm is trained on data for Whites only, it cannot incorporate *any* similarities or differences between White and Black offenders. In other words, there can be no unfairness at this point because Black offenders have yet to be considered. Conformal inference can then be fully appropriate and provide valid finite sample guarantees. But

[15]The votes in random forests are not probabilities and are not consistent estimates of the true conditional probabilities for different outcome classes. However, they can be used to construct valid conformal scores because they are positively associated with the corresponding class probabilities (Gauraha and Spjuth, 2018). The price is some loss of asymptotic efficiency (Gupta et al., 2020). The number of elements in the prediction set can be somewhat larger.

[16]The training data are treated as fixed, and any uncertainty they bring to the conformal analysis is ignored. In that sense, a potentially important source of uncertainty is sidestepped.

different joint predictor distributions still can cause unfairness.

A promising remedy is propensity score weighting introduced above. The weighting is done when the relevant quantiles are computed. Continuing with the 95% conformal prediction set, the .025 and the .975 quantiles are computed using the propensity scores as weights. That will make quantiles computed for Black offenders more like the quantiles computed for White offenders. A fair algorithmic risk tool can result.

In the very unlikely case that those weights are known and do not have to be estimated, the weighting does not change the valid finite sample performance (Tibshirani et al., 2020: pages 6-7). In practice, the propensity scores will be estimated. The weighting process for quantiles does not change, but now the probability claims only are valid asymptotically. One needs, therefore, a substantial number of observations (Bühlmann and Hothorn, 2007: section 9.2). In practice, 1000 observations easily should suffice. The inferential goals are unchanged.[17]

3 The Data

To demonstrate the procedures we have summarized, we analyze a random sample of 300,000 offenders at their arraignment from a particular urban jurisdiction. Because of the random sampling, the data can be treated as IID and exchangeable. When data are IID, they are also exchangeable. Exchangeable data do not have to be IID. For conformal inference, only exchangeability is required. (See appendix A.).[18]

Among those being considered for release at their arraignment, one outcome (coded 1) to be forecasted is whether the individual would be arrested after a release for a crime of violence. The follow-up time was 21 months after release.[19] An absence of such an arrest (coded 0) is the alternative outcome to be forecasted. Predictors include the usual variables routinely available in large jurisdictions. Many were extracted from adult rap sheets and similar information from juvenile records. Biographical variables in-

[17] For a detailed description of conformal prediction under covariate shift, see Tibshirani et al. (2020, section 2.2). For the case of estimated weights, see page 8 of Tibshirani et al. (2020).

[18] Even without random sampling, one might well be able to make an IID case because the vast major of offenders at their arraignment were realized independently of one another.

[19] For reasons related to the ways in which competing risks were defined, 21 months was chosen as the midpoint point between 18 months and 24 months. For this demonstration, the details are unimportant.

cluded race, age, gender, residential zip code, employment information, and marital status. There were overall 70 potential predictors.

In response to stakeholder potential concerns about fairness, 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 surrogate for race. Employment history and marital status were eliminated for similar reasons and also because there were objections to using "life style" measures. Juvenile record 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 might believe that arrests for such crimes could be substantially influenced by police discretion, perhaps motivated by racial animus.

The truth underlying many such concerns is not definitively known, but insofar as the discarded predictors were associated with race, potential biases would remain (Berk, 2009). An alternative strategy, were it acceptable to stakeholders, would be to include all suspect predictors related to race and rely on propensity score weighting to remove racial differences. But in the politics of criminal justice risk assessment, appearances can dominate facts.

In the end, the majority of the predictors were prior arrests for a variety kinds of serious crimes, and the number of counts for various charges at arraignment. The other predictors were whether an individual was currently in probation or parole, age, gender, the age of a first charge as an adult and whether there were earlier arrests in the same year as the current arrest. No doubt, we discarded some potentially useful predictors, but many of those were correlated with the acceptable predictors. Any loss of predictive information may be modest and racial biases introduced could in principle be corrected by weighting. For the analyses to follow, 21 predictors were included.[20]

Consistent with our earlier discussion, 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 simply were determined by the numbers of Black offenders

[20]The two age-related variables and whether there were other arrests with the past year are "dynamic variables." 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 by rap sheets.

and White offenders. Each racial split had at least 40,000 observations such that asymptotic requirements are of minor concern.

4 Fairness Results in the Aggregate

We began by training a stochastic gradient boosting algorithm using the procedure *gbm* from the library *gbm* in the scripting language *R* (Friedman, 2001). 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 differently cases that had different outcomes. All tuning defaults worked satisfactorily except that we chose to construct somewhat more complex fitted values than the defaults allowed.[21] 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.[22]

4.1 Confusion Tables without Adjustment

As described above, confusion tables were computed with test data separately for Black and White offenders. Table 1 is the confusion table for White offenders using the risk algorithm trained on Whites and test data for Whites.[23] Resampling confidence intervals could have been provided for the fairness measures described earlier (Berk, 2020a), but with so many observations, sampling error is not an issue. Moreover, a discussion of how the confidence intervals were computed would be an unnecessary diversion.

For our purposes, the main message is the large impact of the cost ratio. Because false negatives were assessed as 8 times more costly than false pos-

[21] For those familiar with stochastic gradient boosting, 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 regression trees with many recursive partitions of the data.

[22] Because of the random sampling used by the *gbm* algorithm, the number of iterations can vary a bit with each fit of the data. Also, the number of trees can arbitrarily vary about 25% with very little impact.

[23] 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 10% of the target usually lead similar confusion tables.

itives, predictions of violence in Table 1 are dominated by false positives. This follows directly and necessarily from the imposed tradeoffs. Releasing violent offenders is so costly that even a hint of future violence is taken seriously. But then, lots of mistake are made. When the risk algorithm forecasts an arrest for a violent crime, it is wrong 85% of the time. In trade, when the algorithm forecasts no arrest for a violent crime, it is wrong only 5% of the time. This too follows from the imposed 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.

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

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

Forecasts of no violence are a very good bet, but the associated 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.[24]

Table 2: Test Data Confusion Table for Black Offenders Using White-Trained Algorithm (41% Predicted to Fail, 11.3% Actually Fail)

Actual Outcome	No Violence Predicted	Violence Predicted	Classification Error
No Violence	55791	34206 (false positive)	.38
Violence	4137 (false negative)	7357	.35
Forecasting Error	.07	.82	

Table 2 is the confusion table constructed from the test data for Black

[24]In real settings, risk forecasts properly are influenced by many policy-related constraints beyond the preferred tradeoffs between false positives and false negatives. For example, there is usually an upper bound to the number of arraigned offenders who can be detained within existing jail capacity.

offenders using the White-trained boosting algorithm. Tables 1 and 2 are similar. No dramatic fairness concerns surface when the proportions on margins of the two tables are compared. Forecasting errors are virtually the same. For Blacks, the false positive rate is a bit higher, and the false negative rate is a bit lower. But, putting these two modest differences together, implies that overprediction could be a larger problem for Black offenders than White offenders. And indeed, whereas 28% of Whites are predicted to fail post-release, 41% of Black offenders are predicted to fail post-release. This is exactly the sort of disparity that can lead to accusations of racial bias or stated more gently, unfairness.

4.2 Confusion Tables with Adjustment

Clearly, the fault does not lie with the algorithm. White offender and Black offender risks were determined by the same fitted algorithm trained only on White offenders. The algorithmic machinery is exactly the same for each individual. Therefore, an overprediction disparity must be caused by the data. Whites and Blacks must bring at least somewhat different predictors distributions when risks are computed from test data. Berk and Elzarka (2020) recognized this problem and make several efforts to compensate. Their most successful approach was to make the failure base rates for Blacks and Whites more alike, but this is an ad hoc strategy that does not directly address potential disparities in the predictor distributions.

Using a propensity score weighting to compensate for a covariate shift, we take aim directly at the joint predictor distributions for Black and White offenders. The weights for each case are odds ratios defined as

$$\hat{w}_i = \frac{\hat{p}(x_i)}{1 - \hat{p}(x_i)}, \tag{2}$$

where $\hat{p}(x_i)$ is an estimate of $\mathbb{P}(\text{Race} = 1|X = x)$, with Race = 1 an indicator variable denoting that the offender is White (Race = 0 is for Black offenders), and X is a set of predictors (Tibshirani et al., 2020: equation 8). In principle, any classifier can be used for estimation. We again used stochastic gradient boosting procedure *gbm* employing the same predictor variables as before, but with a target cost ratio of 1 to 1. The propensity score weights were applied subsequently to adjust the confusion table for Black offenders. Weights were larger for Black offenders who were more like White offenders. Offenders with larger weights were counted more heavily when confusion table cell tallies were computed.[25]

[25]In R, the procedure *wtd.table* from the library *questionr* was used.

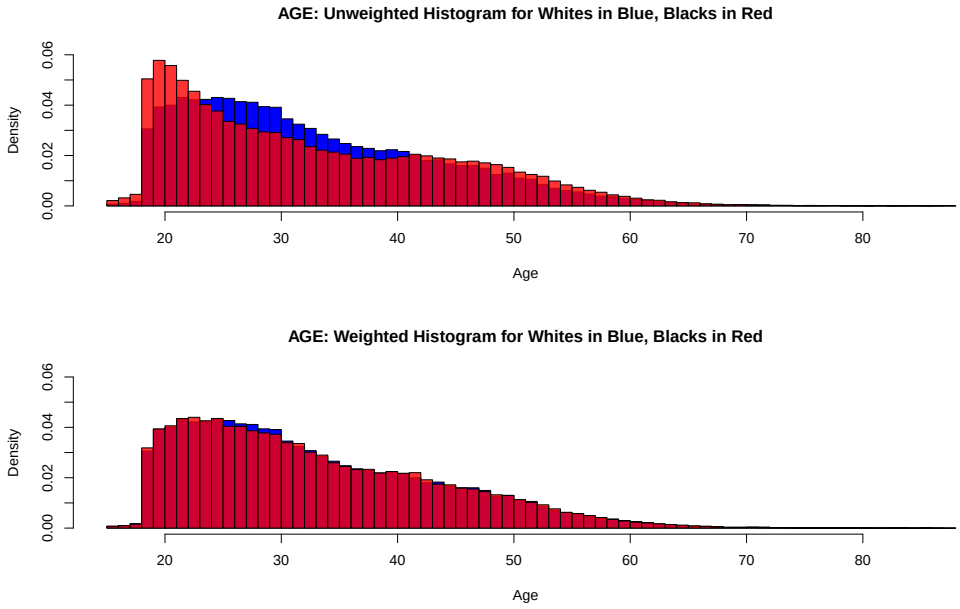

Figure 1: Unweighted and Weighted Histograms for Age

Demonstrations of the impact of such adjustments are displayed in Figures 1, 2, and 3. For each, there are unweighted and weighted overlapping histograms. For the weighted histograms, the entire joint predictor distribution for Blacks was altered. Figure 1 shows the adjustment impact on the age of the offender. Figure 2 shows the adjustment impact on the earliest age at which an offender was charged with a crime. Figure 3 shows the adjustment impact on the number of prior arrests for a crime of violence.[26]

These three predictors are the top three based on the size of each predictor's contribution to the boosting fit of post-release violence.[27] As a group, they account for about 75% of the fit quality, measured as the average contribution to the fit over the ensemble of boosted regression trees (Friedman, 2001). Each contribution is standardized such that the sum of the predictor contributions 100%. For the top three predictors, each variable's contribution was larger than 13%. Below these three, each variable's contribution

[26] We used the R procedure *wtd.hist* in the library *weights* library for each of the weighted histograms. A Black offender who's predictor values are more like White offenders' is "upweighted" so that in effect, the Black offender is counted, say, twice as frequencies for the histogram are computed.

[27] The measures of predictor importance is readily available in the *gbm* output.

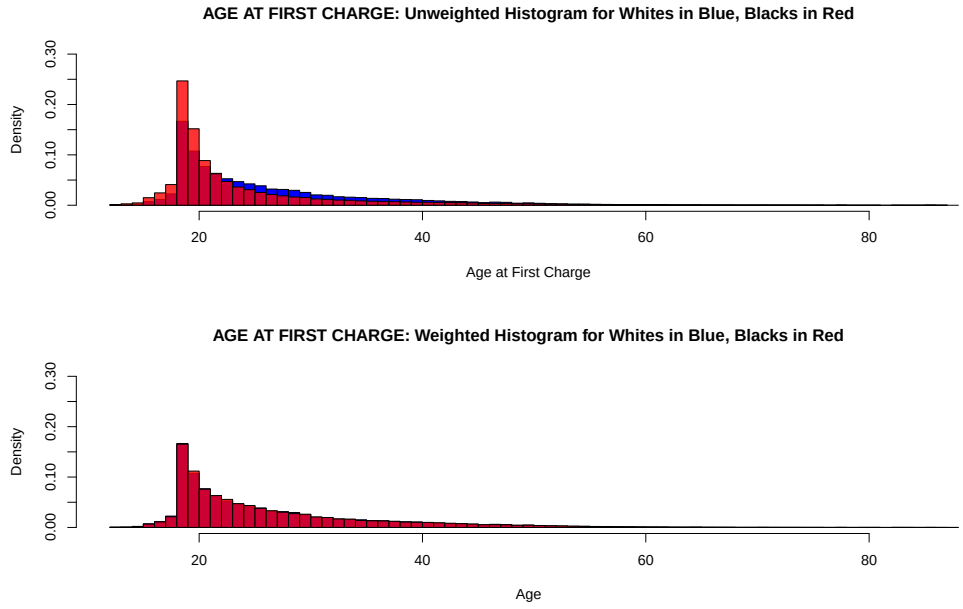

Figure 2: Unweighted and Weighted Histograms for Age at First Adult Charge

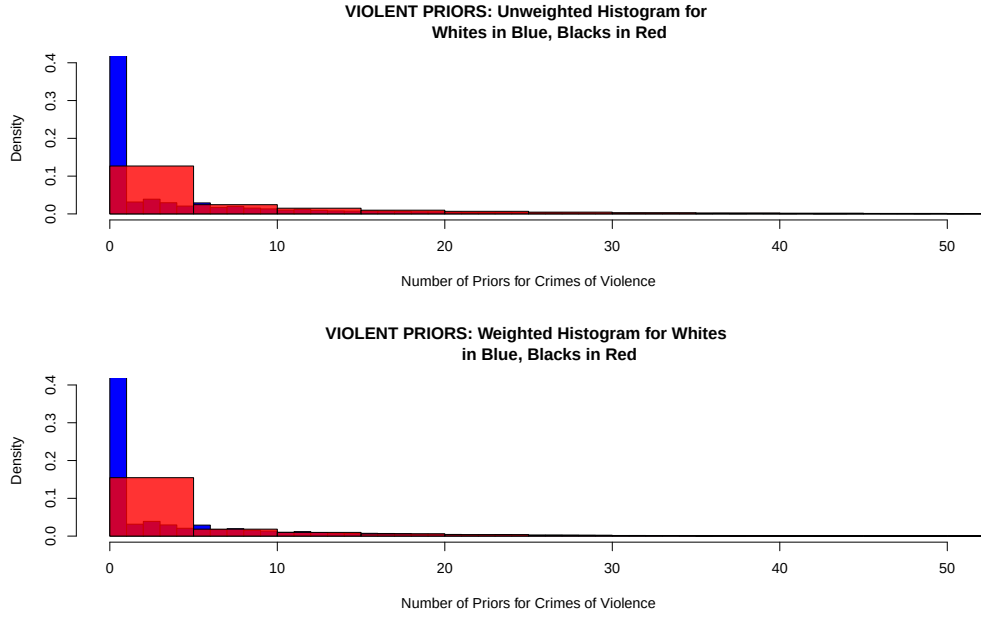

Figure 3: Unweighted and Weighted Histograms for Violent Priors

was under 5%, most less than 1%. In short, a strong clustering of predictors importance indicates these three predictors are most responsible for fit quality.

The top display in Figure 1 shows the unweighted results for an offender's age. The blue distribution is for Whites. The red distribution is for Blacks. Represented in orange is where the Black distribution has a greater density than the White distribution. The bottom display in Figure 1 shows the results when the Black test data are weighted by propensity scores. Overall, the unweighted histograms are similar except for young offenders, where Blacks are relatively more common than Whites. This difference could well foster greater overprediction for Blacks because young offenders commonly are predicted to be higher risk. When the test data for Black offenders are weighted to make the two distribution more alike, the differences between the two distributions virtually disappear.

Figure 2, shows the results for the predictor age at first adult charge. The main disparity between Black and White offenders is that Black offenders are more likely to have their earliest charges at a younger age. This too could help explain a more serious overprediction problem for Blacks.

After weighting by propensity scores, the two distributions overlap nearly perfectly.

In Figure 3, one can see in the top display that the White offenders are relatively more likely than Black offenders to have very few prior arrests for crimes of violence, and often no such arrests at all. However, the weighting shown in the bottom display does not materially help. The two plots are nearly identical. The most visible difference is that the lower left orange rectangle in the weighted histogram is slightly taller.

A very important lesson has been illustrated. Although the number of priors for crimes of violence is a very influential predictor of a post-release arrest for a violent crime (as one might well expect), it is not an influential predictor of race when the propensity scores were calculated. The number of violence priors dropped from the 3rd most important predictor to 10th most important with a fit contribution of less than 2%. The lesson is this: in practice, a substantial and meaningful adjustment for a particular covariate requires that the covariate be an influential predictor of the outcome and also influential when the adjustment weights are computed. If a predictor is effectively unrelated to race, it cannot alter a predictor distribution that varies by race. If it is unrelated to the outcome variable, the adjustment will not affect the subsequent confusion table even if the predictor is related to race.

We can now return to the confusion tables. Because the data are IID, all of the proportions computed from the confusion tables can be interpreted as probabilities. However, because the weights are estimated for the weighted tables, these only have asymptotic guarantees.

Table 3 is the confusion table when the predictors from the test data for Black offenders are adjusted for a covariate shift. The weighting is done with the propensity scores used for the weighted histograms. The weighs were standardized so that the number of Black offenders in the test data is not altered when Table 3 is constructed.

Table 3: Weighted Test Data Confusion Table for Black Offenders Using White-Trained Algorithm (29% Predicted to Fail, 11.3% Actually Fail)

Actual Outcome	No Violence Predicted	Violence Predicted	Classification Error
No Violence	67578	24255 (false positive)	.26
Violence	4549 (false negative)	5157	.46
Forecasting Error	.06	.82	

Table 3 and Table 1 are almost identical and for both, 29% of the of-

fenders are predicted to fail. There is no longer any evidence of unfairness. Should such results materialize when stakeholders are able to examine the confusion tables, it is hard to imagine complaints about inequities.[28]

The primary conclusion from such comparisons is a confirmation that aggregate unfairness results not from the algorithm but from differences in the predictor distributions of the offenders. By training only on White offenders, the algorithm is absolved of responsibility. That by itself leaves unclear the sources of any residual unfairness. Adjusting for a covariate shift identifies those sources and corrects effectively. Because the predictors responsible are identified, one can consider why those particular features of offenders differ by race and also affect confusion table evaluations of the risk algorithm.

However, adjustments for the covariate shift do *not* affect how the gradient boosting classifier performs because propensity score weighting is introduced as the confusion table is constructed; only the confusion table is affected. The prior classification of individual offenders is unchanged. Put another way, we have modified the manner in which aggregate performance is represented and in so doing, have isolated the particular predictors responsible. But overprediction for Black offenders remains. The same would apply to unlabeled realized data for which forecasts were needed. In short, the weighting is at this point a diagnostic procedure not a remedy.

It is difficult to anticipate how well adjustments for a covariate shift would perform with other data from other settings. For this analysis, the two predictors that accounted 51% of the fit for the estimates of risk, accounted for 33% of the fit when the propensity score weights were computed. The same two predictors dominated both. This joint dominance was an important reason for the success when Table 1 and Table 3 were compared. For other data, no such dominance is required, but a set of predictor variables, or highly correlated proxies, must drive both the risk assessment and the weight construction. In practice, the only way to determine whether propensity score weighting can reduce or even removes evidence of unfairness revealed by confusion tables is to try it.

In summary, if training only on White offenders removes racial unfairness in the subsequent confusion table, there need be no concerns about

[28]Nevertheless, they may argue that there is too much overprediction for *all* offenders. The reasoned response would be to alter the cost ratio and make new tradeoffs. For example, if false negatives are made less costly, there will be fewer false positives contributing to overperdiction but more false negatives will increase the possibility of "underprediction." With underprediction, there could be an increase in the number of offenders released who pose a serious threat to public safety.

the impact of a covariate shift. Moreover, there is no evidence that in the aggregate the risk procedures produces unfairness. If there is remaining unfairness in the confusion table, weighting can identify the predictors responsible, presumably providing an opportunity to consider why those particular predictor distributions differ for Black offenders compared to White offenders. But the weighting is not applied to the training data or how the risk algorithm performs. Consequently, another form of risk assessment is necessary. For that, we turn to weighted conformal prediction sets, which has several desirable properties.

5 Results at the Case Level Using Weighted Conformal Prediction Sets

We have so far considered fairness in confusion tables only. Such tables are typically used to provide aggregate fairness measures for the performance of risk algorithms. From a policy perspective, aggregate performance is an appropriate yardstick by which one hopes to judge how well a policy works. But performance on the average provides little information about individual cases. Offenders, whether Black or White, and decision makers probably want to know about the performance of risk forecasts at the individual level and whether racial differences are present.

The gradient boosting classifier and the test data results in Table 3 provide a start. For all offenders forecasted by the classifier to not be arrested for a violent crime, the estimated probability that the forecast is wrong is 0.06; for about 94 out of 100 (i.e. $1 - .06$) arraigned offenders, a forecast of no arrest for a violent crime will be correct. By similar reasoning, a forecast that an offender will be arrested for a crime of violence will be correct with a probability of only 0.18.[29]

But such probability estimates put a thumb on the scale. It is reasonable to treat the training data and trained algorithm as fixed because in risk assessment practice, neither changes when forecasts are made.[30] However, the outcome class is not known when forecasts are obtained – If they were, there would be no need to undertake forecasting. The forecasted outcome

[29] As before, the large number of policy-mandated false positives increases forecasting error substantially.

[30] Although one can certainly imagine employing some form of online learning (Kushner and Yin, 2003), in which the training data and algorithm training were updated, how this would work when the new realizations are unlabeled (i.e., the outcome must be forecasted) is unclear, especially within a conformal inference framework. At the very least, exchangeability could be problematic.

class is an estimate with intrinsic uncertainty. Conditioning on the forecast, ignores that uncertainty, which can lead to overly optimistic assessments of performance.

Moreover, there is reason to be uneasy with forecasting decisions that solely minimize Bayes error. Suppose the gradient boosting classifier estimates that for a given offender the probability of no arrest for a violent crime is 0.49 and the probability of an arrest for a violent crime is 0.51. An arrest for a violent crime necessarily is the forecasted outcome. Yet, the two probabilities are nearly the same; the evidence favoring an arrest is weak. Another random split of the available data into a training dataset and a test dataset, for example, easily could yield a different forecast. A more sensible conclusion might be that the classifier cannot for this offender determine which outcome is substantially more likely. Yet, such uncertainty is not captured when conditioning on the forecasted outcome class. A criminal justice decision maker may not recognize that although the classifier made a call, the results were actually too close to call. There are some ways to address this problem (Berk, 2017), but they are ad hoc and lack formal rigor.

Weighted conformal prediction sets promise a better approach. One does not condition on a particular forecast and one can forecast a prediction set with more than one element. Consider again a 95% conformal prediction set. Recall from section 2.3 that for a given case and two outcome classes, there are four possible inferential outcomes. Two inferential outcomes specify a true class with a probability of .95, one inferential outcome cannot determine which outcome class is the true class, and one inferential outcome treats the case as a possible outlier. The forecasted prediction set depends on applying appropriate conformal procedures that take results from the Bayes classifier but move well beyond those results (Tibshirani, 2020: section 2.2).

For our Black offenders and White offenders, the unweighted 95% conformal prediction set had a lower bound of -.73 and an upper bound of .58. The weighted 95% conformal prediction set had a lower bound of -.72 and an upper bound of .58. The regions defined by the two quantile thresholds are virtually identical for Black and White offenders. Therefore, there is no evidence that differences in joint predictor distributions for Blacks and Whites mattered. For the conformal inference, *training on the White offenders alone was sufficient.* A fair risk algorithm was constructed without any adjustment for a covariate shift, and therefore, no need to assume that the two conditional distributions post arraignment $\mathbb{P}(Y|\mathbf{X})$ were the same. We will proceed, nevertheless, with the weighted results to reinforce under-

standings of how they are used.[31]

The lack of important racial unfairness may be surprising in light of our earlier results, but when conformal scores are used to construct a prediction set, one is working with the quantiles. Information in the scores themselves is collapsed into ranks. Modest differences in fitted risk probabilities between Blacks and Whites, caused by disparities in their joint predictor distributions, often will not matter. For .95 conformal prediction sets, weighting only makes a difference in the immediate neighborhood of .025 or the .975 quantile; most of the conformal scores will have no role in determining the value of the .025 and .975 quantile. In that sense, quantiles can be quite resistant to differences between predictor distributions by race. This is a beneficial feature of the method, not a flaw. Whatever the racial disparities built into risk predictors, it is difficult for those disparities to affect conformal probabilistic claims about the true outcome class. Adjustments for a covariate shift along with its accompanying complications may well be unnecessary.

Table 4: Conformal Prediction Set for 15 Randomly Selected, Test Data Cases for Black Offenders Using the 95% Conformal Prediction Set (1 = An Arrest for a Violence Crime and 0 = No Arrest for Violent Crime)

Case	Conformal Prediction Set
1	1 or 0
2	1
3	1
4	1 or 0
5	1 or 0
6	1 or 0
7	1 or 0
8	0
9	0
10	1 or 0
11	0
12	1 or 0
13	1 or 0
14	1
15	1

There are other advantages of the conformal approach that are perhaps difficult to appreciate at first. Table 4 shows the forecasted outcome class

[31] In practice, it may be preferable to avoid weighting unless the evidence for unfairness is strong. Recall that because the weights are estimated, conformal finite sample guarantees are replaced by conformal asymptotic guarantees.

using the 95% conformal prediction set for 15 cases randomly chosen from the test data. An entry of "1 or 0" indicates that both possible outcomes fell inside the threshold quantiles. One cannot confidentially determine from the data which class is the true class. A little more than half the time, this was the result. Four times an arrest for a violent crime is forecasted, and three times no arrest for a violent crime is forecasted. For the 95% conformal prediction set, forecasts of 1 or forecasts of 0, will be correct with a probability of .95 over cases for which forecasts are sought. The point is that the .95 probability applies to each forecasted prediction set. Whatever the prediction set forecasted, it will be correct with a probability of .95. (i.e. if α is set to 0.05). But any particular prediction set such as {0} or {0,1} might commonly appear, rarely appear, or somewhere in between.[32]

In addition, from these analyses and the 95% conformal prediction set, often there will be no statistically definitive prediction for particular individual cases. More than one outcome class will be in the prediction set. If stakeholders are prepared to accept prediction sets with smaller probabilities, prediction sets with fewer outcome classes become more likely. For example, the majority of forecasts might only include the no arrest outcome, but with a probability of 0.80. Stakeholders have the option of balancing the likely size of prediction set with the associated probability that the prediction set is correct by the value of α chosen.

It is important to emphasize that when a classifier such as stochastic gradient boosting chooses a single outcome class as the forecasted class, the choice is *required* to minimize Bayes loss. The class with the largest predicted probability is the winner, whether the competition with other classes is close or not, and even if the differences between the probabilities could easily be a consequence noise. Conformal prediction sets do not force a single choice, and in that sense, are more responsive to the data. Conformal prediction sets are not less precise. They are more demanding.

5.1 Accuracy for Prediction Sets

For conformal prediction sets, two useful measures accuracy are immediately available. For a given value of α, the probability that the forecasted prediction set is the true prediction set is $1 - \alpha$. The larger the value of $1 - \alpha$, the more certain one can be that the forecast is correct. This is analogous to what can be extracted from the usual confusion table, but when single

[32]The cases for which forecasts are needed must be realized in an exchangeable fashion from the the joint probability distribution responsible from the training and test data. These or comparable requirements apply to all risk assessment and all forecasting.

outcome class is not a forced choice. In addition, for a given value of α, one has a measure of precision. The smaller the prediction set, the more precise the forecast. This too can serve as a measure of accuracy, and there is nothing comparable from a confusion table.

Recall, that the tradeoff between certainty and precision can be controlled by the researcher. One can, for instance, seek more precision at the cost of less certainty. This may not be as painful as it sounds because in policy settings, decision makers are often satisfied with certainty well below ritual probability of .95. For example, a prediction set including only an arrest for a violent crime may require that $1 - \alpha$ be designated as 0.80; the odds are 4 to 1 in favor of an arrest for crime of violence. Those odds may be sufficient for a magistrate sensibly to detain the offender.

The issues can be more subtle if there are three outcomes, such as a misdemeanor arrest, a felony arrest, or no arrest. For $1 - \alpha = .95$, the prediction set might include a felony arrest and a misdemeanor arrest. It is a very good bet that the offender will be arrested while on probation, although whether for a misdemeanor or a felony cannot be determined with a high probability. That might be sufficiently informative for a magistrate to properly detain the offender. In other words, a prediction set with more than a single element is not necessarily problematic if the outcome classes point toward the same decision.[33]

Finally, if for policy reasons, a single outcome class must be selected, one strategy is accept a modest level of certainty (e.g., .75), insofar as a single class falls in the prediction set. This approach is fully consistent with conformal prediction. Another strategy is force the desired precision by choosing the outcome class with the largest fitted probability. This can follow from the Bayes classifier, but it also can follow from conformal prediction sets from well justified ways to compute conformal scores. The price for coercing "perfect" precision is that a proper reading of uncertainty can be muddy.

6 Conclusions

Unfairness can be introduced by the risk assessment methods. For statistical learning tools, unfairness usually is not caused by the algorithm itself, but by the data on which it is trained. There is a very active cottage industry on

[33]Prediction sets with more than two outcome classes can be effectively analyzed by the methods used in this paper, but better methods are being developed that, among other things, promise more precision.

ways to alter the training data and means by which the data are processed, to reduce, and ideally eliminate, unfairness. These efforts are improved when there are clear and encompassing definitions of fairness coupled with a rich understanding of how the data are generated.

Perhaps the major obstacle for the procedures we propose, and for all others that address proper statistical inference for risk assessment, is the nature of the data generation process. If the inferences are model-based, the model must be correct; the model prescribes how the data must be generated. If the model is wrong, the analyst is working with the wrong data. Yet, justifying a particular model specification can be daunting (Freedman, 2009). Alternatively, a model is better seen as an estimator of interesting population functionals (Buja et al., 2019a; 2019b). The estimation target is an acknowledged approximation of the truth. Algorithms are also approximations of the truth. They can be seen as estimation approximations of true response surfaces.

For algorithms and models that are wrong, proper statistical inference depends on IID or at least exchangeable data. The case for IID and/or exchangeable realizations will necessarily depend on subject-matter expertise (Berk, 2020a). One must argue that the data are generated by processes having largely the same underlying properties as probability sampling from a single, very large population. IID would be violated, for example, if data for arraigned cases were collected in a jurisdiction for which arraignment policies and administrative practices changed substantially over the relevant time period; there could be several populations. Perhaps new or revised criminal statutes were implemented. There could be dependence as well if the cases were drawn in clusters from some courtrooms and not others. Then the data would also not be exchangeable and "assume-and-proceed" statistics is not a solution.

The central role of target cost ratios must be acknowledged, and one or more target cost ratios specified. The challenges cannot be sidestepped because failing to address cost ratios means accepting whatever the training data and algorithm determine, whether responsive to the real tradeoffs or not. If not responsive, inappropriate decisions are more likely.

There can be practical complications when our preferred procedures to estimate risk produce uninformative results. For example, the prediction set for a particular offender may empty, implying that the case is a statistical outlier. Under such circumstance, an honest appraisal is that there is very little guidance, and a reasonable response is to rely on other information. A decision might be made to delay any risk assessment until more particulars for the problematic case are collected (Berk and Sorenson, 2016; Madras et

al., 2018a).

Finally, training a risk algorithm on a single more privileged group can formally absolve an algorithm itself from any charges of unfairness. Adjustments for a covariate shift can clean up residual unfairness. However, the Pareto improvement that results must pass political and legal muster before our proposals could properly be implemented. These challenges have yet to be addressed and could well be contentious. One issue is whether there would be under our approach real injuries (Lujan v. Defenders of Wildlife, 1992). One cannot bring suit in federal court in the absence of "injury in fact." Another issue is whether there is a violation of "equal protection" under the fifth and fourteenth amendments to the U.S. Constitution (Coglianese and Lerh, 2017: 1191:1205). In point of fact, a central goal of our research is to make protection more equal.[34]

[34]The analysis steps are easily summarized for two protected groups, one understood to be more disadvantaged than the other. Illustrative Code in R is provided in Appendix B. It is little more than a cobbling together of existing R procedures and can be easily generalized when there are more than two protected groups.

Appendix A: Notes on Conformal Prediction Regions

Much of justification for machine learning statistical inference requires that the observed data are randomly realized independently from the same joint probability distribution. In practice, this can be a challenging requirement. A slightly weaker requirement is that the realized observations are exchangeable. Exchangeability also can be difficult to fulfill in practice, but provides for some alternative data generation mechanisms and a somewhat different suite of inferential procedures. Moreover, depending on the form of inference, asymptotics may not be required for valid inference.

In these notes, we consider the exchangeability option as exploited by conformal prediction regions. The region is called an interval if Y is numeric and a set if Y is categorical. We initially seek an analog to confidence intervals and draw heavily on the work of Lei and colleagues (2018) that, in turn, builds on work for Vovk and colleagues (2005; 2009). We then move on to very recent extensions of conformable prediction sets (e.g., Tibshirani et al., 2020). The technical literature can be difficult because of varying notation, alternatives to traditional concepts, and an evolving literature that has yet to settle on a common narrative. We hope these notes are relatively accessible.

Independent and Identically Realized Observations

Many of the foundational concepts for conformal prediction intervals can be approached initially with concepts based on probability sampling from finite populations. Imagine that for a single random variable Y there is a very large population with N observations. Employing the conceptual equivalent of random sampling *with* replacement, nature generates n realizations of y-values that become the data on hand. We say that the realizations are identically distributed because they are all generated from the same parent population, and they are independent because whether a particular case is realized does not affect the chances that any other case is realized. In common shorthand, the y-values are said to be IID: independently and identically distributed.

The realized data is also exchangeable. Suppose we sample in sequence two cases: case A and case B. Each is sampled with a probability of $1/N$ regardless of the order in which they are sampled. The probability of the sequence $\{AB\}$ is $1/N \times 1/N$, which is the same as for the sequence $\{BA\}$. The order of selection does not matter. Case A and Case B are exchangeable.

It will help the exposition to follow if one imagines the sequence of realized y-values stored in a column vector with n entries with each row in the vector identified by a row number 1 through n. Each row number is sometimes called the row "index" of a realized observation. Case A sampled from the population might in the first row with an index of 1, case B sampled from the same population might be in the second row with an index of 2. But placing B in the first row with an index of 1 and A in the second row with an index of 2 does not matter because the two cases are exchangeable. The same reasoning applies to many sampled cases.

Effectively the same properties follow if the exposition were undertaken with observations realized independently from a hypothetical population of limitless size formally represented by a probability distribution for Y. When there are predictors as well, this is the data generation formulation commonly used in machine learning applications: Cases $(\mathbf{X}_i, Y_i)$ are realized IID data from a joint probability distribution from which a limitless number of observations could be randomly generated. Exchangeability follows, and is fundamental for conformal prediction sets. But exchangeability also can achieved without the assumption of IID realizations.

Exchangeability without IID Realized Observations

Imagine now that the random sampling is done *without* replacement from a very large, finite population. This is sometimes called "simple random sampling." With each new sampled case, the population size is reduced by one. That is, $\mathrm{P}(y_1) = 1/N$, $\mathrm{P}(y_2) = 1/(N-1)$, $\mathrm{P}(y_3) = 1/(N-2)\ldots$. The realized values are not independent because with each realization, the probability of selection changes; the probability of selection for any case depends on the order of selection. The assumption of IID realizations does not hold.

However, sampling without replacement produces exchangeable realized cases. Consider again a very simple example. Suppose we sample two cases: case A followed by case B. Case A has a selection probability of $1/N$, and case B has a selection a probability of $1/(N-1)$. The probability of the selection sequence {AB} is $1/N \times 1/(N-1)$. But it is exactly the same for the reverse sequence {BA}. Order of selection does not matter, and the two realizations A and B are exchangeable. And as before, one usefully can consider the order of selection as row numbers in a vector of realized values.[35]

[35]For any given sample size n under sampling without replacement, the *samples* are independent. And for a given sample size, each possible sample is equally probable.

Some Important Properties of Exchangeable Data

Under exchangeability, case index values have a very useful property. Each case has the same probability of having an index value of 1 as any other case. Each case has the same probability of having an index value of 2 as any other case. And so on. Therefore, if there are 20 exchangeable observations in a dataset, P= (1/20) for each case having an index of 1, or each case having an index of 2 or each case an index value of 3 and so on. This means that the probability distribution of index values is rectangular. From this property, one directly can compute quantiles. For example, the probability that a given case will have an index value greater than 17 is 3/20, or more formally, P(index $> 17) = 0.15$. It follows that under exchangeability, a permutation distribution can serve as a distribution for certain null hypotheses that yield a useful prediction interval or set. This is a key feature of what follows.

Exchangeability can be produced by a variety of data generation mechanisms beyond random sampling without replacement. For example, one can have the equivalent of stratified random sample with replacement. The realizations are still not independent, but they are exchangeable.

Conformal Prediction Sets

Forecasting can be understood as a form of statistical inference in which one computes a point estimate for a value that has not yet been observed.[36] For a conformal prediction region, inferences are being drawn from the exchangeable data on hand to the finite population or probability distribution responsible for the data. The estimation target is the true predicted value in the population or joint probability distribution. Because in this paper we emphasize categorical y-values, we will focus on conformal prediction sets. Y is composed of outcome classes.

Suppose one has n exchangeable test data observations $y_1, y_2, \ldots, y_n$ for a single random variable Y, and one wishes to forecast from a set of predictor variables a new realized value y_{n+1}. No matter how that forecast is estimated, it could important to know the probability that the forecast is correct; the forecasted value is the same as the true value.

An essential step is to define a measure, called a "conformity score," also called "nonconformity measure." For a categorical Y, one simple approach computes for each case the disparity between its actual outcome class and

Routine statistical inference for common parameters, such regression coefficients, is then easily undertaken.

[36]The terms forecasting and prediction will be used interchangeably.

a fitted outcome probability from the risk algorithm. For example, if one outcome class is coded 1 and the other outcome class is coded 0, and from the risk algorithm one has the fitted probability that the outcome class is 1, the conformal score can take two forms: $1 - \hat{p}_i$ or $0 - \hat{p}_i$. These can be seen as case-by-case residuals that measure how well a fitted risk probability conforms to the actual outcome class.[37]

The distribution of conformal scores summarizes a form of heterogeneity derived from the test data and the fitted risk algorithm. Conformity scores, just like the y-values are exchangeable. Quantiles from the distribution of conform scores can, therefore, used to compute probabilities when the scores are ordered from low to high. For example, conformity scores fall below the median score with a probability of .50. More important for our purposes, conformal scores fall within the the 95% conformal prediction set with a probability of .95; they are the middle 95% of the conformal scores The exchangeability of conformal scores justify these inferences.

A conformal prediction interval can be used the test the null hypothesis that a forecasted outcome class is the true class. Conformal scores are constructed using the true outcome class in the data on hand. The conformal score distribution, therefore, represents variation when each true outcome class is compared to its risk algorithm's fitted probabilities. It is, therefore, *the null distribution when the true outcome class is known.* Conformal scores can be seen as test statistics.

One also can compute conformal scores for a forecasted outcome class. Predictor values for case needing a forecast are known. Using the the fitted risk algorithm, a fitted probability is easily computed exactly as before. There are two possible outcome classes: 1 and 0. For each, a conformal score can be computed just as when the actual outcome class was known.

Suppose the hypothesis test's critical value is set at $\alpha = .05$. This means that one will be working with the 95% conformal prediction set (i.e., $1 - \alpha$). From the test's inversion, the 95% conformal prediction set contains the collection of y-values in conformal score form for which the null hypothesis is *not* rejected at the .05 level.

As addressed in the body of the paper, there are four possible inferential results.

- The class coded 1 has a conformal score that falls in the conformal prediction set, but the class coded 0 does not. One can say that the

[37] As will explained shortly, we favor "split" conformal inference for which one uses training data to fit the the risk algorithm and test data to construct conformal scores (Lei et al., 2018).

class coded 1 is the true class with a probability of .95.

- The class coded 0 has a conformal score that falls in the conformal prediction set, but the class coded 1 does not. One can say that the class coded 0 is the true class with a probability of .95.
- Both classes have conformal scores that fall in the conformal prediction set. One cannot conclude which outcome class is the true outcome class.
- Both classes have conformal scores that fall outside of the conformal prediction set. Some treat these cases as outliers that cannot be evaluated properly with the existing data (Guan and Tibshirani, 2019)

6.1 Conformal Prediction Sets for Classification Using Split Samples

It is easy to summarize the steps involved constructing conformal prediction sets that provide uncertainty inferences about forecasted outcome classes. As already noted, we favor the split sample method for reasons provided by Lei and his colleagues (2018).

1. Separate the data into two, random disjoint subsets.
2. Fit a $Y|X$ to the first split. One can use some form of the generalized linear or additive model, a flavor of machine learning, or some other procedure.
3. Obtain the fitted values for the second split using the fitted algorithm that was applied from the first split. From these fitted values and the known outcome class values, construct the conformal scores. Here, we use $1 - \hat{P}_i$ or $0 - \hat{P}_i$ depending on the actual outcome class.
4. Compute the $1 - \alpha$ conformal prediction set.
5. Compute the conformal scores for case(s) needing a forecast. There will be one conformal score for each outcome class value and one such pair for each case.
6. Determine which conformal scores fall inside the conformal prediction intervals to arrive at the results.

These steps apply as well when there are more than two outcome classes, although there will some changes in details. For example, the risk algorithm must be able to handle the multinomial outcome case. Also, with some other changes in details, the outcome can be numeric (Lei et al., 2018).

6.2 A Covariate Shift and Conformal Prediction Sets

Conformal prediction sets with valid finite sample properties require exchangeable data. Suppose there are two available datasets: A and B. Each by itself is exchangeable. However, the predictor distributions differ. Even though for both $\mathbb{P}(Y|\mathbf{X})$ is the same, $\mathbb{P}(\mathbf{X}_A) \neq \mathbb{P}(\mathbf{X}_B)$. Trying to use both datasets in the same conformal analysis will fail because combining the two will preclude exchangeability. When applying split conformal methods, for instance, one might wish to use dataset A for training and dataset B for the the construction of conformal scores. Or more simply, the goal may just be to increase the number of observations being analyzed. How one might properly proceed is discussed by Tibshirani and his colleagues (2020).

For concreteness, suppose one has access to data from hospital A and hospital B. Although age and all other available predictors have in both hospitals the same relationship with whether a patient survives, patients in hospital B are on the average somewhat older and be more likely to be male. Should data from both hospitals be used in the same conformal analysis, exchangeability is lost. The same would apply if the shapes or variances of the two joint predictor distributions differed.

If it is really true that $\mathbb{P}(Y|\mathbf{X})$ is the same in both hospitals, there is a relatively simple solution. One can weight the data from hospital B so that $P(\mathbf{X}_A) \approxeq P(\mathbf{X}_B)$. In practice, the weights will be unknown. But, empirically determining the weights for the joint predictor distribution from hospital B can be addressed as a conventional classification problem.

A binary response variable is defined equal to 1 if an observation comes from hospital A and equal to 0 of an observation comes from hospital B. Pooling the two datasets, a logistic regression, or some other classifier, can be applied with the binary variable as the response, and the common predictors from the two hospitals are regressors. The fitted values easily are transformed into the odds of an observation coming from hospital A compared to hospital B. These odds are used as weights to make the joint predictor distribution from hospital B to be more like the joint predictor distribution for hospital A.

But there is a price. The algorithm used to fit survival and the algorithm used to fit hospital A versus hospital B are likely to wrong. Neither set of

fitted values then converges asymptotically to their true values. However, as long as they converge to approximations of their true fitted values – the key requirement is convergence – there can be valid asymptotic statistical inference. Most popular fitting algorithms are likely to properly converge and valid inferences from conformal prediction sets will remain viable in large samples. But valid inference for small samples available before weighting was introduced is lost.

As an empirical matter, whether the two joint prediction distributions are sufficiently comparable after propensity score adjustments should be examined. Just as in the body of the paper, a good start is to determine the most important predictors for the risk algorithm (e.g., fitting an occurrence of a death). Then, does weighting make the distribution of each such predictor from hospital B sufficiently overlap with the same predictor's distribution from hospital A? Unfortunately, further research is needed to operationally define "sufficiently."

Appendix B: Some Illustrative Code in R Using A Fictitious Dataset

The code include below illustrates the key steps. There are many details to be filled in depending on the data and how the analysis unfolds. Also recent work indicates that one can get results with at least somewhat better statistical properties. The approach taken here is "plain vanilla." In practice, there would be many more predictors

The somewhat stylized code assumes an R data frame with exchangeable data randomly split into disjoint splits, one as training data and one as test data, and each of these split into data for Black offenders and data for White offenders: TrainW, TrainB, TestW, TestB. The specified relative costs are just illustrative.

```
## Fit classifier of choice on Whites only. The Data set has a binary Y
## in the first column follow by three predictors and race.

library(gbm) # Most defaults accepted
CostWeights<-ifelse(TrainW[,1]==1,3,1)) # These are just illustrative
RiskW<-gbm(Y~X1+X2+X3, data=TrainW, n.trees = 500,
            interaction.depth = 5, bag.fraction = .5,
           weight=CostWeights, distribution = "bernoulli")
best.iter<-gbm.perf(RiskW, method = "OOB")
summary(RiskW, n.trees = best.iter)  # Predictor importance

## Get White and Black confusion tables -- use prop.table() to percentage
YhatW<-predict(RiskW,newdata =TestW, type = "response", n.trees = best.iter)
Confusion<-table(TestW[,1],YhatW > .5) # For Whites
YhatB<-predict(RiskW,newdata = TestB, type = "response", n.trees = best.iter)
Confusion<-table(TestB[,1],YhatB > .5) # For Blacks

## Get propensity score weights with Whites = 1 Blacks = 0, no weights
Race<-gbm(Race~X1+X2+X3, data=Test, n.trees = 500,
                  interaction.depth = 3, bag.fraction=.5,
                  distribution = "bernoulli")
            best.iter<-gbm.perf(Race, method = "OOB")
            summary(Race,n.trees = best.iter)
YhatRace<-predict(Race,newdata = TestB, type = "response",n.trees = best.iter)
WeightsB<-YhatRace/(1-YhatRace)

## Get propensity score weighted confusion table for Blacks
library(questionr)
WeightedB<-wtd.table(TestB[,1],YhatB>.5,weights=WeightsB,normwt=T)

## Compute conformal distributions and quantile thresholds
library(Hmisc)
ConformalW<-(TestW[,1]-YhatW) # White Conformal scores
quantile(ConformalW,probs=c(.05,.95)) # 1-alpha=.90
ConformalB<-(TestB[,1])-YhatB  # BlackConformal scores
quantile(ConformalB,probs=c(.05,.95)) # 1-alpha=.90 unweighted
wtd.quantile(ConformalB,weights=WeightsB,probs=c(.05,.95)) # weighted

## A hypothetical forecasted case
Case<-data.frame("X1"=10,"X2"=30,"X3"=25) # New Case
Casehat<-predict(RiskW,newdata=Case,type = "response",n.trees = best.iter)
Conformal1<-(1-Casehat) # Conformal score if Y = 1.
Conformal2<-(0-Casehat) # Conformal score if Y = 0.
```

Figure 4: Illustrative R Code For Split Sample Conformal Prediction Intervals

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.

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. (2020a) *Statistical Learning from a Regression Perspective*, Third Edition, Springer.

Berk, R.A. (2020b) "Artificial Intelligence, Predictive Policing, and Risk Assessment for Law Enforcement." *Annual Review of Criminology*, in press.

Berk, R.A., Heirdari, H., Jabbari, S., Kearns, M., & Roth, A. (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 CriminalJustice Risk Assessment." *Criminology and Public Policy* 2020: 1 – 28.

Berk, R. A., and S.B. Sorenson (2016) "Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions. *Journal of Empirical Legal Studies* 13 1: 95 – 115.

Bühlmann, P. and T. Hothorn (2007) "Boosting Algorithms: Regularization, Prediction and Model Fitting." *Statistical Science* 22 (4): 477 – 505.

Buja, A., Berk, R., Brown, L., George, E., Pitkin, E., Traskin, M., Zhan, K., and L. Zhao (2019a). "Models as Approximations – Part I: A Conspiracy of Nonlinearity and Random Regressors Against Classical Inference in Regression." *Statistical Science* 34(4): 523 – 544.

Buja, A., Berk, R., Brown, L., George, E., Arun Kumar Kuchibhotla, and L. Zhao (2019b). "Models as Approximations – Part II: A General Theory of Model-Robust Regression." *Statistical Science* 34(4): 545 – 565.

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

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].

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.

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

Freedman, D.A. (2009) *Statistical Models* Cambridge University Press.

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

Gauraha, N., and Spjuth (2018) "conformalClassification: A Conformal Prediction R Package for Classification." arXiv:1804.05494v1 [stat.ML].

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.

Gupta, C., Kuchibhotla, A.K., and A.K. Ramdas (2020) "Nested Conformal Prediction and Quantile Out-of-Bag Ensemble Methods." arXIV:1910.51v2 [stat.ME].

Guan, L., and R. Tibshirani (2019) "Prediction and Outlier Detection in Classification Problems." arXiv:1905.004396 [stat: ME]

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

Hardt, M., Price, E., Srebro, N. (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).

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

Imbens, D.W. and D.B. Rubin (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* Cambridge University 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.

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

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

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

Kleinberg, J., Mullainathan, S., & 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 Yu, H. (2017) "Accountable Algorithms." *University of Pennsylvania Law Review* 165 (3): 633 – 705.

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

Lavrakas, P. J. (2008). *Encyclopedia of Survey Research Methods* (Vols. 1-0). Thousand Oaks, CA: Sage Publications.

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

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R.J., and L. Wasserman (2018) "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113 (523): 1094-1111.

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

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]

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.

Oaxaca, R.L. and M.R. Random (1999) "Identification in Detailed Wage Decompositions." *The Review of Economics and Statistics* 81(1): 154 –157.

Rice, J.A. (1995) *Mathematical Statistics and Data Analysis*, second edition. New York: Duxbury Press.

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]

Rosenbaum, P. R., and D.B. Rubin (1983) "The Central Role of the Propensity Score in Observational Studies for Causal Effects." *Biometrika* 70 (1): 41 – 55.

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

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

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

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

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." *TheAnnals of Statistics* 37: 1566 – 1590.

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.

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.