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

Richard Berk
University of Pennsylvania

Arun Kumar Kuchibhotla
Carnegie Mellon University

August 27, 2020

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 adopt a framework from 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 measures of fairness that are effectively free 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 application 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 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 ; Conformal Prediction Set

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 propose another fix for unfairness. Because of its simplicity and apparent effectiveness, there is substantial promise for real criminal justice applications. Unlike most other works, 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 provide, which, in turn, solves problems that the earlier work cannot. Using the foundation of conformal prediction sets (Vovk et al., 2005; 2009; Lei et al., 2018), we offer a statistical justification for risk algorithms that treat 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 forecasting so that there is far better balance between the groups. Valid statistical inference can follow without a reliance on asymptotics. The procedures are illustrated with a sample of 300,000 offenders at arraignment. A didactic discussion of the statistical details is provided in Appendix A.

2 Conceptual Framework and Methods

There are many kind of fairness whose definitions and properties have been thoroughly discussed in the recent literature. For comparisons between legally protected groups, we will focus 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.

Because actual decisions are categorical, 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 parole). There will be no important loss of generality.

A bit more formally, the binary response variable Y has two outcome classes, often coded as 1 or 0. Each forecasted outcome class can be used to characterize risk. The forecasted outcome is a function of a set of covariates $\mathbf{X}$ that may be numeric or categorical (e.g., the number of prior arrests, gender), commonly written as $Y|\mathbf{X}$. Some fitting algorithm such as neural networks is use to obtain $\hat{Y}|\mathbf{X}$. Forecasts may be obtained for new cases that have the same set of predictors $\mathbf{X}$. One uses the estimated structure of $\hat{Y}|\mathbf{X}$ with the same set of predictors and their new predictor values to get new values for $\hat{Y}$.[1]

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 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. Note that one conditions on the actual outcome.

- *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. One conditions on the forecasted outcome and determines the proportion of times the forecast is wrong.

- *Cost Ratio parity* – Are the relative costs of false positives and false negatives the same for each group? The cost ratio determines the way

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

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.[2]

2.2 Developing a Fair Risk Algorithm

Fairness depends on the performance of a risk algorithm and the data used to train it. Some argue that machine 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 addressed apply as well to other groups in a variety of criminal justice settings.

Some of the approaches we take 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. The second depends on the first, but employs rather different statistical tools and tackles rather different practical challenges.

2.2.1 Training the Risk Algorithm on White Offenders

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 $\hat{Y}_{\mathrm{W}}^{\mathrm{Train}} = \hat{f}(\mathbf{X}_{\mathrm{W}}^{\mathrm{Train}})$. Inserting test data into the fitting $\hat{f}$, separate fitted values for Whites and Blacks respectively are $\hat{Y}_{\mathrm{W}}^{\mathrm{Test}} = \hat{f}(\mathbf{X}_{\mathrm{W}}^{\mathrm{Test}})$ and $\hat{Y}_{\mathrm{B}}^{\mathrm{Test}} = \hat{f}(\mathbf{X}_{\mathrm{B}}^{\mathrm{Test}})$. The function is *not* re-estimated

[2] 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 will commits a murder will be seen by many stakeholders as far more costly than failing to accurately identify a prison inmate who after release will become 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.

with the test data; it is fixed after it is estimated with the White training data.

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.*[3]

2.2.2 Adjusting for a Covariate Shift

With the algorithm absolved from blame, one can target the test data to obtain a second potential correction. Despite training the the risk algorithm only on Whites, some forms of unfairness may remain because Black and White offenders can have different predictor distributions. 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, for instance, a larger number or prior arrests than White citizens.

In addition, there are concerns that Black individuals are at greater risk of an 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 the cause, is carried forward by the data used for training risk algorithms. For example, underage Black citizens may be at greater risk of being charged as adults. Unfair risk assessments can be the result.

Should the joint predictor distribution for Black offenders differ from

[3] We assume that consistent with common understandings, 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 could be made worse off, and there would no be gains for Black offenders. This is not likely to be a popular policy option.

the joint predictor distribution for White offenders, one has an example of a “covariate shift” (Tibshirani et al., 2020). In response, 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 by training the algorithm only on Whites, as described in Section 2.2.1. Note that the goal is to make the two joint predictor distributions comparable, not just the predictor means (cf., Oaxaca and Ransom, 1999). Features of predictors beyond means can be related to unfairness.

The adjustment we implement is much like the methods that weight predictor distributions by propensity scores to improve causal inference in observational studies (Imbens and Rubin, 2015: section 12.4.2). We are seeking a form of covariate balance. The “treatment” is the race of the offender, and here one estimates for each case the probability that the offender is White. That probability, transformed into an odds ratio, is used to weight the joint predictor distribution for Black offenders to make it more like the joint predictor distributions for White offenders.

The propensity score adjustment can be very effective, as we show later. It also provides a formal justification for training a risk algorithm only on one of the two protected groups. As discussed next, our approach includes computing valid uncertainty estimates for individual forecasts – the probability that a forecasted outcome class for a given offender is the true outcome class – which requires that the conditional distribution of the response (e.g., arrested or not) $\mathrm{P}(Y|\mathbf{X})$ is the same for both protected groups. Only the predictor distribution $\mathrm{P}(\mathbf{X})$ can differ (Tibshirani et al., 2020: equation 6). By using only Whites to train the risk algorithm, this requirement must hold. In other words, by training the risk algorithm only on whites, we can, as a technical matter, properly proceed.

2.3 Constructing Fair Conformal Prediction Sets

Two applications of propensity score weighting need to be distinguished. The more familiar 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. Insofar as the adjustment succeeds, one can argue that the risk algorithm is producing fair results overall.

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 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).[4]

The test statistic is a conformal score, sometimes called a "nonconformity measure." Loosely speaking, it measures for a given case (e.g., an inmate) 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 $\hat{P}_i(Y = y|\mathbf{X} = \mathbf{x})$.[5] 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 $0 - .3 = . - 3$.

For the test data, the outcome class is known. What does one do about 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.[6]

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

[4] 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.

[5] There are many ways to construct conformal scores (Gupta et al., 2020). The properties of these different methods are an active research area. The conformal score we have used should perform well in our risk assessment setting because we are interested in $\mathrm{P}(Y|\mathbf{X})$.

[6] 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.

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.[7]

- 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 *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 95 out of 100 such forecasts. For the third bullet, 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 of the common estimators of $Y|\mathbf{X}$ that can produce fitted probabilities: logistic regression, neural networks, gradient boosting and more. Moreover, there is no requirement that any such risk probabilities are the true risk probabilities. In modeling parlance, the

[7] 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 in working 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.

fitting procedure's mean function can be (and usually is) misspecified. One must be aware, therefore, that the conformal approach is valid in finite samples, *given* the performance of any algorithm. With a different algorithm, different predictors or different training data, there could be different, but still statistically valid conclusions.[8]

The one required 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. But given the white-trained algorithm, conformal scores can be computed for different groups, and from then on, different joint predictor distributions can cause unfairness. We return to this issue shortly when the data for the empirical application are discussed.

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 score weights. That will make quantiles computed for Black offenders more like the quantiles computed for White offenders.

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 are only 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.

[8] 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.

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.)

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.[9] 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 included 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. 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, and gender. 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. For the analyses to follow, 21 predictors were included.

Consistent with our earlier discussion, the 300,000 cases were randomly

[9] 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.

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 data were used as test data ($N = 150,000$). Racial splits of the training and test data were determined by the numbers of Black offenders and White offenders. Each racial split had at least 40,000 observations. Asymptotic requirements were of no 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 will be arrested for a crime of violence was taken to be 8 times worse than failing to correctly classify an offender who after release will not be 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. The tuning defaults worked satisfactorily except that we chose to construct somewhat more complex fitted values than the defaults allowed.[10] 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.[11]

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.[12] Resampling confidence intervals could have been provided for the fairness measures described earlier (Berk, 2020a), but with so

[10] 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.

[11] 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.

[12] 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.

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 positives, predictions of violence in Table 1 are dominated by false positives. This follows directly and necessarily from the imposed tradeoffs. Releasing violent offenders is 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 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.[13]

Table 2 is the confusion table constructed from the test data for Black offenders using the White-trained boosting algorithm. Tables 1 and 2 are similar. No dramatic fairness concerns surface when the proportions in 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,

[13] 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.

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	

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 adjustments to compensate for a covariate shift, we take aim directly at the joint predictor distributions for Black and White offenders.

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.

Each predictor was chosen as one of the top three based on the size of its contribution to the boosting fit of post-release violence.[14] As a group,

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

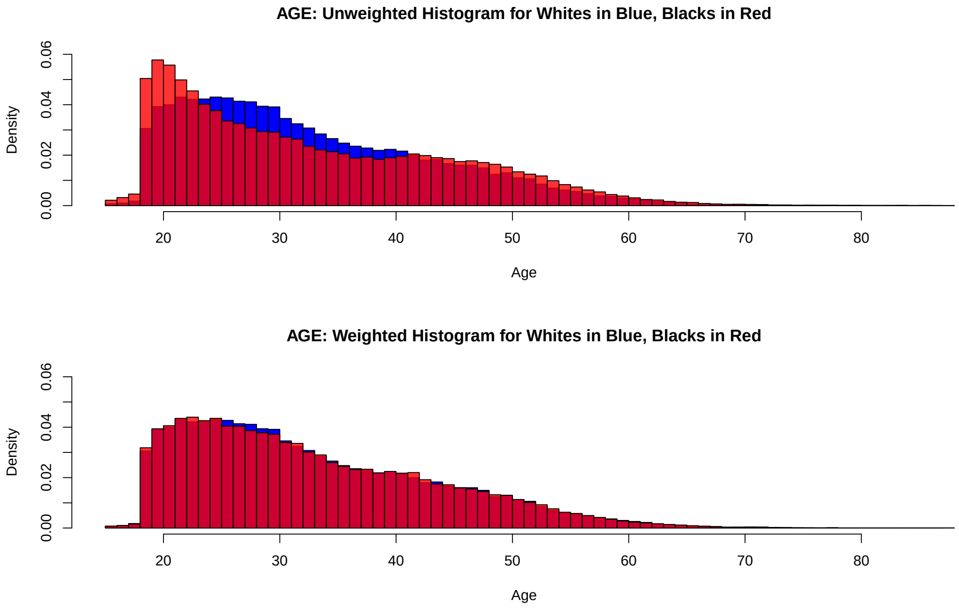


Figure 1: Unweighted and Weighted Histograms for Age

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 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.[15]

Overall, the unweighted histograms are similar except for young offenders, where Blacks are relatively more common than Whites. This difference

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

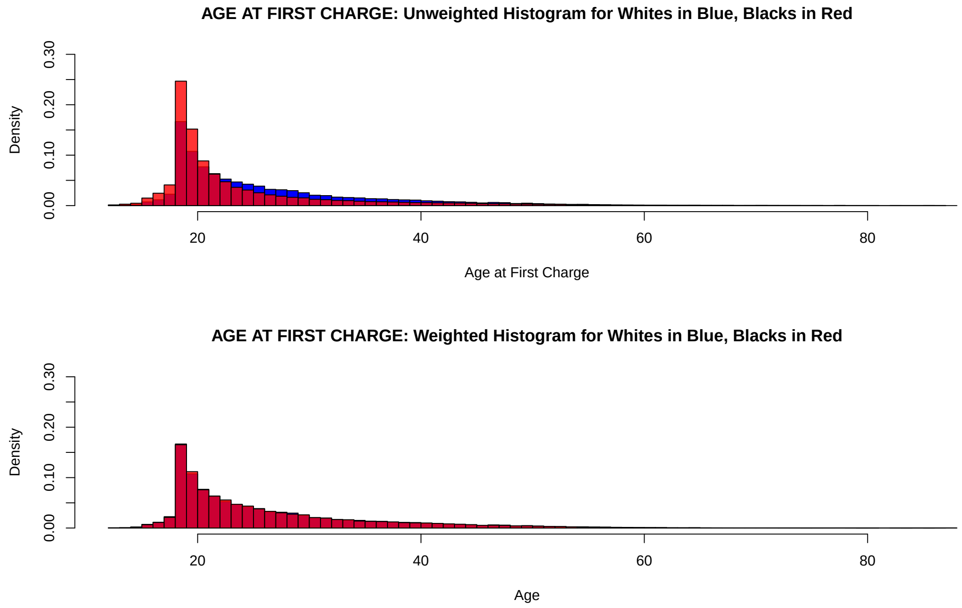


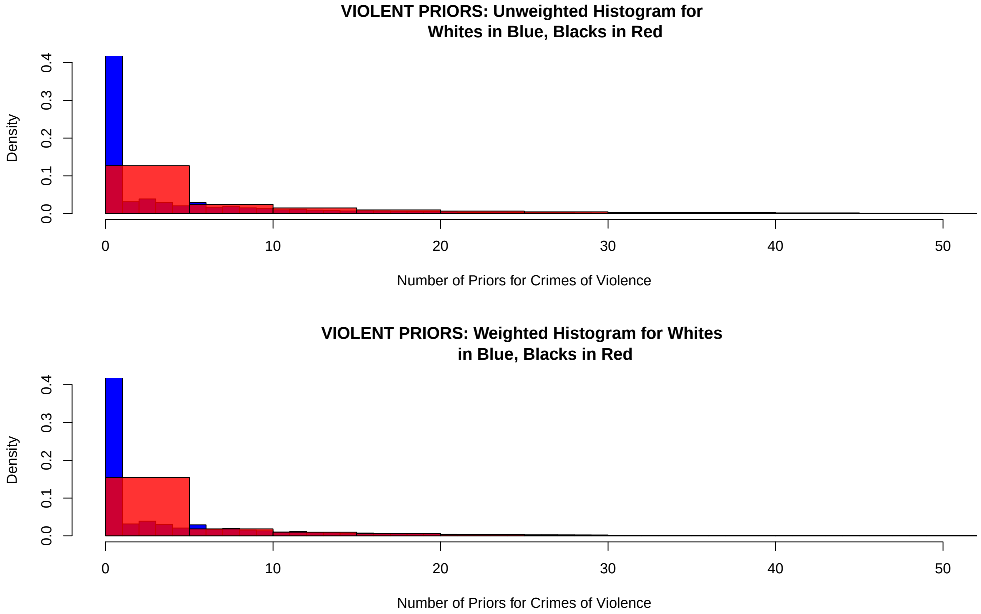
Figure 2: Unweighted and Weighted Histograms for Age at First Charge

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 to be more alike, the differences between the two distributions virtually disappear.

Figure 2, shows the results for the predictor age at first 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 oranger 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-released arrest for a violent crime (as one might well expect), it is not an influential



VIOLENT PRIORS: Unweighted Histogram for Whites in Blue, Blacks in Red

VIOLENT PRIORS: Weighted Histogram for Whites in Blue, Blacks in Red

Figure 3: Unweighted and Weighted Histograms for Violent Priors

predictor of race when the propensity scores are 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 adjustment for a particular covariate requires that the covariate be an influential predictor when the shift weights are computed. If a predictor is effectively unrelated to race, it cannot alter a racial predictor distribution.

We can now return to the confusion tables. 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. We applied the R procedure *wtd.table* in the library *questionr*. 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 and Table 1 are almost identical and for both, 29% of the offenders 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.[16]

[16] Nevertheless, they may argue that there is too much overprediction for *all* offenders.

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	

At the same time, it is difficult to anticipate how well an adjustment for a covariate shift will 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 the same 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 in confusion tables is to try it.

5 Results at the Case Level Using Weighted Conformal Prediction Sets

We have so far considered fairness in confusion tables only. Such tables provide aggregate fairness measures for the performance of risk algorithms. From a policy perspective, aggregate performance is an appropriate yardstick from 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, may want to know about the accuracy of their particular risk forecast and whether racial differences are present. For that, we turn to weighted conformal prediction sets.

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 infer-

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 increasing the possibility of "underprediction." With underprediction, there could be an increase in the number of offenders released who pose a serious treat to public safety.

ential outcomes. Two inferential outcomes specify a true class with a probability of .95, one inferential outcome cannot determine which come class is the true class, and one inferential outcome treats the case as a possible outlier. Which outcome materializes depends on whether the conformal score for a forecasted outcome class falls within the 95% conformal prediction set.

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 two prediction sets are virtually identical. There is no evidence that differences in joint predictor distributions for Blacks and Whites mattered. We will proceed with the weighted results in support fairness, even if just in principle.

This 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 that are caused by disparities in their joint predictor distributions often will not matter. 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. 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.

Table 4 shows the forecasted outcome class 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. One cannot 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.

For these data and these analyses, often there will be no statistically definitive prediction at the level of *individual* cases. One important reason is the large number of false positives caused by the preferred cost ratio. If one could settle for a prediction set that was more narrow, more statistically definitive forecasts would be likely but with less certainty. For example, a 90% prediction set would be more narrow. Consequently, it would be more difficult for both outcomes to have conformal scores falling inside the conformal prediction set. But when a single, forecasted outcome class fell

Table 4: Forecasted Class 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	Forecasted Class
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

inside the prediction set, it would be true outcome class with a probability of .90 rather than .95. Over the long run, claims that the true outcome class had be identified would right only 90% of the time.

6 Conclusions

Unfairness can be introduced by the risk assessment methods. For machine 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 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; 2009b). The estimation target is an acknowledged approximation of the truth. Algorithms can also be seen as estimators of truth approximations. 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 were generated by human activities having largely the same underlying properties as probability sampling from a single, very large population. This 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 courtroom and not others. "Assume-and-proceed" statistics is not a solution.

There also needs to be an appreciation of how the risk procedures will be used. At the aggregate level, confusion tables can be one effective way to address accuracy and fairness for a population of interest. Stakeholders and policymakers can then decide whether performance goals are met sufficiently. But it is also important to properly evaluate performance on a case-by-case basis and in particular, to examine the quality of a given risk forecast. In this paper, we have provided ways to improve fairness at both the aggregate and individual level that rest on sound statistical foundations. We know of no research on risk assessment that simultaneously addresses fairness at both levels of analysis.

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 are practical complications when our procedures produce inconclusive results. Certain cost ratios can cause low forecasting accuracy in the aggregate for certain outcome classes because of an excess of false positives or false negatives. Also, at the individual level, two or more outcome classes may fall in a conformal prediction set or no outcome classes may fall in a conformal prediction set. Under either circumstance, an honest appraisal is that there is limited guidance. A reasonable response is to rely far more heavily on other information. Ideally, a decision that might follow from a risk assessment could be delayed until more particulars for a problematic case are collected (Berk and Sorenson, 2016; Madras et al., 2018a).

Finally, the conformal approach to statistical inference requires that the conditional distribution of risk, given the available covariates, is the same for all relevant protected groups. Training a risk algorithm on a single group formally solves this problem, and if properly framed, absolves the algorithm itself from any charges of 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.

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.[17]

[17] 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.[18] 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.

[18] 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.[19]

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

[19] 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 $P(Y|\mathbf{X})$ is the same, $P(\mathbf{X}_A) \neq \tilde{P}(\mathbf{X}_B)$. 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 $P(Y|\mathbf{X})$ 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."

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. (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.

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).

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.

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.

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.

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.