The Annals of Applied Statistics
2019, Vol. 13, No. 1, 189–220
https://doi.org/10.1214/18-AOAS1201

AN ALGORITHM FOR REMOVING SENSITIVE INFORMATION: APPLICATION TO RACE-INDEPENDENT RECIDIVISM PREDICTION[1]

BY JAMES E. JOHNDROW AND KRISTIAN LUM

Stanford University and Human Rights Data Analysis Group

Predictive modeling is increasingly being employed to assist human decision-makers. One purported advantage of replacing or augmenting human judgment with computer models in high stakes settings—such as sentencing, hiring, policing, college admissions, and parole decisions—is the perceived "neutrality" of computers. It is argued that because computer models do not hold personal prejudice, the predictions they produce will be equally free from prejudice. There is growing recognition that employing algorithms does not remove the potential for bias, and can even amplify it if the training data were generated by a process that is itself biased. In this paper, we provide a probabilistic notion of algorithmic bias. We propose a method to eliminate bias from predictive models by removing all information regarding protected variables from the data to which the models will ultimately be trained. Unlike previous work in this area, our procedure accommodates data on any measurement scale. Motivated by models currently in use in the criminal justice system that inform decisions on pre-trial release and parole, we apply our proposed method to a dataset on the criminal histories of individuals at the time of sentencing to produce "race-neutral" predictions of re-arrest. In the process, we demonstrate that a common approach to creating "race-neutral" models—omitting race as a covariate—still results in racially disparate predictions. We then demonstrate that the application of our proposed method to these data removes racial disparities from predictions with minimal impact on predictive accuracy.

1. Introduction. Statistical and machine learning models are increasingly used to inform high-stakes decisions, including hiring [Hoffman, Kahn and Li (2015)], credit scoring [Khandani, Kim and Lo (2010)], and throughout all stages of the criminal justice system. In the criminal justice context, predictive models of individuals' future behavior are used to inform judges regarding pre-trial release and bail setting, sentencing, and parole [Brennan, Dieterich and Ehret (2009), Phillips, Ferri and Caligiure (2016), Dieterich, Mendoza and Brennan (2016), Cunningham and Sorensen (2006), Dvoskin and Heilbrun (2001), Quinsey et al. (2006), Berk et al. (2009)]. For example, there is an increasing reliance on predictive models to inform judges about the likelihood that a defendant will re-offend

Received March 2017; revised May 2018.

[1] Supported by the Human Rights Data Analysis Group.

Key words and phrases. Algorithmic fairness, criminal justice, neutral predictions, racial bias, recidivism, risk assessment, selection bias.

if released. In this case, data about individual defendants is used to train a model with the objective of predicting future re-offense. The model's prediction for a given defendant is then shown to the judge or parole board presiding over that person's case for use in informing pre-trial release, parole, or sentencing decisions. Given the importance of decisions regarding an individual's personal liberty, it is imperative that any input to the decision-making process—be it a model's prediction or otherwise—be "fair" with respect to legally or socially protected classes such as race, gender, or sexual orientation.

In this paper, we focus on recidivism prediction. The objective is to make predictions regarding an individual's future likelihood of re-offense that are "fair" with respect to that individual's race.[2] Typically, post-release re-offense is measured by re-arrest, and there are many reasons to believe that re-arrest may be a biased measure of re-offense with respect to race. For example, studies suggest that after controlling for criminal behavior, African Americans are more likely than Caucasians to become incarcerated [Bridges and Crutchfield (1988)], and whether walking or driving, African Americans are disproportionately stopped and searched by police [Simoiu, Corbett-Davies and Goel (2016), Rudovsky (2001)]. For drug crimes, African American drug users are arrested at a rate that is several times that of Caucasian drug users despite the fact that African American and Caucasian populations are estimated by public health researchers to use drugs at roughly the same rate [Langan (1995), Mitchell and Caudy (2015)]. Thus, fitting models to data for which certain groups are *observed* committing crime at a disproportionate rate unfairly biases the model's predictions against those groups. In this setting, a model may very successfully predict that an individual is likely to be re-arrested, but the model has no way of disentangling a higher risk of re-arrest due to elevated latent criminality from higher risk due to increased police attention.

If information were available at the local level on the *true rate* of committing crime by race or other protected characteristic, one could introduce corrections (e.g., sampling weights or protected characteristic-specific latent thresholds) into the prediction model to correct for the relative over-sampling of protected demographic groups. However, for a variety of reasons, information on true baseline rates of crime against which to "benchmark" the police data is notoriously difficult to obtain, and different methods for estimating the base rate of committing crime often come to opposite conclusions [Glaser (2014), Alpert, Smith and Dunham (2004)].

As there is a noted dearth of information with which to correct systematic measurement error in arrest data, we adopt a standard of fairness which requires that predictions be independent of protected characteristics. This is sometimes referred to as demographic parity or statistical parity [Dwork et al. (2012)]. Our view is

[2] We use the terminology for racial categorization that is used in the dataset that is the focus of our application. The categories in this dataset are defined as African American, Caucasian, Hispanic, Asian, Native American, and Other.

that absent reliable unbiased samples or benchmarks to correct for sampling bias in arrest data, the most reasonable approach is to assume that the average level of criminality is independent of race. In doing so, we note both the unique circumstances of systematically biased sampling in this setting as well as the fact that independence or demographic parity is one of the set of standard notions of fairness considered in the machine learning literature.

It is worth noting that some authors have—quite reasonably—argued for the insufficiency of demographic parity as a notion of fairness. Most notably, Dwork et al. (2012) outlines how demographic parity fails as a reasonable definition of fairness through three example scenarios outside of the context of criminal risk assessment. Under the first example, Dwork et al. (2012) demonstrate that demographic parity can be achieved even in settings where the relationship between a permitted covariate and the outcome of interest is quite different (or reversed) among protected groups. This example is envisioned in the setting where the modeler does not include or does not have access to information about the protected variable. For example, if full-time employment is positively correlated with re-arrest for white individuals but negatively correlated with re-arrest for black individuals, a model that includes employment status (but not race) may still achieve demographic parity without fully accounting for the group-wise heterogeneous effect. In extreme cases, this can lead to a model that results in predictions that are negatively correlated with the outcome for some groups. In less extreme cases, this can lead to predictions that are much more accurate for one group than the other to the point of rendering the prediction model useless for some groups. In recidivism prediction, while there likely are interactions between the protected variable and the permitted variables, because the permitted covariates are often selected to be well established, theoretically motivated risk factors, we find this scenario unlikely. Furthermore, standard validation techniques to assess the quality of the model would reveal if this were the case, so long as the high-level evaluators had access to the protected variables. For example, assessing accuracy by race—which we do as part of our evaluation—would flag this as a problem. Both other examples that point out the insufficiency of demographic parity in Dwork et al. (2012) center around the risk that bad faith modelers intentionally create predictive models that disadvantage a protected group by explicitly incorporating information about the protected variable into the model building process. Under our proposed framework, all model builders, including those acting in bad faith, would be provided a dataset that is independent of race on which to train their model. Under this strategy, such an individual would not know or be able to infer the protected status of each individual. Our proposed framework would, therefore, preclude the intentional sabotage of individuals based on protected status by design.

Even in the context of risk assessment, some authors who work in this area have argued for the impropriety of demographic parity. For example, in a paper elucidating the relationship between several notions of fairness, Chouldechova (2017) rejects demographic parity for consideration in building recidivism models while suggesting its utility for other topic areas, such as employment or educational admissions settings. However, in the conclusion the author notes that "Throughout this paper we have implicitly operated under the assumption that the observed recidivism outcome Y is a suitable outcome measure for the purpose of assessing the fairness properties of an RPI (recidivism prediction instrument)." The author then cautions that many reported crimes are never "cleared" (i.e., never result in an arrest) and if there is bias in the re-arrest of individuals who re-offend, the evaluation of demographic parity and other fairness notions carried out in the paper may be misleading. This scenario in which there is significant racial bias in re-arrest is specifically that which we seek to address in this paper, and thus the objections of Chouldechova (2017) to demographic parity are not directly relevant to our application. Ultimately, there is significant disagreement among researchers about which notions of fairness are appropriate for which settings. Even if one does not agree with demographic parity for our application, our work can be viewed as furthering an existing area of methodological research and a case study of the effects of applying this standard of fairness in the recidivism risk assessment context.

There has been substantial work to date to create statistical or machine learning models that produce predictions that achieve demographic parity [Kamiran and Calders (2009), Calders and Verwer (2010), Feldman et al. (2015), Adler et al. (2016), Romei and Ruggieri (2014)]. Much of the previous work has primarily focused on settings where the outcome or protected variable is binary. Existing methods, such as propensity score weighting, though not explicitly designed for this task may also prove to be powerful tools to achieve demographic parity, particularly in settings where the protected variable is categorical. The approach we suggest primarily builds on the work of Feldman et al. (2015). These authors propose a procedure to transform a set of covariates $X = (X_1, \dots, X_p)$ to a new set of covariates W such that each variable W_j in W is independent of a categorical protected variable Z. In the special case where $p = 1$, this is sufficient to guarantee that any algorithm trained on W will produce predictions that are independent of Z. The authors give empirical justification for the algorithm when $p > 1$. In addition to not guaranteeing demographic parity when $p > 1$, the algorithm has several limitations. First, it works only for categorical Z, and reasonable performance requires a large number of observations taking each possible level of Z. Although Adler et al. (2016) makes some improvements to the handling of categorical variables, the procedure is not appropriate for continuous Z or Z with many distinct values. Finally, the procedure requires that all of the covariates X_j be continuous variables, further limiting its scope.

To address the described limitations, we approach the problem from a likelihood-based perspective. In this framework, transforming noncontinuous variables as well as transforming variables for which there is little data corresponding to one or more levels of Z are natural. This framework also allows us to make transformed data W that are mutually independent of Z, rather than only pairwise independent, thereby guaranteeing that the predictions of any algorithm trained on the transformed data will be independent of Z. To do so, we define the problem in terms of a chain of conditional models, as is commonly used in multiple imputation [see White, Royston and Wood (2011) for an overview]. Each variable is adjusted by matching its estimated quantile (conditional on the protected variable and all other previously adjusted variables) to the marginal quantiles for that variable. Our approach allows for any number of mixed-scale variables to be adjusted and naturally accommodates one or more protected variables that may be categorical or continuous. This greatly expands the range of datasets that can be transformed and thus expands the universe of problems to which the procedure may be applied.

We apply our method to a dataset pertaining to the criminal justice system in Broward County, Florida. This dataset contains several covariates describing an individual's demographic characteristics and criminal history. The outcome variable of interest is re-arrest within two years of release, a likely biased measure of relapse into criminal behavior. We apply our procedure to render the permitted covariates independent of race, and use both logistic regression and random forest to predict re-arrest. We find that while models fit to the unadjusted data omitting race produce drastically different predictive distributions of the probability of re-offense by race—thus empirically demonstrating the insufficiency of omitting race from the analysis when the goal is statistical parity—equivalent models fit to the data transformed using our procedure produce nearly identical predictive distributions by race. Further, the predictive accuracy of our method decreases only slightly due to the adjustment, a phenomenon we explore in depth. We also find that random forest or logistic regression fit to only seven transformed variables—mostly pertaining to an individual's criminal history—has substantively equivalent predictive power to proprietary models used for recidivism prediction that use a battery of psychological questionnaires and evaluations in addition to information about the individual's criminal past.

2. Setup.

We begin by specifying notational conventions that will be used throughout. In general, random variables are denoted by capital letters and realizations of random variables by lower case. For example, $B = f(A)$ defines a random variable B that is defined as a function of another random variable A, whereas $b = f(a)$ refers to the function f applied to a data point a, giving the value b. When defining functions, we will use lower case letters to specify arguments. For example, $F_{X|Z} : \mathbb{R}^p \times \mathbb{R}^q \to [0, 1]$ defined by $F_{X|Z}(x, z) = \mathbb{P}[X \leq x \mid Z = z]$ specifies a conditional distribution function, while $F_{X|Z}(X, Z)$ is a random variable taking values in [0, 1] defined by applying the transformation $F_{X|Z}$ to (X, Z).

Now, suppose we have a response Y and predictors (Z, X), where Z represent protected characteristics. We take X, Z to be p and q dimensional random vectors with arbitrary measurement scale. Consider a generic prediction rule or model for Y given by

$$\delta : X \to \widehat{Y}. \tag{2.1}$$

Our goal is not to use any information about Z in predicting Y; that is, we want a *fair prediction rule*.[3]

DEFINITION 2.1 (Fair prediction rule). A prediction of the form (2.1) is *fair* with respect to the protected characteristics Z if and only if

$$\widehat{Y} \perp\!\!\!\perp Z. \tag{2.2}$$

In other words, we seek to achieve demographic parity with respect to the protected variables Z. Although f is not a function of Z in (2.1), this is insufficient to guarantee $\widehat{Y} \perp\!\!\!\perp Z$ unless $X \perp\!\!\!\perp Z$. In the overwhelming majority of applications, X and Z are dependent, and thus we must take additional measures to ensure $\widehat{Y}$ is fair.

There is a simple condition that does guarantee fairness. Since $\hat{Y}$ is not a function of Z, we already have $\widehat{Y} \perp\!\!\!\perp Z \mid X$. Since functions of independent random variables are independent, $X \perp\!\!\!\perp Z$ implies $\widehat{Y} \perp\!\!\!\perp Z$. Thus, we seek to define a new random variable W that is independent of Z, while still preserving as much "information" in X as possible. W will then be used in lieu of X to build a model, $\delta : W \to \hat{Y}$, that is guaranteed to be independent of Z.

3. Transformations to independence. Consider random variables (X, Z) with joint distribution $F(x, z)$, and let $\mathbf{X}$ be the collection of all random variables of dimension p. Define $\mathbf{W} \subseteq \mathbf{X}$ to be the set of all random variables of dimension p that are independent of Z and have distribution $G(x)$. For a fixed distance metric $\rho : \mathbf{X} \times \mathbf{X} \to \mathbb{R}^+$ between probability distributions, the goal of our procedure is to find $W \in \mathbf{W}$ that satisfies

$$\rho(X, W) = \inf_{W' \in \mathbf{W}} \rho(X, W'). \tag{3.1}$$

This may or may not be unique, but all such random variables are, informally, as close as possible to X while also being independent of Z. Throughout, we will take ρ to be the Wasserstein-α distance for $\alpha \geq 1$.

The procedure we propose is guaranteed to achieve (3.1) when X is univariate. When X is multivariate, we chain optimal univariate transformation to create a procedure that achieves mutual independence of W from Z. Though the multivariate procedure does not have the optimality guarantee of the univariate procedure, empirically we find it is very successful at achieving independence with minimal distortion of the distribution of X. In the following motivating sections, we assume

[3] We emphasize that the term "fair" is used here in a mathematical context as a shorthand for the independence condition in (2.2). Ultimately, it is up to policymakers and ethicists to determine whether this condition is appropriate in any particular context. However, we argue it is the most appropriate of the existing notions of algorithmic fairness to our motivating application for the reasons outlined in the Introduction.

the joint and conditional distributions of (X, Z) are known, and the desired target distribution, G, is fixed and user-specified. We defer discussion of how to handle unknown F and how to choose G to the end of this section, as the following results hold for any F and G. A more rigorous presentation of the problem and discussion of the theoretical underpinnings of this transformation, including results on the optimality of our proposed procedure, are given in Appendix B.

3.1. *Univariate transformations.* Let $p = 1$ and $F_{X|Z}(x, z)$ be the conditional cumulative distribution function of X given Z. If X is continuous, then the transformation $W = \zeta(X, Z) = G^{\leftarrow}(F_{X|Z}(X, Z))$ defines the optimal transformation of X to W such that W is independent of Z, $W \sim G$, and, informally, W is as similar as possible to X given the independence and distributional constraints, and $G^{\leftarrow}(u) = \inf\{x \in \mathbb{R} : u \leq G(x)\}$ denotes the left-continuous inverse CDF of the target distribution.

For atomic X, the transformation is slightly more complicated. Let $\dot{x} = \{\dot{x}_1, \dot{x}_2, \dots\}$ be the support points of $F_{X|Z}$ ordered such that $\dot{x}_j < \dot{x}_{j+1}$, with associated probabilities $\pi_j = \mathbb{P}[X = \dot{x}_j]$, and define $\nu_j = \sum_{j' \leq j} \pi_j$ with $\nu_0 = 0$. Then, by corollary B.1 the stochastic map $W = \zeta^*(X, Z)$ for $\zeta^*(X, Z) \mid X = \dot{x}_j \sim \text{Uniform}(\nu_{j-1}, \nu_j)$, achieves (3.1). This immediately gives an algorithm for transforming realizations of a univariate X to realizations of W such that $W \perp\!\!\!\perp Z$ with minimal information loss, given in Algorithm 1.

3.1.1. *Continuous example.* Consider the joint model defined by the conditional distributions

$$Z \sim \text{Bern}(0.5),$$

$$X_1 \mid Z \sim N(Z + 4, 1).$$

The goal is to apply the transformation we have defined above to create a new variable, $W_1 \sim G_1$ that is independent of Z and retains as much dependence with the original variable, X_1, as possible. In this example, $F_{X_1|Z}(X_1, Z) = \Phi(X_1 - 4 - Z)$ where $\Phi(\cdot)$ is the standard normal distribution function. Therefore, we define the transformed variable as $W_1 = G_1^{\leftarrow}(\Phi(X_1 - 4 - Z))$. Under this transformation, $W_1 \mid Z \sim G_1$ and, because the distribution is the same for every value of Z, $W_1 \perp\!\!\!\perp Z$, so $W_1 \sim G$ marginally. However, for G_1 any normal distribution, $\text{cor}(W_1, X_1) = 2/\sqrt{5} \approx 0.89$, demonstrating the high level of remaining dependence between X_1 and W_1, as desired.

This transformation is depicted in the left panel of Figure 1 for $G_1(w) = \Phi(w - 4.5)$. Suppose we observe $x_1 = 6$ and $z = 1$. To compute the inner function, we first calculate the conditional quantile, $F_{X_1|Z}(x_1, z) = 0.84$, which is indicated by the height of the dashed curve at $x_1 = 6$. To complete the transformation, we evaluate the inverse distribution function for the target distribution at the previously calculated conditional quantile. That is, we take $G_1^{\leftarrow}(0.84) = 5.5$. So, in doing this transformation, we map the original random variable to the value such that

Algorithm 1: Univariate transformations of variables

```
Data: $\{x_i\}$ and $\{z_{ij}\}$ for $i = 1, \dots, n$ and $j = 1, \dots, q$, where
      $X \mid Z \sim F_{X|Z}(x, z)$, target distribution $G$
Result: $\{w_i\}$ for $i = 1, \dots, n$, where $w_i$ are realization of $W \sim G$ with
      $W \perp\!\!\!\perp Z$.
for $i$ = 1,..., $n$ do
    if $X$ is atomic then
        set $x_i^- = \max\{\dot{x}_k : \dot{x}_k < x_i, k = 1, 2, \dots\}$
        if $x_i^- = \varnothing$ then
            set $x_i^- = -\infty$
        end
        set $\ell(x_i) = F_{X|Z}(x_i^-, z_i)$; $r(x_i) = F_{X,Z}(x_i, z_i)$ where
        sample $u_i = \mathrm{Uniform}(\ell(x_i), r(x_i))$
        set $w_i = G^{\leftarrow}(u_i)$
    end
    if $X$ is continuous then
        set $w_i = G^{\leftarrow}(F_{X|Z}(x_i, z_i))$
    end
end
```

the quantile of the original variable conditional on the protected variable matches the quantile of the transformed variable marginally. The transformed variable can be viewed as a re-scaled measure of how large the original measurement is, after accounting for the value of z.

3.1.2. *Atomic example.* Consider the joint model defined by the alternative set of conditional distributions:

$$Z \sim \mathrm{Bernoulli}(1/2),$$

$$X_2 \mid Z \sim \mathrm{Bernoulli}(p) \qquad \text{where } p = \frac{1}{3} + \frac{1}{3}Z.$$

Under our proposed stochastic transformation, we define $U \sim \mathrm{Uniform}(a, b)$ where $a = F_{X_2|Z}(X_2 - 1, Z)$ and $b = F_{X_2|Z}(X_2, Z)$. More specifically, if $X_2 = 1$, we set $a = (2 - Z)/3$ and $b = 1$. If $X_2 = 0$, we set $a = 0$ and $b = (2 - Z)/3$. The transformation is completed by setting $W_2 = G_2^{\leftarrow}(U)$. Despite achieving $W_2 \perp\!\!\!\perp Z$, dependence between W_2 and X_2 is retained. For example, if we select $G_2 = \mathrm{Bernoulli}(1/2)$, then $\mathrm{cor}(W_2, X_2) = 2/3$.

This transformation is depicted in the right panel of Figure 1. In this example, we observe $x_2 = 1$, so we sample $u \sim \mathrm{Uniform}(1/3, 1)$. Assume we draw $u = 0.4$. This is then transformed to the 0.4 quantile of G_2, which in this case is zero. If

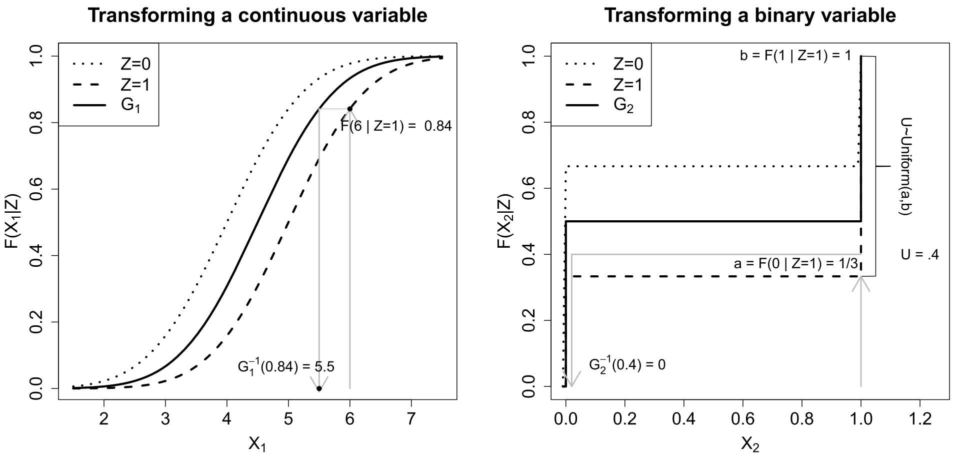

FIG. 1. (*left*) *An example transformation of* $x_1 = 6$ *to* $w_1 = 5.5$ *when* $z = 1$. (*right*) *An example transformation of* $x_2 = 1$ *to* $w_2 = 0$ *when* $z = 1$.

the uniform random draw had been, for example, 0.6, $w_2 = G^{\leftarrow}(0.6)$ would have been one.

3.2. *Multivariate adjustments via chaining*. In the previous section, we demonstrated how to transform a univariate $X \to W$ to obtain independence with Z. This procedure could be applied independently to each covariate X_j to achieve pairwise independence with Z, though it is not guaranteed that the resulting set of independently transformed covariates will achieve mutual independence with Z. Regardless, pairwise adjustments may be desirable to retain interpretability of the covariates. Under a pairwise adjustment, each covariate could be interpreted as a simple Z-adjusted version of itself. For example, if X_j is the number of prior arrests and Z is race, a pairwise adjustment would result in a race-adjusted measure of the number of prior arrests. This section is concerned with extending the above results to achieve mutual independence with Z.

Now, let X be a random vector. We propose to construct an analogous multivariate transformation function $\zeta(\cdot)$ as

$$\begin{aligned} W = \zeta(X, Z) &= \big(\zeta_1(X_1, V^{(1)}), \zeta_2(X_2, V^{(2)}), \dots, \zeta_p(X_p, V^{(p)})\big) \\ &= (W_1, W_2, \dots, W_p), \end{aligned} \tag{3.2}$$

where $V^{(j)} := (Z, W_{1:(j-1)})$ for $j > 1$, $V^{(1)} := Z$, and each $\zeta_j(X_j, V^{(j)})$ is defined as in Section 3.1 such that

$$\zeta_j(X_j, V^{(j)}) \perp\!\!\!\perp (Z, W_{1:(j-1)}).$$

The ordering $X_1, \dots, X_p$ of the X variables is arbitrary, though some orderings may be practically convenient for a given application.

Using basic rules of conditional probability, $p(W \mid Z)$ can be decomposed as

$$p(W \mid Z) = \prod_j p(W_j \mid Z, W_{1:(j-1)}) = \prod_j p\big(\zeta_j(X_j, V^{(j)}) \mid Z, W_{1:(j-1)}\big).$$

Because $\zeta_j(X_j, V^{(j)}) \perp\!\!\!\perp Z, W_{1:(j-1)}$, each element of the product on the right-hand side of the equation can be replaced by $p(\zeta_j(X_j, V^{(j)}) \mid Z, W_{1:(j-1)}) = p(\zeta_j(X_j, V^{(j)})) = p(W_j)$, and the joint distribution reduces to

$$p(W \mid Z) = \prod_j p(W_j),$$

and W is mutually independent of Z. Consequently, we refer to (3.2) as a *transformation to mutual independence*.

3.2.1. *Multivariate chaining toy example.* Now, consider the joint distribution defined by the following conditional specification:

$$\begin{aligned} Z &\sim \mathrm{Bern}(0.5), \\ X_1 \mid Z &\sim N(Z+4, 1), \\ X_2 \mid X_1, Z &\sim \mathrm{Bern}(p), \\ p &= I^{\leftarrow}\big(\Phi(X_1 - Z - 4), 1 + Z, 2 - Z\big), \end{aligned} \tag{3.3}$$

where $I^{\leftarrow}(x, a, b)$ is the inverse cumulative distribution function of a beta distribution with parameters a and b. Under this specification, the marginal distributions of (X_1, Z) and (X_2, Z) are the same as those defined in examples in Sections 3.1.1 and 3.1.2, respectively.

To perform the chained multivariate adjustment to create (W_1, W_2) that are mutually independent of Z, we define the following set of transformations. Put $W_1 = \zeta_1(X_1, Z) = G_1^{\leftarrow}(F_{X_1 \mid Z}(X_1, Z))$—the same as in Section 3.1.1. In the example in Section 3.1.2, we defined $W_2 = G_2^{\leftarrow}(F_{X_2 \mid Z}(X_2, Z))$ to create an adjusted variable independent only of Z. Instead, to create a W_2 such that (W_1, W_2) are mutually independent of Z, we now define $W_2 = G_2^{\leftarrow}(F_{X_2 \mid V^{(2)}}(X_2, V^{(2)}))$. $F_{X_2 \mid V^{(2)}}$ is the conditional distribution function of X_2 given W_1 and Z and, in this case, is not available in closed form.

3.3. *Defining G.* In practice, how one chooses or estimates G is less critical than using the conditional CDF to transform X. It is typical in applied statistics and regression modeling to transform predictors prior to model fitting for computational reasons or to obtain better predictive accuracy, which would neutralize any choice we make for G. On balance, we suggest taking G to be the marginal distribution F_X. This ensures that researchers using the transformed data still have access to the original marginal distribution of the data, which may be of significant value in its own right.

3.4. *Unknown conditional distributions.* In the above, we have assumed that $F_{X|Z}$ and G are known. In practice, the conditional distributions, e.g. $F_{X|Z}$, are typically unknown and must be estimated from the data. In the example we present below, we have found traditional, parametric regression models to be successful at estimating $F_{X|Z}$ if the analyst employs appropriate model selection and fit diagnostic techniques. A more automated approach would likely require more exotic nonparametric models to effectively model the conditional distributions without human input. The key to this approach is reliably estimating the ζ_j, which in turn requires good choices of $\widehat{G}^{\leftarrow}$ and good estimators of $\widehat{F}_{X_j|V^{(j)}}$. Estimation of $F_{X_j|V^{(j)}}$ is an exercise in regression modeling. When Z is low dimensional, it may be appropriate to obtain the conditional through a nonparametric estimate of the joint distribution of Z and X_j. For the particular case of the recidivism data, likelihood-based parametric regression models were more successful. Ultimately, the better the estimator of the conditional distribution $\widehat{F}_{X_j|V^{(j)}}$, the closer to fair any prediction rule $\widehat{y}$ estimated on w, so it is critical to construct these estimators with care.

3.5. *Practical implementation.* In practice, we envision the following workflow. A training data set is pre-processed using the procedure we have so far described. A set of adjusted data sets are released to model developers, who have no access to individual-level information about the protected variable(s). Once a suitable predictive model is selected by the model developers, the parameters of the model would be transmitted to the group administering the risk assessment so that predictions could be made about new people as they are arrested. The set of transformation functions developed in the first stage of the process must also be available to the administrators of the risk assessment tool (but not the model developers), as the covariates for new arrestees would have to undergo the same transformation before being used in the predictive model.

4. Simulation example. In order to illustrate our proposed method, we simulate from the joint model defined in Section 3.2.1, and define the dependent variable as $Y \sim \mathrm{N}(X_1 - X_2 + X_1 Z + X_1 X_2 + X_1 X_2 Z, 1)$. We simulate one realization from this model with sample size $n = 10{,}000$. Realizations from this model are denoted by lowercase variables, z, y, x_1, and x_2.

A density plot of y given z is shown in Figure 2. The goal of our procedure is to produce w_1 and w_2 that are independent of z. Independence manifests as for any prediction rule $\hat{y} = \delta(w_1, w_2)$, the distribution of $\hat{y}$ for $z = 0$ is equal to the distribution of $\hat{y}$ when $z = 1$.

We compare predictions of y given w by comparing three methods for producing w: (1) do no adjustment of x (so $w = x$); (2) perform pairwise transformations to independence to produce w; and (3) perform transformations to mutual independence to produce w. In all cases, we omit z from the set of covariates used to

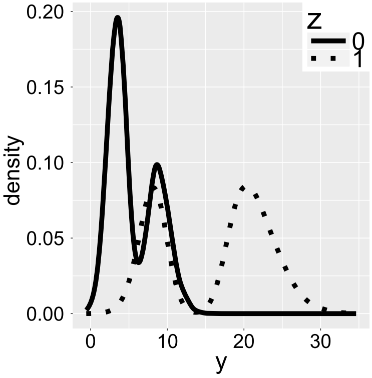

FIG. 2. *A comparison of the distributions of y for each value of z in the simulated data example.*

fit y. In all cases, unlike in the illustrative toy examples, we assume all relevant distributions are unknown and must be estimated from the data, mimicking a real analysis in which the joint distribution of the data is not known.

For procedure (2) in which we perform two independent univariate adjustments, we must estimate the functions $F_{X_1|Z}$ and $F_{X_2|Z}$. To estimate $F_{X_1|Z}$, we fit a linear regression with Gaussian errors of x_1 on z, so that $\widehat{F}_{X_1|Z} = \Phi((x_1 - \hat{x}_1)/\hat{\sigma}^2)$, where $\hat{\sigma}^2$ is the estimated sample variance of $x_1 \mid z$ and $\hat{x}_1$ is the fitted value of $x_1 \mid z$ under our estimated model. We set $\widehat{G}_1^{\leftarrow}$ to be the empirical quantile function of x_1. Then, for each observation $i = 1, \dots, n$, $w_{i1} = \widehat{G}_1^{\leftarrow}(\Phi(x_{i1} - \hat{x}_{i1})/\hat{\sigma}^2))$.

To estimate $F_{X_2|Z}$, we first fit a logistic regression of x_2 on z, resulting in fitted values $\hat{x}_2$. Under this fitted model, $\widehat{F}_{X_2|Z}$ is the CDF of a Bernoulli distribution with parameter $\hat{x}_2$. We then sample $u_i \sim \text{Uniform}(\ell_i, r_i)$ for $\ell_i = \widehat{F}_{X_2|Z}(x_{i2} - 1, \hat{x}_{i2})$ and $r_i = \widehat{F}_{X_2|Z}(x_{i2}, \hat{x}_{i2})$. We again set $\widehat{G}_2^{\leftarrow}$ to be the empirical quantile function of x_2, and $w_{i2} = \widehat{G}_2^{\leftarrow}(u_i)$. This procedure for producing w_1 and w_2 is referred to as "adjusted-pairwise".

For procedure 3, we jointly adjust $x_1, x_2 \to w_1, w_2$. To do this, we make two transformations, the first by transforming X_1 to be independent of Z, exactly as detailed above in procedure (2). To complete the second transformation, we estimate the conditional distribution of X_2 given W_1 and Z by fitting a logistic regression of x_2 on z and w_1, yielding fitted values $\hat{x}_2$. Then $\widehat{F}_{X_2|W_1,Z}$ is given by the CDF of a Bernoulli distribution with parameter $\hat{x}_2$. Similar to the above, we make a stochastic transformation by sampling $u_i \sim \text{Uniform}(\ell_i, r_i)$ with $\ell_i = \widehat{F}_{X_2|Z,W_1}(x_{i2} - 1 \mid \hat{x}_{i2})$ and $r_i = \widehat{F}_{X_2|Z,W_1}(x_{i2} \mid \hat{x}_{i2})$. We again use the empirical quantile function of X_2 as $\widehat{G}_2^{\leftarrow}$ to obtain $w_{2i} = \widehat{G}_2^{\leftarrow}(u_i)$. We refer to this procedure as "adjusted".

Figure 3 shows the empirical distribution of the fitted $\hat{y}$ under each of the three adjustment procedures. Fitted values are calculated as $\hat{y} = w\hat{\beta}$ where $\hat{\beta}$ is the least squares estimate from regression of y on w. From this it is clear from the unadjusted model that omitting z does little to equalize the distributions of $\hat{y}$ conditional

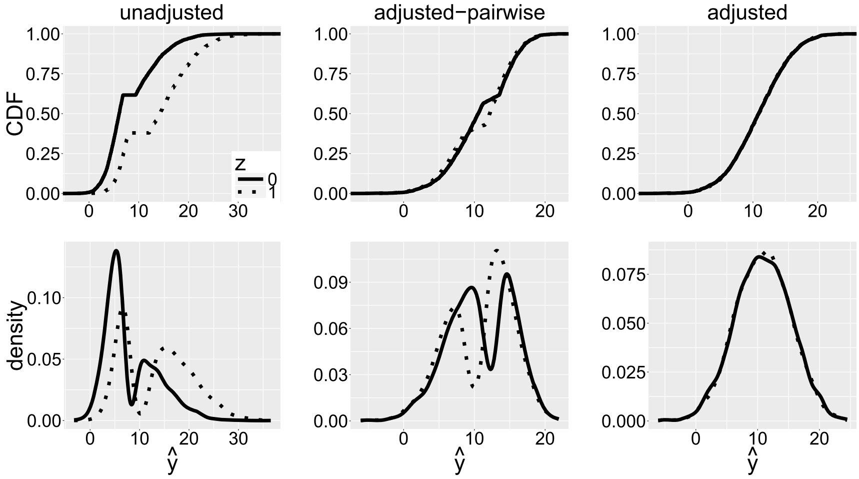

FIG. 3. *The distributions of $\hat{y} \mid z$ under each of the adjustment procedures.*

on z. The pairwise adjustment does reduce some of the discrepancy in the predictive distributions of $\hat{y} \mid z$, but the distribution of $\hat{y} \mid z = 1$ has a much larger right mode than the distribution of $\hat{y} \mid z = 0$. In contrast, the joint adjustment results in nearly indistinguishable predictive distributions, achieving the goal of the procedure.

5. Application: Removing racial dependence in recidivism risk assessment. In 2016, ProPublica released a news article on the use of predictive analytics in recidivism risk assessment [Angwin et al. (2016)]. The focus of the the investigation was on whether risk assessment tools were disproportionately recommending nonrelease for African American defendants. The reporters compiled an extensive dataset from the criminal justice system in Broward County, Florida, combining detailed individual-level criminal histories with predictions from a popular risk assessment tool, COMPAS. COMPAS (an acronym for Correctional Offender Management Profiling for Alternative Sanctions) is a proprietary software tool developed by Northpointe, Inc. that predicts a defendant's likelihood of failing to appear in court, re-offending, and violently re-offending. In order to produce the predictions, a proprietary algorithm is fit to several covariates, including a battery of psychological questions administered at the time of arrest. For each type of prediction, COMPAS produces a decile score (a score between one and ten, in which larger numbers indicate a higher predicted probability of future offense) and a categorical score consisting of three categories—"low", "medium", and "high" risk. In order to assess the accuracy of the recidivism predictions, the ProPublica researchers compared each person's COMPAS prediction to an indicator of whether they had been re-arrested within two years of release.

ProPublica's over-arching conclusion was that the COMPAS tool was "racially biased" based on the observation that of those who were not re-arrested 45% of African Americans were mis-classified by the model as future recidivists, where as only 24% of Caucasian defendants were similarly misclassified [Angwin et al. (2016)]. In a rebuttal, Northpointe asserted that the disparities in the proportion of false positives was entirely due to differing baseline rates of recidivism between African American and Caucasian defendants. They argued that bias should be assessed not in terms of the false positive rate, but rather, in terms of the group-wise positive predictive value or overall predictive accuracy. Using the same data used in ProPublica's analysis, Northpointe and others showed that the predictive accuracy of their model was equivalent for African American and Caucasian defendants [Dieterich, Mendoza and Brennan (2016), Flores, Bechtel and Lowenkamp (2016)]. The disagreement about the interpretation of the predictions of the tool comes down to the fact that the two sides were using different definitions of fairness, both of which cannot simultaneously be achieved when the marginal rate of re-offense differs between groups [Chouldechova (2017), Kleinberg, Mullainathan and Raghavan (2017)].

Ultimately, both definitions assume that re-offense is fairly measured by re-arrest. As we argued in the Introduction, because African Americans are more likely to be re-arrested for re-offending, neither definition seems particularly appropriate for this setting. We note that our understanding of the bias in re-arrest as a measure of re-offense does not come from the observation that the marginal rates of re-arrest differ by group in this dataset, but rather from information outside of this particular dataset—studies of racial disparities in arrests, as discussed in the Introduction. Indeed, differences in the marginal rate of arrest could manifest via processes that are entirely unbiased—most obviously, genuine differences in rates of re-offense, though others are possible. Because re-offense is not observed directly, it is very difficult to know the extent to which the observation of re-offense is biased. Thus, as we argue above, a reasonable way to proceed is to assume that the distribution of risk is independent of race. To this end, we implement the procedure described above to remove all information about race from the covariates we will use for prediction, thus guaranteeing similar distributions of estimated risk by race.

5.1. *Data.* For each defendent in the time period, ProPublica collected several measures of criminal history: the number of misdemeanor, felony, and other charges accrued as a juvenile (denoted respectively by juv_misd_count, juv_fel_count, juv_other_count); the number of adult prior offenses (prior_count); the defendant's sex (sex); and age at the time of the crime (age). These are the covariates that make up x. The dataset also includes the race of the defendant (race), which is our protected variable, z. The response, y, is an indicator of whether the defendant was re-arrested within two years of release. Using this data, the objective is to construct a new dataset w that contains no information about z so that any prediction rule of the form (2.1) applied to the data set will satisfy (2.2).

5.2. *Dependence between race and other covariates in recidivism data.* We begin by assessing dependence between z and x in the data to determine whether transformations to independence are likely to have a meaningful effect. We test for pairwise dependence by discretizing continuous or count variables and summarizing data on pairs of variables in a two-way contingency table. We then compute the G statistic, $M(x_1, x_2) = 2n \sum_{c_1=1}^{d_1} \sum_{c_2=1}^{d_2} \widehat{\pi}_{c_1 c_2} \log[\widehat{\pi}_{c_1 c_2}/(\widehat{\pi}_{c_1 \cdot} \widehat{\pi}_{\cdot c_2})]$, where d_1 and d_2 are the number of unique values of variables x_1, x_2, and $\widehat{\pi}_{c_1 c_2} = n^{-1} \sum_i \mathbb{1}\{x_{i1} = c_1, x_{i2} = c_2\}$, $\widehat{\pi}_{c_1 \cdot} = \sum_{c_2=1}^{d_2} \widehat{\pi}_{c_1 c_2}$, and $\widehat{\pi}_{\cdot c_2} = \sum_{c_1=1}^{d_1} \widehat{\pi}_{c_1 c_2}$ are the empirical cell probabilities of the contingency table. M is a scaled sample estimate of the mutual information between the joint distribution of the discretized variables and the product of their marginal distributions. The G test is in fact a likelihood ratio test of the null hypothesis $H_0 : x_1 \perp\!\!\!\perp x_2$ under the multinomial likelihood, and the test statistic has asymptotically a χ^2 distribution with $(d_1 - 1)(d_2 - 1)$ degrees of freedom under the null hypothesis of independence.

We compute the G statistic for all pairs of variables (z, x_j) consisting of z and one component of x. The p-values of the tests—computed using the asymptotic distribution of the test statistic—are shown in Table 1. There is strong evidence to reject the null hypothesis of independence for all of the pairs, even when adjusting for multiplicity using the method of Benjamini and Hochberg (1995). This indicates that a prediction rule $\delta : x \to \widehat{y}$ is unlikely to be fair for race, and that to guarantee a fair prediction rule we need to estimate and apply a transformation to independence. Put another way, a model which simply excludes race is unlikely to result in fair predictions, as the effect of race will be encapsulated in the estimated effects of each of the variables included in the model.

5.3. *Transformations to independence.* We now estimate maps of the form (3.2) for each x_j in the recidivism data. We first develop conditional density estimates $\widehat{F}_{X_j | V^{(j)}}$ for each x_j. Of the six x_j, one (sex) is binary, one [log(age)—henceforth simply "age"] is continuous, and the other four, which relate to prior

TABLE 1

p values for G tests of the null hypothesis of pairwise independence between race and the indicated variable, either unadjusted for multiplicity, or adjusted using the method of Benjamini and Hochberg (BH)

	Unadjusted	**BH**
sex	8.84E−08	1.06E−07
juv_fel_count	8.00E−21	1.20E−20
juv_misd_count	1.91E−21	3.82E−21
juv_other_count	2.72E−07	2.72E−07
priors_count	6.73E−58	4.04E−57
log(age)	7.22E−49	2.17E−48

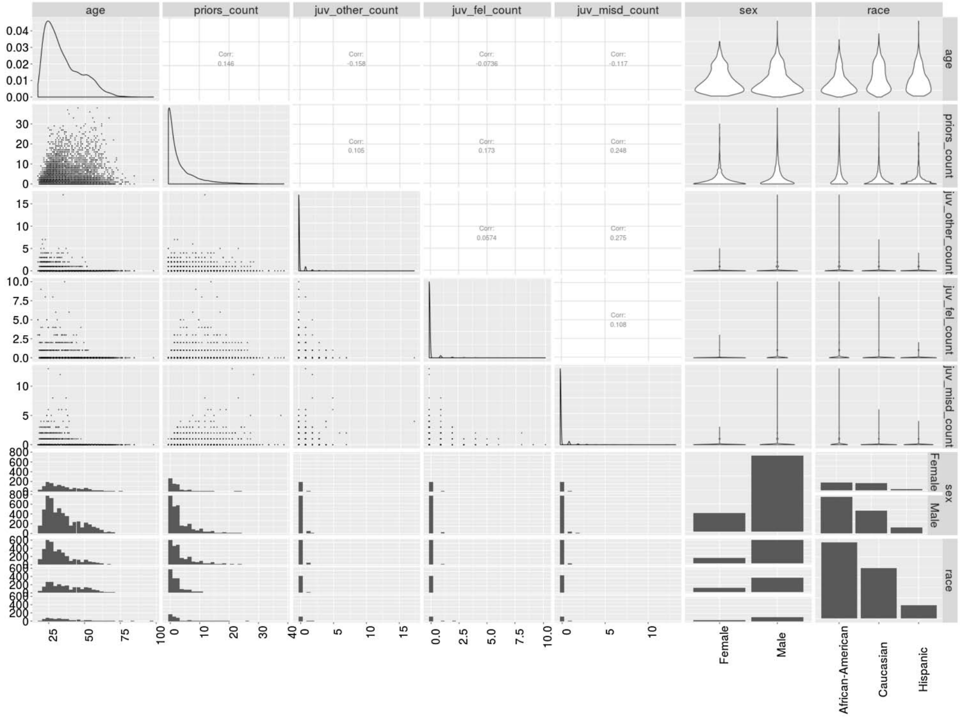

FIG. 4. *Visualizations of marginal and pairwise conditional distributions of covariates. Row and column labels indicate which variables are compared. The diagonal shows marginal distributions*; *upper and lower triangles of the plot matrix show visualizations of the conditional distribution of row variables given column variables.*

criminal record, are counts. A pair plot, showing visualizations of the pairwise joint distributions of each of the covariates and race, is shown in Figure 4. The criminal record variables—juv_misd_count, juv_fel_count, juv_other_count, and priors_count—are highly dispersed counts, and there is evidence of substantial dependence between most pairs of variables.

In constructing the sequence of conditional models, it makes sense to estimate $\widehat{F}_{X|V^{(j)}}$ for the covariates with more complicated marginal distributions first, which facilitates estimation of richer models. Based on Figure 4, we order the variables as: $X_1 =$ age, $X_2 =$ prior_count, $X_3 =$ juv_other_count, $X_4 =$ juv_fel_count, $X_5 =$ juv_misd_count, and $X_6 =$ sex. The protected variable is $Z =$ race. We apply the procedure described in Section 3.2 to estimate a transformation of the form (3.2). In every case, we estimate the marginal distribution of x_j, $\widehat{F}_{x_j}$, using the empirical CDF.

In constructing the chain of conditional models, we always include discretized versions w_1^* and w_2^* of w_1 or w_2, respectively, whenever the w_1 or w_2 are included in the model. This captures nonlinearity in the conditional mean of the other variables. The cutpoints used for discretization are: $\{18, 19, 20, \nu(0.1, w_1), \nu(0.2, w_1), \dots, \nu(1, w_1)\}$ for w_1^* and $\{\nu(0.1, w_2), \nu(0.2, w_2), \dots, \nu(1, w_2)\}$ for w_2^*, where $\nu(\theta, x)$ is the θ-empirical quantile of x. In every case, we estimate $\widehat{G}_j^{\leftarrow}$ using the empirical quantile function of x_j, and we use our estimated $\widehat{F}_{X_j|V^{(j)}}$ and $\widehat{G}_j^{\leftarrow}$ to obtain w_j using Algorithm 1.

The transformation is performed as follows.

1. Estimate $\widehat{F}_{X_1|Z}$ using the empirical CDF of x_1 separately for each value of z. Set $w_1 = \widehat{G}_1^{\leftarrow}(\widehat{F}_{X_1|Z}(x_1 \mid z)$.

2. Set $v^{(2)} = (z, w_1, w_1^*)$ and estimate $\widehat{F}_{x_2|v^{(2)}}$ by zero-inflated negative binomial regression of x_2 on $v^{(2)}$. Set $w_2 = \widehat{G}_2^{\leftarrow}(\widehat{F}_{X_2|V^{(2)}}(x_2 \mid v^{(2)}))$.

3. Set $v^{(3)} = (z, w_1, w_1^*, w_2, w_2^*)$ and estimate $\widehat{F}_{X_3|V^{(3)}}$ using a zero-inflated negative binomial regression of x_3 on $v^{(3)}$. Set $w_3 = \widehat{G}_3^{\leftarrow}(\widehat{F}_{X_3|V^{(3)}}(x_3 \mid v^{(3)}))$.

4. Set $v^{(4)} = (z, w_1, w_1^*, w_2, w_2^*, w_3)$. Estimate $\widehat{F}_{X_4|V^{(4)}}$ using a zero-inflated Poisson regression of $v^{(4)}$ on w_4. Set $w_4 = \widehat{G}_4^{\leftarrow}(\widehat{F}_{X_4|V^{(4)}}(x_4 \mid v^{(4)}))$.

5. Set $v^{(5)} = (z, w_1, w_1^*, w_2, w_2^*, w_3, w_4)$. Estimate $\widehat{F}_{x_5|v^{(4)}}$ using zero-inflated Poisson regression of $v^{(5)}$ on x_5. Set $w_5 = \widehat{G}_5^{\leftarrow}(\widehat{F}_{X_5|V^{(5)}}(x_5 \mid v^{(5)}))$.

6. Set $v^{(6)} = (z, w_1, w_1^*, w_2, w_2^*, w_3, w_4, w_5)$. Estimate $\widehat{F}_{x_6|v^{(6)}}$ using logistic regression of $v^{(6)}$ on x_6. Set $w_6 = \widehat{G}_6^{\leftarrow}(\widehat{F}_{X_6|V^{(6)}}(x_6 \mid v^{(6)}))$.

We repeat the above K times and save each of the transformed datasets, $w^{(k)} = \{w_1, w_2, \dots, w_6\}$ for $k = 1, \dots, K$. Each resulting $w^{(k)}$ is stochastic because all of the $\widehat{F}_{X_j|V^{(j)}}$ are discrete. While any w generated in this way is fair with respect to race, individual predictions depend on the sampled values $u(x) \sim \text{Uniform}(a(x), b(x))$ for all of the discrete variables, and interval estimates of parameters will understate uncertainty resulting from the stochastic nature of the maps ζ_j. Consequently, in generating predictive values for individual subjects or estimating uncertainty in model parameters, we use an average over all K fair datasets. This approach of creating multiple datasets is also used in the privacy settings [Reiter (2005)] and multiple imputation [Rubin (2004), Reiter and Raghunathan (2007)], where a common default value is $K = 10$ [Buuren and Groothuis-Oudshoorn (2011)]. In the fairness setting, we have the additional goal of limiting the effect of stochastic synthetic data w on individual predictions, so we use a larger default value of $K = 50$.

If $\widehat{F}_{X_j|V^{(j)}}$ were the exact conditional distribution $F_{X_j|V^{(j)}}$, then W would satisfy $W \perp\!\!\!\perp Z$. Of course, $\widehat{F}_{X_j|V^{(j)}}$ is an estimate, and thus it will differ from $\widehat{F}_{X_j|V^{(j)}}$ in finite samples, and even asymptotically when $\widehat{F}_{X_j|V^{(j)}}$ is misspecified. Therefore, we evaluate model fit for each conditional model separately, and recommend against applying a "black box" or automated approach to constructing the conditionals. We expect that in most applications, the dimension of X will be relatively small, as is the case in our recidivism application, making it practicable to construct each conditional density estimate carefully.

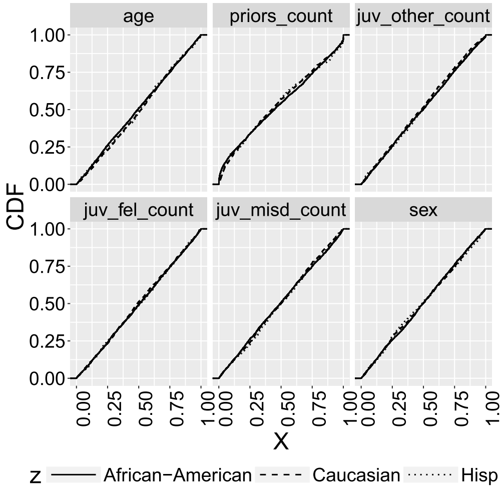

FIG. 5. *Plot of $F_{X_j|V^{(j)}}$ by race for each x_j across* 200 *adjusted datasets.*

Because it is important that the entire conditional distribution is estimated well as opposed to just the conditional expectation, we assess model fit by plotting the fitted conditional CDFs $\widehat{F}_{X_j|V^{(j)}}(x_j, v^{(j)})$ by race. If the fit is good, this should be close to the uniform distribution on the unit interval. Figure 5 gives results for the estimated conditional CDFs by race, all of which are approximately uniform. This is sufficient to guarantee that we have successfully managed to sample w_j as a sample from the marginal distribution of x_j for each race category. However, it is insufficient to guarantee that all information about race has been removed from the transformed dataset, as the model may be badly misspecified, e.g., not enough interaction variables were included in the model. We further analyze the success of the procedure's ability to achieve independence from the protected variable by computing Cramer's V statistics for every pair of variables in the adjusted data, discretized to have 10 unique values (or fewer, if the original variable is discrete with < 10 unique values). Results are shown in Figure 6. Values in the original data are shown for comparison. In most cases, Cramer's V is reduced to near zero in the adjusted data, indicating that we have successfully removed information about race from the adjusted data, at least up to two-way interactions.

5.4. *Predicting recidivism using transformed data.* Using each of the K transformed datasets, we predict re-arrest within two years using random forest (RF). We compare our results to the "unadjusted" model in which all covariates but race are used to explain y. We repeat this analysis using logistic regression in place of RF. The results of the logistic regression analysis are qualitatively very similar to those of RF and are deferred to the Appendix.

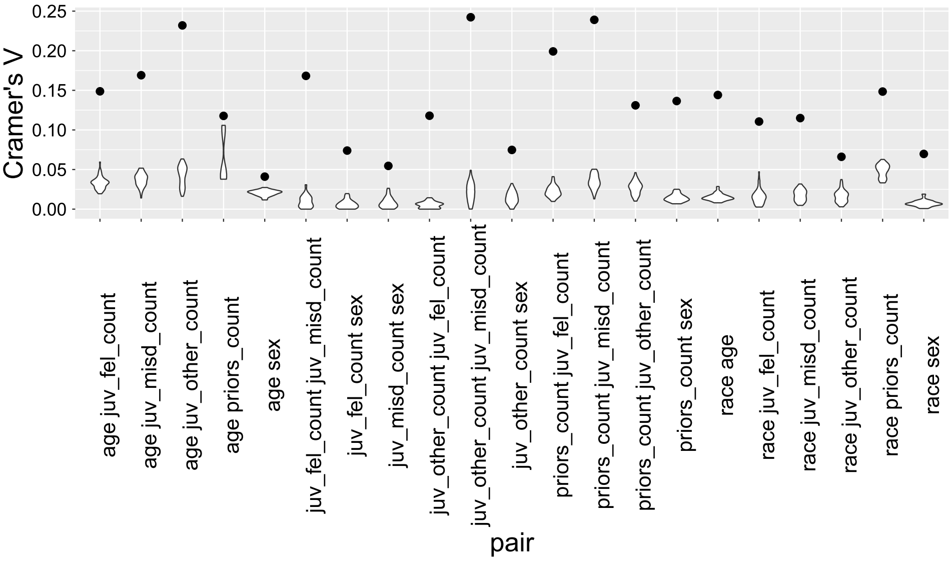

FIG. 6. *Distribution of pairwise Cramer's V for every covariate pair across M transformed datasets for adjusted (white violin plot) and unadjusted datasets (dots).*

Because each variable that we transform is discrete, we also apply Corollary B.1 to create stochastic realizations and similarly produce K pairwise-adjusted datasets. Figure 7 shows the empirical density and CDF of the re-arrest probability for RF trained on data adjusted using our procedure ("adjusted") and trained using the "unadjusted" data. It is clear from the left panels of Figure 7

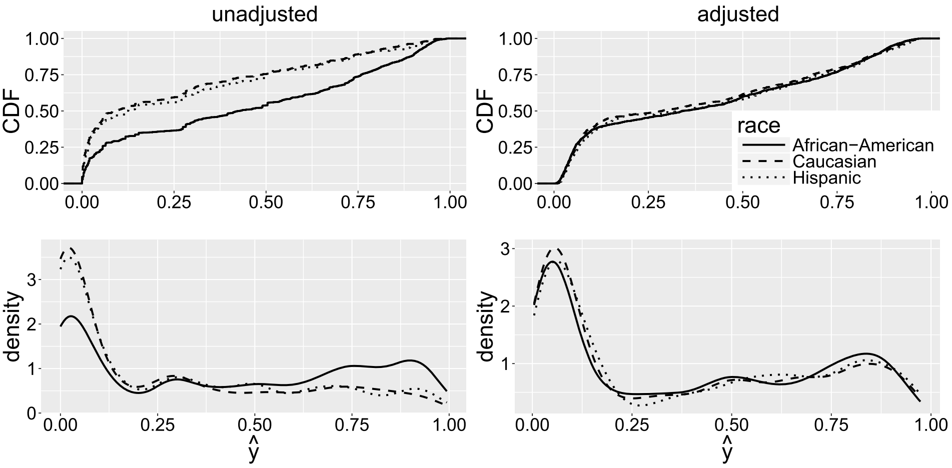

FIG. 7. *Density and CDF of predictions made using random forest by race using adjusted and unadjusted data.*

that when trained on unadjusted data, large differences by race exist in the predictive distribution, with the distribution for African Americans having substantially more mass at probabilities of re-arrest greater than about 0.5. In other words, when trained on unadjusted data omitting race, the model predicts that a larger fraction of the African-American population is at high risk of recidivating than other groups. Predictions made by training RF on data adjusted using our procedure eliminate almost all racial disparities, as evidenced by the nearly identical distributions by race in the two panels on the right.

Having established that predictions made using the transformed data are *fair* under the definition we propose for this context, we now turn to fit assessment. In this case, assessment of how well our model predicts Y using any notion of model performance is not especially well motivated, as Y is a biased measure of the phenomenon it is meant to measure. Nonetheless, we compare how well the predictions from RF fit to the unadjusted, pairwise adjusted, and adjusted datasets perform. In applying our proposed procedure and the pairwise adjustment, some relevant information is lost. Thus, it is expected that the predictive accuracy of a model fit to the adjusted data will be lower than the model trained on unadjusted data. Figure 8 shows the ROC curves for both the predictions from the adjusted, the pairwise adjusted, and unadjusted data. We find that these are not substantially different. For the unadjusted data, the area under the curve (AUC) was 0.72, and for the adjusted data it was 0.71. We note that this AUC is on par with the AUC associated with Northpointe's predictions for this dataset (0.70) as reported in Dieterich, Mendoza and Brennan (2016).

Finally, we compare the various notions of out-of-sample predictive performance across races under the adjusted and unadjusted models using $\hat{p}_i = 0.5$ as a threshold for classification. This is shown in Table 2, which reports accuracy (acc), positive predictive value (ppv), negative predictive value (npv), and false positive rate (fpr) for African Americans, Caucasians, and Hispanics. The positive and negative predictive values exhibit disparities across race using both the

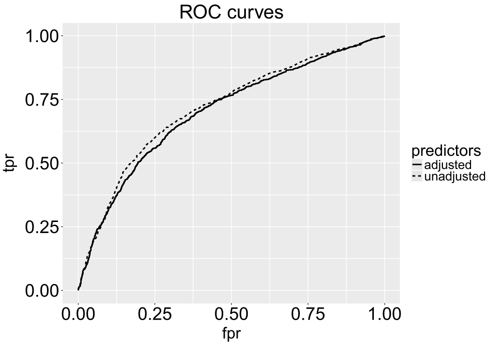

FIG. 8. *ROC for predictions made with random forest using adjusted and unadjusted data.*

TABLE 2

Measures of predictive accuracy for random forest estimated on adjusted and unadjusted datasets

Procedure	**Metric**	**African-American**	**Caucasian**	**Hispanic**	**mad**
Adjusted	ppv	0.73	0.54	0.58	0.06
Unadjusted	ppv	0.71	0.64	0.65	0.02
Adjusted	npv	0.62	0.72	0.79	0.06
Unadjusted	npv	0.64	0.7	0.74	0.03
Adjusted	acc	0.66	0.65	0.7	0.02
Unadjusted	acc	0.67	0.69	0.72	0.02
Adjusted	fpr	0.23	0.28	0.28	0.02
Unadjusted	fpr	0.28	0.14	0.15	0.05

adjusted and unadjusted data, but they are somewhat larger in the adjusted data. To make comparison easy, we show the mean absolute deviation (mad) using the median as the centroid for each metric, that is, mad $= \frac{1}{3}\sum_{i=1}^{3}|x_i - \tilde{x}|$, where $\tilde{x}$ is the median value. This is a measure of how much each performance metric differs across racial group. So, for example, the mad for the first row of Table 2 is $\frac{1}{3}(|0.73 - 0.58| + |0.54 - 0.58| + |0.58 - 0.58|) = 0.06$. The mad increases after adjustment for both ppv and npv. In particular, the positive predictive value is reduced by adjustment for Caucasian and Hispanic people and increased for African-Americans. Conversely, the negative predictive value is reduced by adjustment for African-Americans and increased by adjustment for Caucasians and Hispanics.

On the other hand, adjustment appears to decrease variation by race in false positive rates and only slightly increase variation in accuracy. The mad is unchanged for accuracy, and decreases from 0.05 to 0.02 for fpr. The fpr is decreased for African-Americans and increased for Hispanics and Caucasians. Although compatibility (or lack thereof) between different notions of fairness has not been a focus of this paper, it is interesting that at least in this particular example, mitigating disparate impact actually led to an improvement in the overall similarity of accuracy and false positive rates by race, since variation in false positive rates fell considerably more than variation in accuracy increased after adjustment. Therefore, our procedure need not lead to a deterioration in fairness by all other metrics. This also suggests that optimizing a loss function that incorporates both similarity in false positive rates/accuracy and dependence of the predictive distribution on race may be sensible if deemed socially desirable by policymakers.

5.5. *Implications for risk assessment.* In risk assessment, the predicted probability of recidivism from the fitted model is typically thresholded to produce“high” and “low” risk categories (or potentially more than two risk categories). To illustrate the effect of transformations to independence, we label individuals high risk if their predicted probability of recidivism is greater than 0.5, and otherwise label

TABLE 3

Test set confusion matrices by race with predictions made using random forest on adjusted and unadjusted data

		Unadjusted		**Adjusted**	
Race	**Rearrest**	**Low**	**High**	**Low**	**High**
African-American	no	0.34	0.14	0.37	0.11
African-American	yes	0.20	0.33	0.23	0.29
Caucasian	no	0.53	0.09	0.44	0.18
Caucasian	yes	0.22	0.16	0.17	0.21
Hispanic	no	0.54	0.09	0.46	0.18
Hispanic	yes	0.19	0.18	0.12	0.24

them low risk. We compute confusion matrices containing the empirical probability that individuals re-offend given that they were classified as low or high risk. In addition to giving a sense of the practical impact of the adjustment procedure, this also helps clarify how the overall accuracy can be essentially unchanged by adjustment despite strong dependence between the outcome and race.

For ease of comparison across race, we give proportions rather than raw counts in the confusion matrices, which are shown in Table 3. The accuracy of the predictions are defined as

$$\text{accuracy} = \text{tpp} + \text{tnp},$$

where tpp is defined as true positives divided by total number of observations and tnp is defined as true negatives divided by total number of observations. For African-Americans, the change in tpp of $0.29 - 0.33 = -0.04$ is largely offset by the change in tnp of $0.38 - 0.35 = 0.03$, for a net change of only -0.01 in total accuracy. Opposite trade-offs are observed in the other race categories, with increases in tpp being offset by decreases in tnp for Caucasians and Hispanics. Thus, similar accuracy is achieved by offsetting improvements and degradation of the two metrics, and the direction of the shift differs for African-Americans compared to all others. This is the basic intuition for how roughly equal accuracy is achieved after adjustment.

The confusion matrices also reveal the proportion of individuals that move across the 0.5 threshold by race: the proportion of African-American considered high risk decreases by 0.07, from 0.47 to 0.40; the proportion of Caucasians considered high risk increases by 0.14, from 0.25 to 0.39, and the proportion of Hispanics considered high risk increases by 0.15, from 0.27 to 0.42. It is worth noting that the Hispanic group is by far the smallest of the three, so while this group experiences the largest percent increase in high risk classification, the number of individuals whose status changes as a result of adjustment is relatively small. In any case, the strong dependence of the predicted outcome on race can be viewed as resulting from these differences in rates of re-arrest between the individuals whose status changes after adjustment; these numbers could of course differ at other thresholds. If we accept that African-Americans are more likely to be re-arrested conditional on committing crime, then it seems reasonable to accept this modest degradation in accuracy to obtain race-independent predictions. Of course, one could obtain intermediate outcomes by attenuating rather than fully removing dependence between W and Z. One strategy for achieving attenuated dependence, which may be attractive to practitioners, is to adjust only those variables which are recorded via a subjective determination or a potentially biased process (e.g., number of previous arrests) and leave more objectively derived variables (e.g., age) about which one is sure a "fair" measurement has been made intact.

6. Discussion. We have presented a statistical framework for adjusting a dataset such that models trained to the data will be mutually independent of protected variables. The framework we suggest has extended the existing literature by allowing an arbitrary number of variables of arbitrary type to be both protected and adjusted so long as a suitable conditional model can be found to adequately describe the full conditional distribution of the permitted variables given the protected variables. The extension to allow for the adjustment of discrete variables is itself an advancement, as previous proposals for adjusting of the *training covariates* were only designed for adjusting continuous variables. Our second main contribution is that our method allows the user to make adjustments such that the output dataset is mutually independent of the protected variables, as opposed to pairwise independent. We have tested this procedure on a dataset used for recidivism prediction and demonstrated that, by using a series of chained relatively standard regression models, we are able to produce an adjusted dataset in which all pairs of variables in the dataset are approximately independent. Further, when fitting both random forest and logistic regression models to the data, we have achieved predictive distributions of recidivism that are approximately independent of race—the ultimate goal of the procedure. Even after the adjustment, we observe that the quality of the predictions in terms of overall accuracy like the AUC are on par with methods that are currently in use but do not attempt to adjust for fair predictions. We expect our procedure would also be of value in data privacy and anonymization.

It is often suggested that an equivalent way to accomplish the goal of removing disparate impact would be to simply take the top $p\%$ from each class and designate them as the most risky. However, adjusting the training data has other benefits. By doing the adjustment, a dataset could be released to multiple organizations to build prediction models, and regardless of the details of their model, we would be guaranteed that the predictions would be fair under the definition we support in this case. Model developers could also apply new machine learning or statistical models for which modifications, such as weights, are difficult or computationally costly to introduce without worry that the models they develop would result in race-dependent predictions. Furthermore, this approach limits the risk of malicious actors intentionally biasing the model against protected groups, as model builders would not be able to infer protected characteristics. In addition, if a protected variable is continuous (e.g., protecting parental income in a tool meant to predict success in college for college admissions), simply taking the top $p\%$ within each class is infeasible, as classes would have to be made by discretizing, and people who happened to fall on the lower or upper end of each bin would be unfairly disadvantaged. Lastly, if multiple variables are to be protected, even if each variable has sufficient data in each class, the combination of classes across all variables may not, necessitating an approach like that proposed here.

If deployed in the real world, the approach we have proposed does not obviate the need for good practices with respect to validation and evaluation. We anticipate that this would include checking whether the transformation was successful in removing dependence by performing nonparametric tests for independence or computing summary statistics quantifying dependence on the transformed variables. Here, we have used the G test and Cramér's V for these purposes, but numerous other statistics are available. The nature of dependence among variables is likely to change over time (perhaps, even as a *result* of the application of such a risk assessment model), so it would be important that these tests are routinely applied to ensure that dependence were adequately removed on an ongoing basis. If it were found that the previous algorithm no longer resulted in permissible covariates that were adequately independent of race, the algorithm would need to be retrained and modified to allow for time dynamics in the conditional distribution of covariates given protected variables. Evaluation should also include comparison of the predictive performance of algorithms trained on the untransformed and transformed covariates.

There are several avenues for future work in this area. First, judges are typically the ultimate consumers of predictive risk assessment in criminal justice. Substantial research is necessary to better understand how the presentation format of risk scores affects decision-making by judges. For example, decisions might differ if judges were shown the predicted probability rather than a coarsened measure such as "high risk" versus "low risk". Moreover, if a coarsened risk score is presented, it is likely that the number of categories and the language used to describe each category would affect decisionmaking. In some cases, the highest risk group is in fact more likely than not to remain on good behavior post-release. It is unclear whether judges interpret "high risk" in those cases accurately, or implicitly assume that they are in fact more likely than not to re-offend.

Finally, it is imperative that we engage experts in other fields, as well as the communities most likely to be affected by the model's predictions, to develop mathematical characterizations of fairness that aptly reflect the social or legal meaning of the term. This needs to be done separately for every instantiation of a risk assessment model, as the best mathematical characterization of fairness will likely vary with prediction type and local community understanding of fairness and the goals of implementing the model. While it is important that statisticians and others with related expertise take part in helping those outside our field understand proposed mathematical definition of fairness—for example, independence versus false positive rate versus positive predictive value—ultimately this area of research should be, and is increasingly, undertaken in conjunction with ethicists and policy experts. Until recently, much of the *technical* conversations around these issues seems to be isolated mainly in computer science, machine learning, and statistics, which certainly cannot result in an optimal outcome, since the issue of "what is fairness?" from a legal or ethical perspective clearly lies outside our area of expertise. It is imperative that those doing the technical design sufficiently take cues from scholars who have been studying these issues—particularly in the context of technology—for a long time.

APPENDIX A: OVERVIEW OF NOTIONS OF FAIRNESS

In the academic literature, there are several competing notions of algorithmic or model fairness, an overview of which can be found in Berk (2016). In general, we have found that notions of fairness can be divided into three camps. One school of thought does not focus on a particular metric of fairness, but rather assumes a model will be fair if the protected variable(s) are omitted from the analysis. This is often called "fairness through unawareness" [Kusner et al. (2017)]. In fact, some proprietary software packages used in predictive policing models tout the fairness of their models on the basis of omission of a race variable [Taylor (2015)]. Using this procedure, if the permitted variables are correlated with the protected variables, even if the protected variables are omitted, their effects will remain in the estimated model via their correlation with the permitted variables. The correlated variable may serve as partial proxies for the protected variables.

The other two schools of thought acknowledge that algorithmic fairness is a nontrivial problem, but propose different remedies because they define fairness differently. One area of research defines fairness in terms of equivalence of some measure of predictive accuracy among all classes in a protected variable. For example, Dieterich, Mendoza and Brennan (2016) argue that fairness is defined by similar accuracy and positive predictive value by class. Zafar et al. (2017) defines fairness in terms of equality of misclassification rates across class. A similar notion of fairness was proposed by Hardt et al. (2016), which argues that equivalence of false positive and false negative rates more accurately embodies everyday understanding of what it means to be fair. These two notions of fairness are indeed distinct. For example, Chouldechova (2017) shows the same positive predictive value by protected class and equal false negative and false positive rates cannot both be achieved when the outcome prevalence depends on protected characteristics. Kleinberg, Mullainathan and Raghavan (2017) shows theoretically that these notions of fairness are usually incompatible. A related literature focuses on methods and algorithms for achieving these notions of fairness by optimizing some utility or loss function subject to constraints that express the fairness criterion mathematically [e.g., Dwork et al. (2012), which also applies to achieving the alternative notion of fairness we adopt here].

A third approach that has gained traction primarily in the computer science and machine learning literature defines fairness in terms of disparate impact on a protected class [see Feldman et al. (2015), Barocas and Selbst (2016)]. Under this definition, a model is typically considered fair if differences in the distribution of the model's predictions conditional on the protected variable do not exceed some pre-determined threshold, as measured by some appropriate notion of distance between probability distributions. In most cases, the allowable difference is zero, which is equivalent to the requirement that the predictive distribution is independent of the protected variable. This notion is sometimes called "statistical parity" or "demographic parity." Our paper operates within this definition of fairness and builds upon the extant methodology for mitigating disparate impact.

Methodology for achieving statistical parity has centered on removing information about the protected variables from the training data by transforming the training data. Kamiran and Calders (2009) use a näive Bayes classifier to rank each observation in the training data by its probability of belonging to the "desirable" category.[4] Based on these rankings, the outcome variable in the training data is adjusted until there is no remaining association between the protected variable and the intended outcome variable. This procedure is limited to binary outcome data and the adjusted data is not re-usable in the sense that one could not then use the covariates to estimate relationships with other outcome variables and still be ensured of nondiscriminatory outcomes. Calders and Verwer (2010) presents three algorithms for preventing a model from producing differential predictions by protected class by transforming the training data in accordance to an objective function that is minimized when the predictions from a model fit to the transformed data are independent of the protected variable. In this case also, the methods are restricted to binary protected classes and binary outcome variables. Feldman et al. (2015) propose a method for adjusting or "repairing" the training data such that the user can tune the amount of permissible bias in models fit to the repaired training data. The authors suggest either removing information about the protected variable entirely or adjusting the training data such that the differences in conditional predictive distributions cannot exceed the legal definition of disparate impact. One limitation of this approach noted by the authors is that only continuous-type covariates can be repaired. Further, it is not clear how this procedure could be used to protect a continuous variable without discretizing it, an approach that was taken in

[4] The authors actually optimize for a different notion of fairness that is nonetheless closely related to statistical parity.

Adler et al. (2016). A review and comparison of several more algorithms operating on binary protected and outcome variables can be found in Romei and Ruggieri (2014).

Given that the outcome variable is observed with bias with respect to the protected variable, we believe that the second set of approaches designed with the objective of achieving equivalent predictive accuracy by race are inappropriate for this particular application, as they ultimately rely upon comparing the model's predictions of re-offense to a fundamentally flawed and biased measure of re-offense: re-arrest. The first class of approaches—simply omitting race from the set of covariates used to fit the model—is equally inadequate in this setting, as the covariates that are permitted to be used in the analysis are highly correlated with race.

APPENDIX B: OPTIMAL COUPLING AND TRANSPORT MAPS

Let $(\mathcal{X}, d)$ be a Polish space and $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Borel "cost" function. Let $\mu, \widetilde{\mu}$ be probability measures induced by random variables X, W. The transportation distance with respect to c is defined as

$$\rho_c(\mu, \widetilde{\mu}) \equiv \inf_{\gamma \in \Gamma(\mu, \widetilde{\mu})} \int c(x, w)\, d\gamma(x, w),$$

where γ is a *coupling* of $\mu, \widetilde{\mu}$—a joint distribution on $\mathcal{X} \times \mathcal{X}$ with marginals $\mu, \widetilde{\mu}$—and $\Gamma(\mu, \widetilde{\mu})$ is the space of all couplings of $\mu, \widetilde{\mu}$. The transportation distance is the minimal total cost with respect to c of transporting mass from μ to $\widetilde{\mu}$, and the coupling γ^* achieving the minimal cost is the *optimal coupling*, the solution to the Kantorivich transportation problem. In our context, γ^* tells us how to find W so as to minimize information lost, where information is quantified by c.

A natural choice in our setting is to set $c = d^q(x, w)$, with d the Euclidean norm. If one later uses any method or algorithm based on linear functions of the covariates—such as a generalized linear model—making the Euclidean distance between the original and transformed covariates small will make the loss of predictive accuracy small. This logic can be extended to a broader class of methods, such as kernel methods, by applying our proposed procedure to nonlinear transformations of x.

When $c(x, w) = d^q(x, w)$ for $q \geq 1$, the transportation distance is related to the Wasserstein-q distance by $\mathcal{W}_q^q(\mu, \widetilde{\mu}) = \rho_c(\mu, \widetilde{\mu})$, so the optimal coupling—when it exists—is also the coupling achieving the $\mathcal{W}_q$ distance.

B.1. Univariate transformations. When $\mathcal{X} = \mathbb{R}$ and d is the Euclidean norm, so that $\mu, \widetilde{\mu}$ are associated with distributon functions $F, G : \mathbb{R} \to [0, 1]$,

$$\mathcal{W}_q^q(F, G) = \int_0^1 |F^{\leftarrow}(p) - G^{\leftarrow}(p)|^q \, dp, \tag{B.1}$$

where $F^{\leftarrow}(p) \equiv \sup\{x \in \mathbb{R} : F(x) \leq p\}$ is the left-continuous inverse of F [Dall'Aglio (1956), Mallows (1972), Salvemini (1943), see also Ekisheva and

Houdré (2006)]. (B.1) does not require that F is continuous. This result allows us to define the optimal coupling explicitly in the case where F, G are absolutely continuous with respect to Lebesgue measure.

REMARK B.1 (Optimal coupling on $\mathbb{R}$). When $\mathcal{X} = \mathbb{R}$ with d the Euclidean norm and F, G have densities, the optimal coupling with respect to $c = d^q(x, w)$ for $q \geq 1$ is associated with the map $\zeta(x) = G^{-1}(\zeta^*(x))$ for $\zeta^*(x) = F(x)$.

PROOF.

$$\begin{aligned}\mathbb{E}_F\big[c\big(x, \zeta(x)\big)\big] &= \int_{\mathbb{R}} \{x - G^{-1}\big(F(x)\big)\}^q f(x)\,dx \\ &= \int_{\mathbb{R}} \{F^{-1}\big(F(x)\big) - G^{-1}\big(F(x)\big)\}^q f(x)\,dx \\ &= \int_{[0,1]} \{F^{-1}(p) - G^{-1}(p)\}^q\,dp.\end{aligned}$$

So $\zeta(x)$ achieves the transportation distance, and is therefore associated with the optimal coupling. □

The proof of remark B.1 only required that $\zeta^*(X)$ have a uniform distribution on the unit interval and $F^{\leftarrow}(\zeta^*(X)) = X$ F-almost surely. This suggests how to achieve the Wasserstein distance using random maps when F is atomic.

COROLLARY B.1 (Optimal coupling for atomic F using stochastic maps). *Suppose $\mathcal{X} = \mathbb{R}$ with d the Euclidean norm and F is atomic. Let $\dot{x} = \{\dot{x}_1, \dot{x}_2, \dots\}$ be the support points of F ordered such that $\dot{x}_j < \dot{x}_{j+1}$, with associated probabilities $\pi_j = \mathbb{P}[X = \dot{x}_j]$, and put $\nu_j = \sum_{j' \leq j} \pi_j$. Define a random map $\zeta^*(X)$ by $\zeta^*(X) \mid X = \dot{x}_j \sim \mathrm{Uniform}(\nu_{j-1}, \nu_j)$, with $\nu_0 = 0$. Then the random map $\zeta(X) = G^{\leftarrow}(\zeta^*(X))$ achieves the optimal coupling.*

PROOF. $\zeta^*(X) \sim \mathrm{Uniform}(0, 1)$ marginally and $F^{\leftarrow}(\zeta^*(X)) = X$ a.s. □

In order to achieve $\zeta(X) \perp Z$ within the class of optimal transport maps above, we must have

$$\zeta(X, Z) = G^{\leftarrow}\big(\zeta^*_{X|Z}(X, Z)\big), \tag{B.2}$$

where $\zeta^*_{X|Z}$ is either the conditional distribution $F_{X|Z}(X, Z)$ when F is continuous, or is a random variable constructed as in Corollary B.1 with $\pi_j = \mathbb{P}[X = x_j \mid Z]$ when F is atomic.

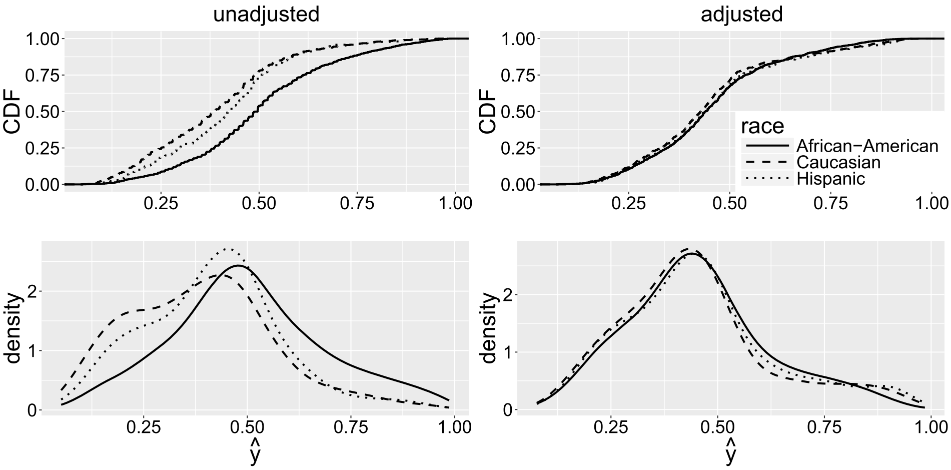

FIG. 9. *The cumulative distribution (top) and the density (bottom) of the predictions by race.*

APPENDIX C: RECIDIVISM PREDICTION USING LOGISTIC REGRESSION

This section presents the results of applying logistic regression to predict Y. In an analysis that mirrors that presented in Section 5.4, we compare a logistic regression model applied to unadjusted data to one applied to data that has been adjusted under the procedure we propose. Figure 9 shows the cumulative distribution and density of the predictions by race. Like we found when using RF, omitting the race variable does little to reduce discrepencies in the distribution of predictions by race. However, a logistic regression model applied to the adjusted datasets result in very similar distributions of fitted values by race. Figure 10 shows the ROC curves for each of the adjustment procedures. In this case also, there is little substantive difference between the methods in terms of this measure of predictive accuracy using this metric.

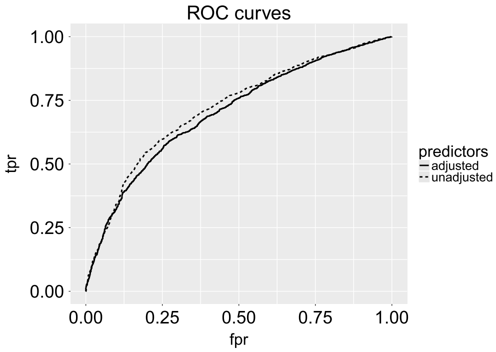

FIG. 10. *ROC curve showing predictive performance of each of the adjustment procedures.*

Acknowledgements. The authors thank David Dunson and Patrick Ball for helpful comments and conversations during the preparation of this manuscript.

REFERENCES

ADLER, P., FALK, C., FRIEDLER, S. A., RYBECK, G., SCHEIDEGGER, C., SMITH, B. and VENKATASUBRAMANIAN, S. (2016). Auditing black-box models for indirect influence. In *IEEE International Conference on Data Mining*.

ALPERT, G. P., SMITH, M. R. and DUNHAM, R. G. (2004). Toward a better benchmark: Assessing the utility of not-at-fault traffic crash data in racial profiling research. *Justice Res. Policy* **6** 43–69.

ANGWIN, J., LARSON, J., MATTU, S. and KIRCHNER, L. (2016). Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*.

BAROCAS, S. and SELBST, A. D. (2016). Big data's disparate impact. *Calif. Law Rev.* **104** 671–732.

BENJAMINI, Y. and HOCHBERG, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B* **57** 289–300. MR1325392

BERK, R. (2016). A primer on fairness in criminal justice risk assessment. *The Criminologist* **41** 6–9.

BERK, R., SHERMAN, L., BARNES, G., KURTZ, E. and AHLMAN, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *J. Roy. Statist. Soc. Ser. A* **172** 191–211. MR2655611

BRENNAN, T., DIETERICH, W. and EHRET, B. (2009). Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Crim. Justice Behav.* **36** 21–40.

BRIDGES, G. S. and CRUTCHFIELD, R. D. (1988). Law, social standing and racial disparities in imprisonment. *Soc. Forces* **66** 699–724.

BUUREN, S. and GROOTHUIS-OUDSHOORN, K. (2011). mice: Multivariate imputation by chained equations in R. *J. Stat. Softw.* **45**.

CALDERS, T. and VERWER, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Min. Knowl. Discov.* **21** 277–292. MR2720507

CHOULDECHOVA, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* **5** 153–163.

CUNNINGHAM, M. D. and SORENSEN, J. R. (2006). Actuarial models for assessing prison violence risk revisions and extensions of the risk assessment scale for prison (RASP). *Assessment* **13** 253–265.

DALL'AGLIO, G. (1956). Sugli estremi dei momenti delle funzioni di ripartizione doppia. *Ann. Sc. Norm. Super. Pisa* (3) **10** 35–74. MR0081577

DIETERICH, W., MENDOZA, C. and BRENNAN, T. (2016). COMPAS risk scales: Demonstrating accuracy equity and predictive parity. Technical report, Northpointe.

DVOSKIN, J. A. and HEILBRUN, K. (2001). Risk assessment and release decision-making: Toward resolving the great debate. *J. Am. Acad. Psychiatry Law* **29** 6–10.

DWORK, C., HARDT, M., PITASSI, T., REINGOLD, O. and ZEMEL, R. (2012). Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* 214–226. ACM, New York. MR3388391

EKISHEVA, S. and HOUDRÉ, C. (2006). Transportation distance and the central limit theorem. ArXiv preprint, Math/0607089.

FELDMAN, M., FRIEDLER, S. A., MOELLER, J., SCHEIDEGGER, C. and VENKATASUBRAMANIAN, S. (2015). Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 259–268. ACM.

FLORES, A. W., BECHTEL, K. and LOWENKAMP, C. T. (2016). False positives, false negatives, and false analyses: A rejoinder to machine bias: There's software used across the country to predict future criminals. And it's biased against blacks. *Fed. Probat.* **80** 38–46.

GLASER, J. (2014). *Suspect Race*: *Causes and Consequences of Racial Profiling*. Oxford Univ. Press, New York.

HARDT, M., PRICE, E., SREBRO, N. et al. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems* 3315–3323.

HOFFMAN, M., KAHN, L. B. and LI, D. (2015). Discretion in hiring. Technical report, National Bureau of Economic Research.

KAMIRAN, F. and CALDERS, T. (2009). Classifying without discriminating. In 2*nd International Conference on Computer*, *Control and Communication* IEEE.

KHANDANI, A. E., KIM, A. J. and LO, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *J*. *Bank*. *Finance*. **34** 2767–2787.

KLEINBERG, J., MULLAINATHAN, S. and RAGHAVAN, M. (2017). Inherent trade-offs in the fair determination of risk scores. In 8*th Innovations in Theoretical Computer Science Conference*. *LIPIcs*. *Leibniz Int*. *Proc*. *Inform*. **67** Art. No. 43, 23. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern. Available at arXiv:1609.05807. MR3754967

KUSNER, M. J., LOFTUS, J., RUSSELL, C. and SILVA, R. (2017). Counterfactual fairness. In *Advances in Neural Information Processing Systems* 4066–4076.

LANGAN, P. A. (1995). The racial disparity in U.S. drug arrests. Bureau of Justice Statistics (BJS) and US Dept. Justice and Office of Justice Programs and United States of America.

MALLOWS, C. L. (1972). A note on asymptotic joint normality. *Ann*. *Math*. *Stat*. **43** 508–515. MR0298812

MITCHELL, O. and CAUDY, M. S. (2015). Examining racial disparities in drug arrests. *Justice Q*. **32** 288–313.

PHILLIPS, M. T., FERRI, R. F. and CALIGIURE, R. P. (2016). Annual report 2014. Technical report, Criminal Justice Agency.

QUINSEY, V. L., HARRIS, G. T., RICE, M. E. and CORMIER, C. A. (2006). *Violent Offenders*: *Appraising and Managing Risk*, 2nd ed. American Psychological Association, Washington, DC.

REITER, J. P. (2005). Releasing multiply imputed, synthetic public use microdata: An illustration and empirical study. *J*. *Roy*. *Statist*. *Soc*. *Ser*. *A* **168** 185–205. MR2113234

REITER, J. P. and RAGHUNATHAN, T. E. (2007). The multiple adaptations of multiple imputation. *J*. *Amer*. *Statist*. *Assoc*. **102** 1462–1471. MR2372542

ROMEI, A. and RUGGIERI, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowl*. *Eng*. *Rev*. **29** 582–638.

RUBIN, D. B. (2004). *Multiple Imputation for Nonresponse in Surveys*. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ. Reprint of the 1987 edition [John Wiley & Sons, Inc., New York]. MR2117498

RUDOVSKY, D. (2001). Law enforcement by stereotypes and serendipity: Racial profiling and stops and searches without cause. *Univ*. *Pa*. *J*. *Const*. *Law* **3** 296.

SALVEMINI, T. (1943). Sul calcolo degli indici di concordanza tra due caratteri quantitativi. *Atti Riun*. *Sci*. *- Soc*. *Ital*. *Stat*..

SIMOIU, C., CORBETT-DAVIES, S. and GOEL, S. (2016). Testing for racial discrimination in police searches of motor vehicles. *SSRN Electron*. *J*. 2811449.

TAYLOR, M. (2015). No one gets hurt: Why the future of crime may be less violent and more insidious. *Calif*. *Mag*. **Summer 2015**.

WHITE, I. R., ROYSTON, P. and WOOD, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Stat*. *Med*. **30** 377–399. MR2758870

ZAFAR, M. B., VALERA, I., RODRIGUEZ, M. G. and GUMMADI, K. P. (2017). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the* 26*th International Conference on World Wide Web* 1171–1180.

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
SEQUOIA HALL
390 SERRA MALL
STANFORD, CALIFORNIA 94305
USA
E-MAIL: johndrow@stanford.edu

HUMAN RIGHTS DATA ANALYSIS GROUP
109 BARTLETT STREET
SAN FRANCISCO, CALIFORNIA 94110
USA
E-MAIL: kl@hrdag.org