Received: 5 March 2020 | Revised: 5 March 2020 | Accepted: 5 March 2020

DOI: 10.1002/bsl.2465

SPECIAL ISSUE ARTICLE

WILEY

Using algorithms to address trade-offs inherent in predicting recidivism

Jennifer Skeem[1] | Christopher Lowenkamp[2]

[1]Schools of Social Welfare and Public Policy, University of California, Berkeley, CA, USA

[2]Administrative Office, U.S. Courts, Washington DC, USA

Correspondence
Jennifer Skeem, University of California, Berkeley, 120 Haviland Hall #7400, Berkeley, CA 94720-7400, USA
Email: jenskeem@berkeley.edu

Abstract

Although risk assessment has increasingly been used as a tool to help reform the criminal justice system, some stakeholders are adamantly opposed to using algorithms. The principal concern is that any benefits achieved by safely reducing rates of incarceration will be offset by costs to racial justice claimed to be inherent in the algorithms themselves. But fairness trade-offs are inherent to the task of predicting recidivism, whether the prediction is made by an algorithm or human. Based on a matched sample of 67,784 Black and White federal supervisees assessed with the Post Conviction Risk Assessment, we compared how three alternative strategies for "debiasing" algorithms affect these trade-offs, using arrest for a violent crime as the criterion . These candidate algorithms all strongly predict violent reoffending (areas under the curve = 0.71–72), but vary in their association with race (r = 0.00–0.21) and shift trade-offs between balance in positive predictive value and false-positive rates. Providing algorithms with access to race (rather than omitting race or "blinding" its effects) can maximize calibration and minimize imbalanced error rates. Implications for policymakers with value preferences for efficiency versus equity are discussed.

1 | INTRODUCTION

Risk assessment instruments are data-driven evaluations of people that characterize the likelihood that they will reoffend in the future by assigning scores to risk factors like criminal history and substance abuse. Over recent years,

The views expressed in this article are those of the authors alone and do not reflect the official position of the Administrative Office of the US Courts.
Jennifer Skeem's work on this article was supported by the Mack Center for Mental Health and Social Conflict, University of California Berkeley.

there has been a resurgence in the use of risk assessment instruments or algorithms in the criminal justice system (Monahan & Skeem, 2014). Although many across the political spectrum extoll the potential contribution of these instruments in helping jurisdictions to reduce mass incarceration without increasing crime rates, others are equally adamant in opposition to incorporating algorithms into criminal justice reform efforts (compare Carter & Shames, 2020 with Pretrial Justice Institute [PJI], 2020). The principal concern is that any benefits achieved using risk assessment instruments will be offset by costs to racial justice claimed to be inherent in the algorithms themselves. Critics argue that algorithms "are derived from data reflecting structural racism and institutional inequity" (PJI, 2020), and speculate that using algorithms to inform human decision-making in the justice system will exacerbate existing racial disparities in incarceration.

Concerns about the legitimacy of specific risk factors and the potential discriminatory effects of using risk assessment instruments are not new – and can be examined empirically. These concerns were addressed in criminal justice research decades ago, during a similar debate about using structured decision-making tools – including sentencing guidelines and risk assessments – to inform sanctioning decisions (e.g., Fisher & Kadane, 1983; Gottfredson & Gottfredson, 1985; Gottfredson & Jarjoura, 1996). Given the recent resurgence in the use of risk assessment in criminal sanctioning (Monahan & Skeem, 2014) – and the rise of big data and predictive analytics – these concerns have become the focus of a new body of research on algorithmic fairness (e.g., Chouldechova, 2017; Corbett-Davies, Pierson, Feller, & Goel, 2016; Kleinberg, Ludwig, Mullainathan, & Sunstein, 2019; Kleinberg, Mullainathan, & Raghavan, 2016; Skeem & Lowenkamp, 2016). It has now become clear that there are mathematical limits "to how fair any algorithm – *or human decision-maker* – can ever be," when predicting recidivism (Corbett-Davies et al., 2016, emphasis added). As explained later, there are multiple definitions of fairness and there are trade-offs between them. These trade-offs are inherent in the prediction task, rather than in algorithms. Risk is a legally relevant consideration for decisions that judges, probation officers, and other professionals make every day about individuals – and trade-offs apply, whether the prediction of reoffending is based on algorithms or on a human's explicit or implicit judgment alone.

In this study, we compare how alternative strategies for "debiasing" risk assessment algorithms affect these trade-offs, based on a large sample of Black and White federal probationers assessed with the Post Conviction Risk Assessment (PCRA) (Johnson, Lowenkamp, VanBenschoten, & Robinson, 2011). Our goal is to illustrate that algorithms make inherent trade-offs between competing values (like crime prevention and racial justice) apparent, and that policymakers can choose among variants of an algorithm to make the trade-off that they find most acceptable, explicit and actionable. Our results include the counterintuitive, but common, finding that access to the protected variable of race can increase both predictive accuracy and racial equity. Our approach is inspired by Kleinberg et al.'s (2019) comprehensive demonstration that well-designed and well-regulated algorithms – especially when provided access to protected variables like race – improve transparency and provide opportunities to detect discrimination that are unavailable when decisions are based solely on unaided human judgment. In a similar vein, Garrett and Monahan (in press) articulate strategies for regulating the use of risk assessment instruments in the criminal justice system. With appropriate regulation and human oversight, Kleinberg et al. (2019) argue, algorithms can "be a potential positive force for equity" (p. 1).

Before detailing the current study's goals and approach, we first describe its context. We begin by noting how risk assessment instruments have been applied to support criminal justice reform efforts, and then highlight different definitions of racial fairness that have emerged in debate about the use of these algorithms.

1.1 | Criminal justice prediction context

Over recent years, jurisdictions across the US have been implementing a variety of efforts designed to reduce incarceration rates that are the highest in the world and that disproportionately affect Black people and communities of color, while at the same time continuing to recognize the need for community safety. Many jurisdictions

are using risk assessment instruments to support their efforts. In fact, risk assessment has been called "the engine that drives" a federal prison reform bill recently signed into law (Garrett, 2018, para. 1). Although strategies for using risk assessment as a tool for reform vary, the overarching logic is straightforward. These algorithms are purpose-built to predict reoffending, and one way to reduce incarceration without increasing crime rates is to accurately identify the people who are least likely to reoffend and release them, supervise them in the community on probation or parole, or abbreviate their period of incarceration (Monahan, 2017; Monahan & Skeem, 2014). Risk assessment can also be used to identify higher risk people and prioritize them for high-quality correctional services, given that these people have been shown to benefit the most from treatment that reduces the likelihood of recidivism (see Skeem & Polaschek, in press).

Judges, probation officers, parole commissioners and others routinely make momentous decisions in a person's life that include consideration of the likelihood that the person will reoffend. Without algorithms, legal actors must rely solely on their own intuitive judgment to estimate an individual's risk. Risk is often relevant, when justice professionals make decisions about who to release from jail before trial, who to release from prison on parole, who can go to low-security versus higher security settings, and who should be prioritized for rehabilitative services (for a review, see Goel, Shroff, Skeem, & Slobogin, in press).

Advocates argue that when risk is a legally relevant consideration, justice professionals should consider risk assessment instruments to improve the consistency, transparency, and accuracy of their decisions (e.g., Monahan, 2017; Neufeld, 2018). First, humans' intuitive predictions of reoffending are opaque, which makes them difficult to challenge as discriminatory, even when they have been implicitly or explicitly influenced by race. By contrast, well-made and well-regulated algorithms can "create new forms of transparency and hence opportunities to detect discrimination that are otherwise unavailable" (Kleinberg et al., 2019, p. 1). Second, over 60 years' worth of research indicates that algorithms are more accurate in predicting violence and other criminal behavior than unaided human predictions, including those made by judges and community supervision officers (e.g., Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2017; Lin, Jung, Goel & Skeem, 2020; Gottfredson, 1999; for a review, see Goel et al., in press).

Nevertheless, some critics oppose the use of risk assessment instruments to scaffold criminal justice reform efforts, based on concerns that algorithms are racially biased and could result in inequitable outcomes for already disadvantaged groups. Race is omitted from these instruments. But in an era of general skepticism about the fairness of algorithms (Courtland, 2018; O'Neill, 2016), critics argue that risk factors included in some risk assessment instruments (e.g., education level, employment, marital status) are correlates of, or even "proxies" for, minority race (Starr, 2014). In the view of Former Attorney General Eric Holder (2014, paragraph 23), the broad use of risk assessment "may exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."

1.2 | Defining and debating racial fairness

1.2.1 | Fairness as balance in error rates for outcomes

These concerns have become increasingly popular among stakeholders in the justice system, based on a widely publicized investigation conducted by the news organization ProPublica. ProPublica claimed that a risk assessment instrument called COMPAS is "biased against blacks" (Angwin, Larson, Mattu, & Kirchner, 2016). Using a publicly available dataset, Angwin et al. (2016) demonstrated that, among defendants who ultimately did not reoffend, Blacks were about twice as likely as Whites to have been misclassified as medium or high risk by the COMPAS (40% vs. 22%). In other words, the "false positive rate" was higher for Black defendants than for White defendants, so the court's use of the COMPAS to inform decisions could inappropriately subject Black defendants to harsh treatment more often

than White defendants. Applying this definition of fairness that focused on balance in error rates for outcomes, the algorithm was biased.

1.2.2 | Fairness as balance in accuracy rates for predictions

Before ProPublica's investigation was released, we had tested the racial fairness of a risk assessment instrument called the Post Conviction Risk Assessment (PCRA; Johnson et al., 2011), based on a sample of over 34,000 people on federal community supervision (Skeem & Lowenkamp, 2016). We focused on a different definition of fairness that is articulated in the Standards for Educational and Psychological Testing (the "Standards;" American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014), a definition that focuses on balance in accuracy rates for predictions.

The Standards' definition of fairness that we tested using traditional nested regression models (Cleary, 1968; Sackett & Bobko, 2010) corresponds to 2017 formal definitions of "calibration" and, to a lesser extent, "positive predictive value" (Chouldechova, 2017). Using these definitions, we found that the PCRA strongly predicted rearrest for a violent offense for both Black and White supervisees and was not biased by race. We demonstrated that a given score on the PCRA meant the same thing, regardless of group membership – each score corresponded to a similar probability of reoffending for both Black and White supervisees. Although PCRA scores were well calibrated by race, we expressed concern that some uses of the PCRA could have disparate impact because Black supervisees obtained modestly higher average PCRA scores than White supervisees.

After ProPublica's investigation was released, it was widely criticized. Both scholars and developers of the COMPAS (e.g., Dieterich, Mendoza, & Brennan, 2016; Flores & Bechtel, 2016) reanalyzed the same public database and argued that the instrument was free of racial bias. Applying definitions of fairness like the ones we had used, these authors found that COMPAS scores were well calibrated (Flores & Bechtel, 2016) and had similar positive predictive values across race at key thresholds (Dieterich et al., 2016). For example, of defendants with a COMPAS score of seven, 60% and 61% of White and Black defendants reoffended, respectively (see Corbett-Davies et al., 2016). Using definitions of fairness that focus on balance in accuracy rates for predictions, then, the algorithm was not biased.

1.3 | Trade-offs inherent in prediction task

Still more analyses of the public COMPAS database demonstrate that when Black and White groups differ in their rates of reoffending, it is impossible to satisfy both of the above fairness definitions at the same time (Corbett-Davies et al., 2017; Chouldechova, 2017; see also Kleinberg et al., 2016). When Black defendants have a higher rate of reoffending than White defendants and an algorithm is well-calibrated to those true outcomes, the algorithm will classify a greater proportion of Black defendants as high risk, and therefore, a greater proportion of Black defendants who ultimately do not reoffend will have been classified as high risk (Corbett-Davies et al., 2017). In other words, the imbalance in false positive rates that ProPublica presents as evidence of racial bias is "a direct consequence" of applying an algorithm that is well-calibrated to a population where the prevalence of recidivism differs by group (Chouldechova, 2017, p. 2).

These trade-offs are inherent to the task of prediction, when rates of reoffending differ across groups. These group differences are common. Based on a 5-year follow-up of 404,638 prisoners released from 30 states, Durose, Cooper, and Snyder (2014) found that 81% of Black people were rearrested, compared with 75% and 73% of Hispanics and non-Hispanic White people, respectively. As Corbett-Davies et al. (2017) note, "the solution is not to eliminate statistical risk assessments. The problems we discuss apply equally to human decision-makers, and humans are additionally biased in ways that machines are not. We must continue to investigate and debate these issues as algorithms play an increasingly prominent role in the criminal justice system."

1.4 | Comparing how alternative approaches to "debiasing" algorithms affect trade-offs

In the current study, we use a matched sample of over 67,784 federal probationers to compare the effect of different candidate algorithms on trade-offs inherent in predicting recidivism, i.e., trade-offs between calibration and positive predictive value on one hand, and false positive and false negative error rates on the other. We focus on rearrest for a violent crime as the outcome variable because it is a relatively unbiased criterion that reflects serious reoffending. For example, Beck and Blumstein (2018) demonstrate that, whether data are drawn from a national victimization survey or national arrest database, the proportion of assailants who are Black across four major types of violent crimes is essentially the same (33% and 32%, respectively). The candidate algorithms that we compare represent regression-based "debiasing" approaches offered decades ago in the criminological literature (e.g., Gottfredson & Gottfredson, 1985; Gottfredson & Jarjoura, 1996), which now have modern equivalents, in a rapidly growing body of research on algorithmic fairness (mostly in computer science; see Kleinberg, Ludwig, Mullainathan, & Rambachan, 2018).

We examine five algorithms based on the PCRA (Johnson et al., 2011). The first two algorithms are provided for the sake of benchmark comparisons. The "race omitted" algorithm includes only the PCRA items as predictors and represents the dominant approach in risk assessment, which is to simply omit race as a legally protected characteristic. By contrast, the "race fitted" algorithm consists of the PCRA items, race, and interactions between each PCRA item and race as predictors (as in Kleinberg et al., 2018). The interaction terms address potential feature bias, or different predictive utility of PCRA items as a function of race (e.g., family problems predicting violent reoffending more strongly for White than for Black participants; Skeem & Lowenkamp, 2016).

The remaining three algorithms represent candidates for debiasing PCRA predictions. Omitting race as a predictor does not purge the algorithm of its effect because the variance that race shares with other predictors "is left behind to affect the weights of variables felt legitimate for inclusion" (Gottfredson & Jarjoura, 1996, p. 54). Race reflects longstanding patterns of social and economic inequality in the US (e.g., differences in schools, neighborhoods, social networks) that can increase the likelihood of criminal behavior (Walker, Spohn, & DeLone, 2011). When race is associated with risk factors, the challenge is to remove its effect while still achieving as much predictive utility as possible. We compare three responses to this challenge:

1 The "proxy eliminated" algorithm represents a relatively conservative response. This approach involves two steps: (I) constructing an algorithm that includes race and generates predictive weights for each variable (with PCRA items plus race as predictors); and then (II) "controlling" for the effects of race by eliminating its variation in the sample of interest (see Gottfredson, Gottfredson, & Gottfredson, 2000; Gottfredson & Jarjoura, 1996). In practical terms, Step II can be achieved by recoding all sample members as the same race (White or Black). Having eliminated variation in race, predictive weights from Step I are used to produce individuals' predicted probabilities of violent reoffending. This approach is meant to ensure that the predictive power of the PCRA items is independent and does not come from the items' correlation with race (i.e., to eliminate "proxy" effects in Step I), while building in a policy control that removes race from an individual's risk estimate (in Step II).
2 The "race eliminated" algorithm represents a relatively liberal response to the debiasing challenge. This algorithm is generated by, first, preprocessing PCRA item scores to entirely remove their association with race and then using the processed PCRA item scores – which are orthogonal to race – as predictors of violent reoffending (see Gottfredson & Jarjoura, 1996; Kleinberg et al., 2018). This approach is designed to produce PCRA risk estimates that are completely uncorrelated with race. This "race eliminated" algorithm removes all variance that race shares with the PCRA items, whereas the "proxy eliminated" algorithm above removes only the violence-predictive variance that race shares with the PCRA items.

3 The "criminal history discount" algorithm represents an exploratory, common-sense approach to debiasing the PCRA. Most of the modest association between race and PCRA total scores is based on the PCRA's criminal history scale (Skeem & Lowenkamp, 2016). Although criminal history robustly predicts recidivism, biases in policing and decision-making about arrests could artificially inflate the number of arrests for Black people. This algorithm reduces Black supervisees' scores on the PCRA prior arrest item before using the processed PCRA items as predictors.

We compare the effect of these candidate 'debiasing' algorithms on inherent trade-offs between prediction-focused accuracy rates (relevant to crime prevention) and outcome-focused error rates (relevant to racial justice). Algorithms quantify these trade-offs but cannot make choices among them; it is up to policymakers to weigh legal, ethical, and value considerations and choose the trade-off that is most acceptable to them (see Kleinberg et al., 2019). This study is meant to demonstrate one approach that researchers and stakeholders could follow, to address concerns about trade-offs in efficiency and equity that are not unique to the criminal justice context (see Kleinberg et al., 2019; Skeem & Lowenkamp, 2016). Although the approach is presented for illustrative purposes, trade-offs in fairness are legitimate and important concerns with the PCRA. The PCRA was developed by the Administrative Office of the US Courts (Probation and Pretrial Services Office) and is administered post-conviction, upon intake to a term of supervised release or probation. A supervisee's PCRA score is used to determine his or her level of supervision. Higher levels of supervision require additional contacts with the assigned probation officer and failure to make those contacts can lead to revocation. When supervisees violate conditions of probation, judges may "revoke a term of supervised release, and require the defendant to serve in prison all or part of the term of supervised release … without credit for time previously served on postrelease supervision" (17 USC §3,583(e)(3). Given connections among risk scores, supervision levels, and the possibility of revocation, it is critical to consider trade-offs inherent in prediction.

2 | METHOD

2.1 | Participants

Participants in this study were drawn from a larger dataset on over 287,595 federal supervisees who were assessed between November 2009 and January 2019. Eligibility criteria were: (a) assessed with the PCRA at least 1 year prior to the collection of follow-up arrest data (to permit a minimum follow-up period, number lost = 37,907); (b) no missing data on PCRA items (to permit computation of alternative algorithms; number lost = 23,773); and (c) race coded as either "Black" or non-Hispanic "White" (to permit the racial comparison of focus; number lost = 69,430). Application of these criteria yielded an eligible pool of 154,485 supervisees.

Within this eligible pool, there were potentially confounding differences between Black and White participants. For example, Blacks were more likely than Whites to be male (86% vs. 79%) and young (age, *M* of 39 vs. 44), and sex and age are robust risk factors for recidivism. The groups also differed in offense type (which can mark differential selection) and federal district (where policing and arrest practices can differ). To isolate the effect of race on risk estimates and recidivism – without creating non-representative groups – we adopted a conservative matching approach. We randomly matched each Black participant to a White participant on age, sex, offense, and district using ccmatch in STATA (Cook, 2015). This process yielded a sample of 67,784 participants: 33,892 Black and 33,892 White. Compared with the larger sample from which it was drawn, the present study sample is similar in demographics, conviction offense, and PCRA scores. As shown in Table 1, the prototypic participant was male, aged 40, and convicted of a drug offense.

TABLE 1 Sample characteristics and rearrest rates

Characteristic	All (*N* = 67,784)	Black (*N* = 33,892)	White (N = 33,892)
Age (years)	40.00 (10.48)	40.00	40.00
Male (%)	86	86	86
Conviction offense[a] (%)			
Drug	48	48	48
Firearms	16	16	16
White collar	17	17	17
Public order	8	8	8
Property	5	5	5
Violence	4	4	4
Sex offense	2	2	2
Arrested any offense (%)	34	38	30
Arrested violent offense (%)	10	12	8
Average follow-up period (days)	1,683 (637)	1,672 (642)	1,694 (633)

Note: Characteristics for the full sample shown here also represent characteristics of the construction and cross-validation subsamples (i.e., the halves of the sample randomly selected for algorithm construction and testing).
[a]Categories with < 5% excluded.

2.2 | PCRA-based risk algorithms

2.2.1 | PCRA psychometrics

The history, development, and predictive utility of the PCRA are detailed elsewhere (see Johnson et al., 2011; Lowenkamp, Holsinger, & Cohen, 2015; Lowenkamp, Johnson, Holsinger, VanBenschoten, & Robinson, 2013). Briefly, the PCRA is an actuarial instrument that explicitly includes variable risk factors and was constructed and validated on large, independent samples of federal supervisees. Items that most strongly predicted recidivism in the construction sample contribute most strongly to total scores (Johnson et al., 2011). Fifteen items are scored and summed to yield a total PCRA risk score that places an individual into one of four risk categories: low, low/moderate, moderate, or high. Each of the 15 items is nested under one of five risk factor domains, four of which are considered changeable (i.e., all but criminal history; Cohen, Lowenkamp, & VanBenschoten, 2016). The domains and items are listed below [except for the first two items listed, items are scored dichotomously (0 or 1)]:

- "Criminal history" includes number of prior arrests (0 = none; 1 = one to two; 2 = three to six; 3 = seven or more), young age (0 = 41+; 1 = 26–40; 2 = under 26), community supervision violations, varied offending pattern, institutional adjustment problems, and violent offense.
- "Employment and education" includes highest grade completed, unstable recent work history, and currently unemployed.
- "Social networks" includes family problems, unmarried, and lack of social support.
- "Substance abuse" includes recent alcohol problems and recent drug problems.
- "Attitudes" is low motivation to change.

The PCRA is reliable and strongly predicts reoffending. Specifically, officers who administer the PCRA must complete a training and certification process. The certification process has been shown to yield high rates of inter-rater agreement in scoring (Lowenkamp et al., 2013). The accuracy of the PCRA in predicting recidivism rivals that of other

well-validated instruments. For example, based on a sample of over 100,000 supervisees, Lowenkamp et al. (2015) found that the PCRA moderately to strongly predicted both rearrest for any crime and rearrest for a violent crime, over up to a 2-year period (AUCs = 0.70–0.77). Finally, scores on the PCRA have been shown to change over time. Of people initially classified as high risk on the PCRA, 47% move to a lower risk classification upon reassessment an average of 9 months later (Cohen et al., 2016). The greatest changes observed were in employment/education and substance abuse. For the present study, we used PCRA scores that were administered by officers when an individual entered supervision, to provide the longest follow-up period possible.

2.2.2 | PCRA algorithms

We used regression-based approaches to create the five algorithmic variants of the PCRA described in the Introduction. The "race omitted" algorithm includes only the 15 PCRA items as predictors, whereas the "race fitted" algorithm includes these 15 PCRA items, race, and race × item interactions as predictors.

The "proxy eliminated" algorithmic approach begins by including the 15 PCRA items and race to generate predictive weights.[1] Race is included in this step to ensure that the PCRA items are free of its contribution. Next, the effect of race on individual predicted probabilities of violent reoffending is controlled by eliminating its variation in the sample of interest (see Gottfredson & Jarjoura, 1996). In practice, this amounts to treating all individuals as "Black," "White," or another value (which is a policy decision; see Gottfredson et al., 2000). For this demonstration, we followed Pope and Sydnor's (2011) recommendation and replaced individual values for race with the sample proportion of Black supervisees (i.e., 0.50) as the policy control, to calculate individuals' predicted probabilities.

The "race eliminated" algorithm was generated by regressing each of the 15 PCRA items on race and retaining the residual values, and then using the residual PCRA item scores as predictors of violent reoffending (see Gottfredson & Jarjoura, 1996; Kleinberg et al., 2018).

The "criminal history discount" algorithm was calculated by first reducing Black supervisees' scores on this domain by 23% (based on a meta-analytic estimate of the effect of race on police decisions to arrest; Kochel, Wilson & Mastrofski, 2011), and then using this revised domain and the remaining PCRA items as predictors.

To calculate PPVs and error rates for the algorithms, we had to select a cutoff score for defining high-risk classifications. Threshold scores for all five algorithms' high-risk classifications were set to align those of the original PCRA, based on predicted probabilities of rearrest. At the lower threshold for high-risk classifications, these probabilities are 63% and 23% for any arrest and violent arrest, respectively (compared with 48% and 15% for moderate classifications; 25% and 6% for low to moderate classifications; and 4% and 0% for low classifications, respectively).

2.3 | Arrest criterion

Data from the National Crime Information Center (NCIC) and Access to Law Enforcement System were used to collect information on arrests. Specifically, Federal Bureau of Investigation (FBI) staff conducted a standard criminal history check on each participant that yielded their entire criminal history. The authors coded the date and types of arrests that occurred after the date of PCRA administration.

The result consisted of two dichotomous measures: arrest for a violent offense and arrest for any offense (excluding technical violations of standard conditions of supervision). Violence was defined using the NCIC definitions (i.e., homicide and related offenses, kidnapping, rape and sexual assault, robbery, assault). Given the importance of using a valid criterion to assess the predictive fairness of algorithms, our outcome analyses focus on the criterion

[1]This proxy-eliminating approach can also be applied with the "race fitted" algorithm as a base in the first step (i.e., including not only race, but interactions between race and the PCRA items). We follow the simpler approach in this demonstration.

of arrest for a violent offense. According to differential selection theory, racial disparities reflect bias in policing and decisions about arrest. This theory applies less to crimes of violence than to crimes that involve drugs (see Blumstein, 2011) and greater police discretion (e.g., "public order" crimes; see Piquero & Brame, 2008). Beck and Blumstein (2018) examined the correspondence between the race of people who police arrested for four major types of violent crime in the US in 2010, and the race of people who victims identified as assailants for those types of crimes in a national victimization survey conducted the same year. The ratio of the percentage of arrestees who were Black (33%) and the percentage of assailants identified by victims as Black (32%) was 0.97 (range 0.91–1.10), providing "strong indication of support for the use of arrest as a proxy indicator of criminal involvement" in violent crime (p. 877). Arrest for a violent crime is, in short, the most accurately recorded and least racially skewed available, and appears to be a valid measure of involvement in violence. In the present sample, the base rate for violent arrest was 10% (12% Black; 8% White, $\chi^2(1) = 282.803$; $p < 0.001$, $\phi = -0.065$) and the base rate for any arrest was 34% (38% Black; 30% White, $\chi^2(1) = 514.15$; $p < 0.001$; $\phi = -0.087$). Black participants were more likely to be arrested than Whites.

As shown in Table 1, all participants were followed for a minimum of 1 year, but the average follow-up period (i.e., time at risk for reoffending) was > 4 years. The difference in the average follow-up time between Black and White participants was statistically significant but small, with Whites being followed an average of 22.44 days longer [$t(67{,}771.2) = -4.58$; $p < 0.001$].

2.4 | Analyses

Regression analyses, which are transparent and relatively easy to implement, were the chief method used to develop and test the algorithms. To avoid overfitting, we used half of the sample to construct each candidate algorithm and the other half of the sample to test the algorithm's performance. Because even trivially small differences can become statistically significant in samples as large as ours (Lin, Lucas, & Shmueli, 2013), we use an alpha level of 0.001 to signal statistical significance and focus on effect sizes in interpreting results.

Our results are meant to represent the performance of alternative PCRA algorithms in the full federal population, across its 94 districts. To address concerns that the data may cluster by district, we used robust standard errors in the regression models to adjust for any heteroscedasticity. Specifically, the variance–covariance estimator with clustering by district was used to address the potential correlation between error terms within districts [STATA vce(cluster); Guitterez & Drukker, 2007; Rogers, 1993].

To test the predictive fairness of alternative PCRA models in a manner that is consistent with the Standards and our past approach (Skeem & Lowenkamp, 2016), we examined whether Black and White groups systematically deviate from a common regression line that relates algorithm scores to the criterion of violent reoffending (i.e., tested whether the groups share intercepts and slopes; Cleary, 1968; see also Sackett & Bobko, 2010). This involves testing the relative fit among three nested regression models: intercept bias would be present if the model that includes the PCRA score alone is outperformed by the model that includes both the PCRA score and race, and slope bias would be present if the model that includes an interaction between race and the PCRA score outperforms the model that includes only race and the PCRA score. We tested the improvement in model fit (i.e., difference between the two models' log likelihood ratios multiplied by 2) for significance using the χ^2 distribution.

To concretize trade-offs between balance in prediction-focused accuracy rates and outcome-focused error rates (following Chouldechova, 2017), we calculate the PPV, FPR and false negative rate (FNR) for high-risk classifications associated with each algorithmic variant of the PCRA, using the threshold scores for high-risk classifications described earlier (see "PCRA algorithms"). The PPV is the percentage of people classified by the algorithm as high risk who were rearrested for a violent offense. The FPR is the percentage of people who were not rearrested for a violent offense but had been classified as high risk by the algorithm. The FNR is the percentage of people who were rearrested for a violent offense but who had not been classified as high risk by the algorithm.

3 | RESULTS

In this section, we first summarize the overall accuracy of the five alternative PCRA algorithms (two benchmark and three 'debiasing' approaches) in predicting rearrest for violence, along with the algorithms' association with race. To permit comparisons with past work, we then present the results of traditional Standards-based tests of the algorithms for slope bias and intercept bias by race, which correspond to calibration and (to a lesser extent) PPV, respectively. Finally, to illustrate how high-risk classifications on alternative algorithms affect trade-offs between predictive accuracy and error rates, we present PPVs, FPRs and FNRs by race for each model. We conclude by illustrating, for each algorithm, how allowing threshold scores for high-risk classifications to vary by race affects these trade-offs.

3.1 | Describing algorithms' overall predictive utility and association with race

To characterize the overall predictive utility of alternative PCRA algorithms, we used a measure of ranking accuracy called the area under the receiver operating curve (AUC). The AUC is widely used to assess the predictive utility of risk assessment instruments because it is readily interpretable and its values are not heavily influenced by base rates of offending, which vary across studies. Given the low base rate of violent reoffending in the present study, the AUC is particularly useful for benchmark comparisons with other research. AUC values range from 0.50 (i.e., accuracy is not improved over chance) to 1.00 (i.e., perfect accuracy). Minimum AUCs of 0.56, 0.64, and 0.71 correspond to "small," "medium," and "large" effect sizes, respectively (see Rice & Harris, 2005).

As shown in Table 2, the AUC values for all five algorithms are large. More importantly, there are no statistically significant differences among the algorithms' AUCs, indicating that their predictive utilities are interchangeable. Regardless of the algorithm used, the AUC values indicate a 71–72% chance that a supervisee randomly selected from those who violently recidivated will obtain a higher PCRA score than one randomly selected from those who did not violently recidivate.

Also shown in Table 2 is the association between supervisees' race and PCRA scores generated by each algorithm. Using Cohen's (1988) conventions for interpreting correlations, the results indicate "small" correlations between race and three PCRA algorithms, i.e., the race-omitted and race-fitted benchmark algorithms and the proxy-eliminated debiasing candidate (with no significant differences among them). As expected – and by design – the race- and criminal history discount PCRA algorithms are essentially unassociated with race; and significantly less associated with race than the other three algorithms.

TABLE 2 Alternative Post Conviction Risk Assessment (PCRA) algorithms: overall predictive utility and association with race

PCRA algorithm	Area under the curve (AUC)	Correlation with Black race
Benchmark comparisons		
Race-omitted	0.721	0.190
Race-fitted	0.722	0.210
Debiasing candidates		
Proxy-eliminated	0.721	0.192
Race-eliminated	0.713	−0.006
Criminal history discount	0.714	0.029

Note: Figures are based on the cross-validation sample (N = 34,021). Although none of the differences among the five models' AUCs is statistically significant, the race-eliminated and criminal history discount algorithms are less associated with race than the other three algorithms.

Together, these results indicate that the five algorithms vary in their association with race – and nevertheless have similar levels of utility in predicting rearrest for a violent crime. This helps address concerns that removing the effect of race – and, in the case of the race-eliminated model, all of its association with other predictors – would compromise the models' predictive utility.

3.2 | Applying Standards-based tests of algorithms' fairness

Next, we used a traditional Standards-based approach to test the calibration and PPV of alternative PCRA algorithms (Skeem & Lowenkamp, 2016). Using the series of three nested regression models described earlier for each alternative algorithm (see the Analyses section), we examined whether Black and White groups systematically deviate from a common regression line that relates the algorithm's scores to the criterion of violent reoffending. The results are summarized in Table 3, which shows differences in model fit used to statistically test for slope bias and intercept bias, along with effect size estimates for the improvement in model fit for the more complex model (R^2 change) and odds ratios for the term of interest added by the more complex model (the algorithm × race interaction for slope bias, and race for intercept bias).

These results in Table 3 highlight three points. First, none of the algorithms manifests significant slope bias. Race does not moderate the strength of the relationship between PCRA algorithms and violent reoffending. Instead, all five algorithms are well calibrated by race – they strongly predict violent recidivism for Black and White participants. Second, all five algorithms manifest statistically significant intercept bias. Across algorithms, intercepts of the relationship between PCRA scores and violent recidivism are higher for Black participants than for White participants, indicating that a given PCRA score "underpredicts" recidivism for Blacks and "overpredicts" recidivism for Whites. Third, the absolute amount of intercept bias is small across algorithms (based on R^2-change values and odds ratios for race), but larger for the race-eliminated and criminal history discount candidates than for the remaining three algorithms, including the proxy-eliminated candidate. The odds ratio for the race-eliminated and criminal history discount algorithms indicate that, after taking PCRA scores into account, Black participants are 1.66 and 1.62 times

TABLE 3 "Test standard" indices of predictive fairness or calibration by race, for alternative Post Conviction Risk Assessment (PCRA) algorithms

	Slope bias			Intercept bias		
PCRA algorithm	ΔR^2, adding interaction	OR for algorithm × race interaction	Improvement in model fit	ΔR^2, adding race	OR for Black race	Improvement in model fit
Benchmark comparisons						
Race-omitted	0.000	0.892	0.328	0.001	1.176	18.526*
Race-fitted	0.000	0.783	1.488	0.001	1.128	10.106*
"Debiasing" candidates						
Proxy-eliminated	0.000	0.999	0.046	0.001	1.176	18.444*
Race-eliminated	0.000	1.027	0.018	0.008	1.655	184.104*
Criminal history discount	0.000	1.337	3.232	0.008	1.622	169.170*

*$p < 0.001$.

Note: Figures are based on the cross-validation sample (N = 34,021). Results reflect comparisons of two sets of nested regression models: models with PCRA scores alone versus PCRA scores plus race (to test intercept bias), and models with the main effects of PCRA scores and race only versus main effects and their interaction (to test slope bias). The improvement in model fit (based on differences in log likelihood ratios) is tested using the χ^2 distribution. ΔR^2 change = the increase in the value of R^2 with the more complex model; OR = odds ratio.

more likely than White participants to be arrested for a violent crime, respectively. This implies that the PPVs of these candidate algorithms are modestly unbalanced by race.

3.3 | Comparing algorithms' effects on trade-offs in PPV and outcome error rates

Finally, we compared the algorithms' PPVs, FPRs, and FNRs – using high risk classifications for the alternative PCRA algorithms that correspond to those of the original PCRA. These classifications differentiate the high-risk group from all others (i.e., moderate-high, moderate, low-moderate, and low). Notably, the base rate of the criterion is low (10%), which limits the absolute magnitude of some estimates even though PCRA algorithms strongly predict violent reoffending. Our focus is on comparing the relative balance that algorithms achieve between Black and White participants in positive predictive value versus error rates.

The results are shown in Table 4. For each algorithm, this table shows PPV, FPR and FNR values for Black and White participants – along with the Black–White difference in percentage points for each value and Cohen's (1988) *h*, an effect size for the distance between the two proportions. By convention, *h* values of 0.20, 0.50 and 0.80 are interpreted as small, medium, and large differences, respectively.

Table 4 reveals three essential points. First, across all five algorithms, imbalances in predictive utility (PPV) or outcome error rates (FPR and FNR) are small or "smaller than small," by conventional interpretive standards. Of course, policymakers may interpret the magnitude of these imbalances as larger, given their values or given how the algorithm would be applied (e.g., to inform relatively low-stakes decisions like service allocation or high-stakes decisions like incarceration). Second, compared with the other three algorithms, the race-eliminated and criminal history discount candidates reverse the direction of imbalanced error rates, so that Whites have higher FPRs than Blacks, and Blacks have higher FNRs than Whites. Third, compared with the other two algorithms, the race-eliminated and criminal history discount candidates – as well as the race-fitted benchmark – show greater imbalance in PPVs and lesser imbalance in error rates.

There are also similarities among the five algorithms. In keeping with tests of intercept bias reported earlier, the algorithms share the same direction of imbalance in PPV, which favors Black participants (though the magnitude of this imbalance is greater for the race-eliminated, criminal history discount, and race-fitted algorithms than the remaining two candidates). Except for the race-fitted algorithm (which shows relatively good balance), the algorithms are also roughly similar in the magnitude (but not direction) of imbalance in FNRs.

Focusing on the debiasing candidates themselves, the proxy-eliminated algorithm departs the least from the benchmark comparisons, particularly the original or race omitted PCRA. This algorithm controls for the predictive variance that race shares with other risk factors, and manifests less imbalance in PPV ($h = 0.06$, higher for Blacks) than in FPRs ($h = 0.20$, higher for Blacks). The race-eliminated algorithm controls for all variance that race shares with other predictors, and manifests more imbalance in PPV ($h = 0.16$, higher for Blacks) than in FPRs ($h = -0.11$; lower for Blacks). The criminal history discount algorithm performs similarly to the race-eliminated algorithm. Together, this suggests that stakeholders who prioritize balance in PPV would opt for the proxy-eliminated candidate (or the race-omitted benchmark), whereas stakeholders who prioritize minimization of FPRs for Blacks would opt for the race-eliminated or criminal history discount algorithm.

3.4 | Illustrating algorithms' trade-offs and the effect of allowing high-risk thresholds to vary by race

The results in the previous section are consistent with Chouldechova's (2017) demonstration that algorithms can be tuned in alternative ways to prioritize fairness in PPV versus error rates. In many cases, Chouldechova (2017) notes, the best approach may be to allow unequal PPVs by group to achieve balance in error rates, which amounts to

TABLE 4 Balance in positive predictive value and error rates by race, for alternative Post Conviction Risk Assessment (PCRA) algorithms

	Positive predictive value				False positive rate				False negative rate			
PCRA Algorithm	**Black**	**White**	**% difference**	***h***	**Black**	**White**	**% difference**	***h***	**Black**	**White**	**% difference**	***h***
Benchmark comparisons												
Race-omitted	23.457	20.107	3.351	0.081	17.848	10.635	7.214	0.208	60.774	70.609	−9.835	−0.208
Race-fitted	26.263	20.349	5.914	0.140	12.445	9.996	2.449	0.078	67.75	71.955	−4.205	−0.092
Debiasing candidates												
Proxy-eliminated	24.279	21.604	2.675	0.064	16.526	9.796	6.730	0.201	61.448	70.928	−9.480	−0.201
Race-eliminated	25.616	19.108	6.508	0.157	11.666	15.317	−3.651	−0.107	71.19	60.269	10.921	0.231
Criminal history discount	25.784	19.263	6.521	0.156	11.832	15.678	−3.846	−0.112	70.521	58.924	11.597	0.243

Note: Positive predictive value is the percentage of people classified by the algorithm as high risk who were rearrested for a violent offense. False positive rate is the percentage of people who were not rearrested for a violent offense but had been classified as high risk by the algorithm. False negative rate is the percentage of people who were rearrested for a violent offense but had not been classified as high risk by the algorithm (who were instead classified as low, low-moderate, moderate, or moderate-high). % difference and *h* = difference in percentage points and Cohen's (2008) *h* effect size for difference between the Black and White groups, respectively.

allowing cutoff scores for risk categories to differ by racial group. Specifically, the cutoff score for classifying Black supervisees as high risk can be set higher than that of White supervisees. This can be done for any risk algorithm.

We applied this approach with each of the five algorithms, based on an adaptation of Kleinberg et al.'s (2018) conceptual framework, method, and graphic for summarizing the results. Conceptually, we distinguish between two types of policy planners. An "efficient planner" is interested only in the PPV of high-risk classifications (i.e., the predicted recidivism of the set of supervisees classified as high risk). An "equitable planner" is interested in both PPV and the racial composition of the high-risk group, with a preference for decreasing the number of supervisees in the high-risk group who are Black.

For a given algorithm, the efficient planner optimizes PPV simply by creating a rank-ordered list of supervisees in their predicted probability of violent reoffending and classifying the top *x*% as "high risk." For this illustration, we define *x* as 16% because this is roughly the proportion of supervisees classified as high risk by the original PCRA, across various algorithms. The equitable planner uses the same algorithm but creates separate lists of White and Black supervisees rank-ordered by their predicted probability of violent reoffending. Holding the cutoff for the high-risk White group constant (to that of the efficient planner), the equitable planner raises the cutoff for Black supervisees to decrease the proportion of the high-risk group that is Black. For this illustration, we focus on changing the racial proportion around an anchor of 50%, given that our sample is half Black and half White. We followed this approach for each of the five algorithms, first rank-ordering individuals by race on the algorithms' predicted probability of reoffending, and then by varying the cutoff for Black supervisees.

3.4.1 | Trade-offs with a single high-risk cutoff

Our results are summarized in Figure 1. In this figure, the *y*-axis is the proportion of the high-risk group that violently reoffended (overall PPV) and the *x*-axis is the proportion of the high-risk group that is Black. The shapes in the figure (i.e., the circle, square, triangle, diamond and *x*) plot the effect of each algorithm when the same cutoff is used for Black and White supervisees.

These shapes are most relevant to the efficient planner, who uses a single cutoff to define the high-risk group. The efficient planner is only interested in maximizing PPV, so would choose the shape with the highest location on the *y*-axis, which is the square representing the race-fitted algorithm. As shown by the square plotted in Figure 1, the high-risk group defined by this algorithm has a PPV of 22.6% and includes 67% Black supervisees. Given that race is legally prohibited as a predictor, though, the efficient planner would instead select the race-omitted or proxy-eliminated algorithm, and would lose little in PPV in doing so. As shown by the square's *y*-axis proximity to the circle and diamond in Figure 1, the race-omitted and proxy-eliminating algorithms have nearly the same PPV as the race-fitted algorithm. In part, this is probably because race is a relatively weak predictor of reoffending, compared with the PCRA items, and rarely moderates the predictive utility of PCRA items (see Skeem & Lowenkamp, 2016).

3.4.2 | Trade-offs with unequal high-risk cutoffs

The lines plotted in Figure 1 are most relevant to the equitable planner, who can use different cutoffs by race to define the high-risk group. The lines show the effect of each algorithm when the cutoff score is changed for Black supervisees to shift their representation in the high-risk group. The slight negative slope of the lines indicates that the PPV of high-risk classifications decreases as the proportion of the high-risk group that is Black increases. For a given algorithm, the higher the threshold used for Black supervisees, the larger the share of the high-risk group that is White, and the lower the share of this high-risk group that recidivates (in keeping with results reported in Table 3, where the PPV for Blacks is higher than Whites; this imbalance increases as the share of Black people in the high-risk group decreases).

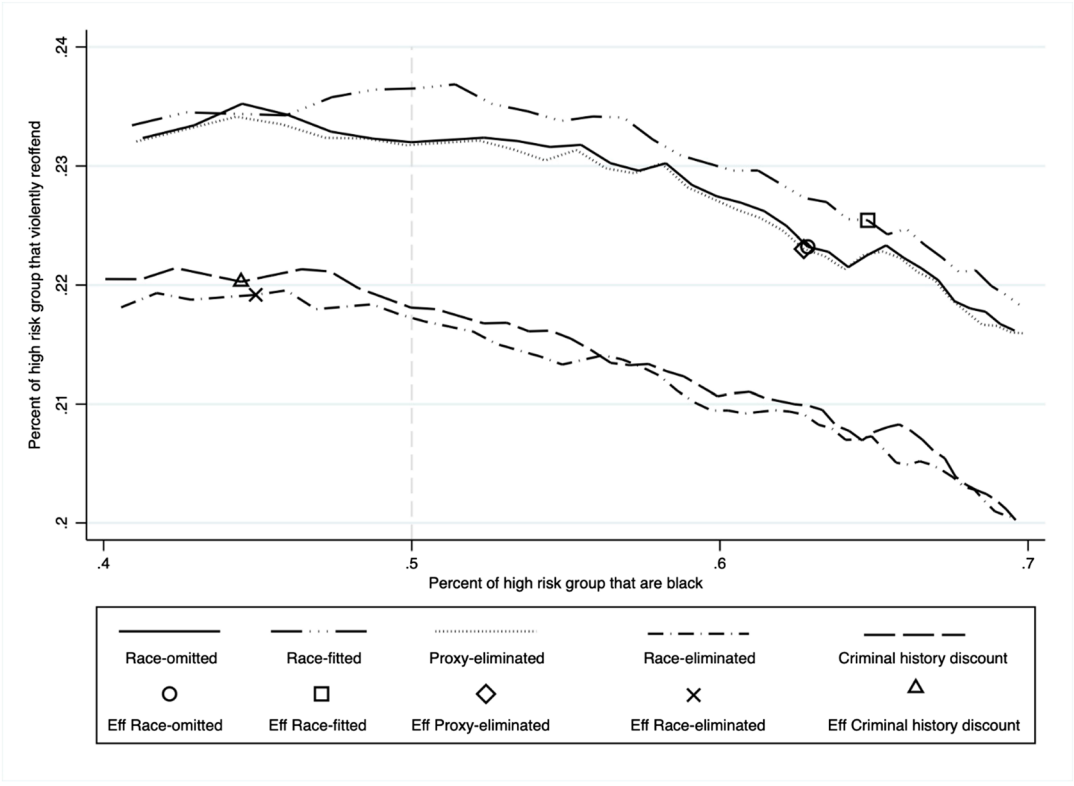

FIGURE 1 Value of high-risk classifications in predicting violence, as racial composition of high-risk group varies [Colour figure can be viewed at wileyonlinelibrary.com]
Note: "Eff"=efficient classifier. The shapes show point estimates of the effect of each algorithm on positive predictive value for the group classified as high risk (the top 16% of predicted probabilities), along with the proportion of the group that is black. The lines show the effect of shifting the proportion of the high risk group that is black (allowing the group size to vary) on the predictive value of the resulting high risk category. Results indicate that the race-fitted, race-omitted, and proxy-eliminated algorithms outperform the algorithms that are completely uncorrelated with race (i.e., race eliminated and criminal history discount algorithms)–in that for any given level of positive predictive value for high risk classifications, these three algorithms maintain higher predictive value while allowing for substantially decreasing the share of the high risk group that is black

The equitable planner is interested in maximizing PPV while minimizing the proportion of the high-risk group classified as Black. In the figure, this translates to the highest location on the *y*-axis with a location on the *x*-axis close to 50%. As shown in Figure 1, like the efficient planner, the equitable planner would select the race-fitted, race-omitted, or proxy-eliminated algorithm, each of which has greater predictive utility than the alternative candidates at any given level of diversity in the high-risk group, including the 50% mark. Using any of these three algorithms with a higher threshold for Black than White supervisees would outperform the two remaining algorithms designed to be unassociated with race. Setting the composition of the high-risk class at 50% Black (in keeping with their representation in this sample), the proportion that would violently reoffend for these three algorithms is > 23%, compared with < 22% for the race-orthogonal and criminal history discount algorithms. Although the percentage point difference appears small, it translates to many potentially averted violent crimes (see Kleinberg et al., 2018 for an analogous interpretation of a 1% difference in a college admissions policy context).

4 | DISCUSSION

In this study, we compared how alternative algorithms affect fairness trade-offs that are inherent in predicting violent reoffending when the rate of reoffending is unequal across racial groups. Our principal conclusion is that providing access to the protected characteristic of race can simultaneously maximize the calibration and positive predictive value of an algorithm (relevant to protecting community safety) while guarding against imbalanced error rates (relevant to achieving racial justice). For the equitable planner defined earlier, our results suggest that the best approach is to allow thresholds for defining the high-risk group to vary by race, using algorithms that are modestly associated with race (race-fitted, race-omitted, or proxy-omitted), rather than algorithms that effectively eliminate this association (race-eliminated and criminal history discount). Although Kleinberg et al. (2018) studied a different problem (fairness in college admission decisions), our results are broadly consistent with their conclusion: "across a wide range of estimation approaches, objective functions, and definitions of fairness, the strategy of blinding the algorithm to race inadvertently detracts from fairness" (p. 26).

The best approach empirically is not necessarily the best approach practically. It can be legally challenging, but not impossible, to base risk assessments on characteristics like race and gender (for reviews, see Goel et al., in press; Slobogin, 2018), at least when those assessments will be considered at sentencing. The algorithms we tested differ in their treatment of race and there is little legal guidance on the nuances among them. It is most clear that race cannot be used as a predictor to estimate an individual's risk of reoffending, which rules out the race-fitted algorithm. As the Supreme Court wrote in *Buck v. Davis* (2017), a death penalty case: "It would be patently unconstitutional for a state to argue that a defendant is liable to be a future danger because of his race." But it is possible that setting different risk thresholds by race is permissible, if doing so promotes predictive accuracy. The most relevant legal guidance on this issue focuses on gender (a characteristic subject to less legal scrutiny than race) and is provided by the Wisconsin Supreme Court in a sentencing case that involved the use of the COMPAS. In *Wisconsin v. Loomis* (2016), the court allowed for COMPAS risk estimates to differ by gender. In rejecting Loomis' claim of gender discrimination, the court reasoned that failure to consider gender would make risk scores less accurate and overestimate the risk that women pose. Similar logic applies in the case of race, when an algorithm manifests intercept bias that can be corrected with different thresholds. Notably, the court also wrote that using risk assessment to inform sentences was a "poor fit," but risk assessment could be used to inform a variety of other sanctioning decisions (e.g., diversion of low-risk people from prison; imposing terms and conditions of probation; *Wisconsin v. Loomis*, 2016). It seems that standards for the latter decisions would be less stringent.

Our secondary conclusion is that "debiasing" approaches like those we tested here should be used with caution, even though they are feasible to implement and could help policymakers realize the trade-offs in prediction that they decide are most acceptable. Although the three candidate algorithms we tested varied in their construction and in their association with race (which ranged from negligible to small), all were strongly predictive of violent reoffending (AUCv values = 71–0.72) and fairly well-calibrated by race (with no significant slope bias). In fact, by these metrics, the candidates performed no differently than the benchmarks. On the surface, this finding mitigates concern that removing the effect of race from an algorithm will eviscerate its predictive utility.

Beneath these interchangeable levels of overall predictive utility, however, lie trade-offs in balancing positive predictive value (PPV) versus false positive error rates. The debiasing candidate that remains modestly correlated with race – the race-aware with policy controls algorithm – is more balanced in PPV than the other candidates, but does not eliminate imbalance in error rates that disfavors Black participants, compared with the benchmarks algorithms. This debiasing candidate, in our view, is the least problematic of the three we tested (but see Gottfredson & Jarjoura's, 1996 cautions about choosing values for the policy control). By contrast, the debiasing candidates that eliminate a correlation with race – the race-eliminated and criminal history discount algorithms – show greater imbalance in PPV and completely reverse error rates that disfavor Black participants (to error rates that disfavor Whites). We advise against using these candidates because they misclassify many individuals' true risk levels for violence, with low-risk Whites classified as high risk, high-risk Blacks classified as

low risk, and both potentially harmed in the process (see Goel et al., in press). In a related sense, these candidates substantially increase intercept bias by race, necessitating different thresholds or risk score interpretations for Black and White participants.

The performance of the benchmark algorithms are worth mentioning. First, the race-omitted algorithm (akin to the original PCRA) performed about as well as the leading debiasing candidate (the race-aware with policy control algorithm). This may be because adding race to PCRA items in a predictive algorithm does little to increase predictive accuracy (Skeem & Lowenkamp, 2016) or to change the weights of the items (by removing predictive variance that race shares with these items, or proxy effects). Second, the race-fitted algorithm performed well: perhaps counterintuitively, this algorithm achieved the greatest racial balance in error rates.

Goel et al. (in press, p. 10) observe that practitioners have long designed tools that adhere to a pragmatic fairness concept: "after constructing risk scores that best capture individual-level risk – and potentially including protected traits to do so – similarly risky individuals are treated similarly, regardless of group membership." This fairness concept amounts to a thresholding approach where practitioners select an acceptable risk level (i.e., predicted probability of offending) for a given decision (e.g., pretrial release) and then apply that threshold to people regardless of their group membership. This approach roughly corresponds to the approach of the efficient planner, described earlier. This pragmatic concept of fairness seems compatible with both legal standards of equity and intuitive notions of social welfare.

But in the criminal justice field, practitioners routinely omit race in developing risk assessment instruments. If the goal is to accurately estimate risk and achieve a pragmatic view of fairness, omitting race without scrutiny is not an option. Accurate risk assessments may not always depend on the inclusion of race. When they do, the legal considerations outlined earlier must be considered, and algorithmic strategies like including race with "policy controls" explored. in press

This study illustrates the effect of candidate "debiasing" algorithms on inherent trade-offs between prediction-focused accuracy rates (relevant to crime prevention) and outcome-focused error rates (relevant to racial justice), when racial groups differ in their reoffending rates. As noted at the outset, algorithms quantify these trade-offs but cannot make choices among them – it is up to policymakers to weigh legal, ethical, and value considerations and choose the trade-off that is most acceptable to them (see Kleinberg et al., 2019). If the pragmatic approach outlined here is followed, imbalance in error rates will follow. Some policymakers may prefer different trade-offs (e.g., minimizing imbalanced false positive rates at the peril of misclassification). Algorithms clearly illustrate the consequences of various options. This is a principal point made by Kleinberg et al. (2018), in their demonstration that well-made and well-regulated algorithms create new transparency and opportunities to detect and address discrimination that are not available, when unaided human judgment is used to assess risk (often implicitly) and make a decision.

Trade-offs inherent in prediction are not avoided by abandoning risk assessment (Corbett-Davies et al., 2016). When rates of reoffending are imbalanced and human judgment is well-calibrated to those true outcomes (whether unaided or structured), the imbalance in false positive rates that inspired ProPublica's critique will materialize. Risk assessments are demonstrably more consistent, transparent, and accurate than unaided human judgment (see the Introduction). The development of algorithms can be regulated to ensure fairness (Kleinberg et al., 2019), as can the implementation of risk assessment instruments (see Garrett & Monahan, in press). Risk is a relevant consideration for a variety of decisions that humans make daily in the criminal justice system, including decisions that implicate the need for community safety and the need to prioritize scarce correctional treatment services to maximize people's likelihood of successful re-entry. Humans who make these decisions are typically aware of the person's race and are subject to forms of racial bias that algorithms are not (Corbett-Davies et al., 2016; Kleinberg et al., 2019). As Goel et al. (in press; p. 15) observe: "all expert testimony – including non-actuarial prediction testimony – is ultimately based on assumptions about the kind of person an offender is, as is the judge's ultimate determination of risk; the key difference, and one that should count as an advantage, is that the instrument displays its stereotyping assumptions on its face."

To our knowledge, only one study has examined the actual impact of risk assessment on racial disparities in the criminal justice system. Stevenson (2018) examined the use of pretrial risk assessment in Kentucky and found that risk assessment was associated with an increase in the proportion of defendants released who were low risk, but not with an increase in racial disparities.

This study is rare. Debate about the use of risk assessment instruments to inform sanctioning decisions has focused almost exclusively on dissecting algorithms, in isolation. These analyses – the present study included – cannot answer the most pressing policy question. Whether the use of risk assessment exacerbates, ameliorates or has no effect on racial disparities is a relative inquiry: risk assessment compared with what existing practices? The comparison typically involves legal professionals' hunches about risk and can involve sentencing guidelines that rely heavily on criminal history, the one variable that largely explains racial disparity in incarceration (Frase, 2009). These comparisons make the relative advantages of well-made and well-regulated risk assessment more apparent. It is possible that using algorithms of the type studied here will reduce racial disparities, compared with a regime that rejects risk entirely and relies on human judgment about deserved punishment.

REFERENCES

American Educational Research Association, American Psychological Association, and National Council on Measurement in Education. (2014). *The standards for educational and psychological testing*. Washington, DC: AERA Publications.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals, and its biased against blacks. Propublica. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Beck, A. J., & Blumstein, A. (2018). Racial disproportionality in U.S. state prisons: Accounting for the effects of racial and ethnic differences in criminal involvement, arrests, sentencing, and time served. *Journal of Quantitative Criminology*, *34*(3), 853–883. https://doi.org/10.1007/s10940-017-9357-6

Blumstein, A. (2011). Crime and rates of incarceration in the U.S. In J. Dvoskin, J. Skeem, R. Novaco, & K. Douglas (Eds.), *Using Social Science to Reduce Violent Offending*. New York: Oxford University Press.

Buck v. Davis, 137 S.Ct. 759 (2017)

Carter, M., & Shames, A. (2020). *A statement from Advancing Pretrial Policy and Research*. Center for Effective Public Policy. https://cepp.com/a-statement-from-advancing-pretrial-policy-and-research-appr/

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, *5*(2), 153–163. https://doi.org/10.1089/big.2016.0047

Cleary, T. A. (1968). Test bias: Prediction of grades of Negro and white students in integrated colleges. *Journal of Educational Measurement*, *5*, 115–124.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences* (2nd ed.). New Jersey: Lawrence Erlbaum.

Cohen, T. H., Lowenkamp, C. T., & VanBenschoten, S. W. (2016). Does change in risk matter? Examining whether changes in offender risk characteristics influence recidivism outcomes. *Criminology & Public Policy*, *15*(2), 263–296. https://doi.org/10.1111/1745-9133.12190

Cook, D. E. (2015). CCMATCH: Stata module to randomly match cases and controlsbased on specified criteria. Version 1.3. www.Danielecook.com

Corbett-Davies, S., Pierson, E., Feller, A., & Goel, S (2016). A computer program used for bail and sentencing decisions was labeled biased against blacks. It's actually not that clear. Washington Post. https://www.washingtonpost.com/news/monkey-cage/wp/2016/10/17/can-an-algorithm-be-racist-our-analysis-is-more-cautious-than-propublicas/

Courtland, R. (2018). Bias detectives: The researchers striving to make algorithms fair. *Nature*, *558*, 357–360. https://doi.org/10.1038/d41586-018-05469-3

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc. https://go.volarisgroup.com/rs/430-MBX-989/images/ProPublica_Commentary_Final_070616.pdf

Durose, M. R., Cooper, A. D., & Snyder, H. N. (2014). *Recidivism of prisoners released in 30 states in 2005: Patterns from 2005 to 2010*. Washington, DC: US Department of Justice, Office of Justice Programs, Bureau of Justice Statistics.

Fisher, F. M., & Kadane, J. B. (1983). Empirically based sentencing guidelines and ethical considerations. In A. Blumstein, J. Cohen, S. Martin, & M. Tonry (Eds.), *Research on sentencing: The search for reform* (pp. 184–193). Washington, DC: National Academy of Sciences.

Flores, A., Bechtel, K., & Lowenkamp, C. (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.". *Federal Probation*, *80*, 38–46.

Frase, R. (2009). What Explains Persistent Racial Disproportionality in Minnesota's Prison and Jail Populations? *Crime and Justice*, *38*, 201–280.

Garrett, B. (2018). The prison reform bill's implementation will be tricky: Here's how to ensure it's a success. Slate. Retrieved from. https://slate.com/news-and-politics/2018/12/prison-reformbill-success.html

Garrett, B., & Monahan, J. (in press). Judging risk. *California Law Review*..

Goel, S., Shroff, R., Skeem, J., & Slobogin, C. (in press). The accuracy, equity, and jurisprudence of criminal risk assessment. In R. Vogel (Ed.), *Research Handbook on Big Data Law*. Cheltenham, UK: Edward Elgar.

Gottfredson, D. M. (1999). *Effects of judges' sentencing decisions on criminal careers*. US Department of Justice, Office of Justice Programs, National Institute of Justice. Available at:. https://www.ncjrs.gov/pdffiles1/nij/178889.pdf, https://doi.org/10.1037/e513192006-001

Gottfredson, S. D., & Gottfredson, D. M. (1985). Screening for risk among parolees: Policy, practice, and method. In D. Farrington, & R. Tarling (Eds.), *Predicting Crime and Delinquency* (pp. 54–77). Albany: State University of New York Press.

Gottfredson, S. D., Gottfredson, D. M., & Gottfredson, M. R. (2000). Risk measures for operational use: Removing invidious predictors. In D. M. Gottfredson's (Ed.), *Juvenile Justice with Eyes Open*. National Center for Juvenile Justice: Pittsburgh, PA.

Gottfredson, S. D., & Jarjoura, G. R. (1996). Race, gender, and guidelines-based decision making. *Journal of Research in Crime and Delinquency*, *33*(1), 49–69. https://doi.org/10.1177/0022427896033001004

Guitterez, R., & Drukker, D. (2007). Stata's cluster-correlated robust variance estimates. Retrieved from http://www.stata.com/support/faqs/statistics/references/

Holder, E. (2014). Attorney General Eric Holder Speaks at the National Association of Criminal Defense Lawyers 57th Annual Meeting. Available at: http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

Johnson, J. L., Lowenkamp, C. T., VanBenschoten, S. W., & Robinson, C. R. (2011). The construction and validation of the Federal Post Conviction Risk Assessment (PCRA). *Federal Probation*, *75*, 16–29.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. (2017). Human decisions and machine predictions. *Quarterly Journal of Economics*, *133*, 237–293. https://doi.org/10.3386/w23180

Kleinberg, J., Ludwig, J., Mullainathan, S., & Rambachan, A. (2018). Algorithmic Fairness. In *AEA Papers and Proceedings* (Vol. 108) (pp. 22–27). Pittsburgh, PA: American Economic Association.

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in The Age Of Algorithms (NBER Working Paper No. 25548). Retrieved from National Bureau of Economic Research website: https://www.nber.org/papers/w25548

Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.

Kochel, T. R., Wilson, D. B., & Matrofski, S. D. (2011). Effect of suspect race on officers arrest decisions. *Criminology*, *49*, 473–512. https://doi.org/10.1111/j.1745-9125.2011.00230.x

Lin, M., Lucas, H. C. Jr., & Shmueli, G. (2013). Research commentary-too big to fail: large samples and the p-value problem. *Information Systems Research*, *24*(4), 906–917.

Lin, Z., Jung, J., Goel, S., & Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, *6*, 1–8. https://doi.org/10.1126/sciadv.aaz0652

Lowenkamp, C. T., Holsinger, A. M., & Cohen, T. H. (2015). PCRA revisited: Testing the validity of the federal Post Conviction Risk Assessment (PCRA). *Psychological Services*, *12*, 149–157. https://doi.org/10.1037/ser0000024

Lowenkamp, C. T., Johnson, J. L., Holsinger, A. M., VanBenschoten, S. W., & Robinson, C. R. (2013). The Federal Post Conviction Risk Assessment (PCRA): A construction and validation study. *Psychological Services*, *10*, 87–96. https://doi.org/10.1037/a0030343

Monahan, J. (2017). In E. Luna (Ed.), *Risk assessment in sentencing*. Reforming Criminal Justice. Retrieved from:. http://academyforjustice.org/volume4/

Monahan, J., & Skeem, J. L. (2014). Risk Redux: The Resurgence of Risk Assessment in Criminal Sanctioning. *Federal Sentencing Reporter*, *26*(3), 158–166.

Neufeld, A. (2018). In defense of risk assessment tools. In *The Marshall Project*. Retrieved from. https://www.themarshallproject.org/2017/10/22/in-defense-of-risk-assessment-tools

O'Neill, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. New York: Crown Publishers.

Piquero, A., & Brame, B. (2008). Assessing the Race-Crime and Ethnicity-Crime Relationship in a Sample of Serious Adolescent Delinquents. *Crime & Delinquency*, *54*(3), 390–422. https://doi.org/10.1177/0011128707307219

Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, *3*(3), 206–231.

Rice, M., & Harris, G. (2005). Comparing effect sizes in follow-up studies: ROC Area, Cohen's d, and r. *Law and Human Behavior*, *29*, 615–620. https://doi.org/10.1007/s10979-005-6832-7

Rogers, W. H. (1993). Regression standard errors in clustered samples. *Stata Technical Bulletin*, *13*, 19–23.

Sackett, P. R., & Bobko, P. (2010). Conceptual and technical issues in conducting and interpreting differential prediction analyses. *Industrial and Organizational Psychology*, *3*, 213–217.

Skeem, J. L., & Lowenkamp, C. T. (2016). Risk, race, and recidivism: Predictive bias and disparate impact. *Criminology*, *54*, 680–712. https://doi.org/10.1111/1745-9125.12123

Skeem, J.L., & Polaschek, D.L.L. (in press). High risk, not hopeless: Correctional intervention for people at high risk for violence. Marquette Law Review.

Slobogin, C. (2018). Principles of risk assessment: Sentencing and policing. *Ohio State Journal of Criminal Law*, *15*, 583–596.

Starr, S. B. (2014). Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, *66*, 842–873.

Stevenson, M. (2018). Assessing risk assessment in action. In *Minnesota Law Review* (Vol. 103) (pp. 303–384). Pretrial Justice Institute (2020). Updated position on pretrial risk assessment tools. https://www.pretrial.org/wp-content/uploads/Risk-Statement-PJI-2020.pdf

Walker, S., Spohn, C., & DeLone, M. (2011). In C. Learning (Ed.), *The Color of Justice: Race, Ethnicity, and Crime in America, 5th*. Belmont, CA: Wadsworth.

Wisconsin v. Loomis, 881 N.W.2d 749 (Wisc. 2016).

How to cite this article: Skeem J, Lowenkamp C. Using algorithms to address trade-offs inherent in predicting recidivism. *Behav Sci Law*. 2020;38:259–278. https://doi.org/10.1002/bsl.2465