Validation and Assessment of Pennsylvania's Risk Assessment Instrument

Pennsylvania Commission on Sentencing

Heinz College System Synthesis Project

May 2019

David Mitre Becerril, Chris Bell, Katie LeFevre, Laurel Lin, Wilson Mui, and Karan Shah | Matt Hannigan, Advisor

Carnegie Mellon University
HeinzCollege
INFORMATION SYSTEMS • PUBLIC POLICY • MANAGEMENT

1. Executive Summary

Since 2011, the Pennsylvania Commission on Sentencing (PCS) has been developing a risk assessment instrument for use during sentencing, as required by the state legislature. In its current form, the risk assessment instrument would be used to identify offenders with a risk profile that is lower or higher than that of the typical offender. If classified as either, the PCS recommends that the court prepare a pre-sentence investigation report (PSI) to provide additional information on the offender. Since the offender has a non-typical risk profile, the additional information can assist the court in determining an appropriate sentence.

This report covers the work of the Heinz System Synthesis team to validate and assess the PCS Sentence Risk Assessment Instrument (SRAI). Since its creation, no outside, independent entity has validated or assessed the PCS logistic regression or risk assessment instrument. As such, the project scope included replicating the logistic regression used for variable selection, the general SRAI, and the Crime Against a Person (CAP) SRAI to ensure the PCS's process was error free. The project team's replicated regression and instruments matched nearly exactly and shows that the creation process of the SRAI was robust.

To assess the performance and fairness of the general and CAP SRAIs, Area Under the Curve of the Receiver Operating Curve (AUC) and 14 additional metrics were calculated. Almost all risk instruments publish their assessments with AUC scores, and, along this measure, both instruments perform moderately well. The additional performance and fairness metrics told a different story for both instruments. For this project, fairness means that different subpopulation groups are treated similarly.The general SRAI's performance was comparable to other risk assessment instruments but predicted that Black offenders were more likely to recidivate than White offenders. The CAP SRAI, however, performed very poorly along both traditional and new metrics. As such, the team's assessment is that while the general SRAI could be implemented to classify low-risk offenders, the CAP SRAI should not be deployed.

Next, attempts were made to improve the general SRAI by adjusting the high- and low-risk cutoff range on the risk scale, changing the weights applied to each risk category, and removing gender as a risk factor in the instrument. On request from the PCS, the team focused on assessment and optimization of the instrument rather than experimenting with alternative risk categories. To improve the general SRAI, the team

recommends that the high-risk cutoff be increased by at least two points, at which point there are statistically significant increases on accuracy and fairness.

Finally, the team explored the implications of removing gender as a feature from the general SRAI. Preliminary results indicate that the performance, fairness, and accuracy of the instrument either stay the same or improve. Notably, the fairness metrics between male and female offenders improve substantially. Similar to the modification of the general SRAI, the team also recommends increasing the high-risk cutoff to ten, for greater accuracy and fairness without substantially reducing performance. While further analysis is needed to determine the exact implications of removing gender, the general SRAI appears to perform well without it.

2. Introduction

Risk assessment instruments are used throughout the criminal justice system to aid judicial decision making. However, their use can generate considerable public concern regarding accountability and fairness. If proper care is not taken, risk assessment instruments could perpetuate any systemic biases already in the criminal justice process. Therefore, any risk assessment instrument and its underlying data must be thoroughly tested to ensure predictive accuracy treats all offenders fairly.

The Pennsylvania Commission on Sentencing (PCS) has been developing a sentence risk assessment instrument (SRAI) in response to a 2010 state legislative mandate. The SRAI is intended to assist Pennsylvania judges during the sentencing process by identifying offenders who are above or below typical risk profiles with respect to recidivism, and thus may warrant the use of a presentence investigation report (PSI) for a fully informed decision before imposing a sentence.

To date, the SRAI has not been evaluated or examined by an independent entity. A team of graduate students from Carnegie Mellon University's Heinz College was tasked with reviewing and assessing the predictive validity of the SRAI and identifying any biases based on offender demographics. Assessments are conducted on both risk scales of the SRAIs developed by the PCS. The general recidivism risk scale classifies offenders on their risk of recidivating all crime (general SRAI), and the crime against a person risk scale classifies offenders on their risk of recidivating a crime against a person (CAP SRAI). Potential modifications to improve the performance and accuracy of the general SRAI are also explored. Finally, recommendations are provided on the usage and improvements of both the general SRAI and the CAP SRAI.

3. Pennsylvania Risk Assessment Instrument

Purpose of Risk Assessment Tool

The PCS was established in 1973 with the goal to create more consistent, uniform, and fair statewide sentencing practices. In 2008, the Commonwealth of Pennsylvania's legislature required the PCS to adopt new guidelines for parole and resentencing, and in 2010 the legislature passed Act 95 which mandated the development of a validated risk assessment tool.[1] Act 95 defined the following requirements:

- Adopt a risk assessment instrument to be used at sentencing,
- Consider the risk of re-offense and threat to public safety,
- Help determine if the offender is a candidate for alternative sentencing programs, and
- Develop an empirically based worksheet using factors predicting recidivism.

The PCS started its work in the summer of 2010 with Phase I, and continues through today in Phase III. The SRAI is intended to be used after an opening plea or trial. Judges are given a Sentence Risk Assessment Summary report, which displays the risk categories and possible risk points, the offender's risk points, and the risk score scale. An offender is classified as low, typical, or high risk, with low- and high-risk classifications made relative to the typical Pennsylvania offender. If an offender is classified as low or high risk, the PCS recommends the court obtain a PSI to provide additional information. The instrument does not dictate that a PSI be pulled nor recommend any specific sentencing types or lengths.[2]

Progress of the PCS

Since 2011, the PCS has released 18 reports covering three phases of the development of the risk assessment instrument.[3] Phase I began with the search for predictive factors of recidivism through bivariate and multivariate regression; selection of the Burgess method as a classification framework; validation of the results on different samples; analysis of the impact that risk assessment scores may have on low-risk offenders; and research on best practices to communicate assessment risk scores to judges.

[1] Interim Report 8, Communicating Risk at Sentencing.
[2] Risk Assessment Project Phase III The Development and Validation of the Proposed Risk Assessment Scales.
[3] Full reports available at http://pcs.la.psu.edu/publications-and-research/risk-assessment.

In Phase II, separate risk assessment instruments were designed and validated for offenders at each Offense Gravity Score (OGS) level. The same factors of recidivism were used in this analysis for accuracy and simplicity, even though not every factor was statistically significant for each OGS-specific risk instrument.

In Phase III, the SRAI was split into two risk instruments scales: one to classify general recidivism cases and one to classify crime against a person recidivism cases. In addition, the PCS conducted a racial impact report and switched to using convictions rather than arrests for both input factors and outcomes based on public feedback. The PCS risk assessment instrument is currently in Phase III.

Predicting Recidivism

The PCS chose the unweighted Burgess method as a classification method to predict whether an offender is likely to recidivate in the future. With the unweighted Burgess method, factors related to the outcome variable of recidivism are selected after analysis of a multivariate regression. Factors found to be predictive are assigned unit weights. The weights of all the factors are then summed to create a risk score. Cutoffs for low and high risk were then chosen to identify offenders with risk scores that were one standard deviation below and above the mean risk score. The unweighted Burgess method's strength lies in its simplicity but its performance is susceptible from interrelationships between the predictors.[4]

Selecting Predictors

In Phase I, the PCS researched and selected the primary variables related to recidivism. Potential variables were discovered through a literature review of criminology research and other risk assessment instruments. A bivariate analysis between each variable and recidivism was conducted to select the most significant features related to recidivism. The chosen risk categories are: Age, Number of Prior Arrests, Gender, Prior Offense Type, Multiple Current Convictions, County, Prior Juvenile Adjudication, Prior Record Score, and Current Offense Type.

During Phases II and III, additional policy considerations altered the variables included. County was removed, multiple current convictions were altered to only count those in

[4] Don Gottfredson and Howard Snyder, "The Mathematics of Risk Classification: Changing Data into Valid Instruments for Juvenile Courts," 2005, www.ojp.usdoj.gov/ojjdp.

the judicial proceeding (JP), prior record score was dropped, and arrests were changed to convictions.

In the current versions of the general SRAI and CAP SRAI, the following risk categories are used: Gender, Age, Current Conviction Offense Type, Multiple Current Convictions in JP, Number of Prior Convictions, Prior Conviction Offense Type, and Prior Juvenile Adjudication.

Creating a Risk Scale

After the predictors of recidivism were selected, the PCS assigned weights, or risk points, to each predictor. Gender, Multiple Current Convictions in JP, and Prior Juvenile Adjudication are binary (0 or 1), while the other factors assign different risk points depending on the offender's characteristic. For example, current age is separated into five groups, with offenders younger than twenty-one receiving five risk points and offenders older than forty-nine receiving zero. Current Conviction Offense Type is weighted by the severity of the offense, and Number of Prior Convictions is weighted by the number of prior convictions, separated by buckets. Prior Conviction Offense Type is binary between one or negative one, with negative one assigned for a prior firearm or weapon conviction.

The risk points are summed and compared against a risk scale, which ranges from 0 to 18 and in order of increasing perceived risk. For the general SRAI, the PCS chose these cutoffs: 0 to 4 points is low risk, 5 to 9 points is typical risk, and 10 to 18 is high risk. These cutoffs were chosen so that the typical risk range covers the one standard deviation above and below the average risk score. For the CAP SRAI, the risk scale ranges from 0 to 17, with low-risk classified as a score between 0 and 3, typical risk between 4 and 8, and high risk above 9.

4. History of Risk Assessments and Decision Making

Errors in Human Decision Making

Human decision making is inherently flawed. Clinical judgments or actuarial tools are common solutions to address decision errors and inform predictions. While clinical judgments are based on professional experience, such as the opinion of a psychologist, actuarial tools are empirically based.

Since humans are subject to decision errors and biases, actuarial predictions are more accurate because humans "make different decisions at different times about the same problem."[5] Krauss et al. (2004) argues that "in most decision-making situations … actuarially developed predictions outperform human judgments."[6] He summarizes the common cognitive errors in individual decision making, or predictive judgments, as follows:

- Failing to consider the normal rate at which an event is likely to occur.
- Combining and considering "factors in a way that is subjectively appealing rather than empirically derived."
- "Failing to incorporate the notion that outliers have a tendency to subsequently return toward the mean of sample."
- "The tendency to make decisions or judge information in a manner that fits our preconceived categories or stereotypes of a situation."
- "Placing excessive weight on vivid or easily retrievable information."[7]

Actuarial tools improve human decision making. One of the biggest advantages of algorithms is its consistency. Since someone with the same circumstances and offenses will be treated the same by an algorithm, it removes the problems of human decision making and bias.

Using an algorithm removes some of the flaws in human judgment but it can also introduce new flaws. To better understand bias in predictive tools, researchers developed fairness metrics to help quantify the level of bias in algorithmic outcomes. Beyond measures of accuracy and false positive rates used to assess fairness, there

[5] Ruback et al., "Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis," page 47.
[6] Krauss, "Adjusting Risk of Recidivism: Do Judicial Departures Worsen or Improve Recidivism Prediction under the Federal Sentencing Guidelines?" page 732.
[7] Ibid., page 736.

are other metrics such as statistical or demographic parity, conditional procedure accuracy equality, conditional use accuracy or predictive value equality, and treatment equality. If there is no difference between the fairness metrics of a subpopulation, by gender or race for example, then total fairness has been achieved.

The accuracy of judges is difficult to measure because there is no benchmark for the accuracy of judges to compare to the accuracy of the algorithm. Kleinberg et al. (2018) created an algorithm to assess the risk of flight from a dataset of New York City cases and then compared the algorithm's assessment of the offenders to the judges in order to determine whether algorithm predictions are better than judges decisions. The authors argue that the algorithm "does no worse than the judges (and typically much better) on the outcome."[8] Kleinberg et al. (2018) found that judges struggled with high-risk cases and release many high-risk offenders. "The riskiest 1% of defendants have a predicted risk of 62.6% yet are released at a rate of 48.5% rate."[9] The authors found that high-risk offenders were more likely to be released if their current charge was minor and that judges were more likely to detain a low-risk offender if their current charge was more serious. They concluded that judges overweight the importance of the current charge.[10]

Further complicating the use of risk assessments is that the risk tolerance of judges is unknown. Which means that the crime risk level judges are willing to accept is unknown and different between judges. This is part of why judges' decisions are noisy.

Kleinberg et al. (2018) also suggest that judges could be mispredicting a defendant's ability to pay when setting bail, rather than considering the risk. They caution that in prediction, "biases arise when omitted variables correlate with the payoffs."[11] In contrast, causal inference is biased when omitted variables are correlated with outcomes.

If the algorithm outperforms the decision maker in one single dimension, it does not mean that the decision maker is automatically incorrect, or that algorithms can improve their decision making. The authors also recognize limitations of algorithms in that judges' decisions may be informed by variables unseen by the algorithm, such as tattoos, or injuries. In addition, when evaluating risk assessments, the authors suggest considering the decision makers' goals, because decisions that appear to be bad may simply reflect different goals. In the context of criminal justice, it could be that many

[8] Kleinberg et al., "Human Decisions and Machine Predictions."
[9] Ibid., page 13.
[10] Ibid., page 27.
[11] Ibid., page 5.

offenders are released, including some high-risk offenders, to reduce the number of people incarcerated.

History of Risk Assessment Instruments

Assessing an individual's future risk to society has played a role in the criminal justice system for over a century. Early assessments were based on an individual person or groups' professional judgment. The first parole and probation boards rose to prominence in the late nineteenth century and intermediate sentencing, which relied on a judge to determine release, during the early twentieth century. In the mid-twentieth century, the theory of sentencing moved towards the rehabilitation of offenders.[12] Judges were given greater freedom to shape sentences directly to the individual offender's circumstances, acknowledging that each person's history and situation is different. However, this switch also allowed judges to exert greater influence on the amount of bias in the criminal justice system.

Risk assessments in sentencing fell out of prominence with the return to a retribution theory of justice during the 1970s and early 1980s. Instead of conferring a punishment based on future risk or rehabilitation, sentences were based on an offender's 'blameworthiness' of past deeds. To address the supposed bias in sentencing, lawmakers sought to take some sentencing power away from judges through standardizing sentencing practices with the Sentencing Reform Act in 1984.[13] Two years later, the Anti-Drug Abuse Act of 1986 set mandatory minimums for many drug-related crimes, further creating standardization in sentencing. In the following decades, these actions helped create the mass incarceration phenomena in the United States.

In response to the mass incarceration of Americans, which vastly increased prison costs and destroyed communities, evidenced-based practices gained prominence during the 1990s. This movement revitalized the interest in actuarial risk assessments, which used large data sets to determine factors related to recidivism. Through classifying offenders according to risk, limited rehabilitation and imprisonment resources could be more efficiently allocated.

---

[12] John Monahan and Jennifer L. Skeem, "Risk Assessment in Criminal Sentencing," *Ssrn*, 2016, https://doi.org/10.1146/annurev-clinpsy-021815-092945.

[13] Danielle Kehl, Priscilla Guo, and Samuel Kessler, "Algorithms in the Criminal Justice System: Assessing the Use of Risk Assessments in Sentencing," *Responsive Communities.*, 2017, 1–36, https://cyber.harvard.edu/publications/2017/07/Algorithms.

Today, risk assessment instruments are used at a variety of phases in the criminal justice system. The most general are risk-needs assessments (RNA), which are used during sentencing or in prison rehabilitation programs. Another common use of RNAs is in pre-trial detention or release to determine whether an offender needs to be held in jail before trial. Finally, multiple instruments are used in sentencing to help inform judges of an offender's risk of recidivism. While rarely used to determine the exact sentence an offender should receive, they are often designed to identify offenders who may be candidates for alternative treatment.[14]

Risk Assessment Instrument Generations

Risk assessment instruments can be generally classified into four generations, with each generation growing in sophistication. First generation risk assessment tools are professional judgment, described above, which is based solely on the presiding judge's opinion. In the evidence-based movement, judges were provided information on both dynamic and static risk factors. Dynamic risk factors are those that can change over time, such as employment status and drug use. Static factors are factors that are immutable, such as sex and criminal history.

Second generation risk assessment instruments are primarily actuarial tools that largely abandoned many dynamic factors in favor of solely static factors. At this stage, dynamic factors may not accurately reflect the progress an offender has made over time. Second generation instruments are the first to take a list of risk factors and translate them into a risk scale, ascribing a numeric risk score to each offender.

Third generation risk assessment instruments build on the shortcomings of only using static factors by analyzing the interplay between static and dynamic factors. They may also rely on structured interviews to gather assessment data. Unlike second generation instruments, third generation instruments attempt to balance data-driven decisions with criminology theory and research that recognize offenders' behavior may change over time.

Fourth generation risk assessment instruments expand upon third generation instruments by integrating case management and risk reduction techniques into the sentencing process. These instruments aim to move past simply assessing risk and instead towards rehabilitation and supervision.[15]

[14] Sarah L Desmarais and Jay P Singh, "Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States: An Empirical Guide," 2013.
[15] Desmarais and Singh.

Risk Assessment Instruments in Other States

Risk assessments, or risk and needs assessments, are used in other states to inform different aspects and stages of the criminal justice system, such as pre-trial decisions, probation, and sentencing.[16] Most states use already existing assessments such as ORAS, PSA, COMPAS, and LSI-R, but a few states (Alaska, Missouri, North Carolina, Ohio, Pennsylvania, and Virginia) have developed their own state-wide custom assessments.[17] Sometimes an assessment is used in a jurisdiction below the state level, such as a county.

Validating the development and results of the instrument and evaluating it for errors or bias is part of best practices. Ideally, the evaluation would be done by a third-party. Validation is the process of checking the accuracy of something. In any system, validating results helps ensure the process is working as intended. Checking the accuracy or validity of the instrument is important in order to assess how well it is working. Especially in the context of sentencing, the potential impact of the instrument requires validation. The project team should be able to independently replicate the results of the PCS. If the commission's results are replicable, the credibility of the instrument creation process is strengthened.

In other states' validations, Area Under the Curve of the Receiver Operating Curve (AUC) was commonly used to test the validity of risk assessment instruments. AUC measures the instrument's ability to separate the persons between classes. This is done by plotting the True Positive Rate against the False Positive Rate.

Previous work has classified the following range of AUC values; a weak AUC = 0.55, a moderate AUC = 0.63, and a strong AUC = 0.71. These cutoffs were used in the University of Cincinnati's validation of ORAS.[18] Based on these values, the PCS SRAI has a moderate AUC of 0.66. This means that there is a 66% chance that the SRAI will be able to identify high-risk offenders from low-risk offenders. Table 4.1 shows the AUC scores for other instruments.

---

[16] Douglas et al., "Risk assessment tools in criminal justice and forensic psychiatry: The need for better data" page 134.
[17] Cravez, "Pretrial Risk Assessment Tool Developed for Alaska," page 2; and Electronic Privacy Information Center, "EPIC - Algorithms in the Criminal Justice System."
[18] Latessa et al., The Ohio Risk Assessment System (ORAS), report, Corrections Institute, University of Cincinnati, October 31, 2017, page 28; and State of Alabama, "Alabama Department of Corrections Minimum Standards for Community Punishment and Corrections Programs" 2016, page 17 and 28.

Table 4.1 Other Instruments' AUC Scores

State	Instrument	AUC Score
PA	PCS	0.66
VA	VPRAI	0.66
OH	ORAS	0.61
Various	LSI-R	0.64
Various	COMPAS	0.67
Various	PSA	0.65

Both Ohio and Virginia developed instruments custom to their state. While the majority of validation studies only reported overall AUC scores, the validation studies for Ohio's and Virginia's risk instruments included AUC scores by subpopulation. The remainder of this section will outline when Virginia and Ohio use their assessments, how they were developed, and information about validation.

Ohio

Ohio's Department of Rehabilitation and Correction contracted with the University of Cincinnati, Center for Criminal Justice Research to develop the Ohio Risk Assessment System (ORAS).[19] ORAS is a fourth generation instrument used by at least four other states (Alabama, Indiana, Massachusetts, and Texas).[20] It has tools which are similar to the PCS SRAIs in that it uses the Burgess method to classify risk categories.

The ORAS is comprised of six assessment tools, each for different phases of the criminal justice system. The Community Supervision Tool (CST) and the Community Supervision Screening Tool (CSST) are the most comparable to the PCS in the criminal justice process. The outcome variable these instruments generally assess is risk of arrest for new crimes, but the pre-trial assessment tool also assesses the risk of failure-to-appear.

The instruments examine both static and dynamic factors. To identify risk factors for some of the instruments, Ohio offenders were interviewed and monitored for one year to

[19] "Ohio Risk Assessment System," ODRC.
[20] Latessa et al., The Ohio Risk Assessment System (ORAS), report; and State of Alabama, "Alabama Department of Corrections Minimum Standards for Community Punishment and Corrections Programs," 17.

measure recidivism. Other data on recidivism was also used.[21] Information from the interviews and surveys of those who recidivate informed the instruments. Both of those components combine to generate a score for each offender. That score is then categorized into either three or four groups. Each instrument uses different factors, and the CST includes factors related to criminal history; education employment, and financial situation; family and social support; neighborhood problems, substance use; peer associations; criminal attitudes and behavior patterns.

Items associated with recidivism were then scored to create scales communicating the risk of recidivism. They used a modified Burgess method to assign point values.[22]

Validation of the tool was supported by the R-value of the logistic regression of risk score to recidivism as well as AUC scores. ORAS defines recidivism as arrests for a new crime. Interestingly, the CST validation results suggested creating different risk cutoffs for males and females.[23]

ORAS has been adopted by Indiana and Texas, and was examined in Massachusetts but required further validation.

Virginia

The Virginia Pretrial Risk Assessment Instrument Revised (VPRAI) was developed and validated by Luminosity, Inc through the Virginia Department of Criminal Justice Services.[24] The instrument is used to inform judicial decisions on bail by predicting the risk of failure to appear and danger to the community. It is most relevant to the PCS instrument in that it considers the danger to the community.

The variables were identified from personal interviews, arrest warrants and other related data, defendant references, and prior criminal justice supervision records. Statistically significant variables were later used in a logistic regression to identify the significant predictors of pretrial outcomes. The risk factors were then assigned weights, which led to scores and corresponded to risk levels. The VPRAI has five risk levels.

[21] Other data such as case files, public record searches, and arrest data.
[22] Edward J. Latessa, Richard Lemke, Matthew Makarios, and Paula Smith, "The Creation and Validation of the Ohio Risk Assessment System (ORAS)," Federal Probation 74, no. 1 (June 2010), 17-18.
[23] The state of Alabama does have a separate risk needs assessment for women, see the FY 2017 Alabama Department of Corrections Annual report at http://www.doc.state.al.us/docs/AnnualRpts/2017AnnualReport.pdf.
[24] Danner et al., “Race and Gender Neutral Pretrial Risk Assessment, Release Recommendations, and Supervision: VPRAI and PRAXIS Revised.” Luminosity, Inc., Risk Assessment, November 2016.

The VPRAI's validity was re-examined through descriptive, bivariate, and multivariate analysis, such as chi-square and AUC.[25] Race and gender neutrality of the current model were also assessed. The instrument assesses the likelihood of pretrial failure, including technical violations.

The eight VPRAI risk factors are primary charge type, pending charges, criminal history, two or more failures to appear, two or more violent convictions, length at current residence, employed/primary caregiver, and history of drug abuse.

To examine the VPRAI's race neutrality, the minority ethnic and racial groups were combined into one group called People of Color. Outcomes were then compared to Whites. Of the eight risk factors, only "Lived at residence for less than one year" was not statistically significant. The AUC for White offenders was higher than offenders of color, and the difference was statistically significant. However, when the risk factors are weighted, summed, and collapsed into risk levels, the difference in the AUC values lose statistical significance. As a whole, the VPRAI can be considered neutral between People of Color and Whites for pretrial failure.

The VPRAI is gender neutral according to the logistic regression and comparing the risk levels between males and females in the sample. The two risk factors of "Two or more violent convictions" and "lived at residences less than one year" were not statistically for females and the report noted that assigning weights to these factors could overclassify pretrial failure risk for females.

The study uses the VPRAI with different weights for each factor, rather than equal weighting, and examines the outcomes with regards to race and gender. The analysis concludes the VPRAI does not have predictive bias and considers it neutral to gender and race.

Comparing VPRAI and ORAS to the SRAI

The ORAS-CST and CSST are the most similar to the PCS, and there are some relevant aspects of the VPRAI. Comparatively, the ORAS validation had a smaller sample size than the VPRAI.

[25] Danner et al., "Race and Gender Neutral Pretrial Risk Assessment, Release Recommendations, and Supervision: VPRAI and PRAXIS Revised," pages 2-3.

VPRAI provides a good comparison to the PCS instrument because it takes into consideration public safety. It also does not use gender and still achieves a moderate AUC score. Though the primary purpose of the instrument is failure to appear, it is able to accurately predict technical violations. The VPRAI's revalidation and effort to make improvements is also an example of good practice. As was done in the VPRAI revalidation, the PCS' general SRAI and CAP SRAI AUC's for race and gender was calculated and compared. The impact of adding weights to risk factors was also analyzed.

The ORAS and VPRAI provide important benchmarks to assess the general SRAI and CAP SRAI. Their validation and evaluation provide insight into methods to assess the PCS' SRAIs. Section 7 compares the overall and subpopulation AUC scores of these instruments.

5. Exploratory Data Analysis

Overview of Dataset

The dataset from the PCS consists of 131,064 observations, each representing an offender who was convicted of one or more offenses between 2004 and 2006, and where recidivism information on three-years post-release is available. Each observation is a partial view of the conviction and is only the most serious, non-DUI offense of the first judicial proceeding of the offender in this time period; a conviction may include many offenses.

In preparation for the assessment on instrument performance and fairness, an exploratory analysis was performed with four characteristics: race, gender, age, and conviction offense type. Characterizing the population by race and gender is important because those two variables define subpopulations and are assessed for fairness. These characteristics are also fundamental attributes of the offender population, and the instrument should perform similarly on similar populations, thus supporting its external validity.

It should be noted that this dataset is over a decade old; changes in demographics, as well as criminal justice reform, may have fundamentally changed the characterization of the offending population. If the risk assessment instrument is used in the future, validation on a more current dataset may need to be performed to ensure the instrument continues to have external validity.

Exploratory Data Analysis

The distribution of race is dominated by two ethnic groups: White and Black offenders, who make up 60% and 31% of the offender population, respectively. This is followed by Hispanic, which composes 7% of the population, and all other races comprising 2%. Because the racial distribution is unbalanced, and the instrument was developed on this dataset, it will be tuned to identify features of White and Black offenders. Figure 5.1 and Table 5.1 show the racial breakdown by proportion and by count.

Figure 5.1 Racial Breakdown

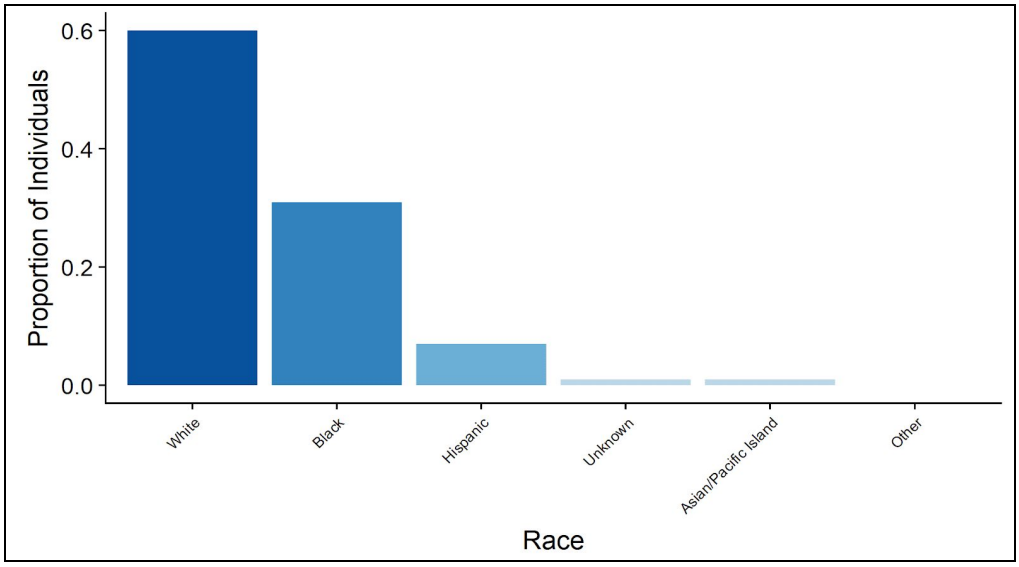


Table 5.1 Offender Racial Breakdown

Offender Race	Count	Proportion
White	77,997	60%
Black	40,949	31%
Hispanic	9,452	7%
Unknown	1,349	1%
Asian/Pacific Island	1,023	1%
Other	229	0%

Males make up about 80% of the offender population, totally roughly 104,000 individuals. Males were also roughly 50% more likely to recidivate than females, and comprise nearly 84% of the recidivator population (Table 5.2).

Table 5.2 Offender Gender Breakdown

Offender Gender	Count	Proportion
Male Offender Population	104,240	79.5%
Female Offender Population	26,824	20.5%
Recidivator Population Percent Male	36,054	83.6%
Recidivator Population Percent Female	7,068	16.4%
Percentage of men who recidivate	-	34.6%
Percentage of women who recidivate	-	26.3%

Age is a critical category within the risk assessment instrument because an offender can be assessed up to five points for being younger than 21. The offender population skews young, with a median age of 28 (Figure 5.2). The age profile of the two largest racial groups (Black and White) matches the overall population age profile, and both groups have median ages between 27 and 27 and 29 (Figure 5.3).

Figure 5.2 Distribution of Age

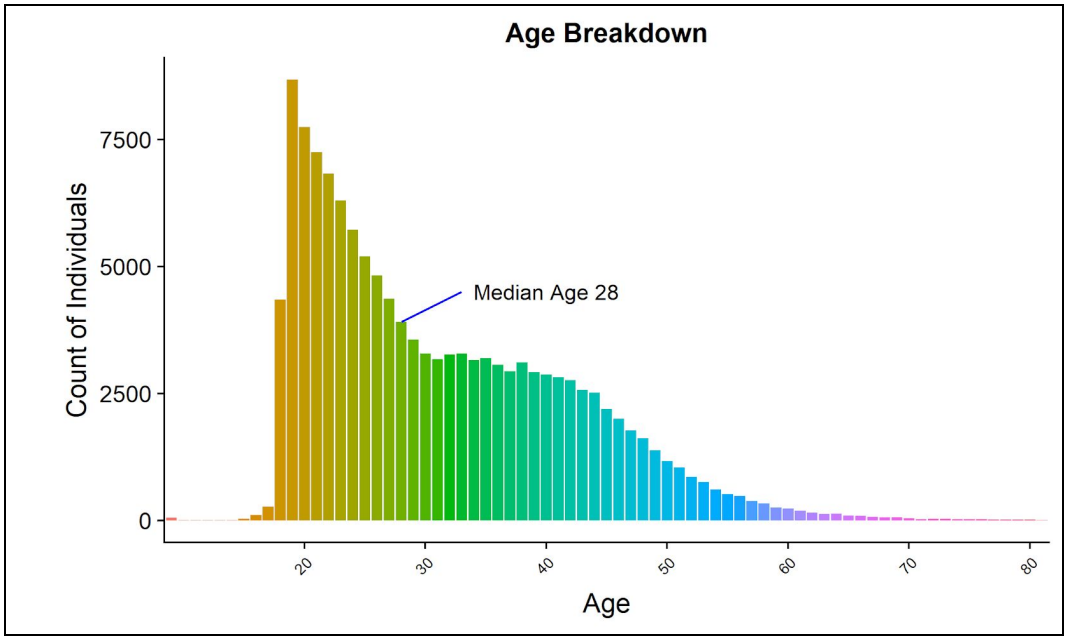
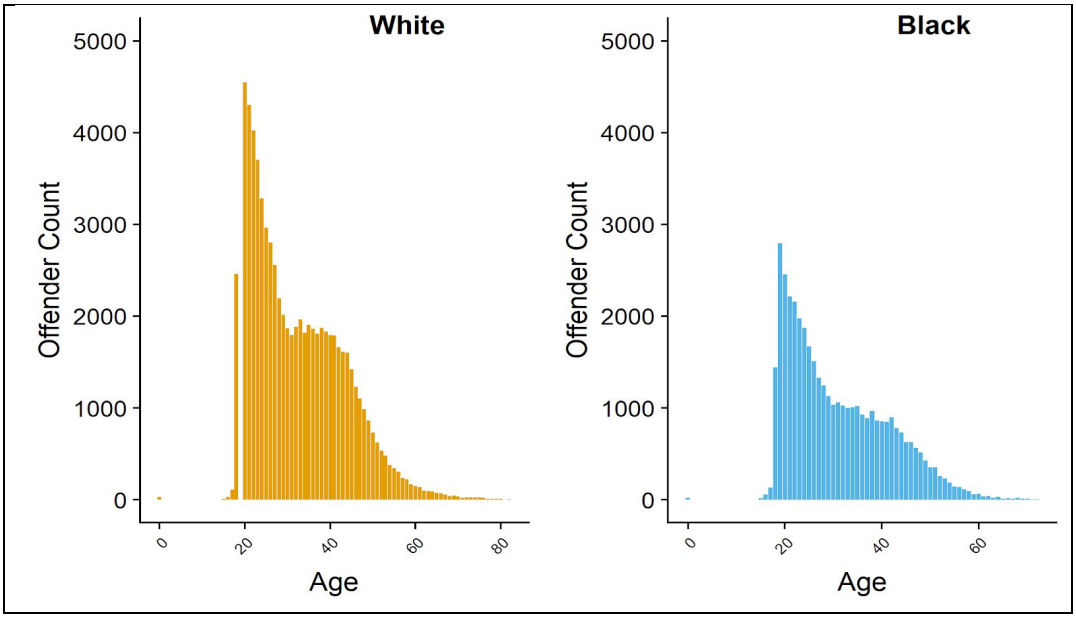
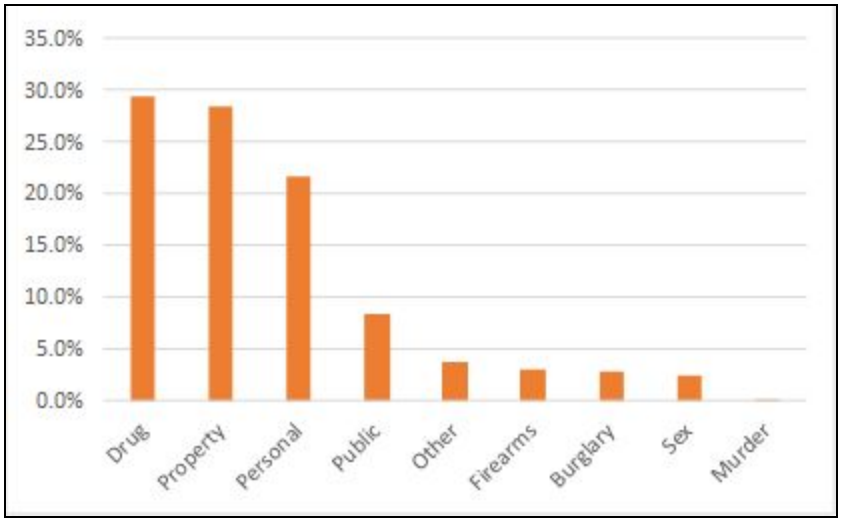


Figure 5.3 Age Breakdown of White and Black Offenders



Of the current conviction types, 29.4% are drug-related, 28.5% are property-related, and 21.7% are personal-related (Figure 5.4). However, this is not the offense distribution because the dataset only includes the most serious offense of the first judicial proceeding. Furthermore, the criteria of seriousness are not well defined, therefore, the dataset may not reflect the true distribution of offense types.

Figure 5.4 Crime Type Breakdown



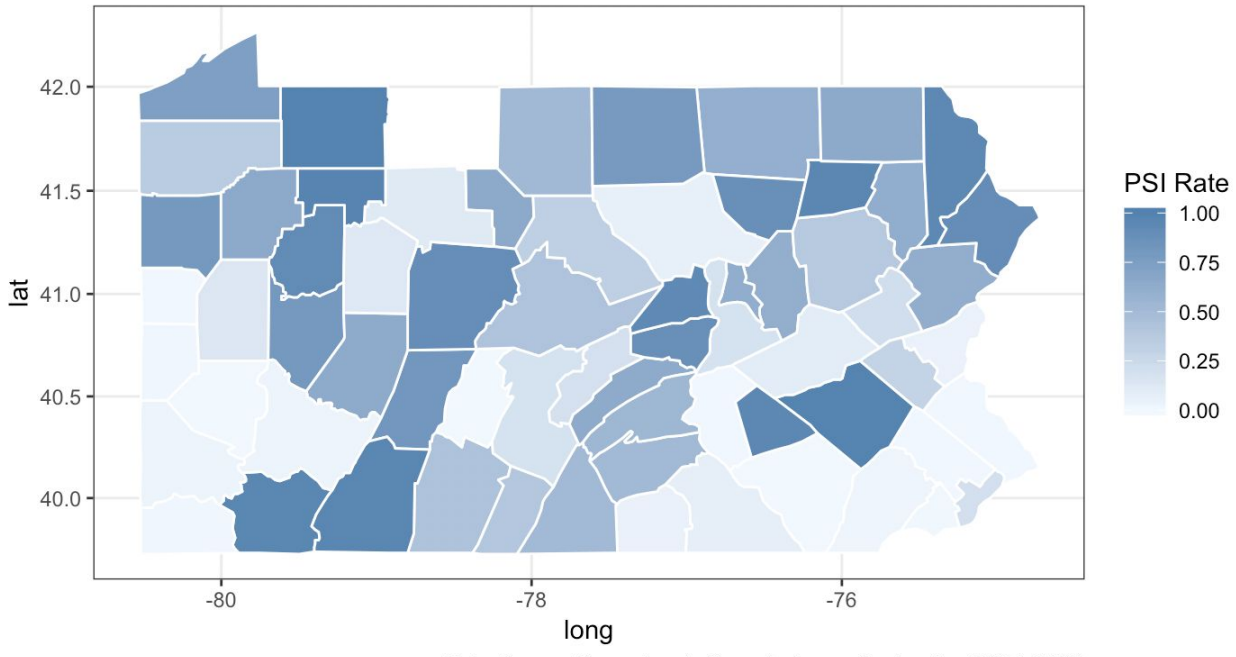
PSI Usage Statistics

For an offender classified as high or low risk, the PCS recommends judges seek additional information through a PSI. Since the instrument encourages PSI use, it is important to understand how the expected rate of PSI usage across the state to change if the instruments recommendations are followed. One metrics calibrating usage of PSI is frequency, which is defined as

$$PSI\ Rate\ =\ \frac{Offenders\ Completed\ PSI}{Total\ Number\ of\ Offenders}$$

A higher PSI rate indicates more usage of PSI, while a lower PSI rate indicates less usage of PSI. According to the data provided by the PCS, 31,823 total PSIs were completed in Pennsylvania from 2004 to 2006, corresponding with a state PSI rate of 24.2%. If the judicial processes are similar across regions, PSI rates should be close to the state level. However, Figure 5.5 shows the Pennsylvania PSI rates by county and reveals inconsistent county PSI usage.

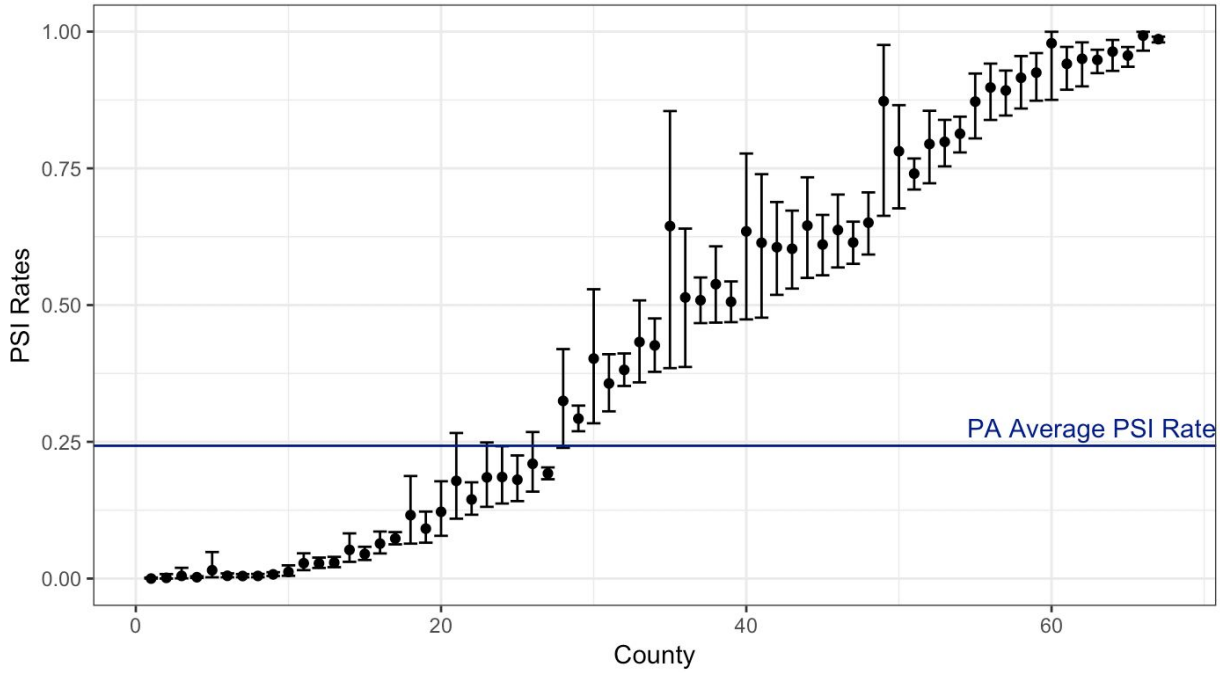
Figure 5.5 Pennsylvania PSI Rate by County



Data Source: Pennsylvania Commission on Sentencing (2004-2006)

To determine whether the differences in PSI rates are significant, the 95% confidence interval (CI) of the PSI rate for each county is calculated. As shown in Figure 5.6, PSI rates by county range from zero to one. The CIs on both tails are narrow, which indicates maximum and minimum PSI rates are based on large enough samples and are not by chance alone.

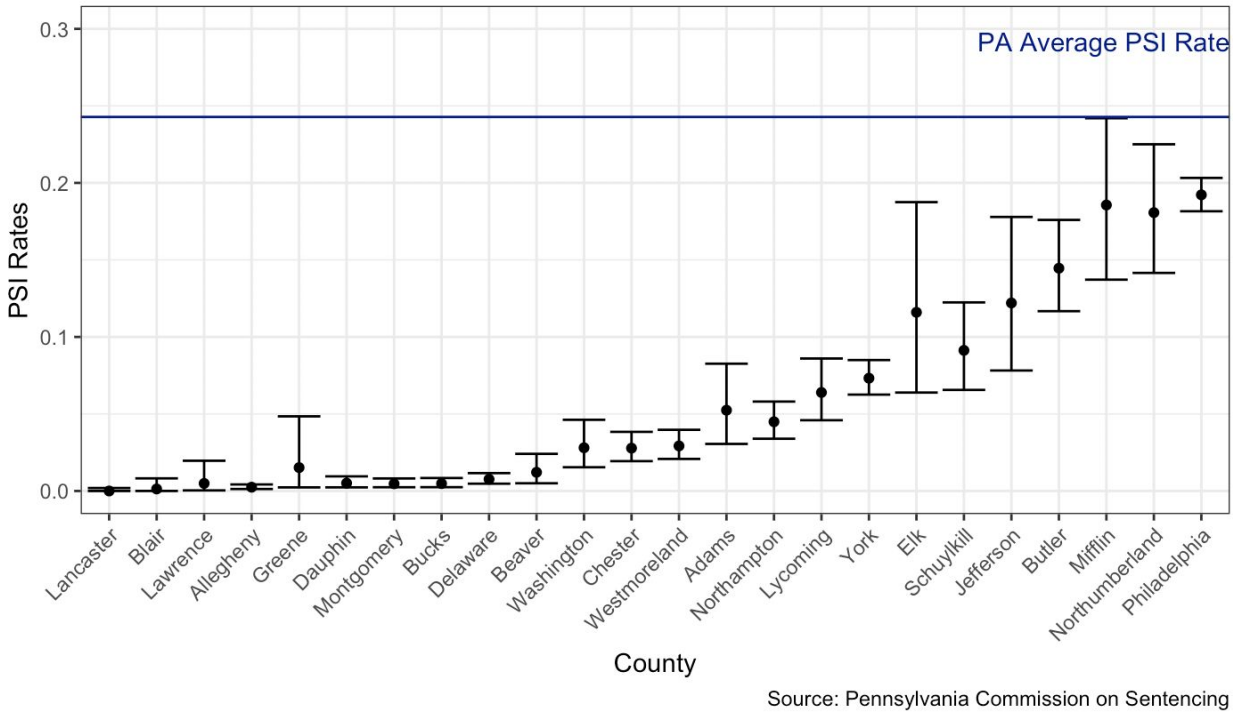
Figure 5.6 PSI Rates Compared to PA Average



Source: Pennsylvania Commission on Sentencing

Figure 5.7 zooms in on the counties with PSI rates lower than the state average. These counties have significantly lower usage of PSI compared with the state average, even with the upper bounds of their confidence levels.

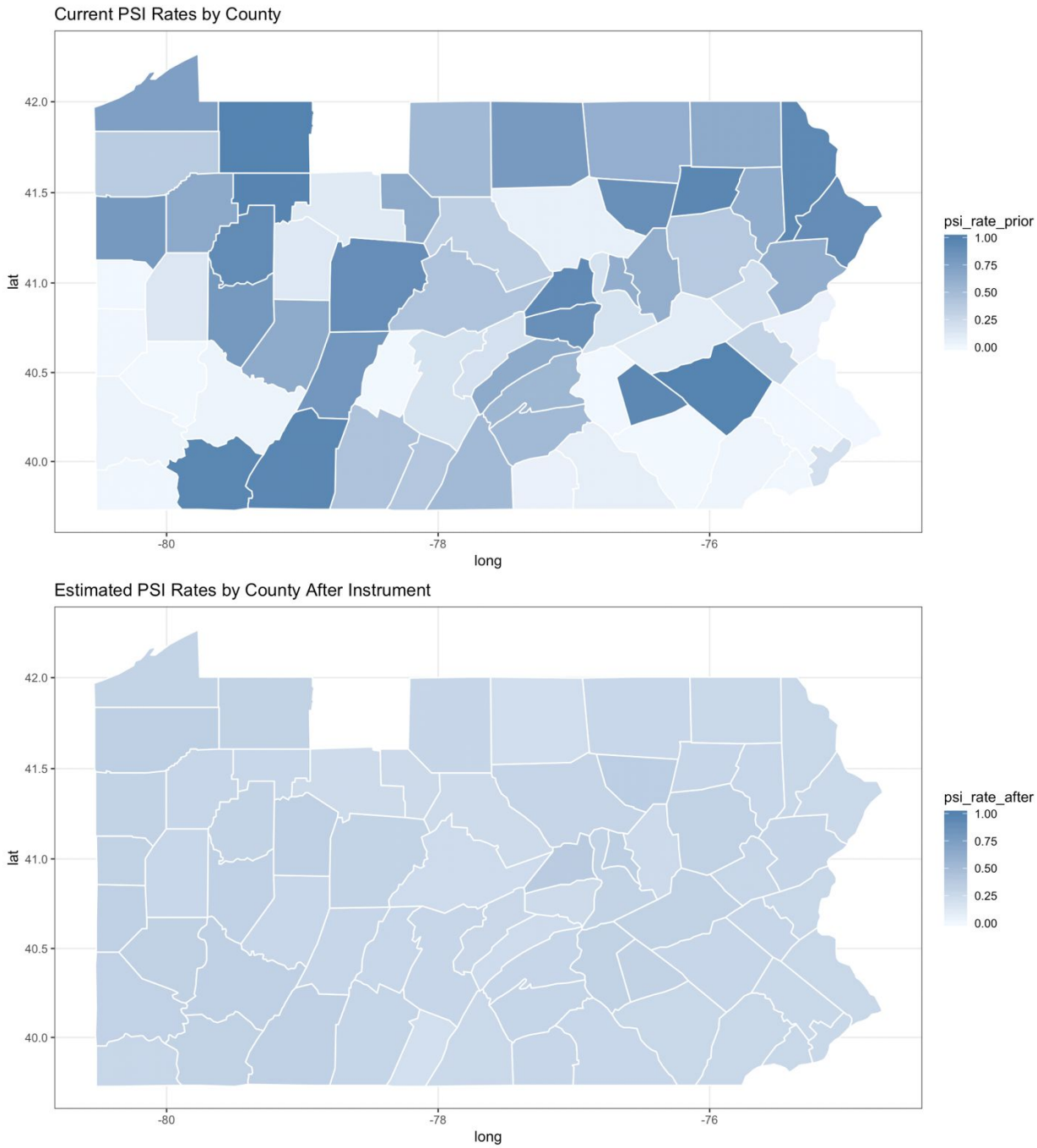
Figure 5.7 PSI Rates Lower than PA Average



Source: Pennsylvania Commission on Sentencing

If PSIs were to completed following the rate at which the instrument identifies high- or low-risk offenders, the PSI rates across counties will be more consistent (Figure 5.8). The total PSI number would increase to 36,336 for the same offenders from 2004 to 2006 and the overall PSI rate of Pennsylvania would be 27.7%, which indicates more labor hours.

Figure 5.8 Comparison of PSI Rates Before and After the Instrument



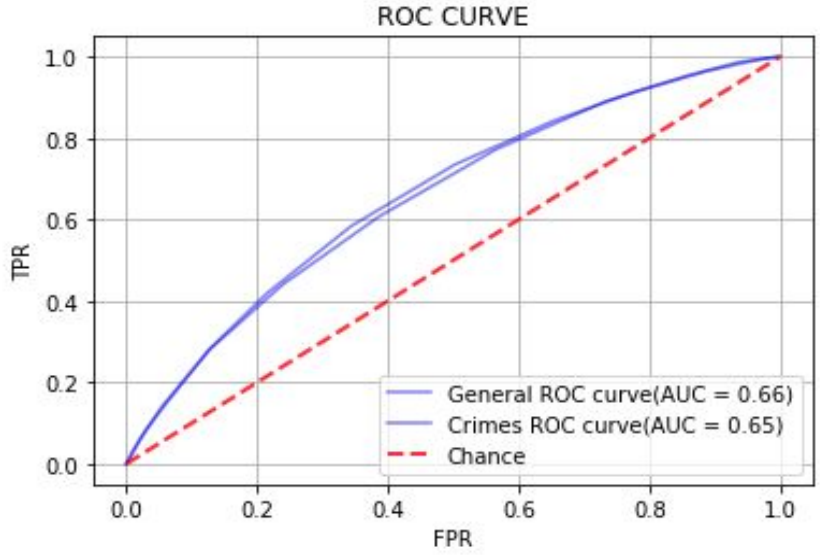
6. Replication

The first stage of assessment is replication, which ensures that the PCS's SRAI development procedure is errorless. Both general and CAP SRAIs were replicated in terms of logistic regression models and risk results. AUC scores were only calculated for general recidivism.

Logistic regression models were used by the PCS to select features that were significantly correlated with reconviction within three years. The coefficients and significance levels replicated were close to those provided by the PCS, with acceptable nuances as shown in Table A6.1 and A6.2 in the Appendix.

The risk scores and corresponding risk categories based on the original cutoffs were also successfully replicated. An AUC score was used by the PCS to validate the instrument's performance, and the team was able to replicate the AUC calculated by the PCS (Figure 6.1).

Figure 6.1 ROC and AUC Scores by Project Team



Our replication verified three important steps in the PCS's process, development, application, and validation. The aligned results indicate that the PCS's procedure was errorless.

7. Performance and Fairness

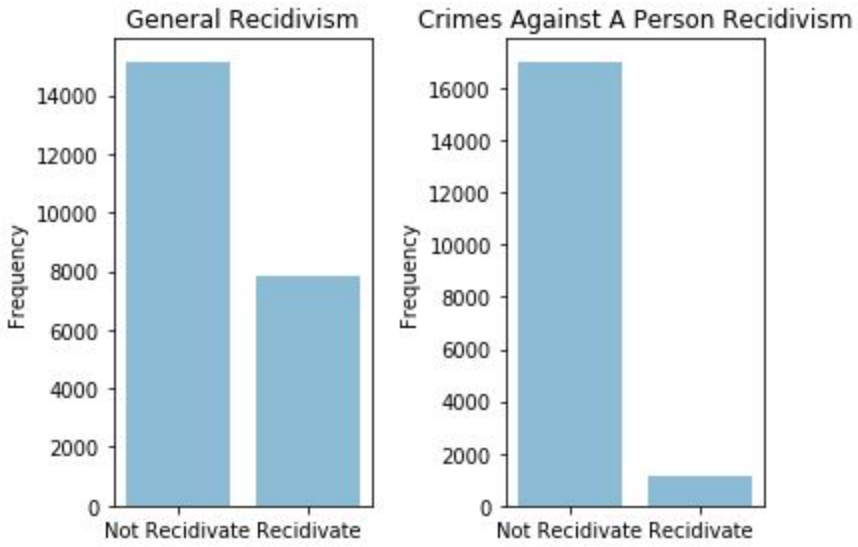
Performance in Predictive Modeling

Limitations of Accuracy

Imbalanced data refers to a skewed distribution of the target variable. In imbalanced datasets, the majority class of the target variable dominates the minority class in terms of the number of members within each class. Many times, accuracy is used as the metric to determine which classifier, or algorithm, performs better on such data. However, accuracy is not an appropriate metric. For example, in fraud detection, if 1% of the samples are fraudulent and 99% are truthful, the default majority classifier which classifies based on the majority would classify all the examples to be truthful. This would result in an accuracy of 0.99 without having to classify a single fraudulent example correctly. This is because accuracy places a greater weight on the majority class, rather than the minority class.

From Figure 7.1, it is clear that the non-recidivating class outweighs the recidivating class in the validation dataset. For the general SRAI, there are only 8,000 individuals who recidivate, which is roughly half the number of individuals who do not recidivate. In the CAP SRAI, there are only about 2,000 individuals that recidivate, which is an eighth of the number of individuals who do not recidivate. Therefore, in both cases, accuracy alone should not be used to determine the performance and fairness of the data.

Figure 7.1 Imbalanced Data



Why Use Performance Metrics?

Since accuracy is skewed by imbalanced data, a common strategy is to use metrics that are not affected by unbalanced datasets such as precision, sensitivity, and specificity. Sensitivity can be interpreted as the accuracy of the positive examples, and specificity can be interpreted as the accuracy of the negative examples. Precision can be interpreted as the measure of correctness. These formulations for the general SRAI and CAP SRAI can be observed in Table 7.1 below.

Popular evaluation metrics, such as the G-mean and F1-score, combine other performance metrics such as sensitivity and specificity along with precision to assess whether a classifier is balanced in its predictions of the majority and minority class. The G-mean is the square root of the sensitivity times the specificity. Even if the majority class is correctly classified, the G-mean metric will be relatively low if the prediction for the minority class if poor. On the other hand, the F1-score equals sensitivity multiplied by two and multiplied by precision and then divided by the sum of sensitivity and precision. The advantage of the F1-score metric is that the generalized form allows for the weighting of precision or sensitivity. Both metrics can be used to extensively evaluate imbalanced datasets.

AUC-ROC curves (AUC curves) can also be used to assess the performance of a model that is based on an imbalanced dataset because it is independent of the prevalence rate. The closer the curve is toward the left corner of the plot, the greater a classifier's ability to differentiate between the opposing classes. By using multiple performance metrics along with ROC curves, models can be numerically and visually compared with one another. In addition, by using multiple metrics, models can be chosen based on varying interests such as weighting the significance of the models in comparison to the specificity of the models. Therefore, a combination of formulaic and graphical metrics can be used to assess the performance of a classifier on an imbalanced dataset.

Methodology

In order to calculate the different performance metrics for both the general SRAI and CAP SRAI, confusion matrices were constructed on the validation data. Individuals in the high-risk categories were considered correctly classified if they recidivated, and individuals in the low-risk categories were considered correctly classified if they did not

recidivate. Confusion matrices were calculated for both the general recidivism instrument and the crime against a person instrument (Table 7.1)

Table 7.1 Confusion Matrices of Recidivism

General		
	Prediction	
Truth	No	Yes
No	9,660	1,842
Yes	5,501	6,020

Crime Against A Person		
	Prediction	
Truth	No	Yes
No	9,014	160
Yes	8,024	944

Note: Estimations based on the validation dataset.

Only based on AUC, the general SRAI and CAP SRAI perform moderately well with AUC scores of 0.66 and 0.65 respectively (see Figure 6.1 in previous section). They also perform similarly compared to the two instruments used in other states identified earlier, ORAS and VPRAI. When broken down by subpopulations, both the general SRAI and CAP SRAI do not see large drops in AUC scores, either compared to the overall AUC score or between subpopulation AUC scores. This is in comparison to ORAS, which has a much higher AUC for White offenders and female offenders than for Black offenders or offenders of color and male offenders, respectively (Table 7.2).

Table 7.2 AUC Scores by Subpopulations

State and Instrument	Overall	White	People of Color/Black	Male	Female
Ohio - ORAS	0.60	0.64	0.54	0.56	0.64
Virginia - VPRAI Revised	0.68	0.70	0.66	0.69	0.68
Pennsylvania - PCS Gen.	0.66	0.66	0.66	0.66	0.65
Pennsylvania - PCS CAP	0.65	0.66	0.65	0.65	0.64

Bootstrap

The Bootstrap Percentile Method was used to determine whether the differences in performance metrics between races and genders were significant. Bootstrapping is a

resampling technique used to estimate statistics on a population by repeatedly sampling a dataset with replacement. Given the assumption that the offender population does not change over time, the offenders convicted between 2004 and 2006 should be a sample of the whole offender population, from which population differences between subpopulations can be approximated. To do so 500 samples a twentieth the size of the 2004-2006 dataset were drawn. The differences were calculated for performance metrics between races and genders with each sample, and 95% confidence intervals were obtained by taking the lower bound 2.5% and upper bound 97.5% percentiles of 500 estimates. If the range of the confidence interval of a difference did not include zero, the difference is considered statistically significant.

Performance Assessment

The metrics in Table 7.3 were calculated by using the confusion matrices represented in Table 7.1. In addition to the performance metrics, the high error rate and low error rate are included. The high error rate takes the count of discrepancies between the true classification and the predicted classification of the high-risk group over the total number of individuals in the high-risk group. The low error rate is calculated similarly to the high error rate, but for the low-risk category. When assessing performance for the SRAIs then, high values for high (low) error rate indicates poor performance at classifying high- (low-) risk offenders.

Table 7.3 Metric Values

Metric	Formula	General SRAI	CAP SRAI
Accuracy	(TP+TN) / (TP+TN+FP+FN)	0.681	0.549
Precision	(TP) / (TP+FP)	0.523	0.105
Sensitivity	(TP) / (TP+FN)	0.766	0.855
Specificity	(TN) / (TN+FP)	0.637	0.529
F1-Score	(2 * Precision * Sensitivity) / (Sensitivity + Precision)	0.621	0.187
G-Means	(Sensitivity * Specificity) ** (1/2)	0.698	0.673
High Error Rate (low tail)	# Count(Prediction ≠ True)	0.477	0.895
Low Error Rate (low tail)	# Count(Prediction ≠ True)	0.160	0.017

The performance metrics were calculated for both the general SRAI and CAP SRAI. Since the data for the crime against a person outcome is noisier, fewer conclusions relative to the general SRAI can be made between the metrics in the two columns. The CAP SRAI clearly has a larger high-error-rate in comparison to the low error rate. This means that the CAP SRAI performs worse in identifying high-risk individuals than low-risk individuals. Furthermore, the CAP SRAI high error rate is extremely large.

The large high error rate is large because most individuals do not recidivate a crime against a person. Individuals who are classified as low risk are almost always predicted correctly. Interestingly the number of individuals identified in both the high-risk and low-risk categories for the CAP SRAI is the same at approximately 9,000 individuals in each risk category group. However, out of 9,174 low-risk individuals, 9,014 individuals are predicted correctly, whereas out of 8,968 high-risk individuals, only 944 individuals are predicted correctly. This explains why sensitivity is so high (0.855) compared to the specificity (0.529).

Additional Considerations

While performance metrics can help determine whether a dataset is imbalanced, they do not resolve the imbalance. They just provide a better representation of the underlying data and results. To resolve the imbalance, other techniques could be considered such as upsampling, downsampling, or SMOTE (i.e. Synthetic minority over-sampling technique). Upsampling is a technique that randomly repeats the number of records of the minority class until the proportion of the minority and majority class are equivalent. Downsampling reduces the majority class until the proportion of the minority and majority class are equivalent. While these techniques can moderate the imbalance in the data, upsampling is preferred over downsampling because, in downsampling, the underlying distributions may be disturbed by reducing the majority class count. However, both techniques aren't without limitations and neither were pursued as part of this project.

Fairness in Predictive Modeling

A data-driven tool can involuntarily reproduce and reinforce existing biases or introduce new ones.[26] Such biases can have long-lasting impacts on people's lives. Therefore, it is important that risk assessment instruments are fair.

A common strategy for dealing with unfairness is to exclude indicators of group membership (e.g., gender, race). However, this approach may not remove the bias if other variables correlate with the group membership.[27] It is also common to rely only on the AUC to assess the predictive validity of an instrument. The AUC measures the instrument's ability to separate observations between classes, but it does not assess the instrument's ability to accurately predict a positive class or measure fairness between groups.[28] Therefore, different metrics are necessary to measure how fair the instrument is.

Contrary to accuracy or precision, which are defined, there is no single, accepted concept of fairness in the field. Fairness is assessed by examining how the instrument treats different protected or vulnerable groups.[29] Two high-level definitions of fairness exist. Statistical fairness means that, on average, different protected groups should be treated similarly. Individual fairness means that "similar individuals should be treated similarly".[30]

The PCS SRAI satisfies individual fairness as two offenders of the same age, gender, and criminal history receive the same risk score. To assess if the PCS SRAI is statistically fair, the project team used five different fairness metrics proposed in the assessment literature (Table 7.4).[31] When the difference in all the fairness metrics between the subpopulation groups is zero, then the instrument has achieved total fairness.

The fairness metrics use the results from the confusion matrix, and are not algorithm-specific. The metrics can be estimated for any instrument. In fact, most of the metrics are the same ones used for estimating the performance of any instrument, such as accuracy, and true positive and negative rates. The difference is that each

[26] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.
[27] R, Berk. et al., Fairness in Criminal Justice Risk Assessments: The State of the Art.
[28] J. P. Singh, Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer: Performance indicator primer.
[29] R. Berk, et al.
[30] A. Chouldechova, and A., Roth, The Frontiers of Fairness in Machine Learning.
[31] R. Berk, et al. A. Chouldechova.

performance metric is estimated and compared between two or more subpopulation groups. For instance, they are estimated separately for the White and Non-White offenders and then, compared.

In summary, there are five fairness metrics. Overall accuracy measures how well the instrument correctly identifies or excludes a condition. Demographic parity measures the proportion of persons that are predicted to have a condition, this should be the same across groups. Conditional procedure accuracy equality is achieved when the true positive and true negative rates, although different between them, are the same across groups. Predictive value equality is achieved when, conditional on the prediction of a positive (negative) outcome, the probability of success (failure) is the same across groups. Treatment equality means that the cost of falsely classifying a person is the same across groups.

Table 7.4 Fairness Metrics

Category	Description	Formula
Overall Accuracy	Measures how well the Instrument correctly identifies or excludes a condition	(TP+TN)/Total
Demographic Parity	Estimates the proportion of persons that are predicted to (not) have a condition.	(TP+FP)/Total; (FN+TN)/Total
Conditional Procedure Accuracy Equality	Measures the proportion of true positives (true negatives) that are correctly identified as positives (negatives). Also known as true positive (true negative) rate or sensitivity (specificity)	TP/(TP+FN); TN/(FP+TN)
Conditional Use Accuracy or Predictive Value Equality	Measures the proportions of positive (negative) results that are true positive (true negative) results.	TP/(TP+FP); TN/(FN+TN)
Treatment Equality	Estimates the ratio of false negative and false positives.	FP/FN

Note: TP: true positive; TN: true negative; FP: false positive; FN: false positive. Total: total sample

The fairness metrics can be calculated for any protected group. Based on the PCS's interests, these were estimated based on the offenders' race and gender (see the two following sections). When the recidivism rate differs between two subpopulation groups, the false positive and negative rates will be different across those groups.[32] Therefore,

[32] A. Chouldechova.

there will not be strict equality across some fairness metrics. To better conduct the fairness assessment, the recidivism rate was estimated across each protected group over the whole dataset (including both the validation and development set). The confusion matrices and fairness metric across each protected group was computed using the validation set.

General SRAI Fairness Assessment

Race

The PCS classified the offenders' race as White and Others,[33] Black, or Hispanic offenders. The fairness assessment was conducted using these same categories. Table 7.5 shows that the recidivism rates vary across the three subpopulation groups. The Black population has the highest recidivism rate (36.4%), followed by the White population (31.5%), and finally the Hispanic population (29.0%).

Table 7.5 Recidivism Rate by Race Group

Race	Recidivate	Observations	Rate
White and Others	No	54,854	68.5%
White and Others	Yes	25,254	31.5%
Black	No	26,380	63.6%
Black	Yes	15,124	36.4%
Hispanic	No	6,708	71.0%
Hispanic	Yes	2,744	29.0%

Note: Estimations based on the whole dataset (validation and development set).

The confusion matrices provide the inputs for estimating the fairness metrics for each race group (Table 7.6). The confusion matrix was estimated for the three race groups, but only the White and Black groups are analyzed as 91% of the offenders were included in these two groups.

[33] Others includes unknown race, Asian/Pacific Islander, and other races. These races groups accounts for two percent of the total offenders. Therefore, excluding "Others" from the White and Others category will not change the assessment and main findings of the fairness assessment.

Table 7.6 Confusion Matrices by Offenders' Race

White and Others		
	Prediction	
Truth	No	Yes
No	6,775	1,290
Yes	3,154	3,463

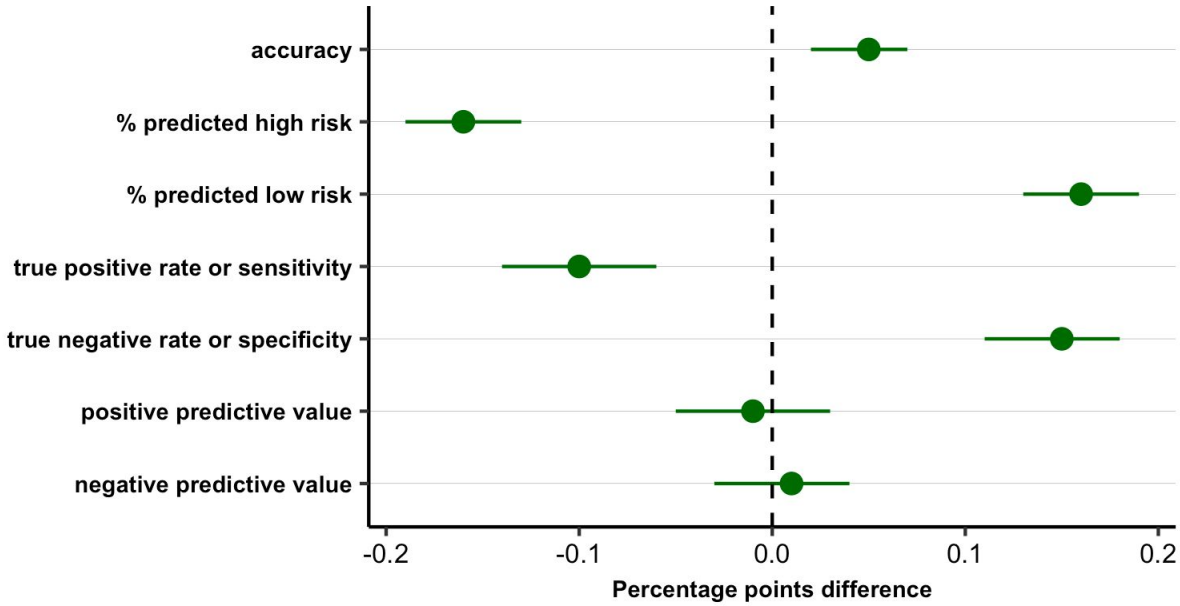
Black		
	Prediction	
Truth	No	Yes
No	2,258	451
Yes	1,954	2,219

Hispanic		
	Prediction	
Truth	No	Yes
No	627	101
Yes	393	338

Note: Estimations based on the validation dataset.

Figure 7.2 shows the differences in the fairness metrics between White and Black offenders and the confidence interval of their differences built using the bootstrapping method.[34] If the confidence interval range includes the zero, then there are no statistical differences between both population groups. Appendix A7.1 includes the point estimates of the fairness metrics for each race group.

Figure 7.2 Fairness Metrics Differences Between White and Black Offenders



Note: The difference is estimated as White minus Black. 95% confidence intervals. Estimations based on the validation dataset.

[34] The treatment equality was excluded from any data visualization because it ranged above one. Table A7.1 includes the point estimate of all the fairness metrics, including the treatment equality.

The accuracy is five percentage points higher for the White offender population. Seventy percent of the White offenders are correctly identified as high and low risk. For Black offenders, this statistic goes down to 65%. The demographic parity includes the percentage of offenders classified as high and low risk. The general SRAI classifies 61% of the Black offenders as high risk, when the true recidivism rate is 36%. Similarly, 45% of the White offenders are identified as high risk, with 31% being their true recidivism rate. The instrument is making mistakes for both race groups.

The conditional procedure accuracy equality includes the true positive and negative rates. Here the team found two main insights. First, the true positive rate is 10 percentage points higher for the Black offender population. Meanwhile, 73% of the White offenders who recidivate are correctly classified as high risk. This figure goes up to 83% for Black offenders. Secondly, the true negative rate is 14 percentage points higher for the White population: 68% of the White offenders who do not recidivate are correctly classified as low risk, the figure goes down to 54% for the Black offenders. It seems that the general SRAI is biased as more White offenders who do not recidivate are classified as low risk in comparison to the Black offenders. However, this result is a direct consequence of having different recidivism rates across groups and an instrument that does not perfectly separate both classes (recidivate vs not recidivate).

The positive and negative predictive values are equal across the two subpopulation groups. This result means that the probability that an offender, either White or Black, classified as high risk is truly an offender is 53%. Similarly, the probability that an offender classified as low risk is truly a non-offender is 83%. These values mean that after accounting for the prevalence rate for each race group, the probability of being identified as high risk (or low risk) is the same across both race groups.[35]

The treatment equality, which includes the ratio of false positive and negative, does not hold across groups.[36] It is higher for Black offenders than for White offenders. Therefore, in relative terms, there are more false positives than false negatives for Black offenders compared to Whites.

In summary, the general SRAI does not achieve total fairness as only the positive and negative predictive values are equal across White and Black offenders.

[35] The positive and negative predictive values takes into account the recidivism rate as they can be derived using the Bayes Theorem.
[36] Treatment equality metric is not shown in Figure 7.2 but included in Table A7.1.

Gender

The offenders' gender is classified as female or male. Table 7.7 shows that men have a higher recidivism rate than women. One out of every three male offenders recidivates, while only one of every four women recidivates.

Table 7.7 Recidivism Rate by Offender Gender

Gender	Recidivate	Observations	Rate
Male	No	68,186	65.4%
Male	Yes	36,054	34.6%
Female	No	19,756	73.7%
Female	Yes	7,068	26.3%

Note: Estimations based on the whole dataset (validation and development set).

To estimate the fairness metrics, two confusion matrices in Table 7.8 were used.

Table 7.8 Confusion Matrices by Offender Gender

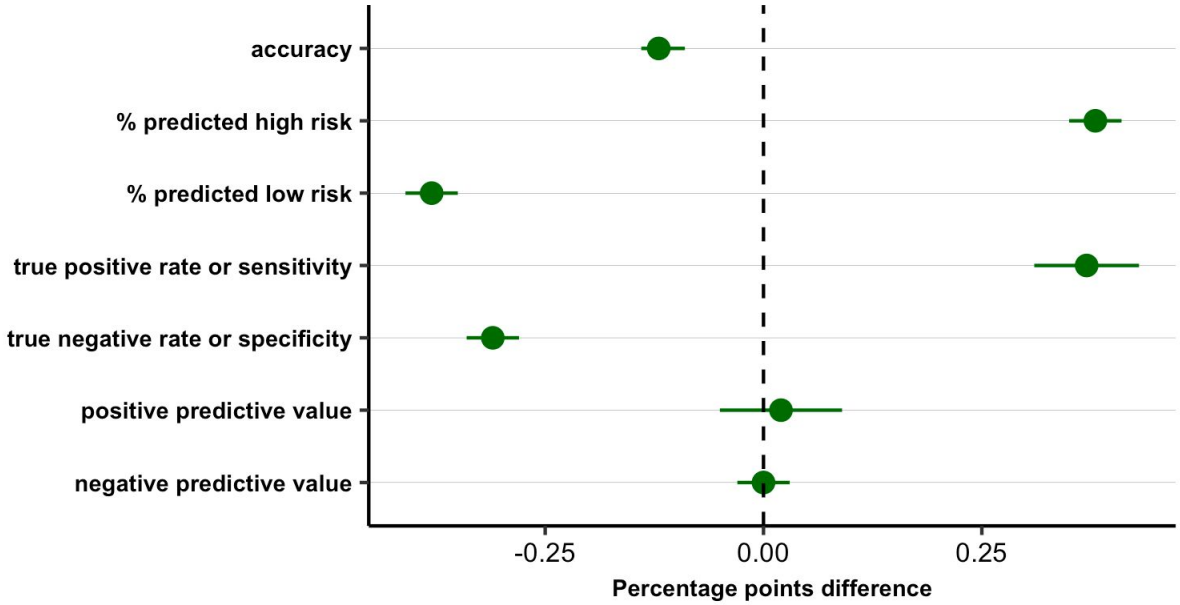
Female		
	Prediction	
Truth	No	Yes
No	3,455	654
Yes	533	540

Male		
	Prediction	
Truth	No	Yes
No	6,205	1,188
Yes	4,968	5,480

Note: Estimations based on the validation dataset.

Figure 7.3 shows that accuracy is 12 percentage points higher for women than men (77% vs 65%). The general SRAI classifies 21% and 59% of the female and male population as high risk, respectively, when the true recidivism rate is 26% and 35%. The general SRAI is systematically making more mistakes for men offenders, meaning that the demographic parity does not hold between genders.

Figure 7.3 Fairness Metrics Differences Between Male and Female Offenders



Note: The difference is estimated as male minus female. 95% confidence intervals. Estimations based on the validation dataset.

The conditional procedure accuracy equality also shows differences between gender groups. The true positive rate is 37 percentage points higher for male offenders. Eighty-two percent of the male offenders who recidivate were classified as high risk. This value is 45% for female offenders. In contrast, the true negative rate is 29 percentage points higher for women, as 56% of the male offenders who do not recidivate are correctly classified as low risk and 87% of the female offenders who do not recidivate are correctly classified as low risk. These two results imply that the instrument is better for predicting high-risk males and low-risk female offenders.

The predictive value equality shows that the positive and negative predictive values are similar between both genders groups. There are no statistical differences across gender groups. These metrics show that conditional on the prediction of being high risk (or low risk), the probability of recidivating (or not recidivating) is the same across male and female offenders. Finally, the treatment equality implies that, in relative terms, there are more false positives than false negatives for male offenders than for female offenders.[37]

[37] The treatment inequality across genders is shown in Table A7.2.

In short, only the positive and negative predictive values are equal across male and female offenders. The instrument does not achieve total fairness, which can be explained as there are different recidivism rates across both gender groups.

CAP SRAI Fairness Assessment

Race

The offenders' race is classified as White and others, Black, or Hispanic. Table 7.9 shows that the prevalence rate for recidivating for a crime against a person is low: 4.8% for White and others offenders, 5.9% for Black offenders, and 4.9% for Hispanic offenders. The confusion matrices in Table 7.10 show the inputs used for estimating the fairness metrics for each race group. The confusion matrix for each race group was calculated, but only the White and Black offenders are analyzed in detail.

Table 7.9 Recidivism Rate for Crime Against a Person by Race Group

Race	Recidivate	Observations	Rate
White and Others	No	76,262	95.2%
White and Others	Yes	3,846	4.8%
Black	No	39,052	94.1%
Black	Yes	2,452	5.9%
Hispanic	No	8,988	95.1%
Hispanic	Yes	464	4.9%

Note: Estimations based on the whole dataset (validation and development set).

Table 7.10 Confusion Matrices for Crime Against a Person by Offenders' Race

White and Others		
	Prediction	
Truth	No	Yes
No	5,638	95
Yes	4,718	540

Black		
	Prediction	
Truth	No	Yes
No	2,763	51
Yes	2,803	345

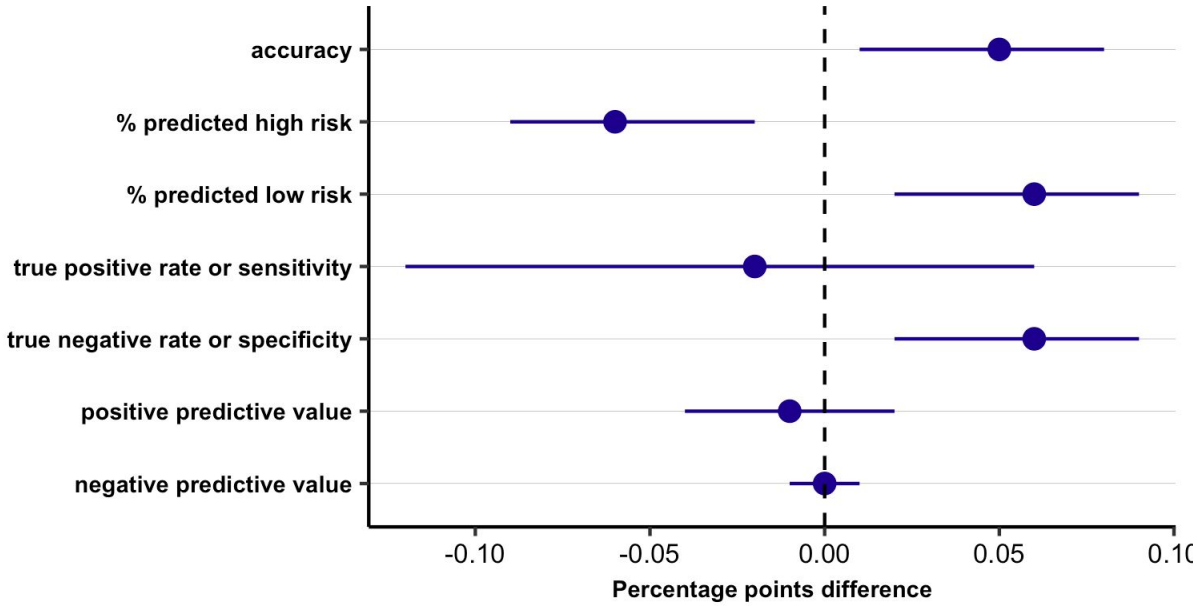
Hispanic		
	Prediction	
Truth	No	Yes
No	703	14
Yes	503	59

Note: Estimations based on the validation dataset.

Figure 7.4 shows that the accuracy for White and Black offenders is small (five percentage points) but statistically significant among the two groups: 56% and 51%, respectively. The accuracy is slightly better than a random guess.

The CAP SRAI classifies more Black offenders than White offenders as high risk (48% and 54% respectively). The difference between both groups is small. However, the true recidivism rate for a crime against a person for the White and Black population is 4.8% and 5.9%, respectively. Therefore, the predicted percentage of offenders is 10 times higher than the true recidivism rate.

Figure 7.4 Fairness metrics differences between White and Black offenders



Note: The difference is estimated as White minus Black. 95% confidence intervals. Estimations based on the validation dataset.

The conditional procedure accuracy equality shows that the difference between the true positive rates across race groups is not statistically significant. Therefore, a similar percentage of White and Black offenders who recidivate are classified as high risk (85% and 87% respectively). However, the true negative rate is small but statistically different across groups. Among the White population, 54% of the offenders who do not recidivate were correctly classified as low risk. This figure is 49% among Black offenders.

The predictive value equality shows that the positive and negatives predictive values are similar between females and male offenders. Said otherwise, conditional on the

prediction of being classified as high risk (or low risk), the probability of recidivating for a crime against a person (or not recidivating) is the same across both gender groups. The treatment equality, which includes the ratio of false positive and negative is statistically equal between Black and White offenders.[38] Therefore, in relative terms, there is the same number of false positives and false negatives for both race groups.

If the CAP SRAI is used only for identifying offenders with a low-risk of recidivating for a crime against a person, only the percentage of offenders predicted as low risk and the negative predictive value should be considered. Then, the differences between White and Black offenders are small or statistically insignificant. These results seem to suggest that the CAP SRAI is unbiased. However, the performance metrics show low predictive power. The CAP SRAI seems to be fair because it has low predictive power for identifying even low-risk offenders.

Gender

The offenders' gender is classified as female or male. Table 7.11 shows that men have twice the probability of recidivating for a crime against a person than women (5.8% vs 2.6%).

Table 7.11 Recidivism Rate for Crime Against a Person by Offender Gender

Gender	Recidivate	Observations	Rate
Male	No	98,171	94.2%
Male	Yes	6,069	5.8%
Female	No	26,131	97.4%
Female	Yes	693	2.6%

Note: Estimations based on the whole dataset (validation and development set).

The confusion matrices shown in Table 7.12 were used to estimate the fairness metrics.

[38] Table A7.3 includes the point estimate of the treatment equality for both race groups.

Table 7.12 Confusion Matrices for Crime Against a Person by Offender Gender

Female		
	Prediction	
Truth	No	Yes
No	3,926	45
Yes	312	19

Male		
	Prediction	
Truth	No	Yes
No	5,088	115
Yes	7,712	925

Note: Estimations based on the validation dataset.

Figure 7.5 shows the fairness metrics for male and female offenders. The accuracy is higher for female offenders by about 50 percentage points compared to male offenders (43% vs. 92%). The percentage of offenders predicted as high risk is 54 percentage points higher for male offenders than female offenders (62% vs. 8%). These values are far away from the true recidivism rate.

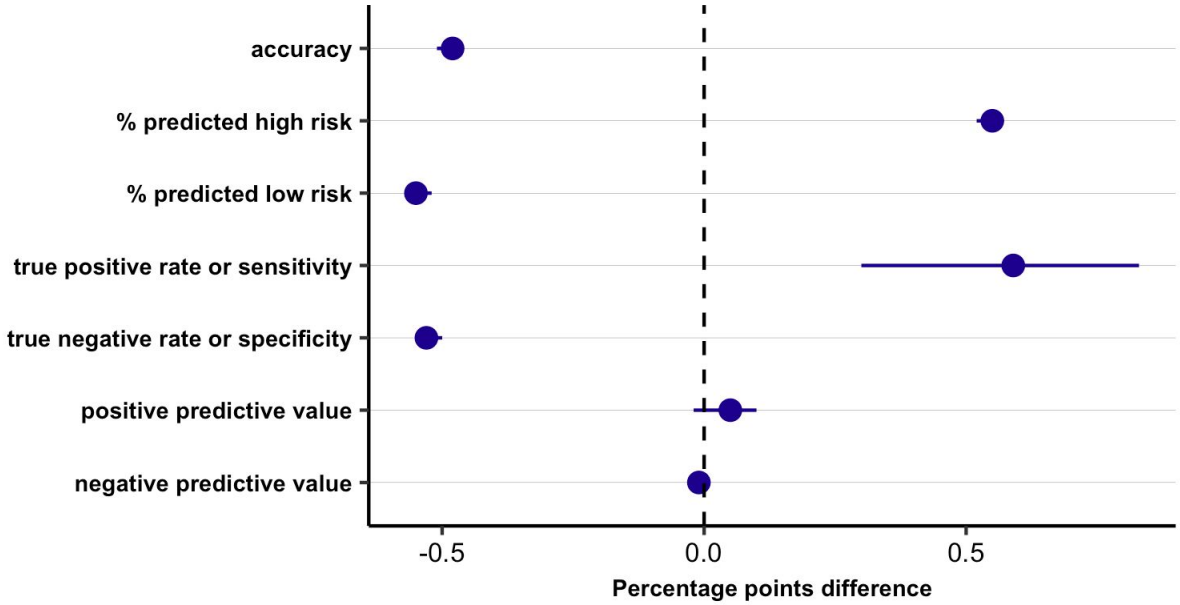
The true positive rate is almost 60 percentage points higher for male offenders than female offenders (89% and 30%). Which means that more male offenders who recidivate were correctly classified as high risk. Conversely, the true negative rate is 53 percentage points higher for female offenders (93% vs. 40%). More female offenders who did not recidivate were correctly classified as low risk.

The predictive value equality, which includes the positive and negative predictive values, shows no statistical differences between male and females offenders. Conditionally on being identified as low risk, both gender groups have a 98% probability of not recidivating for a crime against a person. Conditional on being identified as high risk, there is an 11% and 6% probability for male and female offenders, respectively, of recidivating for a crime against a person. These differences are not statistically significant, so conditionally on being identified as high risk, male and female offenders have the same probability of recidivating for a crime against a person.

Finally, the treatment equality is different between male and female offenders.[39] In relative terms, there are 10 times more false positives than false negatives for male offenders than for female offenders.

[39] Table A7.4 includes the point estimate of the treatment equality for both gender groups.

Figure 7.5 Fairness metrics differences between Male and Female offenders



Note: The difference is estimated as Male minus Female. 95% confidence intervals. Estimations based on the validation dataset.

8. Potential Improvements

Sensitivity Analysis of Risk Cutoffs

Currently, the general SRAI identifies low-risk offenders as those who score 0-4 points on the risk scale, typical risk offenders as those who score 5-9 points, and high-risk offenders as those who score 10-18 points. The sensitivity analysis analyzed how the performance and fairness metrics changed when the cutoffs were changed from the default values.

Table 8.1 Summary of Cutoff Sensitivity Analysis

Option	Cutoff Low	Cutoff High	Accuracy	F1	F1 Difference	Treatment Difference	High Risk Population
A	0-3	9-18	✘	✔	✔	✘	28.0%
B	0-3	10-18	✘	✔	✔	✘	17.0%
X	0-4	10-18	- (Default) -				17.0%
C	0-4	11-18	✔	✘	✘	✔	9.8%
D	0-5	11-18	✔	✘	✘	✔	9.8%
E	0-5	14-18	✔	✘	✔	✔	0.8%
F	0-6	13-18	✔	✘	✔	✔	2.0%

The sensitivity analysis results demonstrated that accuracy worsened when the low-risk cutoff was reduced to 0-3 (Table 8.1, options 1, 2), which effectively reduced the low-risk group size. Conversely, accuracy improved when the high-risk cutoff was increased (Table 8.1, options 3, 4, 5, 6).

The improvement of the instrument's accuracy score was the result of the instrument being more selective of who it deemed high risk. Increasing the high-risk cutoff to 11 improved overall accuracy by five percentage points, to 73, but captured only 27% of the population. A cutoff increase to 12 improved the instrument accuracy to 78%, with a captured population of 22% (Table 8.2). Fewer offenders are assigned to the high-risk group when using these alternative cutoffs, but those who were assigned were more likely to recidivate.

Table 8.2 Accuracy and Group Size by Cutoffs

Cutoff	Accuracy	Group Size	Accuracy*group size
L(0-4) H(10-18)	0.68	0.35	0.24
L(0-4) H(11-18)	0.73	0.27	0.20
L(0-4) H(12-18)	0.78	0.22	0.17

Fairness metrics between White and Black offenders improved when the cutoffs for either the low-risk or the high-risk group increased (Table 8.3). These changes effectively increased the size of the low-risk group, and decreased that of the high-risk group. The green cells in Table 8.3 indicate the metrics which improved from the baseline values at the default risk cutoffs.

Table 8.3 Metric Differences Between Black and White Offenders at Alternative Cutoffs

Differences in Percentage Points		Original L (0-4), H(10-18)	L (0-4), H (11-18)	L (0-4), H (12-18)	L (0-5), H (10-18)	L (0-5), H (11-18)
Overall accuracy	Accuracy	0.05	0.03	0.02	0.04	0.02
Demographic parity	% predicted high risk	0.16	0.13	0.07	0.07	0.09
	% predicted low risk	0.16	0.13	0.07	0.07	0.09
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.10	0.11	0.09	0.04	0.12
	True negative rate or specificity	0.14	0.10	0.04	0.08	0.06
Predictive Value Equality	Positive predictive value	0.01	0.01	0.03	0.08	0.01
	Negative predictive value	0.01	0.01	0.01	0.01	0.01
Treatment Equality	Ratio of false positive and negative	1.89	0.82	0.22	0.48	0.38

No Pareto efficient alternative exists. That is, no alternative cutoffs improved all metrics; the improvement of some metrics corresponded to a worsening of others. However, overall accuracy, as well as racial fairness, improved when either the low-risk cutoff or the high-risk cutoff was increased. These improvements were the result of increasing the population designated low risk and decreasing the population designated high risk.

Optimizing Risk Category Weights

Another strategy to improve the SRAI involved adjusting the weights associated with each of the seven risk categories. These weights were determined using a linear programming strategy that optimized the instrument to achieve the highest accuracy.

The results showed that no global optimal solution could be found, but that many local optimal solutions existed. That is, many sets of weights on the risk categories could result in improved accuracy, but no single set dominated (Table 8.4). Further analysis showed that, similar to the findings in the sensitivity analysis, no Pareto efficient solution exists and improvements in accuracy resulted in lower F1-scores. Furthermore, analysis of the results showed that the optimization model improved accuracy by setting weights so fewer individuals were captured in the high-risk group. The underlying mechanism that improved instrument performance was therefore the reduction in the number of individuals classified in the high-risk group, which could be achieved by either shrinking the range of the high-risk group or by adjusting the risk category weights.

Table 8.4 Local Optimal Weights and Results from Optimization

Solution	Gender	Age	Crime Category	Multiple Charges	Prior Convictions	Sum Prior Offenses	Juvenile Record	Accuracy	F1
Default	1	1	1	1	1	1	1	0.69	0.62
1	1.47	0.48	1.13	0.88	0.54	0.22	0.72	0.83	0.09
2	1.21	0.83	0.82	0.27	0.66	0.24	0.17	0.83	0.18
3	1.28	0.50	1.58	0.04	0.62	0.23	0.74	0.83	0.22

It should be noted that optimizing the instrument by adjusting the risk category weights gives more flexibility in determining risk group membership, but the interpretability of these weights decreases. This is because the weights are derived through a statistical method that reduces the likelihood of misidentifying recidivators, and do not have a sociological basis. Additionally, similar improvements can be achieved by adjusting risk cutoffs by optimizing the weight, but this lacks interpretability. Although this model can be further adjusted to improve interpretability by adding constraints such as linearity or integrality to the weights, any solutions found can perform no better than the current solution.

Optimizing Risk Factor Values

The final optimization strategy explored adjusting the values assigned to each risk factor. This strategy gives greater flexibility in how the model associates risk values to offender characteristics and behavior. Whereas optimizing weights at the risk category level allows adjustment in seven decision variables, optimizing at the risk factor level allows adjustment in 31 decision variables. This added complexity makes computation challenging and the limited results from this optimization have not been validated. However, the preliminary results suggest that even with greater flexibility, accuracy could not be significantly improved from optimizing at the risk category level or from adjusting the risk cutoffs. Furthermore, the mechanism by which this optimization strategy improved accuracy was also by reducing the size of the high-risk group.

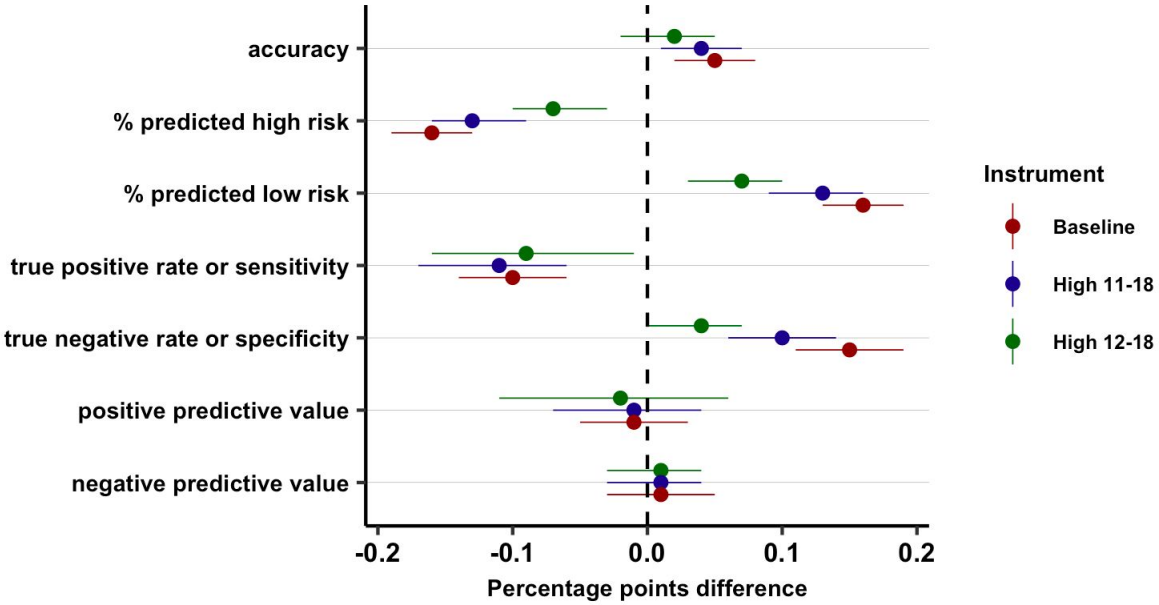
Optimizing at the risk factor level was rejected as an optimization strategy because of the added complexity and lack of significant improvement in performance as compared to either optimization at the risk category level, or adjustments to the risk cutoffs.

Adjusting Risk Cutoffs and Assessing Fairness

The sensitivity analysis showed that moving the risk cutoffs could improve some performance and fairness metrics in comparison to the current cutoffs. Additional analysis was performed to quantify the improvements in fairness along gender and race as the high-risk cutoffs were raised (and the low-risk range was kept at 0 to 4). Performance metrics were calculated for White and Black offenders, as well as for males and females. The difference between subpopulation metrics quantified the fairness between groups, and a bootstrapping strategy was used to give confidence intervals of these differences and determine statistical significance.

Figures 8.1 and 8.2 show the difference, along with the 95% confidence interval, between the metrics of each subpopulation along race and gender respectively. With respect to race, as the high-risk cutoff increased, the difference between White and Black offender metrics trend towards zero, indicating the instrument is treating both subpopulations equally (Appendix A8.1, A8.3). This is indicated by the points closer to the zero vertical line. Although a correlation between higher high-risk cutoffs and fairness is apparent, most metrics do not achieve parity at statistically significant levels until the cutoff is raised to 12.

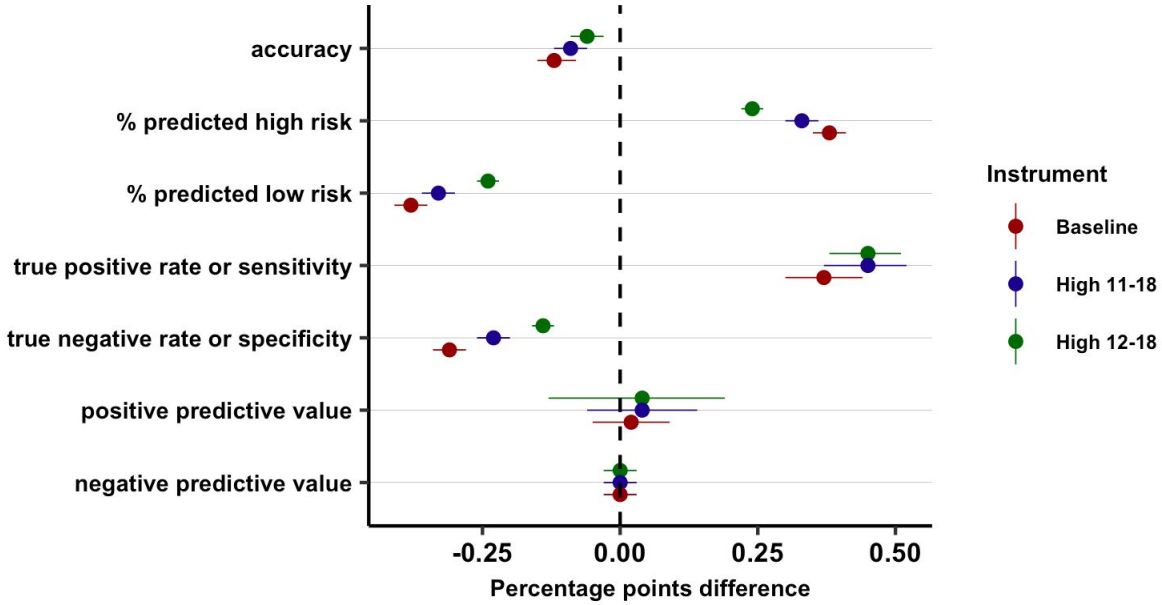
Figure 8.1 Fairness Metric Differences Between White and Black Offenders



Note: The difference is estimated as White minus Black offenders, with 95% confidence intervals

With respect to gender, the difference between male and female metrics also trended to zero as the high-risk cutoff was increased, indicating improved fairness. However, even at a high-risk cutoff of 12, males and females had statistically significant differences on the order of half a percentage point in most metrics (Appendix, A8.2, A8.4).

Figure 8.2 Fairness Metric Differences Between Gender



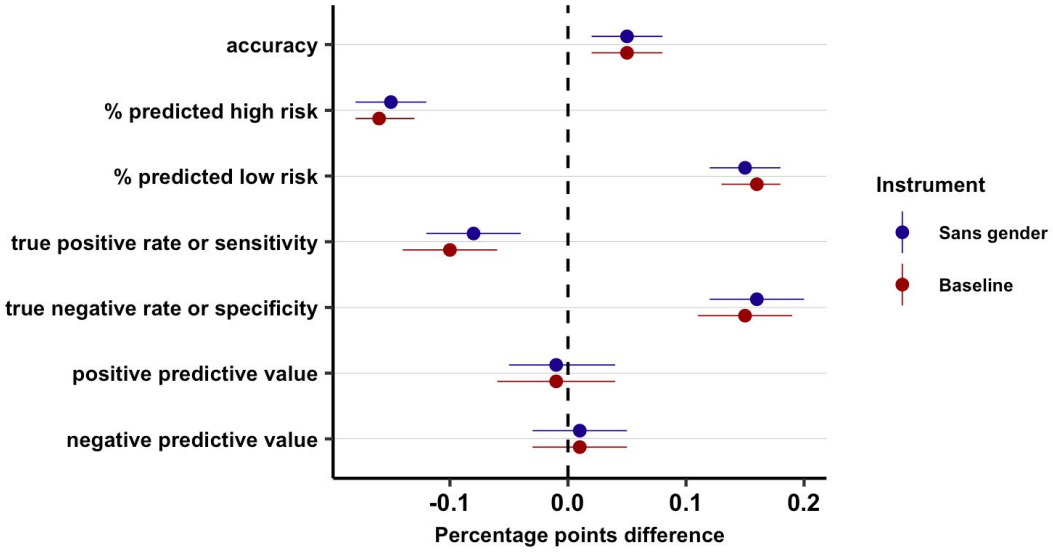
Note: The difference is estimated as male minus female, with 95% confidence intervals.

Removing Gender

The instrument's performance was analyzed after the removal of gender as a risk factor. This modification was done to better understand the implications of removing gender in case of legislative action that mandated its removal. The removal of gender is equivalent to reducing all male offender risk scores by one point, and keeping female offender risk scores unchanged. This reduces the maximum possible risk score to 17. If all risk cutoffs were shifted one point lower, 0-3 for low and 9-17 for high, the instrument performed on par with the default instrument on most metrics, better on sensitivity, and worse on specificity. If the risk cutoffs remained unchanged, 0-4 for low and 10-17 for high, accuracy would increase to 74 percent, six percentage points higher than the default instrument with gender (Table 8.5).

Fairness metrics between Black and White offenders improved or did no worse when gender was removed and the cutoffs were shifted one point lower (Figure 8.3). These improvements were also found when the cutoffs remained unchanged, at 0-4 for low and 10-17 for high (Appendix A8.4).

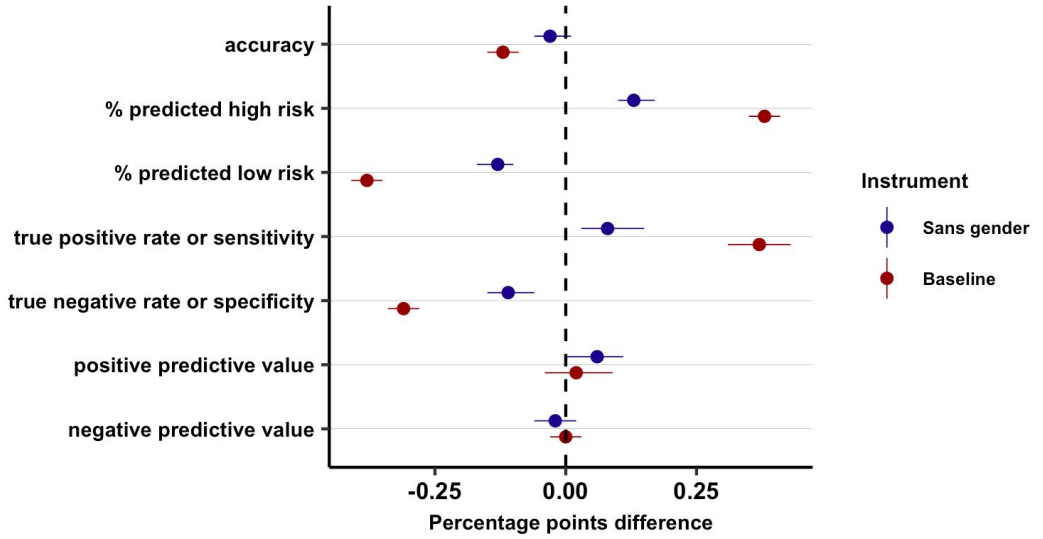
Figure 8.3 Fairness Metric Differences Between White and Black Offenders



Note: The difference is estimated as male minus female, with 95% confidence intervals.

Fairness metrics for gender improved after the removal of gender as a risk factor. Figure 8.4 shows the improvements in gender fairness between the default instrument as the baseline and the modified model with risk cutoffs set at 0-3 for low, and 9-17 for high. Points closer to the vertical zero line indicate better fairness between gender. The figure shows statistically significant improvements in most metrics, and near parity for others.

Figure 8.4 Fairness Metric Differences Between Gender



Note: The difference is estimated as male minus female, with 95% confidence intervals.

After the removal of gender, the performance of the general SRAI at two risk cutoffs was considered. The first shifted the risk cutoffs to 0-3 for low, and 9-17 for high, and the second left the risk cutoffs unchanged at 0-4 for low, and 10-17 for high. Analysis of the performance metrics of both options showed that neither solution was dominant over the other (Table 8.5).

Table 8.5 Performance Metrics Sans Gender

Metric	Original L (0-4), H (10-18)	Sans Gender L (0-3), H (9-17)	Sans Gender L (0-4), H (10-17)
Accuracy	0.68	0.67	0.74
Precision	0.52	0.51	0.54
Sensitivity	0.77	0.82	0.58
Specificity	0.64	0.58	0.81
F1-Score	0.62	0.63	0.56
G-Means	0.70	0.69	0.68

9. Summary and Recommendations

Evaluating risk assessment instruments can be difficult because there is no standard for ‘acceptable’ values of accuracy, fairness, and performance. The most common metrics reported, accuracy and AUC, cannot communicate how the instrument performs and treats different subpopulations. These metrics do, however, represent the most common set of benchmarks to compare the PCS SRAIs against. Based on these standards, the SRAI falls within industry standards.

Replication of the logistic regression, SRAI, and the results of the general and crime against a person recidivism scales was successful and matched the PCS’s initial results. This external validation of the PCS model and instruments shows that their process was errorless and robust.

The general SRAI performs moderately well in terms of AUC and performance, although there are some differences in how it treats Black and White offenders. These differences are larger for classification of high-risk offenders than low-risk offenders.

The CAP SRAI performs poorly based on performance metrics, such as F1-score, and treats of Black and White offenders and male and female offenders differently. While different base rates between offender subpopulations impact the performance of both risk scales, the CAP SRAI is especially impacted by the very low likelihood of recidivism for a crime against a person.

Recommendation 1: Restrict General SRAI Usage

On the most common performance metrics, the general SRAI performs similarly to risk instruments used in other states. Its overall AUC is similar to its peers at 0.66. When broken out by subpopulation, its performance is consistent, scoring a 0.66 and 0.66 for both White and Black offenders and 0.66 and 0.65 for male and female offenders, respectively.

Additional performance and fairness assessment metrics revealed some of the weaknesses in the general SRAI. On overall accuracy, the general SRAI scores relatively high at 0.68, but the error rate when classifying high-risk offenders is much higher than when classifying low-risk offenders. In addition, there are statistically significant differences in how the general SRAI treats Black and White offenders, as well as male and female offenders. As the previous analysis showed (see Figures 7.2

and 7.3), the general SRAI does not treat subpopulations equally. Due to the different base rates in recidivism between each subpopulation, however, it is practically unattainable to achieve both a high level of accuracy and fairness.

Based on these results, the first recommendation is to only use the general SRAI in its current form to identify low-risk offenders. The general SRAI more accurately predicts low-risk offenders than high-risk offenders. With a high error rate of 0.48, the general SRAI is correctly classifying an offender as high risk only 48% of the time. However, its low error rate is 0.16, meaning it correctly classify low-risk offenders 84% of the time. Without alteration, the general SRAI should therefore be used only to identify low-risk offenders.

Recommendation 2: Do Not Implement CAP SRAI

Similar to the general SRAI, the CAP SRAI performs well on AUC, with an overall score of 0.65. However, it deviates greatly on the additional performance and fairness metrics (see Figures 7.4 and 7.5). This is due to its poor performance in correctly classifying high-risk offenders. As the risk of recidivism for a crime against a person is extremely low, the CAP SRAI scores highly on metrics that assess its low-risk classification, such as sensitivity. Its success at classifying low-risk offenders pushes up its overall accuracy and AUC scores.

With its poor performance metrics and poor ability to predict high-risk offenders, the second recommendation is to not utilize the CAP SRAI in full nor in part (such as to only identify low-risk offenders). The CAP SRAI's overall poor performance on metrics such as precision, specificity, and F1-score suggests that the entire instrument is of low quality. This is in comparison to the general SRAI, whose performance metrics performed reasonably well but had issues of fairness between subpopulations. Using the CAP SRAI to classify only low-risk offenders when the entire instrument performs poorly is akin to cherry-picking specific results. Without alteration, the CAP SRAI should not be implemented.

Recommendation 3: Alter Cutoffs for the General SRAI

The third recommendation is to move the high-risk cutoff from 10 to 12 to improve the performance of the general SRAI. While this results in the general SRAI classifying fewer people overall, its accuracy and fairness metrics both increase. With this alteration, the general SRAI could then be used to classify both low- and high-risk offenders.

As the previous analysis has shown, increasing the high-risk cutoff and keeping the low-risk cutoff the same decreases the overall performance of the general SRAI, as shown in the F1-score. However, through classifying fewer offenders as high risk, the general SRAI treats subpopulations more fairly. These changes are only statistically significant when the high-risk cutoff increases by at least two points. Additional increase in the high-risk cutoff reduces the overall population classified, but it is also less error-prone.

Recommendation 4: Remove Gender and Alter Cutoffs for General SRAI

A major concern of the PCS is that any risk assessment instrument that scores offenders differently based on gender would not pass the Constitution of Pennsylvania's equal protection clause. While gender is a common and significant predictor of recidivism in the literature and other risk assessment instruments, removing it has no measurable statistical impact on the accuracy and fairness of the general SRAI. As previously shown, the improvements in fairness were particularly significant for female offenders.

The fourth and final recommendation is to remove gender from the general SRAI. Future analyses should determine exactly how removing gender improves the instrument beyond altering the low or high-risk cutoffs by one point. On first glance, accuracy improves because the general SRAI is classifying fewer offenders overall. It is also classifying fewer females, who are less represented in the offender population and have a lower base rate of recidivism.

Additional analysis is required to determine the exact ramifications of removing gender from the instrument. However, preliminary results show that gender can be removed from the general SRAI without any statistically different results on its accuracy or performance while also achieving statistically significant increases in fairness between male and female offenders.

References

Section 3: Pennsylvania Risk Assessment Instrument

Gonzales, Alberto R., Tracy A. Henke, and J. Robert Flores. “Changing Data into Valid Instruments for Juvenile Courts The Mathematics of Risk Classification: Report.” Washington, D.C., 2005. www.ojp.usdoj.gov/ojjdp.

“Interim Report 8 Communicating Risk at Sentencing.” State College, PA, 2014.

“Risk Assessment Project Phase III: The Development and Validation of the Proposed Risk Assessment Scales.” State College, PA, 2018. http://pcs.la.psu.edu/publications-and-research/risk-assessment/phase-iii-reports/development-and-validation-of-the-risk-assessment-scale/view.

Section 4: History of Risk Assessment Instruments and Decision Making

Cravez, Pamela. "Pretrial Risk Assessment Tool Developed for Alaska." Alaska Justice Forum, Winter 2018, 34, no. 3. Accessed February 15, 2019. https://scholarworks.alaska.edu/bitstream/handle/11122/8087/ajf.343.winter2018.online.pdf?sequence=2.

Danner, Monda; VanNostrand Marie; and Spruance, Lisa. “Race and Gender Neutral Pretrial Risk Assessment, Release Recommendations, and Supervision: VPRAI and PRAXIS Revised.” Luminosity, Inc., Risk Assessment, November 2016. Accessed February 15, 2019. https://www.ncsc.org/~/media/Microsites/Files/PJCC/Danner VanNostrand Spruance 2016 VPRAI Praxis revised.ashx.

Douglas, T; Pugh, J; Singh, I; Savulescu, J; and Fazel, S. “Risk Assessment Tools in Criminal Justice and Forensic Psychiatry: The Need for Better Data” European Psychiatry: The Journal of the Association of European Psychiatrists vol. 42 (2017), 134-137. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5408162/#bib0005.

Electronic Privacy Information Center, "EPIC - Algorithms in the Criminal Justice System," Electronic Privacy Information Center. Accessed February 15, 2019. https://epic.org/algorithmic-transparency/crim-justice/.

Kleinberg, Jon; Lakkaraju, Jimabindu; Leskovec, Jure; Ludwig, Jens; and Mullainathan, Sendhil. "Human Decisions and Machine Predictions," Author Manuscript, February 2018. doi:10.3386/w23180.

Krauss, Daniel. "Adjusting Risk of Recidivism: Do Judicial Departures Worsen or Improve Recidivism Prediction under the Federal Sentencing Guidelines?" Behavioral Sciences and the Law, 22, no. 6 (2004): 731-750.

Latessa, Edward; Lemke Richard; Makarios, Matthew; Smith, Paula; and Lowenkamp. Christopher. "The Creation and Validation of the Ohio Risk Assessment System (ORAS)," Federal Probation, vol. 74, no. 1 (June 2010): 16-22.

Latessa, Edward; Manchak, Sarah; Lux, Jennifer; Newsome, Jamie; Lugo, Melissa, and Papp, Jordan. “The Ohio Risk Assessment System (ORAS): A Re-Validation and Inter-Rater Reliability Study,” Draft Final Report, Corrections Institute, University of Cincinnati, October 31, 2017. Accessed February 15, 2019, https://www.uc.edu/content/dam/cech/centers/ccjr/docs/ORAS_ReValidation_Final_Report-2.9.18.pdf.

"Ohio Risk Assessment System," ODRC. Accessed February 15, 2019, https://drc.ohio.gov/oras.

Ruback, Barr; Kempinen, Cynthia; Tinik, Leigh; and Knoth, Lauren. "Communicating Risk Information at Criminal Sentencing in Pennsylvania: An Experimental Analysis," Federal Probation, vol. 80, no. 2 (September 2016): 47-56.

Sarah L Desmarais and Jay P Singh, “Risk Assessment Instruments Validated and Implemented in Correctional Settings in the United States: An Empirical Guide,” 2013.

United State of America, State of Alabama, Department of Corrections, Community Corrections Division, “Alabama Department of Corrections Minimum Standards for Community Punishment and Corrections Programs,” 2016. Accessed February 15, 2019, http://www.doc.state.al.us/docs/AlabamaMinimumStandardsforCCP.pdf.

United State of America, State of Virginia, Department of Criminal Justice Services, “Pretrial Risk Assessment in Virginia,” Luminosity, 2009. Accessed February 15, 2019, https://www.dcjs.virginia.gov/sites/dcjs.virginia.gov/files/publications/corrections/virginia-pretrial-risk-assessment-report.pdf.

Section 7: Performance and Fairness Metrics

Chouldechova, Alexandra. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments,” arXiv:1610.07524 (2016), https://arxiv.org/abs/1610.07524v1.

Chouldechova, Alexandra; Roth, Aaron. “The Frontiers of Fairness in Machine Learning,” arXiv:1810.08810 (2018), https://arxiv.org/abs/1810.08810.

Singh, Jay P. "Predictive Validity Performance Indicators in Violence Risk Assessment: A Methodological Primer." Behavioral Sciences & the Law 31, no. 1 (2013): 8-22. doi:10.1002/bsl.2052, https://doi.org/10.1002/bsl.2052.

Berk, Richard, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. "Fairness in Criminal Justice Risk Assessments." Sociological Methods & Research, May 30, 2018. doi:10.1177/0049124118782533.

Appendix

Table A6.1 Full Sample Logistic Regression Predicting Conviction Recidivism for Any Crime during 3-Year follow-up

	Odds Ratio	
	PCS	**Heinz**
Black	1.147***	1.143***
Hispanic	0.889**	0.886***
Male	1.319***	1.318***
Semi- Urban	0.955	0.952
Rural	0.918*	0.917*
Philadelphia	0.909*	0.908*
<21	5.150***	5.166***
21-25	3.187***	3.189***
26-29	2.393***	2.392***
30-39	2.070***	2.067***
40-49	1.584***	1.583***
Personal	0.842***	0.865***
Sex	0.580***	0.590***
Drug	0.881***	0.898***
Firearms/Weapons	0.978	1.002
Other	0.865***	0.926**
Current Convictions	1.090***	1.092***
1 Prior Conviction	1.608***	1.608***
2-3 Prior Convictions	2.097***	2.095***
4-5 Prior Convictions	2.650***	2.648***
6+ Prior Convictions	3.709***	3.714***
Prior Personal/Sex conviction	0.972	0.97
Prior Property conviction	1.131***	1.137***
Prior Drug conviction	1.064*	1.064*
Prior Firearm/Weapon conviction	0.833***	0.832***
Prior Public Order conviction	1.073*	1.072*
Prior Public Adm conviction	1.182***	1.179***
Prior DUI conviction	1.056	1.053

Juvenile Adjudication (Yes)	1.420***	1.418***
OGS 5/8	0.941**	0.943*
OGS 9/14	0.746***	0.747***
_cons	0.104***	0.102***
N	65,532	65,532
AIC	77522.687	77540.61

Note: Full Sample Logistic Regression Predicting 3-Year Conviction Recidivism for Any Crime (using convictions) among 2004-2006 Development Dataset.

Table A6.2 Full Sample Logistic Regression Predicting Conviction Recidivism for a Crime Against a Person during 3-Year follow-up

	Odds Ratio	
	PCS	**Heinz**
Black	1.274***	1.281***
Hispanic	1.005	1.01
Male	1.886***	1.874***
Semi-urban	1.053	1.053
Rural	1.097	1.096
Philadelphia	1.146	1.149
<21	3.719***	3.697***
21-25	2.301***	2.294***
26-29	1.839***	1.836***
30-39	1.559***	1.558***
40-49	1.364**	1.362**
Personal/Sex	1.641***	1.680***
Drug	0.826***	0.846**
Firearm/Weapon	1.524***	1.550***
Other	1.243***	1.242***
Multiple Current Convictions	1.051	1.048
1 Prior Conv	1.434***	1.434***
2-3 Prior Conv	1.609***	1.610***
4-5 Prior Conv	1.835***	1.836***
6+ Prior Conv	1.822***	1.822***
Prior Personal conv	1.577***	1.577***

Prior Property conv	0.962	0.96
Prior Drug conv	0.844**	0.844**
Prior Firearm/Weapon conv	0.948	0.947
Prior Public Order conv	1.315***	1.313***
Prior Public Adm conv	1.141*	1.143*
Prior DUI conv	1.068	1.068
Prior Juvenile Adjudication	1.351***	1.350***
OGS 5/8	0.889**	0.867***
OGS 9/14	0.586***	0.583***
_cons	0.009***	0.009***
N	65,532	65,532
AIC	25686.935	25682.41

Note: Full Sample Logistic Regression Predicting 3-Year Conviction Recidivism for Any Crime (using convictions) among 2004-2006 Development Dataset.

Table A7.1 Fairness Metrics Based on Race for General SRAI

Category	Metric	White and Others*	Black	Statistical Difference
Overall accuracy	Accuracy	0.70	0.65	Yes
Demographic parity	Percentage predicted high-risk	0.45	0.61	Yes
	Percentage predicted low-risk	0.55	0.39	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.73	0.83	Yes
	True negative rate or specificity	0.68	0.54	Yes
Predictive Value Equality	Positive predictive value	0.52	0.53	No
	Negative predictive value	0.84	0.83	No
Treatment Equality	Ratio of false positive and negative	2.44	4.33	Yes

Note: Estimations made on the validation set. Statistical difference was estimated using bootstrapping.
*White includes other races except for Hispanic.

Table A7.2 Fairness Metrics Based on Gender for General SRAI

Category	Metric	Male	Female	Statistical Difference
Overall accuracy	Accuracy	0.65	0.77	Yes
Demographic parity	Percentage predicted high-risk	0.59	0.21	Yes
	Percentage predicted low-risk	0.41	0.79	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.82	0.45	Yes
	True negative rate or specificity	0.56	0.87	Yes
Predictive Value Equality	Positive predictive value	0.52	0.50	No
	Negative predictive value	0.84	0.84	No
Treatment Equality	Ratio of false positive and negative	4.18	0.81	Yes

Note: Estimations made on the validation set. Statistical difference was estimated using bootstrapping.

Table A7.3 Fairness Metrics Based on Race for CAP SRAI

Category	Metric	White and Others*	Black	Statistical Difference
Overall accuracy	Accuracy	0.56	0.51	Yes
Demographic parity	Percentage predicted high-risk	0.48	0.54	Yes
	Percentage predicted low-risk	0.52	0.46	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.85	0.87	No
	True negative rate or specificity	0.54	0.49	Yes
Predictive Value Equality	Positive predictive value	0.10	0.11	No
	Negative predictive value	0.98	0.98	No
Treatment Equality	Ratio of false positive and negative	49.66	54.96	No

Note: Estimations made on the validation set. Statistical difference was estimated using bootstrapping.
*White includes other races except for Hispanic.

Table A7.4 Fairness Metrics Based on Gender for CAP SRAI

Category	Metric	Male	Female	Statistical Difference
Overall accuracy	Accuracy	0.43	0.92	Yes
Demographic parity	Percentage predicted high-risk	0.62	0.08	Yes
	Percentage predicted low-risk	0.38	0.92	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.89	0.30	Yes
	True negative rate or specificity	0.4	0.93	Yes
Predictive Value Equality	Positive predictive value	0.11	0.06	No
	Negative predictive value	0.98	0.99	No
Treatment Equality	Ratio of false positive and negative	67.06	6.93	Yes

Note: Estimations made on the validation set. Statistical difference was estimated using bootstrapping.

Table A8.1 Fairness Metrics Based on Race for General SRAI with Alternative Cutoffs

Category	Metric	White and Others	Black	Statistical Difference
Overall accuracy	Accuracy	0.74	0.71	No
Demographic parity	Percentage predicted high-risk	0.32	0.45	Yes
	Percentage predicted low-risk	0.68	0.55	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.62	0.73	Yes
	True negative rate or specificity	0.79	0.69	Yes
Predictive Value Equality	Positive predictive value	0.54	0.55	No
	Negative predictive value	0.84	0.83	No
Treatment Equality	Ratio of false positive and negative	1.38	2.20	Yes

Note: Estimations made on the validation set. Confidence intervals were built using the bootstrapping method. Risk scores go from 0 to 4 for low risk and from 11 to 18 for high risk.

Table A8.2 Fairness Metrics Based on Gender for General SRAI with Alternative Cutoffs

Category	Metric	Male	Female	Statistical Difference
Overall accuracy	Accuracy	0.71	0.80	Yes
Demographic parity	Percentage predicted high-risk	0.45	0.11	Yes
	Percentage predicted low-risk	0.55	0.89	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.73	0.29	Yes
	True negative rate or specificity	0.70	0.93	Yes
Predictive Value Equality	Positive predictive value	0.54	0.50	No
	Negative predictive value	0.84	0.84	No
Treatment Equality	Ratio of false positive and negative	2.28	0.40	Yes

Note: Estimations made on the validation set. Confidence intervals were built using the bootstrapping method. Risk scores go from 0 to 4 for low risk and from 11 to 18 for high risk.

Table A8.3 Fairness Metrics Based on Race for General SRAI with Alternative Cutoffs

Category	Metric	White and Others	Black	Statistical Difference
Overall accuracy	Accuracy	0.79	0.77	No
Demographic parity	Percentage predicted high-risk	0.20	0.27	Yes
	Percentage predicted low-risk	0.8	0.73	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.47	0.56	No
	True negative rate or specificity	0.89	0.85	No
Predictive Value Equality	Positive predictive value	0.57	0.60	No
	Negative predictive value	0.84	0.83	No
Treatment Equality	Ratio of false positive and negative	0.66	0.88	No

Note: Estimations made on the validation set. Confidence intervals were built using the bootstrapping method. Risk scores go from 0 to 4 for low risk and from 12 to 18 for high risk.

Table A8.4 Fairness Metrics Based on Gender for General SRAI with Alternative Cutoffs

Metric	Category	Male	Female	Statistical Difference
Overall accuracy	Accuracy	0.76	0.83	Yes
Demographic parity	Percentage predicted high-risk	0.29	0.05	Yes
	Percentage predicted low-risk	0.71	0.95	Yes
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.59	0.14	Yes
	True negative rate or specificity	0.83	0.97	Yes
Predictive Value Equality	Positive predictive value	0.58	0.54	No
	Negative predictive value	0.84	0.84	No
Treatment Equality	Ratio of false positive and negative	1.06	0.14	Yes

Note: Estimations made on the validation set. Confidence intervals were built using the bootstrapping method. Risk scores go from 0 to 4 for low risk and from 12 to 18 for high risk.

Table A8.5 Fairness Metrics Based on Race for General SRAI Sans Gender

Category	Metric	Original L (0-4), H(10-18)	L (0-4), H (10-17)	L (0-3), H (9-17)
Overall accuracy	Accuracy	0.05	0.03	0.04
Demographic parity	Percentage predicted high-risk	0.16	0.11	0.16
	Percentage predicted low-risk	0.16	0.11	0.16
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.10	0.12	0.09
	True negative rate or specificity	0.14	0.07	0.15
Predictive Value Equality	Positive predictive value	0.01	0.03	0.02
	Negative predictive value	0.01	0.01	0.01
Treatment Equality	Ratio of false positive and negative	1.89	0.49	2.58

Table A8.6 Fairness Metrics Based on Gender for General SRAI Sans Gender

Category	Metric	Original L (0-4), H (10-18)	L (0-3), H (9-17)	L (0-4), H (10-17)
Overall accuracy	Accuracy	0.12	0.03	0.05
Demographic parity	Percentage predicted high-risk	0.38	0.14	0.12
	Percentage predicted low-risk	0.38	0.14	0.12
Conditional Procedure Accuracy Equality	True positive rate or sensitivity	0.37	0.10	0.13
	True negative rate or specificity	0.31	0.11	0.08
Predictive Value Equality	Positive predictive value	0.02	0.06	0.05
	Negative predictive value	0.00	0.01	0.03
Treatment Equality	Ratio of false positive and negative	3.37	1.32	0.35
Captured population		0.36	0.33	0.36