**Methodological Appendix**

This appendix provides more detail information about the techniques used to construct the algorithmic alternatives of the IRAS-PAT risk estimates. Specifically, we used logistic regression to create five algorithmic alternatives of the IRAS-PAT risk estimates, based on an adaptation of Skeem and Lowenkamp's (2020) algorithmic corrections for debiasing assessments. With each algorithm, we produced predicted probabilities separately for all outcomes to be used as an alternative estimate to the IRAS-PAT risk estimates during our reevaluation of differential prediction by race in IRAS-PAT assessments. Below, we outline the process we used to create *algorithm total scores* and *algorithm risk levels:*

* *Algorithm 1* *total scores* represented the “Race Omitted” algorithm. We generated predicted probabilities by regressing each misconduct outcome onto the seven IRAS-PAT items. This approach represents a “practice as usual” strategy, which omits race in the estimation of risk given it is a legally protected class group.
* *Algorithm 2 total scores* represented the “Race Fitted” algorithm. We generated predicted probabilities by regressing each misconduct outcome onto the main effects of the seven IRAS-PAT items and race and their interactions. By including the interaction, this approach adjusted for and removed from risk estimates any variability attributable to differences in item-level predictive accuracy in the prediction of misconduct outcomes.
* *Algorithm 3 total scores* represented the “Proxy Eliminated” algorithm. First, we regressed each misconduct outcome onto the seven IRAS-PAT items and race to generate regression coefficients (i.e., predictive weights) for each predictor. Race was included in the model to ensure the IRAS-PAT items’ regression coefficients were adjusted by its contribution. Second, we eliminated the variation of race in our sample by recoding all defendants as the same race (e.g., all “Black”, all “White”, or another fixed value). Consistent with prior approaches (Pope & Sydnor, 2011; Skeem & Lowenkamp, 2020), we choose to recode the values of race with the proportion of Black defendants in our sample (i.e., 0.19). Third, we calculated predicted probabilities in a logistic regression model by using the equation,

where *P* is the probability of misconduct, *B*0 and *B*1…*B*n represent the unstandardized regression coefficients generated in Step 1, and X1 … Xn represent the seven IRAS-PAT item values and the fixed value of race, which was created in Step 2. By including predictive weights for each item generated in Step 1 when calculating the alternative risk estimates, this approach ensured the unique, predictive strength of IRAS-PAT items was not influenced by the items’ association with race, while also instituting a policy control that removed race from the estimation of risk given it is a legally protected class group.

* *Algorithm 4 total scores* represented the “Race Eliminated” algorithm. First, we regressed each IRAS-PAT item onto race to generate residual values for each item. Second, we generated predicted probabilities by regressing each outcome onto the seven residual IRAS-PAT items. In practice, this removed all variance that race shared with the IRAS-PAT items when predicting risk of misconduct.
* *Algorithm 5 total scores* represented the “Criminal History Discount” algorithm. Point-biserial correlations showed the IRAS-PAT criminal history domain was positively associated with race (*r*[3,537] = .07, *p* ≤ .001. Criminal history is a well-established predictor of recidivism (Cottle et al., 2001; Gendreau et al., 1996), but strong evidence exists of racial disparities in police decision-making to arrest (Ousey & Lee, 2008; Tapia, 2011). A meta-analysis of studies investigating the effect of race on the police decision to arrest found people of color had a higher probability of being arrested (.26) compared to White suspects (.20; Kochel et al., 2011). Based on these differential rates, we reduced Black defendant’s criminal history domain scores by 30% to adjust for biases in police decision-making. Next, we generated predicted probabilities by regressing each pretrial misconduct outcome onto the discounted criminal history domain and remaining items.

*Algorithm risk levels* (Low; Moderate; High) were created based on the total score variant. For a given algorithm, we created a rank-ordered list of defendants’ predicted probabilities for each misconduct outcome and classified the top 12% as High risk, the bottom 43% as Low risk, and the remaining as Moderate risk. These cutoff thresholds coincided with the proportion of defendants within each IRAS-PAT risk level (see Table 1 in main article).

**References**

Cottle, C. C., Lee, R. J., & Heilburn, K. (2001). The prediction of criminal recidivism in juveniles: A Meta-analysis. *Criminal Justice and Behavior*, *28*(3), 367–394. https://doi.org/10.1177/0093854801028003005

Gendreau, P., Little, T., & Goggin, C. (1996). A meta-analysis of the predictors of adult offender recidivism: What works! *Criminology*, *34*(4), 575–608. https://doi.org/10.1111/j.1745-9125.1996.tb01220.x

Kochel, T. R., Wilson, D. B., & Mastrofski, S. D. (2011). Effect of suspect race on officers’ arrest decisions. *Criminology*, *49*(2), 473–512. https://doi.org/10.1111/j.1745-9125.2011.00230.x

Ousey, G. C., & Lee, M. R. (2008). Racial disparity in formal social control: An investigation of alternative explanations of arrest rate inequality. *Journal of Research in Crime and Delinquency*, *45*(3), 322–355. https://doi.org/10.1177/0022427808317575

Pope, D. G., & Sydnor, J. R. (2011). Implementing anti-discrimination policies in statistical profiling models. *American Economic Journal: Economic Policy*, *3*(3), 206–231. https://doi.org/10.1257/pol.3.3.206

Skeem, J. L., & Lowenkamp, C. T. (2020). Using algorithms to address trade‐offs inherent in predicting recidivism. *Behavioral Sciences & the Law*, *38*(3), 259–278. https://doi.org/10.1002/bsl.2465

Tapia, M. (2011). Gang membership and race as risk factors for juvenile arrest. *Journal of Research in Crime and Delinquency*, *48*(3), 364–395. https://doi.org/10.1177/0022427810393013

**Appendix A.** Confusion Matrix Exemplar

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Actual Outcome** | | **Accuracy Rate Equations** |
| **Test Prediction** | Failure – Positive Event | Success – Negative Event |
| Failure Predicted – Positive Test | *TP*  True Positives | *FP*  False Positives | *TP / (TP + FP)*  Positive Predictive Value |
| Success Predicted – Negative Test | *FN*  False Negatives | *TN*  True Negatives | *TN / (FN + TN)*  Negative Predictive Value |
| **Error Rate Equations** | *FN / (TP + FN)*  False Negative Rate | *FP / (FP + TN)*  False Positive Rate |  |

**Appendix B.** Confusion Matrices for Pretrial Misconduct Outcomes and IRAS-PAT Risk Estimates

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Overall** | | |  | **Black Defendants** | | |  | **White Defendants** | | |
|  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |
| Any FTA - No | 2,838 | 321 |  | Any FTA - No | 554 | 59 |  | Any FTA - No | 2,284 | 262 |
| Any FTA - Yes | 275 | 105 |  | Any FTA - Yes | 64 | 12 |  | Any FTA - Yes | 211 | 93 |
| FPR: 10.16% |  |  |  | FPR: 9.62% |  |  |  | FPR: 10.29% |  |  |
| FNR: 72.37% |  |  |  | FNR: 84.21% |  |  |  | FNR: 69.41% |  |  |
| PPV: 24.65% |  |  |  | PPV: 16.90% |  |  |  | PPV: 26.20% |  |  |
| NPV: 91.17% |  |  |  | NPV: 89.64% |  |  |  | NPV: 91.54% |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Overall** | | |  | **Black Defendants** | | |  | **White Defendants** | | |
|  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |
| Any new arrest - No | 2,664 | 290 |  | Any new arrest - No | 507 | 49 |  | Any new arrest - No | 2,157 | 241 |
| Any new arrest - Yes | 449 | 136 |  | Any new arrest - Yes | 111 | 22 |  | Any new arrest - Yes | 338 | 114 |
| FPR: 9.82% |  |  |  | FPR: 8.81% |  |  |  | FPR: 10.05% |  |  |
| FNR: 76.75% |  |  |  | FNR: 83.46% |  |  |  | FNR: 74.78% |  |  |
| PPV: 31.92% |  |  |  | PPV: 30.99% |  |  |  | PPV: 32.11% |  |  |
| NPV: 85.58% |  |  |  | NPV: 82.04% |  |  |  | NPV: 86.45% |  |  |

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Overall** | | |  | **Black Defendants** | | |  | **White Defendants** | | |
|  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |  |  | Not High Risk | High Risk |
| Any arrest - No | 2,274 | 171 |  | Any arrest - No | 406 | 34 |  | Any arrest - No | 1,868 | 137 |
| Any arrest - Yes | 839 | 255 |  | Any arrest - Yes | 212 | 37 |  | Any arrest - Yes | 627 | 218 |
| FPR: 6.99% |  |  |  | FPR: 7.73% |  |  |  | FPR: 6.83% |  |  |
| FNR: 76.69% |  |  |  | FNR: 85.14% |  |  |  | FNR: 74.20% |  |  |
| PPV: 59.86% |  |  |  | PPV: 52.11% |  |  |  | PPV: 61.41% |  |  |
| NPV: 73.05% |  |  |  | NPV: 65.70% |  |  |  | NPV: 74.87% |  |  |

*Note:* Indictors of statistical bias require a single cutoff threshold. As a result, we dichotomized the ordinal risk levels (i.e., Low; Moderate; High) in the IRAS-PAT into *Not High Risk* (defendants classified as Low/Moderate risk) and *High Risk* (defendants classified as High risk).

**Appendix C.** Change in Accuracy and Error Rates between IRAS-PAT Original Risk Estimates and Race Fitted Algorithms

Diagram

Description automatically generated

*Note:* Indictors of statistical bias require a single cutoff threshold. As a result, we dichotomized the ordinal risk levels (i.e., Low; Moderate; High) in the IRAS-PAT into *Not High Risk* (defendants classified as Low/Moderate risk) and *High Risk* (defendants classified as High risk).

**Appendix D.** AUC Values by Pretrial Misconduct Outcome and Race, using Race Fitted Algorithms

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Pretrial Outcomes** | **Black**  *n* = 688 | | |  | **White**  *n* =2,847 | | |
| AUC (SE) | 95% CI | Convention |  | AUC (SE) | 95% CI | Convention |
| Algorithm 2 total score |  |  |  |  |  |  |  |
| Any FTA | 0.66 (0.03) | [0.60, 0.72] | Good |  | 0.74 (0.01) | [0.72, 0.77] | Excellent |
| Any new arrest | 0.62 (0.03) | [0.56, 0.67] | Fair |  | 0.70 (0.01) | [0.67, 0.72] | Good |
| Any arrest | 0.65 (0.02) | [0.61, 0.69] | Good |  | 0.74 (0.01) | [0.72, 0.76] | Excellent |
| Algorithm 2 risk level |  |  |  |  |  |  |  |
| Any FTA | 0.58 (0.03) | [0.53, 0.63] | Fair |  | 0.72 (0.01) | [0.69, 0.74] | Excellent |
| Any new arrest | 0.56 (0.02) | [0.52, 0.59] | Fair |  | 0.67 (0.01) | [0.64, 0.69] | Good |
| Any arrest | 0.56 (0.01) | [0.53, 0.59] | Fair |  | 0.71 (0.01) | [0.69, 0.73] | Excellent |

**Appendix E.** Sensitivity analysis

We conducted additional sets of analyses to address specific criticisms that results of differential prediction in risk assessments would be due to baseline differences between groups or unequal sample sizes between groups, given the small number of Black defendants in the sample. First, to adjust for baseline differences between groups, we conducted propensity score matching using MatchIt in R and specifying a full matching procedure (Ho et al., 2011; Stuart & Green, 2008). Propensity scores measure the probability that a given individual will belong to a specific group (e.g., Black or White defendant) given known characteristics of that individual. For the purposes of this analysis, we matched White defendants to Black defendants based on county, age, gender, IRAS-PAT total score, time in the community, highest charge level, and charge types. Charge types were selected based on prevalence in the overall sample (i.e., > 10% of defendants). Weights generated from the propensity score matching procedure were then used in multivariable models (i.e., weighted models). Second, to address criticisms that findings would reflect unequal sample sizes between White and Black defendants, we developed a stratified sample of White defendants. This process involved collecting a random sample of White defendants from each of the individual county samples based on the number of Black defendants from each original county sample. The stratified sample shows whether the results are independent of sample size differences between Black and White defendants. These analytic decisions are discussed below. Model summaries are presented in Table F-1 to F-3, which include odds ratios from the unweighted sample. Significant (*p* < .05) odds ratios are bolded.

|  |  |
| --- | --- |
| **Summary of Analytic Considerations** | |
| **Regression Without the Use of Propensity Scores (Unweighted)** | Multivariable models unweighted by propensity scores, which do not address concerns about unbalanced groups, are helpful because they provide a baseline model for comparison to other models that integrate estimation and sampling considerations. |
| **Using Propensity Scores as Weights in a Regression** | Multivariable models weighted by propensity scores statistically balance individuals on a specific set of covariates based on their likelihood of being in a group. Black defendants were assigned a propensity score of 1, whereas White defendants were assigned a score above or below 1 that was calculated in the propensity score analysis based on the observed covariates. We then weighted estimations using these propensity scores. |
| **Regression with a Stratified Sample** | To balance the racial groups based on sample size, we developed a stratified sample of White defendants. We randomly selected White defendants within each county to create a sample of White defendants that was equal to the proportion of Black defendants within the respective county. This resulted in a stratified sample of 689 White defendants with the original 689 Black defendants (*N* = 1,378). |

**Table E-1.** Summary of Logistic Regression Models of IRAS-PAT Risk Estimates Predicting Any FTA, by Analytic Consideration

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Odds Ratio Status for Any FTA** | | |
| **Predictor** |
| **Unweighted**  *N* = 3,539 | **Weighted**  *N*= 3,539 | **Stratified**  *N =* 1,378 |
| **Model 1** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | 1.04 | 0.80 | **1.46** |
| ***Block 2*** |  |  |  |
| Black (White) | 1.05 | 0.82 | 1.37 |
| Total score | **1.47** | **1.48** | **1.38** |
| ***Block 3*** |  |  |  |
| Black X Total score | **0.79** | **0.78** | **0.74** |
| **Model 2** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | 1.04 | 0.80 | **1.46** |
| ***Block 2*** |  |  |  |
| Black (White) | 1.03 | 0.82 | 1.37 |
| Risk level (Low) |  |  |  |
| Moderate | **3.27** | **2.65** | **2.65** |
| High | **7.15** | **6.80** | **3.96** |
| ***Block 3*** |  |  |  |
| Black X Moderate | **0.43** | 0.58 | **0.33** |
| Black X High | **0.26** | **0.29** | 0.30 |

**Table E-2.** Summary of Logistic Regression Models of IRAS-PAT Risk Estimates Predicting Any New Arrest, by Analytic Consideration

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Odds Ratio Status for Any New Arrest** | | |
| **Predictor** |
| **Unweighted**  *N* = 3,539 | **Weighted**  *N*= 3,539 | **Stratified**  *N =* 1,378 |
| **Model 1** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | **1.27** | 1.04 | **1.36** |
| ***Block 2*** |  |  |  |
| Black (White) | **1.28** | 1.05 | 1.29 |
| Total score | **1.37** | **1.29** | **1.26** |
| ***Block 3*** |  |  |  |
| Black X Total score | 0.89 | 0.96 | 0.97 |
| **Model 2** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | **1.27** | 1.04 | **1.36** |
| ***Block 2*** |  |  |  |
| Black (White) | **1.27** | 1.05 | 1.30 |
| Risk level (Low) |  |  |  |
| Moderate | **2.49** | **2.26** | **1.94** |
| High | **4.84** | **3.27** | **2.65** |
| ***Block 3*** |  |  |  |
| Black X Moderate | **0.59** | 0.68 | 0.72 |
| Black X High | **0.50** | 0.81 | 1.15 |

**Table E-3**. Summary of Logistic Regression Models of IRAS-PAT Risk Estimates Predicting Any Arrest, by Analytic Consideration

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Odds Ratio Status for Any Arrest** | | |
| **Predictor** |
| **Unweighted**  *N* = 3,539 | **Weighted**  *N*= 3,539 | **Stratified**  *N =* 1,378 |
| **Model 1** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | **1.34** | 1.11 | **1.55** |
| ***Block 2*** |  |  |  |
| Black (White) | **1.36** | 1.14 | **1.46** |
| Total score | **1.53** | **1.55** | **1.39** |
| ***Block 3*** |  |  |  |
| Black X Total score | **0.82** | **0.80** | **0.87** |
| **Model 2** |  |  |  |
| ***Block 1*** |  |  |  |
| Black (White) | **1.34** | 1.11 | **1.55** |
| ***Block 2*** |  |  |  |
| Black (White) | **1.37** | 1.15 | **1.48** |
| Risk level (Low) |  |  |  |
| Moderate | **3.12** | **3.01** | **2.63** |
| High | **7.97** | **7.90** | **4.08** |
| ***Block 3*** |  |  |  |
| Black X Moderate | **0.60** | 0.63 | 0.62 |
| Black X High | **0.33** | **0.33** | 0.60 |

**References**

Ho, D., Stuart, E. A., King, G., & Imai, K. (2011). MatchIt: Nonparametric preprocessing for parametric causal inference. *Journal of Statistical Software*, *42*(8). https://doi.org/10.18637/jss.v042.i08

Stuart, E. A., & Green, K. M. (2008). Using full matching to estimate causal effects in nonexperimental studies: Examining the relationship between adolescent marijuana use and adult outcomes. *Developmental Psychology*, *44*(2), 395–406. https://doi.org/10.1037/0012-1649.44.2.395