Measuring Racial Discrimination in Algorithms: Online Appendix

David Arnold, UCSD
Will Dobbie, Harvard Kennedy & NBER
Peter Hull, University of Chicago & NBER

January 2021

I. Algorithm Estimation Details

This appendix details our baseline algorithmic predictions of pretrial misconduct risk. We use a gradient boosted decision tree model, based on the model that Kleinberg et al. (2018) develop for the NYC pretrial system. We use the same feature set X_i, which includes a total of 38 variables summarizing prior criminal history, charge characteristics, and demographic variables such as the age of the defendant. The outcome variable Y_i is an indicator for pretrial misconduct, defined as either a failure to appear or being arrested for a new crime.

Gradient boosting is an ensemble method that aggregates several "weak learners" in an iterative fashion. Here, the weak learners are decision trees, which divide the data through a sequence of binary splits based on the feature set. The algorithm averages multiple decision trees built sequentially on the data, with subsequent iterations up-weighting the observations predicted most poorly by the preceding sequence of trees. The complexity of the gradient boosting algorithm depends on the "depth" of each tree and a "shrinkage" parameter which governs how different trees are averaged together.

Following Kleinberg et al. (2018), we choose the model hyperparameters by k-fold cross-validation with five folds. We first select a random 80 percent sample of released defendants, which we take as the training dataset. Applying the cross-validation procedure to this dataset yields an optimal tree depth of 4 and shrinkage parameter of 0.05. We then use the full training set and the remaining 20 percent of released defendants (the test dataset) to fit the gradient boosted decision tree model with these hyperparameters. Finally, we apply the model to the full sample (including defendants detained before trial) to compute risk predictions $\hat{Y}_i$. We use the complete sample to estimate algorithmic discrimination for consistency with Arnold, Dobbie and Hull (2020) and to maximize precision.

II. Parameter Estimation Details

This appendix details our estimation of the four race-specific parameters $\theta = \{\mu_w, \mu_b, \rho_w, \rho_b\}$ and the discrimination measure Δ. As in Arnold, Dobbie and Hull (2020), we first estimate a series of judge- and race-specific estimates of release rates $\hat{\pi}_{jr}$, released misconduct rates $\hat{\mu}_{jr}$, and second moments of algorithmic release recommendations and misconduct potential among released defendants $\hat{\rho}_{jr}$. We obtain the $\hat{\pi}_{jr}$ by regressing release indicators D_i on judge-by-race interactions, controlling for court-by-time fixed effects (the level at which judges are as-good-as-randomly assigned in our sample; see Arnold, Dobbie and Hull (2020) for details). We obtain the $\hat{\mu}_{jr}$ and $\hat{\rho}_{jr}$ by regressing pretrial misconduct Y_i and its interaction with algorithmic recommendations $T_i Y_i$ on the same regressors in the subsample of released defendants. We then use these estimates to estimate mean risk μ_r and the second moments ρ_r by the vertical intercept, at one, of race-specific judge-level local linear regressions

of the $\hat{\mu}_{jr}$ and $\hat{\rho}_{jr}$ on the $\hat{\pi}_{jr}$. The local linear regressions use a Gaussian kernel with race-specific rule-of-thumb bandwidths, as in Arnold, Dobbie and Hull (2020). Finally, we use estimates of the four race-specific parameters as inputs for the formulas derived in the main text to estimate Δ.

We repeat this estimation procedure for a set of 20 release rate thresholds τ, yielding the range of algorithmic discrimination estimates in Figure 1 of the main text. We obtain pointwise 95% confidence intervals on this range by a clustered parametric bootstrap procedure, as in Arnold, Dobbie and Hull (2020). This bootstrap procedure draws from the asymptotic distribution of estimation error in the $\hat{\pi}_{jr}$, $\hat{\mu}_{jr}$, and $\hat{\rho}_{jr}$ estimates, where we allow for two-way clustering at the individual and judge level. We then recompute the extrapolation-based estimates of θ to estimate the estimation error distribution of Δ at each release rate.

Appendix Table A.1 reports our baseline local linear estimates of θ and Δ as well as alternative estimates based on simple linear and quadratic extrapolations. These results are broadly similar.

III. Alternative Regression-Based Risk Predictions

This appendix measures algorithmic discrimination in a simpler regression-based prediction of pretrial misconduct risk, inspired by the Laura and John Arnold Foundation Public Safety Assessment tool (LJAF PSA). The LJAF PSA is used in a number of states and cities to assist bail judges in making pretrial release decisions. LJAF PSA scores are based on nine defendant and case observables: the defendant's age; an indicator for a violent crime charge; an indicator for a pending charge at the time of offense; indicators for a prior misdemeanor, felony, or violent crime conviction; the number of previous failures to appear over the last two years; an indicator for a failure to appear more than two years ago; and an indicator for prior incarceration.

We construct ordinary least squares risk predictions $\hat{Y}_i$ by regressing, in the sample of released defendants, an indicator for pretrial misconduct (either a failure to appear or being rearrested for a new crime) on a set of observed characteristics based on the LJAF PSA inputs. For most characteristics, we are able to match the inputs exactly. We do not observe whether a defendant has a pending charge, however, so we exclude this input. We also do not observe prior incarceration, so we instead use an indicator for prior arrest. As with the main algorithmic predictions, we use these $\hat{Y}_i$ and a range of risk thresholds τ to form release recommendations $T_i = \mathbf{1}[\hat{Y}_i < \tau]$ for all defendants in the sample.

Panel A of Appendix Figure A.1 shows our extrapolation-based estimation of the key race-specific second moments ρ_w and ρ_b when using the regression-based prediction of pretrial misconduct. Panel B of Appendix Figure A.1 plots the corresponding range of estimated measures of algorithmic discrimination for the regression-based prediction of pretrial misconduct. As with our baseline gradient boosted decision tree algorithm, the regression-based algorithmic recommendations yield similar second-moment estimates for white and Black defendants (of around 0.2) at the average release rate in NYC (73 percent). These estimates and the common mean risk estimates yield a 6.7 percentage point disparity in the recommended release rates of white and Black defendants with the same potential for pretrial misconduct. This discriminatory disparity is a large share (73.6 percent) of the unadjusted release rate disparity in algorithmic recommendations (9.1 percentage points), and a similar share as with our baseline gradient boosted decision tree algorithm. We again find algorithmic discrimination over a wide range of potential release rates, with the estimated Δ statistically distinguishable from zero at all but the highest release rates.

IV. Alternative Discrimination Measures

This appendix shows how our estimates of race-specific parameters $\{\mu_w, \mu_b, \rho_w, \rho_b\}$ can be used to construct alternative measures of algorithmic discrimination in the NYC pretrial setting. We first estimate race-specific covariances of misconduct potential Y_i^* and algorithmic release recommendations T_i. We then estimate racial disparities in true- and false-negative rates, δ_r^T and δ_r^F, which enter our average discrimination measure Δ. A racial equality in false-negative rates can be seen as satisfying what is known in the computer science literature as "equality of opportunity" (Hardt, Price and Srebro, 2016), meaning that "qualified" white and Black defendants without pretrial misconduct potential are released at the same rate. We also show that our estimates can be used to detect departures from what is known in the computer science literature as "sufficiency" (Zafar et al., 2017), and what Kleinberg, Mullainathan and Raghavan (2017) refer to as "calibration," meaning the racial equality of positive and negative predictive values.

Appendix Figure A.2 first plots our estimates of race-specific covariances of misconduct potential and algorithmic release recommendations across a range of release rates. These estimates are obtained by $Cov(Y_i^*, T_i \mid R_i) = \rho_{R_i} - \mu_{R_i} \times E[T_i \mid R_i]$. We tend to find a stronger (more negative) covariance for Black defendants than white defendants. This results from the fact that we estimate a higher mean risk μ_{R_i} for Black defendants than for white defendants, while we tend to obtain similar estimates of the second moment ρ_{R_i} and somewhat higher release rates $E[T_i \mid R_i]$ for white defendants.

Appendix Figure A.3 next plots our estimates of racial disparities in true- and false-negative rates. These estimates are obtained by the formulas for δ_r^T and δ_r^F in the main text. We find a disparity in false-negative rates that is large and roughly constant across different release rates, where white defendants with misconduct potential tend to be released at a higher rate than Black defendants with misconduct potential. In contrast, the racial disparity in true-negative rates (i.e., the release rate differential among defendants without misconduct potential) is only statistically significantly different from zero at low release rates. These results suggest that a measure of "inequality of opportunity" (that $\delta_w^T - \delta_b^T \neq 0$) could fail to detect overall racial discrimination (that $\Delta \neq 0$) in this setting.

Finally, Appendix Figure A.4 plots estimates of algorithmic "insufficiency." Sufficiency is defined by the equality of negative and positive predictive values across race, or equivalently the independence of Y_i^* and R_i given T_i. Paralleling our main discrimination measure Δ, we define insufficiency as:

$$\Sigma = E[E[Y_i^* \mid R_i = w, T_i] - E[Y_i^* \mid R_i = b, T_i]] \tag{1}$$

Here, the inner difference compares the misconduct rate for white and Black defendants, holding fixed the algorithmic recommendation T_i. The outer expectation averages this comparison over the recommendation distribution. A finding of $\Sigma < 0$ indicates that white individuals tend to be less risky than Black defendants with identical algorithmic recommendations. As with Δ, this measure can be decomposed, as:

$$\Sigma = (\sigma_w^R - \sigma_b^R)E[T_i] + (\sigma_w^D - \sigma_b^D)(1 - E[T_i]) \tag{2}$$

where $\sigma_r^R = E[Y_i^* \mid R_i = r, T_i = 1]$ and $\sigma_r^D = E[Y_i^* \mid R_i = r, T_i = 0]$ are the misconduct rate among released and detained individuals, respectively, of race r.

To estimate Σ, we use this decomposition and the fact that:

$$\sigma_r^R = 1 - \frac{E[Y_i^* T_i \mid R_i = r]}{E[T_i \mid R_i = r]} = 1 - \frac{\rho_r}{E[T_i \mid R_i = r]} \tag{3}$$

$$\sigma_r^D = 1 - \frac{E[Y_i^* (1 - T_i) \mid R_i = r]}{E[(1 - T_i) \mid R_i = r]} = 1 - \frac{\mu_r - \rho_r}{1 - E[T_i \mid R_i = r]} \tag{4}$$

Panel A of Appendix Figure A.4 shows a generally negative Σ across a wide range of algorithmic release rates when we use our estimates of first- and second-moments as inputs to these formulas, meaning that white defendants tend to have lower pretrial misconduct rates than Black defendants conditional on the algorithm's release recommendation. Panel B of Appendix Figure A.4 shows that the average misconduct rate disparities in Panel A are driven by large but noisy racial disparities among detained defendants and smaller but more precise racial disparities among released defendants.

V. Quantifying the Bias from Selective Labels

This appendix compares our estimates of algorithmic discrimination Δ with a potentially biased measure estimated in the selected subsample of defendants released before trial. We compare Δ to:

$$\begin{aligned} \Delta^S &= E[E[T_i|R_i = w, Y_i^*, D_i = 1] - E[T_i|R_i = b, Y_i^*, D_i = 1] \mid D_i = 1] \\ &= E[E[T_i|R_i = w, Y_i, D_i = 1] - E[T_i|R_i = b, Y_i, D_i = 1] \mid D_i = 1], \end{aligned} \tag{5}$$

where we use the fact that $Y_i^* = Y_i$ in the released ($D_i = 1$) subsample. Unlike Δ, this Δ^S measure can be computed directly from the observed data. Generally, $\Delta \neq \Delta^S$ when the release decisions that reveal misconduct potential Y_i^* are not as-good-as-random within race. Thus, a comparison of Δ and Δ^S quantifies the potential bias in observational measures of algorithmic discrimination.

Appendix Figure A.5 plots estimates of Δ^S against our main algorithmic discrimination estimates Δ. We find these two estimates are broadly similar and not statistically distinguishable across most of the range in algorithmic release rates. Thus, while in principle the selective labels problem can induce bias in observable measures of algorithmic discrimination, we find by computing Δ in this setting that the scope for such bias is small.

References

Arnold, David, Will Dobbie, and Peter Hull. 2020. "Measuring Racial Discrimination in Bail Decisions." *NBER Working Paper No. 26999.*

Hardt, Moritz, Eric Price, and Nathan Srebro. 2016. "Equality of Opportunity in Supervised Learning." *Proceedings of the 30th Conference on Neural Information Processing Systems*, 3323–3331.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. "Human Decisions and Machine Predictions." *Quarterly Journal of Economics*, 133(1): 237–293.

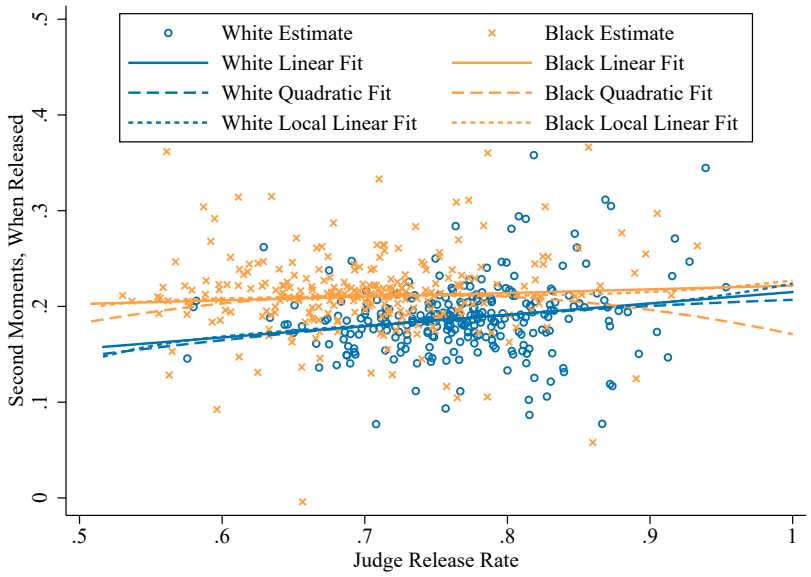
Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan. 2017. "Inherent Trade-Offs in Algorithmic Fairness." *Proceedings of Innovations in Theoretical Computer Science*, 43:1–43:23.

Zafar, Muhammad Bilal, Isabel Valera, Manuel Gomez Rodriguez, and Krishna Gummadi. 2017. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment." *Proceedings of the 26th International Conference on World Wide Web.*

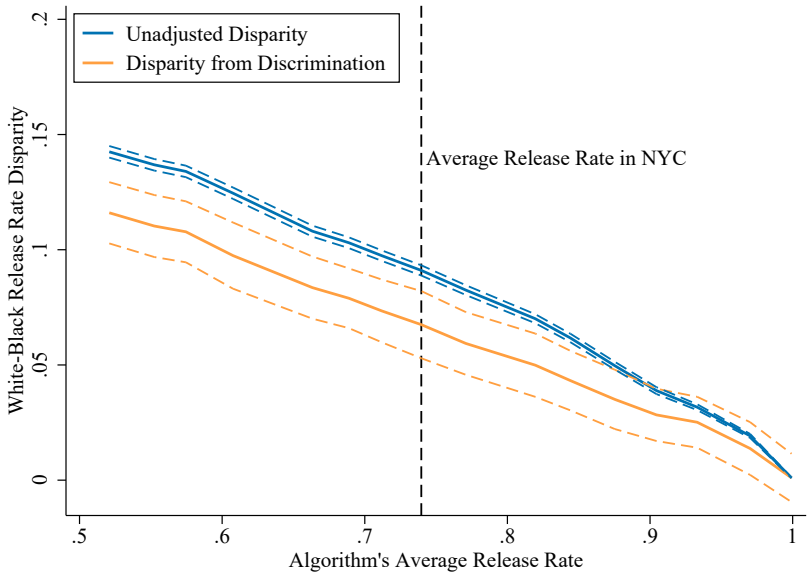
A Appendix Figures and Tables

Figure A.1: Estimating Discrimination for Regression-Based Risk Predictions

Panel A. Extrapolated Second Moments

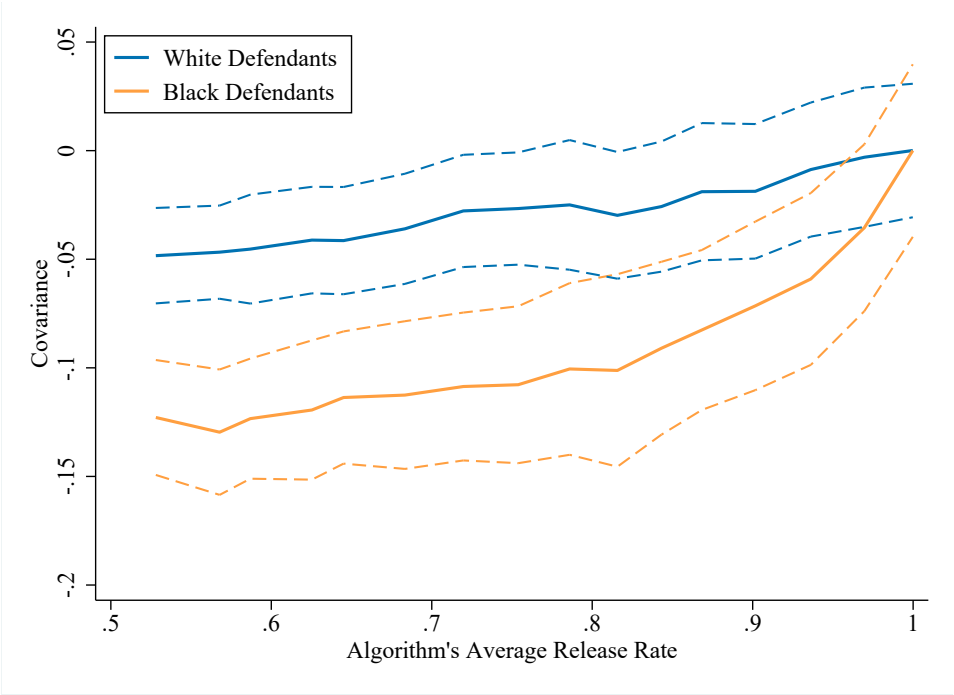


Panel B. Algorithmic Discrimination



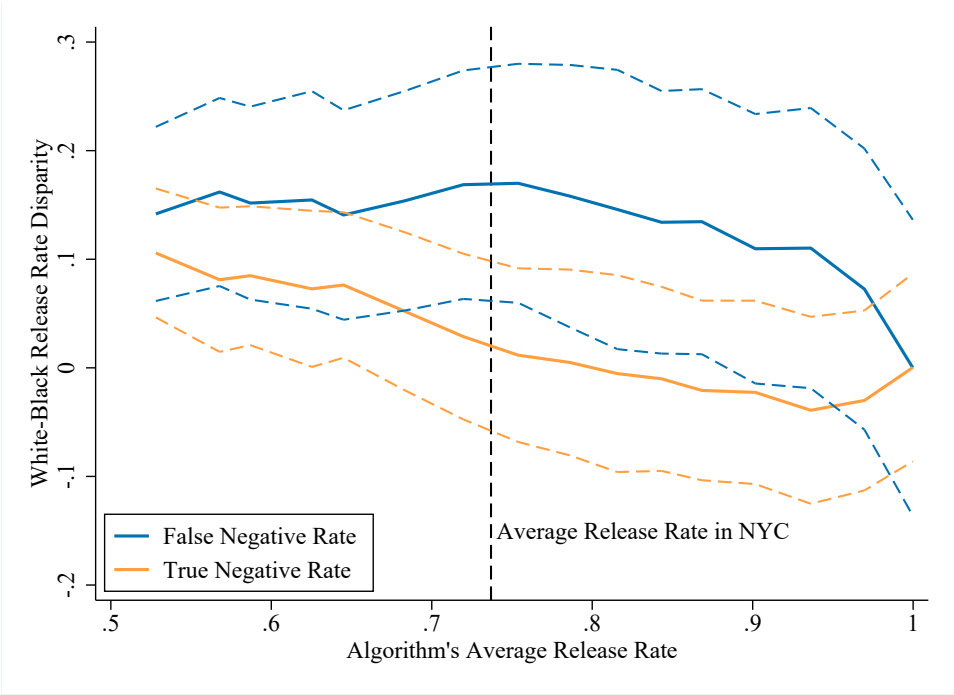
Notes. Panel A plots race-specific release rates for the 268 judges in our sample against race-specific second moments of pretrial misconduct and regression-based risk predictions among released defendants. All estimates adjust for court-by-time fixed effects (the level at which judges are as-good-as-randomly assigned), and recommendations are calibrated to the average release rate in NYC. The two curves plot the fitted values of race-specific local linear regressions that inverse-weight by the variance of the estimated released second moments and use a Gaussian kernel with a race-specific rule-of-thumb bandwidth. Panel B plots estimates of algorithmic discrimination and unconditional racial disparities in algorithmic recommendations for different average release rates. The discrimination estimates use extrapolated first- and second-moment estimates, as in Panel A and as described in the text. Dashed lines indicate pointwise 95% confidence intervals obtained from the bootstrapping procedure described above.

Figure A.2: Covariance of Pretrial Misconduct and Algorithmic Release Recommendations



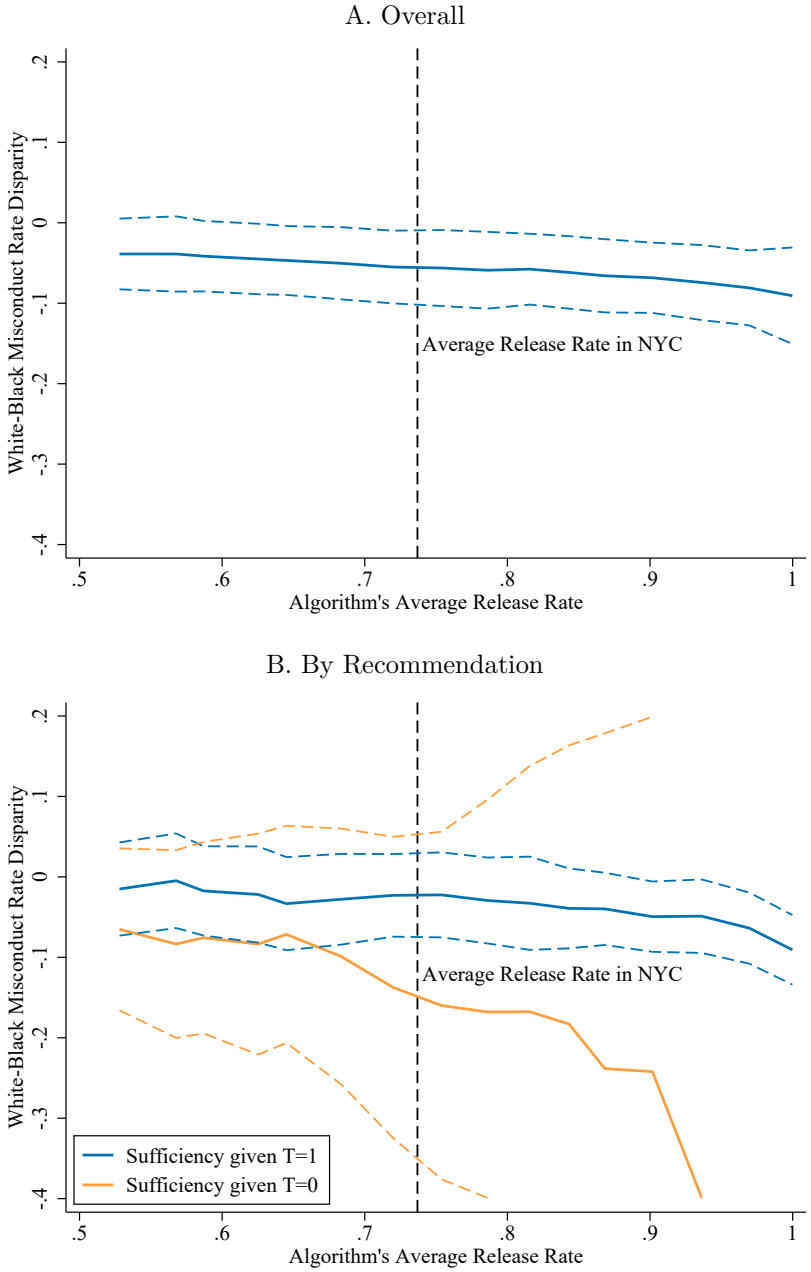
Notes. This figure plots the range of race-specific covariance between pretrial misconduct potential and algorithmic release recommendations for different average release rates. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Covariances are computed by using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described above.

Figure A.3: Decomposition of Algorithmic Discrmination



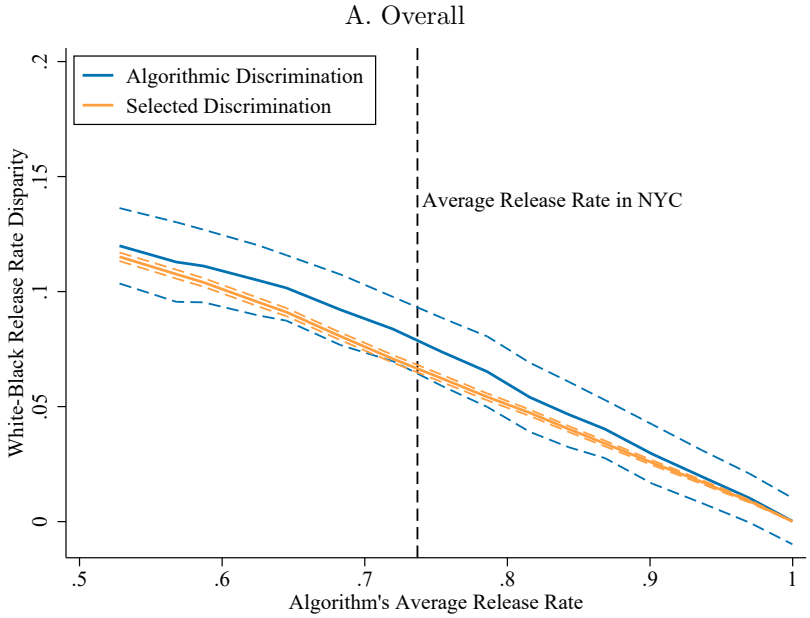
Notes. This figure plots the range of racial disparities in true and false negative rates, for different average release rates, which make up the disparities due to racial discrimination. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described above.

Figure A.4: Measuring Discrimination by Sufficiency of Algorithmic Recommendations



Notes. This figure plots the range of racial disparities in average positive and negative predictive values, or the insufficiency of algorithmic release rate recommendations, for different average release rates. Panel A shows the overall racial disparity while Panel B shows disparities separately by the algorithm's recommendation. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described above.

Figure A.5: Quantifying the Bias from Selective Labels



Notes. This figure plots our main discrimination estimates against a potentially biased measure of discrimination that is estimated on the subsample of defendants released before trial. Algorithmic recommendations are from our baseline gradient-boosted decision tree model. Disparities from discrimination are computed as described in the text, using local linear estimates of the race-specific first and second moments. Dashed lines indicate pointwise 95 percent confidence intervals, computed by the bootstrapping procedure described above.

Table A.1: Parameter and Discrimination Estimates

	Linear Extrapolation	Quadratic Extrapolation	Local Linear Extrapolation
Panel A: Mean Misconduct Risk	(1)	(2)	(3)
White Defendants	0.338	0.319	0.346
	(0.007)	(0.022)	(0.016)
Black Defendants	0.400	0.394	0.436
	(0.006)	(0.020)	(0.016)
Panel B: Misconduct/Recommendation Second Moment			
White Defendants	0.207	0.215	0.226
	(0.006)	(0.019)	(0.012)
Black Defendants	0.202	0.160	0.213
	(0.006)	(0.016)	(0.017)
Panel C: Algorithmic Discrimination			
Release Rate Disparity	0.086	0.080	0.079
	(0.003)	(0.011)	(0.007)

Notes. Panels A and B of this table summarize estimates of race-specific mean risk and second moments of misconduct potential and the algorithmic release recommendation from different extrapolations of quasi-experimental variation. Panel C reports corresponding estimates of algorithmic discrimination, as defined in the text. Column 1 uses a linear extrapolation of the variation, while column 2 uses a quadratic extrapolation and column 3 uses a local linear extrapolation with a Gaussian kernel and a rule-of-thumb bandwidth. Robust standard errors are obtained by the bootstrapping procedure described above and appear in parentheses.