Measuring Racial Discrimination in Bail Decisions

David Arnold, Will Dobbie, and Peter Hull

Ljubica Ristovska & Sagar Saxena

November 21, 2022

Goal 1: measure disparate impact

- U.S. anti-discrimination law: illegal to discriminate on the basis of race, sex, color, religion, and national origin
- Disparate treatment
  - Discriminatory policy/practice if motivated by discriminatory purpose
  - E.g., audit studies help identify disparate treatment instances
  - Requires proof of intent
- Disparate impact
  - Discriminatory policy/practice if leads to adverse impacts on a protected class and the decision-maker cannot provide a substantial legitimate justification
  - Holds decision-makers accountable for "direct discrimination" from considering a protected characteristic and "indirect discrimination" from considering characteristics unrelated to a protected class that nevertheless lead to an adverse impact
- Ideal statistical test for disparate impact: compare treatment of different protected groups with identical potential for achieving a given outcome of interest
- No model of decision-making needed; need quasi-random assignment

Goal 2: decompose disparate impact

- Decompose disparate impact into racial bias, incorrect stereotypes (prediction errors), and accurate statistical discrimination
  - Standard approaches use outcome-based tests which can identify taste-based discrimination à la Becker (1957) but not statistical discrimination
- Need a structural model of judge decision making
  - Can use the estimated model to run counterfactuals with policies aimed at reducing disparate impact

Setting: U.S. pretrial system (NYC)

- Bail judges set bail conditions at an arraignment hearing shortly after arrest
  - Bail conditions: release on recognizance, cash bail, supervised release program, deny bail
- Objective: release most defendants while minimizing risk of pretrial misconduct
- Judge has info on current offense, prior criminal record, release recommendation from non-profit based on 6-item checklist
  - In NYC, judges asked to only consider failure to appear (FTA) not risk of new criminal activity
  - Many defendants in NYC do not have bail set due to case dismissal or desk appearance which does not lead to arraignment hearing
- In many jurisdictions, including NYC, the case assignment process generates quasi-random variation in assigned judges conditional on court-by-time FEs
  - Rotation calendar system to assign judges to arraignment shifts in five courthouses
  - Paper verifies conditional random assignment via OLS regressions of residualized, leave-one-out judge leniency (average release rate) measure on defendant and case characteristics

Data

- Nov 1 2008 - Nov 1 2013 arraignments in NYC
- Only felony and misdemeanor cases ’
- Keep cases that make it to arraingment hearing and thus have a judge assigned
- Focus on white vs. Black

Summary statistics

Table 1: Descriptive Statistics

	All Defendants	White Defendants	Black Defendants
Panel A: Pretrial Release	(1)	(2)	(3)
Released Before Trial	0.730	0.767	0.695
Share ROR	0.852	0.852	0.851
Share Money Bail	0.144	0.144	0.145
Share Other Bail Type	0.004	0.004	0.004
Share Remanded	0.000	0.000	0.000
Panel B: Defendant Characteristics			
White	0.478	1.000	0.000
Male	0.821	0.839	0.804
Age at Arrest	31.97	32.06	31.89
Prior Rearrest	0.229	0.204	0.253
Prior FTA	0.103	0.087	0.117
Panel C: Charge Characteristics			
Number of Charges	1.150	1.184	1.118
Felony Charge	0.362	0.355	0.368
Misdemeanor Charge	0.638	0.645	0.632
Any Drug Charge	0.256	0.257	0.256
Any DUI Charge	0.046	0.067	0.027
Any Violent Charge	0.143	0.124	0.160
Any Property Charge	0.136	0.127	0.144
Panel D: Pretrial Misconduct, When Released			
Pretrial Misconduct	0.299	0.266	0.332
Share Rearrest Only	0.499	0.498	0.499
Share FTA Only	0.281	0.296	0.269
Share Rearrest and FTA	0.220	0.205	0.232
Total Cases	595,186	284,598	310,588
Cases with Defendant Released	434,201	218,256	215,945

Notes. This table summarizes the NYC analysis sample. The sample consists of bail hearings that were quasi-randomly assigned judges between November 1, 2008 and November 1, 2013, as described in the text. Information on demographics and criminal outcomes is derived from court records as described in the text. Pretrial release is defined as meeting the bail conditions set by the first assigned bail judge. ROR (released on recognizance) is defined as being released without any conditions. FTA (failure to appear) is defined as failing to appear at a mandated court date.

Setup (in potential outcomes)

- Each individual i in a population has a latent binary state $Y_i^* \in \{0, 1\}$, not observed by decision-maker nor econometrician
  - Paper has extensions for multivalued and continuous outcomes
  - Remain agnostic as to why differences in Y_i^* exist – can be affected by discrimination at other points of the system
  - Here, Y_i^* denotes pretrial misconduct
- Race $R_i \in \{w, b\}$
- Decision makers j make decisions $D_{ij} \in \{0, 1\}$ for each individual
  - Important: D_{ij} is the *potential* decision of decision-maker j
  - Here, decision-maker is a judge
- Objective of decision maker: "align" D_{ij} with Y_i^*
  - In case of bail, objective is to set $D_{ij} = 1$ (release) if $Y_i^* = 0$ (will not commit pretrial misconduct)

Disparate impact: formal definition

- Recall: disparate impact = differences in treatment between protected classes conditional on misconduct potential
- "Correct classification": difference in release rate between white and Black individuals not at risk of comitting misconduct for judge j:

$$\Delta_{j0} = \mathbb{E}[D_{ij} \mid R_i = w, Y_i^* = 0] - \mathbb{E}[D_{ij} \mid R_i = b, Y_i^* = 0]$$

- "Incorrect classification": difference in release rate between white and Black individuals at risk of comitting misconduct for judge j:

$$\Delta_{j1} = \mathbb{E}[D_{ij} \mid R_i = w, Y_i^* = 1] - \mathbb{E}[D_{ij} \mid R_i = b, Y_i^* = 1]$$

Disparate impact: formal definition

- Average disparate impact for judge j is a weighted average of correct and incorrect classifications, weights are average misconduct risk in population $\bar{\mu} = \mathbb{E}[Y_i^*]$:

$$\Delta_j = \Delta_{j0}(1 - \bar{\mu}) + \Delta_{j1}\bar{\mu}$$

  - Also interpretable as the expected level of discrimination for judge j when population risk of misconduct is unknown
  - System-wide level of discrimination can be recovered as a case-weighted average of Δ_j's
  - Does not capture discriminatory intent, just discriminatory effects of decisions

- If $\Delta_j > 0$, judge j discriminates against Black defendants
- If $\Delta_j < 0$, judge j discriminates against white defendants
- If $\Delta_j = 0$, judge j does not discriminate
  - Does not discriminate conditional on other potentially discriminatory policies, systems, conditions

Observed data

- Suppose $Z_{ij} = 1$ if defendant i is assigned to judge j
- $D_i = \sum_j Z_{ij} D_{ij}$ is defendant i's *observed* release status
- $Y_i = D_i Y_i^*$ is *observed* outcome (whether pretrial misconduct occurred)
  - Only observe Y_i^* when $D_i = 1$
- Econometrician observes $(R_i, Z_{i1}, \ldots, Z_{iJ}, D_i, Y_i)$
  - Denote whether defendant is white as $W_i = \mathbb{1}[R_i = w]$
- Assume complete random assignment of judges to defendants: $Z_{ij} \perp (R_i, D_{ij}, Y_i^*)$

Observational disparity analysis: benchmarking regression

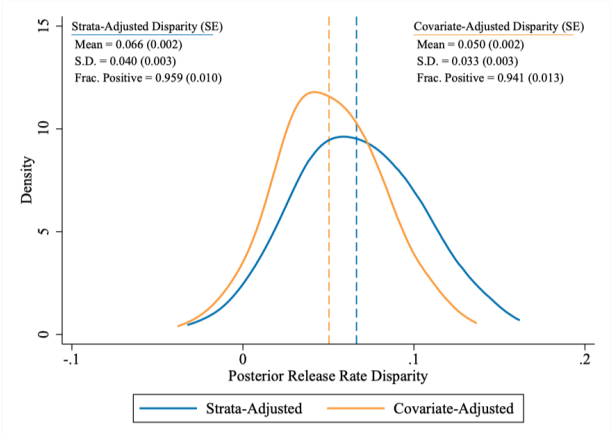
- Regress release decision on judge x race FEs and judge FE:

$$D_i = \sum_j \alpha_j W_i Z_{ij} + \sum_j \phi_j Z_{ij} + \varepsilon_i$$

- $\alpha_j = \mathbb{E}[D_i \mid R_i = w, Z_{ij} = 1] - \mathbb{E}[D_i \mid R_i = b, Z_{ij} = 1]$
  - α_j captures differences in release rate for white relative to Black defendants for judge j

Benchmarking regression in the data

Figure 1: Observational Release Rate Disparities



Notes. This figure plots the posterior distribution of observational release rate disparities for the 268 judges in our sample. We estimate disparities by OLS regressions of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects. The strata-adjusted disparity regression controls only for the main judge fixed effects and court-by-time fixed effects. The covariate-adjusted disparity regression adds the baseline controls from Table 2. The distribution of judge disparities, and fractions of positive disparities, are computed from these estimates as posterior average effects; see Appendix B.3 for details. Means and standard deviations refer to the estimated prior distribution.

Omitted variable bias

- Because $Z_{ij} \perp (R_i, D_{ij}, Y_i^*)$, can re-write α_j as a weighted average of potential outcomes for white vs. black individuals:

$$\alpha_j = (\delta_{jw0}(1-\mu_w) + \delta_{jw1}\mu_w) - (\delta_{jb0}(1-\mu_b) + \delta_{jb1}\mu_b)$$

  - $\delta_{jry} = \mathbb{E}[D_{ij} \mid R_i = r, Y_i^* = y]$ and $\mu_r = \mathbb{E}[Y_i^* \mid R_i = r]$
  - Weights are averages of race-specific misconduct risk

- Disparate impact Δ_j can be written as a slightly different weighted average of potential outcomes for white vs. black individuals

$$\Delta_j = (\delta_{jw0}(1-\bar{\mu}) + \delta_{jw1}\bar{\mu}) - (\delta_{jb0}(1-\bar{\mu}) + \delta_{jb1}\bar{\mu})$$

  - $\bar{\mu} = \mathbb{E}[Y_i^*] = p_w\mu_w + p_b\mu_b$
  - Weights are average population misconduct risk

Omitted variable bias

- Define omitted variable bias (OVB) in benchmarking regression as $\xi_j = \alpha_j - \Delta_j$
  - Upward bias in α_j if $\xi_j > 0$ and downward bias if $\xi_j < 0$
- $\xi_j = [(\delta_{jw0} - \delta_{jw1})p_b + (\delta_{jb0} - \delta_{jb1})p_w] \times (\mu_b - \mu_w)$
- Unlikely in practice, but no OVB ($\xi_j = 0$) if either of the following hold:
  - $\delta_{jr0} = \delta_{jr1}$ for each r: judge decisions are uncorrelated with misconduct potential
  - $\mu_b = \mu_w$: misconduct potential does not differ by race

Recovering Δ_j using quasi-experimental methods

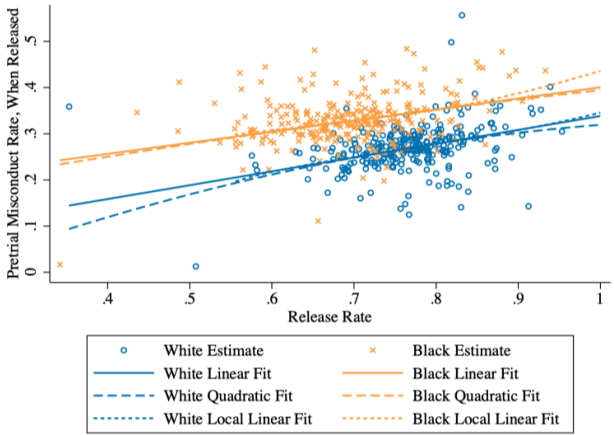
- Important: (a) does not require a model of decision-making and (b) requires quasi-random assignment of decision-makers to individuals
- Key idea: when have quasi-random assignment of decision-makers to individuals, measuring Δ_j reduces to estimating race-specific average misconduct risk
- Given an estimate of μ_r can purge observational estimates α_j's of OVB
- $\Delta_j = \mathbb{E}[\Omega_i D_i \mid R_i = w, Z_{ij} = 1] - \mathbb{E}[\Omega_i D_i \mid R_i = b, Z_{ij} = 1]$
  - $\Omega_i = (1 - Y_i)\frac{1-\bar{\mu}}{1-\mu_{R_i}} + Y_i \frac{\bar{\mu}}{\mu_{R_i}}$

Intuition for estimating μ_r

- Suppose you have a supremely lenient judge j^* who releases everyone, regardless of misconduct risk or race
- If this judge is randomly assigned, then their observed race-specific release rates approximate race-specific average misconduct risk in the full population (μ_j)
- In practice, rarely observe such lenient decision makers, and thus use model-based or statistical extrapolations of release and misconduct rate across randomly assigned decision-makers, evaluated at leniency close to 1

Extrapolation estimates

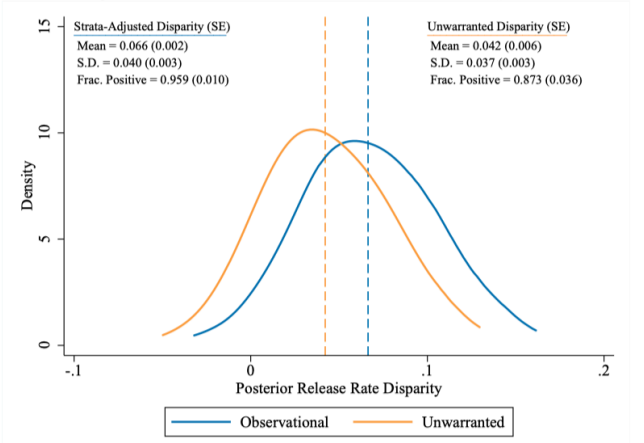
Figure 2: Judge-Specific Release Rates and Conditional Misconduct Rates



Notes. This figure plots race-specific release rates for the 268 judges in our sample against rates of pretrial misconduct among the set of released defendants. All estimates adjust for court-by-time fixed effects. The figure also plots race-specific linear, quadratic, and local linear curves of best fit, obtained from judge-level regressions that inverse-weight by the variance of the estimated misconduct rate among released defendants. The local linear regressions use a Gaussian kernel with a race-specific rule-of-thumb bandwidth.

Correcting for OVB

Figure 3: Observational and Unwarranted Release Rate Disparities



Notes. This figure plots the posterior distribution of observational and unwarranted release rate disparities for the 268 judges in our sample. Strata-adjusted disparities are estimated by the coefficients of an OLS regression of an indicator for pretrial release on white×judge fixed effects, controlling for judge main effects and court-by-time fixed effects. Unwarranted disparities are estimated as described in Section 5, using the local linear extrapolations from Figure 2 to estimate the mean risk of each race. The distribution of judge disparities, and fractions of positive disparities, are computed from these estimates as posterior average effects; see Appendix B.3 for details. Means and standard deviations refer to the estimated prior distribution.

Decomposing Disparate Impact

A model of judge decision making

- Judge j releases defendant i if the benefit of doing so exceeds the cost

$$D_{ij} = \mathbb{1}\left[\pi_{jR_i} \geq p_j(\nu_{ij}, R_i)\right]$$

  - π_{jR_i} is some subjective benefit of releasing the defendant
  - $p_j(\nu_{ij}, R_i) = \Pr(Y_i^* = 1 | \nu_{ij}, R_i)$ where ν_{ij} is a noisy signal of pretrial misconduct potential
  - $\nu_{ij} = Y_i^* + \eta_{ij}, \eta_{ij} \sim N(0, \sigma_{jr}^2)$

Outcome based tests

- Standard recent approaches such as Arnold, Dobbie, and Young (2018) test for discrimination using marginal outcome based tests derived from Becker (1957)
- Under this model, a judge is deemed to have a taste for discrimination if the perceived benefit of releasing white defendants is greater than the perceived benefit of releasing Black defendants i.e. $\pi_{jw} > \pi_{jb}$
- If the judge is not biased i.e. $\pi_{jw} = \pi_{jb}$, we expect the "cost" of marginal defendants from both racial groups to be the same
  - Can test whether these costs (or observed pretrial misconduct) is the same for the marginal defendants
  - Higher pretrial misconduct for marginal white defendants is evidence of taste-based discrimination

Decomposing disparate impact

- It is possible for disparate impact to be nonzero even if outcome-based tests show no evidence of discrimination
- This is because outcome based tests compare defendants with the same perceived risk of pretrial misconduct but the perceived risk for a given defendant might be affected by racial differences in average risk or signal variance
- The authors introduce a hierarchical MTE model to estimate mean risk parameters μ_r and the means and variances of parameters associated with the benefit and signal quality used in their model of judge decision making
- Estimates suggest that both bias and statistical discrimination drive disparate impact

Counterfactuals

- In counterfactuals, force (some or all) judges to adjust their race-specific leniencies to the point where their racial disparities are eliminated
  - Policy instrument: race-specific release rate quotas
- Find that targeting the most discriminatory NYC judges can reduce the average level of discrimination by 36%, while targeting all can essentially eliminate discrimination

Discussion: applications to healthcare

- Has potential for applications in healthcare
- Decision-makers in healthcare: hospitals, insurers, physicians, other healthcare providers
- Lots of binary decisions to be made: test, treat, admit, refer, transplant
- Quasi-random assignment also important
  - Quasi-random assignment of emergency cases to hospitals via ambulances
  - Quasi-random assignment of patients to providers within an ER
- Is the objective function of the decision maker "simple" enough to fit this framework?
  - In other words, can we define a unidimensional risk measure in all of these settings or are tradeoffs more complicated?
- To what extent do race-specific risk distributions vary in healthcare settings?
  - If not, then statistical discrimination is minimal and outcome-based tests can detect objects of interest (taste-based discrimination and prediction error)
  - If statistical discrimination is important, then this framework is better suited for detecting disparate impact