Measuring Racial Discrimination in Bail Decisions

Discussion: Matthew Gentzkow

Racial disparities in the judicial system sit at the very top of the policy agenda

Identifying sources of these disparities is an essential step in addressing them

Similar methodological issues arise in many other contexts (health care, education, housing, etc.)

Existing Approaches

- **Gaps in decision rates** (e.g., Gelman et al. 2007; Abrams et al. 2012)
- **Gaps in outcomes** (e.g., Becker 1993; Knowles et al. 2001)
- **“Judges design”** under strong monotonicity assumption (e.g., Arnold et al. 2018; Marx 2018)

This Paper

- Rich data & clean setting
- Identify discrimination under weak assumptions
- Structural model to further decompose drivers

- Results suggest substantial share of racial gap is due to discrimination, both statistical and preference-based

- **Important, convincing, and methodologically rich**

Outline

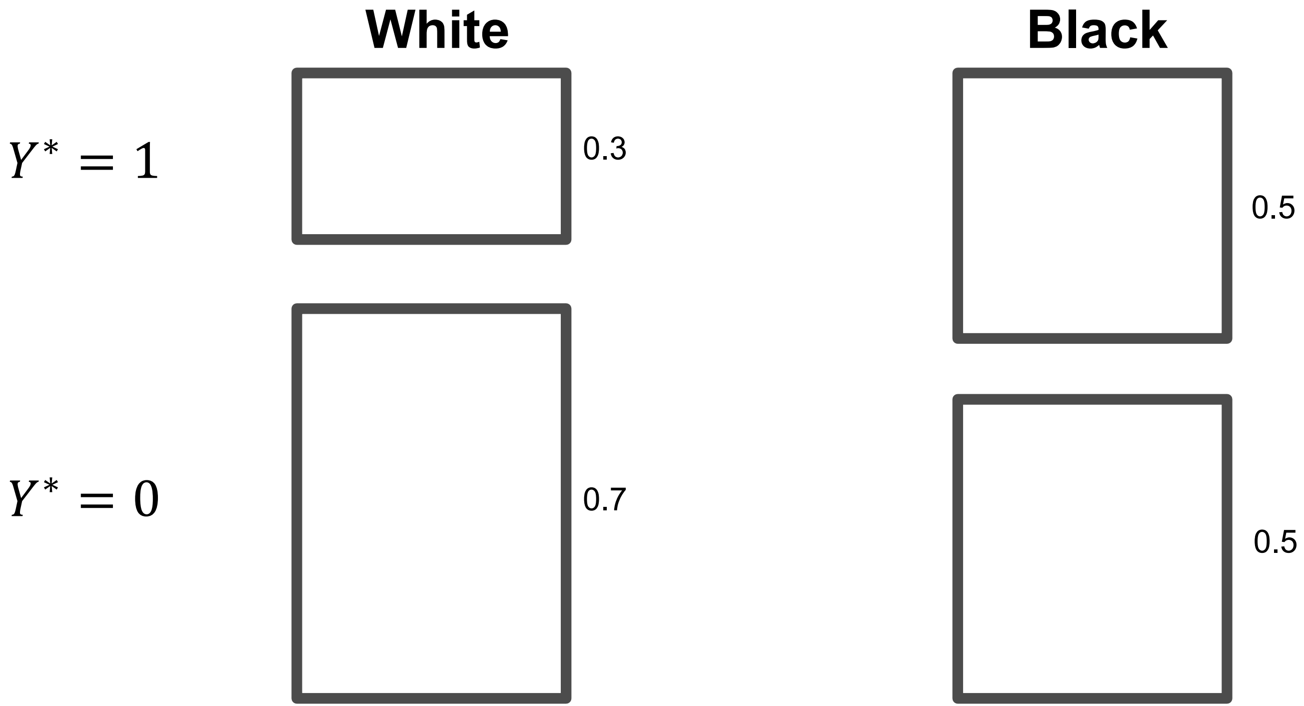
1. Objectives
2. Identification
3. Observables

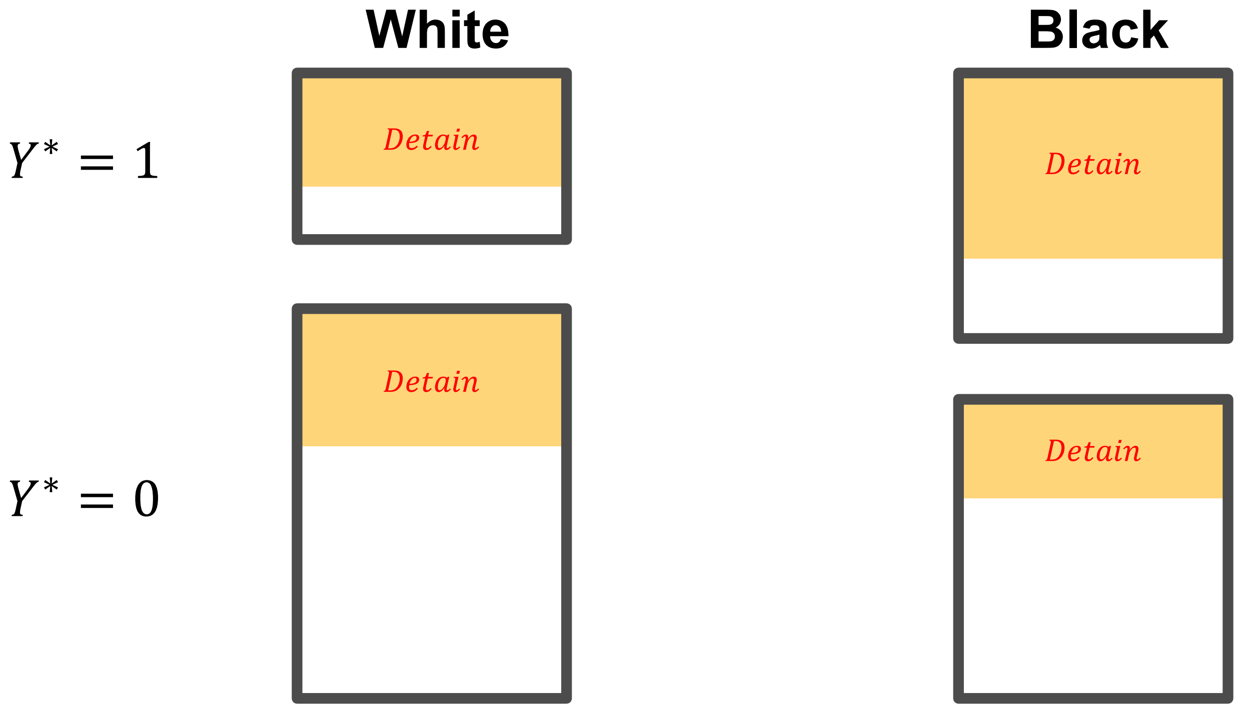
Objectives

Y^*: Potential outcome
D: Decision
R: Race
ν: Agent's information
$p(\nu; R)$: Agent's posterior
X: Observables

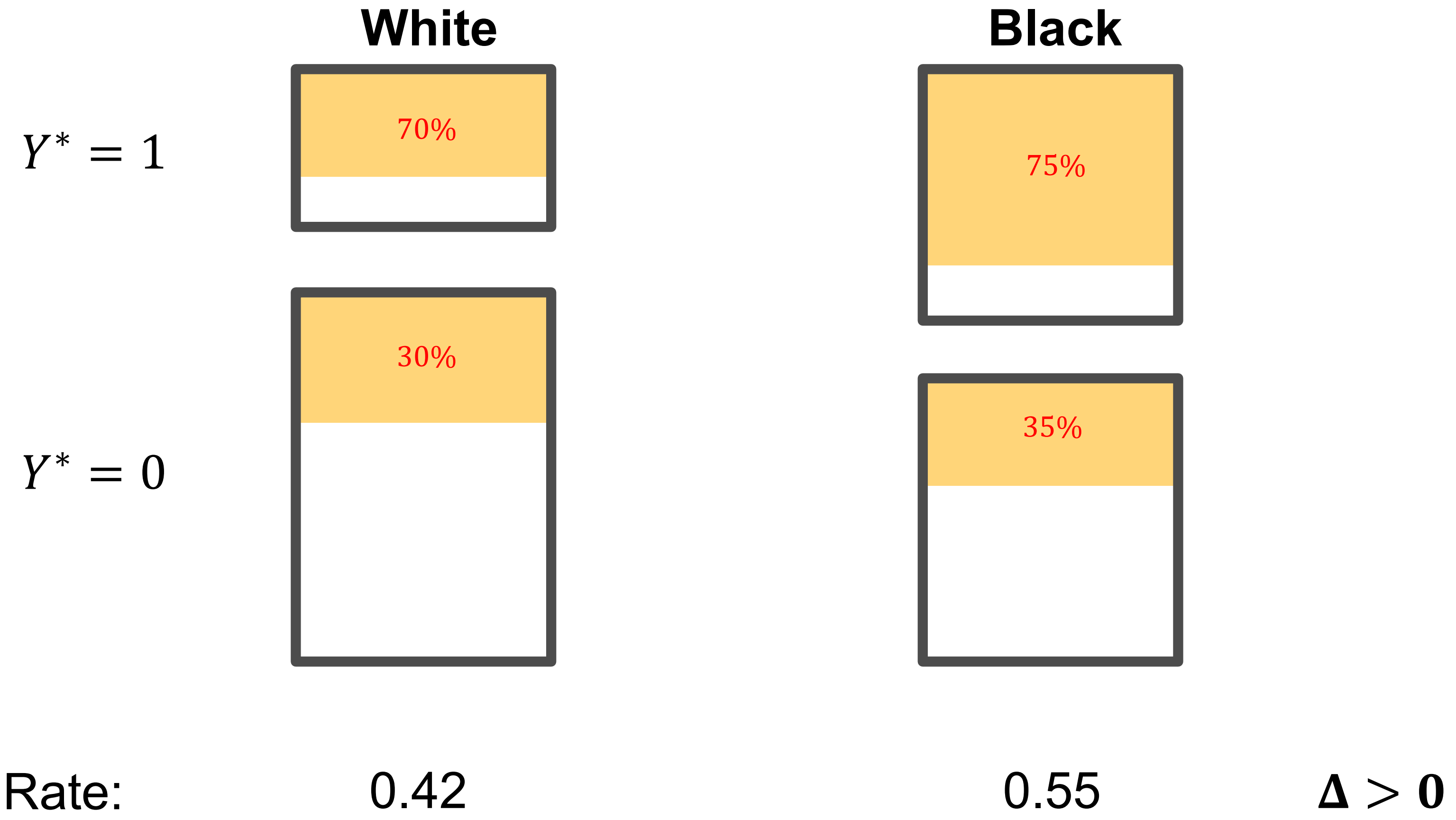
Discrimination

$$\Delta = E\left[E[D|Y^*, R = w] - E[D|Y^*, R = b]\right]$$


White
Black
$Y^* = 1$
$Y^* = 0$
0.3
0.7
0.5
0.5


White
Black
$Y^* = 1$
$Y^* = 0$
Detain
Detain
Detain
Detain

Rate: 0.42 0.5 $\mathbf{\Delta = 0}$


White
Black
$Y^* = 1$
70%
75%
30%
35%
$Y^* = 0$
Rate:
0.42
0.55
$\boldsymbol{\Delta > 0}$

Rate: 0.42 0.45 $\boldsymbol{\Delta < 0}$

Other Objectives

Discrimination: $E[D|Y^*, R = w] - E[D|Y^*, R = b]$

Bias: $E[D|\boldsymbol{p}, R = w] - E[D|\boldsymbol{p}, R = b]$

Race blindness: $E[D|\boldsymbol{\nu}, R = w] - E[D|\boldsymbol{\nu}, R = b]$

Disparity: $E[D|\boldsymbol{X}, R = w] - E[D|\boldsymbol{X}, R = b]$

Tradeoffs

- Can't in general be both non-discriminatory and unbiased
  - Unbiased rule generally leads to different $E(D|Y^*)$
  - See, e.g., Kleinberg et al. 2017
  - Note that efficient → unbiased

- Can't in general be both unbiased and race-blind
  - $p(\nu, R)$ generally differs by R for given ν

- Hard to be both non-discriminatory and race-blind
  - Unless ν effectively orthogonal to R

Legal Standards

Equal protection clause requires “discriminatory intent or purpose”

“Estimation of Δ_j is… a **necessary** first step to establish unconstitutional behavior, though it may not be sufficient absent proof of discriminatory intent” (p. 6)

Are we sure?

What about a policy that explicitly denied bail to blacks in crimes where white base rate is higher, so as to produce $\Delta = 0$?

Bottom line:

This paper takes a strong stand in favor of non-discrimination as the right objective

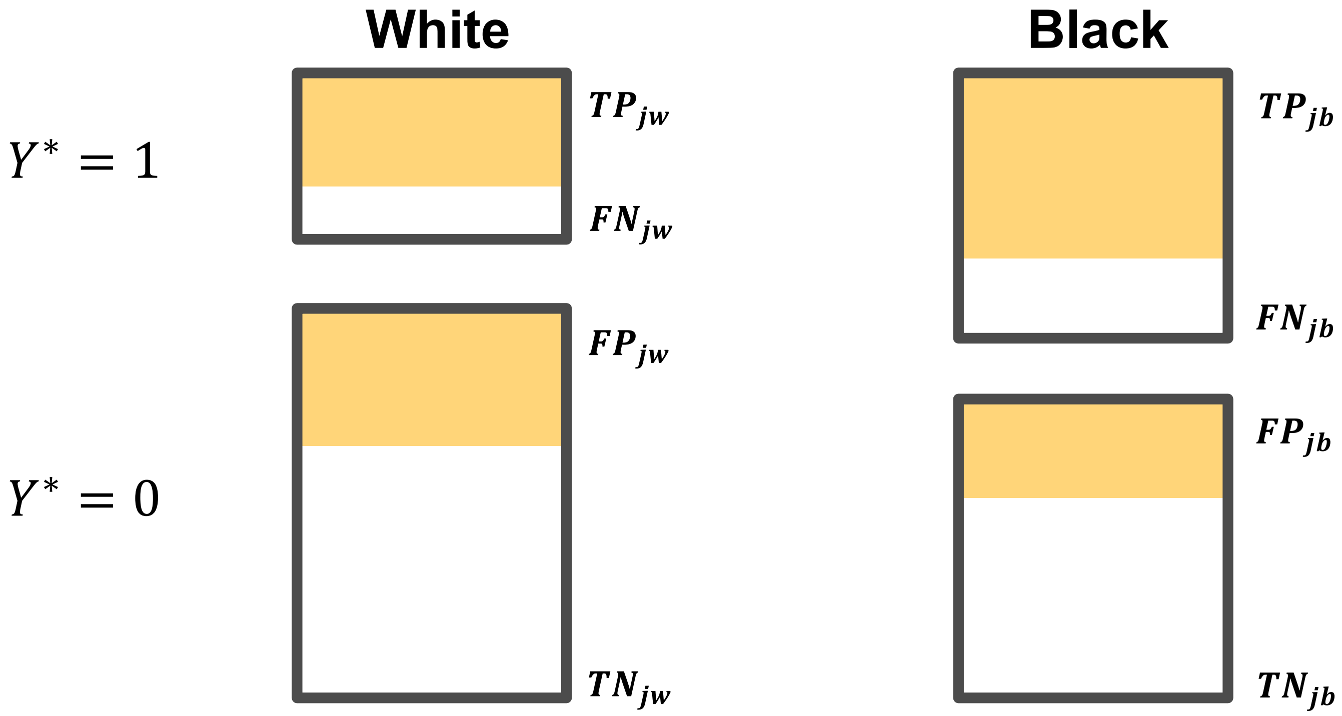
I think the difficult ethical / legal / conceptual issues are more complex than the paper suggests

Identification

Classification Matrix

	$Y^* = 1$	$Y^* = 0$
$D = 1$	TP_{jr}	FP_{jr}
$D = 0$	FN_{jr}	TN_{jr}

$$\Delta_j = f(TP_j, FP_j, FN_j, TN_j)$$


White
Black
$Y^* = 1$
$Y^* = 0$
TP_{jw}
FN_{jw}
FP_{jw}
TN_{jw}
TP_{jb}
FN_{jb}
FP_{jb}
TN_{jb}

One-Sided Selection

	$Y^* = 1$	$Y^* = 0$
$D = 1$	TP_{jr}	FP_{jr}
$D = 0$	FN_{jr}	TN_{jr}

- Only observe true value of Y^* if not detained
- Thus, observed data identify FN_{jr}, TN_{jr}, and $\delta_{jr} = TP_{jr} + FP_{jr}\ \forall j, r$
- 3 moments, 4 unknowns (w/ 1 of each linearly dependent)

- **Need 1 more moment for identification**

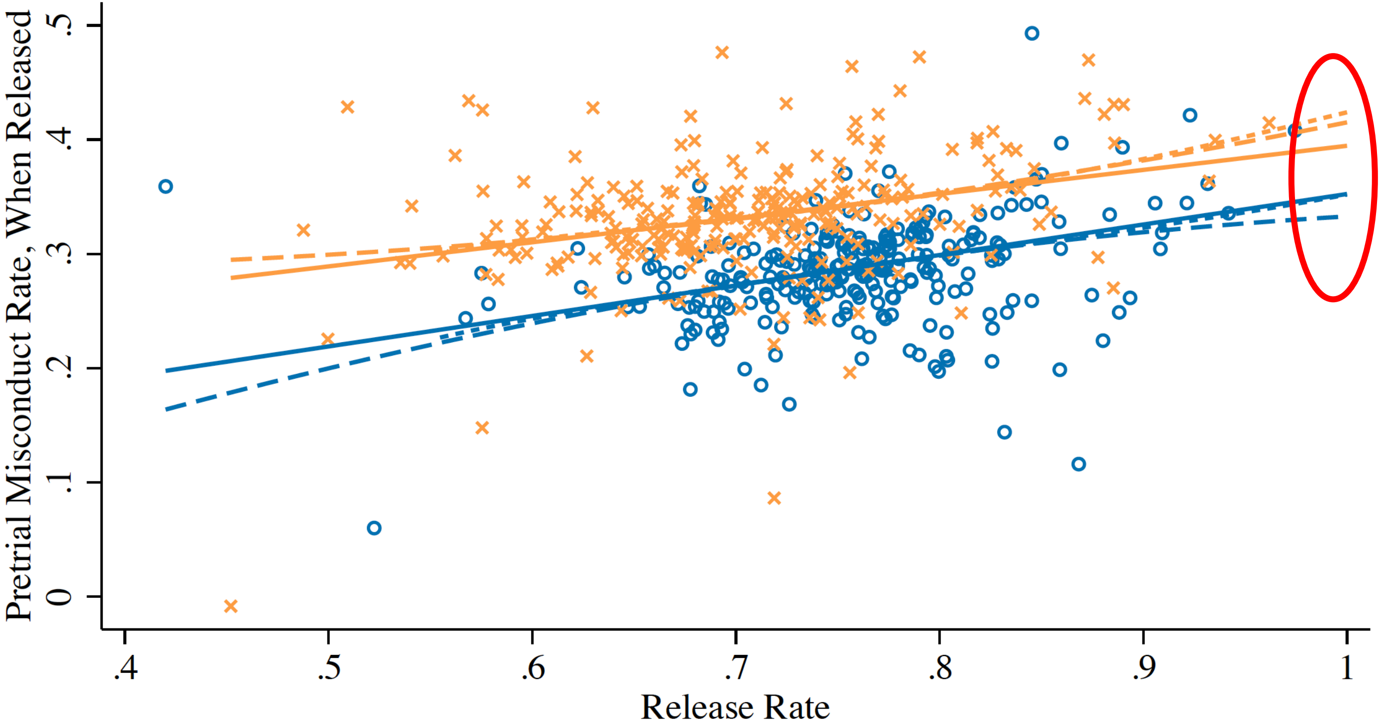
- E.g, sufficient to observe misconduct rates

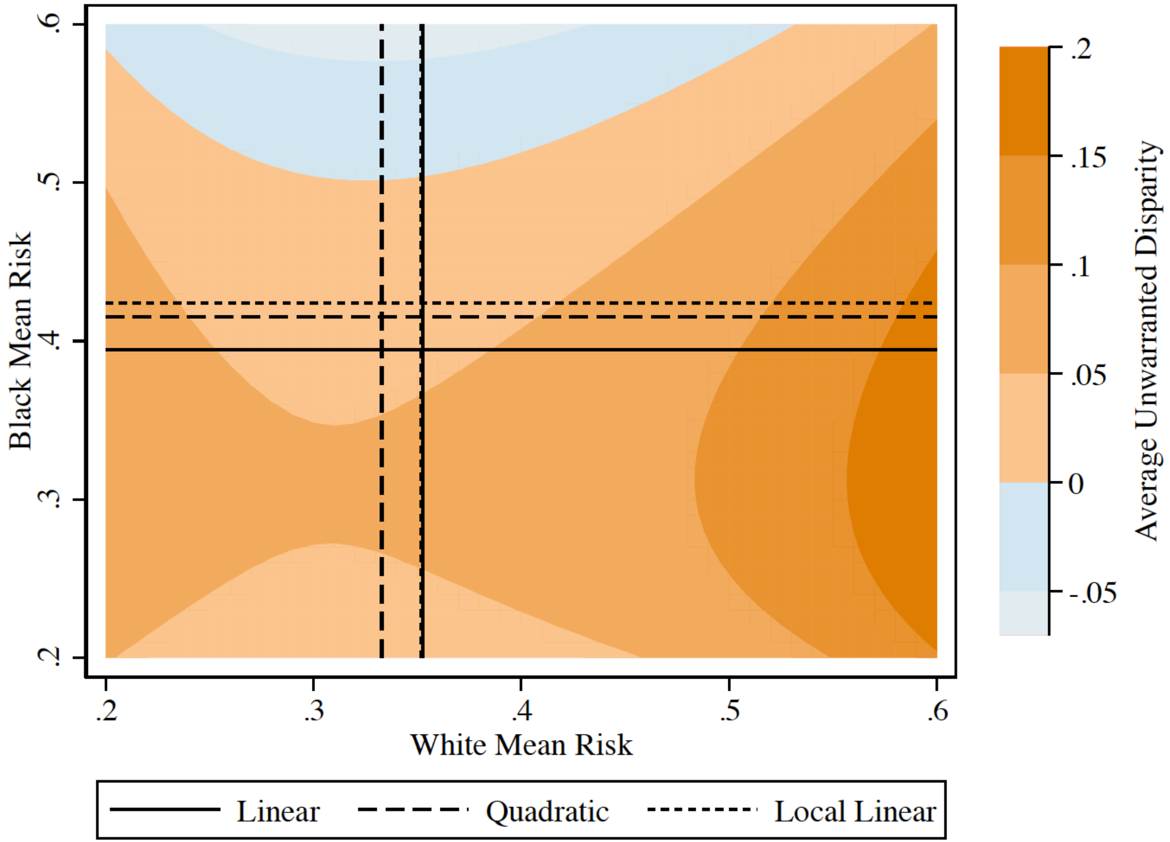
$$\mu_{jr} = TP_{jr} + FN_{jr}$$

- Random assignment: $\mu_{jr} = \mu_r \forall j$

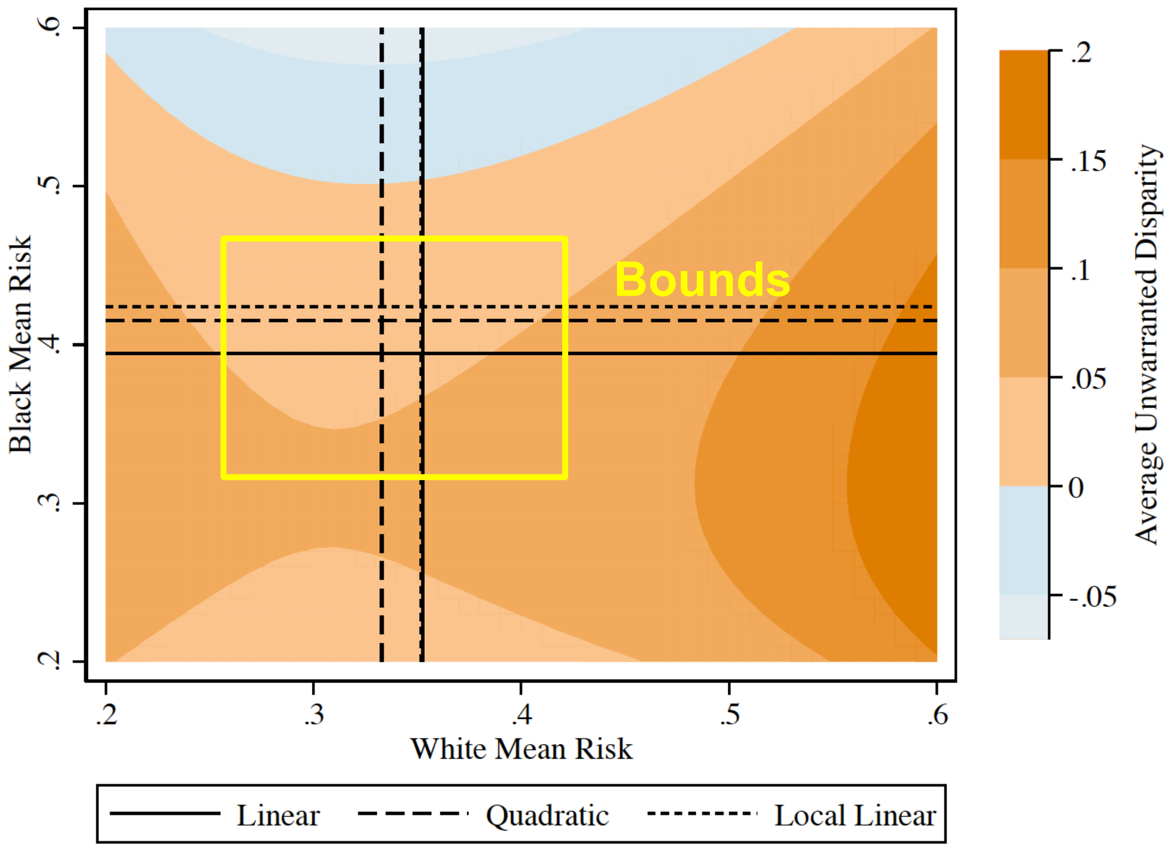
- **All we need is an estimate of $\boldsymbol{\mu_W, \mu_B}$!**

"Identification at infinity"





Black Mean Risk
.2
.3
.4
.5
.6
White Mean Risk
.2
.3
.4
.5
.6
Average Unwarranted Disparity
-.05
0
.05
.1
.15
.2
Linear
Quadratic
Local Linear



Bounds
Black Mean Risk
.2
.3
.4
.5
.6
White Mean Risk
.2
.3
.4
.5
.6
Average Unwarranted Disparity
-.05
0
.05
.1
.15
.2
Linear
Quadratic
Local Linear

Observables

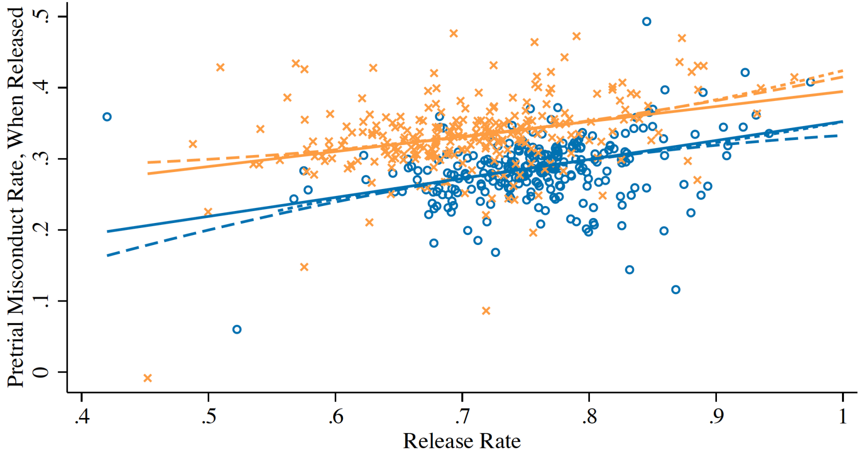
Approach

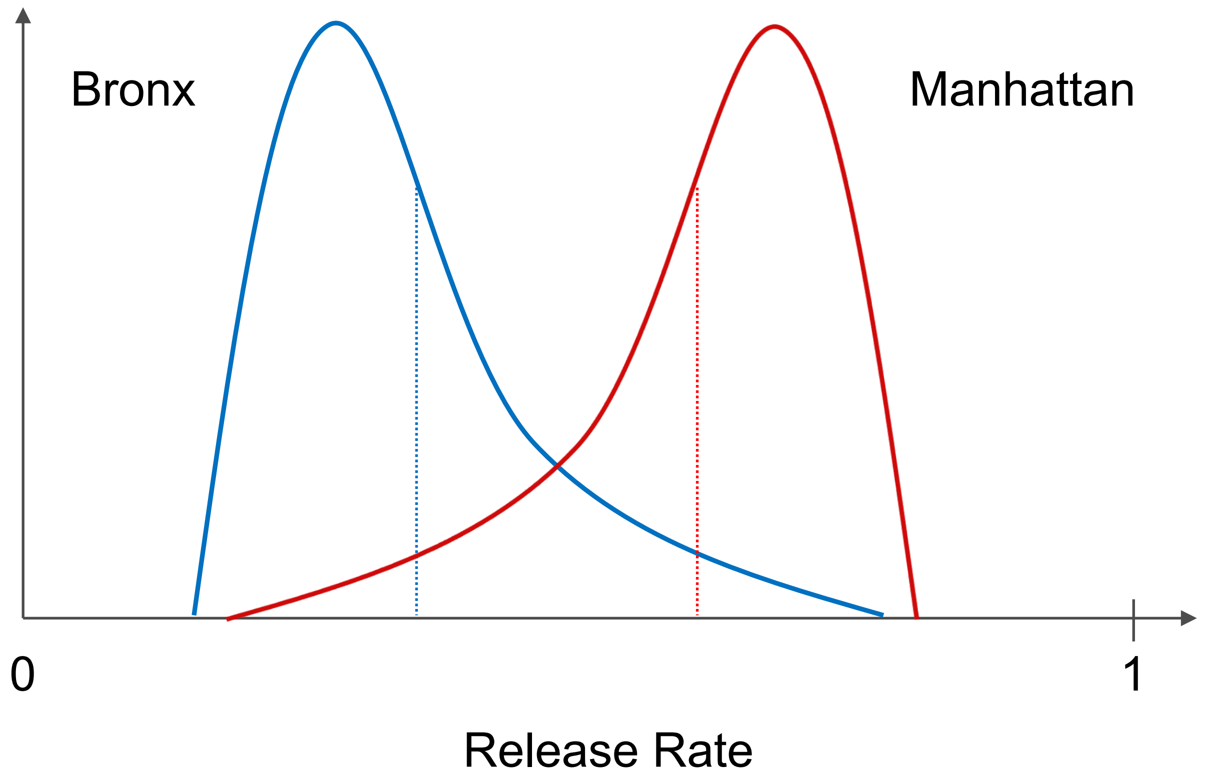
- Partial observed release and misconduct rates for observables X (court $\times$ time, type of crime, etc.)
- Analyze residuals as if they were true release and misconduct rates for a single population, both in reduced-form and structural analysis

- n.b. Chan et al. (2020) take essentially the same approach
- n.b. This is an issue of estimation not identification

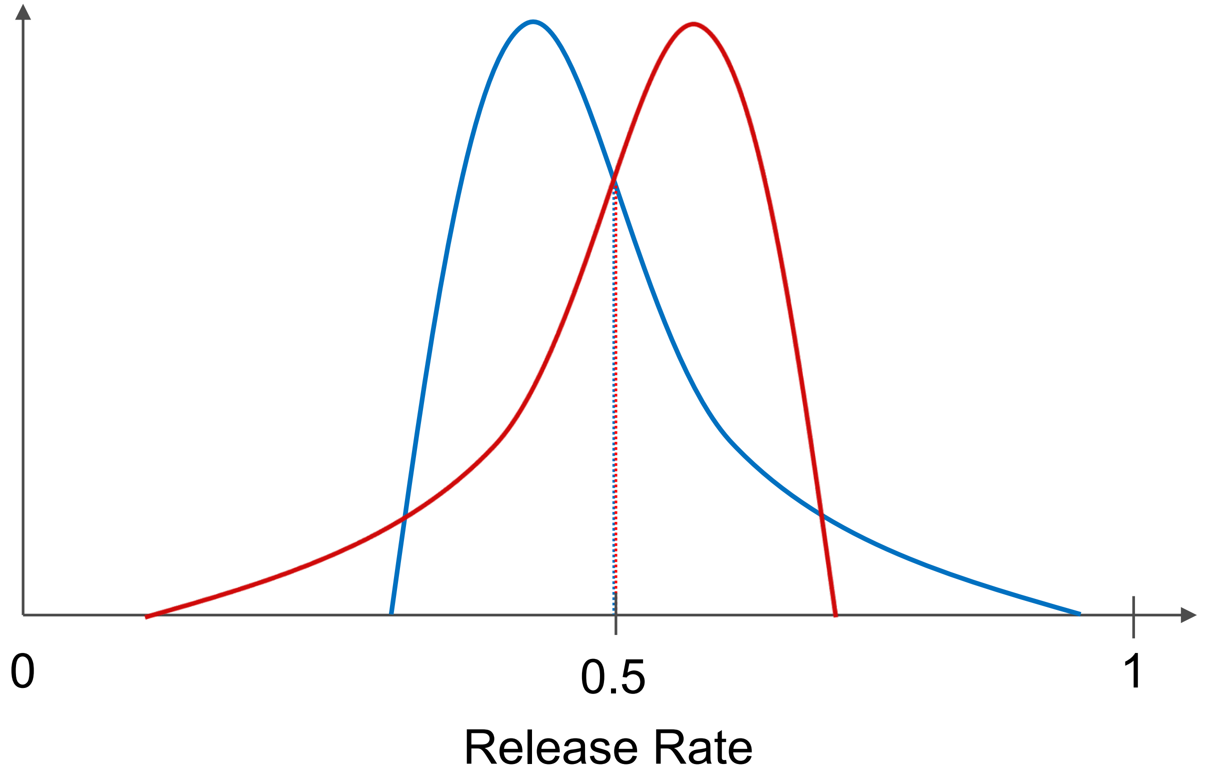
Example of an implication:

Release rates below are not actually what is observed
Might be no judge in the data who releases 90% of defendants!




Bronx
Manhattan
0
1
Release Rate

Residuals



Assumption (FN 18)

$$E\left[Y_i^* \middle| D_{ij} = 1, R_i, X_i\right] = \psi_{jr} + X_i'\gamma$$
$$E\left[D_{ij} \middle| R_i, X_i\right] = \phi_{jr} + X_i'\beta$$

Note that coefficients (γ, β) do not vary by j or i

Rules out judges who treat different X's differently (e.g., we both release shoplifters but I detain 80% of violent criminals whereas you detain 60%)

Likely to be inconsistent with other assumptions…

Consider a particular

$$E\left[D_{ij}\middle|r, X_i\right] = \phi_{jr} + X_i'\beta$$

Let $\widetilde{X}$ be the value such that $\phi_{jr} + \tilde{X}\,\beta = 1$

Then we must have

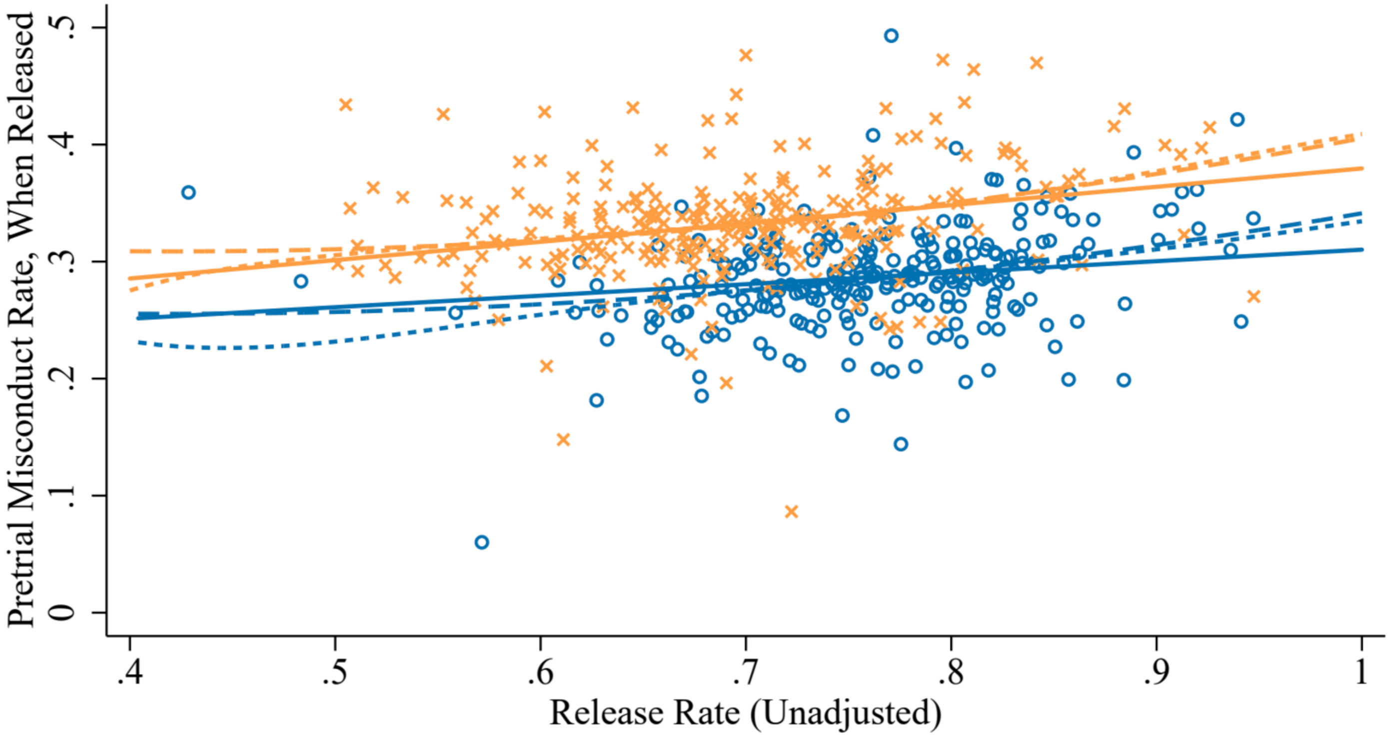
$$E\left[Y_i^*\middle|D_{ij} = 1, r, \tilde{X}\right] = \psi_{jr} + \tilde{X}\,\gamma = \mu_r(\tilde{X}\,)$$

which does not vary by j

Therefore, we must have $\psi_{jr} = \psi_{kr}$ whenever ϕ_{jr}= ϕ_{kr}

Why I'm Not So Worried in this Case

- Unadjusted release rates are high (mean = 73%)
- Robustness checks fit model separately by location
- And…


Pretrial Misconduct Rate, When Released
0
.1
.2
.3
.4
.5
.4
.5
.6
.7
.8
.9
1
Release Rate (Unadjusted)

An important priority for future work in this space should be to incorporate observables explicitly and coherently