Supplementary Materials to "Identifying Prediction Mistakes in Observational Data"

Ashesh Rambachan

October 14, 2021

F Expected Utility Maximization Behavior with Continuous Characteristics

In this section, I extend the setting described in Section 2 to allow for the characteristics $X \in \mathcal{X}$ to be continuously distributed. I focus this extension on the case of a screening decision for simplicity. The outcome $Y^* \in \mathcal{Y}$, the choices $C \in \mathcal{C} := \{0, 1\}$ and the characteristics $W \in \mathcal{W}$ are still finite. I now allow the characteristics $X \in \mathcal{X} \subseteq \mathbb{R}^{d_x}$ to be continuously distributed. The random vector $(W, X, C, Y^*) \sim P$ is defined over $\mathcal{W} \times \mathcal{X} \times \mathcal{C} \times \mathcal{Y}$ and summarizes the joint distribution of the characteristics, choices and latent outcome. I assume the joint distribution P admits a density $p(w, x, c, y^*)$ that is continuous in x at each value $(w, c, y^*) \in \mathcal{W} \times \mathcal{C} \times \mathcal{Y}$ and satisfies $p(w, x) > 0$ for all $\mathcal{W} \times \mathcal{X}$.

The researcher observes the characteristics (W, X) and the decision maker's choice C in each decision, but only observes the outcome Y^* if the decision maker selected $C = 1$. Therefore, the researcher observes the joint distribution $(W, X, C, Y) \sim P$, where $Y := Y^* \cdot 1\{C = 1\}$. The researcher places bounds on the unobservable choice-dependent outcome probabilities by specifying a family of conditional densities over (x, y^*) conditional on $W = w, C = c$, denoted by $\mathcal{B}_{c,w}$. Whenever the decision maker selects $C = 1$, the researcher observes (W, X, C, Y^*), and so $\mathcal{B}_{1,w}$ is a singleton that only contains the observable density $p(x, y^* \mid C = 1, W = w)$. Over the choice $c = 0$, the set $\mathcal{B}_{0,w}$ is a set of joint densities $\tilde{p}(x, y^* \mid C = 0, W = w)$ that (i) $p(x, y^* \mid C = 0, W = w) \in \mathcal{B}_{0,w}$, and (ii) $\sum_{y^* \in \mathcal{Y}} \tilde{p}(x, y^* \mid C = 0, W = w) = p(X = x \mid C = 0, W = w)$ for all $\tilde{p}(x, y^* \mid C = 0, W = w) \in \mathcal{B}_{0,w}$.

The expected utility maximization model requires minimal modification to account for the continuous characteristics. The definition of a utility function and private information is unchanged. Definition 3 is extended to ask whether there exists a random vector $(W, X, C, V, Y^*) \sim Q$ that admits a density $q(w, x, v, c, y^*)$ that is consistent with the observable data ("Data Consistency") by replacing the probability mass functions with the appropriate probability density functions. Analogously, the characterization of expected utility maximization behavior also extends directly.

Theorem F.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ and $\widetilde{p}(x, y^* \mid C = 0, W = w) \in \mathcal{B}_{0,w}$ for all $w \in \mathcal{W}$ such that*

$$\mathbb{E}_Q \left[U(c, Y^*; W) \mid C = c, W = w, X = x \right] \geq \mathbb{E}_Q \left[U(c', Y^*; W) \mid C = c, W = w, X = x \right],$$

for all $c \in \{0, 1\}$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_c(w, x) > 0$ and $c' \neq c$, where $\mathbb{E}_Q[\cdot]$ is the expectation under Q with density q satisfying

$$q(w, x, 1, y^*) = p(w, x, 1, y^*),$$

$$q(w, x, 0, y^*) = \widetilde{p}(x, y^* \mid C = 0, W = w)p(C = 0, W = w).$$

Proof. The proof of this result is analogous to the proof of Theorem 2.1 Towards this, I first extend Lemma E.1 to the case with continuous characteristics (which analyzes the case with multiple choices and $|\mathcal{C}| = N_c$). Throughout, I write $p_{C,Y^*}(c, y^* \mid w, x) := P(C = c, Y^* = y^* \mid W = w, X = x)$ as shorthand, where notation such as $p_{X,Y^*}(x, y^* \mid w, c)$ is defined analogously.

Lemma F.1. *The decision maker's choices are consistent with expected utility maximization behavior if and only if there exists a utility function $U \in \mathcal{U}$ that satisfies*

i. For all $c \in \mathcal{C}^y$, $(w, x) \in \mathcal{W} \times \mathcal{X}$ and $c' \neq c$,

$$\sum_{y^* \in \mathcal{Y}} p_{C,Y^*}(c, y^* \mid w, x)U(c, y^*; w) \geq \sum_{y^* \in \mathcal{Y}} p_{C,Y^*}(c, y^* \mid w, x)U(c', y^*; w).$$

ii. For all $c \in \mathcal{C} \setminus \mathcal{C}^y$ and $w \in \mathcal{W}$, there exists $\widetilde{p}_{C,Y^*}(\cdot \mid w, c) \in \mathcal{B}_{w,c}$ such that

$$\sum_{y^* \in \mathcal{Y}} \widetilde{p}_{C,Y^*}(c, y^* \mid w, x)U(c, y^*; w) \sum_{y^* \in \mathcal{Y}} \widetilde{p}_{C,Y^*}(c, y^* \mid w, x)U(c', y^*; w)$$

for all $x \in \mathcal{X}$ and $c' \neq c$, where $\widetilde{p}_{C,Y^*}(c, y^* \mid w, x) = \widetilde{p}_{X,Y^*}(x, y^* \mid w, c)p_{W,C}(w, c)/p_{W,X}(w, x)$.

Proof of Necessity for Lemma F.1: The proof of necessity follows the same argument as the proof of necessity of Lemma E.1 below by replacing the probability mass function Q with the density q. □

Proof of Sufficiency for Lemma F.1: The proof of sufficiency follows the proof of sufficiency of Lemma E.1 below by again simply replacing all probability mass functions with the appropriate density function. □

Theorem F.1 then follows directly from Lemma F.1 by considering the special case with $\mathcal{C} = \{0, 1\}$ and $\mathcal{C}^y = \{1\}$. □

Theorem F.1 can be applied to analyze the special case in which the latent outcome Y^* is binary. The bounds on the unobservable choice-dependent outcome probability $\mathcal{B}_{0,w}$ are simply joint densities $\tilde{p}(x, Y^* = 0 \mid W = w, C = 0)$, $\tilde{p}(x, Y^* = 1 \mid W = w, C = 0)$ that are continuous in $x \in \mathcal{X}$ and satisfy $p(x, Y^* = 0 \mid W = w, C = 0) + p(x, Y^* = 1 \mid W = w, C = 0) = p(x \mid W = w, C = 0)$. From Theorem F.1, the decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function U if and only if for all $w \in \mathcal{W}$ there exists $U(0, 0; w) < 0, U(1, 1; w) < 0$ satisfying

$$\sup_{x \in \mathcal{X}^1(w)} p(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{U(0, 0; w)}{U(0, 0; w) + U(1, 1; w)} \tag{12}$$

$$\frac{U(0, 0; w)}{U(0, 0; w) + U(1, 1; w)} \leq \inf_{x \in \mathcal{X}^0(w)} \overline{p}(Y^* = 1 \mid C = 0, W = w, X = x), \tag{13}$$

where $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w, x) > 0\}$, $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w, x) > 0\}$ and $\overline{p}(Y^* = 1 \mid C = 0, W = w, X = x)$ is the upper bound on $\tilde{p}(x, Y^* = 1 \mid W = w, C = 0)/p(x \mid$

$W = w, C = 0)$ over densities satisfying the bounds $\mathcal{B}_{c=0,w}$. This provides a sharp characterization of the identified set of strict preference utility functions in terms of "intersection bounds." Therefore, researchers test whether a decision maker's choices are consistent with expected utility maximization behavior at accurate beliefs using inference procedures developed in, for example, Chernozhukov, Lee and Rosen (2013).

Finally, Theorem F.1 can also be simplified through dimension reduction over the continuously distributed characteristics $X \in \mathcal{X}$. Consider functions $D_w \colon \mathcal{X} \to \{1, \ldots, N_d\}$ for each $w \in \mathcal{W}$ that partition the characteristic space into level sets $\{x \in \mathcal{X} \colon D_w(x) = d\}$. In a binary screening decision, if the decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function U that satisfies an exclusion restriction on the characteristics X, then

$$\max_{d \in \{1,\ldots,d_w\}} P(Y = 1 \mid C = 1, W = w, D_w(X) = d) \leq \frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \tag{14}$$

$$\frac{U(0,0;w)}{U(0,0;w) + U(1,1;w)} \leq \overline{P}(Y = 1 \mid C = 1, W = w, D_w(X) = d) \tag{15}$$

holds for all $w \in \mathcal{W}$.

G Alternative Bounds on the Missing Data

In the main text, I discussed how researchers can construct bounds on the missing data using a randomly assigned instrument. I now discuss alternative assumptions under which researchers can construct bounds on the missing data. I define these alternative assumptions for the leading case of a screening decision with a binary outcome.

G.1 Direct Imputation

The simplest empirical strategy for constructing bounds on the unobservable choice-dependent outcome probabilities is "direct imputation." In a binary screening decision, direct imputation uses the observable $P_{Y^*}(1 \mid 1, w, x)$ to bound the unobservable $P_{Y^*}(1 \mid 0, w, x)$.

Assumption G.1. For each $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $0 < \pi_1(w, x) < 1$, there exists $\kappa_{w,x} \geq 0$ satisfying

$$P_{Y^*}(1 \mid 1, w, x) \leq P_{Y^*}(1 \mid 0, w, x) \leq (1 + \kappa_{w,x}) P_{Y^*}(1 \mid 1, w, x).$$

The parameter $\kappa_{w,x} \geq 0$ specifies how different the unobservable choice-dependent outcome probability may be relative to the observable choice-dependent outcome probability. In pretrial release decisions, setting $\kappa_{w,x} = 1$ means that the researcher is willing to assume that the conditional probability of pretrial misconduct among detained defendants is no more than two times the conditional probability of pretrial misconduct among release defendants. Such bounding assumptions are used in, for example, Kleinberg et al. (2018*a*), and Jung et al. (2020*a*).

In practice, the researcher may wish to test whether the decision maker is making systematic prediction mistakes under various choices of the parameter $\kappa_{w,x}$, and thereby conduct a sensitivity analysis of how robust the behavioral conclusions are to various assumptions about the unobservable choice-dependent outcome probabilities. I illustrate such a sensitivity analysis in Appendix I,

where I report how the fraction of judges for whom we can reject expected utility maximization behavior varies as the parameter $\kappa_{w,x}$ varies in the New York City pretrial release setting.

Finally, Assumption G.1 has a natural interpretation under the expected utility maximization model. The parameter $\kappa_{w,x}$ bounds the average informativeness of the decision maker's private information $V \in \mathcal{V}$.

Proposition G.1. *Consider a screening decision with a binary outcome and suppose Assumption G.1 holds. If the decision maker's choices are consistent with expected utility maximization behavior at some private information $V \in \mathcal{V}$ and joint distribution $(W, X, V, C, Y^*) \sim Q$, then for each $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $0 < \pi_1(w, x) < 1$ and $0 < P_{Y^*}(1 \mid 1, w, x) < 1$*

a. $1 \leq \frac{Q(C=0|Y^*=1,W=w,X=x)/Q(C=1|Y^*=1,W=w,X=x)}{Q(C=0|W=w,X=x)/Q(C=1|W=w,X=x)} \leq 1 + \kappa_{w,x}$,

b. $1 - \kappa_{w,x}\frac{P_{Y^*}(1|1,w,x)}{P_{Y^*}(0|1,w,x)} \leq \frac{Q(C=0|Y^*=0,W=w,X=x)/Q(C=1|Y^*=0,W=w,X=x)}{Q(C=0|W=w,X=x)/Q(C=1|W=w,X=x)} \leq 1.$

Proof. Notice that

$$\frac{Q(C = 0 \mid Y^* = 1, W = w, X = x)/Q(C = 1 \mid Y^* = 1, W = w, X = x)}{Q(C = 0 \mid W = w, X = x)/Q(C = 1 \mid W = w, X = x)}$$

$$= \frac{Q(Y^* = 1 \mid C = 0, W = w, X = x)}{Q(Y^* = 1 \mid C = 1, W = w, X = x)}.$$

Since the decision maker's choices are consistent with expected utility maximization behavior, $(W, X, V, C, Y^*) \sim Q$ satisfies the data consistency condition in Definition 3 at some $\tilde{P}_{Y^*}(\cdot \mid 0, w, x)$ satisfying the bounds in Assumption G.1 for each $(w, x) \in \mathcal{W} \times \mathcal{X}$. Therefore, $Q(Y^* = 1 \mid C = 0, W = w, X = x) = \tilde{P}_{Y^*}(1 \mid 0, w, x)$ and it immediately follows that $\frac{Q(Y^*=1|C=0,W=w,X=x)}{Q(Y^*=1|C=1,W=w,X=x)} = \frac{\tilde{P}_{Y^*}(1|0,w,x)}{P_{Y^*}(1|1,w,x)} \in [1, 1 + \kappa_{w,x}]$ under Assumption G.1. This proves (a). To show (b), notice that the bounds in Assumption G.1 imply that

$$P_{Y^*}(0 \mid 1, w, x) - \kappa_{w,x} P_{Y^*}(1 \mid 1, w, x) \leq P_{Y^*}(0 \mid 0, w, x) \leq P_{Y^*}(0 \mid C = 1, w, x).$$

Moreover, as before, we also have that

$$\frac{Q(C = 0 \mid Y^* = 0, W = w, X = x)/Q(C = 1 \mid Y^* = 0, W = w, X = x)}{Q(C = 0 \mid W = w, X = x)/Q(C = 1 \mid W = w, X = x)} = \frac{\tilde{P}_{Y^*}(0 \mid 0, w, x)}{P_{Y^*}(0 \mid 1, w, x)}$$

and (b) then follows immediately. □

The direct imputation bounds imply bounds on the relative odds ratio of the decision maker's choice probabilities conditional on the outcome and the characteristics relative to their choice probabilities conditional on only the characteristics. This places a bound on the average informativeness of the decision maker's private information under the expected utility maximization model since

$$Q(C = 1 \mid Y^* = 1, W = w, X = x) = \mathbb{E}_Q\left[Q(C = 1 \mid V = v, W = w, X = x) \mid Y^* = 1, W = w, X = x\right]$$
$$Q(C = 1 \mid Y^* = 0, W = w, X = x) = \mathbb{E}_Q\left[Q(C = 1 \mid V = v, W = w, X = x) \mid Y^* = 0, W = w, X = x\right]$$

under the Information Set condition in Definition 3. In this sense, the direct imputation bounds are related to classic approaches for modelling violations of unconfoundedness in causal inference such as Rosenbaum (2002), which model violations of unconfoundedness by postulating that there exists some unobserved characteristics V that governs selection and places bounds on the magnitude of the relative odds ratio of the propensity score conditional on V and the observable characteristics versus the propensity score conditional on just the observable characteristics. See Imbens (2003), which develops a tractable parametric model for such a violation of unconfoundedness in a treatment assignment problem. Kallus, Mao and Zhou (2018) and Yadlowsky et al. (2020) derive bounds on the conditional average treatment effect and average treatment effect under related models for violations of unconfoundedness, and provide methods for inference on the derived bounds.

G.2 Proxy Outcomes

In some empirical applications, the researcher may observe an additional proxy outcome $\tilde{Y} \in \tilde{\mathcal{Y}}$, which does not suffer from the missing data problem but is correlated with the outcome $Y^* \in \mathcal{Y}$. By specifying bounds on the relationship between the proxy outcome $\tilde{Y}$ and the outcome Y^*, the researcher may construct bounds on the unobservable choice-dependent outcome probabilities.

Proxy outcomes are common in medical and consumer lending settings. For example, Mullainathan and Obermeyer (2021) observe each patient's longer term health outcomes regardless of whether a stress test for a heart attack was conducted during a particular emergency room visit. A patient's longer term health outcomes are related to whether the patient actually had a heart attack, no matter the testing decisions of doctors. Similarly, Chan, Gentzkow and Yu (2021) observe whether each patient had a future pneumonia diagnosis within one week of an initial examination, regardless of whether a doctor ordered an MRI at the initial examination. Future pneumonia diagnoses may be a useful proxy for whether the doctor failed to correctly diagnose pneumonia during the initial examination. In mortgage approvals, Blattner and Nelson (2021) observe each loan applicant's default performance on other credit products, regardless of whether each loan applicant was approved for a mortgage. A loan applicant's default performance on other credit products is related with whether they would have defaulted on the mortgage.

Assumption G.2 (Proxy Outcomes)**.** There exists a proxy outcome $\widetilde{Y} \in \widetilde{\mathcal{Y}}$, and the researcher observes the joint distribution $(W, X, C, \widetilde{Y}, Y^* \cdot C) \sim P$. Assume $P(W = w, X = X, \widetilde{Y} = \widetilde{y}) > 0$ for all $(w, x, \widetilde{y}) \in \mathcal{W} \times \mathcal{X} \times \widetilde{\mathcal{Y}}$.

Over decisions in which the decision maker selected $C = 1$, the researcher observes the joint distribution of the proxy outcome and the outcome $P(\tilde{Y} = \tilde{y}, Y^* = y^* \mid C = 1, W = w, X = x)$. By placing assumptions on how the joint distribution of the proxy outcome and the outcome conditional on $C = 0$ is bounded by the observable joint distribution of the proxy outcome and the outcome conditional on $C = 1$, the researcher can construct bounds on the unobservable choice-dependent outcome probabilities of the form given in Assumption 2.1. Let $P_{Y^*}(y^* \mid \tilde{y}, c, w, x) := P(Y^* = y^* \mid \tilde{Y} = \tilde{y}, C = c, W = w, X = x)$, $P_{\tilde{Y}}(\tilde{y} \mid c, w, x) := P(\tilde{Y} = \tilde{y} \mid C = c, W = w, X = x)$, and $\pi_c(\tilde{y}, w, x) := P(C = c \mid \tilde{Y} = \tilde{y}, W = w, X = x)$.

Assumption G.3 (Proxy Bounds)**.** For each $(w, x, \tilde{y}) \in \mathcal{W} \times \mathcal{X} \times \tilde{\mathcal{Y}}$ satisfying $0 < \pi_1(\tilde{y}, w, x) < 1$,

there exists bounds $\underline{\kappa}_{\tilde{y},w,x}, \overline{\kappa}_{\tilde{y},w,x} \geq 0$ satisfying

$$(1 - \underline{\kappa}_{\tilde{y},w,x}) P_{Y^*}(1 \mid \tilde{y}, 1, w, x) \leq P_{Y^*}(1 \mid \tilde{y}, 0, w, x) \leq (1 + \overline{\kappa}_{\tilde{y},w,x}) P_{Y^*}(1 \mid \tilde{y}, 1, w, x).$$

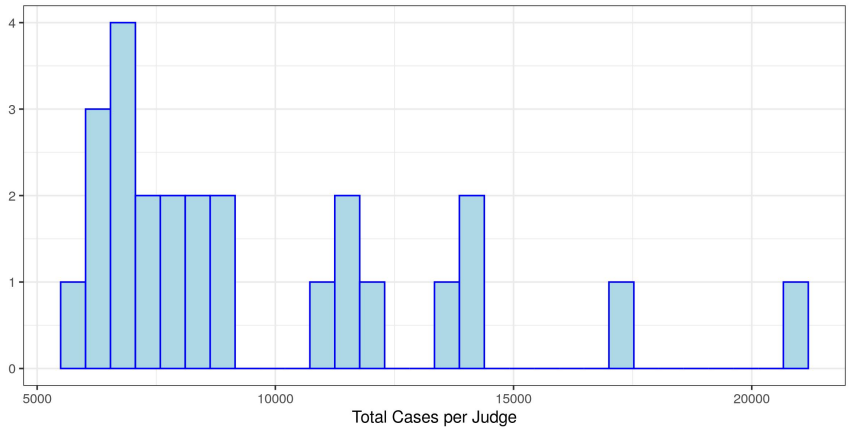
Proposition G.2. *Consider a binary screening decision in which Assumptions G.2-G.3 hold. For each* $(w, x) \in \mathcal{W} \times \mathcal{X}$*,*

$$\sum_{\tilde{y} \in \tilde{\mathcal{Y}}} (1 + \underline{\kappa}_{\tilde{y},w,x}) P_{Y^*}(1 \mid \tilde{y}, 1, w, x) P_{\tilde{Y}}(\tilde{y} \mid 0, w, x) \leq P_{Y^*}(1 \mid 0, w, x),$$
$$P_{Y^*}(1 \mid 0, w, x) \leq \sum_{\tilde{y} \in \tilde{\mathcal{Y}}} (1 + \overline{\kappa}_{\tilde{y},w,x}) P_{Y^*}(1 \mid \tilde{y}, 1, w, x) P_{\tilde{Y}}(\tilde{y} \mid 0, w, x).$$

The advantage of this approach is that it may be easier to express domain-specific knowledge through the use of proxy outcomes. For example, in the mortgage approvals setting in Blattner and Nelson (2021), the proxy bounds summarize the extent to which the mortgage default rate among accepted applicants that also defaulted on other credit products differs from the counterfactual mortgage default rate among rejected applicants that also defaulted on other credit products. In the medical testing setting in Mullainathan and Obermeyer (2021), the proxy bounds summarize the extent to which heart attack rate among tested patients that went on to die within 30 days of their emergency room visit differs from the heart attack rate among untested patients that went on to die within 30 days of their emergency room visit.

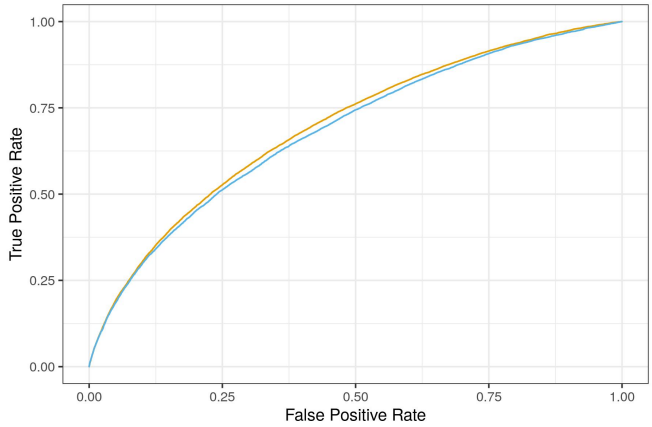
H Summary Figures and Tables for New York City Pretrial Release

Figure S1: Histogram of number of cases heard by each judge in the top 25 judges.

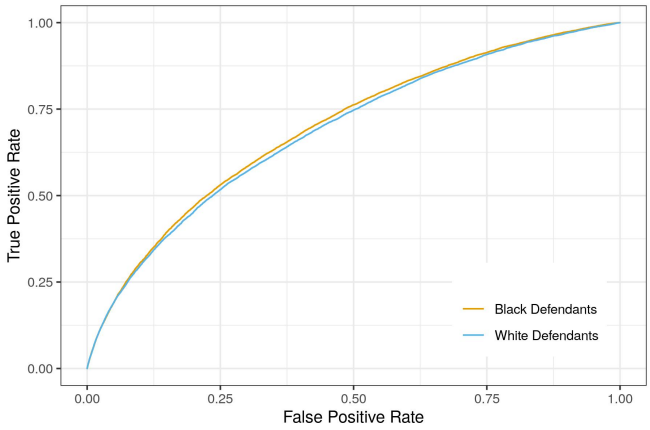


Notes: This figure plots a histogram of the number of cases heard per judge in the top 25 judges that are the focus of my empirical analysis in the New York City pretrial release data. Every judge in the top 25 made at least 5,000 pretrial release decisions over the sample period. See Section 5.2 for further details. Source: Rambachan and Ludwig (2021).

Figure S2: Receiver-operating characteristic (ROC) curves for ensemble prediction functions



(a) Race-by-age cells



(b) Race-by-felony charge cells

Notes: This figure plots the Receiver-Operating Characteristic (ROC) curves for the ensemble prediction that predicts failure to appear among defendants that were released by the top 25 judges. It reports the ROC curve for the ensemble prediction function constructed within race-by-age cells and race-by-felony charge cells separately. Age is binarized into young and older defendants, where older defendants are defined as defendants older than 25 years. The ensemble prediction function is constructed over cases heard by the remaining bail judges and evaluated out-of-sample on cases heard by the top 25 judges. The ensemble prediction function averages the predictions of a random forest, which is estimated using the the `R` package `ranger` at the default hyperparameter values (Wright and Ziegler, 2017), and an elastic net model, whose hyperparameters are tuned using three-fold cross-validation. The ROC curve plots the false positive rate on the x-axis and the true positive rate on the y-axis. The out-of-sample area under the curve (AUC) on all defendants equals 0.693 for the ensemble prediction function constructed over race-by-age cells and 0.694 for the ensemble prediction function constructed over race-by-felony cells. See Section 5.3.1 for further details. Source: Rambachan and Ludwig (2021).

Table S1: Summary statistics comparing the main estimation sample and cases heard by the top 25 judges, broken out by defendant race.

	All Defendants		White Defendants		Black Defendants	
	Estimation Sample	Top Judges	Estimation Sample	Top Judges	Estimation Sample	Top Judges
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.720	0.736	0.757	0.777	0.687	0.699
Defendant Characteristics						
White	0.475	0.481	1.000	1.000	0.000	0.000
Female	0.173	0.173	0.154	0.152	0.190	0.192
Age at Arrest	31.95	31.75	32.03	31.88	31.87	31.63
Arrest Charge						
Number of Charges	1.152	1.167	1.187	1.217	1.119	1.121
Felony Charge	0.372	0.367	0.367	0.356	0.376	0.377
Any Drug Charge	0.253	0.224	0.253	0.217	0.253	0.230
Any DUI Charge	0.047	0.049	0.070	0.072	0.027	0.027
Any Violent Crime Charge	0.375	0.395	0.358	0.379	0.390	0.410
Property Charge	0.130	0.132	0.122	0.123	0.138	0.140
Defendant Priors						
Any FTA	0.516	0.497	0.443	0.419	0.582	0.570
Number of FTAs	2.177	2.034	1.633	1.492	2.670	2.537
Any Misdemeanor Arrest	0.683	0.667	0.615	0.596	0.744	0.734
Any Misdemeanor Conviction	0.383	0.368	0.334	0.315	0.427	0.418
Any Felony Arrest	0.581	0.566	0.503	0.482	0.652	0.644
Any Felony Conviction	0.285	0.271	0.234	0.215	0.331	0.323
Any Violent Felony Arrest	0.398	0.387	0.306	0.292	0.481	0.476
Any Violent Felony Conviction	0.119	0.114	0.084	0.078	0.150	0.147
Total Cases	569,256	243,118	270,704	117,073	298,552	126,045

Notes: This table provides summary statistics about defendant and case characteristics for the main estimation sample and the cases heard by the top 25 judges in the New York City pretrial release data for all defendants and separately by the race of the defendant. See Section 5.2 for further discussion. Source: Rambachan and Ludwig (2021).

Table S2: Summary statistics for released and detained defendants in the main estimation sample and for cases heard by the top 25 judges

	All Defendants		Released Defendants		Detained Defendants	
	Estimation Sample	Top Judges	Estimation Sample	Top Judges	Estimation Sample	Top Judges
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.720	0.736	1.000	1.000	0.000	0.000
Defendant Characteristics						
White	0.475	0.481	0.499	0.508	0.412	0.407
Female	0.173	0.173	0.199	0.197	0.107	0.106
Age at Arrest	31.95	31.75	31.22	31.20	33.82	33.29
Arrest Charge						
Number of Charges	1.152	1.167	1.148	1.162	1.161	1.182
Felony Charge	0.372	0.367	0.288	0.288	0.588	0.586
Any Drug Charge	0.253	0.224	0.229	0.204	0.314	0.279
Any DUI Charge	0.047	0.049	0.062	0.063	0.010	0.010
Any Violent Crime Charge	0.375	0.395	0.388	0.409	0.341	0.355
Property Charge	0.130	0.132	0.115	0.114	0.171	0.181
Defendant Priors						
Any FTA	0.516	0.497	0.409	0.395	0.793	0.784
Number of FTAs	2.177	2.034	1.362	1.295	4.284	4.103
Any Misdemeanor Arrest	0.683	0.667	0.610	0.598	0.871	0.863
Any Misdemeanor Conviction	0.383	0.368	0.284	0.278	0.637	0.621
Any Felony Arrest	0.581	0.566	0.487	0.477	0.824	0.814
Any Felony Conviction	0.285	0.271	0.200	0.194	0.505	0.487
Any Violent Felony Arrest	0.398	0.387	0.315	0.309	0.614	0.608
Any Violent Felony Conviction	0.119	0.114	0.081	0.080	0.216	0.210
Total Cases	569,256	243,118	410,394	179,143	158,862	63,975

Notes: This table provides summary statistics about defendant and case characteristics for the main estimation sample and the cases heard by the top 25 judges in the New York City pretrial release data for all defendants and separately by whether the defendant was released or detained. See Section 5.2 for further discussion. Source: Rambachan and Ludwig (2021).

Table S3: Summary statistics of misconduct rates among released defendants in the main estimation sample and cases heard by the top 25 judges.

	All Defendants		White Defendants		Black Defendants	
	Estimation Sample	Top Judges	Estimation Sample	Top Judges	Estimation Sample	Top Judges
	(1)	(2)	(3)	(4)	(5)	(6)
Failure to Appear (FTA)	0.151	0.146	0.135	0.131	0.167	0.161
Rearrest (NCA)	0.261	0.257	0.230	0.225	0.292	0.289
Any Misconduct	0.331	0.324	0.297	0.290	0.366	0.359
Total Cases	410,394	179,143	205,174	91,026	205,220	88,117

Notes: This table summarizes the observed misconduct rates among released defendants for the main estimation sample and the cases heard by the top 25 judges in the New York City pretrial release data for all defendants and separately by the race of the defendant. See Section 5.2 for further discussion. Source: Rambachan and Ludwig (2021).

Table S4: Summary statistics in the universe of all cases subject to a pretrial release decision and the main estimation sample in the New York City pretrial release data, broken out by defendant race.

	All Defendants		White Defendants		Black Defendants	
	Full Sample	Estimation Sample	Full Sample	Estimation Sample	Full Sample	Estimation Sample
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.736	0.720	0.765	0.757	0.691	0.687
Defendant Characteristics						
White	0.457	0.475	1.000	1.000	0.000	0.000
Female	0.169	0.173	0.153	0.154	0.184	0.190
Age at Arrest	32.06	31.95	32.06	32.03	31.88	31.87
Arrest Charge						
Number of Charges	1.165	1.152	1.176	1.187	1.114	1.119
Felony Charge	0.335	0.372	0.332	0.367	0.346	0.376
Any Drug Charge	0.244	0.253	0.251	0.253	0.252	0.253
Any DUI Charge	0.053	0.047	0.074	0.070	0.028	0.027
Any Violent Crime Charge	0.365	0.375	0.348	0.358	0.380	0.390
Property Charge	0.135	0.130	0.127	0.122	0.145	0.138
Defendant Priors						
Any FTA	0.499	0.516	0.442	0.443	0.586	0.582
Number of FTAs	2.099	2.177	1.635	1.633	2.707	2.670
Any Misdemeanor Arrest	0.668	0.683	0.616	0.615	0.747	0.744
Any Misdemeanor Conviction	0.371	0.383	0.335	0.334	0.430	0.427
Any Felony Arrest	0.565	0.581	0.502	0.503	0.654	0.652
Any Felony Conviction	0.273	0.285	0.232	0.234	0.334	0.331
Any Violent Felony Arrest	0.384	0.398	0.306	0.306	0.484	0.481
Any Violent Felony Conviction	0.114	0.119	0.084	0.084	0.152	0.150
Total Cases	758,027	569,256	347,006	270,704	370,793	298,552

Notes: This table provides summary statistics about defendant and case characteristics for the sample of all cases subject to a pretrial release decision and the main estimation sample in the New York City pretrial release data, broken out for all defendants and by the race of the defendant. Cases in the main estimation sample appear to have more severe charges than the main estimation sample. See Section 5.2 for further discussion. Source: Rambachan and Ludwig (2021).

Table S5: Summary statistics for released and detained defendants in the universe of all cases subject to a pretrial release decision and the main estimation sample in the New York City pretrial release data.

	All Defendants		Released Defendants		Detained Defendants	
	Full Sample	Estimation Sample	Full Sample	Estimation Sample	Full Sample	Estimation Sample
	(1)	(2)	(3)	(4)	(5)	(6)
Released before trial	0.736	0.720	1.000	1.000	0.000	0.000
Defendant Characteristics						
White	0.457	0.475	0.476	0.499	0.406	0.412
Female	0.169	0.173	0.192	0.199	0.105	0.107
Age at Arrest	32.06	31.95	31.41	31.22	33.90	33.82
Arrest Charge						
Number of Charges	1.165	1.152	1.166	1.148	1.161	1.161
Felony Charge	0.335	0.372	0.258	0.288	0.549	0.588
Any Drug Charge	0.244	0.253	0.221	0.229	0.307	0.314
Any DUI Charge	0.053	0.047	0.068	0.062	0.011	0.010
Any Violent Crime Charge	0.365	0.375	0.377	0.388	0.333	0.341
Property Charge	0.135	0.130	0.119	0.115	0.179	0.171
Defendant Priors						
Any FTA	0.499	0.516	0.394	0.409	0.792	0.793
Number of FTAs	2.099	2.177	1.310	1.362	4.304	4.284
Any Misdemeanor Arrest	0.668	0.683	0.596	0.610	0.870	0.871
Any Misdemeanor Conviction	0.371	0.383	0.275	0.284	0.639	0.637
Any Felony Arrest	0.565	0.581	0.472	0.487	0.823	0.824
Any Felony Conviction	0.273	0.285	0.190	0.200	0.503	0.505
Any Violent Felony Arrest	0.384	0.398	0.302	0.315	0.612	0.614
Any Violent Felony Conviction	0.114	0.119	0.077	0.081	0.216	0.216
Total Cases	758,027	569,256	558,167	410,394	199,860	158,862

Notes: This table provides summary statistics about defendant and case characteristics for the sample of all cases subject to a pretrial release decision and the main estimation sample in the New York City pretrial release data, broken out for all defendants and by whether the defendant was released or detained. The main estimation sample has a slightly lower release rate than the universe of cases. See Section 5.2 for further discussion. Source: Rambachan and Ludwig (2021).

Table S6: Balance check estimates for the quasi-random assignment of judges for all defendants and by defendant race.

	All Defendants (1)	White Defendants (2)	Black Defendants (3)
Defendant Characteristics			
Black	−0.00011		
	(0.00008)		
Female	0.000003	0.00005	−0.00003
	(0.00013)	(0.00017)	(0.00017)
Age	−0.00001	−0.00002	−0.000002
	(0.000003)	(0.00001)	(0.000004)
Arrest Charge			
Number of Charges	−0.000003	−0.000003	0.000003
	(0.00001)	(0.00001)	(0.00003)
Felony Charge	0.00009	−0.00012	0.00027
	(0.00015)	(0.00017)	(0.00018)
Any Drug Charge	−0.00012	−0.00010	−0.00013
	(0.00013)	(0.00018)	(0.00016)
Any Violent Crime Charge	−0.00004	−0.00013	0.00004
	(0.00010)	(0.00015)	(0.00014)
Any Property Charge	−0.00033	−0.00029	−0.00035
	(0.00016)	(0.00019)	(0.00025)
Any DUI Charge	0.00044	0.00039	0.00028
	(0.00024)	(0.00027)	(0.00039)
Defendant Priors			
Prior FTA	−0.00004	−0.00011	0.00003
	(0.00010)	(0.00016)	(0.00013)
Prior Misdemeanor Arrest	0.00007	0.00003	0.00011
	(0.00010)	(0.00013)	(0.00015)
Prior Felony Arrest	0.00006	0.00003	0.00009
	(0.00014)	(0.00021)	(0.00019)
Prior Violent Felony Arrest	−0.00013	−0.00008	−0.00016
	(0.00011)	(0.00019)	(0.00016)
Prior Misdemeanor Conviction	0.00016	0.00021	0.00011
	(0.00013)	(0.00017)	(0.00016)
Prior Felony Conviction	−0.00019	0.00011	−0.00040
	(0.00012)	(0.00018)	(0.00015)
Prior Violent Felony Conviction	−0.00008	−0.00024	0.00002
	(0.00015)	(0.00021)	(0.00020)
Joint p-value	0.06953	0.15131	0.41840
Court × Time FE	✓	✓	✓
Cases	569,256	270,704	298,552

Notes: This table reports OLS estimates for regressions of the constructed judge leniency measure on various defendant and case characteristics in the main estimation sample. These regressions are estimated over all defendants and separately by the race of the defendant. Standard errors, reported in parentheses, are clustered at the defendant and judge level. The joint p-value is based on the F-statistic for whether all defendant and case characteristics are jointly significant. See Section 5.3.2 for further details. Source: Rambachan and Ludwig (2021).

Table S7: Balance check estimates for the quasi-random assignment of judges by defendant race and age.

	White Defendants		Black Defendants	
	Young (1)	Older (2)	Young (3)	Older (4)
Defendant Characteristics				
Female	−0.00008	0.00017	−0.00007	−0.00005
	(0.00025)	(0.00019)	(0.00024)	(0.00024)
Age	−0.000004	−0.00001	−0.00006	−0.00001
	(0.00004)	(0.00001)	(0.00003)	(0.00001)
Arrest Charge				
Number of Charges	−0.00002	−0.000003	−0.00002	0.00001
	(0.00003)	(0.000005)	(0.00006)	(0.00003)
Felony Charge	0.00002	−0.00024	0.00019	0.00033
	(0.00023)	(0.00019)	(0.00023)	(0.00022)
Any Drug Charge	−0.00033	0.00004	−0.00046	0.00004
	(0.00033)	(0.00022)	(0.00025)	(0.00020)
Any Violent Crime Charge	−0.00025	−0.00010	−0.00016	0.00018
	(0.00026)	(0.00019)	(0.00024)	(0.00018)
Any Property Charge	−0.00005	−0.00046	−0.00017	−0.00045
	(0.00034)	(0.00023)	(0.00031)	(0.00029)
Any DUI Charge	0.00021	0.00042	−0.00160	0.00062
	(0.00045)	(0.00030)	(0.00072)	(0.00044)
Defendant Priors				
Prior FTA	−0.00013	−0.00015	0.00034	−0.00021
	(0.00026)	(0.00021)	(0.00022)	(0.00020)
Prior Misdemeanor Arrest	0.00026	−0.00018	−0.00008	0.00034
	(0.00021)	(0.00017)	(0.00022)	(0.00022)
Prior Felony Arrest	−0.00008	0.00018	0.00035	−0.00025
	(0.00026)	(0.00027)	(0.00030)	(0.00024)
Prior Violent Felony Arrest	−0.00024	−0.00001	−0.00020	−0.00019
	(0.00030)	(0.00023)	(0.00025)	(0.00021)
Prior Misdemeanor Conviction	0.00040	0.00023	0.00040	0.00004
	(0.00029)	(0.00025)	(0.00028)	(0.00018)
Prior Felony Conviction	0.00052	0.00005	−0.00094	−0.00016
	(0.00049)	(0.00019)	(0.00033)	(0.00017)
Prior Violent Felony Conviction	−0.00029	−0.00020	0.00113**	−0.00012
	(0.00077)	(0.00022)	(0.00054)	(0.00021)
Joint p-value	0.85104	0.44370	0.038862	0.16062
Court × Time FE	✓	✓	✓	✓
Cases	99,536	171,168	119,156	179,396

Notes: This table reports OLS estimates for regressions of the constructed judge leniency measure on various defendant and case characteristics in the main estimation sample. These regressions are estimated separately over subsamples defined on the race and age of the defendant, where "young" is defined as less than or equal to 25 years and "old" is defined as older than 25 years. Standard errors, reported in parentheses, are clustered at the defendant and judge level. The joint p-value is based on the F-statistic for whether all defendant and case characteristics are jointly significant. See Section 5.3.2 for further details. Source: Rambachan and Ludwig (2021).

Table S8: Balance check estimates for the quasi-random assignment of judges by defendant race and felony charge.

	White Defendants		Black Defendants	
	Felony Charge (1)	No Felony Charge (2)	Felony Charge (3)	No Felony Charge (4)
Defendant Characteristics				
Female	0.00003	0.00001	−0.00003	−0.00004
	(0.00023)	(0.00021)	(0.00026)	(0.00021)
Age	−0.00002	−0.00001	0.000004	−0.000004
	(0.00001)	(0.00001)	(0.00001)	(0.00001)
Arrest Charge				
Number of Charges	−0.000002	−0.00004	−0.000005	0.00003
	(0.00001)	(0.00003)	(0.00003)	(0.00007)
Any Drug Charge	−0.00022	−0.00008	−0.00012	−0.00008
	(0.00028)	(0.00024)	(0.00031)	(0.00023)
Any Violent Crime Charge	−0.00043	0.00001	0.00038	−0.00013
	(0.00030)	(0.00018)	(0.00026)	(0.00017)
Any Property Charge	−0.00038	−0.00038	0.00023	−0.00070
	(0.00027)	(0.00028)	(0.00029)	(0.00035)
Any DUI Charge	0.00047	0.00049	0.00100	0.00012
	(0.00057)	(0.00030)	(0.00093)	(0.00042)
Defendant Priors				
Prior FTA	−0.00014	−0.00005	0.00012	−0.00003
	(0.00023)	(0.00020)	(0.00024)	(0.00015)
Prior Misdemeanor Arrest	0.00024	−0.00012	0.00009	0.00010
	(0.00025)	(0.00017)	(0.00028)	(0.00018)
Prior Felony Arrest	−0.00007	−0.000005	−0.00043	0.00040
	(0.00036)	(0.00023)	(0.00032)	(0.00022)
Prior Violent Felony Arrest	−0.00042	0.00012	−0.00001	−0.00020
	(0.00029)	(0.00021)	(0.00025)	(0.00018)
Prior Misdemeanor Conviction	−0.00009	0.00050	0.00042	−0.00013
	(0.00030)	(0.00021)	(0.00027)	(0.00017)
Prior Felony Conviction	0.00010	0.00024	−0.00040	−0.00041
	(0.00034)	(0.00023)	(0.00025)	(0.00019)
Prior Violent Felony Conviction	0.00040	−0.00084	−0.00004	0.0000001
	(0.00036)	(0.00030)	(0.00028)	(0.00024)
Joint p-value	0.05623	0.27401	0.24607	0.24712
Court × Time FE	✓	✓	✓	✓
Cases	99,463	171,241	112,517	186,035

Notes: This table reports OLS estimates for regressions of the constructed judge leniency measure on various defendant and case characteristics. These regressions are estimated separately over subsamples defined on the race of the defendant and whether the defendant was charged with a felony offense. Standard errors, reported in parentheses, are clustered at the defendant and judge level. The joint p-value is based on the F-statistic for whether all defendant and case characteristics are jointly significant. See Section 3.3.2 for further details. Source: Rambachan and Ludwig (2021).

I Additional Empirical Results for New York City Pretrial Release

I now present additional empirical results on the behavior of judges in the New York City pretrial release system.

I.1 Welfare Effects of Automation Policies: Race-by-Felony Charge Cells

Section 6.2 of the main text compared the total expected social welfare under the observed release decisions by judges in new York City against the total expected social welfare under counterfactual algorithmic decisions, conducting this exercise over race-by-age cells and deciles of predicted failure to appear risk. In this section of the Supplement, I report the results of the same analysis over race-by-felony charge cells and deciles of predicted failure to appear risk for completeness and find analogous results as reported in the main text.

I.1.1 Automating Judges Who Make Prediction Mistakes

Figure S3a plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces judges who were found to make detectable prediction mistakes against the observed release decisions of these judges.

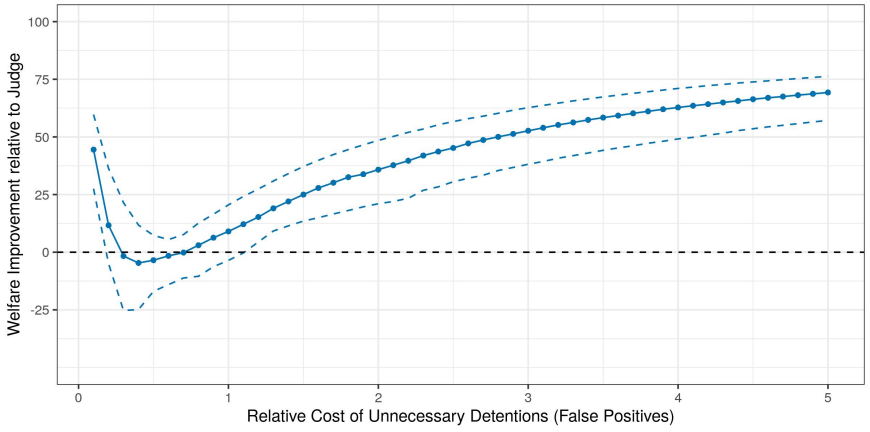
For most values of the social welfare function, the algorithmic decision rule dominates the observed choices of these judges, but for social welfare costs of unnecessary detentions ranging over $U^*(0,0) \in [0.3, 0.7]$, the algorithmic decision rule either leads to no improvement or strictly lowers worst-case expected total social welfare relative to the judges' observed decisions.

Figure S3b therefore plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that only corrects detectable prediction mistakes at the tails of the predicted failure to appear risk distribution against the observed release decisions of these judges. As found in the main text, the algorithmic decision rule that only corrects detectable prediction mistakes appears to weakly dominate the observed release decisions of judges, no matter the value of the social welfare function.

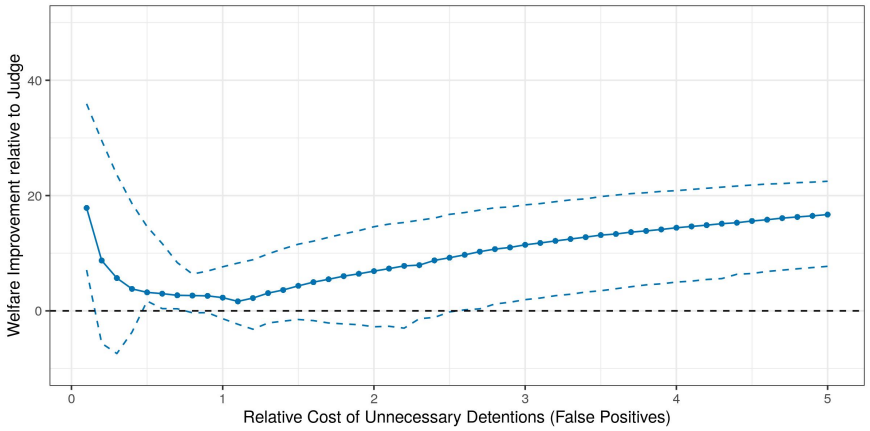
I.1.2 Automating Judges Who Do Not Make Prediction Mistakes

I next compare welfare effects of automating the release decisions of judges whose choices were found to be consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk. Figure S4 plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces these judges against their observed release decisions. As in the main text, I find that automating these judge's release decisions may strictly lower worst-case expected total social welfare for a large range of social welfare costs of unnecessary detentions,

Figure S3: Ratio of total expected social welfare under algorithmic decision rules relative to observed release decisions of judges that make detectable prediction mistakes over race-by-felony charge cells.



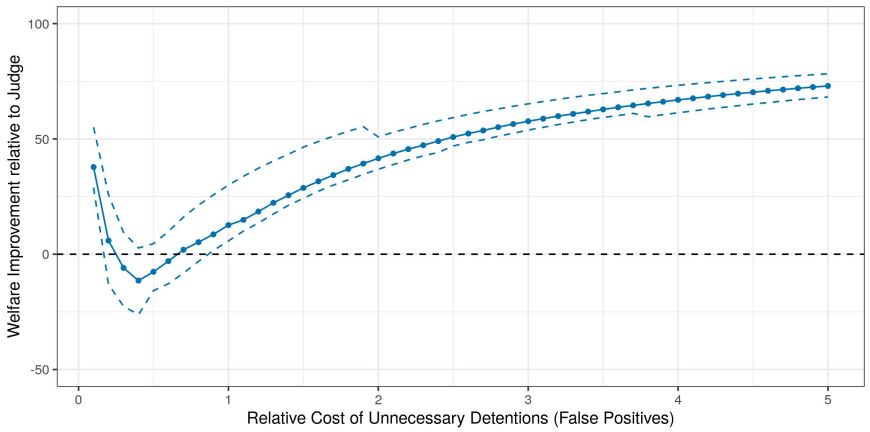
(a) Welfare improvement of full automation decision rule



(b) Welfare improvement of decision rule that corrects prediction mistakes

Notes: This figure reports the change in worst-case total expected social welfare under two algorithmic decision rules against the judge's observed release decisions among judges who were found to make detectable prediction mistakes. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-felony cells and deciles of predicted failure to appear risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear to in court $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges that make mistakes, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 of the main text and Supplement I.1 for further details. Source: Rambachan and Ludwig (2021).

Figure S4: Ratio of total expected social welfare under full automation decision rule relative to observed decisions of judges that do not make detectable prediction mistakes over race-by-felony charge cells.



Notes: This figure reports the change in worst-case total expected social welfare under the algorithmic decision rule that fully automates decision-making against the judge's observed release decisions among judges whose choices were consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk. Worst case total expected social welfare under each decision rule is computed by first constructing a 95% confidence interval for total expected social welfare under the decision rule, and reporting smallest value that lies in the confidence interval. These decisions rules are constructed and evaluated over race-by-felony cells and deciles of predicted risk. The x-axis plots the relative social welfare cost of detaining a defendant that would not fail to appear to in court $U^*(0,0)$ (i.e., an unnecessary detention). The solid line plots the median change across judges that make mistakes, and the dashed lines report the minimum and maximum change across judges. See Section 6.2 of the main text and Supplement I.1 for further details. Source: Rambachan and Ludwig (2021).

I.2 Identifying Prediction Mistakes: Direct Imputation

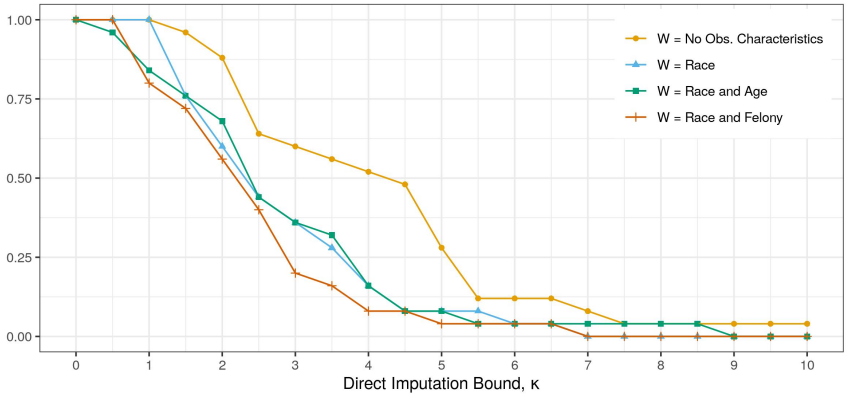
Sections 5-4 of the main text tested whether the pretrial release decisions of judges in New York City were consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk under various exclusion restrictions on their preferences by constructing bounds on the failure to appear rate of detained defendants using the quasi-random assignment of judges.

I now show how the same test may be conducted using direct imputation (Supplement G.1) to construct bounds on the failure to appear rate of detained defendants. The key input in constructing the direct imputation bounds is the parameter $\kappa_{w,d} \geq 0$ for each value $W = w$, $D_w(X) = d$, which bounds the failure to appear rate among detained defendants relative to the observed failure to appear rate among released defendants. I assume that $\kappa_{w,d} \equiv \kappa$ does not vary across values $W = w, D_w(X) = d$ and conduct my analysis under various assumptions on the magnitude of $\kappa \geq 0$. Comparing how the results change as $\kappa \geq 0$ varies can be interpreted as a sensitivity analysis on the informativeness of the judges' private information: how do conclusions about behavior change as we allow judges to have more accurate private information?

What Fraction of Judges Make Prediction Mistakes? Using the direct imputation bounds, I test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at strict preferences under preferences that (i) do not depend on any observable characteristics, (ii) depend on only the defendant's race, (iii) depend on both the defendant's race and age, or (iv) depend on on the defendant's race and whether the defendant was charged with a felony offense. I test the inequalities in Corollary 3.4 across deciles of predicted risk with each possible W cell. As in the main text, I include inequalities that compare the observed failure to appear rate among released defendants at predicted risk deciles six to ten against the direct imputation bounds on the failure to appear rate among detained defendants at predicted risk deciles one to five. I construct the variance-covariance matrix of the moments using the empirical bootstrap conditional on the observable characteristics W and predicted risk decile $D_w(X)$ on the cases assigned to a particular judge. Figure S5 reports the fraction of judges in the top 25 for whom we can reject expected utility maximization behavior at strict preferences under various assumption on which observable characteristics W affect the utility function. The adjusted rejection rate reports the fraction of rejections after multiple hypotheses using the Holm-Bonferroni step down procedure, which controls the family-wise error rate at the 5% level.

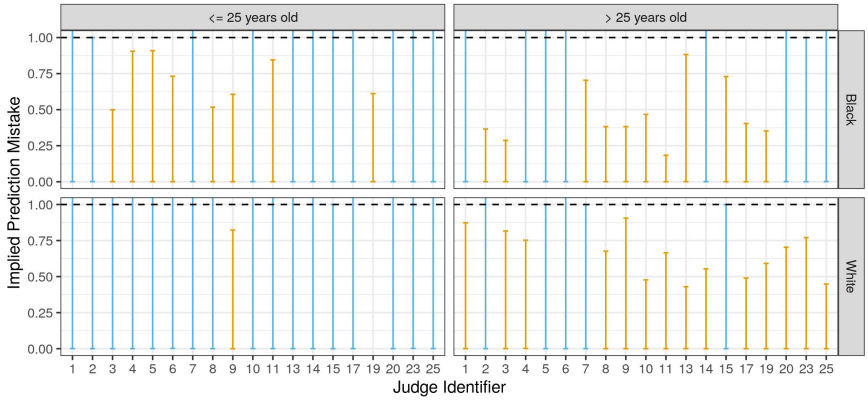
What Types of Prediction Mistakes are Being Made? I next apply the identification results in Section 4 to analyze the types of prediction mistakes based on observable characteristics made by judges in the New York City pretrial release data, constructing bounds on the unobservable failure to appear rate among detained defendants using direct imputation. Figure S6a reports 95% confidence intervals for the identified set of values $\delta(w, d)/\delta(w, d')$ between the highest d and lowest decile d' of predicted risk within each race-by-age W cell using the direct imputation bounds with $\kappa = 2$. Figure S6b plots the 95% confidence intervals for the identified set on the same object within each race-by-felony charge W cell. As in found in Section 5.5, judges appear to underreact to predictable variation in failure to appear risk. Whenever these bounds are informative, they lie strictly below one.

Figure S5: Fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about failure to appear risk using direct imputation bounds.

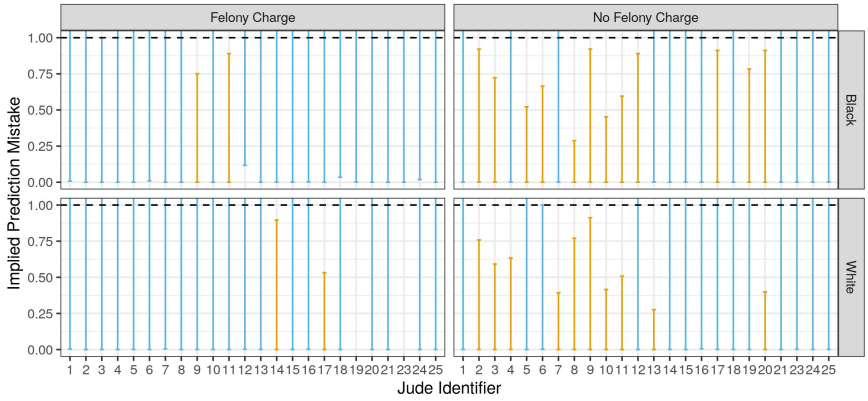


Notes: This figure summarizes the results for testing whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at strict preference utility $U(c, y^*; w)$ that (i) do not depend on any observable characteristics, (ii) depend on the defendant's race, (iii) depend on both the defendant's race and age, and (iv) depend on both the defendant's race and whether the defendant was charged with a felony offense. Bounds on the failure to appear rate among detained defendants are constructed using direct imputation (see Supplement G.1) for $\kappa = \{0, 1, \ldots, 10\}$. I first construct the unadjusted rejection rate by testing whether the pretrial release decisions of each judge in the top 25 are consistent with the moment inequalities in Corollary 3.4 at the 5% level using the conditional least-favorable hybrid test. I construct the variance-covariance matrix of the moments using the empirical bootstrap conditional on the payoff-relevant characteristics W and the predicted risk decile $D_w(X)$. The adjusted rejection rate reports the fraction of rejections after correcting for multiple hypothesis testing using the Holm-Bonferroni step down procedure, which controls the family-wise error rate at the 5% level. This is the same procedure described in Section 5.4. The adjusted rejection rate reports the fraction of rejections correcting for multiple hypotheses using the Holm-Bonferroni step-down procedure, which controls the family-wise error rate at the 5% level. Source: Rambachan and Ludwig (2021).

Figure S6: 95% confidence intervals for the implied prediction mistake of failure to appear risk between the highest and lowest predicted failure to appear risk deciles using direct imputation bounds with $\kappa = 2$.



(a) Race-by-age W cells



(b) Race-by-felony charge W cells

Notes: This figures plots the 95% confidence interval for the identified set on the implied prediction mistake $\delta(w,d)/\delta(w,d')$ between the highest predicted failure to appear risk decile d and the lowest predicted failure to appear risk decile d' within each race-by-age cell and race-by-felony charge cell. The bounds on the failure to appear rate among detained defendants are constructed using direct imputation with $\kappa = 2$ (Section G.1) and for each judge in the top 25 whose choices are inconsistent with expected utility maximization behavior at these bounds (Figure S5). These confidence intervals are constructed by first constructing a 95% joint confidence interval for a judge's reweighed utility threshold $\tau(w,d), \tau(w,d')$ using test inversion based on the moment inequalities in Theorem 4.2, and then constructing the implied prediction mistake $\delta(w,d)/\delta(w,d')$ associated with each pair $\tau(w,d), \tau(w,d')$ in the joint confidence set (Corollary 4.1). See Section 4.2 for theoretical details on the implied prediction mistake. Source: Rambachan and Ludwig (2021).

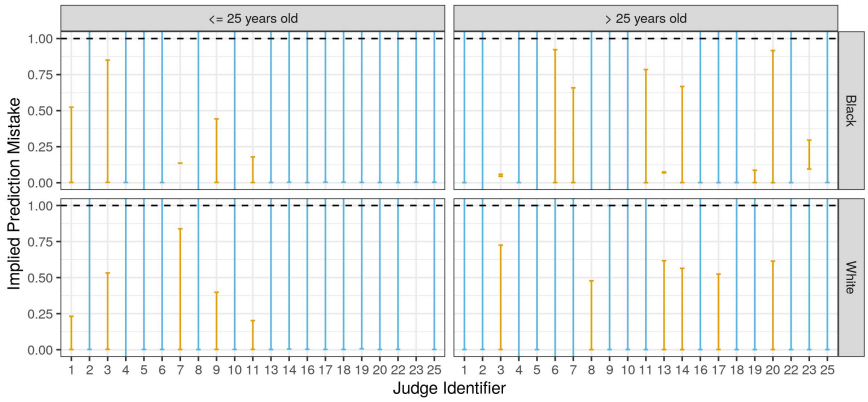
I.3 Defining the Outcome to be Any Pretrial Misconduct

Section 5 of the main text considered an extension to my baseline empirical results on detectable prediction mistakes in the New York City pretrial release system that defined the outcome of interest to be whether a defendant would commit "any pretrial misconduct" (i.e., either fail to appear in court or be re-arrested for a new crime). I now characterize the extent to which judges' predictions of any pretrial misconduct are systematically biased using the identification results in Section 4 of the main text.

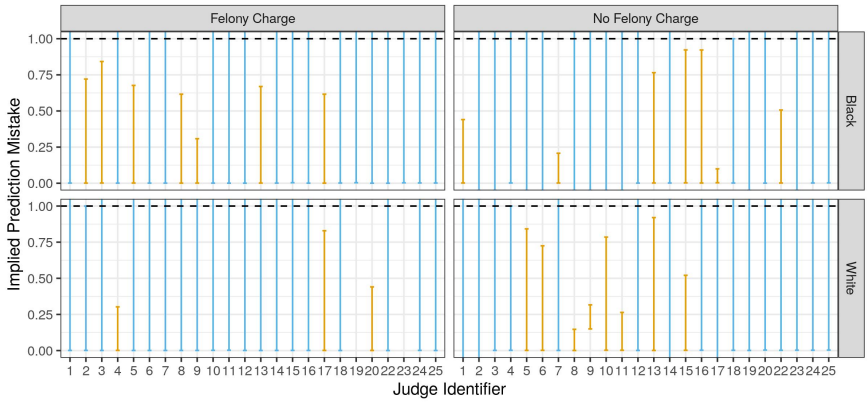
What Types of Prediction Mistakes are Being Made? Figure S7a reports 95% confidence intervals for the identified set of the implied prediction mistake $\delta(w,d)/\delta(w,d')$ between the highest d and lowest decile d' of predicted pretrial misconduct risk within each race-by-age W cell. Figure S7b plots the 95% confidence intervals for the identified set on the same object within each race-by-felony charge W cell. Judges appear to systematically underreact to predictable variation in pretrial misconduct risk between defendants at the tails of the pretrial misconduct risk distribution. Whenever these bounds are informative, they lie strictly below one.

Furthermore, among judges whose choices are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk, Table S9 reports the location of the studentized maximal violation of the revealed preference inequalities and shows the fraction of judges for whom the maximal violation occurs over the tails of the predicted distribution (deciles 1-2, 9-10) or the middle of the predicted risk distribution (deciles 3-8) for black and white defendants respectively. I again find that maximal violations of the revealed preference inequalities mainly occur over defendants that lie in the tails of the predicted risk distribution, and furthermore the majority occur over black defendants at the tails of the predicted risk distribution.

Figure S7: 95% confidence intervals for the implied prediction mistakes of any pretrial misconduct risk between the highest and lowest predicted any pretrial misconduct risk deciles



(a) Race-by-age W cells



(b) Race-by-felony charge W cells

Notes: This figures plots the 95% confidence interval for the identified set on $\delta(w,d)/\delta(w,d')$ between the highest predicted any pretrial misconduct risk decile d and the lowest predicted any pretrial misconduct risk decile d' within each race-by-age cell and race-by-felony charge cell. The outcome Y^* is whether the defendant would commit any pretrial misconduct upon release (i.e., either fail to appear in court or be re-arrested for a new crime). Bounds on the any pretrial misconduct rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3.2). Confidence intervals are constructed for each judge whose choices are inconsistent with expected utility maximization at these bounds (Table A1). These confidence intervals are constructed by first constructing a 95% joint confidence interval for a judge's reweighed utility threshold $\tau(w,d), \tau(w,d')$ using test inversion based on the moment inequalities in Theorem 4.2, and then constructing the implied prediction mistake $\delta(w,d)/\delta(w,d')$ associated with each pair $\tau(w,d), \tau(w,d')$ in the joint confidence set (Corollary 4.1). See Section 4.2 for theoretical details on the implied prediction mistake. Source: Rambachan and Ludwig (2021).

Table S9: Location of the maximum studentized violation of revealed preference inequalities among judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about any pretrial misconduct risk.

	Utility Functions $U(c, y; w)$	
	Race and Age	Race and Felony Charge
Unadjusted Rejection Rate	84%	98%
White Defendants		
Middle Deciles	0.00%	0.00%
Tail Deciles	4.76%	4.16%
Black Defendants		
Middle Deciles	9.52%	16.66%
Tail Deciles	85.71%	79.16%

Notes: This table summarizes the location of the maximum studentized violation of revealed preference inequalities among judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about pretrial misconduct risk and preferences that depend on both the defendant's race and age as well as the defendant's race and whether the defendant was charged with a felony. Bounds on the failure to appear rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3.2). Among judge's whose release decision violate the revealed preference inequalities at the 5% level, I report the fraction of judges for whom the maximal studentized violation occurs among white and black defendants at the tails of the any pretrial misconduct predicted risk distribution (deciles 1-2, 9-10) and at the middle of the any pretrial misconduct predicted risk distribution (deciles 3-8). The outcome Y^* is whether the defendant would commit any pretrial misconduct upon release (i.e., either fail to appear in court or be re-arrested for a new crime). Source: Rambachan and Ludwig (2021).

I.4 Alternative Pretrial Release Definition

In Section 5 of the main text, I tested whether the pretrial release decisions of judges in New York City were consistent with expected utility maximization behavior at accurate beliefs about failure to appear risk under various exclusion restrictions on their preferences. To do so, I collapsed the pretrial release decision into a binary choice of simply whether to release or detain the defendant.

However, in practice, judges in New York City choose what bail conditions and monetary amount to set for a defendant. Defendants may either be "released on recognizance" (i.e., automatically released without any bail conditions) or the judge may set some monetary bail, in which case the defendant is only released if they can post the set bail amount. To account for this, I now extend my baseline empirical implementation by defining a judge's choice as whether to release the defendant on recognize.

I.4.1 Identification Result for Alternative Pretrial Release Definition

To develop this extension, I first apply the identification results in Section 2 to analyze this modified decision problem. Let $C \in \{0,1\}$ denote whether the judge chose "release on recognizance" ($C = 1$). Let $W \in \mathcal{W}$ denote the directly payoff relevant defendant characteristics and $X \in \mathcal{X}$ denote the excluded defendant characteristics as before. The latent outcome is now defined as the pair (R^*, Y^*), where $R^* \in \{0,1\}$ denotes whether the defendant would satisfy the monetary bail condition set by the judge (i.e., would the defendant be able to pay the bail amount set by the judge?) and $Y^* \in \{0,1\}$ is whether the defendant would fail to appear in court if released. Let $R \in \{0,1\}$ denote the observed release decision. The observed release decision satisfies $R = C + (1-C)R^*$, meaning that the defendant is released if the judge selects release on recognizance or the judge sets monetary bail conditions and the defendant satisfies them.

I assume that the judge's utility function takes the same form as in Section 3 of the main text. The judge receives some payoff if a defendant is released that goes on to fail to appear in court or a defendant is detained that would not fail to appear in court. That is, I consider the set of utility functions satisfying $U(c, r^*, y^*; w) = U(r, y^*; w)$, $U(0,1;w) = 0, U(1,0;w) = 0$ and $U(0,0;w) < 0, U(1,1;w) < 0$.

I apply Theorem 2.1 to derive conditions under which the judge's choices are consistent with expected utility maximization behavior at accurate beliefs about both failure to appear risk and the ability of defendant's to meet the bail conditions. As in the main text, for each $w \in \mathcal{W}$, define $\mathcal{X}^1(w) := \{x \in \mathcal{X} \colon \pi_1(w,x) > 0\}$ and $\mathcal{X}^0(w) := \{x \in \mathcal{X} \colon \pi_0(w,x) > 0\}$.

Proposition I.1. *Assume $P_{Y^*}(1 \mid 1, w, x) < 1$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_1(w,x) > 0$ and $P(R = 0 \mid C = 0, W = w, X = x) > 0$ for all $(w,x) \in \mathcal{W} \times \mathcal{X}$ with $\pi_0(w,x) > 0$. The decision maker's choices are consistent with expected utility maximization behavior at some strict preference utility function if and only if for all $w \in \mathcal{W}$*

$$\max_{x \in \mathcal{X}^1(w)} P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \min_{x \in \mathcal{X}^0(w)} P(Y^* = 1 \mid R = 0, C = 0, W = w, X = x).$$

Proof. The inequalities in Theorem 2.1 imply that the judge's choices are consistent with expected utility maximization behavior at accurate beliefs if and only if

(1) for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(C = 1 \mid W = w, X = x) > 0$.

$$P(Y^* = 1 \mid C = 1, W = w, X = x) \leq \frac{-U(0, 0; w)}{-U(0, 0; w) - U(1, 1; w)}$$

(2) for all $(w, x) \in \mathcal{W} \times \mathcal{X}$ with $P(C = 0 \mid W = w, X = x) > 0$,

$$P(Y^* = 1, R = 1 \mid C = 0)U(1, 1; w) + P(Y^* = 0, R = 0 \mid C = 0)U(0, 0; w) \geq$$

$$P(Y^* = 1 \mid C = 0, W = w, X = x)U(1, 1; w).$$

The conditions in (2) may be re-arranged as

$$P(Y^* = 0, R = 0 \mid C = 0, W = w, X = x)U(0, 0; w) \geq P(Y^* = 1, R = 0 \mid C = 0, W = w, X = x)U(1, 1),$$

where $P(Y^* = 0, R = 0 \mid C = 0, W = w, X = x) = P(R = 0 \mid C = 0, W = w, X = x) - P(Y^* = 1, R = 0 \mid C = 0, W = w, X = x)$. Substituting this in and re-arranging then delivers that

$$P(Y^* = 1, R = 0 \mid C = 0, W = w, X = x) \left(-U(0, 0; w) - U(1, 1; w)\right) \geq$$

$$-P(R = 0 \mid C = 0, W = w, X = x)U(0, 0; w).$$

The result is then immediate. □

The judge's choices are inconsistent with expected utility maximization behavior at accurate beliefs about both failure to appear risk and the ability of defendant's to satisfy the monetary bail conditions if and only if the maximal failure to appear rate among defendants that were released on recognizance is less than the minimal bound on the failure to rate among defendants that could not satisfy their monetary bail conditions. The same dimension reduction techniques may be applied from the main text to reduce the number of moment inequalities that must be tested.

I.4.2 Empirical Implementation and Results

To apply this identification result, I test whether the implied revealed preference inequalities in Proposition I.1 are satisfied over the deciles of predicted failure to appear risk that were constructed in Section 5 of the main text. I use the quasi-random assignment of judges to cases to construct bounds on the unobservable failure to appear rate among detained defendants that could not satisfy their monetary bail conditions. The only modification is that I now estimate the observed failure to appear rate among defendants that were released on recognizance.

The results are summarized in Table S10 below. I find that at least 32% of judges make detectable prediction mistakes about failure to appear risk and the ability of defendants to satisfy their monetary bail conditions. This suggests that a large fraction of judges are making prediction mistakes in their joint predictions of failure to appear risk and the ability of defendants to satisfy their monetary bail conditions in their release on recognizance vs. monetary bail decisions.

Table S10: Estimated lower bound on the fraction of judges whose "release on recognizance" decisions are inconsistent with expected utility maximization behavior at accurate beliefs about behavior under bail conditions and failure to appear risk given defendant characteristics.

	Utility Functions			
	No Characteristics	Race	Race + Age	Race + Felony Charge
Adjusted Rejection Rate	32%	32%	32%	52%

Notes: This table summarizes the results of the robustness exercise to assess whether the "release on recognizance" vs monetary bail decisions of judges are consistent with expected utility maximization behavior at strict preference utility functions that either (i) do not depend on any characteristics, (ii) depend on the defendant's race, (iii) depend on both the defendant's race and age, and (iv) depend on both the defendant's race and whether the defendant was charged with a felony offense. The outcome is defined to be whether the defendant would be released under the chosen bail condition (i.e., either the judge decides to release the defendant on recognizance or the defendant satisfies the bail conditions set by the judge) and FTA if released. I first construct the unadjusted rejection rate by testing whether the pretrial release decisions of each judge in the top 25 are consistent with the moment inequalities in Corollary 3.4 at the 5% level using the conditional least-favorable hybrid test using the same procedure described in Section 5.4. The adjusted rejection rate reports the fraction of rejections after multiple hypothesis correction using the Holm-Bonferroni step down procedure. See Section I.4 for discussion. Source: Rambachan and Ludwig (2021).