Supplementary Materials to "Identifying Prediction Mistakes in Observational Data"

Ashesh Rambachan

October 3, 2022

This online supplement contains additional theoretical results and additional empirical results for the paper "Identifying Prediction Mistakes in Observational Data" by Ashesh Rambachan.

F Direct imputation 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. For exposition, I define these alternative assumptions for a screening decision with a binary outcome $Y^* = Y_1^* \in \{0, 1\}$.

Direct imputation uses the observed $P_1(1 \mid x) := P(Y^* = 1 \mid C = 1, X = x)$ to bound the unobserved $P_0(1 \mid x) := P(Y^* = 1 \mid C = 0, X = x)$.

Assumption 3. For each $x \in \mathcal{X}$ with $0 < \pi_1(x) < 1$, there exists $\kappa_x \geq 0$ satisfying

$$P_1(1 \mid x) \leq P_0(1 \mid x) \leq (1 + \kappa_x) P_1(1 \mid x).$$

The parameter $\kappa_x \geq 0$ specifies how different the unobservable choice-dependent outcome probability may be relative to the observable choice-dependent outcome probability. In pretrial release, setting $\kappa_x = 1$ means 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).

In practice, the researcher may wish to test whether the decision maker is making systematic prediction mistakes under various choices of the parameter κ_x, and thereby conduct a sensitivity analysis of how robust the behavioral conclusions are to various assumptions about the unobservable choice-dependent outcome probabilities. Supplement H illustrates such a sensitivity analysis in, reporting how the fraction of judges for whom we can reject expected utility maximization behavior varies as the parameter κ_x varies.

Finally, Assumption 3 has a natural interpretation under the expected utility maximization model. The parameter κ_x bounds the average informativeness of the decision maker's private information $V \in \mathcal{V}$.

Proposition F.1. *Consider a screening decision with a binary outcome $Y^* \in \{0, 1\}$ and suppose Assumption 3 holds. If the decision maker's choices are consistent with expected utility maximization behavior at $u \in \mathcal{U}$ and $(X, V, C, Y^*) \sim Q$, then for each $x \in \mathcal{X}$ with $0 < \pi_1(x) < 1$ and $0 < P_1(1 \mid x) < 1$*

a. $1 \leq \frac{Q(C=0 \mid Y^*=1, X=x)/Q(C=1 \mid Y^*=1, X=x)}{Q(C=0 \mid X=x)/Q(C=1 \mid X=x)} \leq 1 + \kappa_x,$

b. $1 - \kappa_x \frac{P_1(1 \mid x)}{P_1(0 \mid x)} \leq \frac{Q(C=0 \mid Y^*=0, X=x)/Q(C=1 \mid Y^*=0, X=x)}{Q(C=0 \mid X=x)/Q(C=1 \mid X=x)} \leq 1.$

Proof. Notice that

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

Since the decision maker's choices are consistent with expected utility maximization behavior, $(X, V, C, Y^*) \sim Q$ satisfies the Data Consistency condition in Definition 2 at some $\tilde{P}_0(\cdot \mid x)$ satisfying the bounds in Assumption 3 for each $x \in \mathcal{X}$. Therefore, $Q(Y^* = 1 \mid C = 0, X = x) = \tilde{P}_0(1 \mid x)$ and it immediately follows that $\frac{Q(Y^*=1|C=0,X=x)}{Q(Y^*=1|C=1,X=x)} = \frac{\tilde{P}_0(1|x)}{P_1(1|x)} \in [1, 1+\kappa_x]$ under Assumption 3. This proves (a). To show (b), notice that the bounds in Assumption 3 imply that $P_1(0 \mid x) - \kappa_{w,x} P_1(1 \mid x) \leq P_0(0 \mid x) \leq P_1(0 \mid x)$. As before,

$$\frac{Q(C=0 \mid Y^*=0, X=x)/Q(C=1 \mid Y^*=0, X=x)}{Q(C=0 \mid X=x)/Q(C=1 \mid X=x)} = \frac{\tilde{P}_0(0 \mid x)}{P_1(0 \mid x)}$$

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 bounds the average informativeness of the decision maker's private information since

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

under the Information Set condition in Definition 2. The direct imputation bounds are therefore related to 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 V that governs selection and places bounds on the magnitude of the relative odds ratio of the propensity score conditional on (V, X) versus the propensity score conditional X. Imbens (2003) develops a parametric model for such a violation of unconfoundedness in a treatment assignment problem.

G Summary tables for New York City pretrial release

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 5.1 of the main text for further discussion.

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 5.1 of the main text for further discussion.

Table S3: Balance check estimates for the quasi-random assignment of judges by defendant race and age.

	White Defendants		Black Defendants	
	Young	Older	Young	Older
	(1)	(2)	(3)	(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 of the main text for further details.

Table S4: Balance check estimates for the quasi-random assignment of judges by defendant race and felony charge.

	White Defendants		Black Defendants	
	Felony Charge	No Felony Charge	Felony Charge	No Felony Charge
	(1)	(2)	(3)	(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 5.3 of the main text for further details.

H 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.

H.1 Welfare effects of automation policies: race-by-felony charge cells

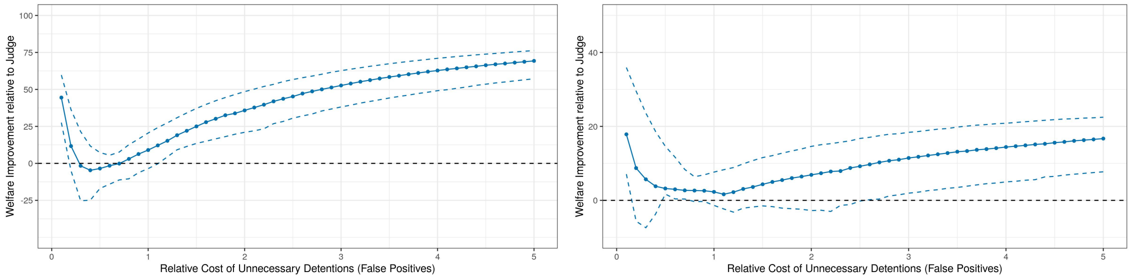
Section 6 of the main text compared the total expected social welfare under the observed release decisions by judges in new York City against the total expected social welfare under counterfactual algorithmic decisions, conducting this exercise over race-by-age cells and deciles of predicted failure 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.

Figure S1a plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces judges who were found to make systematic prediction mistakes against the 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 $|\tilde{u}| \in [0.3, 0.7]$ (where recall $u^*_{0,1} = -\tilde{u}/|1 + \tilde{u}|$), 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 S1b therefore plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that only corrects systematic 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 systematic prediction mistakes weakly dominates the observed release decisions of judges, no matter the value of the social welfare function.

I finally 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 S2 plots the improvement in worst-case total expected social welfare under the algorithmic decision rule that fully replaces these judges against their 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 range of social welfare costs of unnecessary detentions.

Figure S1: Comparison of algorithmic decision rule relative to observed release decisions of judges that make detectable prediction mistakes over race-by-felony charge cells.

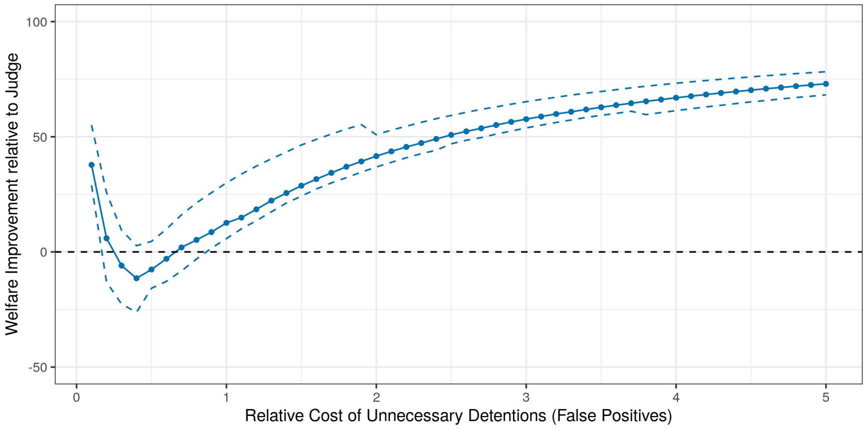


(a) Full automation decision rule

(b) 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 in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1 + \tilde{u}|$. 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 of the main text and Supplement H.1 for further details.

Figure S2: Comparison of algorithmic 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 in court $|\tilde{u}|$ (i.e., an unnecessary detention), where $u^*_{0,1} = -\tilde{u}/|1+\tilde{u}|$. 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 of the main text and Supplement H.1 for further details.

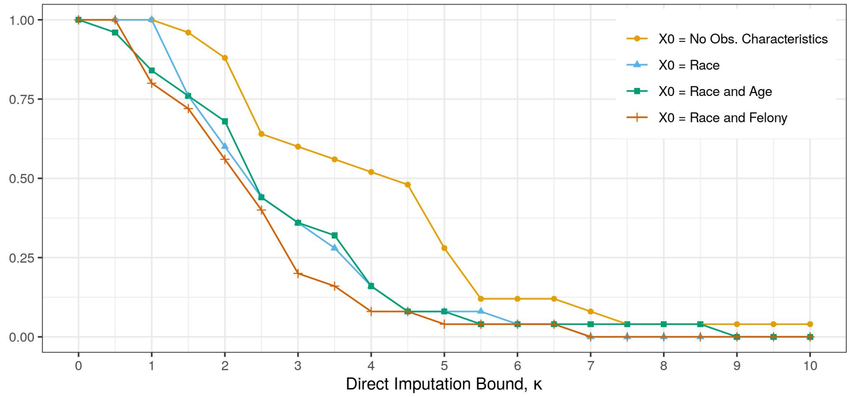
H.2 Identifying prediction mistakes using direct imputation

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

I show how the same test may be conducted using direct imputation (Supplement F) to construct bounds on the failure to appear rate of detained defendants. In this case, the parameter $\kappa_{x_0,d} \geq 0$ for $X_0 = x$, $D(X) = d$ bounds the failure to appear rate among detained defendants relative to the failure to appear rate among released defendants. I assume $\kappa_{x_0,d} \equiv \kappa$ does not vary across values $X_0 = x_0, D(X) = d$ and report results as $\kappa \geq 0$ varies. Comparing how results change as $\kappa \geq 0$ varies is a sensitivity analysis on how conclusions about behavior change as we allow judges to have more accurate private information.

What fraction of judges make prediction mistakes? I test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at some linear utility function that (i) does not depend on any observable characteristics, (ii) depends on the defendant's race, (iii) depends on both the defendant's race and age, or (iv) depends on the defendant's race and whether the defendant was charged with a felony offense. I test the inequalities in Proposition 5.1 across deciles of predicted risk with each possible X_0 cell. Figure S3 reports the fraction of judges in the top 25 for whom we can reject expected utility maximization at accurate beliefs under various assumption on which observable characteristics X_0 affect the utility function.

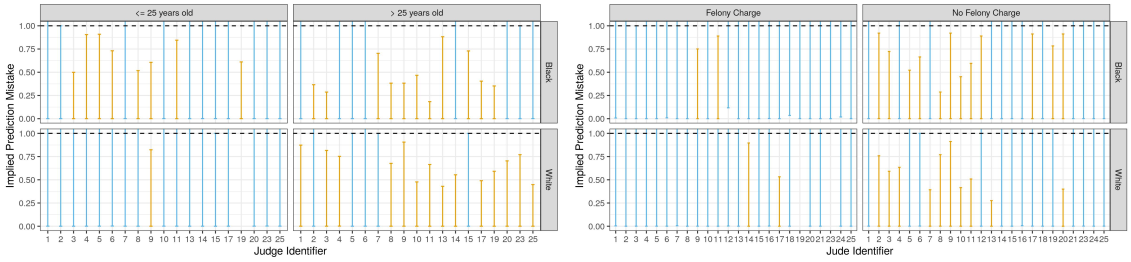
Figure S3: Fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about failure to appear risk using direct imputation bounds.



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 $u(c, y^*; x_0)$ 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 F) for $\kappa = \{0, 1, \ldots, 10\}$. The adjusted rejection rate reports the fraction of rejections after a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level.

Bounding prediction mistakes based on defendant characteristics I apply the identification results in Section 4.2 to analyze the types of prediction mistakes based on observable characteristics made by judges in the New York City pretrial release data, constructing direct imputation bounds on the failure to appear rate among detained defendants. Figure S4(a) reports 95% confidence intervals for the identified set of values $\delta(x_0, d)/\delta(x_0, d')$ 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 S4(b) plots the same results for each race-by-felony charge X_0 cell. As in the main text, judges appear to underreact to predictable variation in failure to appear risk. Whenever informative, these bounds lie strictly below one.

Figure S4: 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 X_0 cells

(b) Race-by-felony charge X_0 cells

Notes: This figures plots the 95% confidence interval for the identified set on the implied prediction mistake $\delta(x_0, d)/\delta(x_0, d')$ 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 $\kappa = 2$ (Supplement F) and for each judge in the top 25 whose choices are inconsistent with expected utility maximization behavior at these bounds. See Section 4.2 for theoretical details on the implied prediction mistake.

H.3 Defining the outcome to be any pretrial misconduct

Given the stated objectives of the NYC pretrial system, the main text initially defined the outcome $Y^* = Y_1^* \in \{0, 1\}$ to be whether a defendant would fail to appear in court. I now define the outcome of interest $Y^* = Y_1^* \in \{0, 1\}$ 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). Section **??** of the main text considered an extension to my baseline empirical results 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 report the fraction of judges that make systematic prediction mistake about pretrial misconduct risk given defendant characteristics, characterize the costs and shares of these systematic prediction mistakes using the identification results in Section 4.1, and bound the extent to which judges' predictions of any pretrial misconduct are systematically biased using the identification results in Section 4.2.

What fraction of judges make prediction mistakes? As in the main text, I test whether the release decisions of each judge in the top 25 are consistent with expected utility maximization behavior at some linear utility function that (i) does not depend on any observable characteristics, (ii) depends on the defendant's race, (iii) depends on both the defendant's race and age, or (iv) depends on the defendant's race and whether the defendant was charged with a felony offense. Table S5 shows that the pretrial release decisions of at least 64% of judges are inconsistent with expected utility maximization at accurate beliefs about pretrial misconduct risk and some linear utility function satisfying the conjectured exclusion restriction.

Table S5: Estimated fraction of judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about any pretrial misconduct risk given defendant characteristics.

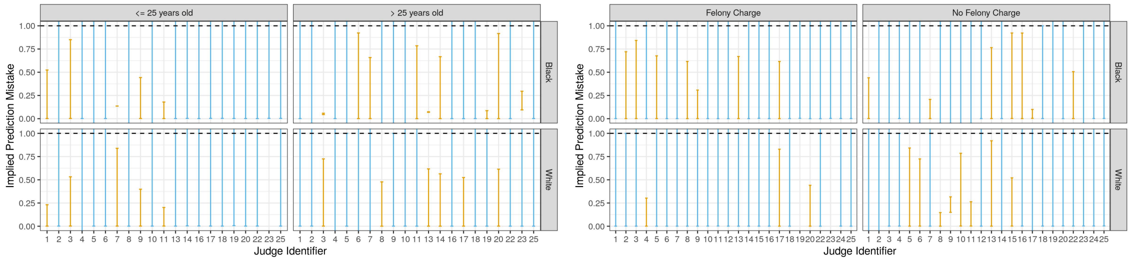
	Utility Functions $u(c, y^*; x_0)$			
	No Characteristics	Race	Race + Age	Race + Felony Charge
Adjusted Rejection Rate	76%	72%	64%	92%

Notes: This table summarizes the results for testing for misrankings in the release decisions of each judge in the top 25 at linear utility functions $u(c, y^*; x_0)$ that (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 $Y^* = Y_1^* \in \{0, 1\}$ is whether the defendant would commit any pretrial misconduct (i.e., either fail to appear in court or be re-arrested for a new crime) upon release. Bounds on the any pretrial misconduct rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3). The adjusted rejection rate reports the fraction of rejections after a multiple hypothesis testing correction that controls the family-wise error rate at the 5% level.

Bounding prediction mistakes based on defendant characteristics Figure S5a reports 95% confidence intervals for the identified set of the implied prediction mistake $\delta(x_0, d)/\delta(x_0, d')$ between the highest d and lowest decile d' of predicted pretrial misconduct risk within each race-by-age X_0 cell. Figure S5b plots the 95% confidence intervals for the identified set on the same object within each race-by-felony charge X_0 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 S6 reports the location of the largest studentized misranking and shows the fraction of judges for whom the largest misranking 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 the largest misrankings 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 S5: 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 X_0 cells

(b) Race-by-felony charge X_0 cells

Notes: This figures plots the 95% confidence interval for the identified set on $\delta(x_0, d)/\delta(x_0, d')$ 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_1^* 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). See Section 4.2 for theoretical details on the implied prediction mistake.

H.4 Alternative pretrial release definition

In Section 5 of the main text, I collapsed the pretrial release decision into a binary choice of simply whether to release or detain the defendant. 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. I extend my baseline empirical implementation by defining a judge's choice as whether to release the defendant on recognizance.

Let $C \in \{0, 1\}$ denote whether the judge "released on recognizance" ($C = 1$). The latent outcome is the pair $Y^* = (R^*, Y_1^*)$, where $R^* \in \{0, 1\}$ denotes whether the defendant would satisfy the monetary bail condition set by the judge and $Y_1^* \in \{0, 1\}$ 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 the defendant is released if the judge selects release on recognizance or sets monetary bail conditions and the defendant satisfies them. I assume the judge receives payoffs if a defendant is released and fails to appear in court or a defendant is detained and would not fail to appear in court. That is, I consider the set of

Table S6: Location of the largest misranking among judges whose release decisions are inconsistent with expected utility maximization behavior at accurate beliefs about any pretrial misconduct risk.

	Utility Functions $u(c, y; x_0)$	
	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: TThis table summarizes the location of the largest (studentized) misranking in Proposition 5.1 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. The outcome Y_1^* 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 failure to appear rate among detained defendants are constructed using the judge leniency instrument (see Section 5.3).

utility functions $\mathcal{U}$ satisfying $u(c, r^*, y_1^*; x_0) = u(r, y_1^*; x_0)$, where $u(0,1;x_0) = 0, u(1,0;x_0) = 0$, $u(0,0;x_0) < 0, u(1,1;x_0) < 0$, and $|u(0,0;x_0) + u(1,1;x_0)| = 1$.

I apply Theorem C.1 to derive conditions under which the judge's choices are consistent with expected utility maximization at accurate beliefs about both failure to appear risk and the ability of defendant's to meet the bail conditions. For each $x_0 \in \mathcal{X}_0$, define $\Pi_1(x_0) := \{x_1 \in \mathcal{X}_1 \colon \pi_1(x_0, x_1) > 0\}$ and $\Pi_0(x_0) := \{x_1 \in \mathcal{X}_1 \colon \pi_0(x_0, x_1) > 0\}$.

Proposition H.1. *Assume $P_1(1 \mid x) < 1$ for all $x \in \mathcal{X}$ with $\pi_1(x) > 0$ and $P(R = 0 \mid C = 0, X = x) > 0$ for all $x \in \mathcal{X}$ with $\pi_0(x) > 0$. The decision maker's choices are consistent with expected utility maximization behavior at some $u \in \mathcal{U}$ if and only if, for all $x_0 \in \mathcal{X}_0$,*

$$\max_{x_1 \in \Pi_1(x_0)} P(Y_1^* = 1 \mid C = 1, X = (x_0, x_1)) \leq \min_{x_1 \in \Pi_0(x_0)} P(Y_1^* = 1 \mid R = 0, C = 0, X = (x_0, x_1)).$$

Proof. The inequalities in Theorem C.1 imply that the judge's choices are consistent with expected utility maximization behavior at accurate beliefs if and only if

(1) for all $x \in \mathcal{X}$ with $\pi_1(x) > 0$, $P(Y_1^* = 1 \mid C = 1, X = x) \leq -u(0,0;w)$.

(2) for all $x \in \mathcal{X}$ with $\pi_0(x) > 0$,

$$P(Y_1^* = 1, R = 1 \mid C = 0, X = x)u(1,1;x_0) + P(Y^* = 0, R = 0 \mid C = 0, X = x)u(0,0;x_0) \geq$$
$$P(Y_1^* = 1 \mid C = 0, X = x)u(1,1;x_0).$$

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

$$P(Y_1^* = 0, R = 0 \mid C = 0, X = x)u(0, 0; x_0) \geq P(Y_1^* = 1, R = 0 \mid C = 0, X = x)u(1, 1; x_0),$$

where $P(Y_1^* = 0, R = 0 \mid C = 0, X = x) = P(R = 0 \mid C = 0, X = x) - P(Y_1^* = 1, R = 0 \mid C = 0, X = x)$. Substituting and re-arranging then delivers

$$P(Y_1^* = 1, R = 0 \mid C = 0, X = x)\left(-u(0, 0; x_0) - u(1, 1; x_0)\right) \geq$$

$$-P(R = 0 \mid C = 0, X = x)u(0, 0; x_0).$$

The result is then immediate. □

I test whether the implied revealed preference inequalities in Proposition H.1 are satisfied over the deciles of predicted failure to appear risk that were constructed in Section 5.2 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. I now estimate the observed failure to appear rate only among defendants that were released on recognizance. The results are summarized in Table S7 below. I find that at least 32% of judges make systematic prediction mistakes about failure to appear risk and the ability of defendants to satisfy their monetary bail conditions.

Table S7: 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 $u(r, y_1^*; x_0)$			
	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. See Section H.4 for discussion.