HUMAN DECISIONS AND MACHINE PREDICTIONS
Online Appendix

Jon Kleinberg
Himabindu Lakkaraju
Jure Leskovec
Jens Ludwig
Sendhil Mullainathan

August 11, 2017

Appendix

A National replication

While New York is a particularly important city given its size, it is still just one jurisdiction. Perhaps the results reported so far are due to something idiosyncratic about our dataset, or to New York's criminal justice system? We presented some results above using a small, older dataset drawn from eight jurisdictions in the 1970s (within the context of trying to rule out omitted payoffs bias) that begins to answer this question. Here we more systematically address whether the results can be replicated in other places and datasets.

We use a large dataset assembled by the US Department of Justice (DOJ) that captures data on felony defendants from 40 urban counties from across the US over the period 1990-2009. The dataset was collected as part of the state court processing statistics series, formerly called the National Pretrial Reporting Program (see USDOJ, 1990-2009). DOJ identified the 75 most populous counties in the US, sub-sampled 40 of them, and then collected data intended to be representative of all felony cases that were filed in these counties during a selected set of days during May of each study year, which included selected years between 1990 and 2009. The jurisdictions from which data were collected as part of this dataset are as follows (state followed by counties): Arizona (Maricopa, Pima); California (Los Angeles, Orange, San Bernardino, Ventura); Connecticut (Hartford); Florida (Broward, Miami-Dade, Hillsborough; Orange); Hawaii (Honolulu); Illinois (Cook); Indiana (Marion); Maryland (Baltimore, Montgomery, Prince George); Michigan (Oakland, Wayne); Missouri (Saint Louis); New Jersey (Essex, Middlesex); New York (Bronx, Kings, Nassau, New York, Suffolk); North Carolina (Wake); Ohio (Cuyahoga, Franklin, Hamilton); Tennessee (Shelby); Texas (Dallas, El Paso, Harris, Tarrant); Utah (Salt Lake City); Washington (King); and Wisconsin (Milwaukee). Unlike New York, most of these jurisdictions ask judges to focus on public safety as well as flight risk, rather than just flight risk, as part of pre-trial bond court hearings. So we train our algorithm on an outcome defined as committing either failure to appear or re-arrest.

The DOJ dataset contains a total of 151,461 observations. Descriptive statistics for this dataset are in Appendix Table A.14. The dataset includes sampling weights that account for the higher probability of selection for large jurisdictions and the fact that the dataset captured felony cases over just 1 week for the largest counties and for a full month for the smallest counties. Our results are unweighted; results with weights are fairly similar. The DOJ dataset unfortunately does not include judge identifiers.

The top panel of Appendix Table A.15 shows that just as in the New York City data, judges in this national DOJ data also release defendants with very high levels of predictable risk. Those in the top 1% of the predicted risk distribution have an observed crime rate of 76.6%, and yet are released by judges at a rate of 52.5%. The release rate for this high-risk set is not that much lower than the overall release rate of 61.6%, although the observed crime rate is much higher than the sample average. (Given the DOJ dataset's panel structure, we identify the top 1% of the predicted risk distribution within each county-year cell, and then report the average observed crime rate among that group.)

The bottom panel of Appendix Table A.15 shows the results of carrying out our policy simulation for the potential gains from letting the algorithm rank-order all defendants and then forming a simple rule that detains in descending order of predicted risk. As we did with our NYC data, we randomly partition the dataset into a training, imputation and test set and fit boosted decision trees in the training and imputation sets to rank-order defendants for release and help estimate outcomes under our algorithm's counter-factual decision rule. As we did with the policy simulation with the NYC data, for this exercise we define 'crime rate' here as the number of crimes (FTA or re-arrests) divided by the total number of defendants who pass through bond court, so that we can meaningfully compare crime rates at different potential release rates.

Compared to the judges, a release rule based on machine learning predictions would let us reduce the crime rate by 18.8% holding the release rate constant, or, holding the crime rate constant, we could reduce the jail rate by 24.5%. These gains come from constraining the algorithm to have the same release rate within each county-year cell as judges, to avoid giving the algorithm an unfair advantage in being able to make release decisions the local judges could never make - for example releasing all defendants in low-crime places or time periods, and detaining all defendants in high-crime places or times.

We also find in the national data that the 'predicted judge' outperforms the judge. Though the lack of judge identifiers prevents us from doing a contraction test, all of the NYC findings that we are able to re-estimate in the national data appear to replicate.

B Empirical evidence for conditionally random assignment of cases to judges in New York City

Several of the analyses reported in our paper take advantage of quasi-random assignment of cases to judges in the New York City dataset. This appendix provides additional details on our statistical tests for whether this assignment process is actually as-good-as random.

As noted in the data section we construct 577 borough, year, month and day of week 'cells' in the New York City data where we have at least five judges, which together account for 56.5% of our sample. The summary statistics for this sample are presented in Table A.3.

To define judge leniency, within each cell we calculate the release rate of each judge and then divide cases up into 'judge leniency quintiles' based on the leniency of the judge that hears their case.

We examine balance across these leniency quintiles in the projection of the FTA variable onto the baseline characteristics. That is, we regress the FTA rate of defendant (i) whose case is heard by judge (j) in cell (c), y_{ijc}, against the defendant's baseline characteristics, x_{ijc}, and then retain the predicted value, $\hat{y}_{ijc}$. This essentially creates an index of all the baseline defendant background and case characteristics, weighted in proportion to the strength of their relationship to the main outcome we are examining in our analysis (failure to meet pre-trial release conditions).

Then separately for each borough, year, month and day of week cell, we regress this predicted value against a set of indicators for within-cell judge leniency quintile, Q_{jc}, and calculate the F-test statistic for the null hypothesis that the judge leniency indicators are jointly zero. Call this F_c.

$$\hat{y}_{ijc} = \beta_0 + \beta_1 Q_{jc} + \epsilon_{ijc}$$

We then randomly permute the judge leniency quintiles across cases M=1,000 times within each cell. This provides us with a distribution of F-test statistics calculated under the null hypothesis of no relationship between judge leniency and defendant characteristics, F^*_{ck} for k=1,...M. Then for each cell we calculate:

$$P_c = \frac{1}{M} \sum_k 1(F^*_{ck} > F_c)$$

If defendant characteristics were systematically related to judge leniency, we would expect to see a concentration of our F-test statistics F_c with low p-values. Yet Figure III shows that the histogram of P_c values across the 577 cells in our analysis sample does not show unusual mass at low p-values. In Appendix Figure A.1 we see this is similar in each of the individual boroughs in New York, not just citywide.

While these analyses suggest balance in average defendant characteristics, in principle one might worry about the lingering possibility of imbalance at the high end of the $\hat{y}$ distribution which might be particularly relevant for our comparison of inter-judge detention differences on the margin. One way to address this concern is to identify the $\hat{y}$ threshold that corresponds to the full-sample judge release rate of 73.7%, or $P[\hat{y}_{ijc} < \hat{y}^*]] = .737$. When we construct an indicator for each observation in our sample, $D_{ijc} = 1$ if $\hat{y}_{ijc} < \hat{y}^*$, and re-do our balance test as described above, the results are similar. This remains true if we set an even higher $\hat{y}$ threshold to identify cases in the top decile of the predicted risk distribution instead.

C Data and Methods for Analyzing Detention Rates to Married or Employed Defendants

As noted in the main text, our New York City dataset does not include information about the defendant's family status or employment status, so we cannot compare the algorithm's detention rate for such defendants compared to current judge decisions in New York. (Pre-trial services in New York City does collect such information as part of their interviews of defendants prior to bond court hearings, but we do not have access to those survey results.)

To examine this question we relied instead on a national dataset assembled by Toborg (1981, 1997) that captured information on a sample of 3,488 pre-trial defendants between 1976 and 1978 drawn from eight jurisdictions. The jurisdictions are: Baltimore City, Maryland; Baltimore County, Maryland; Washington, DC; Dade County (Miami), Florida; Jefferson County (Louisville), Kentucky; Pima County (Tucson), Arizona; Santa Cruz County, California; and Santa Clara County (San Jose), California.

Since judges in most jurisdictions are asked to focus on safety as well as flight risk, our outcome is an index equal to one if the defendant was either re-arrested or FTA'd. As with the New York dataset the key predictors are current charge and prior record; the algorithm uses age but no other socio-demographic factors, such as defendant's marital or employment status.

And as with our analysis of the New York City data, we randomly divide these data into a 40% test set, a 40% imputation set, and a 20% test set. We use 5-fold cross-validation to select the optimal level of complexity for a gradient boosted decision tree algorithm. In the analyses presented in the main paper, we set observations with missing values (coded as NA in the dataset) for the married or employed variables to zero. Unfortunately the dataset's documentation does not provide much detail about how to interpret missing data. In Appendix Table A.6 we show that the results are qualitatively similar when we ignore these observations entirely.

D Additional robustness checks

In this appendix we consider other potential sources of confounding beyond selective labels and omitted payoffs bias.

D.1 Jail capacity constraints

Besides omitting preferences that drive judicial decisions, we might also have omitted a particularly important constraint that binds judge's decisions: jail capacity. This could prevent the judge (but not the algorithm) from putting high-risk people in jail during times when the local jail is filled up. A simple example illustrates the key concern: Suppose that on day one the local jail is at 100% capacity while on day two the jail is at 50% capacity. The judge wishes to release 50% each day. But in practice the judge would release 100% of defendants on day one because of jail capacity constraints, including many high-risk defendants the judge wishes they could have detained if the jail had room. If the number of cases is the same each day we would attribute to the judge an overall release rate of 75%, and so have the machine learner identify the lowest-risk 75% of cases each day and 'release' them - even though in practice this would not have been possible on day 1 in our example.

Appendix Table A.7 shows that even after accounting for this concern, we continue to see large potential social-welfare gains from releasing defendants using machine rather than judge predictions. We re-train our algorithm in the New York data, but now constraining the algorithm to have the same release rate as what we see over each three month window in New York City as a whole. The results are very similar to what we see in our main analysis in terms of either the average crime rate of the predictably-riskiest 1% that the algorithm can identify among the judge's released set, or in terms of the potential gains from the re-ranking policy simulation. The same finding holds if we constrain the algorithm to not exceed the release rate observed within each quarter-year, specific to the New York City borough in which each case is being heard.

D.2 Algorithmic (in)stability?

A different potential concern is that our analysis overstates the potential gains of the algorithm relative to the judges because the algorithm is unstable - that is, changing over time in ways that attenuate the potential gains of the algorithm relative to the judge decisions. This could lead us to over-state the potential gains in the future from adopting an algorithmic release rule, since so far we have been comparing the relative performance of the algorithm to the judges averaged over our entire sample period.

Yet as shown in Appendix Table A.7, there are no signs that our algorithm is particularly unstable. To check this, we re-train the algorithm using just data from the first few years of our sample period (2008 through 2012). Instead of evaluating the algorithm's performance on a randomly-selected set of cases drawn from the same time period over which the algorithm was trained, we now evaluate the algorithm using data from the first five months of data in 2013. (As a reminder, we have taken out the data from May through November 2013 for our final hold-out set so the first part of 2013 are the latest observations available in our working dataset. The results are similar if we use the last six months of the larger dataset from the final hold-out, as we discuss below.) The gains from relying on the algorithm's predictions rather than current judge decisions are, if anything, even stronger when we predict out over time compared to our main analyses.

D.3 Human data mining?

A final potential concern is that our algorithm performs well only because after much trial and error we have stumbled across the one model specification that dominates the judges - that is, our results are due to some form of inappropriate human data mining. As noted above, one way we guard against this is by forming a true hold-out set of $203,338$ cases that remained in a 'lock box' until this final draft of the paper.

Both quasi-contraction and re-ranking results produce similar results in this 'lock box' as in the 'preliminary' hold-out set (Appendix Table A.8). For example, released defendants in the top 1% of the FTA risk distribution fail to appear in court 55.9% of the time, very similar to the analogous 56.3% in the preliminary hold-out set. For predicted re-arrest risk, the top 1% is re-arrested at a 39.5% rate in the true hold-out set compared to 44.3% in the preliminary hold-out, while the top percentile in predicted risk for re-arrest for a violent crime specifically offends at a 25.6% rate in the true hold-out compared to 25.3%. And, in the lock-box set, re-ranking reduces crime by 24.2% at the judge's release rate, very similar to the 24.7% in the preliminary hold-out set.

The results are also similar in each hold-out subsample: a random subset of defendants; a random selection of judges, and all defendants heard by those judges; and the last six months of our sample period.

Appendix Tables

Table A.1: Linear Projection of Algorithm's FTA Risk Prediction

Fraction of ML Prediction Variance Explained by Linear Projection					
Adjusted R^2	0.455	0.493	0.468	0.470	0.514
F-Tests of Joint Significance of Categorical Controls					
Control Dummies	*p-values on F-tests*				
Arrest County	<.0001	<.0001	<.0001	<.0001	<.0001
Arrest Month	<.0001	<.0001	<.0001	<.0001	<.0001
Detailed Arrest Charge		<.0001			<.0001
Detailed Prior Arrest			<.0001		<.0001
Detailed Prior Conviction				<.0001	<.0001
Selected Coefficients					
Age	−0.003	-0.004	-0.003	-0.003	-0.003
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Current Arrest					
Number of		-0.00001			-0.00001
Arrest Charges		(0.00002)			(0.00002)
Felony		-0.046			-0.047
		(0.082)			(0.080)
Misdemeanor		-0.018			-0.021
		(0.082)			(0.080)
Violent Felony		-0.024			-0.025
		(0.001)			(0.001)
Drug Charge		0.022			0.021
		(0.0004)			(0.0004)
Firearm		-0.020			-0.018
		(0.001)			(0.001)

Continued on next page

Table A.1 – continued from previous page

Selected Coefficients					
Defendant Priors					
FTAs	0.020	0.019	0.023	0.023	0.023
	(0.0001)	(0.00005)	(0.0001)	(0.0001)	(0.0001)
Felony Arrests			-0.004		-0.001
			(0.0001)		(0.0001)
Felony Convictions				-0.009	-0.005
				(0.0002)	(0.0002)
Misdemeanor Arrests			-0.001		0.002
			(0.00004)		(0.0001)
Misdemeanor Convictions				-0.003	-0.004
				(0.00004)	(0.0001)
Violent Felony Arrests			0.002		-0.002
			(0.0002)		(0.0002)
Violent Felony Convictions				0.011	0.015
				(0.0005)	(0.001)
Drug Arrests			0.002		-0.003
			(0.0001)		(0.0001)
Drug Convictions				0.005	0.008
				(0.0001)	(0.0002)
Firearm Arrests			-0.004		-0.004
			(0.0004)		(0.0004)
Firearm Convictions				-0.013	-0.008
				(0.001)	(0.001)
Observations	221,876	221,876	221,876	221,876	221,876

Notes: This Table shows the share of variation in the machine learning algorithm's predictions that can be explained by a linear functional form. Each column reports the results from a regression that has the algorithm's predicted crime risk for each case in our NYC training set as the dependent variable, with various measures of current offense and prior criminal record as explanatory variables.

Table A.2: Linear Projection of Algorithm's Any Crime (FTA or Rearrest) Prediction

Fraction of ML Prediction Variance Explained by Linear Projection					
Adjusted R^2	0.548	0.557	0.594	0.571	0.643
F-Tests of Joint Significance of Categorical Controls					
Control Dummies			*p-values on F-tests*		
Arrest County	<.0001	<.0001	<.0001	<.0001	<.0001
Arrest Month	<.0001	<.0001	<.0001	<.0001	<.0001
Detailed Arrest Charge		<.0001			<.0001
Detailed Prior Arrest			<.0001		<.0001
Detailed Prior Conviction				<.0001	<.0001
Selected Coefficients					
Age	-0.007	-0.007	-0.008	-0.008	-0.008
	(0.00002)	(0.00002)	(0.00002)	(0.00002)	(0.00002)
Current Arrest					
Number of		0.00001			0.00003
		(0.00003)			(0.00003)
Felony		-0.025			-0.058
		(0.120)			(0.108)
Misdemeanor		-0.021			-0.054
		(0.120)			(0.108)
Violent Felony		-0.009			-0.008
		(0.001)			(0.001)
Drug Charge		0.038			0.022
		(0.001)			(0.001)
Firearm		0.004			0.001
		(0.002)			(0.001)

Continued on next page

Table A.2 – continued from previous page

	Selected Coefficients				
Defendant Priors					
FTAs	0.034	0.033	0.017	0.030	0.015
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
Felony Arrests			0.005		0.012
			(0.0001)		(0.0001)
Felony Convictions				0.010	-0.003
				(0.0003)	(0.0003)
Misdemeanor Arrests			0.004		0.016
			(0.0001)		(0.0001)
Misdemeanor Convictions				-0.002	-0.015
				(0.0001)	(0.0001)
Violent Felony Arrests			0.005		-0.004
			(0.0002)		(0.0002)
Violent Felony Convictions				0.023	0.018
				(0.001)	(0.001)
Drug Arrests			0.003		-0.003
			(0.0001)		(0.0001)
Drug Convictions				0.007	0.009
				(0.0001)	(0.0002)
Firearm Arrests			0.004		-0.001
			(0.0005)		(0.0005)
Firearm Convictions				-0.013	-0.019
				(0.001)	(0.001)
Observations	221,876	221,876	221,876	221,876	221,876

Notes: This Table shows the share of variation in the machine learning algorithm's predictions of Any Crime (= FTA or Rearrest) that can be explained by a linear functional form. Each column reports the results from a regression that has the algorithm's predicted crime risk for each case in our NYC training set as the dependent variable, with various measures of current offense and prior criminal record as explanatory variables.

Table A.3: Summary Statistics for Contraction Subsample of New York Data

	Full Subsample	Judge Releases	Judge Details	P-value
Sample Size	313,601	230,704	82,897	
Release Rate	.7357	1.000	.00	
Outcomes				
Failure to Appear (FTA)	.1512	.1512		
Arrest (NCA)	.2587	.2587		
Violent Crime (NVCA)	.0370	.0370		
Murder, Rape, Robbery (NMRR)	.0187	.0187		
Defendant Characteristics				
Age	31.97	31.30	33.83	<.0001
Male	.8319	.8087	.8967	<.0001
White	.1207	.1336	.0849	<.0001
African American	.4913	.4603	.5775	<.0001
Hispanic	.3352	.3413	.3180	<.0001
Arrest County				
Brooklyn	.2844	.2823	.2900	<.0001
Bronx	.2279	.2237	.2395	<.0001
Manhattan	.2570	.2461	.2875	<.0001
Queens	.2013	.2161	.1600	<.0001
Staten Island	.0295	.0318	.0230	<.0001

Continued on next page

Table A.3 – continued from previous page

	Full Subsample	Judge Releases	Judge Detains	P-value
Arrest Charge				
Violent Crime				
Violent Felony	.1493	.1208	.2287	<.0001
Murder, Rape, Robbery	.0592	.0401	.1123	<.0001
Aggravated Assault	.0858	.0872	.0818	<.0001
Simple Assault	.2133	.2425	.1321	<.0001
Property Crime				
Burglary	.0205	.0127	.0421	<.0001
Larceny	.0724	.0644	.0947	<.0001
MV Theft	.0065	.0058	.0085	<.0001
Arson	.0006	.0003	.0014	<.0001
Fraud	.0701	.0766	.0518	<.0001
Other Crime				
Weapons	.0518	.0502	.0562	<.0001
Sex Offenses	.0090	.0088	.0097	.0155
Prostitution	.0139	.0161	.0079	<.0001
DUI	.0474	.0615	.0081	<.0001
Other	.1385	.1446	.1215	<.0001
Gun Charge	.0337	.0214	.0678	<.0001
Drug Crime				
Drug Felony	.1427	.1189	.2088	<.0001
Drug Misdemeanor	.1125	.1138	.1087	.0001
Defendant Priors				
FTAs	2.0928	1.3007	4.2971	<.0001
Felony Arrests	3.182	2.118	6.141	<.0001
Felony Convictions	.6206	.3906	1.261	<.0001
Misdemeanor Arrests	5.127	3.34	10.10	<.0001
Misdemeanor Convictions	3.128	1.555	7.506	<.0001
Violent Felony Arrests	1.018	.7069	1.883	<.0001
Violent Felony Convictions	.1536	.1014	.2987	<.0001
Drug Arrests	3.205	2.142	6.162	<.0001
Felony Drug Convictions	.2759	.1785	.5468	<.0001
Misdemeanor Drug Convictions	1.048	.5392	2.465	<.0001
Gun Arrests	.2200	.1685	.3635	<.0001
Gun Convictions	.0467	.0365	.0750	.0001

Notes: This table shows descriptive statistics overall and by judge release decision for the subsample of cases from our full New York City dataset that we use in our "contraction" exercise (such as in Figure V). The probability values at right are for pair-wise comparison of the equivalence of the mean values for the released versus detained defendants.

Table A.4: Summary Statistics by Leniency Quintile in Contraction Subsample of New York Data

	Quintile				
	First	Second	Third	Fourth	Fifth
Sample Size	44700	55021	63145	81277	69458
Release Rate	.62	.70	.73	.75	.83
Outcomes					
Failure to Appear	.09	.10	.11	.11	.13
Arrest	.15	.18	.19	.19	.23
Violent Crime	.02	.03	.03	.03	.03
Murder, Rape, Robbery	.01	.01	.01	.01	.02
Defendant Characteristics					
Age	32.27	32.04	31.91	31.91	31.86
Male	.84	.83	.83	.83	.83
White	.11	.11	.11	.14	.12
African American	.50	.49	.50	.49	.48
Hispanic	.34	.35	.34	.32	.35
Arrest County					
Brooklyn	.29	.27	.29	.30	.28
Bronx	.22	.28	.23	.18	.25
Manhattan	.25	.23	.26	.28	.25
Queens	.21	.22	.21	.17	.21
Staten Island	.02	.01	.01	.07	.02

Continued on next page

Table A.4 – continued from previous page

	Quintile				
	First	Second	Third	Fourth	Fifth
Arrest Charge					
Felony	.47	.46	.46	.46	.44
Misdemeanor	.53	.54	.54	.54	.56
VFO	.16	.15	.15	.15	.14
Drug	.27	.26	.25	.25	.25
Drug Felony	.15	.15	.14	.14	.14
Drug Misdemeanor	.12	.11	.11	.11	.12
Gun Charge	.04	.03	.03	.03	.03
Murder, Rape, Robbery	.07	.06	.06	.06	.05
Aggravated Assault	.08	.08	.09	.09	.09
Burglary	.02	.02	.02	.02	.02
Larceny	.08	.07	.07	.08	.07
MV Theft	.01	.01	.01	.01	.01
Arson	.00	.00	.00	.00	.00
Drug Sale	.08	.08	.08	.07	.07
Drug Possession	.14	.14	.13	.13	.13
Weapons	.05	.05	.05	.05	.05
Sex Offenses	.01	.01	.01	.01	.01
Prostitution	.01	.01	.01	.01	.01
Fraud	.07	.07	.07	.07	.07
Simple Assault	.20	.21	.21	.22	.22
DUI	.04	.05	.05	.05	.05
Other	.13	.14	.14	.14	.14

Continued on next page

Table A.4 – continued from previous page

	Quintile				
	First	Second	Third	Fourth	Fifth
Defendant Priors					
FTAs	2.35	2.19	2.08	2.04	1.92
Felony Arrests	3.57	3.30	3.15	3.11	2.95
Felony Convictions	.71	.64	.61	.60	.57
Misdemeanor Arrests	5.76	5.38	5.07	5.01	4.72
Misdemeanor Convictions	3.69	3.31	3.10	3.04	2.75
Violent Felony Arrests	1.13	1.05	1.00	1.00	.95
Violent Felony Convictions	.17	.16	.15	.15	.14
Drug Arrests	3.58	3.34	3.16	3.13	2.99
Felony Drug Convictions	.31	.29	.27	.27	.25
Misdemeanor Drug Convictions	1.23	1.10	1.03	1.02	.95
Gun Arrests	.25	.22	.21	.22	.21
Gun Convictions	.05	.05	.05	.05	.04

Notes: This table shows descriptive statistics by judge quintiles cases that serve as our New York City analysis of contraction as in Figure V.

Table A.5: Omitted Payoff Bias–Employment and Marital Status
Results from Re-Ranking in 1970s National Dataset

Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population Married	Employed	Either
Distribution of Defendants (Base Rate)			.23	.48	.56
Judge	.22	.00	.19	.28	.39
	(.02)		(.04)	(.04)	(.05)
Algorithm					
Usual Ranking	.20	-9.76	.22	.42	.50
	(.02)		(.04)	(.05)	(.05)
Match Judge on Married	.20	-9.44	.19	.42	.49
	(.02)		(.04)	(.05)	(.05)
Match Judge on Employed	.20	-9.11	.18	.28	.36
	(.02)		(.04)	(.04)	(.05)
Match Base Rate on Married	.20	-9.76	.22	.42	.50
	(.02)		(.04)	(.05)	(.05)
Match Base Rate on Employed	.20	-9.76	.22	.42	.50
	(.02)		(.04)	(.05)	(.05)

Notes: This Table reports results from an older, smaller national dataset of 685 observations that contains information on defendant's marital and employment status. It shows the potential gains of the algorithmic release rule relative to the judge at the judge's release rate with respect to crime reductions and share of the jail population that is employed, married or both. The first row shows the share of the defendant population overall that is employed, married or both. The second row shows the results of the observed judge decisions. The third row shows the results of the usual algorithmic re-ranking release rule, which does not use employment or marital status in predicting defendant risk and makes no post-prediction adjustments to account for either. In the fourth and fifth rows we adjust the algorithm's ranking of defendants for detention to ensure that the share of the jail population that is married or employed are no higher than those under current judge decisions. The sixth and seventh rows constraint the algorithmic release rule's jail population to have no higher share of married or employed than that of the general defendant pool.

Table A.6: Omitted Payoff Bias–Employment and Marital Status
Results from Re-Ranking in 1970s National Dataset
Dropping Data with Missing Observations

Release Rule	Crime Rate	Drop Relative to Judge	Percentage of Jail Population Married	Employed	Either
Distribution of Defendants (Base Rate)			.34	.52	.63
Judge	.23	.00	.28	.33	.46
	(.02)		(.06)	(.06)	(.06)
Algorithm					
Usual Ranking	.20	-11.33	.31	.43	.54
	(.02)		(.06)	(.06)	(.06)
Match Judge on Married	.20	-1.55	.28	.43	.52
	(.02)		(.06)	(.06)	(.06)
Match Judge on Employed	.20	-1.98	.28	.33	.44
	(.02)		(.06)	(.06)	(.06)
Match Base Rate on Married	.20	-11.33	.31	.43	.54
	(.02)		(.06)	(.06)	(.06)
Match Base Rate on Employed	.20	-11.33	.31	.43	.54
	(.02)		(.06)	(.06)	(.06)

Notes: Table reports results from an older, smaller national dataset that contains information on defendant's marital and employment status. In this Table we drop all data where either of these variables is coded as "NA". It shows the potential gains of the algorithmic release rule relative to the judge at the judge's release rate with respect to crime reductions and share of the jail population that is employed, married or both. The first row shows the share of the defendant population overall that is employed, married or both. The second row shows the results of the observed judge decisions. The third row shows the results of the usual algorithmic re-ranking release rule, which does not use employment or marital status in predicting defendant risk and makes no post-prediction adjustments to account for either. In the fourth and fifth rows we adjust the algorithm's ranking of defendants for detention to ensure that the share of the jail population that is married or employed are no higher than those under current judge decisions. The fifth and sixth rows constraint the algorithmic release rule's jail population to have no higher share of married or employed than that of the general defendant pool.

Table A.7: Robustness Checks for NYC Quasi-Contraction and Re-Ranking Results

	Predictably	*To Achieve Judge's*	
	Riskiest 1%	*Crime Rate*	*Release Rate*
	Crime Rate	Release Rate	Crime Rate
Full Data	.5636	.1531	.0854
	(.0149)	(.0011)	(.0008)
By Borough	.5361	.1543	.086
	(.015)	(.0011)	(.0008)
By Quarter-Year	.556	.1539	.0854
	(.0149)	(.0011)	(.0008)
By Borough-Quarter-Year	.5533	.1558	.0861
	(.0149)	(.0011)	(.0008)
Train on 2008 - 2012	.6274	.1335	.0851
Test on 2013	(.0237)	(.0017)	(.0014)

Notes: Table reports various robustness checks for the re-ranking policy simulation. Each row represents a different robustness check. For each robustness check, in the first column, the predictably riskiest 1% is displayed. The second column displays the release rate needed to achieve the judge's crime rate. The third column displays the crime rate achieved at the judge's release rate. The first row shows the full sample results from the rest of the paper. The second row constrains the release rule to release the same fraction as judges do in each borough. The third row constrains the release rule to release the same fraction as judges do in each quarter-year. The fourth row constrains the release rule to release the same fraction as judges do in each borough-quarter-year cell. The final row trains the algorithm using 2008-2012 data and evaluates its on 2013 data.

Table A.8: Results on Lock Box Hold Out Subset of the New York City Data

Panel A: Outcomes for the 1% Predicted Riskiest

		Outcome Algorithm Evaluated On				
		Failure to Appear	Any Other Crime	Violent Crime	Murder Rape and Robbery	All Crimes
	Base Rate	.1471	.2513	.0360	.0176	.3200
Outcome Algorithm Trained On	Failure to Appear	.5590 (.0128)	.5797 (.0127)	.0734 (.0067)	.0507 (.0057)	.7292 (.0115)
	Any Other Crime	.3949 (.0126)	.6744 (.0121)	.1154 (.0083)	.0774 (.0069)	.7452 (.0113)
	Violent Crime	.2575 (.0113)	.5991 (.0127)	.2235 (.0108)	.1534 (.0093)	.6578 (.0123)
	Murder, Rape and Robbery	.2555 (.0113)	.5964 (.0127)	.2155 (.0106)	.1521 (.0093)	.6511 (.0123)
	All Crimes	.4837 (.0129)	.6965 (.0119)	.1241 (.0085)	.0841 (.0072)	.7859 (.0106)

Continued on next page

Table A.8 – continued from previous page

Panel B: Effect of Re-Ranking on Other Outcomes

		Outcome Algorithm Evaluated On				
		Failure to Appear	Any Other Crime	Violent Crime	Murder Rape and Robbery	All Crimes
	Base Rate	.1084	.1852	.0265	.0130	.2358
Outcome Algorithm Trained On	Failure to Appear	.0822	.1657	.0227	.0113	.2087
		(.0006)	(.0008)	(.0003)	(.0002)	(.0009)
	Percentage Gain	-24.21%	-1.562%	-14.31%	-13.66%	-11.47%
	Any Other Crime	.0928	.1534	.0182	.0076	.2036
		(.0006)	(.0008)	(.0003)	(.0002)	(.0009)
	Percentage Gain	-14.37%	-17.19%	-31.41%	-41.86%	-13.65%
	Violent Crime	.1066	.1685	.0151	.0054	.2208
		(.0007)	(.0008)	(.0003)	(.0002)	(.0009)
	Percentage Gain	-1.660%	-9.045%	-43.16%	-58.40%	-6.34%
	Murder, Rape and Robbery	.1057	.1668	.0155	.0055	.2194
		(.0007)	(.0008)	(.0003)	(.0002)	(.0009)
	Percentage Gain	-2.54%	-9.981%	-41.76%	-57.90%	-6.964 %
	All Crimes	.0881	.1542	.0190	.0082	.2023
		(.0006)	(.0008)	(.0003)	(.0002)	(.0009)
	Percentage Gain	-18.74%	-16.78%	-28.25%	-36.80%	-14.21%

Notes: This Table replicates the results in Table V on the lockbox hold-out set, that was untouched until the editorial process at a journal. The top panel reports the observed crime rate for the riskiest 1% of defendants by the algorithm's predicted risk, for different measures of crime using algorithms trained on different crime measures. The first row shows base rates for each type of crime across the columns. In the second row we train the algorithm on failure to appear (FTA) and show for the 1% defendants with highest predicted risk who are observed to commit each different form of crime across the columns. The remaining rows show the results for the top 1% predicted riskiest for an algorithm trained on different forms of crime. The bottom panel shows the potential gains of the algorithmic re-ranking release rule versus the judges (at the judges observed release rate) for each measure of crime shown across the rows, for an algorithm trained on each measure of crime shown in each row.

Table A.9: Disposition of Cases by Risk Decile in New York Data

Risk Decile	N	Crime Risk	Bail Amount Mean	Bail Amount Median	Release Rate	Fraction of Defendants Who Were: ROR	Bail	Remanded
1	11094	.0519	16744.24	2500	.8949	.7964	.1963	.0073
		(.0021)	(104.0)		(.0029)	(.0038)	(.0038)	(.0008)
2	11094	.0701	16117.13	2500	.8838	.7825	.2089	.0086
		(.0024)	(101.4)		(.0030)	(.0039)	(.0039)	(.0009)
3	11094	.0867	14742.57	2500	.8399	.7303	.2582	.0114
		(.0027)	(90.9)		(.0035)	(.0042)	(.0042)	(.0010)
4	11096	.1043	12492.68	2500	.8172	.7083	.2818	.0099
		(.0029)	(87.1)		(.0037)	(.0043)	(.0043)	(.0009)
5	11091	.1229	11759.95	2500	.7925	.6844	.3012	.0143
		(.0031)	(84.5)		(.0039)	(.0044)	(.0044)	(.0011)
6	11094	.1467	10126.39	2500	.7530	.6451	.3435	.0114
		(.0034)	(80.0)		(.0041)	(.0045)	(.0045)	(.0010)
7	11093	.1798	8546.30	2500	.6726	.5608	.4214	.0178
		(.0036)	(72.3)		(.0045)	(.0047)	(.0047)	(.0013)
8	11094	.2241	6852.95	2000	.6108	.4966	.4853	.0181
		(.0040)	(67.3)		(.0046)	(.0047)	(.0047)	(.0013)
9	11094	.2874	5189.83	1500	.5693	.4650	.5208	.0142
		(.0043)	(65.9)		(.0047)	(.0047)	(.0047)	(.0011)
10	11094	.4316	2657.39	1000	.5326	.4569	.5304	.0127
		(.0047)	(65.5)		(.0047)	(.0047)	(.0047)	(.0011)

Notes: Table reports, by predicted risk decile, how defendants cases are disposed of.

Table A.10: ROR and Bail Distribution by Judge Leniency Quintile
in Contraction Subset of New York Data

Leniency Quintile	Release Rate	Percentage who get ROR	Remanded	Bail	Bail Distribution (When Set) By Percentile: 10	25	50	75	90
5	.8285	.7264	.0087	.2650	500	1000	1500	5000	15000
	(.0014)	(.0017)	(.0004)	(.0017)					
4	.7509	.6451	.0137	.3411	500	1000	2000	5000	15000
	(.0015)	(.0017)	(.0004)	(.0017)					
3	.7298	.6294	.0122	.3584	500	1000	2000	5000	15000
	(.0018)	(.0019)	(.0004)	(.0019)					
2	.6978	.5973	.0133	.3894	500	1000	2000	5000	15000
	(.0020)	(.0021)	(.0005)	(.0021)					
1	.6185	.5097	.0169	.4734	500	1000	2000	5000	15000
	(.0023)	(.0024)	(.0006)	(.0024)					

Notes: Table reports, by leniency quintile, the fraction of cases that are ROR, remanded and given bail. For those cases given bail the distribution of bail amount by percentile is shown.

Table A.11: Cases where Judge Predictions Differ From Crime Predictions

		Sample		Test col.
	Full	Low $\hat{J}(x)$	High $\hat{J}(x)$	(2)=(3)
		Low $m(x)$	High $m(x)$	*p-value*
	(1)	(2)	(3)	(4)
Sample Size	110938	1578	743	
Release Rate	0.74	0.40	0.97	
Outcomes				
Failure to Appear	0.15	0.08	0.28	
Arrest	0.26	0.25	0.35	
Violent Crime	0.04	0.05	0.05	
Murder, Rape, Robbery	0.02	0.01	0.03	
Defendant Characteristics				
Age	31.97	36.05	19.07	0.00
Male	0.83	0.94	0.60	0.00
White	0.13	0.08	0.10	0.08
African American	0.49	0.56	0.59	0.16
Hispanic	0.33	0.34	0.27	0.00
Arrest County				
Brooklyn	0.29	0.23	0.41	0.00
Bronx	0.22	0.27	0.04	0.00
Manhattan	0.25	0.19	0.47	0.00
Queens	0.19	0.26	0.07	0.00
Staten Island	0.04	0.05	0.01	0.00

Continued on next page

Table A.11 – continued from previous page

		Sample		
	Full	Low $\hat{J}(x)$ Low $m(x)$	High $\hat{J}(x)$ High $m(x)$	Pr((2)=(3)) *p-value*
	(1)	(2)	(3)	(4)
Arrest Charge				
Felony	0.45	0.96	0.00	0.00
Misdemeanor	0.55	0.04	1.00	0.00
VFO	0.15	0.73	0.00	0.00
Drug	0.25	0.23	0.10	0.00
Drug Felony	0.14	0.23	0.00	0.00
Drug Misdemeanor	0.11	0.00	0.10	0.00
Gun Charge	0.03	0.33	0.00	0.00
Murder, Rape, Robbery	0.06	0.32	0.00	0.00
Aggravated Assault	0.09	0.12	0.00	0.00
Burglary	0.02	0.04	0.00	0.00
Larceny	0.07	0.02	0.29	0.00
MV Theft	0.01	0.01	0.00	0.00
Arson	0.00	0.01	0.00	0.00
Drug Sale	0.08	0.08	0.01	0.00
Drug Possession	0.13	0.11	0.05	0.00
Weapons	0.05	0.11	0.01	0.00
Sex Offenses	0.01	0.02	0.00	0.00
Prostitution	0.01	0.00	0.23	0.00
Fraud	0.07	0.01	0.30	0.00
Simple Assault	0.21	0.04	0.00	0.00
DUI	0.05	0.00	0.03	0.00
Other	0.14	0.10	0.08	0.10

Continued on next page

Table A.11 – continued from previous page

	Full	Sample Low $\hat{J}(x)$ Low $m(x)$	High $\hat{J}(x)$ High $m(x)$	Pr((2)=(3)) *p-value*
	(1)	(2)	(3)	(4)
Defendant Priors				
FTAs	2.09	1.09	0.10	0.00
Felony Arrests	3.16	5.02	0.16	0.00
Felony Convictions	0.61	1.48	0.00	0.00
Misdemeanor Arrests	5.11	4.96	0.74	0.00
Misdemeanor Convictions	3.12	3.30	0.01	0.00
Violent Felony Arrests	1.01	1.89	0.09	0.00
Violent Felony Convictions	0.15	0.44	0.00	0.00
Drug Arrests	3.18	3.22	0.38	0.00
Felony Drug Convictions	0.27	0.54	0.00	0.00
Misdemeanor Drug Convictions	1.04	0.78	0.00	0.00
Gun Arrests	0.22	0.57	0.02	0.00
Gun Convictions	0.05	0.16	0.01	0.00

Notes: This table shows descriptive statistics for cases for the defendants with low $\rho^{\hat{J}}$ and highρ^{m}, unlikely to be released by judge but released by the algorithm (column 3) and high $\rho^{\hat{J}}$ and low ρ^{m} unlikely to be released by judge but released by the algorithm (column 4).

Table A.12: Decomposing Judicial Error

		Release Rule				
		Algorithm ($m(x)$)		$E[m(x)\|\hat{J(x)}]$	Predicted Judge ($\hat{J}$)	Judge
		Overlap	% Total Gain	Overlap	Overlap	Overlap
Release Rule	Algorithm ($m(x)$)	100%	100%	76.86%	74.85%	70.34%
	$E[m(x)\|\hat{J(x)}]$	76.86%	36.39%	100%	85.9%	76.42%
	Predicted Judge ($\hat{J}$)	74.85%	25.86%	85.9%	100%	80.2%
	Judge	70.34%	0%	76.42%	80.2%	100%

Table A.13: Comparing Judge Release to Predicted Judge Release Decisions

	Released by			
	Judge	$\hat{J}$	Both	Judge & not $\hat{J}$
Sample Size	81725	81725	70744	10981
Outcomes				
Failure to Appear	0.15	0.14	0.14	0.25
Arrest	0.26	0.23	0.23	0.44
Violent Crime	0.04	0.03	0.03	0.06
Murder, Rape, Robbery	0.02	0.02	0.02	0.03
Defendant Characteristics				
Age	31.31	30.78	30.74	34.94
Male	0.81	0.81	0.80	0.86
White	0.14	0.15	0.15	0.09
African American	0.46	0.45	0.44	0.58
Hispanic	0.34	0.34	0.34	0.32
Arrest County				
Brooklyn	0.29	0.28	0.28	0.31
Bronx	0.22	0.22	0.21	0.26
Manhattan	0.24	0.24	0.24	0.26
Queens	0.21	0.21	0.22	0.14
Staten Island	0.05	0.05	0.05	0.03

Continued on next page

Table A.13 – continued from previous page

	Released by			
	Judge	$\hat{J}$	Both	Judge & not $\hat{J}$
Arrest Charge				
Felony	0.39	0.37	0.36	0.65
Misdemeanor	0.61	0.63	0.64	0.35
VFO	0.12	0.11	0.10	0.23
Drug	0.23	0.22	0.21	0.36
Drug Felony	0.12	0.10	0.10	0.25
Drug Misdemeanor	0.11	0.12	0.12	0.11
Gun Charge	0.02	0.02	0.01	0.07
Murder, Rape, Robbery	0.04	0.03	0.03	0.11
Aggravated Assault	0.09	0.09	0.09	0.08
Burglary	0.01	0.01	0.01	0.04
Larceny	0.07	0.06	0.06	0.10
MV Theft	0.01	0.01	0.01	0.01
Arson	0.00	0.00	0.00	0.00
Drug Sale	0.06	0.05	0.05	0.15
Drug Possession	0.13	0.13	0.12	0.16
Weapons	0.05	0.05	0.05	0.05
Sex Offenses	0.01	0.01	0.01	0.01
Prostitution	0.02	0.02	0.02	0.01
Fraud	0.08	0.08	0.08	0.05
Simple Assault	0.24	0.26	0.26	0.11
DUI	0.06	0.06	0.07	0.00
Other	0.15	0.15	0.15	0.12

Continued on next page

Table A.13 – continued from previous page

	Released by			
	Judge	$\hat{J}$	Both	Judge & not $\hat{J}$
Defendant Priors				
FTAs	1.30	0.85	0.73	4.98
Felony Arrests	2.10	1.52	1.35	6.98
Felony Convictions	0.38	0.25	0.22	1.45
Misdemeanor Arrests	3.33	2.37	2.15	10.93
Misdemeanor Convictions	1.55	0.66	0.57	7.90
Violent Felony Arrests	0.71	0.54	0.48	2.13
Violent Felony Convictions	0.10	0.07	0.06	0.34
Drug Arrests	2.12	1.54	1.39	6.82
Felony Drug Convictions	0.17	0.12	0.10	0.65
Misdemeanor Drug Convictions	0.53	0.24	0.20	2.66
Gun Arrests	0.17	0.14	0.13	0.41
Gun Convictions	0.04	0.03	0.03	0.08

Notes: This table shows descriptive statistics for cases released by the predicted judge $\rho^{\hat{j}}$ vs those actually released by judges ρ^{j}.

Continued on next page

Table A.14 – continued from previous page

Full Sample Judge Releases Judge Details P-value

Table A.14: Summary Statistics for National Sample

	Full Sample	Judge Releases	Judge Detains	P-value
Sample Size	140,538	85,189	55,349	
Release Rate	.6062	1.0000	.0000	
Outcomes				
Failure to Appear (FTA)	.2065	.2065		
Arrest (NCA)	.1601	.1601		
Violent Crime (NVCA)	.0179	.0179		
Murder, Rape, Robbery (NMRR)	.0094	.0155	.0000	
Defendant Characteristics				
Age	30.3691	30.1134	30.7626	<.0001
Male	.8300	.7948	.8842	<.0001
White	.2457	.2702	.2080	<.0001
African American	.3832	.3767	.3933	<.0001
Hispanic	.2188	.1927	.2590	<.0001
Arrest Charge				
Violent Crime				
Violent Felony	.2488	.2215	.2909	
Murder, Rape, Robbery	.0908	.0615	.1359	<.0001
Aggravated Assault	.1191	.1211	.1160	.0039
Other Violent	.0389	.0389	.0390	.9165
Property Crime				
Burglary	.0866	.0682	.1150	<.0001
Larceny	.0934	.1033	.0780	<.0001
MV Theft	.0306	.0237	.0412	<.0001
Fraud	.0288	.0374	.0154	<.0001
Other Crime				
Weapons	.0312	.0328	.0287	<.0001
DUI	.0296	.0352	.0210	<.0001
Gun Charge	.0000	.0000	.0000	
Other	.0344	.0343	.0345	.8020
Drug Crime				
Drug	.3471	.3635	.3219	
Drug Felony	.3471	.3635	.3219	<.0001
Drug Sale	.1702	.1740	.1643	<.0001
Drug Possession	.1769	.1895	.1576	<.0001
Drug Misdeamenor	.0000	.0000	.0000	

Continued on next page

Table A.14 – continued from previous page

	Full Sample	Judge Releases	Judge Detains	P-value
Defendant Priors				
FTAs	.3125	.2586	.3955	<.0001
Felony Arrests	2.9355	2.2686	3.9618	<.0001
Felony Convictions	1.1862	.8170	1.7544	<.0001
Misdemeanor Arrests	3.1286	2.6030	3.9375	<.0001
Misdemeanor Convictions	1.6808	1.2591	2.3298	<.0001
Violent Felony Convictions	.1046	.0735	.1526	<.0001

Notes: This table shows descriptive statistics overall and by judge release decision for the 140,538 cases that serve as our national analysis dataset. The probability values at right are for pair-wise comparison of the equivalence of the mean values for the released versus detained defendants.

Table A.15: Replicating Main Findings in DOJ National Data Set

Panel A: Quasi-Contraction in Released Defendants		
	Crime Rate	Release Rate
Riskiest 1%	.7655	.5252
	(.0196)	(.0424)
Average	.3062	.6155
	(.0039)	(.0041)

Panel B: Re-Ranking All Defendants by Risk		
	Crime Rate	Release Rate
Judge	.1885	.6155
	(.0033)	(.0041)
ML at Judge Release Rate	.1531	
	(.0031)	
ML at Judge Crime		.7097
		(.0039)

Notes: The table above presents results from applying the predictions of a machine learning algorithm to a national dataset of 151,461 felony defendants assembled by the US Department of Justice (DOJ) for 40 urban counties over a 20 year period. The first panel reports crime rates (crimes per released defendant) by predicted risk among the set of defendants released by the judge. In the first row we identify the top 1% of the predicted risk distribution within each county-year cell and then report the observed rate of crime, which is defined here as either failing to appear in court or re-arrest for a new offense. The second row shows the average crime rate within the set of released defendants. In the bottom panel we report the results of our policy simulation of re-ranking, where we construct a release rule that detains people in descending order of predicted crime risk. We first present the results of the judge's decision, the release rate, and the crime rate (defined now as crimes per defendant who passed through bond court), followed by the crime rate that results from the algorithm's rule evaluated at the judge's release rate, followed by the release rate that could be achieved by the algorithmic rule evaluated at the same crime rate that the judges achieve.

A Appendix Figures

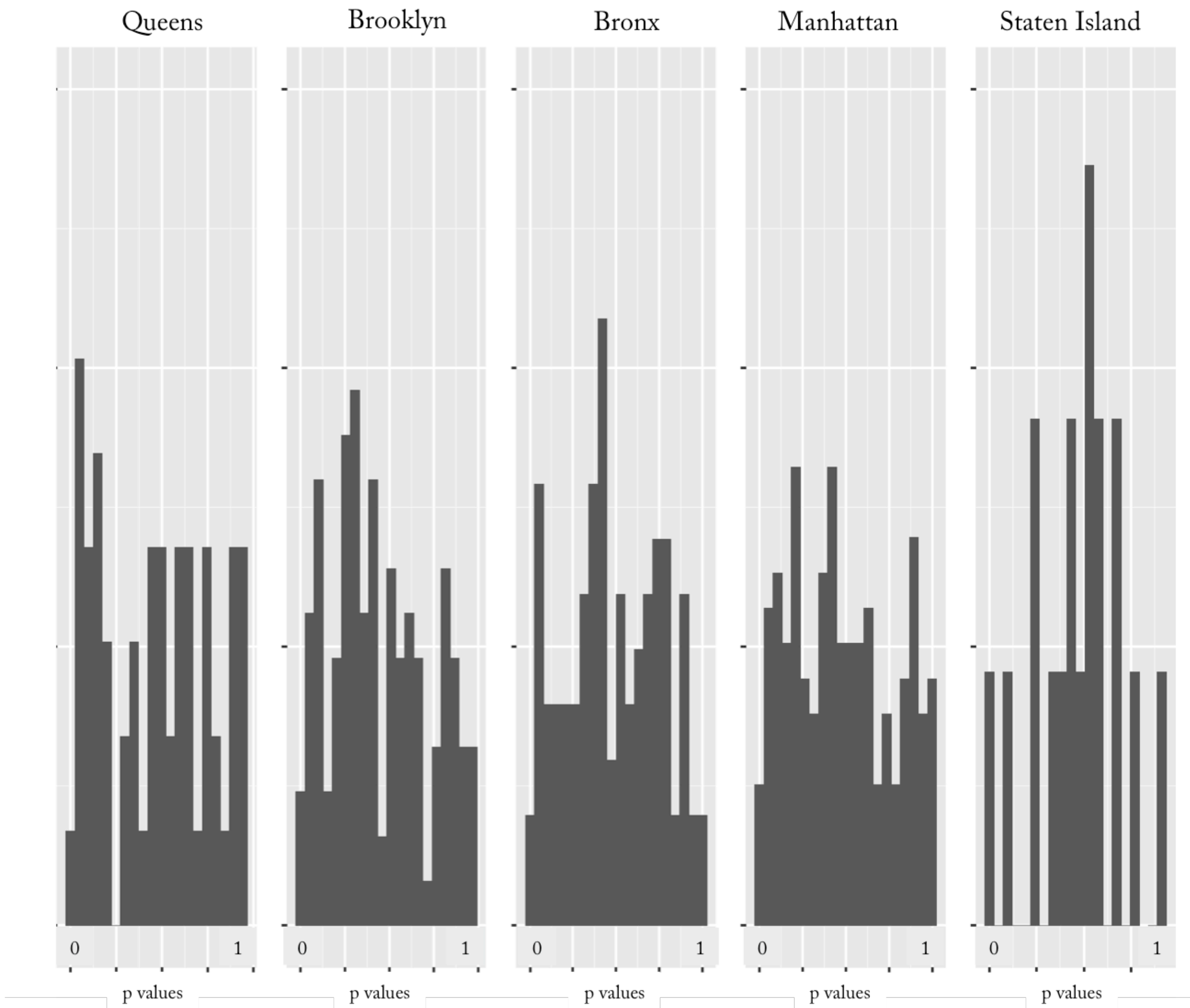

Figure A.1: Testing Quasi-Random Assignment of Defendants Across Leniency Quintiles
Borough by Borough Distribution of p-values for Balance Tests in Contraction Sample

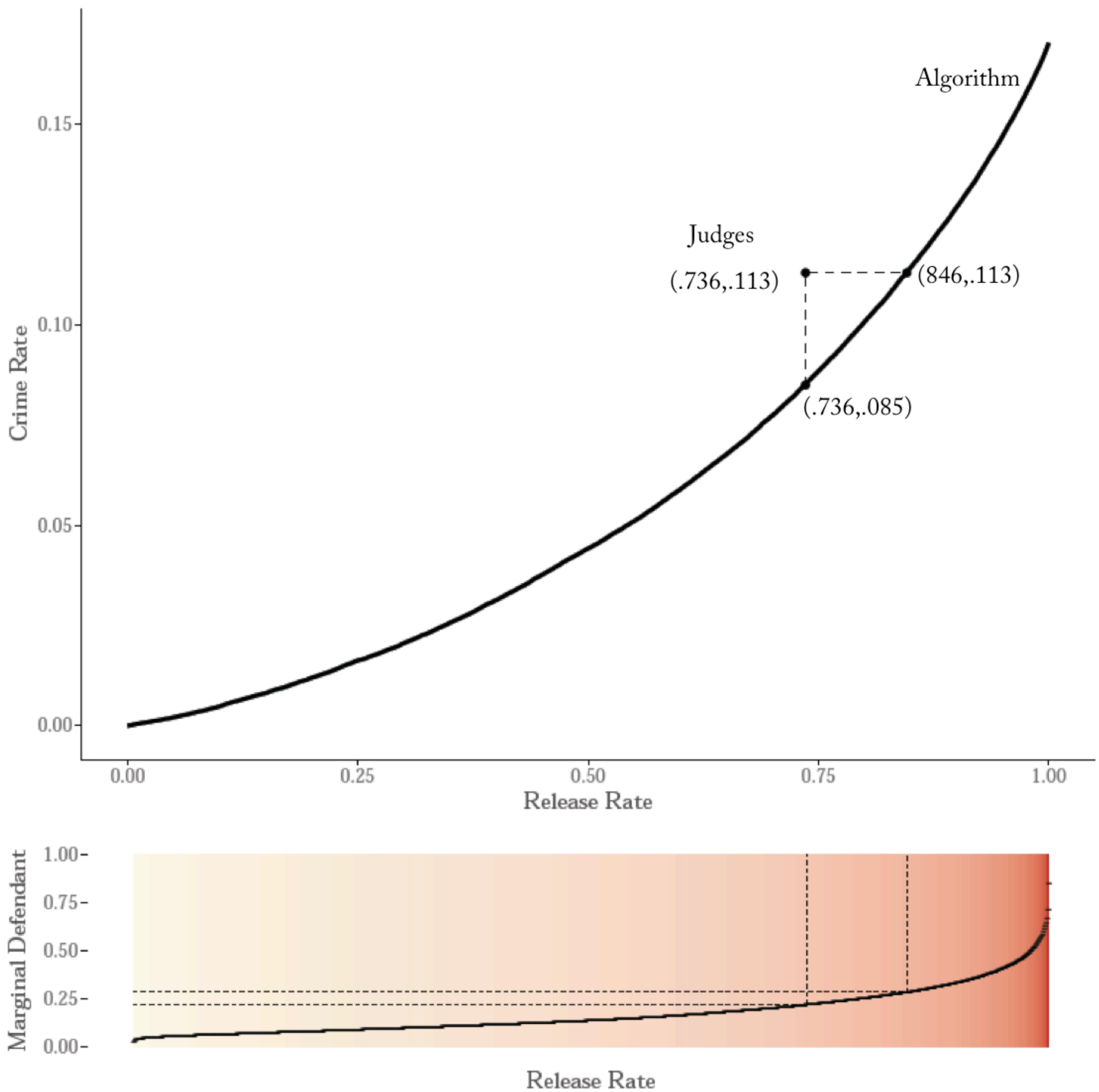

Figure A.2: The Algorithmic Crime Rate - Release Rate Tradeoff
When Defendants are Released According to Predicted Risk

Notes: The curve in the top panel shows the crime rate and release rate combinations that would be possible in NYC if judges were given a risk tool that could re-rank all defendants by their predicted crime risk and recommend them for detention in order of risk. Since we would like a crime rate that can be meaningfully compared across release rates, the y-axis shows the ratio of crimes committed by released defendants to the *total* number of defendants, not just the number released.The curve shows what gains would be possible relative to actual current judge decisions, assuming perfect compliance with the new tool. The curve in the bottom panel shows the risk level of the marginal person detained at each possible release rate under the algorithmic release rule.

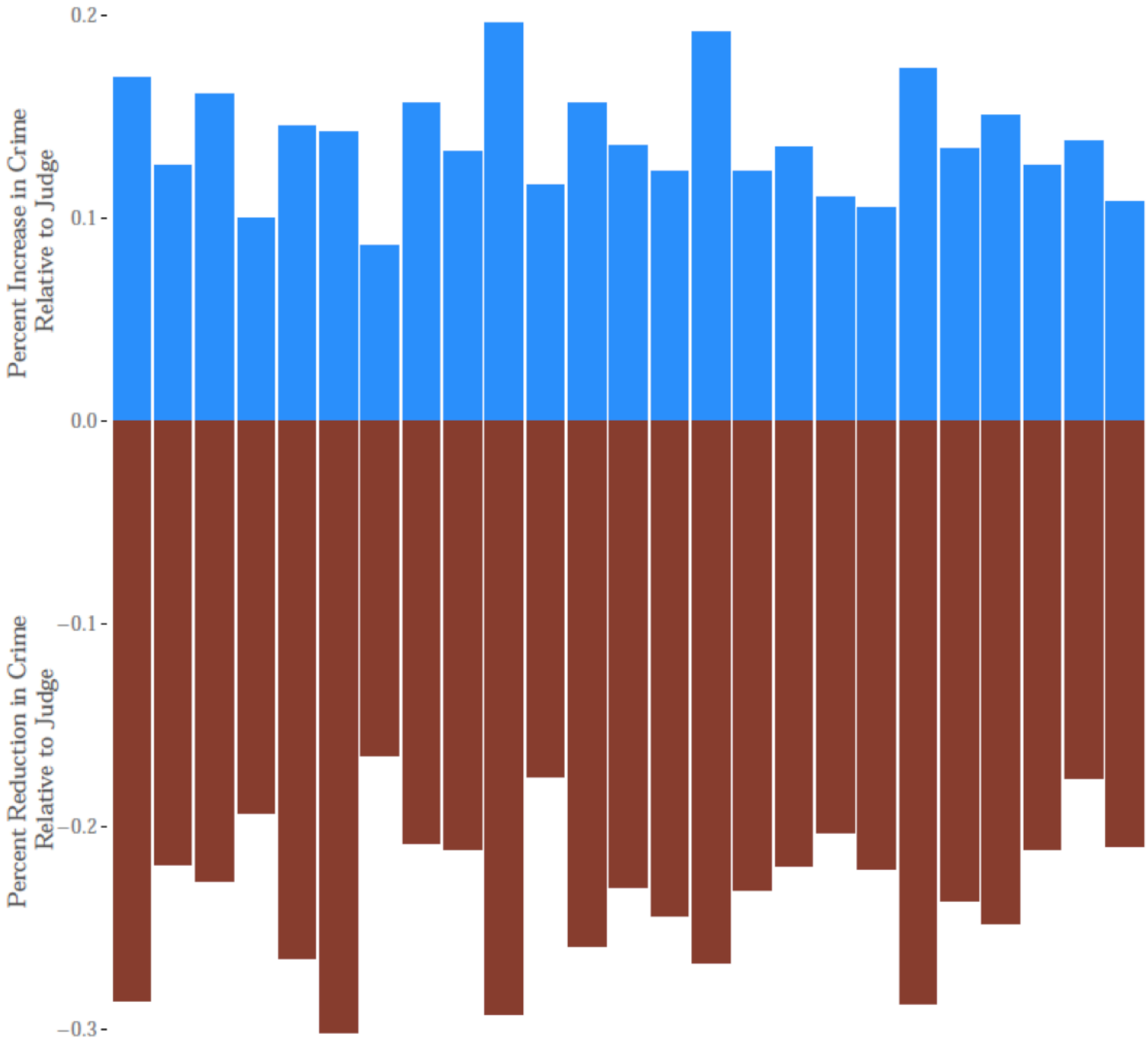

Figure A.3: Gains from Algorithmic Re-Ranking over Individual Judges

Notes: This figure shows the potential gains from an algorithmic release rule versus current judge decisions separately for each of the 25 judges in our NYC dataset with the largest caseloads within our test set. We use the algorithm trained on data from all judges and cases in the training set, and then compare potential gains of the algorithm versus each of these judges one at a time (each bar represents comparison to a separate judge) in terms of gains in release rates holding crime rate constant (top panel) and reduction in crime rates holding release rates constant (bottom panel).

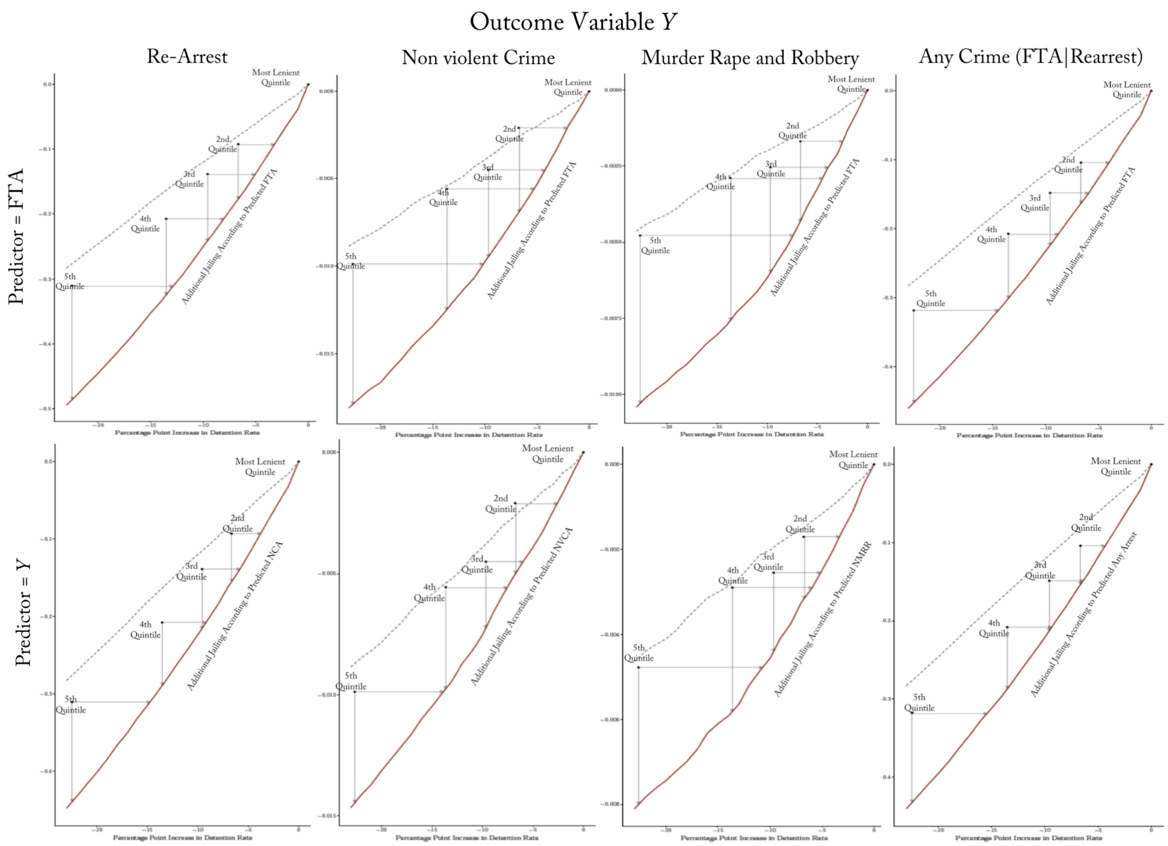

Figure A.4: Effect of Contraction on Other Outcomes

Notes: This figure presents results of the contraction exercise as in Figure V but using outcomes other than Failure to Appear. Each column uses a different outcome: rearrest (NCA), non-violent crime (NVCA), murder rape and robbery (NMRR) and any crime (NCA or FTA). In the first row, contraction is done using the FTA predictor exactly as in Figure V—additional defendants are jailed according to predicted flight risk. In the second row, contraction takes place using a $\hat{Y}$ predictor (trained on a separate training sample) where Y is the dependent variable in question, so an NCA predictor is used in column 1, row 2 for example.

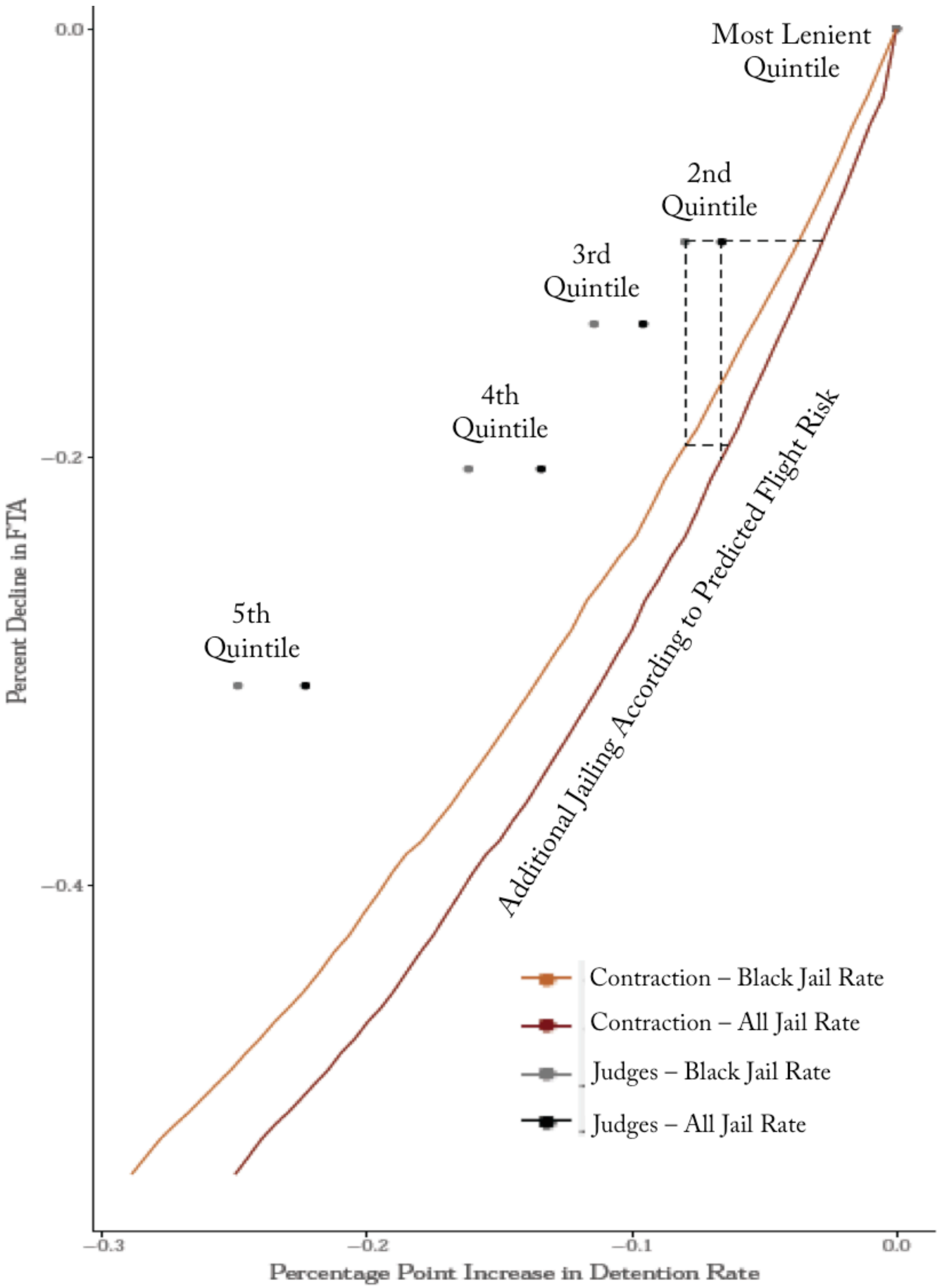

Figure A.5: Effect of Contraction on Racial Bias

Notes: This figure presents results of the contraction exercise as in Figure V but focusing on jailing rates of blacks separately. We show both the black and total jailing rates for both judges and the contraction algorithm.

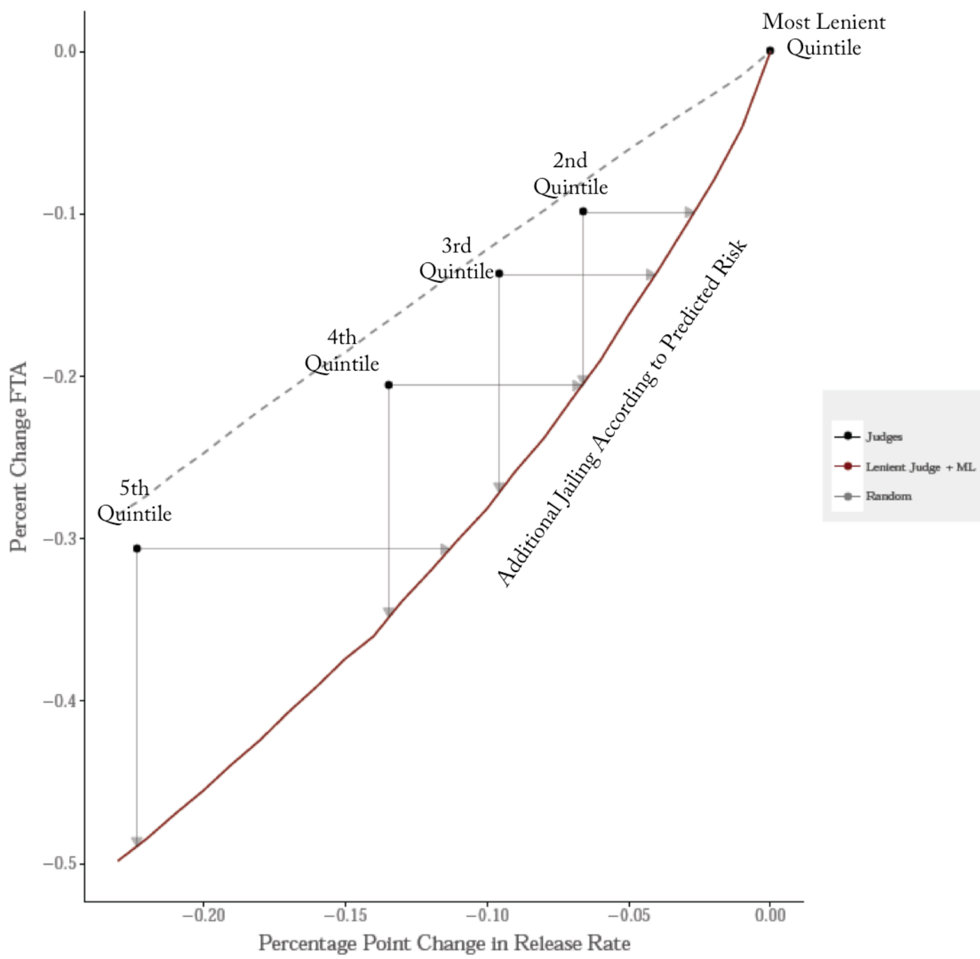

Figure A.6: Effect of Contraction when Algorithm can only Jail only ROR Defendants

Notes: This figure looks at performance when additional defendants are jailed according to a predictive model of who judges would jail as in Figure V, except it allows the algorithm to only jail from the subset of defendants whom judges ROR. It compares crime rates and release rates to the decisions of stricter judges. The right-most point in the graph represents the release rate of the most lenient quintile of judges, with the crime rate that results. The red line shows the crime reductions that realize if we jail additional defendants from this pool according to predicted behavior of judges. By comparison, the light dashed line shows the decline in crime (as a percentage of the lenient quintile's crime rate, shown on the y-axis) that results from randomly selecting additional defendants to detain from within the lenient quintile's released cases, with the change in release rate relative to the lenient quintile shown on the x-axis. As another comparison, the green curve shows the crime rate / release rate tradeoff that comes from jailing additional defendants within the lenient quintile's released set in descending order of the algorithm's predicted crime risk. The four points on the graph show the crime rate / release rate outcomes that are observed for the actual decisions made by the second through fifth most lenient quintile judges, who see similar caseloads on average to those of the most lenient quintile judges.

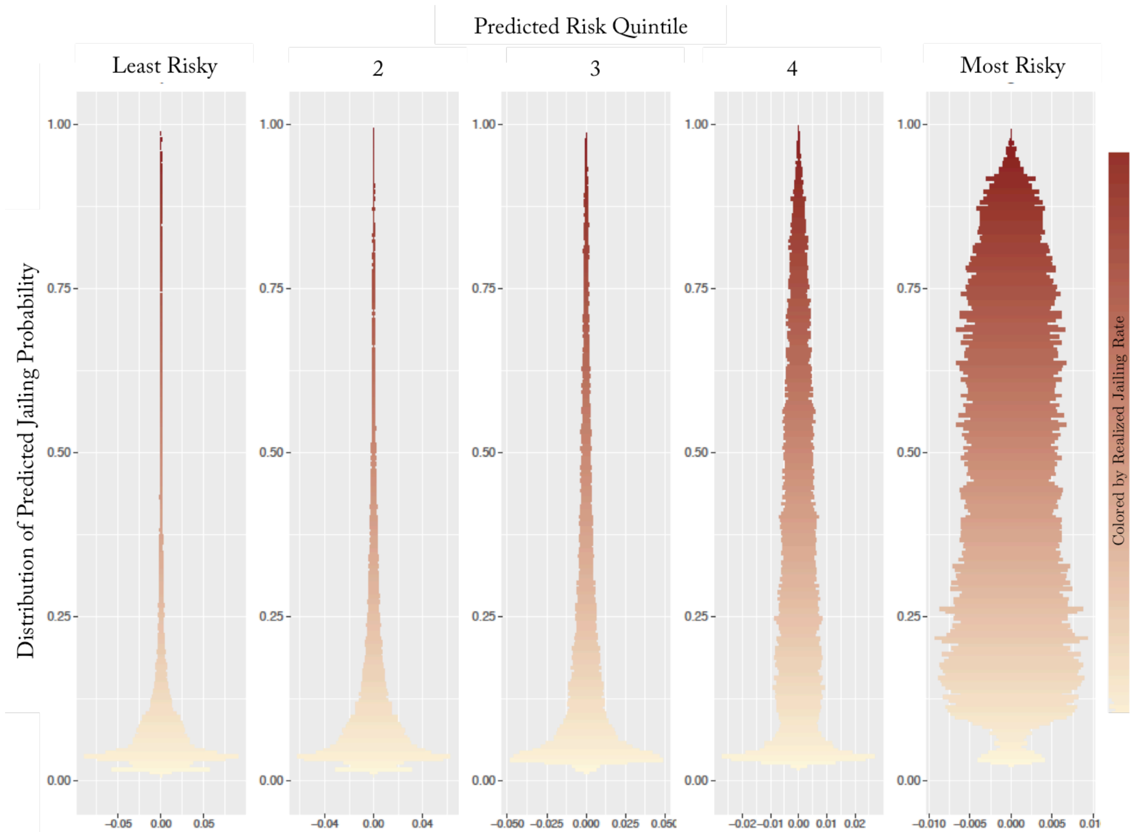

Figure A.7: How Do Release Decisions Change with Crime Risk?
Distribution of Case-by-Case Jailing Probabilities by Risk Quintile

Notes: This figure shows the relative "spread" in the predicted judge jail probabilities for cases in our NYC test set grouped together by the algorithm's predicted crime risk. In each predicted risk quintile we graph the distribution (using a volcano plot) of the conditional distribution of predicted judge jailing probabilities. We further color each point by the realized release rate at each rate.

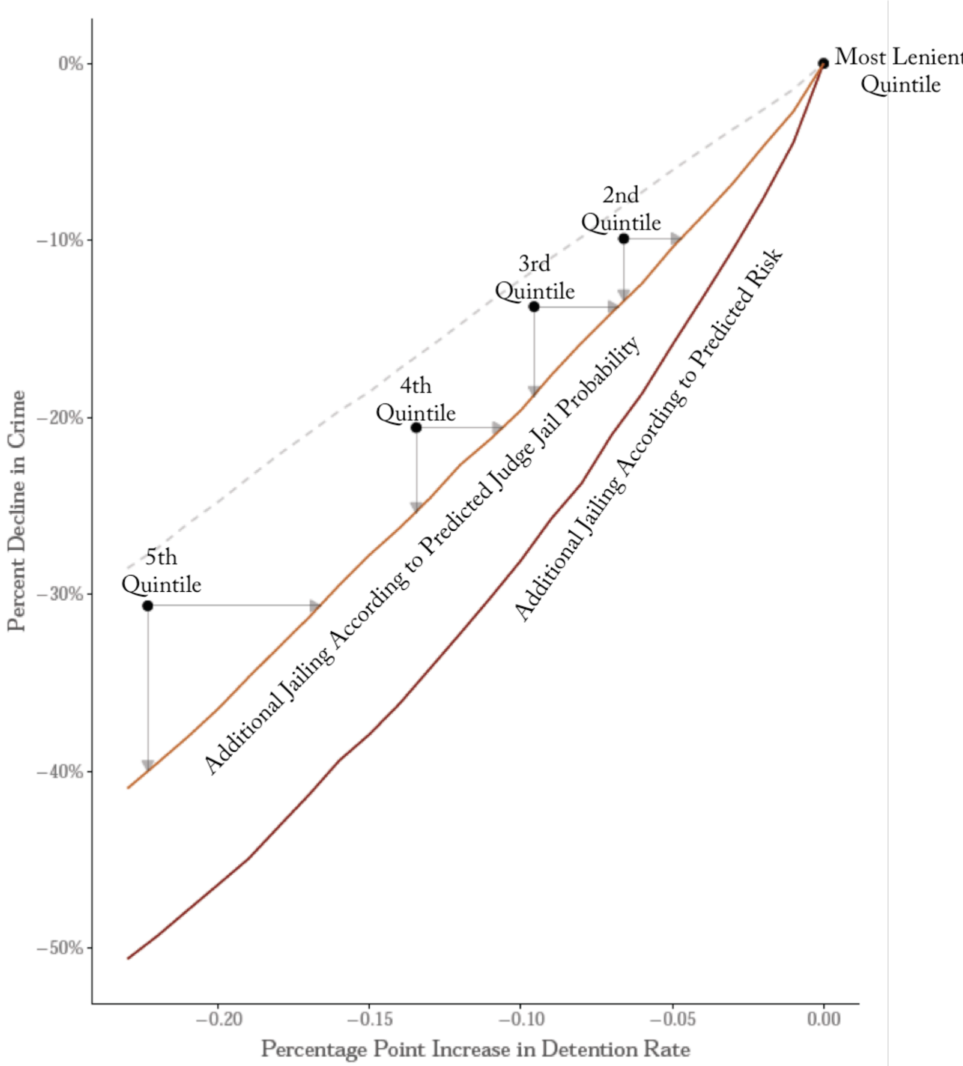

Figure A.8: Comparing Judges to Predicted Judges
Effect of Jailing Additional Defendants According to Predicted Jail Probabilities

Notes: This figure compares the change in crime rates and release rates that could be achieved by jailing additional defendants using the algorithm's predicted crime risk compared to the decisions of stricter judges. The right-most point in the graph represents the release rate of the most lenient quintile of judges, with the crime rate that results. The light dashed line shows the decline in crime (as a percentage of the lenient quintile's crime rate, shown on the y-axis) that results from randomly selecting additional defendants to detain from within the lenient quintile's released cases, with the change in release rate relative to the lenient quintile shown on the x-axis. The red curve shows the crime rate / release rate tradeoff that comes from jailing additional defendants within the lenient quintile's released set in descending order of the algorithm's predicted crime risk. The additional curve on the graph shows the crime rate / release rate outcomes we would get from jailing additional defendants within the lenient quintile judges' caseloads in descending order of an algorithm's predicted probability that the judges jail a given defendant. The four points on the graph show the crime rate / release rate outcomes that are observed for the actual decisions made by the second through fifth most lenient quintile judges, who see similar caseloads on average to those of the most lenient quintile judges.

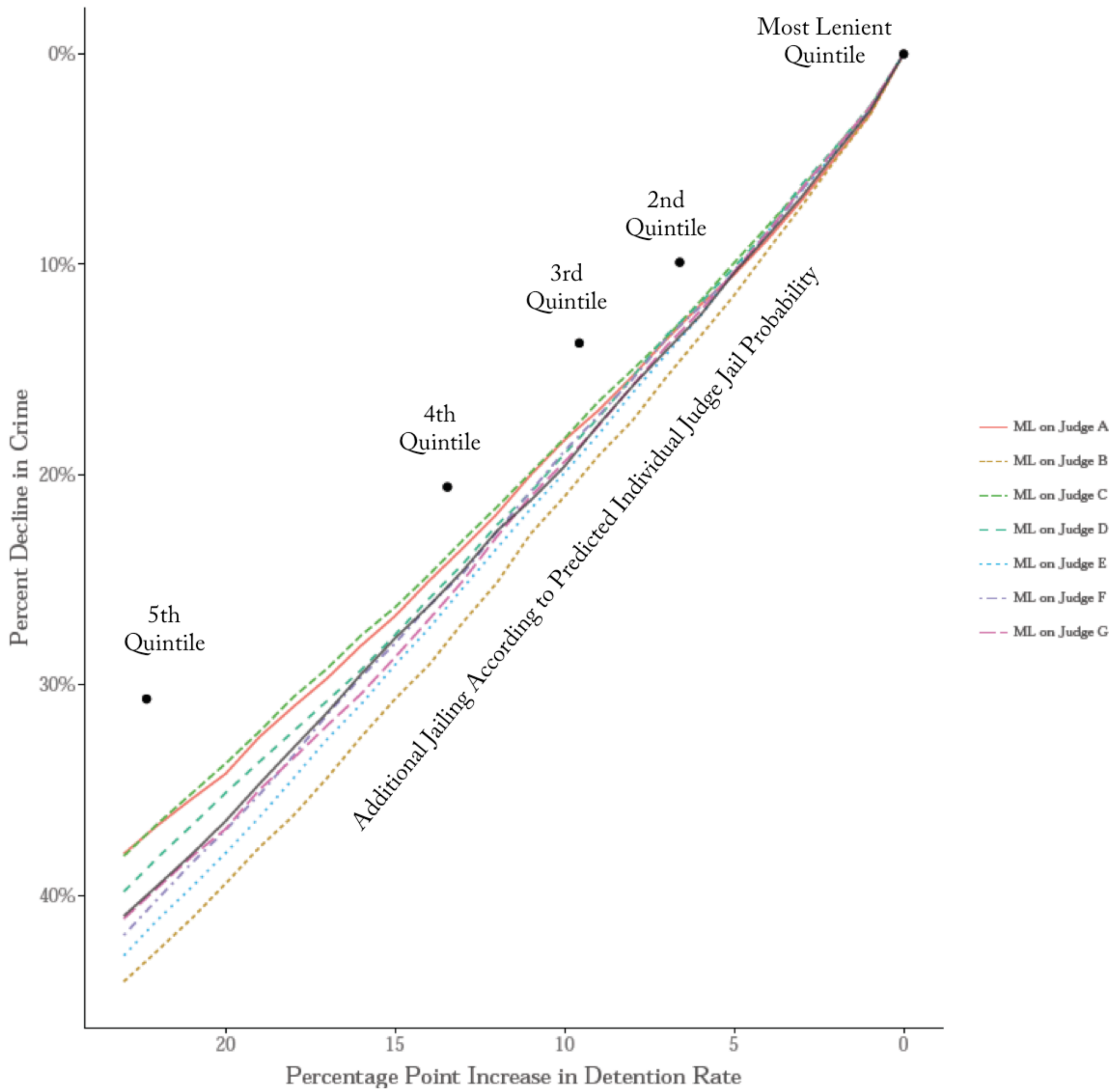

Figure A.9: Effect of Detaining Based on Individual Judge Predictions

Notes: This figure looks at performance when additional defendants are jailed according to a predictive model of individual judge's decisions. Each line shows performance of each of seven judges with the highest caseloads. The four points on the graph show the crime rate / release rate outcomes that are observed for the actual decisions made by the second through fifth most lenient quintile judges, who see similar caseloads on average to those of the most lenient quintile judges.