ANALYZING HUMAN DECISIONS AND MACHINE PREDICTIONS IN BAIL DECISION MAKING

Jon Kleinberg
Cornell University

Himabindu Lakkaraju
Harvard University

Jure Leskovec
Stanford University

Jens Ludwig
University of Chicago

Sendhil Mullainathan
University of Chicago

ABSTRACT

Can machine learning (ML) models outperform human decision makers in settings involving high stakes decisions? Bail decisions provide a good test case. Millions of times each year, judges make jail-or-release decisions that hinge on a prediction of what a defendant would do if released. The concreteness of the prediction task combined with the volume of data available makes this a promising machine-learning application. Yet comparing the algorithm to judges proves complicated. First, the available data are generated by prior judge decisions. We only observe the crime outcomes for released defendants, not for those who were detained. This makes it hard to evaluate counterfactual decision rules based on algorithmic predictions. To address the aforementioned challenges, we outline novel econometric strategies to evaluate ML models in the presence of missing counterfactuals. We leverage these strategies and carry out a first of its kind, large scale study to compare human and algorithmic decision-making on a dataset of about 1.5 million bail decisions. Policy simulations on this data demonstrate that ML models can achieve crime reductions up to 24.7% with no change in jailing rates, or alternatively jailing rate reductions up to 41.9% with no increase in crime rates. Our analysis also reveals that these reductions are persistent across all categories of crime including violent crimes and that these reductions can be achieved while simultaneously reducing racial disparities. We also develop statistical tests to understand the reasons for the performance gap between human and algorithmic decision making. More specifically, we discover that unmeasured confounders are unduly influence human decisions. These results suggest that while machine learning can be valuable, realizing this value requires integrating these tools into an economic framework: being clear about the link between predictions and decisions, and constructing unbiased decision counterfactuals.

Humans Decisions vs Machine Predictions: Large Scale Analysis of Bail Decision-Making

Here, we present a large scale analysis of human and algorithmic performance on the task of bail decision-making. We carry out our analysis on a dataset of about 1.5 million cases which corresponds to all the arrests made in New York City between 2008 and 2013. We organize the rest of the write up as follows: First, we describe in detail the dataset that we employ for our analysis. Next, we discuss how to compare human decisions and algorithmic predictions when the data is selectively labeled i.e., we only observe the crime outcomes of defendants released by judges. To this end, we introduce a novel approach called *contraction.* We then describe the novel diagnostic tests that we propose to better understand the factors that contribute to errors in human judgment. Lastly, we conclude our discussion by addressing a critical question - *are racial disparities aggravated when making decisions using machine learning algorithms?*

Dataset Description

Our dataset comprises of all arrests made in New York City between November 1, 2008 and November 1, 2013. This includes information pertaining to 1,460,462 cases and captures much of the information available to the judge at the time of the bail hearing, such as the instant offense and the rap sheet, which includes prior failures to appear (FTAs).

The dataset also includes the outcome of each case (release/detain), FTA, and any re-arrest prior to resolution of the case. The only measure of defendant demographics we use to train our algorithms is age.[1]

Of the initial sample, $758,027$ were released on bail and so relevant for our analysis.[2] We then carry out further preprocessing which leaves us with a working dataset of 554,689 cases. Table 2 (Appendix) presents descriptive statistics for our analysis sample. As is true in the criminal justice systems of many American cities, males (83.2%) and minorities (48.8% African-American, 33.3% Hispanic) are over-represented. A total of 36.2% of our sample was arrested for some sort of violent crime (nearly two-thirds of which are simple assaults), 17.1% for property crimes, and 25.5% for drug crimes, and the remaining arrests are for a mix of miscellaneous offenses such as driving under the influence, weapons, and prostitution. 73.6% of defendants were released prior to adjudication. Of those released, 15.2% fail to appear, while 25.8% are re-arrested prior to adjudication, 3.7% are arrested for a violent crime, and 1.9% are arrested for the most serious possible violent crimes (murder, rape, and robbery).

Table 2 (Appendix) also makes clear judges are paying some attention to defendant characteristics in deciding who to release, since the average values differ by release status. Exactly how good judges are in making these decisions relative to an algorithm's predictions is our focus for the rest of the section.

Comparing Human Decisions and Machine Predictions on NYC Bail Data

We randomly partition our working dataset of 554,689 observations into a 80% training set of 443,751 observations which we use to train our predictive model and a 20% hold-out test set of 110,938 observations on which we evaluate the model's performance. We employ gradient boosted trees as our predictive model and we denote it by *M*. The resulting AUC of the model *M* on the hold-out test set is 0.707.

To compare the performance of the predictive model *M* with that of human judges, we propose a novel technique called *contraction* technique. To illustrate the contraction technique, consider the setting where each judge decides on bail petitions of 100 defendants. Let us say our goal is to compare the performance of a black box model with that of some judge j' who releases 70% of the defendants who appear before him. In order to compare the model performance with j', we run the black box model on the set of defendants judged by the most lenient judge q who releases, say 90%, of the defendants. We achieve this by constraining the black box model to detain the same 10 defendants who were detained by q thus avoiding the missing labels. In addition to these 10 defendants, we allow the black box model to detain another 20 defendants deemed as highest risk by the model. We then compute the crime rate on the remaining 70 defendants who are released by the model. Since the outcome labels (crime/no crime) of all of these defendants are observed, the crime rate can be easily computed from the data. Note that our contraction technique exploits the following characteristics of the data and the problem setting: presence of multiple decision makers, random assignment of defendants to decision makers, and heterogeneity of release rates across decision makers.

Figure 1 highlights the performance of both human judges as well as the predictive model *M*. Note that the performance of the predictive model *M* is computed using the aforementioned contraction technique. The solid red curve shows the crime reductions achieved by the predictive model relative to the crime rate of the most lenient quintile of judges. By comparison, the light dashed line shows the decline in crime (as a percentage of the lenient quintile's crime rate) that results from randomly selecting additional defendants to detain from within the lenient quintile's released cases. The four black points on the graph show the crime rate and the corresponding release rate 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.

We find that the predictive model *M* achieves significant gains over human judges. The second quintile of judges reduce crime by 9.9% by increasing the detention rate by 6.6 percentage points. The predictive model, on the other hand, manages to achieve the same crime reduction by increasing the detention rate by only 2.8 percentage points, or equivalently by increasing the detention rate by 6.6 percentage points, the predictive model could have reduced crime by 20.1%. Such significant gains can be observed even with other leniency quintiles. In summary, the predictive model could result in 24.7% fewer crimes while maintaining the same detention rate as that of the judges (0.264)[3], Alternatively, to produce the same crime rate as judges currently do, the predictive model could jail 41.8% fewer people.

[1] Previous research demonstrates a strong age patterning to criminal behavior, and courts have generally found consideration of age to be legally acceptable.

[2] We exclude 272,381 desk appearance tickets, as well as the 295,314 cases disposed of at arraignment, the 131,731 cases that were adjourned in contemplation of dismissal, and eliminate some duplicate cases.

[3] This detention rate is computed across all the 554,689 cases. Refer to Table 2 for details. Note that the detention rate of the most lenient quintile of judges is 0.171

Understanding Judges' Mis-predictions

The previous results suggests that judges are mis-predicting. We now attempt to understand why they are mis-predicting. This exercise can help shed light into what judges are getting wrong, and more generally highlights the potential of machine learning tools to help test theories of human decision-making and behavior, not just solve policy problems.

Which cases are hard? Our goal here is to understand where in the risk distribution judges are having the most trouble. While *ex ante* the answer is not obvious, looking at other domains could potentially provide us with some initial intuition. For example in education, studies find that principals do a good job identifying which teachers are in the tails of the performance distribution - the very high performers and the very low performers- but have a hard time distinguishing among teachers in the middle of the distribution Jacob and Lefgren [2005].

We can examine this question in our data by investigating where in the predicted-risk distribution judges have the most uncertainty. What we observe is just a binary indicator of whether the judges released a given defendant i which does not convey much about when judges are most uncertain. However we can learn more about this by predicting the behavior of the judges. That is, we can train a new model to predict not the behavior of the defendants, but the decisions of the judges *i.e.*create a *predicted judge* denoted by $\hat{J}$.

In Figure 2 we sort all defendants in our test set into quintiles based on the model *M*'s predicted risk of crime as shown along the x-axis. Then for each quintile of predicted defendant risk, we show the distribution of the predicted judge $\hat{J}$'s jailing probabilities. Additionally, we color each slice of the distribution by the actual jailing rate in that slice. We see from this that in each quintile of risk, those with highest predicted jailing rates are in fact jailed more: our predictions of the judge have some predictive power.

More importantly, we can also see in Figure 2 that cases in the lowest-risk quintiles have a much narrower distribution spread of predicted jailing probability. These cases appear relatively easy - or at least judges behave consistently. In contrast, there is substantial dispersion in predicted jailing probabilities for the highest-risk cases; some high-risk defendants have high jailing probabilities while other cases with similarly high risk have low probabilities of jail—that is, they are treated as if they are low risk. Thus, in contrast to what we might have expected, judges in this case do not seem to struggle most with the middle. Instead they seem to struggle most with one tail - the high risk cases.

Are Judges Misusing Unobservables? Our prediction of the judge's release decisions can also give us some sense for why judges might be mis-predicting. Recall that judge decision pertaining to a defendant i is a function of both observables $\boldsymbol{x}_i$ (like criminal record) and factors $\boldsymbol{z}_i$ not observable to the predictive models (like the defendant's demeanor in the courtroom). The behavioral science literature suggests that this extra information could be part of the problem and may be the reason judges do poorly relative to the algorithm. For example, few studies suggest that highly salient interpersonal information (such as the degree of eye contact that is made) can be over-weighted Danziger et al. [2011]; and that less salient but more informative data (like past behaviors) can be under-weighted.

If $\boldsymbol{z}$ is adding useful signal to judges in helping them predict defendant risk in our bail application, then judge should be be able to outperform predicted judge $\hat{J}$. If z simply adds noise then the reverse should be true. In essence, we would like to compare the judge to the predicted judge. We will do this in the same way we compared in Section the judge to the predictive model M. As before, we utilize the contraction technique and begin with the set of cases released by the most lenient quintile judges. We then jail additional defendants as we predict the judges would - jailing first those defendants with the highest jailing probability as per the predicted judge $\hat{J}$ (the difference with our previous comparison to the judges is that we had earlier jailed defendants by the model M's predicted crime risk).

Figure 3 shows the results of this exercise. We plot the performance of judges in each leniency quintile. We then show what crime rate would result if we jailed as the predicted judge would. We also include for comparison here what would happen if we jailed according to predicted crime output by model M. We see here clearly that the predicted judge does better than judges themselves. In fact, simply jailing according to $\hat{J}$ gets us roughly halfway towards the gain we had from jailing according to predicted risk. Our results taken together suggests one reason why judges might be mis-predicting: Their actual decisions are noisy relative to the predicted judge $(\hat{J})$. In particular, this "noise" appears to be due to unobservable variables, which unduly influence these decisions.

Are Machines Aggravating Racial Inequity?

One of the most important concerns that has been raised by the possibility of using data-driven predictions to inform criminal justice decisions is racial inequity. Even though we do not make race or ethnicity available to the machine learning model, it is possible the model winds up using these factors inadvertently - if other predictors are correlated with race or ethnicity. As then-Attorney General Eric Holder noted in a 2014 speech, "we need to be sure the use of aggregate data analysis won't have unintended consequences," cautioning that these tools have the potential to "exacerbate unwarranted and unjust disparities that are already far too common in our criminal justice system and in our society."[4] To some this outcome is a foregone conclusion; for example law professor Bernard Harcourt has argued that "using risk-assessment tools is going to significantly aggravate the unacceptable racial disparities in our criminal justice system" (Starr [2014], Harcourt [2014].

Table 1 compares the performance of judge decisions to different versions of our predictive model M. The first row shows the racial composition of the defendant pool as a whole. The second row shows the outcomes of judge decisions. At the current release rate, the crime rate (the ratio of crimes to total defendants)equals 11.3%, and 88.9% of inmates in jail are members of minority race or ethnic groups. The second row presents the results of our predictive model M, which does not have access to information about defendant race or ethnicity. As noted above, the resulting release rule lets us reduce crime by 24.7% at a given release rate, with a jail population that would turn out to have about the same share minority as we see under current judge decisions, equal to 90.1%. There is a slight increase by a few percentage points in the share of the jail population that is black, which is mostly offset by a small decline in the share Hispanic. But these are very small differences given the sampling variability in our test set.

The remaining rows of Table 1 show that it is possible to explicitly constrain the predictive model to ensure no increase in racial disparities in the jail with very little impact on the model's performance in reducing crime. For example the third row shows what happens if we use our model M's predicted risk to rank order defendants separately by their race and ethnic group (white, black and Hispanic). We then detain defendants in descending order of risk but stop detaining black and Hispanic defendants once we have hit the exact number of each group detained by the judges, to ensure that the minority composition of the jail population is no higher under the model's release rule compared to the judges. Compared to the crime gains we achieved using the algorithm's usual ranking-and-release rule, ensuring that we do no worse than matching the judges on racial disparities in the jail population leads to almost no loss in terms of the potential gain in reduced crime.

The second-to-last row shows what happens if we constrain the model to be race neutral (that is, to ensure the share of each minority group in the jail population is no higher than their share within the general pool of defendants). The final row goes one step further and constrains the algorithm to ensure that the share of the jail population that is either black or Hispanic is no higher than either the share the judge jails or the share within the general defendant pool. In both cases we see that it is possible to reduce the share of the jail population that is minority - that is, reduce racial disparities within the current criminal justice system - while simultaneously reducing crime rates relative to human judges.

References

S. Danziger, J. Levav, and L. Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

B. E. Harcourt. Risk as a proxy for race: The dangers of risk assessment. *Federal Sentencing Reporter*, 27:237, 2014.

B. A. Jacob and L. Lefgren. Principals as agents: Subjective performance measurement in education. Technical report, National Bureau of Economic Research, 2005.

S. B. Starr. Evidence-based sentencing and the scientific rationalization of discrimination. *Stanford Law Review*, 66: 803, 2014.

[4] http://www.justice.gov/opa/speech/attorney-general-eric-holder-speaks-national-association-criminal-defense-lawyers-57th

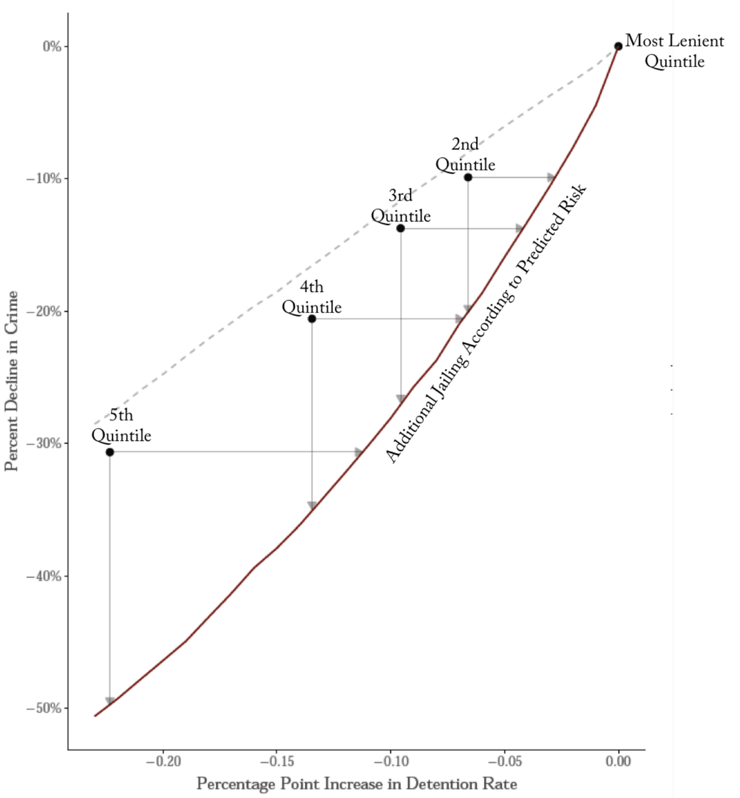

Figure 1: Comparing impact of detaining in order of predicted risk To what judges achieve

Notes: This figure presents the comparison of human and machine performance on NYC bail data. The performance of the predictive model *M* (gradient boosted trees) is computed using the contraction technique. The x-axis plots the detention rates relative to the release rate of the most lenient quintile of judges. The y-axis shows the percentage decline in crime rates relative to the crime rate of the most lenient quintile of judges. The red line shows the crime reductions achieved by predictive model as computed by the contraction technique. 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. 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.

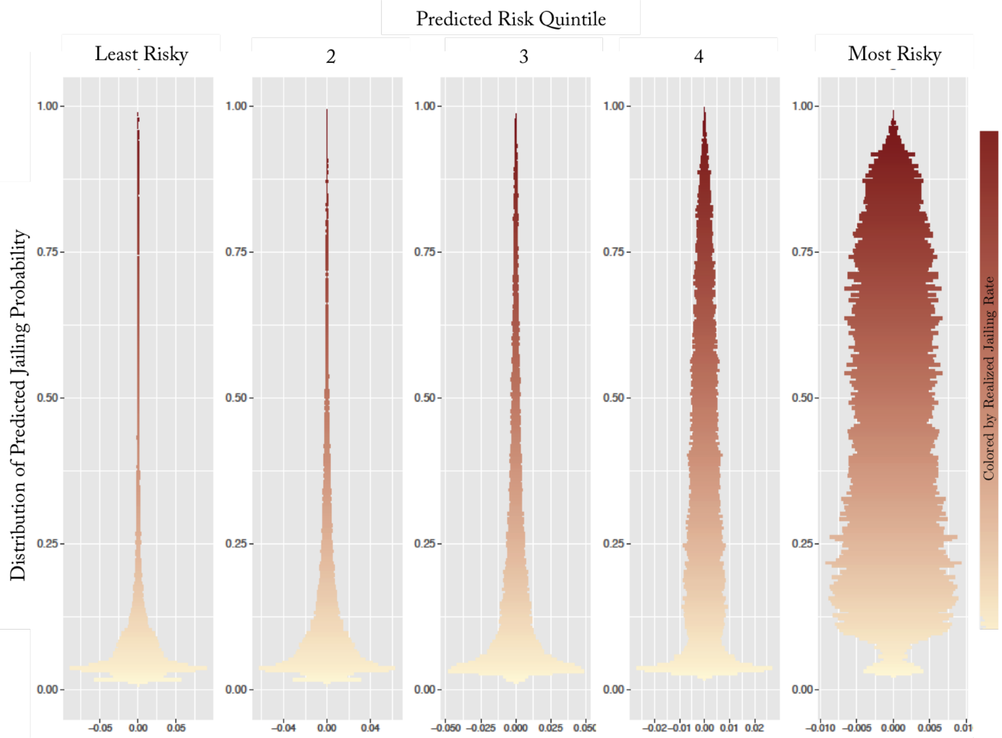

Figure 2: Distribution of predicted jail probabilities (predicted judge $\hat{J}$) by predicted crime risk (predictive model M).

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 predicted judge jailing probabilities. We further color each point by the realized release rate at each rate.

Table 1: Racial fairness.

	Crime Rate	Drop Relative	Percentage of Jail Population		
Release Rule		to Judge	Black	Hispanic	Minority
Distribution of Defendants (Base Rate)			.4877	.3318	.8195
Judge	.1134 (.0010)	0%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Predictive Model M					
Usual Ranking	.0854 (.0008)	-24.68%	.5984 (.0029)	.3023 (.0027)	.9007 (.0017)
Match Judge on Race	.0855 (.0008)	-24.64%	.573 (.0029)	.3162 (.0027)	.8892 (.0018)
Equal Release Rates for all Races	.0873 (.0008)	-23.02%	.4877 (.0029)	.3318 (.0028)	.8195 (.0023)
Match Lower of Base Rate or Judge	.0876 (.0008)	-22.74%	.4877 (.0029)	.3162 (.0027)	.8039 (.0023)

Notes: The first row shows the share of the defendant population overall that is black or Hispanic. The second row shows the results of the observed judge decisions. The third row shows the results of the predictive model M, which does not use race in predicting defendant risk and makes no post-prediction adjustments to account for race. In the fourth row we adjust the model's ranking of defendants to ensure that the share of the jail population that is black and Hispanic are no higher than those under current judge decisions. The next row constrains the model's jail population to have no higher share of black or Hispanic than that of the general defendant pool, while the final row constrains the model's jail population to have no higher share black or Hispanic than either the judge decisions or the overall defendant pool.

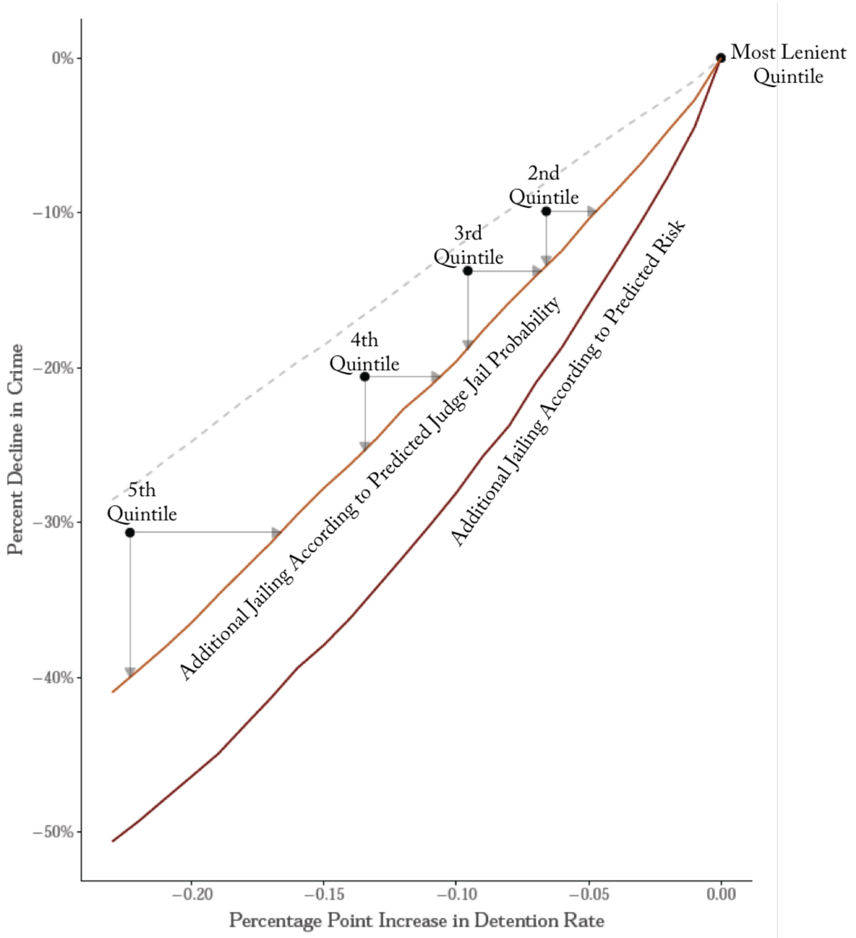

Figure 3: Judge vs Predicted Judge $\hat{J}$ vs Predictive Model M

Notes: 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 model M's predicted crime risk. The orange 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 predicted judge $\hat{J}$'s jailing probability. 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.

Appendix

Table 2: Summary statistics.

	Full Sample	Judge Releases	Judge Detains	P-value
Sample Size	554,689	408,283	146,406	
Release Rate	.7361	1.0000	0.00	
Outcomes				
Failure to Appear (FTA)	.1521	.1521		
Arrest (NCA)	.2581	.2581		
Violent Crime (NVCA)	.0372	.0372		
Murder, Rape, Robbery (NMRR)	.0187	.0187		
Defendant Characteristics				
Age	31.98	31.32	33.84	<.0001
Male	.8315	.8086	.8955	<.0001
White	.1273	.1407	.0897	<.0001
African American	.4884	.4578	.5737	<.0001
Hispanic	.3327	.3383	.3172	<.0001
Arrest County				
Brooklyn	.2901	.2889	.2937	.0006
Bronx	.2221	.2172	.2356	<.0001
Manhattan	.2507	.2398	.2813	<.0001
Queens	.1927	.2067	.1535	<.0001
Staten Island	.0440	.0471	.0356	<.0001
Arrest Charge				
Violent Crime				
Violent Felony	.1478	.1193	.2272	<.0001
Murder, Rape, Robbery	.0581	.0391	.1110	<.0001
Aggravated Assault	.0853	.0867	.0812	<.0001
Simple Assault	.2144	.2434	.1335	<.0001
Property Crime				
Burglary	.0206	.0125	.0433	<.0001
Larceny	.0738	.0659	.0959	<.0001
MV Theft	.0067	.0060	.0087	<.0001
Arson	.0006	.0003	.0014	<.0001
Fraud	.0696	.0763	.0507	<.0001
Other Crime				
Weapons	.0515	.0502	.0552	<.0001
Sex Offenses	.0089	.0086	.0096	.0009
Prostitution	.0139	.0161	.0078	<.0001
DUI	.0475	.0615	.0084	<.0001
Other	.1375	.1433	.1216	<.0001
Gun Charge	.0335	.0213	.0674	<.0001
Drug Crime				
Drug Felony	.1411	.1175	.2067	<.0001
Drug Misdemeanor	.1142	.1156	.1105	<.0001
Defendant Priors				
FTAs	2.093	1.305	4.288	<.0001
Felony Arrests	3.177	2.119	6.127	<.0001
Felony Convictions	.6157	.3879	1.251	<.0001
Misdemeanor Arrests	5.119	3.349	10.06	<.0001
Misdemeanor Convictions	3.122	1.562	7.473	<.0001
Violent Felony Arrests	1.017	.7084	1.879	<.0001
Violent Felony Convictions	.1521	.1007	.2955	<.0001
Drug Arrests	3.205	2.144	6.163	<.0001
Felony Drug Convictions	.2741	.1778	.5429	<.0001
Misdemeanor Drug Convictions	1.049	.5408	2.465	<.0001
Gun Arrests	.2194	.1678	.3632	<.0001
Gun Convictions	.0462	.0362	.0741	<.0001

Continued on next page

Table 2 – continued from previous page

	Full Sample	Judge Releases	Judge Detains	P-value

Notes: This table shows descriptive statistics of 554,689 cases that serve as our New York City analysis dataset. The p-values in the right most column are for pair-wise comparison of the equivalence of the mean values for the released versus detained defendants.