Guiding Prosecutorial Decisions with an Interpretable Statistical Model

Zhiyuan Lin
Stanford University
zylin@cs.stanford.edu

Alex Chohlas-Wood
Stanford University
alexcw@stanford.edu

Sharad Goel
Stanford University
scgoel@stanford.edu

ABSTRACT

After a felony arrest, many American jurisdictions hold individuals for several days while police officers investigate the incident and prosecutors decide whether to press criminal charges. This pre-arraignment detention can both preserve public safety and reduce the need for officers to seek out and re-arrest individuals who are ultimately charged with a crime. Such detention, however, also comes at a high social and financial cost to those who are never charged but still incarcerated. In one of the first large-scale empirical analyses of pre-arraignment detention, we examine police reports and charging decisions for approximately 30,000 felony arrests in a major American city between 2012 and 2017. We find that 45% of arrested individuals are never charged for any crime but still typically spend one or more nights in jail before being released. In an effort to reduce such incarceration, we develop a statistical model to help prosecutors identify cases soon after arrest that are likely to be ultimately dismissed. By carrying out an early review of five such candidate cases per day, we estimate that prosecutors could potentially reduce pre-arraignment incarceration for ultimately dismissed cases by 35%. To facilitate implementation and transparency, our model to prioritize cases for early review is designed as a simple, weighted checklist. We show that this heuristic strategy achieves comparable performance to traditional, black-box machine learning models.

CCS CONCEPTS

• **Applied computing** → **Law, social and behavioral sciences**; • **Computing methodologies** → *Machine learning*; *Learning linear models.*

KEYWORDS

criminal justice, interpretable machine learning, policy evaluation, prosecutorial decision making, propensity score matching

ACM Reference Format:
Zhiyuan Lin, Alex Chohlas-Wood, and Sharad Goel. 2019. Guiding Prosecutorial Decisions with an Interpretable Statistical Model . In *AAAI/ACM Conference on AI, Ethics, and Society (AIES '19), January 27–28, 2019, Honolulu, HI, USA.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3306618.3314235

AIES '19, January 27–28, 2019, Honolulu, HI, USA

ACM ISBN 978-1-4503-6324-2/19/01…$15.00
https://doi.org/10.1145/3306618.3314235

1 INTRODUCTION

Many American jurisdictions detain individuals after a felony arrest while they conduct a preliminary investigation of the incident. In some states, this detention period is legally allowed to extend up to 48 weekday hours, but can stretch to more than 100 total hours if weekends or holidays occur in the middle of the detention period. Police departments use this detention window to gather evidence for the local prosecuting attorney, and at the conclusion of the investigation period, the prosecutor's office reviews the available evidence and decides whether or not to prosecute the case. Such pre-arraignment detention serves two key purposes: first, to protect public safety by confining individuals who may pose a threat to the community; and second, to reduce the need for police officers to seek out and re-arrest individuals who are ultimately charged. Such detention, however, also exacts high social and financial costs. Even brief incarceration may lead to job loss, inadequate support for dependents, strains on family relationships, and social stigma [8, 9, 13, 19]. These detentions may also undermine public trust in law enforcement and disrupt community activity, particularly among those who are arrested but never charged with any crime.

Here we carry out one of the first extensive examinations of pre-arraignment incarceration by analyzing a detailed dataset of approximately 30,000 felony arrests in a large American city between 2012 and 2017.[1] We find that 45% of felony arrestees in our dataset are never charged by the prosecutor, though they often spend several nights in jail before that decision is made. This high rate of dismissal may in part be the result of differing standards of evidence applied for arrest and prosecution. In particular, though the legal standard for arrest is *probable cause*, prosecuting attorneys will typically only pursue a case when they believe they can establish an individual's guilt *beyond a reasonable doubt*, a considerably higher legal bar.

To safely reduce pre-arraignment incarceration, we propose a three-step, algorithmically assisted strategy. First, based on preliminary arrest reports, we estimate each felony arrestee's likelihood of eventual dismissal via a simple statistical model that takes the form of a weighted checklist. The model's simplicity means that arrestees can be scored manually, without the need for computing resources, and that policymakers can more easily audit the scoring procedure. Second, we use these scores to create a short list of arrestees each day that have the highest likelihood of eventual dismissal. Third, prosecutors, under our proposed strategy, carry out an early review of these candidate cases, and dismiss those deemed unlikely to meet the bar for charging. Though prosecutors may still choose to wait until the 48-hour deadline to issue the actual charging decision, we find that most of the pertinent information is available within the

[1] We are unable to disclose the name of the city due to our agreement with the jurisdiction.

first few hours after arrest, facilitating early review and preliminary dismissal.

We estimate that prosecutors in the jurisdiction we consider could often handle 5–10 additional cases per day at existing staffing levels. By carrying out early review of five cases per day, we estimate that prosecutors could reduce pre-arraignment incarceration time for ultimately dismissed felony arrestees by up to 35%. We further consider the effect of early review and release on public safety, and find that the effect of extended pre-arraignment detention on recidivism is negligible.

Though pre-arraignment detention serves an important role in the criminal justice system, it often imposes a heavy burden on individuals and communities. By designing a simple and transparent algorithmic system to identify candidates for early dismissal, we seek to reduce the harms of detention while retaining its benefits.

2 BACKGROUND

To frame our analysis, we start by briefly outlining the steps leading up to a prosecutor's charging decision. We then review some of the related literature on constructing and evaluating algorithms to guide decisions in the criminal justice system.

2.1 From Arrest to Arraignment

After an arrest is made by the city's police department, an officer brings the arrested individual to county jail. The county sheriff books the individual on charges proposed by the arresting officer and enters information about the incident into the sheriff's database. This booking information is typically available to the prosecutor's office within a few hours of arrest, and forms the basis of our early review system. In theory, an individual can post bail during this post-arrest detention period, though in practice most arrestees remain in custody while awaiting a charging decision. Following booking, the arresting officer completes one or more subsequent reports, including a narrative description of the incident. The jurisdiction we consider has 48 weekday hours to gather this information and make a decision to either charge or dismiss the arrestee.

A small group of *reviewing attorneys* at the prosecutor's office assesses these incoming arrests and makes a final charging decision. The prosecutor's office reviews approximately 20 felony cases daily—as we describe in Section 3 below, our analysis focuses exclusively on felonies—and maintains surplus capacity to handle spikes in workload. A decision to charge generally indicates a belief that sufficient evidence exists to demonstrate the individual committed the alleged crime beyond a reasonable doubt. In such cases, the arrestee would face their charges at an arraignment hearing the day after their charging decision has been filed. If the reviewing attorney believes there is insufficient evidence to proceed, or otherwise feels that prosecution would not serve the interest of justice, the prosecutor's office will dismiss the case.

2.2 Related Work

The use of statistical algorithms during the charging process is quite limited. There is, however, a long history of using related *risk assessment tools* in other parts of the criminal justice system [1, 3]. For example, *pretrial* risk assessments estimate a defendant's likelihood of engaging in future criminal activity or of failing to appear at trial, and are now widely used to inform judicial release decisions at arraignment.

To identify cases that are likely to be eventually dismissed, one could train traditional, black-box machine learning models. Such models are designed to achieve optimal accuracy, but are often difficult to interpret and to explain. Particularly in criminal justice applications, such opaqueness can impede adoption and sow mistrust. The newly active subfield of interpretable machine learning has sought to develop predictive algorithms that are both accurate and explainable [10, 31]. We apply a simple regress-and-round procedure that was recently shown to perform on par with traditional machine learning methods on a variety of prediction tasks [17].

In theory, algorithms can combat explicit and implicit bias in unguided human decisions. However, researchers and policymakers have also shown that statistical tools can themselves exacerbate inequities by inadvertently encoding biases in the training data or through otherwise poor design [2, 5–7, 14, 21, 22, 26, 27], a possibility we consider below.

3 DATA DESCRIPTION & EXPLORATION

We use detailed police and prosecutor data to develop our proposed early review system. Specifically, we consider booking information, police reports, and charging decisions for nearly every felony arrest of an adult in our partner jurisdiction from 2012 to 2017. For each case in our dataset, we have: the date, time, and basic arrestee demographic information; a list of initial charges proposed by the arresting officer; and the date-stamped charging decision by the prosecutor's office. Subsequent reports may be later filed by the police department to provide additional detail and clarification.

Starting from this corpus of 33,944 felony cases, we make several filtering decisions to facilitate our analysis. First, we exclude cases involving probation violations or outstanding warrants from other jurisdictions, as these cases are handled through a separate review process. Second, we similarly exclude murder and rape cases since they are likewise handled through a different procedure, given the severity of these crimes. Finally, we restrict to arrestees that are at least 18 years old at the start of our observation period, to ensure we can construct full criminal histories over the period we analyze. After this filtering, we are left with 26,606 cases for our primary analysis.

We reserve data from 2012, at the beginning of our observation period, so that we can construct a 1-year criminal record for each arrestee; and we reserve data from 2017, at the conclusion of the observation period, so that we can calculate 1-year recidivism rates for each individual. Our statistical modeling is thus restricted to 18,712 cases between 2013 and 2016.

Reviewing attorneys may make dismissal decisions throughout the day, but these are only communicated to the sheriff once per day, at approximately 4pm. The sheriff requires several hours to process these decisions, typically releasing arrestees around 9pm. Decisions in our dataset are only recorded with a date, not a time, and so we use these 4pm and 9pm times as approximations for decision and release in all cases.

Among the universe of cases we consider, 47% are never charged. The majority of eventually dismissed arrestees spend multiple

nights in jail awaiting a decision Indeed, 96% of eventually dismissed individuals spend at least 24 hours in detention, and 30% spend at least 72 hours. This delay is largely driven by the fact that reviewing attorneys typically wait until the police department files its final summary packet, which usually comes near the decision deadline. The final summary packet can in theory provide reviewing attorneys with new evidence, but in practice its primary function is to summarize the previously filed reports. Excepting the case summary packet, 90% of cases receive all associated reports and revisions within eight hours of booking. Our proposed process accommodates the possibility of additional information appearing in the summary packet by delaying the actual charging decision until the charging deadline.

4 OPTIMIZING RELEASE DECISIONS

To reduce pre-arraignment incarceration, we propose the following statistical strategy. First, when an arrestee is booked, we wait eight hours so that most case information can be filed by the police department. This window also serves as a "cooling off" period for the arrestee—often an implicit policy requirement before an individual can be released. After eight hours, we consider the case to be "review-ready". Each morning, review-ready cases are sorted by their estimated probability of dismissal according to the output of a predictive model (described below). Reviewing attorneys start the day by evaluating cases as normal, until the day's necessary workload is complete (e.g., by making final charging decisions on cases for which the police department's summary packet has been received). When extra capacity becomes available—as occurs on most days, according to the prosecutor's office—reviewing attorneys would begin an early review of the cases most likely to be dismissed, in descending order of dismissal probability. Such scores could be automatically generated, or even tallied in a matter of minutes, by individuals who manage the intake process. This is in contrast to a full charging decision review, which typically takes attorneys 1-2 hours per case. Thus, we estimate our model would save intake attorneys a significant amount of time ranking and highlighting cases to be reviewed.

If the attorney believes the case is indeed likely to be ultimately dismissed, they would request the sheriff release the arrestee that evening. After receiving the police department's summary packet for that case (typically a day or two later), the reviewing attorney would again review it and make a final charging decision. If, based on the new information, the reviewing attorney decides charges should be brought, a warrant could be issued and the individual re-arrested by the police.

Our proposed strategy hinges on having accurate model predictions of dismissal probabilities for each case. As discussed above, we would also like our model to be simple and interpretable, both to foster trust in the system and to ease adoption. Although our simple model is only designed to rank cases, the order in which cases are reviewed would still affect incarceration time. For example, many cases ranked with lower scores may never be reviewed early due to capacity constraints. Therefore, it is important for the model to remain simple and interpretable, so that (as with any impactful policy) it can be thoroughly examined and trusted by users and any interested stakeholder. Simple models may also facilitate implementation, given that they do not require much IT infrastructure—in the extreme, they can be calculated using only pen and paper. We next describe three methods for constructing these predictive models. The first two use traditional machine learning methods (to serve as benchmarks), and the third is a simple, weighted checklist. All three models are trained on data from 2013–2014 and evaluated on data from 2015–2016, with the training set having 10,218 cases and the test set having 8,494 cases.

Factor	Points
Initial points	4
Each previously filed felony case in last 12 months	-1
Gang-related	-1
Elderly victim	-1
Each assault-related charge*	-1
Each drug-related charge	-1
Each theft-related charge	-2
Incident involves exactly 2 charges	-2
Incident involves 3 or more charges	-4

*Does not include domestic violence

Table 1: A simple model for estimating the likelihood a case is dismissed. The prosecutor's office could manually score incoming cases by subtracting points for any matching case attributes that appear in the list. Higher final scores indicate a higher likelihood of dismissal. One can arbitrarily set the initial points to signal which cases should be reviewed early. For example, one could inform attorneys to prioritize cases with a positive final score, and then adjust the initial points to alter the proportion of cases prioritized.

Our benchmark machine learning models use L^1-regularized logistic regression (lasso) and gradient boosted decision trees (GBDT). Boosted models such as GBDT are considered best-in-class for many prediction tasks [4, 20, 24]. Both models predicted the ultimate charging decision for each case based on all information available within the first eight hours after booking. The detailed list of case information features is described in the appendix. We further extract information from the written narratives via regular expression text matching, including types of evidence collected (e.g., video footage of the incident), whether the arresting officer had a body-worn camera, and whether witnesses were interviewed. Finally, we used a 50-dimensional GloVe [23] text embedding of the complete police narrative. On the test set, we found that the lasso and GBDT models achieved 82% and 83% ROC AUC, respectively, indicating good predictive performance.

We next describe our method for creating a simple, interpretable model, which is based on the regress-and-round procedure outlined by Jung et al. [17]. First, we selected a subset of the predictive features used for the complex models through a combination of step-wise feature selection and consultation with domain experts. In particular, we used the number of alleged charges; the number of assault-related charges; the number of theft-related charges; the number of drug-related charges; the number of domestic violence charges; the number of filed felony cases in which the arrestee was involved in the last 12 months; whether the incident involved

a gun or knife; whether the victim was elderly; and whether the incident was gang-related. On this restricted feature set, we fit an L^1-regularized logistic regression model. Following Goel et al. [12], we constrained all model coefficients (except the intercept) to be non-positive, as the presence of any factor in our list of features should decrease the likelihood an arrestee is dismissed. Finally, we rescaled and rounded all the fitted coefficients $\hat{\beta}_i$ (except the intercept) to be integers in the range $[-M, M]$. Specifically, we set $M = 5$, and defined integer weights

$$w_i = \text{Round}\left(\frac{M\hat{\beta}_i}{\max_j(|\hat{\beta}_j|)}\right). \quad (1)$$

Despite its simplicity, the resulting model (shown in Table 1) achieves 75% AUC on the test set. For comparison, recall that the GBDT model trained on all available features, including information extracted from the narrative reports, obtained 83% AUC.

To apply this model, the prosecutor's office could either manually or programmatically score incoming cases, subtracting points for each matching case attribute in Table 1. A higher final score indicates a case is more likely to be dismissed, and so should be prioritized for early review. Depending on their workload each day, reviewing attorneys would start with cases having the highest score and work their way down the list.

We note that assault-related cases tend to receive higher scores—and are thus more likely to be reviewed for early release—than theft-related cases. This pattern illustrates an important point: the model estimates dismissal probability, not case severity. With assault-related allegations, there is often an absence of physical evidence or witnesses to establish guilt beyond a reasonable doubt, leading to dismissal. Similarly, intake attorneys may be more inclined to prosecute cases with elderly victims, given the particularly egregious nature of these incidents. It bears emphasis that our statistical model only identifies cases for early review, rather than automatically determining whether an arrestee should be released. As such, prosecuting attorneys must still carefully evaluate the evidence on a case-by-case basis before determining the appropriate course of action.

5 EVALUATION

Given our statistical model described above, we now carry out a detailed study of its potential effects if deployed. First we consider the calibration of the model across groups defined by race and gender. We then estimate our proposed strategy's effect on incarceration and public safety.

5.1 Model Calibration

We start by examining the mapping from model scores to actual dismissal rates, disaggregated by an arrestee's race. To deal with data sparsity, dismissal rates are estimated via a logistic regression model, fit separately for each group. The top panel of Figure 1 shows that cases with similar scores have similar dismissal rates, regardless of the arrestee's race, an important property of equitable models [6]. For example, across all race groups, cases with a score of 4 are dismissed just under 80% of the time.

The middle panel of Figure 1 shows a similar plot broken down by gender. In contrast to race, we see that among men and women with

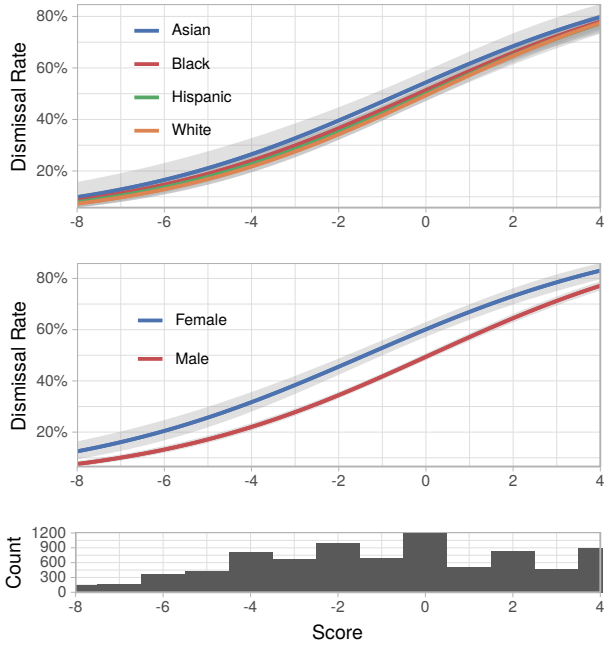

Figure 1: Estimated dismissal rates by model score (see Table 1), disaggregated by an arrestee's race (top) and gender (middle), with gray bands indicating 95% confidence regions. The bottom panel shows the distribution of individuals across scores. Scores appear to be well-calibrated by race; but we also see noticeable gender disparities, indicating that among men and women with the same score, women are dismissed at higher rates.

the same model score, women are dismissed at consistently higher rates. For example, a male arrestee with a score of 2 is dismissed about 65% of the time, but a female arrestee with the same score is dismissed about 73% of the time. Said another way, a female arrestee with a score of -1 is dismissed at about the same rate as a male arrestee with a score of 1.

This gender miscalibration could result from two qualitatively different mechanisms, each having different policy implications. On the one hand, past prosecutorial decisions may have been relatively harsh against men (or, equivalently, relatively lenient toward women). That is, given the exact same facts, prosecutors might have unjustifiably dismissed female arrestees at higher rates than male arrestees. If true, the miscalibration we see would reflect bias in the training data, and in this case, one could potentially use the model in Table 1 to mitigate past prejudice.

On the other hand, controlling for model score, cases involving women might have different fact patterns than those involving men, and may thus be dismissed at legitimately higher rates. For example, among alleged assault cases, those involving female arrestees might, on average, be less severe, prompting prosecutors to dismiss these cases more often. An analogous pattern often occurs with risk assessment scores for recidivism, where women have been found to reoffend less often than men with similar criminal histories [6, 26]. In this case, one solution is to develop gender-specific

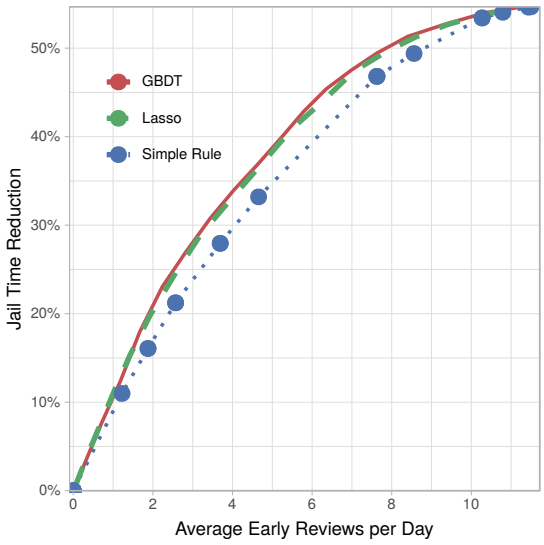

Figure 2: Jail-time reduction for eventually dismissed arrestees as a function of the number of early reviews conducted and the prediction model. The simple rule (in Table 1) performs on par to the lasso and GBDT models. (The simple rule estimates are piece-wise linear, since the model produces discrete scores.) Using simple rules, early review of approximately five cases per day could reduce jail time for eventually dismissed arrestees by approximately 35%.

prediction models, or to simply add points based on gender (e.g., one could assign cases involving a female arrestee two additional points). Though somewhat controversial [28], there is precedent for this approach, with the State of Wisconsin using gender-specific risk assessment tools to guide sentencing decisions. Indeed, the Wisconsin State Supreme Court approved of this choice, writing that "if the inclusion of gender [in the model] promotes accuracy, it serves the interests of institutions and defendants, rather than a discriminatory purpose" [29]. With the data available to us, it is difficult to definitively identify and address the source of the miscalibration. This issue, however, would need to be studied further and resolved before deploying any such tool.

5.2 Jail-Time Reduction & Recidivism

We next estimate the efficacy of early review on incarceration, focusing on jail-time reduction for those arrestees who are never charged with any crime. We consider this subpopulation for three reasons: (1) if individuals who are ultimately charged are released early, they may need to be re-arrested to ensure their appearance at arraignment proceedings, often requiring significant time and effort by law enforcement, tempering the benefits of early release; (2) many charged arrestees are eventually convicted and receive credit for time served during their pre-arraignment detention periods, again moderating the benefits of releasing such individuals early; and (3) the social costs of incarceration may be particularly large for those who are never charged, as it is more likely they did not in fact commit the alleged crime, and they may thus lose more trust in law enforcement if detained for an extended period.

To facilitate estimation, we assume that if a case flagged for early review is eventually dismissed, then the reviewing attorney would also order the arrestee be released after early review. As discussed above, most information is available for nearly all cases by the proposed early review time (i.e., eight hours after booking), bolstering the plausibility of this assumption. We also assume that without such an early review system, arrestees would remain incarcerated while their cases are being considered.

The results of our analysis, under the assumptions above, are shown in Figure 2. To generate the plot, we further assume that each day reviewing attorneys assess all cases above a pre-determined threshold t. For example, they might review all cases with at least $t = 80\%$ chance of dismisssal. This threshold determines one point on each model's curve, each with a corresponding average workload (plotted on the x-axis) and expected jail-time reduction (plotted on the y-axis) at that threshold t. We then trace out curves for all three of our prediction models by varying the review threshold t.

The plot demonstrates two key points. First, the performance of the simple rule (described in Table 1) is quite similar to the performance of the traditional, black-box machine learning models. We can thus obtain the benefits of simplicity and transparency without significantly sacrificing performance. Second, all three models—including our simple heuristic—can significantly reduce jail times. For example, if a prosecutor's office used our simple model in Table 1, we estimate an early review of approximately five cases per day could reduce jail time for eventually dismissed arrestees by about 35%. Similar reductions are achieved within groups defined by race and gender (not shown in the plot).

We note that reviewing attorneys currently handle approximately 22 felony cases per day during the week and 1 felony case per day on the weekends. Based on discussions with our partner jurisdiction, we estimate that the office can handle approximately 5–10 additional cases per day on average (both during the week and on the weekend), in order to respond to spikes in workload. Since prosecuting attorneys must always be ready to handle an influx of cases, they cannot easily assist with activities that require large blocks of time, like litigation. Early review is thus particularly well-suited to make use of spare capacity, and it would not generally displace other work.

We also investigate the potential effects of early review and release on public safety. We specifically examine the effect of extended detention on re-arrest within one year. To start, we note that 33% of those who spend one night in detention recidivate within one year, while only 30% of those who spend two or more nights recidiviate within one year, suggesting a possible public safety benefit of detention. However, those arrestees who spend one night may be qualitatively different than those who are detained for multiple nights—making it difficult to determine the causal effect of detention from such statistics alone.

We account for this issue by using propensity score matching [16]. First, we estimate the probability that each arrestee spends exactly one night in jail, based on all available case and criminal history information—these are the propensity scores. We then match each arrestee who spent two or more nights in jail with one having a similar propensity score who spent one night in jail. After matching,

we find that key covariates are well balanced across the two groups. On the matched sample, we find that one-year recidivism rates are nearly identical: 33% for those who spent one night, and 32% for those who spent at least two nights. The difference in recidivism rates is not statistically significant, with a 95% confidence interval of (-2%, 4%). These results suggest that extended pre-arraignment detention has little impact on recidivism.[2]

6 DISCUSSION

By analyzing a large, detailed dataset of police reports and charging decisions, we developed a statistically informed strategy for reducing pre-arraignment detention. In theory, this proposed intervention could substantially reduce jail time for individuals who are arrested but never charged with a crime, and this reduction could potentially be accomplished without hiring additional staff or adversely affecting public safety. In practice, however, there are several challenging issues that policymakers must consider before undertaking such an initiative, three of which we briefly discuss below.

First, reviewing attorneys may not in fact be able to make early release decisions reliably. Our analysis assumes that attorneys have the requisite information to anticipate eventual charging decisions soon after arrest. Supporting this assumption, we find that detailed information about an incident, including the responding officer's first report, is nearly always recorded and accessible to prosecutors within the first eight hours of booking. Further, our statistical models are able to predict final prosecutorial decisions based on that information with reasonably high accuracy, again indicating that much of the relevant information is available early on. Nonetheless, at least some information—like the police department's final summary packet—is not available until close to the decision deadline, potentially hampering early decision making. On a related note, prosecutors, police officers, and the community more broadly would have to consider the costs of releasing individuals who are later charged, including possible risks to public safety and potential costs for the police department in serving and executing arrest warrants.

Second, policymakers would need to carefully examine the statistical procedure for selecting cases for early review, and ensure that it does not inadvertently recapitulate historical inequities. The simplicity of our proposed model facilitates such an audit. We further find that our simple scoring system is calibrated across race groups, an important first step. However, we also find that women are consistently dismissed at higher rates than men with the same score, pointing to potential problems. As we discussed in the model calibration subsection, this disparity may result from past decisions that were biased against men; alternatively, it could indicate that gender has a legitimately valuable predictive role in identifying cases to review. Before deploying such a system it is important to better understand and address this phenomenon. One strategy is to manually audit cases with the same score to determine whether there are genuine differences in the recorded fact patterns.

Finally, one must consider the potentially complicated impacts of such an intervention on various actors in the criminal justice system, including police officers, prosecutors, perpetrators, and victims. For example, police officers—upon learning the structure of the early review system—may alter their behavior to focus on cases most likely to be prosecuted. On the other hand, even though our model is only designed to rank cases for review, prosecutors' decisions may be influenced by knowledge of this ranking. Although this could have benefits by aligning police and prosecutor decision making, it could also be harmful if certain types of crimes are no longer investigated. For example, knowing that individuals accused of domestic violence may be released early, officers may de-prioritize those cases, in turn harming victims and potentially emboldening abusers. We note, though, that such a cascade effect may never arise, as officers likely already know that such cases are often dismissed, early review still means arrestees spend at least one day in jail, and statutory guidelines mandate certain allegations are always investigated. Prosecutors might also change their behavior, potentially by too quickly releasing arrestees identified via the statistical model for review, rather than carefully examining the facts of each case. In addition to these three challenges, is also possible that prosecutor's offices would not support our proposed policy, as they do not bear the cost of pre-arraignment incarceration, and may support the notion that such incarceration serves a purpose, even if a case is eventually dismissed.

The statistical strategy we outline in this paper marks one of the first instances of an algorithmic intervention for pre-arraignment detention. Our proposed policy could allow jurisdictions to more closely hew to the principle of presumed innocence, reducing incarceration for arrestees who are never charged with any crime, while preserving the opportunity for local law enforcement to intervene where necessary. As with all policy interventions, one must carefully consider both the direct and indirect impacts of deploying a system for early review and release of arrestees. Such attention is doubly important for algorithmic policies, whose effects can be hard to predict, and for policies pertaining to the criminal justice system, which involves particularly vulnerable populations. However, a well-designed and evaluated algorithmic policy to reduce pre-arraignment incarceration can also bring considerable benefits, both for individuals and for communities. We hope this work is a first step toward achieving that goal.

[2]In the Appendix, we conduct *offline policy evaluation* by applying classical sensitivity analysis techniques from the causal inference literature [17]. That analysis likewise indicates that early review and release would not meaningfully affect recidivism, consistent with the matching results described here.

REFERENCES

[1] Richard Berk. 2012. *Criminal justice forecasts of risk: a machine learning approach.* Springer Science & Business Media.

[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2017. Fairness in Criminal Justice Risk Assessments: The State of the Art. (2017). Working paper available at https://arxiv.org/abs/1703.09207.

[3] Ernest W Burgess. 1928. Factors determining success or failure on parole. *The workings of the indeterminate sentence law and the parole system in Illinois* (1928), 221–234.

[4] Tianqi Chen and Carlos Guestrin. 2016. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*. ACM, 785–794.

[5] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 5, 2 (2017), 153–163.

[6] Sam Corbett-Davies and Sharad Goel. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *arXiv preprint arXiv:1808.00023* (2018).

[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 797–806.

[8] Stephanie Holmes Didwania. 2018. The Immediate Consequences of Pretrial Detention: Evidence from Federal Criminal Cases. *https://ssrn.com/abstract=2809818 or http://dx.doi.org/10.2139/ssrn.2809818* (2018).

[9] Will Dobbie, Jacob Goldin, and Crystal S. Yang. 2018. The Effects of Pretrial Detention on Conviction, Future Crime, and Employment: Evidence from Randomly Assigned Judges. *American Economic Review* 108, 2 (February 2018), 201–40. https://doi.org/10.1257/aer.20161503

[10] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[11] Miroslav Dudík, John Langford, and Lihong Li. 2011. Doubly Robust Policy Evaluation and Learning. *ICML* (2011).

[12] Sharad Goel, Justin M Rao, and Ravi Shroff. 2016. Precinct or Prejudice? Understanding Racial Disparities in New York City's Stop-and-Frisk Policy. *Annals of Applied Statistics* 10, 1 (2016), 365–394.

[13] Arpit Gupta, Christopher Hansman, and Ethan Frenchman. 2016. The Heavy Costs of High Bail: Evidence from Judge Randomization. *The Journal of Legal Studies* 45, 2 (2016), 471–505. https://doi.org/10.1086/688907 arXiv:https://doi.org/10.1086/688907

[14] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances In Neural Information Processing Systems*. 3315–3323.

[15] Jennifer L Hill. 2011. Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20, 1 (2011), 217–240.

[16] Guido W Imbens and Donald B Rubin. 2015. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press.

[17] Jongbin Jung, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G Goldstein. 2017. Simple rules for complex decisions. *arXiv preprint arXiv:1702.04690* (2017).

[18] Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. 2018. Algorithmic Decision Making in the Presence of Unmeasured Confounding. *arXiv preprint arXiv:1805.01868* (2018).

[19] Emily Leslie and Nolan G. Pope. 2017. The Unintended Impact of Pretrial Detention on Case Outcomes: Evidence from New York City Arraignments. *The Journal of Law and Economics* 60, 3 (2017), 529–557. https://doi.org/10.1086/695285 arXiv:https://doi.org/10.1086/695285

[20] Zhiyuan Lin, Tim Althoff, and Jure Leskovec. 2018. I'll Be Back: On the Multiple Lives of Users of a Mobile Activity Tracking Application. In *Proceedings of the International World-Wide Web Conference. International WWW Conference*, Vol. 2018. NIH Public Access, 1501.

[21] John Monahan, Jennifer Skeem, and Christopher Lowenkamp. 2017. Age, risk assessment, and sanctioning: Overestimating the old, underestimating the young. *Law and human behavior* 41, 2 (2017), 191.

[22] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy.* Broadway Books.

[23] Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

[24] Byron P Roe, Hai-Jun Yang, Ji Zhu, Yong Liu, Ion Stancu, and Gordon McGregor. 2005. Boosted decision trees as an alternative to artificial neural networks for particle identification. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment* 543, 2-3 (2005), 577–584.

[25] Paul R Rosenbaum and Donald B Rubin. 1983. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society. Series B (Methodological)* (1983), 212–218.

[26] Jennifer Skeem, John Monahan, and Christopher Lowenkamp. 2016. Gender, risk assessment, and sanctioning: The cost of treating women like men. *Law and human behavior* 40, 5 (2016), 580.

[27] Jennifer L Skeem and Christopher T Lowenkamp. 2016. Risk, race, and recidivism: predictive bias and disparate impact. *Criminology* 54, 4 (2016), 680–712.

[28] Sonja B Starr. 2014. Evidence-based sentencing and the scientific rationalization of discrimination. *Stan. L. Rev.* 66 (2014), 803.

[29] State v. Loomis. 2016. 881 N.W.2d 749 (Wis. 2016).

[30] David Word, Charles Coleman, Robert Nunziata, and Robert Kominski. 2008. Demographic aspects of surnames from census 2000. http://www2.census.gov/topics/genealogy/2000surnames/surnames.pdf

[31] Jiaming Zeng, Berk Ustun, and Cynthia Rudin. 2016. Interpretable classification models for recidivism prediction. *Journal of the Royal Statistical Society: Series A (Statistics in Society)* (2016).

Appendices

Appendix A DATA DESCRIPTION AND PROCESSING

We include here additional details about the data we use and our processing pipeline. In the jurisdiction we study, decision dates for a large fraction of cases (54%) are recorded as occurring after the legally allowed deadline. To investigate this issue, we conducted a manual audit of approximately 100 cases, tracking decisions in real-time and comparing them to the electronic record. We determined the discrepancy was due to a data logging issue, with the vast majority of such cases filed shortly before the deadline. We thus assume throughout our analysis that all decisions recorded as occurring late were made on the day of the deadline. We also note that our results are qualitatively similar if we restrict our analysis to only those cases with decision dates officially recorded as occurring by the deadline.

About 15% of the resident population identifies as Hispanic, but only 0.3% of arrestees in our dataset are listed as Hispanic. A manual review of surnames indicated that many ostensibly Hispanic arrestees were instead listed as non-Hispanic white. We addressed this issue by imputing Hispanic ethnicity from surnames, using a dataset from the U.S. Census Bureau that estimates the racial and ethnic distribution of people with a given name [30].[3] We classify an individual as Hispanic if at least 75% of people with his or her surname identify as Hispanic. After this recoding, 16% of arrestees are labeled Hispanic, more in line with expectations.

The statutory deadline for making charging decisions is 48 weekday hours. Many people, however, might be held for longer if weekends or holidays fall during this period s . In particular, as discussed in the main text, over 30% of arrestees who are eventually released stay in custody for over 72 hours.

Appendix B IMPACT ON RECIDIVISM

Our analysis suggests that our proposed early review strategy would not substantially affect recidivism. Below we provide additional details regarding our offline policy evaluation method. We also examine the sensitivity of our results to unmeasured confounding.

[3] http://www.census.gov/topics/population/genealogy/data/2000_surnames.html

Offline policy evaluation. In addition to the matching technique we use to estimate the effect of our proposed early review strategy on recidivism, we further estimate this effect via *offline policy evaluation* [11, 17, 18]. To do so, we first train a GBDT model that predicts recidivism based on all available case covariates and the number of nights an arrestee spent in jail. Then, based on this fitted model, we estimate each arrestee's (couterfactual) recidivism probability if that arrestee were to have spent one night in jail—a straightforward and popular technique called response surface modeling [15]. We now combine these counterfactual estimates $\hat{y}_i(1)$ with the observed recidivism outcomes y_i to assess the overall effect of early review and release. Let π denote a specific early review strategy (e.g., one corresponding to a particular daily workload), with $\pi_i \in \{0, 1\}$ indicating whether arrestee i is ultimately released early. Then the estimated recidivism rate $\hat{R}_\pi$ under policy π is:

$$\frac{1}{N}\sum_{i=1}^{N} \pi_i \hat{y}_i(1) + (1 - \pi_i) y_i, \tag{2}$$

where N is the total number of arrestees. In other words, we simply use the observed recidivism outcome y_i when the policy π does not alter arrestee i's detention status, and we use the estimated recidivism outcome $\hat{y}_i(1)$ when the policy results in i being released early.

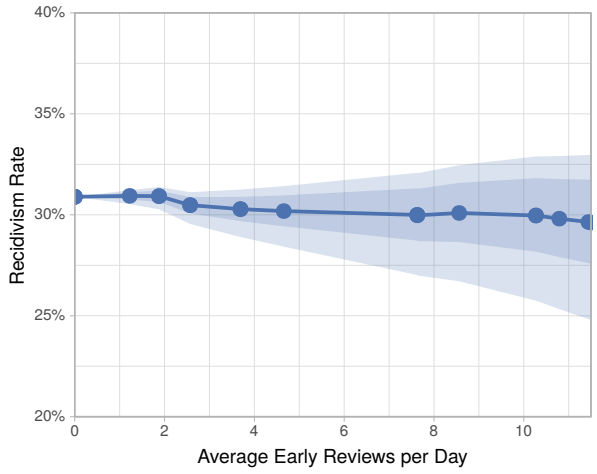

Figure A1: Estimated recidivism rates for various review and release policies, indexed by the average number of cases reviewed each day. Bands indicate plausible ranges of estimates after accounting for unmeasured confounding via sensitivity analysis. The darker band assumes the hidden confounder can alter individual outcomes by a factor of two, and the lighter band assumes the confounder can alter outcomes by a factor of three.

The solid line in Figure A1 shows estimated recidivism rates under various review and release policies derived from the heuristic scoring rule in Table 1. The x-axis indexes such policies by their workload, with higher workloads corresponding to greater jail-time reductions (as shown in Figure 2). The y-axis shows estimated recidivism $\hat{R}_\pi$ under each such policy. The flatness of the line suggests that early review and release would not meaningfully affect recidivism, consistent with our matching results.

Sensitivity analysis. In both our matching analysis and our offline policy evaluation, we implicitly assumed that there was no unmeasured confounding. In theory, however, it is possible that those who spend one night in jail are systematically different from those who spend two or more nights in jail, in ways that are not recorded in the data but which nonetheless affect recidivism. To account for this possibility, we apply the classical sensitivity method of Rosenbaum and Rubin [25], which was recently adapted to the setting of offline policy evaluation by Jung et al. [17]. At a high level, we first assume there is an unobserved covariate $u \in \{0, 1\}$ that affects both an arrestee's detention length and also that individual's recidivism rate. For example, u might indicate whether the arrestee has an extensive criminal record in another jurisdiction, information that is plausibly available to reviewing attorneys—and that might affect detention duration—but is not recorded in our data. We then explore how our recidivism estimates change as we alter three key parameters: (1) the probability that $u = 1$; (2) the effect of u on the arrestee's likelihood to recidivate; and (3) the effect of u on the arrestee's likelihood to spend exactly one night in jail.

The bands in Figure A1 show plausible ranges of recidivism across policies for two sensitivity regimes. In the first, indicated by the darker band, we assume that the unmeasured confounder u can alter an arrestee's odds of spending one night in jail by a factor of two, and can also alter an arrestee's odds of recidivism by a factor of two. The second, more extreme regime, indicated by the lighter band, assumes that u can alter these quantities by a factor of three. In both cases, we allow $\Pr(u = 1)$ to vary freely from 0 to 1. As expected, the bands widen toward the right of the plot, corresponding to policies that deviate more sharply from the status quo, and whose effects are thus harder to predict. However, for policies that require reviewing approximately five cases per day—as we believe is feasible at current staffing levels—our sensitivity analysis suggests recidivism might at most change by a few percentage points. Thus, even accounting for possible unmeasured confounding, it appears that our proposed early review strategy would not dramatically affect recidivism rates.

Appendix C CASE INFORMATION FEATURES

Case information features used by GBDT and L^1-regularized logistic regression include arrest month, day-of-week, and time-of-day; time to deadline; location(s) where the incident occurred (out of 11 areas); arrestee's age; 1-year arrest and charge history for the arrestee; the full list of alleged charges; the number of alleged charges; the number of assault-related charges; the number of theft-related charges; the number of drug-related charges; the number of filed felony cases involving the arrestee in the last 12 months; whether the incident involved a weapon; type of weapon used (if any); whether the victim was elderly; whether the case is gang-related; the number of incident reports already filed; the number of people involved in the incident; and the length of the police narrative.