Personalized Risk Assessments in the Criminal Justice System

Sharad Goel, Justin M. Rao and Ravi Shroff*

Decisions throughout the criminal justice system suffer from real and perceived inequities and inefficiencies. In an effort to address these concerns, judges and correctional officers have increasingly turned to statistical risk assessment tools to help improve bail, sentencing, and parole decisions. At a high-level, it is straightforward to develop and apply these tools. Historical, individual-level data are first used to fit a predictive model to estimate the likelihood of a particular adverse outcome. The fitted model is then used to compute personalized risk scores, which are in turn used to inform decisions.

Statistical risk assessment has a long history in criminal justice, dating back to parole decisions in the 1920's. More recently, Berk et al. (2009) developed a risk assessment tool for the Philadelphia Adult Probation and Parole Department (APPD) to sort offenders into three categories: the highest-risk offenders are considered likely to commit a subsequent violent offense; medium-risk offenders are equally likely to commit a new crime, but in a nonviolent way; and low-risk offenders are unlikely to break the law again. These risk categories determine the terms of parole, with, for example, those deemed most risky required to check in with their case officers more often than those considered low-risk. A randomized controlled trial by the APPD established that the system reduced the burden on parolees without significantly increasing rates of reoffense. Similarly, judges in more than two dozen jurisdictions throughout the United States now use a risk assessment tool to determine which defendants to detain and which to release prior to trial. Developed by The Laura and John

*Goel: Stanford University, Stanford, CA 94305 (e-mail: scgoel@stanford.edu); Rao: Microsoft Research, Redmond, WA 98052 (e-mail: justin.rao@microsoft.com); Shroff: New York University, New York, NY 10012 (e-mail: ravi.shroff@nyu.edu). We thank Richard Berk and Maya Perelman for their helpful comments.

Arnold Foundation, and based on a statistical analysis of 1.5 million cases, the tool identifies nine factors that predict whether a pretrial defendant will commit a new crime or will fail to reappear in court if released before trial.

Here we examine New York City's stop-and-frisk program, and propose two new use cases for personalized risk assessments. First, we show that risk assessment tools can help police officers make considerably better real-time stop decisions. Whereas traditional applications rely on detailed administrative data—such as an individual's criminal record—patrolling officers have little information to work with, and so it is perhaps surprising that a statistical approach is effective in this setting. Second, we show that risk assessment tools can not only be used to make upcoming decisions, but also to audit past actions. We argue that a significant fraction of New York City police stops were conducted on the basis of relatively weak evidence, in possible violation of constitutional protections. For a more complete treatment of these ideas, see Goel, Rao and Shroff (2016).

1 Improving Police Decisions with Real-Time Risk Assessments

Over the last decade, New York City residents have been stopped and briefly detained by the police millions of times in an effort to get weapons, drugs, and other contraband off the streets. A common complaint against the city's controversial stop-and-frisk program is that the tactic is used indiscriminately, with little focus on stopping those most likely to be engaged in criminal activity. To illustrate the potential benefits of personalized risk assessment tools for law enforcement, we outline and evaluate a statistically informed procedure for making real-time stop decisions. This work builds on past risk assessment tools for aiding officers in the field (Berk, He and Sorenson, 2005).

After each police stop, NYPD officers are required to record various aspects of the encounter, including demographic characteristics of the suspect, the time and location of the

stop, the suspected crime, the rationale for the stop (e.g., whether the suspect was wearing clothing common in the commission of a crime), and the eventual outcome (e.g., whether the suspect was arrested). Responses are subsequently standardized, compiled, and released annually to the public. Here we consider the 2.9 million stops recorded between January 1, 2008 and December 31, 2012, focusing on the 760,502 instances in which an individual was stopped for suspicion of criminal possession of a weapon (CPW), by far the most commonly occurring suspected crime in our dataset.

To build a risk assessment tool, we fit a logistic regression model to the historical stop-and-frisk data, using information available immediately before a CPW stop is made to estimate the likelihood a suspect has a weapon. Specifically, we include in the model indicator variables for the suspect's demographics (sex, race, and build); whether the stop occurred on public transit, in public housing, or neither; whether the stop occurred inside or outside; the date and time of the stop (month, day of week, and time of day, binned into disjoint four-hour blocks); one or more recorded reasons for the stop (e.g., "furtive movements" and "high crime area"); whether the stop was the result of a radio run; and whether the officer was in uniform. We additionally include continuous variables for the year, suspect's height, weight, and age, and the time for which the officer observed the suspect before stopping him or her. We also include two location-specific features: (1) indicator variables for the precinct where the stop occurred; and (2) the historical local hit rate of stops (i.e., the percentage of CPW stops in the vicinity during the previous year that turned up a weapon). Finally, we include in the model all pairwise interactions between these variables, resulting in 7,705 predictive features. To avoid look-ahead bias, we train the model on the 301,513 CPW stops from 2009–2010 (with data from 2008 additionally used to compute local hit rates), and then generate out-of-sample predictions for the 288,158 CPW stops from 2011–2012. Given the large number of stops and features, we fit the logistic regression model with stochastic gradient descent, a highly scalable method popular in the machine learning community for its speed and low use of memory.

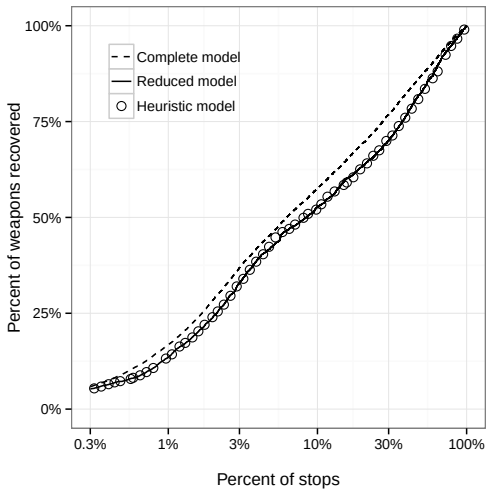

Figure 1: Percent of weapons recovered as a function of the percent of stops conducted, under three risk assessment models. A small number of well-chosen stops is sufficient to recover most weapons.

This procedure yields the following family of stop rules. For any threshold $p > 0$, stop an individual if: (1) the individual would have been stopped under the usual stop-and-frisk practice; and (2) the probability of recovering a weapon, as estimated under the model, is at least p. The first condition is important since the model is trained only on stops that in fact occurred, and so it may not generalize to the population at large. One can thus think of this strategy as a two-step procedure, where an officer first relies on his or her usual training to determine whom to possibly stop, and then checks whether the model-estimated risk exceeds a pre-specified threshold, set perhaps by the city or police department. The higher the threshold p, the fewer people stopped, but also the fewer weapons recovered.

To evaluate the performance of this approach, we first use the model (trained on 2008–2010 data) to estimate the ex-ante likelihood that each CPW stop in 2011–2012 would turn up a weapon. We then rank stops in descending order by this likelihood, with the stops deemed most likely to result in finding a weapon accordingly appearing at the top of the list. Because we know whether or not a weapon was ultimately found on the suspect, we can estimate how many weapons one would have recovered had only the top x-percent of stops been conducted.

The dashed line in Figure 1 shows this relationship, plotting the percent of stops conducted against the number of weapons recovered, where we normalize the number of recovered weapons on the y-axis by the total number of weapons recovered in all CPW stops during this period. Remarkably, only 6% of stops are needed to recover 50% of weapons, and only 58% are necessary to recover 90% of weapons. Because the vast majority of CPW stops have little chance of turning up a weapon, one can significantly curtail the stop-and-frisk program while still reaping many of its benefits.

2 From Complex Models to Simple Heuristics for Decision-Making

The stop strategy outlined above is conceptually simple but may be difficult to implement in practice. Officers cannot simply evaluate a complex statistical model in their heads when deciding whether or not to stop a suspect (although technology, such as a handheld computer, could help with this). Further, it seems unlikely that police departments would adopt opaque machine learning models to inform stop decisions.[1] To address these difficulties, we draw on a large body of work that has found simple, transparent, and interpretable heuristics often work as well as complex statistical models (Gigerenzer and Todd, 1999).

As before, we start by using logistic regression to estimate the likelihood of recovering a weapon in a CPW stop. This time, however, we use only the 18 normalized stop circumstances officers already consider (e.g., "suspicious bulge" and "furtive movements"), indicator variables for each of the 77 precincts, and indicator variables for the three location types (public housing, transit, and neither); we do not include interactions. To further reduce model complexity and increase interpretability, we constrain the 18 coefficients corresponding to stop reasons to be non-negative. This non-negativity constraint captures the

[1] Aside from being generally difficult to understand and explain, there may even be legal hurdles to employing such "black box" methods (Ferguson, 2013).

intuitively reasonably assumption that—all else equal—the 18 stop factors only increase the likelihood an individual has a weapon. Using the 2009–2010 CPW data, we thus fit the *reduced model*,

$$\Pr(y_i = 1) = \text{logit}^{-1}\left(\sum_{j=1}^{18} \alpha_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}\right) \tag{1}$$

where y_i indicates whether a weapon was found, $a_{j,i}$, $b_{k,i}$ and $c_{\ell,i}$ indicate stop reason, precinct and location type, respectively, and we constrain $\alpha_j \geq 0$. We find that only 5 of the 18 stop circumstances have positive weight (the remaining 13 are identically zero): (1) suspicious object; (2) sights and sounds of criminal activity; (3) suspicious bulge; (4) witness report; and (5) ongoing investigation. Notably, all five circumstances are directly tied to criminal activity, and the more subjective conditions (e.g., "furtive movements") drop out of the model.

This reduced model is more transparent and interpretable than the complete statistical model of Section 1, but it is still cumbersome to evaluate on the fly. We simplify the expression in two steps. To implement a threshold-based stopping procedure—as in Section 1—we need not compute the actual probability of recovering a weapon, but can instead compute a stop score that is monotonically related to the probability. Accordingly, our first simplification is to ignore the inverse logistic transformation in (1), and assign to each stop a score equal to the sum of the relevant coefficients. Second, we round the five coefficients for the stop circumstances to the nearest integer (we leave the precinct and location-type coefficients unaltered). This rounding results in three non-zero coefficients for the stop reasons: "suspicious object" (value = 3), "sights and sounds of criminal activity" (value = 1), and "suspicious bulge" (value = 1). Letting $\tilde{\alpha}_j$ denote the rounded coefficients, and reindexing $\tilde{\alpha}_j$ so that the first three terms correspond to the non-zero values, the score S_i for the i-th stop is,

$$S_i = \sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} + \sum_{k=1}^{77} \beta_k b_{k,i} + \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}.$$

Finally, suppose we have selected a stop threshold T, then the stop condition $S_i \geq T$ is equivalent to

$$\sum_{j=1}^{3} \tilde{\alpha}_j a_{j,i} \geq T_r$$

where

$$T_r = T - \sum_{k=1}^{77} \beta_k b_{k,i} - \sum_{\ell=1}^{3} \gamma_\ell c_{\ell,i}.$$

Accordingly, to assess the likelihood a potential CPW stop will lead to the recovery of a weapon, officers simply need to add at most three small, positive integers, and check whether their sum exceeds a fixed threshold T_r for the area they are patrolling. Since officers commonly patrol only a single area during a shift, this procedure is particularly straightforward to carry out in practice.

The tradeoff between the number of people stopped and the number of weapons recovered is controlled by the threshold T, which in turn determines area-specific thresholds T_r. Figure 1 shows this relationship. Despite its simplicity, the heuristic stop strategy performs surprisingly well, with just 8% of stops required to recover 50% of weapons. The heuristic approach does function somewhat worse than the complete model of Section 1—which requires only 6% of stops to recover the majority of weapons—but in exchange it offers greater transparency, interpretability, and ease of use.

3 A Statistical Approach to Fourth Amendment Violations

Risk assessment tools are designed to help make better decisions (e.g., whether or not to stop and question an individual). These same methods, we now argue, can also be used to audit past actions.

Due to the Fourth Amendment protection against unreasonable search and seizure, police stops must be based on "reasonable suspicion" (i.e., specific, articulable facts that would lead

a reasonable person to suspect criminal activity is afoot). As a point of contrast, random searches, regardless of whether they are an effective deterrent, are generally prohibited outside of circumscribed contexts. In the landmark stop-and-frisk case, Floyd v. City of New York (2013), legal experts hand-classified each potential stop justification as reasonable or not. That analysis resulted in 6% of all stops (including non-CPW stops) classified as unreasonable, but the presiding judge believed the classification overly conservative and suggested the true number of stops lacking reasonable suspicion was likely considerably higher.

Statistical risk assessments offer an alternative, intriguing possibility for directly determining whether stops are justified. Namely, one can use a predictive model to summarize the available information in terms of the likelihood of stop success, and then interpret "reasonable suspicion" to mean this ex-ante likelihood is suitably high (above, say, 1%). Taking this approach, we find that 43% of CPW stops had less than 1% chance of turning up a weapon. Moreover, we find striking racial disparities. Whereas 49% of blacks stopped under suspicion of CPW had less than 1% chance of in fact possessing a weapon, the corresponding fraction for Hispanics is 34%, and is just 19% for stopped whites. The courts have yet to quantify the standard of reasonable suspicion in terms of precise probabilistic thresholds, and so it is unclear whether these stops indeed violate the Fourth Amendment. At the very least, though, our results indicate that a substantial fraction of stops—particularly those involving blacks and Hispanics—were conducted on the basis of relatively little evidence.

4 Discussion

To date, risk assessment tools primarily have been applied to bail, sentencing, and parole decisions. Such statistical methods, however, have the potential to bring greater efficiency, equity, and transparency to the criminal justice system more broadly. In addition to guiding and auditing police stops, risk assessments could, for example, inform decisions on which cases to prosecute, which evidentiary tests to run, and which plea bargains to offer.

The use of risk assessment tools raises complex legal and ethical questions. At the heart of many of these issues is whether such methods adequately capture each individual's specific circumstance. On these grounds, Eric Holder, former Attorney General of the United States, has been critical of risk assessment tools, arguing that "[e]qual justice can only mean individualized justice, with charges, convictions, and sentences befitting the conduct of each defendant and the particular crime he or she commits." Holder further warns that insufficiently personalized risk assessments—those, for example, that rely only on one's education level, socioeconomic background, or neighborhood—may exacerbate unjust and unwarranted disparities. Importantly, even statistical tools that perform well overall may still fail to properly consider the relevant information in each and every case, potentially violating an individual's due process rights. To mitigate such worries, the courts have generally held that risk assessment tools may inform, but not completely replace, judicial determinations. Of course, it is not clear that human discretion improves decisions (Danziger, Levav and Avnaim-Pesso, 2011), and such intervention may even plausibly lead to worse results.

Aside from the legal and ethical challenges, risk assessment tools face myriad technical and practical obstacles. Most notably, risk models are almost always trained on non-representative data, and so may not generalize to the broader population to which they are applied (Hajian and Domingo-Ferrer, 2013). For example, to estimate the likelihood of a defendant failing to appear at trial if released on bail, one may only consider data on those who were indeed released, and this endogeneity could in turn lead to spurious assessments for those who were not. Similarly, as populations change over time, static decision-making rules may lose their efficacy. Defendants, for example, may learn to game standardized risk assessment tools. Finally, the training data may contain systematic errors, intentional or not, which may lead to biased predictions. To wit, there is compelling evidence that New York City police officers do not always accurately record the basis for their stops.

Despite these hurdles, risk assessment tools can help improve and evaluate decisions throughout the criminal justice system. Looking ahead, we hope scholars and practitioners

continue to explore and adopt such statistical methods.

References

Berk, Richard, Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay Ahlman. 2009. "Forecasting Murder within a Population of Probationers and Parolees: A High Stakes Application of Statistical Learning." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 172(1): 191–211.

Berk, Richard, Yan He, and Susan Sorenson. 2005. "Developing a Practical Forecasting Screener for Domestic Violence Incidents." *Evaluation Review*, 29(4): 358–383.

Danziger, Shai, Jonathan Levav, and Liora Avnaim-Pesso. 2011. "Extraneous Factors in Judicial Decisions." *Proceedings of the National Academy of Sciences*, 108(17): 6889–6892.

Ferguson, Andrew Guthrie. 2013. "Predictive Policing and Reasonable Suspicion." *Emory Law Journal*, 62(2): 259.

Gigerenzer, Gerd, and Peter M Todd. 1999. *Simple Heuristics that Make Us Smart.* Oxford University Press, USA.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016. "Precinct or Predjudice: Understanding Racial Disparities in New York City's Stop-and-Frisk Policy." *Annals of Applied Statistics*.

Hajian, Sara, and Josep Domingo-Ferrer. 2013. "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining." *Knowledge and Data Engineering, IEEE Transactions on*, 25(7): 1445–1459.