NBER WORKING PAPER SERIES

PREDICTING POLICE MISCONDUCT

Greg Stoddard
Dylan J. Fitzpatrick
Jens Ludwig

Working Paper 32432
http://www.nber.org/papers/w32432

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
May 2024

This project was supported by the Irving Harris Foundation, Griffin Catalyst, the Joyce Foundation, the Charles Koch Foundation, the MacArthur Foundation, the Motorola Solutions Foundation and the U.S. Department Of Justice's Bureau of Justice Assistance. We are grateful to the Chicago Police Department as well as the Independent Monitoring team for the Chicago Police Department's consent decree with the Illinois Attorney General for their partnership and guidance, particularly Barbara West, Bob Boik, Jack Kenter and Rick Peplinski. This work was made possible by support from the University of Chicago Crime Lab, especially Roseanna Ander, Anthony Berglund, Ellen Dunn, Nicole Gillespie, Maggie Goodrich, Katie Larsen, Sandy Jo MacArthur, Zoe Russek, and Diamond Thompson. Thanks to Danielle Allen, Hye Chang, John Greer, Emma Nechamkin, and Alex Williamson for their outstanding data science work. For thoughtful feedback we thank Aaron Chalfin, Oeindrila Dube, Zubin Jelveh, Max Kapustin, John Rappaport, Kyle Rozema, and Max Schanzenbach. Findings, opinions and any errors here are those of the authors alone and do not necessarily reflect the views of the Chicago Police Department. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

Predicting Police Misconduct
Greg Stoddard, Dylan J. Fitzpatrick, and Jens Ludwig
NBER Working Paper No. 32432
May 2024
JEL No. C0,K0

ABSTRACT

Whether police misconduct can be prevented depends partly on whether it can be predicted. We show police misconduct is partially predictable and that estimated misconduct risk is not simply an artifact of measurement error or a proxy for officer activity. We also show many officers at risk of on-duty misconduct have elevated off-duty risk too, suggesting a potential link between accountability and officer wellness. We show that targeting preventive interventions even with a simple prediction model – number of past complaints, which is not as predictive as machine learning but lower-cost to deploy – has marginal value of public funds of infinity.

Greg Stoddard
University of Chicago Crime Lab
190 S. LaSalle, Suite 2600
Chicago, IL 60603
gstoddard@uchicago.edu

Dylan J. Fitzpatrick
University of Chicago Crime Lab
190 S. LaSalle, Suite 2600
Chicago, IL 60603
djfitzpa@uchicago.edu

Jens Ludwig
Harris School of Public Policy
University of Chicago
1307 East 60th Street
Chicago, IL 60637
and NBER
jludwig@uchicago.edu

Section 1: Introduction

How can policing in America be improved? That question has always been important for a sector with an annual budget of over $100 billion and that employs over 700,000 officers. But its importance has only increased on the heels of a widely publicized series of police uses of force, particularly against Black Americans, new evidence of racial bias in policing and other criminal justice decisions (Arnold et al. 2020 , Fryer 2020 , Goncalves and Mello 2021, Hoekstra and Sloan 2020), declining trust in police (Washburn 2023, Nadeem 2022), low morale among officers,[2] and rising gun violence (CDC 2020, Gramlich 2023), all collectively culminating in growing calls for change (Olander 2023, Subramanian & Arzy 2021, Rogers & Kanno-Youngs 2021).

Most of the public discussion seems to focus on what to do *after* some tragedy has happened. For example, after the killing of Laquan MacDonald by former Chicago Police Department (CPD) officer Jason Van Dyke, 90% of the articles published in the *Chicago Tribune* focused on some *after-the-fact* issue like a possible cover-up, how long it took the case to go to trial, or whether the police union should have given Van Dyke a job. Only 22% of articles included any mention of whether this tragedy could have been *prevented* in the first place.

Whether these tragedies are *preventable* surely depends at least in part on the degree to which they are *predictable*. Some previous studies suggest there may be some predictive signal in police administrative data about risk of future misconduct (Carton et al., 2016, Rozema and Schanzenbach 2019), which in principle means a predictive algorithm could be used to target preventive interventions. But there remains debate about the benefits and costs of such systems. One concern is reliance on administrative data collected and maintained by police departments themselves. Are we learning something about police misconduct, or instead about problems with

[2] https://www.npr.org/2021/06/24/1009578809/cops-say-low-morale-and-department-scrutiny-are-driving-them-away-from-the-job

the police data? Is the level of predictability high enough to be of any practical policy value (Chalfin and Kaplan, 2021)? Is what looks like 'misconduct risk' simply a proxy for 'activity' (Worden, Harris, McLean 2014; Rozema and Schanzenbach 2019)? Are there any interventions actually capable of preventing misconduct that are worth targeting? And are the benefits from predictive targeting of interventions large enough to justify the resources required to deploy such systems (Walker, Alpert and Kenney, 2001)? This is all set against a backdrop of many civil rights organizations expressing deep skepticism about the use of predictive algorithms in any area of public policy, but particularly within the criminal justice system.[3]

This paper seeks to answer these open questions and consider the potential for social impact from predicting police misconduct. We draw on detailed data from CPD obtained as part of an effort by our research center, the University of Chicago Crime Lab, to help implement elements of the consent decree between CPD and the Illinois Attorney General's Office.[4] Specifically, the consent decree requires that CPD implement an early intervention system (EIS), which uses data to identify officers at elevated risk of misconduct. For that purpose we apply different statistical models, including machine learning, to police administrative data.[5] Our main conclusion is that policies to predict and prevent police misconduct wind up comparing favorably to other candidate policy interventions on criteria like the marginal value of public funds (MVPF) (Hendren and Sprung-Keyser, 2020), as seems to often be the case with algorithmic policies (Ludwig, Mullainathan and Rambachan, 2024a,b).

We first find that police misconduct does indeed have some predictable structure. We examined two types of misconduct: on-duty events (like sustained complaints of excessive force)

[3] For example https://civilrightsdocs.info/pdf/criminal-justice/Pretrial-Risk-Assessment-Full.pdf
[4] The use of an EIS is typically required by the US Department of Justice in consent decrees. The decision of what risk factors to include and how to weigh them is typically informed by a combination of expert judgment and legal negotiations between various stakeholders, such as city leaders, police unions, and police department leadership.
[5] The methods and results presented in this paper are similar but distinct from the methods implemented in the Chicago Police Department. All of the results from this paper are qualitatively similar to what was done in practice.

and off-duty events (complaints of domestic violence, off-duty altercations, etc). The level of predictability of police misconduct is similar to what we see for predicting other human behavior.[6] For example, the highest-risk officers are indeed at greatly elevated risk: Those in the top 1% of the predicted risk distribution are 6.7 times more likely for on-duty misconduct than the average officer, and 6.2 times more likely to have off-duty misconduct. However we also confirm the Chalfin and Kaplan (2021) finding that the highest-risk officers account for a modest share of all misconduct. It turns out just a modest share of officers are very high risk, so the large majority of other officers (who are not zero risk) account for a large share of total outcomes – a version of the 'prevention paradox' from epidemiology.[7]

This predictability does not seem to be simply an artifact of measurement error in the police data. We capitalize on a 'natural experiment' that seems to have increased the quality of police misconduct data after 2016, as reflected by an increase in the share of citizen complaints against officers that are sustained. While an algorithm trained during the post-2016 period (higher-data-quality) captures more signal, even an algorithm trained during the pre-2016 period still seems to capture useful signal as evaluated using the higher-quality post-2016 data.[8]

What predicts misconduct? The key driver of risk turns out to be an accumulated pattern of prior events (even if some of those events are seemingly minor) rather than having a single serious prior event. That is, information about the severity of a prior event - such as whether a misconduct complaint was sustained or the amount of money associated with a lawsuit payout - adds modest signal over simple counts of prior events. Relatedly, even past *unsustained*

---

[6] For instance, a study by Desmarais, Zottola and Lowder (2020) of six pretrial risk assessments that try to predict defendant behavior (re-arrest, skipping court) found AUC values that range from .65 to .73. Chouldechova et al (2018) find an AUC of .8 when predicting whether a child will be removed from a home during a child welfare investigation. Hastings, Howison, and Inman (2019) find an AUC of .8 predicting risk of prescription opioid misuse.

[7] For instance, even with a machine learning model built on an extensive set of risk factors and a dataset spanning ten years, the top 5% of officers (roughly 600 officers) by predicted risk only account for around 22% of all on-duty misconduct that occur over a two year period.

[8] We also show below that this is not merely a mechanical artifact of the different base rates across periods.

complaints have signal about future misconduct; that is, the predictive model implies that officers who are accumulating a number of unsustained complaints are at high risk of a sustained complaint in the future. These findings raise normative policy questions analogous to those arising when predicting criminal behavior among private citizens (should we use just convictions, or also arrests?) and have implications for what events might be captured by police administrative data systems or registries that help vet candidates for lateral moves (e.g. Grunwald and Rappaport, 2020).

These findings also point to ways of reducing the costs of building and deploying a predictive risk model without losing substantial amounts of predictive signal. We show that even simple summary statistics of an officer's past events - like a count of prior complaints the past two years ('rank by complaints') – capture substantial signal about risk of on-duty misconduct. This leads to the optimistic practical conclusion that resource-constrained departments unable to invest the funding (or the time) to build their own algorithm can still capitalize on the benefits of targeting preventive interventions. This may be particularly important for smaller departments, which otherwise would find it difficult to build machine learning models both because of the fixed costs (invariant to department size) and limited sample of officers to train algorithms on. This is not a trivial issue given that around 60% of all police killings of civilians in the US happen in departments with fewer than 500 officers.[9]

A different type of potential cost with such risk models is the concern that what looks like 'risk' may simply be a proxy for 'activity' (Worden, Harris, McLean 2014; Rozema and Schanzenbach 2019). The concern is that officers will alter their behavior to avoid being labeled

[9] This statistic was computed using data from Mapping Police Violence (https://mappingpoliceviolence.org) on police killings and the 2014 Law Enforcement Management and Statistics (LEMAS) survey conducted by the Bureau of Justice Statistics. Code to replicate this result can be found with our replication materials https://github.com/uchicago-urbanlabs-crimelab/predicting-police-misconduct.

high risk, and purposely avoid police activities that society wants and expects. Without taking a normative position about what policing 'outputs' society *should* want, we show that less than half of the variation in predictable risk of on-duty misconduct can be attributed to the most commonly used and widely available measures of policing activity (arrests, guns recovered, etc). That is, while many high-risk officers have high amounts of activity, most high-activity officers are not high risk. Removing the correlation between risk and activity has a fairly modest effect on the identification of which officers are high-risk.

A different potential cost is that of algorithmic bias in the risk models, e.g. that the risk models exacerbate possible bias in reporting misconduct by officers from different race or ethnic groups. A common test for bias is calibration – that is, whether the relationship between predicted and actual values of the outcome looks similar across groups. We find the algorithm passes this type of calibration test for our measures of both on-duty and off-duty misconduct.

The ability to accurately and fairly predict risk is only useful if there are effective interventions that could be targeted with predictive algorithms, which we show there are - including behavioral-science-informed training (Owens et al 2018; Dube, MacArthur & Shah 2023). We also show that risk for on- and off-duty misconduct turns out to be highly correlated.[10] This suggests the hypothesis that efforts to prevent on-duty misconduct might benefit from interventions that not only directly target job-related risk factors, but target off-duty risk factors like trauma or substance abuse as well (Asmundson and Stapleton, 2008, Ziobrowski et al 2023).

We now have the building blocks to quantify the social welfare gains of targeting some preventive intervention using predicted misconduct. We calculate the marginal value of public funds (MVPF) from Hendren and Sprung-Keyser (2020), defined as the public's willingness to

[10] Rozema and Schanzenbach (2016) also show that on-duty and off-duty events are correlated. Our result builds on that finding by showing that the *predictable* part of each outcome (as opposed to idiosyncratic component) is correlated, which has implications for the ability to identify ex ante officers at risk for both types of misconduct.

pay for the policy divided by the net cost to government. We assume a training intervention capable of reducing misconduct by 20% (see Dube et al., 2023). We show that relative to random targeting, a very simple prediction model - rank officers by past complaints - yields an MVPF value of infinity: positive benefits to society from reduced misconduct winds up saving the government money (reduced lawsuit payouts and investigation costs), a 'free lunch' partly because the model itself is so low-cost to build and deploy. Whether similarly favorable MVPF values hold with a full-blown machine learning model depends partly on a policy decision that can't be answered absent a real-live deployed algorithm (what share of officers the department decides to flag as high risk) and partly on a currently unknown parameter (how many departments and time periods the prediction algorithm could generalize to).

There are naturally a large number of additional open questions about the exact benefits and costs of different specific types of predictive models and preventive interventions to reduce police misconduct, which are beyond the scope of our paper to answer. But the findings presented here suggest at the very least that these questions are worth exploring.

Section 2: Data and Methods

A. Data

We worked with CPD to assemble a data extract from Chicago's administrative data systems[11] that provided a rich set of factors about an officer's activity and assignments, including: citizen complaints (which include information about the officer being complained against as well as the nature of the alleged misconduct; these may include for instance excessive use of force, false arrest, etc.), internal complaints (filed by CPD supervisors against officers,

[11] The dataset used for this research was provided to us under data sharing agreements with the Chicago Police Department as part of our effort to help them build and implement an early intervention system for officers at high risk of misconduct. While we cannot reshare this data, we have replicated the main results of this analysis using public data from the New York City Police Department. The data and code for that replication can be found at https://github.com/uchicago-urbanlabs-crimelab/predicting-police-misconduct

which may include failure to fill out a report or insubordination), use of force reports (which are supposed to be filled out for all uses of force ranging from emergency handcuffing up to use of their service weapon), attendance, measures of activity including arrests and drugs and guns confiscated, a record of where each officer has worked, and a measure of how long they have been on the job.

From these datasets we created explanatory variables, or 'features', that measure an officer's history of use of force, complaints, arrests, and attendance over the past one, two and five years. We used different time horizons to allow the models to weigh prior events differently depending on when they occurred and to handle cases where officers had less than five years of history at the time of prediction. We also included features for an officer's most recent assignment (e.g. what unit they work in and what their role is) and their years of experience as of the time of prediction. See the appendix for more details on feature construction.

Using this data, we assembled a panel dataset that covers the period from 2010 to 2018, where each year includes observations for all officers active in that year.[12] The total number of person-year observations in our data is N=113,768, with each year having 12,000-13,000 officers. Each year's observations includes officers that were active in that year, their risk factors from the prior five years, and a set of outcome measures based on misconduct that an officer was involved in over the next two years.[13] We refer to the years leading up to and including year T as the *observation period*, and to years T+1 and T+2 as the *outcome period*. For example, the 2012 observation period measures an officer's activity up through the end of 2012 and includes

[12] We use the term 'officer' in the colloquial sense, meaning any sworn employee (i.e. non-civilian). The decision to use the entire sample of the police department, as opposed to restricting the sample by years of experience or nature of assignment, differs from some prior work (Chalfin and Kaplan 2021, Rozema and Schanzenbach 2019). In the appendix we show that our results are not sensitive to the choice of using the full population or a subpopulation - in fact, we show that the use of the full population allows us to more accurately predict future misconduct.

[13] In the appendix we show risk models produce highly similar risk rankings with a 1, 2, or 4 year outcome period.

outcome measures for whether the officer had misconduct in 2013 or 2014. See Table 1 for a summary of key statistics of the dataset and machine learning model performance.

The misconduct outcomes that we predict in this paper vary across two dimensions. The first dimension is whether the complaint alleges misconduct that was *on-duty* (allegations of harm while an officer was carrying out a policing function, such as excessive force or a wrongful arrest of excessive force, verbal abuse, an improper stop/search, or that involve a wrongful arrest[14]) or *off-duty* (these include domestic incidents, complaints involving drugs or alcohol, or off-duty altercations).[15] During our outcome period 1.9% of officers are involved in on-duty misconduct and 5% are involved in off-duty misconduct.

The second dimension is whether we define outcomes using *all complaints* of a given type or just *sustained complaints*. A complaint is sustained when the investigation proves the allegations happened and the officer's conduct was out-of-policy (e.g., excessive force). Most complaints are not sustained - between 2010 and 2018 there was an average of 2,915 on-duty complaints per year and 3.1% of them were sustained. Over that same time period, there was an average of 499 off-duty complaints per year with a sustained rate of 21.7%.

We examined the choice of outcome empirically by constructing statistical models to predict each outcome – sustained on-duty, all on-duty, sustained off-duty, and all off-duty – and comparing their performances. The model built to predict all off-duty complaints is a strictly better model than the model that predicts sustained off-duty complaints because it is more accurate at predicting *all* off-duty complaints as well as *sustained* off-duty complaints (details of

[14] The exact list of complaint categories is excessive force, improper arrest or search, verbal abuse, coercion, search warrant incident, arrest and lockup incident, bribery or official corruption, and weapon discharge. In the appendix, we show that our results are not sensitive to whether we use this broader index of activity or if we specifically limit to sustained complaints of excessive force.

[15] The exact list of categories includes domestic incidents, drug and alcohol complaints, conduct unbecoming of an officer violations (while off-duty), criminal misconduct, and sexual misconduct.

this analysis are shown in the appendix). Hence, we focus on the "all off-duty" model and outcome (shortened to just off-duty for brevity) for the remainder of the paper.

The comparison between the sustained on-duty model and the all on-duty model is less clear because neither model dominates the other; the sustained on-duty model is a slightly better predictor of sustained on-duty complaints, while the all on-duty model is a better predictor of all on-duty complaints. One noteworthy feature of the sustained on-duty model is that its risk scores are significantly less correlated with policing activity (a point we discuss more below). In our main exhibits we will focus on presenting results for the sustained on-duty model, but show all main results for the sustained on-duty model and the all on-duty model in the appendix.

B. Methods

We constructed risk models to predict misconduct using all of the available data, essentially asking the machine learning algorithms to estimate the chance that an officer has a misconduct outcome in the next two years[16] based on the factors known about an officer at the time of prediction[17]. We generate predictions (risk estimates) for each observation via *cross-fitting*[18], a technique that iteratively partitions the data into train-test splits so that each observation receives an out-of-sample prediction, i.e. a risk estimate from models that did not use that observation in the training procedure.[19] Specifically, each iteration of the cross-fitting

[16] The date is based on when the incident occurred, not the date that the complaint was filed or the investigatory finding was issued. This is a key distinction because complaint investigations can take years to reach a conclusion.
[17] This "predict the future" set-up, and out-of-sample prediction more generally, differs from prior studies (e.g. Jain, Sinclair, & Papachristos 2022; Cubit 2023) that instead characterize patterns in the data through time T but do not test whether those patterns hold in future time periods or among different officers.
[18] The use of cross-fitting has been recently studied and popularized in research on the application of machine learning in econometrics, e.g. Chernozhukov et al (2018).
[19] The standard evaluation procedure in machine learning is to use an 80/20 train/test split where 80% of the data is used to train the model and the remaining 20% is used to evaluate the model. We deviate from that approach here because reserving only 20% of the dataset size will create small sample issues when we evaluate how well we can flag a small group - less than 5% - of risky officers. The use of cross-fitting enables the use of the entire dataset for evaluation, thus reducing the issues induced by small sample sizes in the evaluation stage.

procedure randomly partitions the dataset into three sets - P1, P2, P3 - by officer ID[20]. We then train three models, each with one partition held-out from the training procedure, denoted m_{-P1}, m_{-P2}, m_{-P3}. To get out-of-sample predictions for each observation $X_{i,t,}$ we use the model in which $X_{i,t}$ was part of the hold-out set, i.e. $m_{-P(i)}$ where P(i) denotes the partition that officer i's observations belong to. Finally, we repeat this procedure J=10 times to reduce error from Monte Carlo variation in the data partitioning. Letting the superscript j denote the models that were generated from iteration j, the final prediction for each observation $X_{i,t}$ can be written as:

$$\hat{p}_{i,t} = \frac{1}{J}\sum_{j=1}^{J} m^{j}_{-P_i}(X_{i,t})$$

We use a machine learning algorithm known as gradient-boosted trees[21] (GBT) that is capable of modeling highly non-linear and interactive functional forms. The building block of GBT is the decision tree, which captures non-linearity and interactions by iteratively splitting the training data into subgroups to maximize the homogeneity of the subgroups with respect to the outcome variable. Trees can fit the data quite well (low bias) but can be too sensitive to learning the idiosyncratic noise in a given dataset (high variance). GBT combats that variance by adding multiple trees together, but rather than simply averaging independently-built trees (as with random forest), GBT sequentially builds trees with each subsequent tree designed to correct the prediction errors of the previous ones. Gradient-boosting is consistently one of the best performing machine learning models (Grinsztajn, Oyallon, & Varoquaux 2022).[22]

[20] Each partition is formed by randomly sampling a third of officers (without replacement) and including all observations from those officers in that partition, so no officer has observations in both train and test.
[21] We use scikit-learn's implementation (Pedregosa et al 2011c) Histogram-based Gradient Boosting Classifier, an extension of the original gradient boosting algorithm that has better computational performance for large datasets. See the appendix for a full-description of the machine learning methods.
[22] We also tested random forests and regularized logistic regression and found that gradient-boosting was the most accurate model. Results are shown in the appendix.

All machine learning algorithms face a model complexity tradeoff– if the models are too complex (e.g. the gradient boosting ensemble has too many trees, the trees use too many subgroups, etc), they can overfit to the training data. If the models are not complex enough, they can fail to capture the true relationship between the covariates and predicted outcome. Model complexity is controlled by a set of hyperparameters (e.g. the number of trees in the gradient boosting ensemble or the maximum number of leaf nodes that any tree can have) that, ideally, should be selected based on empirical risk minimization. More formally, let f denote a type of machine learning algorithm (eg gradient boosting), T denote the training data, θ denote a set of hyperparameters, and f_θ^T denote the machine learning model produced by the training algorithm when using hyperparameters θ and training data T. Then the optimal hyperparameters are the ones that minimize the cross-validated empirical loss, i.e.

$$\theta^* = \underset{\theta \in \Theta}{\operatorname{argmin}} \frac{1}{N} \sum_{i \in T} L(y_i, f_\theta^{T \setminus K(i)}(X_i))$$

The classic approach to solving the above minimization problem is an exhaustive search over hyperparameter configurations but can be infeasible for algorithms that have many hyperparameters (as most modern machine learning algorithms do). Instead, the typical practice is to only search over the 2-4 hyperparameters believed to be most influential (i.e. restricting the size of θ), which creates the risk of choosing sub-optimal hyperparameters and reduced model accuracy (Weerts, Mueller, and Vanschoren 2020).

To ensure that our results are not sensitive to under-optimization, we conduct a sensitivity test in which we use a modern hyperparameter tuning algorithm to optimize over the space of hyperparameter configurations more efficiently. This tuning algorithm, known as FLAML, (fast and lightweight automated machine learning; Wang, Wu, Weimer, and Zhu 2021) searches the

space of hyperparameter configurations through a weighted random sampling technique that proportionally samples and tests new configurations based on an estimate of "accuracy gain per computation time" – decreasing the time required to find a good set of hyperparameters, which means we can optimize over all hyperparameters rather than just a subset. Our sensitivity test shows that the models produced via the standard versus more advanced tuning method are similar, showing our results are not driven by under-optimization (see appendix for more details).

Section 3: Is Risk Predictable?

In what follows we first show there is some predictable structure to both types of misconduct, on- and off-duty, and then present some additional results that try to speak to how much of this predictability is true 'signal' for misconduct versus measurement error.

A. Results

We first measure the predictive accuracy of these risk models using the standard area under the receiver-operating curve (abbreviated to ROC-AUC or AUC). Intuitively, AUC captures the probability that a randomly selected Y=1 case in the new out-of-sample (OOS) validation dataset has a higher predicted value by the algorithm than a randomly selected Y=0 case. A model that produced random predictions (captured no signal at all) would have an AUC of .5, while a perfect predictor would have an AUC of 1. The on-duty risk model has an out-of-sample AUC of .752 and the off-duty risk has an out-of-sample AUC of .682. As noted in the introduction, these AUC values fall within the range of AUC values encountered in other human behavior prediction problems.

A different way to gauge predictive accuracy that is perhaps more intuitive and helps readers 'see' a bit more of the data is to compare predicted to realized misconduct rates across the predicted risk distribution. That is, we estimate the likelihood the officer has the misconduct

outcome in the outcome period for each officer-year observation and then rank officers by their risk estimate within each year. We then bin officer-year observations in a given year by their predicted risk and examine observed misconduct rates in the outcome period. A random (all noise) predictor would essentially randomly rank officers and so produce a uniform misconduct rate across the predicted risk distribution. The better the prediction, the steeper the 'slope' in the relationship between predicted risk and observed misconduct rates.

We find that officers with the highest levels of predicted risk are significantly more likely to have a misconduct outcome than the average officer. The top two plots in Figure 1 show the statistical accuracy of the model that predicts off-duty misconduct (1a and 1b) and the bottom two plots show the accuracy of the on-duty model (1c and 1d). Figures 1a and 1c show that the estimated risk scores are predictive of actual misconduct because the rate of actual misconduct (shown on the y-axis) increases with risk percentile (shown on the x-axis). By contrast, non-predictive risk models would yield a flat relationship between percentiles and outcome rates, with each percentile having an outcome rate at the population average (dashed lines in the chart).

Figures 1b and 1d "zoom in" on just officers in the top 10% of risk and show that officers at the very top of the risk distribution have a greatly elevated rate of misconduct relative to the department average. Figure 1b shows that the top 1% of officers by estimated off-duty risk (approximately 120 officers) have future off-duty misconduct at a rate of 29.9%, which is roughly 6.2x the average rate in the department. Put another way, if we chose a random group of 120 officers each year, we'd only expect about 6 of them to have an off-duty complaint in the following two years; if we chose the 120 officers with the highest risk estimates, we see that closer to 35-40 of them will have an off-duty complaint in the next two years. Figure 1d shows

the top 1% of officers by predicted on-duty misconduct risk engage in future on-duty misconduct at a rate of 12.3%, which is around 6.7x the base rate.[23]

The other key result, shown most clearly in Figures 1b and 1d, is that predictable risk is concentrated in a very small group of officers, and drops off quickly as we move down the ranking of predicted risk. The rate of future off-duty misconduct for officers at the 96th percentile of predictive risk is less than half of the rate of officers at the 99th percentile. As we move lower in the risk distribution, the predicted risk / realized misconduct gradient flattens out significantly. There are just small differences in outcome rates between officers at the lowest end of risk and officers all the way up to the 80th percentile of predicted risk. We find similar patterns for the on-duty model, with the noted exception that the outcome rate at the 99th percentile is not as starkly elevated as with the off-duty model.

The shape of this curve supports the argument from Chalfin and Kaplan (2021) that the officers with the highest estimated risk will only account for a moderate fraction of officers who actually have on-duty or off-duty misconduct. This metric is known as recall, which is the fraction of positive instances (officers who had a misconduct allegation) that were predicted to be positive instances, and is shown for both models in Figure 2. In this context, if we flag officers in the top 5% of predicted risk, recall is the percentage of officers with an on-duty misconduct that were in the top 5% of predicted risk. The recall of a model is defined both by statistical accuracy - how well the available data can predict the event - and the threshold for what constitutes “high-risk”. For instance, Figure 2 shows that if we set the high-risk threshold at the 95th percentile of risk, then the on-duty model would have flagged 22% of the officers who

[23] The evaluation of the future misconduct rates for the on-duty and off-duty risk models are only evaluated against the specific outcomes that they were designed to predict for these calculations. For instance, if an officer flagged for on-duty misconduct engages in future off-duty misconduct, we would not count that as a successful prediction. This narrow evaluation is hence a conservative estimate of the future misconduct rates of higher-risk officers. In the appendix, we evaluate the accuracy of these risk models against a more generalized notion of misconduct and find that 30-40% of officers flagged by either risk model engage in some sort of future misconduct.

actually had an on-duty misconduct outcome. And at that same flagging rate, the off-duty model would have correctly flagged 19% of officers who actually had an off-duty misconduct outcome. The fact that the bulk of people are low-risk, and in aggregate account for most misconduct, is not unique to policing, but rather yet another example of the 'prevention paradox' (Rose 1981). We will show below that risk prediction is useful, but these results make clear it is not a panacea.

B. Measurement error

One concern with these results is that police administrative data do not capture misconduct; they capture reported and recorded misconduct. To see what can go wrong consider the framework from Mullainathan and Obermeier (2017), where measured misconduct equals true misconduct plus measurement error, $Y_i = Y_i^* + \Delta_i$. So algorithmic prediction yields:

$$\hat{Y}_i \approx E[Y_i|X_i] = E[Y^*_i|X_i] + E[\Delta_i|X_i]$$

If the measurement error in misconduct is mean zero and uncorrelated with the predictors, the last term in the equation above equals zero and the result of measurement error is to simply add random noise to the data that reduces predictive accuracy. But there's no guarantee that's the case. Imagine, for example, that the chances that a complaint against an officer is sustained is not random, but instead depends on, say, how much 'clout' the officer has (within the department or city government, etc.) In principle, 'clout' could be even *more* predictable from officer characteristics than misconduct. If 'clout' were highly correlated with the predictors but only weakly correlated with the misconduct outcome (so the last term below is small or even zero) we could inadvertently wind up with a 'clout predictor' rather than a misconduct predictor.

$$Var(\hat{Y}_i) = Var(\hat{Y}^*_i) + Var(\hat{\Delta}_i) + 2\,Cov(\hat{Y}^*_i, \hat{\Delta}_i)$$

To explore this possibility we make use of a 'natural experiment' that plausibly changed the degree of measurement error in our measure of on-duty sustained complaints. Beginning in 2017, the number of sustained on-duty complaints and the likelihood that any given on-duty complaint was sustained rose substantially (these trends are plotted in Figure 3). The best explanation for these changes is an increase in the rate that 'true misconduct' results in a sustained complaint, e.g. an increase in $P(\widehat{Y}=1|Y^{*}=1)$. In principle these changes could alternatively be driven by an increase in the amount of true misconduct but arguing against that possibility is data showing that the number of use of force reports, officer-involved shootings, and total number of complaints also decreased (also shown in Figure 3).[24]

We have, in other words, a natural experiment in which we plausibly have two data periods that seem to differ dramatically in the relative degree of measurement error, Δ. We use this natural experiment to test how much of the variation in estimated risk is driven by true risk versus 'clout' by creating two on-duty risk models - an "early model" which is trained only on the earlier time periods (outcome periods up to and including 2015-2016) and a "late model" which is trained only on the later time periods (outcome periods of 2016-2017 and later). We then compare the predictions made by each model on the later period data to test whether the

[24] To see the issue in a different way consider the following simple model. Let M be a 0/1 variable that denotes whether an officer has committed misconduct over a time period and let R be a 0/1 variable for whether that officer had a reported complaint against them in that period. We assume that the investigation process has no "false positives", eg P(S=1|R=0) = 0, which is reasonable given the investigation procedures and burden of proof. The data are consistent with at least one of three shifts having occurred: (1) The rate of reporting true misconduct, P(R=1|M=1), went up; (2) The rate of sustaining reported misconduct, P(S=1|R=1,M=1), went up; or (3) the amount of misconduct, P(M=1), went up. If either of the first two things happened, then P(S=1|M=1) goes up and there's less measurement error between reported and true misconduct. The only case in which measurement error does not decrease is if the amount of misconduct increased (P(M=1) increased). That seems unlikely given that we also observe a decrease in the total number of complaints, a decrease in use of force, and a decrease in use of force with a firearm. While it's theoretically possible for those three things to have decreased while true misconduct to have increased, it seems less likely than the alternative explanation that investigations became more efficient with the roll-out of body-worn cameras over 2016 and 2017 and the overhaul of the city agency that investigates citizen complaints that occurred in 2017. Çubukçu et al (2023), for instance, documents that the roll-out of body-worn cameras increased the chance that a complaint ends in a sustained finding.

increased measurement error in the earlier data causes the model to flag different officers than the late model, and whether those differences influence accuracy.

This analysis shows that the measurement error in the earlier data does indeed cause the earlier model to 'miss' a group of high-risk officers that are identified by the later model. Table 2 shows a crosstab of officer-observations from 2017 based on whether officers would have been flagged by both models, only the early model, or only the late model. Officers flagged by the late model but not by the early model engage in future misconduct at a rate of 22.1%, which is significantly higher than the 6% rate of future misconduct among officers flagged by the early model but not by the late one (the base rate of misconduct in this period is 2.9%). In other words, the measurement error in the earlier data causes the model to under-estimate risk for a group of officers that are otherwise identified as high-risk by the later model.[25] Nonetheless, the early model still identifies officers at elevated risk. Those flagged by both models have the highest rate of future misconduct (27%), while even the officers flagged only by the early model have a rate of future misconduct that's twice as high as average (6% versus the base rate of 2.9%).

Taken together, this analysis shows measurement error is a concern but not a fatal one. Models built on data with reduced measurement error outperform models built on data with more measurement error, but even those models still find a useful degree of predictive signal.

Section 4: What predicts risk?

We now turn to the question of what are the best predictors of misconduct. While this is a natural question, it is also a conceptually difficult one because of the well-known statistical phenomenon of *model multiplicity* - many different machine learning models can perform

[25] In the appendix we show that the superior performance of the late model when evaluated in the late period is not just due to 'data drift' given that when we repeat the 'early vs late' experiment for the off-duty outcome, we don't find a stark contrast in performance between early and late models. We also show the later models do not perform better simply because of a higher base rate of Y=1 cases, which we show through an experiment where we randomly flip a subset of y=1 cases to y=0 so that the base rates are the same in the modified late data and the early data.

similarly even when they look and behave differently from another (also sometimes called “the Rashomon effect”). This introduces the complication that feature importance depends heavily on the selection of the specific model, even though there may be many equally good models. Given this conceptual challenge, we highlight general principles that, taken together, suggest that relatively simple models may suffer surprisingly modest losses in predictive accuracy relative to more complicated ones.

A. Prediction doesn’t require detailed features

Our first finding is that very detailed features (covariates) about prior events add only modest additional predictive signals beyond simple features. For each predicted outcome, we constructed three models with different levels of feature granularity. The “simple” model only used counts of events like number of prior complaints, number of prior uses of force, etc. The “intermediate” model also had access to prior event counts broken into coarse categories like “prior enforcement-related complaints” or “prior low-level use of force”. Finally, the “complex” model has access to very fine-grained counts based on specific details of past events (eg “prior sustained complaints involving excessive force where the officer used a firearm”, “prior uses of force where the officer used an open-handed strike, etc”). We found that the “intermediate” models achieved nearly the same level of predictive accuracy as the “complex” models for both on-duty and off-duty misconduct (see Table 3 for details). So while separating complaints into broad categories like “on-duty” vs “off-duty” is important for prediction, the additional effort to collect and process data about specific details of the complaint (what specific type of excessive force was it, how many officers were named on the complaint, what were the outcomes of each of the specific allegations) seems to yield few benefits in terms of predictive accuracy.

The finding that specific details of prior events do not carry substantial additional signal has two important implications. First, it suggests that simple policies - like flagging officers with the most prior events - might have a reasonable level of predictive accuracy while being much simpler to implement. We return to this question below. The second implication is that the outcome of a complaint investigation seems not to provide much information about future risk, e.g. that a pattern of unsustained complaints are predictive of future sustained complaints, which has potential implications for policies on data collection and retention. We examine this next.

B. Sustained complaints versus all complaints

One important decision in policing, and hence one important policy debate, is how to treat records of complaints that were not sustained (proven true) by the investigative process. Many departments limit access to or destroy these records, often arguing that retaining such records risks harming officers' careers based on events that may have never happened. The counter-argument is that the complaint investigation process is far from perfect and so many true complaints are not sustained. Just like in the justice system, complaint investigations start with the presumption of innocence, so the investigation needs to prove by a preponderance of evidence that the alleged events happened and that the officer's actions were out-of-policy.[26] Most departments have a relatively low sustain rate – Chicago sustains 3% of on-duty complaints, which is in line with national studies that estimate a sustain rate of less than 10% (Hickman and Poore 2016).

We find that non-sustained complaints are predictive of both on-duty and off-duty misconduct.[27] We arrived at these conclusions by training two different machine learning models

[26] For example, an excessive force investigation needs to prove that the officer not only most likely used force, but that their actions were excessive given the circumstances.

[27] These findings build on the work of Rozema and Schanzenbach (2019) that shows all allegations, even those that were not investigated, are predictors of future litigation. One potential drawback of using lawsuits as the outcome is that lawsuits might reflect misconduct, but are not a perfect signal of true misconduct because cities might settle to

– one model that used features from *all* prior complaints and another that only used features from prior *sustained* complaints (each model only uses features from prior complaints in order to test the predictive value of non-sustained complaints). The results of this exercise, shown in Table 4, demonstrate that our ability to predict misconduct decreases by nearly 50% only using sustained complaints. This finding replicates on public data from NYPD (see appendix), suggesting that this finding is not simply an artifact of Chicago's complaint investigation process.

Our results demonstrate that restricting access to records of non-sustained complaints comes at a large cost to statistical accuracy. This fact should be weighed against the various other considerations that arise in retaining and using non-sustained complaints, a decision that raises a variety of larger normative questions that are beyond the scope of the present paper.

C. Focus on patterns, not events

Together, these findings suggest a simple heuristic for identifying the riskiest officers is to "focus on patterns, not events" - identifying risky officers seems to be more about finding people with many prior events rather than finding people with a single prior serious or egregious event. To illustrate this point, we compared two risk policies - flagging any officer that has a sustained complaint in the prior five years, which results in about 7% of officers being flagged each year, or flagging an equally-sized group of officers with the most complaints in the past five years (e.g. flagging the top 7% of officers when ranked by number of complaints, whether those complaints are sustained or not). Officers flagged by the sustained complaint policy are significantly less likely to have a sustained complaint in the future relative to officers who would be flagged by the rank-by-all-complaints policy (3% vs 5%; full analysis in the appendix).

avoid incurring the cost of litigation. Our finding that non-sustained complaints predicts future sustained complaints thus eliminates the potential explanation that non-sustained complaints were only predicting spurious lawsuits.

The “patterns, not events” heuristic also applies to other potential prediction outcomes that we can measure in the public data from a different department, NYPD. We first similarly show that the group of officers with the most prior complaints are more likely to have a future sustained complaint than the group of officers with a prior sustained complaint. We also show that officers with the most prior lawsuits are more likely to have a future expensive lawsuit than officers with a prior expensive lawsuit. Overall, it appears to be a robust pattern that officers with a pattern of many events (whether they are sustained or not, and regardless of the severity of the behavior alleged in the complaint) are more likely to be involved in future adverse outcomes relative to officers with a serious event in their past.

Section 5: Is prediction practical?

The previous section showed that the prediction tools are not a panacea but nonetheless can identify a set of high-risk officers who could disproportionately benefit from targeting preventive interventions. But the question of whether prediction of police misconduct is useful depends not only on an assessment of benefits, but also of costs. We present results here that speak at least directionally or qualitatively to some of these costs and highlight ways in which they might be mitigated in some instances.

A. Machine learning versus simple rules

One potential concern about use of data and predictive analytics is the cost of building and/or setting up these systems. While academic policy analysts typically argue for allocation of resources using some type of benefit-cost analysis, in practice policymakers in the real world often behave as if they face hard budget constraints - which is presumably part of the motivation for shifting from benefit-cost ratios to assess policies to the marginal value of public finds (Hendren and Sprung-Keyser, 2020). Machine learning tools typically incur costs relative to

human intuition, whether the department builds its own predictive tool from scratch or buys one from a vendor. The good news is those costs can be mitigated with only modest-to-moderate loss of predictive accuracy by using simple prediction rules instead (building on the findings from the previous section). This possibility may be particularly valuable for smaller departments.

We analyze a simple alternative to using a complex machine learning risk model to flag the highest-risk officers: flaggings officers with the highest number of prior complaints instead - 'rank-by-complaints' (RBC). We compared the machine learning models and the rank-by-complaints policy by flagging the top 5% officers (which is about 632 officers) by either predicted risk or number of complaints over the prior two years[28], and comparing the recall (the fraction of officers that committed misconduct that were flagged ahead of time) of each method (results are shown in Table 5). While RBC does not perform as well as either risk model - the recall for on-duty misconduct is about 25% lower than the on-duty risk model and the recall for off-duty misconduct is about 30% lower than the off-duty risk model - it's notable that such a simple policy is competitive with a machine learning model that has access to a much wider variety of data and can combine data in more complex ways.[29] The predictive accuracy of RBC is a useful finding because development costs of machine learning tools are not just about money - they're about time as well, specifically the opportunity cost of having a predictive model only with a delay, foregoing chances to predict and prevent misconduct during development.

RBC may be particularly useful for small departments, which account for the vast majority of all police departments in the US and a majority of all police killings of civilians.[30] In

[28] If there are ties when flagging the top 5% of officers by the prior number of complaints, we break those ties randomly and report the average performance metric over 10 iterations of random tie-breaking.
[29] We repeated the comparison between machine learning models and simple ranking policies using public data from NYPD and similarly found that the ranking policies compare favorably to the machine learning models. Details of this analysis can be found in the appendix.
[30] Specifically, 63% of police killings as recorded by the Mapping Police Violence project are committed by departments with fewer than 500 officers. See appendix for more details of this analysis.

our experience the costs of building predictive models tend to be independent of the size of the jurisdiction in which the tool will be built. So the machine learning development costs will be much larger as a share of the total department budget for the police department in Galena, Illinois (population 3,300) than in Chicago, Illinois (population 2.7 million). Moreover the quality of custom-built machine learning models will, on average, be less accurate for smaller departments. The accuracy of machine learning models depends both on the number of observations and the number of covariates per observation - with larger data sizes enabling the machine learning models to capture more complex (and accurate) functional forms. In the appendix, we show that subsampling the Chicago data to match the scale of smaller departments degrades the performance of the ML models substantially, suggesting that smaller departments may not have enough data to develop a model that outperforms a simple policy.

B. Risk and police activity

One frequently-cited concern about the use of officer risk systems is that they conflate 'risk' with 'activity' because complaints and use of force are, it is sometimes argued, necessary byproducts of routine police work. Hence, if activity or assignment is not controlled for in some way, risk systems will simply flag active officers, and in turn, disincentivize policing activity (see for example Worden, Harris, and McLean 2014).

We tested these concerns by examining the effects of removing the correlation between risk scores and activity/assignment through a residualization procedure, and found that the original risk scores and the activity-adjusted risk scores mostly flag the same officers. We regressed risk scores against a set of measures that capture the unit an officer is assigned to, their role in that unit (e.g. police officer, sergeant, etc), and their policing activity[31] over the prior years

[31] There is deep disagreement about what constitutes 'good police activity'. Without taking a normative stand on that question, our method can accommodate different definitions by changing the activity measures that are used in this regression. We show for example that our conclusions hold whether we define activity broadly (all arrests, street stops, guns recovered, and department awards) or more narrowly (only felony arrests, guns recovered, and awards).

(number of arrests, stops, etc). This regression yields an officer's expected risk score given their assignment and activity. We then construct 'residualized risk scores' by taking the difference between the original risk scores and the expected risk score from the activity/assignment regression (i.e. the residual from the regression of risk scores on activity and assignment). We then compared which officers would be flagged if we had used the residualized risk scores instead of the original risk scores - essentially flagging officers whose level of risk is most in excess of what we would've expected given their activity and assignment.[32]

We found that most officers who are flagged by the original models are also flagged by the residualized models, implying that their level of risk is high even after adjusting for activity and assignment. Table 6 shows that 70% of officers flagged by the on-duty misconduct model are also flagged the residualized on–duty model. Moreover, when the residualized model and the original model disagree, we find that the original model more accurately predicts future misconduct. Put another way, adjusting for activity and assignment does very little, and when it does, it makes the models less accurate.

One reason that residualization has little effect is that the risk scores themselves have only a low to moderate correlation with activity and assignment. In the appendix, we show that activity and assignment explain only 41% of the variation in on-duty risk scores and 18% of the variation in off-duty risk scores. These findings align with prior work (Rozema and Schanzenbach 2019) that shows that high-risk officers who switch districts continue to receive complaints at higher rates after moving to a new district, suggesting that their prior circumstances weren't the primary cause of their elevated complaint rate. The low correlation

[32] One potential concern with this method is that risk and activity/assignment relationship could be endogenous, e.g. riskier officers may choose to police in a different way than lower risk ones or choose to work in certain assignments. Hence this method only yields an upper-bound on the mechanical relationship between risk and activity. This strengthens the finding since we find a low-to-moderate relationship even with this upper bound.

between risk scores and activity/assignment is due partly to our choice of predicting future sustained on-duty complaints, rather than all future on-duty complaints. A model that predicts any future on-duty complaint has a significantly higher correlation with activity/assignment (see appendix). In sum, if the nature of the algorithm could be made transparent to officers the risk of disincentivizing activity would seem to be modest.[33]

C. Algorithmic Bias

A final type of concern is whether the use of data to identify officers at elevated risk for future misconduct might exacerbate different types of biases that might occur within the department. This has become a major concern with algorithms for public policy in general, including with race and criminal justice specifically (see for example Ludwig and Mullainathan 2021 for a review). While race was not included as a factor in any of our risk models, it is still possible that the algorithms may perform differently across officer race/ethnicity.

We analyzed that possibility here by examining how model performance varies by race, shown in Tables 7a and 7b. The on-duty model has roughly equal performance across officer race - White, Black, and Hispanic officers are flagged at roughly equal rates and the rate of future on-duty misconduct is roughly the same across race/ethnicity among flagged officers. The off-duty model is more likely to flag Black officers than White or Hispanic officers, but the rate of future off-duty misconduct among flagged officers is roughly equal across race/ethnicity. That is, the flags are equally accurate for all flagged officers (Black vs. non-Black) suggesting the difference in off-duty flag rates are not just artifacts of algorithmic mistakes.

Whether the observed differences in the off-duty model reflect true differences in rates of off-duty misconduct or instead reflect biased data cannot easily be determined. One possibility is

[33] Note that if front-line police officers don't understand that risk is not highly correlated with activity, it would be possible for them to respond by reducing activity to avoid being flagged by a risk prediction model even if risk and activity are not highly correlated in the model in reality.

that the bar for making an off-duty complaint against a Black officer is lower than for other officers. We don't find evidence of this concern in our data (recognizing that we cannot test this theory perfectly given the limits of the data); off-duty complaints received by both flagged Black and flagged White officers are sustained at a rate of 28.3% and 28.9%.

Ultimately, the differences that we observe in off-duty misconduct rates by officer race is an important subject for future research. While prior research has documented the role of complainant demographics on complaint investigations (Headley et al., 2020) the particular issue of *off-duty* complaints has not been studied to the best of our knowledge. To the extent that the risk models inform the routing of a helpful support or service that reduces the likelihood of future off-duty misconduct, it is possible that the difference in flagging rates could potentially serve to reduce the racial differences in the rate of off-duty misconduct.

Section 6: Are there useful interventions to target?

Using data to target resources only makes sense if there's something useful to target. While too little is currently known about effective interventions in this area, there is some encouraging evidence accumulating from a recent series of RCTs. For example, Owens, Weisburd, Amendola and Alpert (2018) studied the effects of an intervention that involved having a supervisor review a recent case with an officer to get them to reflect more on their thinking and decision making during the event. They find a short-term reduction (six weeks out) in use of force of as much as 50%, no detectable change in citizen complaints against the officers, and a reduction in the number of arrests that officers make of around one-sixth. Dube, MacArthur and Shah (2023) find that a behavioral-science informed intervention that gets officers to recognize their potential to misconstrue situations out in the field and to essentially 'stop, look and listen' before they act reduces use of force by 22% and also reduces discretionary arrests that may have limited public safety value. Both interventions are effective for some

time-limited period; data could be potentially used to prioritize which officers get relatively more frequent 'boosters' of this type of training.

Our predictive models themselves may suggest additional types of interventions that could be helpful. A clue along those lines comes from our finding that predictions of on-duty and off-duty misconduct wind up being highly correlated.[34] Many of the officers at highest risk of on-duty misconduct are *also* at elevated risk for future *off-duty* misconduct. We analyzed the top 1% of officers by predicted risk of an on-duty misconduct - around 120 officers each year. Even though these officers are flagged based on their estimated risk of future on-duty misconduct, we find that they are also at elevated risk for off-duty misconduct. Approximately 17.6% of officers in this group are involved in off-duty misconduct in the two years following being flagged - a rate 3.7x higher than the department average. Moreover this elevated rate is not just driven by a few officers in this group. Table 8 shows the distribution of off-duty risk among officers at highest risk of on-duty misconduct (the top 2% on-duty risk). Off-duty risk is categorized based on risk relative to the department average. A risk level below the department average is classified as "Low", risk levels between 1-2x the department average is "Average", between 2-3x the department average is categorized as "Elevated", and more than 3x the department average is "High". Close to 70% of the high risk group for on-duty misconduct complaints fall into the category for elevated or high risk for future off-duty misconduct.[35]

The overlap between on–duty and off-duty risk has two important implications for the design of preventive interventions. First, efforts to address misconduct and efforts to improve officer wellness could benefit from being part of the same conversation. While previous research

[34] This finding echoes that of Rozema and Schanzenbach (2019) who find that actual on- and off-duty complaints are correlated. To see why our finding is subtly but importantly different, return to the notation described above where Y=f(X)+e, or misconduct is a function of some predictable structure and noise. It is possible that Rozema and Schanzenbach's correlation is due to correlation in the unpredictable part of the outcome, but our findings suggest instead that there is a strong correlation in the predictable parts of both outcomes.

[35] The converse is not true. Most officers at the top of the off-duty risk distribution have average on-duty risk.

has already shown that officers face higher rates of stress, exposure to traumatic incidents, and PTSD (Asmundson and Stapleton, 2008), the results we present here suggest this set of challenges that might normally fall into the category of 'officer wellness' may also be relevant for efforts to reduce on-duty misconduct and promote police legitimacy as well.

To be clear, these results do no *not* imply that *all* serious misconduct stems from issues in an officer's off-duty life, nor do they prove that interventions that help off-duty behavior would necessarily improve on-duty behavior (or vice versa). However, the results do suggest that some officers face multiple dimensions of risk simultaneously and efforts to improve officer wellness (out-of-work factors) could be a useful part of the efforts to improve risk and misconduct management. This would seem to be an important open question for future research.

The second implication is that a one-size-fits-all approach towards interventions and support may not be the best strategy. It might have seemed logical that officers at high-risk for on-duty misconduct would be good candidates for some sort of, say, police training. And that could well be true for many officers at highest risk for on-duty problems. But that may not be true for the nearly one-third of this group of officers who are at elevated risk for on-duty problems and are also at high-risk for off-duty outcomes. They may need a different approach. It is possible that departments will need to not only have a suite of interventions to address multiple underlying 'root causes,' but also design processes to figure out what sort of intervention works best for which types of officers.

While it would be too much to say social and medical science has figured out how to perfectly solve every out-of-work challenge people face in life, it would also be wrong to say that *nothing* is known about how to help people with these life problems. There is for example evidence of effective interventions for problems like substance use (Beaulieu et al., 2021),

trauma (Watkins et al., 2018), depression (Cujpers et al., 2020), and anxiety (Bandelow et al., 2022). Of course there are challenges with ensuring that adequate treatment capacity is available and of adequate quality, and that patients will seek out or adhere to treatment, problems that are beyond the scope of the present paper to solve. But for present purposes the key point is that there are indeed interventions with evidence of effectiveness that target relevant risk factors for on- and off-duty police misconduct, which in turn suggests the value of complementary policies (like predictive models) to target them in ways that maximize the social good they accomplish.

Section 7: MVPF Calculations for Misconduct Predictor

We now have the necessary building blocks to quantify the social welfare gains of targeting some preventive intervention using misconduct prediction. We assume a training intervention capable of reducing misconduct by 20% (see Dube et al., 2023). From Hendren and Sprung-Keyser (2020) the marginal value of public funds is defined as:

$$MVPF = \Delta W / (\Delta E - \Delta C)$$

where ΔW is the value of the policy impact on affected people (i.e., willingness to pay), ΔE is the up-front government expenditure required to build the algorithm, and ΔC is savings to government spending achieved by the policy. We show that for the very simple, low-cost prediction model discussed above, ‘rank-by-complaints’ (RBC), the estimated value of MVPF is infinity (the policy on net reduces government costs - it’s a ‘free lunch’ compared to a benchmark of random targeting) partly because this model has such low cost to build and deploy. Whether building a full-blown machine learning model also generates such favorable MVPF values is harder to say at the present time.

A. A rank-by-complaints predictor

We showed above that a simple RBC predictor of police misconduct captures important predictive signal about an officer's future risk of misconduct. While RBC is not as predictive as a full-blown machine learning model, it has the great advantage of allowing any department that has a reasonably well-functioning administrative data infrastructure to implement RBC quickly and cheaply; from our past work with different government agencies our best estimate for RBC is on the order of $\Delta E=\$500,000$. The analysis presented earlier in the paper suggests the public's willingness to pay for a misconduct predictor that captures signal about misconduct should be positive, $\Delta W>0$, since the predictor seems to capture true signal about misconduct (not just either activity or measurement error in the police data) and does not seem to exacerbate racial bias, at least of the types we are able to examine with the data we have here.

Deploying a tool like this within the Chicago PD with its 13,000 officers would prevent an additional ~$336,000 dollars in misconduct-related costs per year (relative to random targeting) if CPD flagged the top 1% of highest-risk officers for the training intervention, an additional $551,000 dollars if flagging the highest-risk 2%, and an additional $882,000 dollars if flagging the highest-risk 5%. In the appendix we describe how we derive our estimate of government savings per misconduct event (complaints or lawsuits) prevented. Because the cost of the RBC predictor is a one-time fixed cost, the implication is that for RBC, the MVPF calculation is infinite for a department of Chicago's size - that is, government savings outweigh development costs of $500,000 - even if the department was very selective in how many officers it flags. For example at a 2% flagging rate the RBC predictor pays for itself within the first year, and even at a 1% flagging rate the RBC predictor has paid for itself in the second year.

B. Machine Learning Predictors

As described above, machine learning models are more predictively accurate than RBC but also more costly to build and deploy. So relative to RBC, the public's willingness to pay, ΔW, should be larger, and that the government's cost savings ΔC from reduced lawsuit payouts and reduced investigation costs should also be larger (since more instances of misconduct are now being prevented). But relative to RBC, the build cost of a machine learning algorithm is also higher. From our own work with different government agencies in the past, we think a defensible upper-bound for this cost is on the order of something like ΔE=$5 million.[36]

Whether this type of model yields a MVPF value as favorable as that of RBC will depend on a number of factors that are currently hard to determine, partly because some of these factors will be application-specific and partly because some hinge on two open questions. The first is what share of officers will be flagged to receive the intervention. This is unavoidably application-specific since it is a policy question that some combination of police department and city government leadership chooses. The second is the degree to which the algorithm's predictions generalize across time and space. The more stable in time the underlying data generating process, the longer the algorithm once built can be deployed and operate effectively - that is, the less frequent it will have to be rebuilt and so the longer is the period of time over which the fixed build costs can be amortized and accrue benefits in the form of reduced misconduct. Similarly, the build costs of the algorithm will also depend on the degree to which the algorithm's predictions are context-dependent versus very generalizable across location, since that determines whether departments would need to build their own machine learning models versus could use one 'off the shelf' built with data from some other jurisdiction.

[36] Our team helped build and deploy a new pretrial release algorithm for New York City, for which we calculate an upper bound cost of $4 million (Ludwig, Mullainathan, and Rambachan, 2024). Our figure of $5 million is an even more conservative upper bound for the costs of deploying a misconduct predictor.

Table 9 shows that for a department the size of Chicago, building a machine learning algorithm from scratch would only yield a MVPF=infinity (like RBC) if the algorithm could be deployed for at least five years, the department flags a sizable number of officers (5% or more), and the build cost of the algorithm turned out to be somewhat lower than our upper bound estimate of $5 million. One way in which the algorithm costs could wind up less than $5 million - perhaps far less - would be if a single algorithm could be built that applies to multiple jurisdictions. If there were even just a few other jurisdictions sufficiently 'Chicago-like' in terms of their data generating process, a single algorithm that could be built for multiple cities would achieve a MVPF of infinity with larger benefits to the public than the RBC predictor achieves. Even if each algorithm is deployed in just a single jurisdiction, if the department was of NYPD's size the MVPF of the machine learning model again achieves infinity (because of the larger scale in terms of number of officers affected and number of misconducts prevented).

C. Alternative Benchmarks for Targeting

One potential concern with our MVPF calculations is that random targeting of a preventive intervention might be too pessimistic a bar to clear; that is, what if whatever status quo procedure departments use to target preventive interventions today is better-than-random? To examine the sensitivity of our analysis to that benchmark, we replicated our MVPF analyses using a baseline that flags officers with the most prior *sustained* complaints. This policy proxies the implicit procedures of many departments that focus attention on officers after a serious event has occurred. With this as a baseline, the estimated per-year savings when flagging at the 2% level (for example) drops from $638,000 to $372,000 for ML targeting and from $551,000 to $285,000 for RBC targeting. Even with this different baseline, the RBC policy yields an infinite

MVPF at any flagging rate if it runs for at least 3 years of operation (see the appendix for full calculations).

Section 8: Conclusion

The causes of police misconduct remain the topic of ongoing debate. Many explanations point to 'macro' factors like structural racism and lack of transparency and democratic oversight. Other candidate explanations fall more directly under the control of the police department, like the potential failure to hire the right officers (Chalfin et al., 2016) or hire a sufficiently diverse set of officers (McCrary, 2007, Hoekstra and Sloan, 2022), inadequate supervision of officers, inadequate training, and untreated mental health challenges. To the extent to which it is possible to prevent misconduct in the first place, failure to do so harms not only those directly affected, but also undermines public trust in law enforcement (and perhaps government itself).

Nor is the current status quo obviously good for police officers themselves, either. Alongside the recent decline in public trust of the police we have also seen a decline in morale among police. Officers are reportedly leaving the profession in greater numbers and departments are finding it difficult to recruit.[37] The very nature of police work requires exposure to stress and trauma (Violanti et al, 2017). A recent survey of the Dallas Police Department found that a quarter of respondents had positive screening results for mental illness symptoms (Jetelina et al, 2020). More officers die by suicide than in the line of duty.[38]

This is all to say that anything capable of preventing police misconduct would generate important benefits in a policy area that is often referred to as the 'new civil rights movement.'[39]

---

[37] https://www.washingtonpost.com/national-security/2023/05/27/police-vacancies-hiring-recruiting-reform/

[38] Over the last ten years, the FBI LEOKA (Law Enforcement Officers Killed and Assaulted) program reports between 100-120 officers killed in the line of duty per year (which includes deaths from assaults and from accidents). Blue Help, a non-profit that tracks law enforcement suicide, reports 170-228 suicides per year.

[39] See for example https://www.politico.com/newsletters/the-recast/2021/05/25/george-floyd-death-anniversary-civil-rights-492986

The current approach implicitly relies largely on a theory of deterrence, in the sense that most police departments seem to rely mostly on reactive, after-the-fact responses (some combination of disciplinary actions, retraining or job reassignment). Were it possible to predict misconduct risk in advance, perhaps it would be possible to prevent more misconduct in the first place (recognizing Chalfin and Kaplan, 2021's point that prediction and prevention is not a panacea).

A key contribution of our work is to show that predicting misconduct to target preventive interventions has very favorable social welfare gains - a simple prediction model (rank by complaints) generates benefits to the public and saves the government money, yielding a MVPF of infinity. We show these results are not sensitive to the choice of outcome variable (for example, use of lawsuits as in Rozema and Schanzenbach versus use of sustained complaints as here), that predictable structure does not *seem* to be simply an artifact of non-random measurement error in these data themselves, and that risk is not just a proxy for officer activity. Misconduct is predictable enough to have practical net benefits.

Our findings may also help inform a number of pragmatic policy or implementation questions. The ability to use simple rank-by-complaint models means even small departments can capitalize on predictive modeling. Our findings also speak to the fear of some observers that flagging officers based on misconduct risk might inadvertently disincentivize officers from doing their jobs. Nor does it appear that risk would exacerbate potential racial biases in the data, as we observe the models perform similarly across officer race and ethnicity. Or consider our finding that a good mental model for identifiable risk is to "focus on patterns, not specific events." Many policies about data access, use, and retention tend to focus on prior serious events - like sustained complaints, officer-involved shootings, or officers being fired for cause, while de-emphasizing or discarding seemingly low-level events like non-sustained complaints or less serious allegation

types. And many departments limit access to or require the destruction of non-sustained complaint records.[40] Our results suggest these reporting criteria will have the consequence of causing the databases to support less accurate predictions relative to databases that record *all* misconduct.

These results also have implications for a widely-cited police reform proposal: creation of police misconduct databases, which are intended to address 'wandering officers' (Grunwald and Rappaport, 2017) who get fired from one department (or leave in the middle of an investigation), find employment at another department, and continue to engage in problematic behavior.[41] Designers of these databases face difficult tradeoffs, which can be seen in the call for a national misconduct database in an early 2023 executive order issued by President Biden. The proposed national database limits the reporting requirement to sustained complaints and discipline resulting from serious misconduct. This narrow scope reduces the chance that an officer's career prospects are impacted by a potentially false claim of misconduct. Our results show the tradeoff of this decision: the proposed national database is sacrificing a degree of predictive signal by only asking departments to report sustained complaints where the allegations were considered "serious". While the exact loss in accuracy is impossible to estimate without data from more departments, our results from Chicago (and public NYPD data) suggest it could be substantial.

A key priority for future work is to better understand what types of preventative interventions of the sort that would be targeted by predictive models are most useful in practice. We show that there are some evidence-based prevention interventions, particularly

[40] In a review of 178 police union contracts, Rushin (2017) finds that at least 87 have provisions that limit consideration of disciplinary history, sometimes requiring destruction of those records as soon as six months after the conduct occurred. These provisions are typically structured to restrict access to records of non-sustained complaints and/or low-level complaints on a relatively short time horizon, while sometimes requiring the destruction or expungement of more serious records over a longer time horizon.

[41] One prominent example was the hiring of Timothy Loehmann, the former Cleveland police officer who shot and killed 12-year old Tamir Rice, by the Tioga Police Department in Pennsylvania.

behavioral-science-informed police training interventions (Owens et al., 2018, Dube, MacArthur and Shah, 2023), that could be targeted. We also show that there is correlation between predicted on-duty versus off-duty risk, which suggests that policies that address off-duty challenges officers face could potentially have secondary benefits of reducing on-duty misconduct. The good news is that there is an accumulating body of evidence on interventions to address substance abuse, trauma, depression and anxiety, but most of that evidence comes from studies of civilians rather than of police officers specifically. The usefulness of the available preventive interventions will, at the end of the day, be the rate-limiting step in shaping how data and predictive analytics assist the effort to prevent police misconduct.

References

Arnold, D., Dobbie, W.S., and Hull, P. (2020) Measuring racial discrimination in bail decisions. Working Paper 26999, *National Bureau of Economic Research.*

Ariel, B., Farrar, W. A., & Sutherland, A. (2015). The effect of police body-worn cameras on use of force and citizens' complaints against the police: A randomized controlled trial. *Journal of Quantitative Criminology*, 31, 509-535.

Asmundson, G. J. & Stapleton, J. A. (2008). Associations between dimensions of anxiety sensitivity and PTSD symptom clusters in active‐duty police officers. *Cognitive Behaviour Therapy, 37(2)*, 66-75.

Ba, B., & Rivera, R. (2019). The effect of police oversight on crime and allegations of misconduct: evidence from Chicago." *U of Penn, Institute for Law & Econ Research Paper,* 19-42

Beaulieu, M., Tremblay, J., Baudry, C., Pearson, J., and Bertrand, K.. (2021). A systematic review and meta-analysis of the efficacy of the long-term treatment and support of substance use disorders. *Social Science and Medicine,* 285.

Braga, A., Coldren, J. R. Jr., Sousa, W., Rodriguez, D., & Alper, O. (2017). The benefits of body-worn cameras: new findings from a randomized controlled trial at the Las Vegas Metropolitan Police. Arlington, VA: CNA.

Brenan, M. (2020, 8). *Amid Pandemic, Confidence in Key US Institutions Surges*. Gallup.com . https://news.gallup.com/poll/317135/amid-pandemic-confidence-key-institutions-surges.aspx

Carton, S., Helsby, J. E., Joseph, K., Mahmud, A. S., Park, Y., Walsh, J., Cody, C., Patterson, C. E., Haynes, L., & Ghani, R. (2016). Identifying Police Officers at Risk of Adverse Events. *Knowledge Discovery and Data Mining.*

CDC/National Center for Health Statistics. (2022, March 1). *Firearm Mortality by State*. CDC.gov . https://www.cdc.gov/nchs/pressroom/sosmap/firearm_mortality/firearm.htm

Chalfin, A. (2022). Policing and Public Safety *Arnold Ventures Public Safety Series.*

Chalfin, Aaron, Oren Danieli, Andrew Hillis, Zubin Jelveh, Michael Luca, Jens Ludwig and Sendhil Mullainathan (2016) "Productivity and selection of human capital with machine learning." *American Economic Review, Papers & Proceedings*. 106(5): 124-7.

Chalfin, A. and Kaplan, J., 2021. How many complaints against police officers can be abated by incapacitating a few "bad apples?". *Criminology & Public Policy, 20(2), pp.351-370.*

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters.

Chouldechova, A., Benavides-Prado, D., Fialko, O., & Vaithianathan, R. (2018). A Case Study of Algorithm-Assisted Decision Making in Child Maltreatment Hotline Screening Decisions. *Conference on Fairness, Accountability and Transparency,* 134–148.

Cubitt, T. I. (2023). Using network analytics to improve targeted disruption of police misconduct. Police Quarterly, 26(1), 24-53.

Çubukçu, S., Sahin, N. M., Tekin, E., & Topalli, V. (2021). Body-Worn Cameras and Adjudication of Citizen Complaints of Police Misconduct. *Social Science Research Network.*

Cuijpers, Pim, Argyris Stringaris, and Miranda Wolpert (2020) "Treatment outcomes for depression: challenges and opportunities." The Lancet.

Desmarais, S. L., Zottola, S. A., Clarke, S. C., & Lowder, E. M. (2021). Predictive Validity of Pretrial Risk Assessments: A Systematic Review of the Literature. *Criminal Justice and Behavior.*

Dube, O., MacArthur, S. J., & Shah, A. K. (2023). A cognitive view of policing (No. w31651). National Bureau of Economic Research.

Einav, L., Finkelstein, A., Mullainathan, S., & Obermeyer, Z. (2018). Predictive modeling of US health care spending in late life. *Science*, 360(6396), 1462-1465.

Fryer, Roland G (2020) An Empirical Analysis of Racial Differences in Police Use of Force: A Response *Journal of Political Economy*, 128 (10), 4003–4008.

Goncalves, Felipe and Steven Mello (2021) "A few bad apples? Racial bias in policing," *American Economic Review*, 111 (5), 1406–41.

Gramlich, John (2023). *What the data say about gun deaths in the United States.* https://www.pewresearch.org/short-reads/2023/04/26/what-the-data-says-about-gun-deaths-in-the-u-s/

Grinsztajn, L., Oyallon, E., & Varoquaux, G. (2022). "Why do tree-based models still outperform deep learning on tabular data?" *ArXiv (Cornell University).*

Grunwald, B., & Rappaport, J. (2020). The Wandering Officer. *Yale Law Journal.*

Hastings, J. S., Howison, M., & Inman, S. (2020). Predicting high-risk opioid prescriptions before they are given. *Proceedings of the National Academy of Sciences of the United States of America*, 117(4), 1917–1923.

Headley, A. M., D'Alessio, S. J., & Stolzenberg, L. (2020). The effect of a complainant's race and ethnicity on dispositional outcome in police misconduct cases in Chicago. *Race and justice, 10(1)*, 43-61.

Hendren, Nathaniel and Ben Sprung-Keyser (2020) "A unified welfare analysis of government policies." *Quarterly Journal of Economics*. 135(3): 1209-1318.

Hendren, Nathaniel and Ben Sprung-Keyser (2022) "The case for using the MVPF in empirical welfare analysis." Cambridge, MA: NBER working paper 30029.

Hickman, M. J., & Poore, J. E. (2016). National data on citizen complaints about police use of force: Data quality concerns and the potential (mis) use of statistical evidence to address police agency conduct. Criminal Justice Policy Review, 27(5), 455-479.

Hoekstra, Mark and CarlyWill Sloan (2022) "Does race matter for police use of force? Evidence from 911 calls." *American Economic Review*. 112(3): 827-60.

Jain, A., Sinclair, R., & Papachristos, A. V. (2022). Identifying misconduct-committing officer crews in the Chicago police department. Plos one, 17(5), e0267217.

Jetelina, K. K., Molsberry, R., Gonzalez, J. M. R., Beauchamp, A. M., & Hall, T. (2020). Prevalence of Mental Illness and Mental Health Care Use Among Police Officers. *JAMA Network Open*, 3(10), e2019658.

Jordan, Andrew and Kim, Taeho, Strengthening Police Oversight: the Impacts of Misconduct Investigators on Police Officer Behavior (December 30, 2022). Available at SSRN: https://ssrn.com/abstract=4099052 or http://dx.doi.org/10.2139/ssrn.4099052

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., ... & Liu, T. Y. (2017). Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 30.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Lescovec, Jens Ludwig and Sendhil Mullainathan (2018) "Human decisions and machine predictions." *Quarterly Journal of Economics.* 133(1): 237-293.

Ludwig, J., & Mullainathan, S. (2021). Fragile algorithms and fallible decision-makers: lessons from the justice system. *Journal of Economic Perspectives, 35(4)*, 71-96.

Ludwig, J., Mullainathan S. and Rambachan, A. (2024a) "The unreasonable cost-effectiveness of algorithms." *AEA Papers & Proceedings*.

Ludwig, J., Mullainathan S., and Rambachan, A. (2024b) "The unreasonable cost-effectiveness of algorithms." Cambridge, MA: NBER working paper.

Lum, C., Koper, C. S., Wilson, D. B., Stoltz, M., Goodier, M., Eggins, E., ... & Mazerolle, L. (2020). Body-worn cameras' effects on police officers and citizen behavior: A systematic review. *Campbell Systematic Reviews*, 16(3), Article-number.

McCrary, Justin (2007) "The effect of court-ordered hiring quotas on the composition and quality of police." *American Economic Review*. 97(1): 318-353.

Mullainathan, S., & Obermeyer, Z. (2017). Does machine learning automate moral hazard and error?. *American Economic Review, 107(5)*, 476-480.

Nadeem, R. (2022, February 15). *Americans' trust in scientists, other groups declines*. Pew Research Center Science & Society. https://www.pewresearch.org/science/2022/02/15/americans-trust-in-scientists-other-groups-declines/

Office of the Associate Director of Communication. (2022, June 6). *Firearm Deaths Grow, Disparities Widen*. CDC.gov . https://www.cdc.gov/vitalsigns/firearm-deaths/index.html

Olander, O. (2023, February 7). *Biden on police killings: 'We can't turn away'*. POLITICO. https://www.politico.com/news/2023/02/07/biden-police-killings-state-of-the-union-00081718

Owens, E., Weisburd, D., Amendola, K. L., & Alpert, G. P. (2018). Can you build a better cop? Experimental evidence on supervision, training, and policing in the community. *Criminology & Public Policy*, 17(1), 41-87.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011c). Scikit-learn: Machine Learning in Python. *HAL (Le Centre Pour La Communication Scientifique Directe)*

Probst, P., & Boulesteix, A. L. (2017). To tune or not to tune the number of trees in random forest. The Journal of Machine Learning Research, 18(1), 6673-6690.

Rabanser, S., Günnemann, S., & Lipton, Z. (2019). Failing loudly: An empirical study of methods for detecting dataset shift. *Advances in Neural Information Processing Systems*, *32*.

Rambachan, Ashesh (2023) "Identifying prediction mistakes in observational data." MIT economics working paper.

Rogers, K. & Kanno-Youngs, Z. (2021, November 30). *After another police shooting, Biden urges calm. Activists want answers. (Published 2021)*. The New York Times https://www.nytimes.com/2021/04/13/us/politics/biden-daunte-wright-police-shooting.html

Rozema, K., & Schanzenbach, M. (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy*, 11(2), 225-268.

Rushin, S. (2017). Police Union Contracts. *Duke Law Journal,* 66(6), 1191–1266.

Sierra‐Arévalo, M., & Papachristos, A. (2021). Bad apples and incredible certitude. Criminology & Public Policy, 20(2), 371-381.

Subramanian, R., & Arzy, L. (2021, May 21). *State policing reforms since George Floyd's murder*. Brennan Center for Justice. https://www.brennancenter.org/our-work/research-reports/state-policing-reforms-george-floyds-murder

Walker, Samuel, Geoffrey P. Alpert and Dennis J. Kenney (2001) *Early Warning Systems: Responding to the Problem Officer*. US Department of Justice, Office of Justice Programs, National Institute of Justice Research in Brief. https://www.ojp.gov/pdffiles1/nij/188565.pdf

Wang, C., Wu, Q., Weimer, M., & Zhu, E. (2021). Flaml: A fast and lightweight automl library. *Proceedings of Machine Learning and Systems*, *3*, 434-447.

Washburn, E. (2023, February 6). *America Less Confident in Police than Ever Before: A Look at the Numbers*. Forbes.com . https://www.forbes.com/sites/emilywashburn/2023/02/03/america-less-confident-in-police-than-ever-before-a-look-at-the-numbers/?sh=5f3a02a46afb

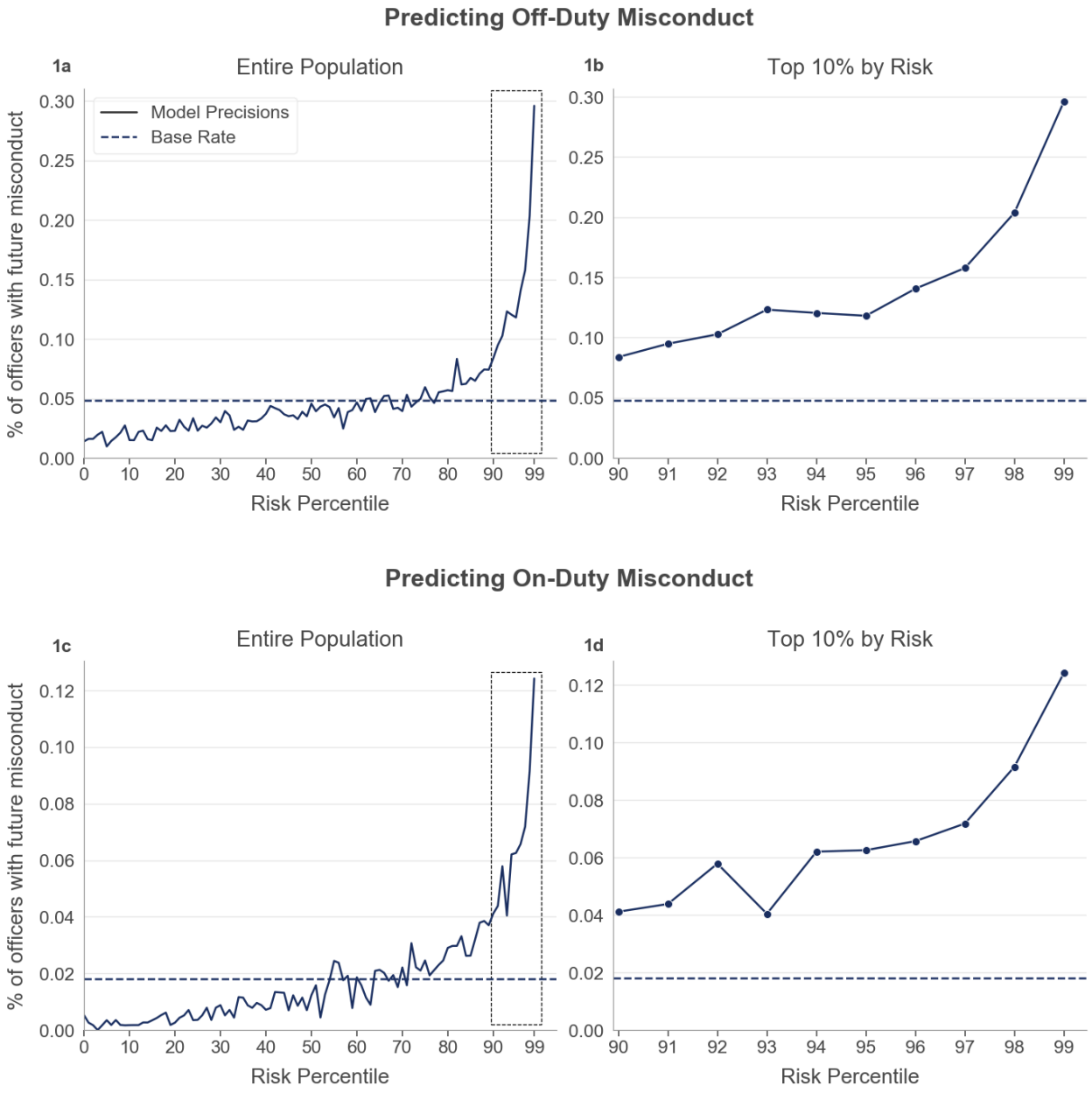
Watkins, Laura E., Kelsey R. Sprang and Barbara O. Rothbaum (2018) "Treating PTSD: A review of evidence-based psychotherapy interventions." *Frontiers in Behavioral Neuroscience*.

Weerts, H. J., Mueller, A. C., & Vanschoren, J. (2020). Importance of tuning hyperparameters of machine learning algorithms. *arXiv preprint arXiv:2007.07588*.

Worden, Robert, Christopher Harris, and Sarah J. McLean. (2014) Risk Assessment and Risk Management in Policing. *Policing: An International Journal of Police Strategies & Management 37.2* 239-258.

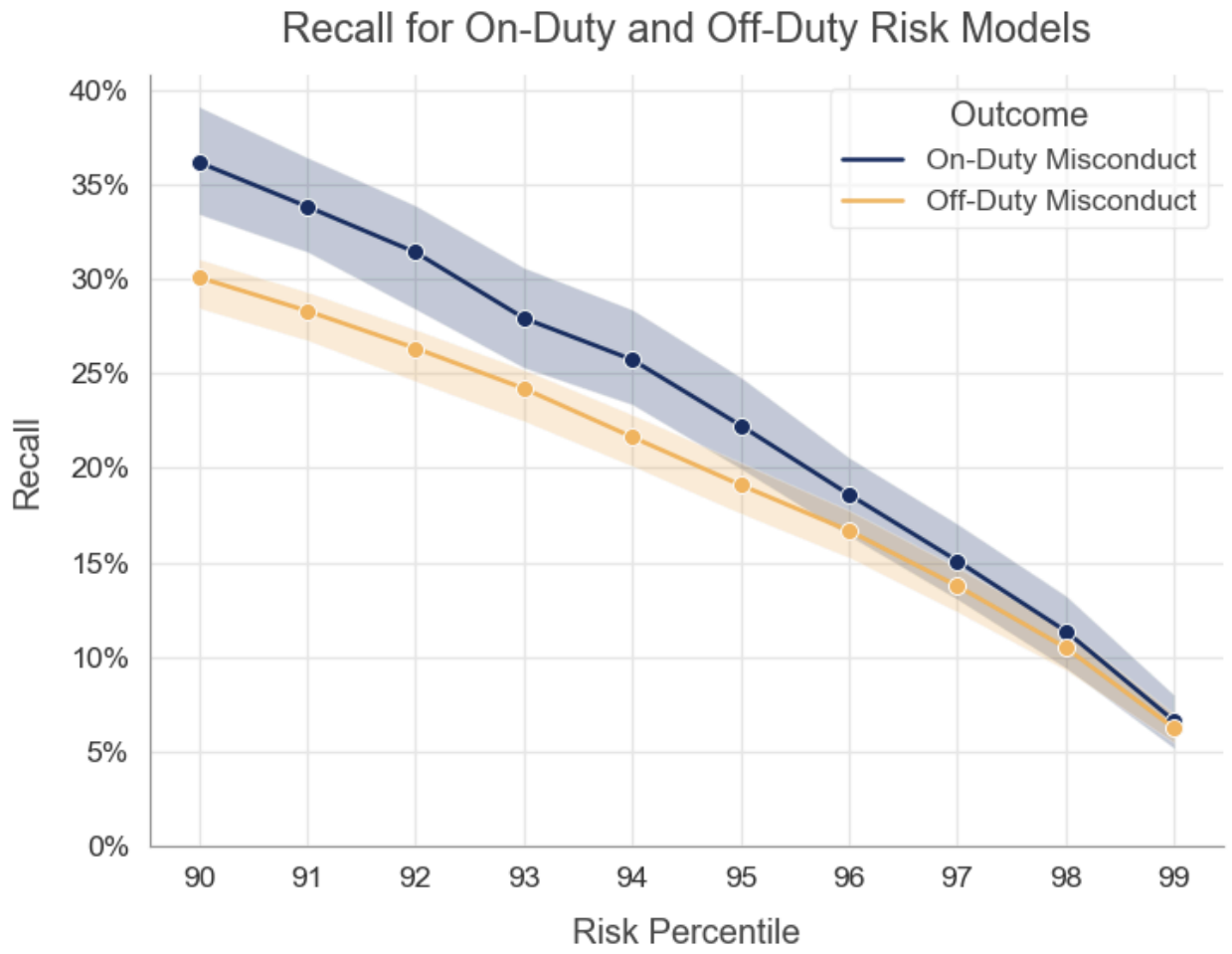
Ziobrowski, H. N., Cui, R., Ross, E. L., Liu, H., Puac-Polanco, V., Turner, B., ... & Kessler, R. C. (2023). Development of a model to predict psychotherapy response for depression among Veterans. *Psychological Medicine, 53(8)*, 3591-3600.

Figure 1: Rates of misconduct by risk percentiles



Notes: Figure 1a shows the rate of future off-duty misconduct as a function of predicted off-duty risk percentile. Misconduct rate is defined as the fraction of officer-year observation in each risk percentile that are involved in misconduct within two years of being flagged. Predicted risk percentiles are defined based on each year and then pooled across all years. Each point in the plots represents exactly 1% of observations, e.g. the point at the 99th percentile represents observations between the 99th and 100th percentile, the point at the 98th percentile represents observations between the 98th and 99th percentile, etc. Figure 1b "zooms" in to highlight misconduct rates among the top 10% of officers by predicted risk. Figure 1c shows the rate of future on-duty misconduct as a function of predicted on-duty risk percentile, and Figure 1d zooms in on the top 10% of officers by predicted on-duty risk.

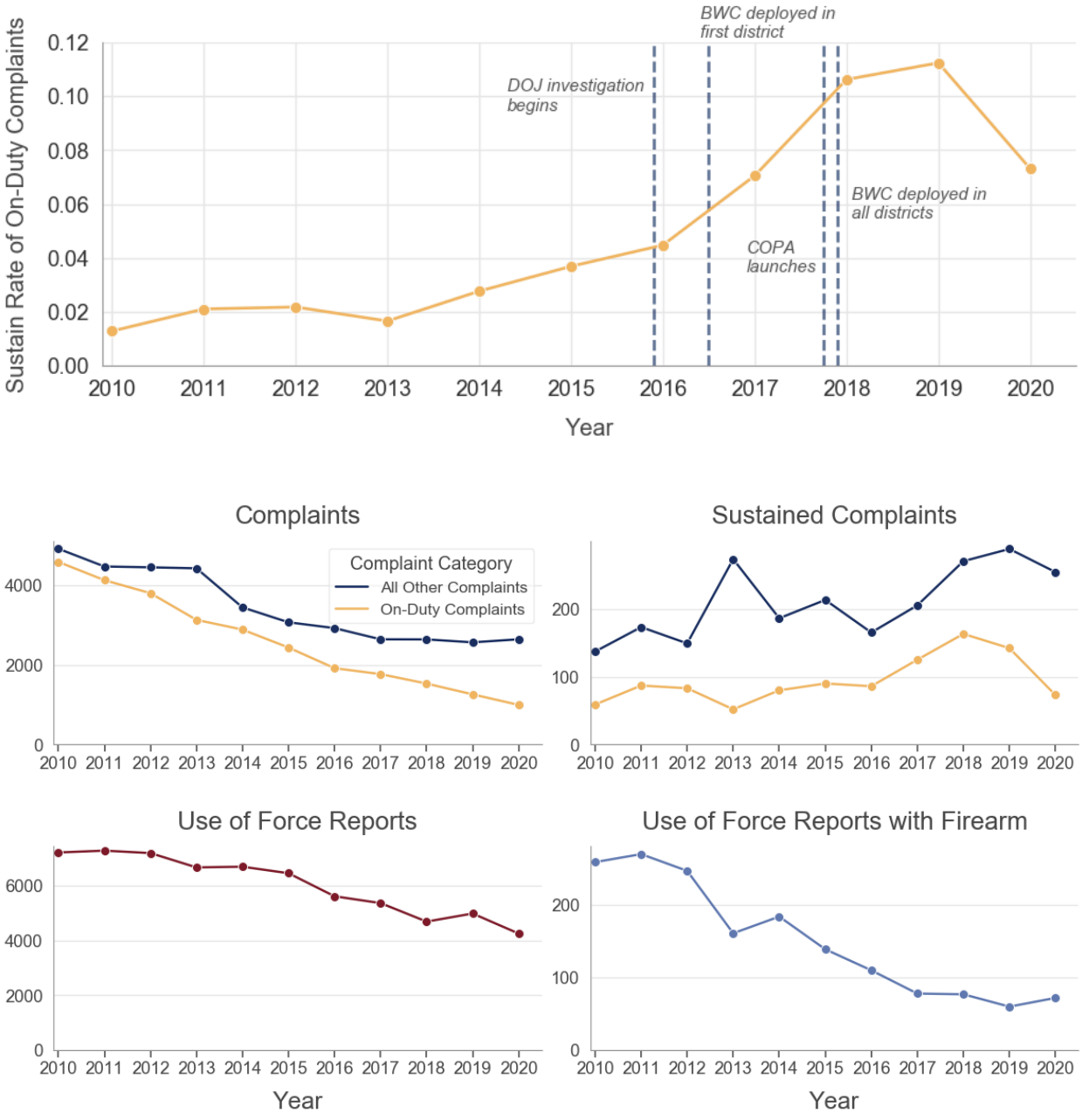
Figure 2: Recall of risk models



Note: Recall of the on-duty and off-duty risk models as a function of risk percentile. Recall measures the fraction of officers who were involved in misconduct in the outcome period (years T+1 or T+2) that were flagged in the observation period (year T). Each point on these curves show the recall of the risk models if the threshold for high–risk flagging were set at that percentile. For example, the plot shows that setting the high-risk threshold at 95% - flagging the top 5% of officers - would result in a recall of 19% for the off-duty model and a recall of 22% for the on-duty model.

The shaded regions indicated the 95% confidence interval for these recall statistics. Confidence intervals were formed by bootstrapping the dataset at the officer level 100 times, computing the recall @ X% for each bootstrap replicate, and reporting the values at the 2.5th and 97.5th percentile of the empirical distribution of recall values.

Figure 3: Change in complaints over time



Note: Figure 3a shows the percentage of complaints that were filed each year that resulted in a sustained finding, broken out by the type of complaint. The sustain rate of on-duty complaints (e.g. complaints of excessive force, wrongful arrest, etc) are consistently rising over time but the pace of that trend increases significantly from 2016 to 2017. The timing of that acceleration aligns with a number of reforms that occurred in Chicago following the release of video footage that showed a CPD officer shooting and killing a Black teenager named Lacquan McDonald. In particular, CPD started their deployment of

body-worn cameras in mid-2016 and there was an overhaul to the civilian agency that investigates certain types of complaints in mid-2017.

Figure 3b shows that the number of complaints filed against CPD officers has been steadily declining since 2010. Figure 3c shows that the number of sustained on-duty complaints is relatively constant from 2010-2016 but increases beginning in 2017. The rise in the sustain rate and the number of sustained complaints around 2017 is consistent with an increase in the likelihood that a true instance of misconduct results in a sustained finding.

The other possible explanation for a simultaneous rise in the number of sustained complaints and sustain rate is that the total amount of true misconduct increased beginning in 2017. This seems less likely given that we also observe a drop in the total number of complaints (3b), a drop in use of force reports (3d), and a drop in use of force reports where a firearm was used (3e). While the drop in use of force could have been driven by under-reporting, use of a firearm is significantly harder to under-report, suggesting that the drop in use of force is a real phenomenon rather than an artifact of under-reporting.

Table 1: Summary of dataset and models

Number of observations	113,768	
Years covered	2010-2018	
Number of observations per year	12,000-13,000	
	On-duty misconduct	Off-duty misconduct
Base Rate *Percentage of officers who have an instance of misconduct during the two-year outcome period*	1.9%	4.8%
Precision @ 2% *Percentage of officers in the top 2% of estimated risk who have an instance of misconduct during the two-year outcome period*	10.8% [9.0,12.4]	24.8% [22.3, 27.6]
Lift @ 2% *Ratio of precision@2% to the base rate*	5.9x [5.1,6.8]	5.2x [4.6, 5.7]
Recall @ 2% *Percentage of officers who have an instance of on-duty misconduct during the outcome period that were in the top 2% of risk at the time of prediction*	11.4% [9.3,13.3]	10.4% [9.3, 11.5]
Precision @ 5%	8.4% [7.6,9.3]	18.3% [16.9, 19.9]
Lift @ 5%	4.6x [4.1, 5.0]	3.8x [3.5, 4.1]
Recall @ 5%	22.2% [20.0, 24.8]	19.1% [17.6, 20.3]
ROC-AUC	.752 [.740, .768]	.682 [.671, .688]

Notes: Summary of the dataset, the two predicted outcomes, and the predictive accuracy of the machine learning models that predict those outcomes. Performance metrics for the model are computed for each year and then averaged over all years in the dataset. Confidence intervals for each performance metric were constructed by bootstrapping the dataset at the officer level 100 times, measuring the performance metric for each bootstrap replicate, and reporting the values at the 2.5th and 97.5th percentile.

Table 2: Effects of measurement error on risk models

	Rate of future on-duty misconduct	
	Not flagged by late model	Flagged by late model
Not flagged by early model	2.4% (11,885 officers)	22.1% (181 officers)
Flagged by early model	6.1% (181 officers)	27.3% (66 officers)

Notes: Comparison of officers flagged by the 'early' model and 'late' model during the 'late' periods. As argued in Section 3, data from before 2017 has more under-reporting of true misconduct and hence a larger degree of measurement error between the true outcome (whether an officer engaged in on-duty misconduct) and the reported outcome (whether an officer had a sustained on-duty complaint). We tested the effect of that measurement error by training one model only on data from pre-2017 (the early model) and another model on data from post-2017 (the late model). The above table shows the rate of future on-duty misconduct and number of officers in groups defined by whether they would have been flagged by both models (bottom right cell), just the late model (top right cell), just the early model (bottom left cell), or neither model (top left cell). All observations are from 2017.

This analysis shows that the early model (the one built on data with more measurement error) misses a group of officers with significantly elevated risk. Officers who were flagged by just the late model are 7-8x more likely to engage in future misconduct than the average officer. The measurement error in the earlier data causes the early model to underestimate risk for this group.

On the other hand, this analysis shows that the early model still identifies a relatively high-risk group of officers despite the measurement error. The highest risk group of officers are those who are flagged by both the early and the late model - those officers are roughly 9-10x more likely to engage in future on-duty misconduct relative to the average officer. Thus, despite the higher degree of measurement error, the early model still identifies the highest risk group of officers. And even the officers who are only flagged by the early model are still twice as likely to engage in future misconduct relative to the average officer.

Table 3: Model accuracy by feature complexity

		Recall@5%	
Feature complexity	**Number of Features**	On-duty misconduct	Off-duty misconduct
Simple	30	21%	16.4%
Intermediate	150	21.2%	18.7%
Complex	800	22.2%	19.1%

Notes: The statistical accuracy of machine learning models as a function of the granularity of features used by each model. The 'simple' model only uses coarse counts of events from an officer's past such as "number of complaints in past five years" or "number of use of force reports in past two years", etc. The features in the intermediate model include slightly more detail about prior events such as "prior off-duty complaints" or "prior level 1 use of force" (force level, according to CPD policy, is a measurement of how severe an officer's physical tactics were, with level 1 being the lowest and 3 being the highest). Finally, the complex model includes very granular features about the specific nature of prior events and the outcomes of associated investigations (where relevant) like "prior sustained complaints of excessive force in which the officer was alleged to have injured the subject" or "prior use of force where the officer used their taser".

These results show that very fine-grained features - like those included in the complex model - do not add much predictive value above and beyond the coarse categorization scheme used in the 'intermediate' features. From an engineering perspective, these results suggest that model developers can avoid the time it takes to collect and process these very fine-grained features without having to sacrifice statistical accuracy. From a policy perspective, these results suggest that very simple policy (like ranking by the total number of prior complaints) may be a viable alternative to full-blown risk models (a point that we explore in Section 5). Furthermore, the fact that categorizing complaints based on the outcome of a prior investigation (e.g. was the complaint sustained or not) does not add much predictive signal, suggests that even prior unsustained complaints carry useful information about future risk (a point that is further discussed in Section 4).

Table 4: Effect of limiting to sustained complaints

	Recall@5%	
Risk Model	On-duty misconduct	Off-duty misconduct
All complaints	16.13%	16.41%
Only sustained complaints	8.58%	8.7%

Notes: The effects of removing non-sustained complaints from an officer's record on predictive accuracy. In this experiment, we built two risk models for each outcome- one risk model used features derived from all prior complaints made against an officer and the other could only use features from prior sustained complaints. No other other features were included in these models in order to highlight the role of complaints. The "sustained only" models are significantly less accurate than the "all complaint" models - the recall@5% of the 'sustained only' models being approximately half of the recall@5% of the 'all complaints' model. These results show that even records of non-sustained complaints carry useful predictive signals. This finding also replicates on public data from NYPD (shown in the appendix).

This finding has implications for the design of data collection and risk management systems, particularly because police departments vary widely in whether and how they retain records of nonsustained complaints. For instance, many have called for the creation of police misconduct databases to help departments avoid hiring an officer who has a history of misconduct from other departments (known as the `wandering officer` Grunwald and Rappaport, 2017). However many of these databases, including the national database proposed by the Biden Administration in early 2023, only include sustained complaints. This narrow scope reduces the chance that an officer's career prospects are impacted by a potentially false claim of misconduct. Our results show the tradeoff of this decision: the proposed national database is sacrificing a degree of predictive signal by only asking departments to report sustained complaints where the allegations were considered "serious". Whether the benefits of the increased accuracy outweigh the potential costs of including nonsustained complaint records depends on a variety of normative questions that are beyond the scope of the present paper.

Table 5: Risk Models vs Rank-by-complaints

Comparison of risk models and rank-by-complaints		
	Recall and Annual True Positives @5%	
	On-Duty	Off-Duty
ML model	22.2% Annual true positives = 53	19.1% Annual true positives = 116
Rank by complaints in past two years	16.4% Annual true positives = 36	13.1% Annual true positives = 79

Notes: The statistical accuracy of machine learning models compared to the statistical accuracy of a "rank by prior complaints" policy, where accuracy is measured by recall (the percentage of officers who committed misconduct in years T+1 or T+2 who flagged in year T) what flagging the top 5% of officers by estimated risk or the number of prior complaints (ties are broken randomly when flagging by prior complaints in order to flag exactly 5% of officers, and then results are averaged over 10 iterations of tie-breaking). We also show the number of true positives (officers who committed misconduct in years T+1 or T+2 who flagged in year T). While the machine learning models outperform the RBC policy, it's notable that such a simple policy achieves a level of accuracy that is in the ballpark of much more complex models.

Table 6: Comparison of original and residualized risk models

Comparison of the fully-residualized and original on-duty misconduct risk model, flagging at 5%		
	Number of officers per year	% of officer with on-duty misconduct in the outcome period
Flagged by both models	456	11.6%
Flagged only by original model	176	10%
Flagged only by residualized model	176	6.3%

Notes: Comparison between the original on-duty risk model and the residualized on-duty risk model. The flags for the residualized risk model are computed by first regressing on-duty risk score on an officer's activity measures and assignment, and then ranking officers by the difference between their actual risk score and the predicted risk score from the activity and assignment regression. The table shows that most officers (72%) flagged by the original risk models are also flagged by the residualized risk model (specifically 456 of the 632 (456 + 176) officers are flagged by the original model are also flagged by the residualized model), showing that their estimated level of risk is high even after accounting for their activity and assignment. Furthermore, the results show that officers who are only flagged by the residualized model have a lower rate of future on-duty misconduct relative to officers flagged only by the original risk model. In sum, removing the correlation between risk and activity/assignment has a minor effect on who is flagged, and where it does make a difference, it results in models that are less accurate at predicting future on-duty misconduct.

Table 7a: Performance of on-duty risk model by officer race/ethnicity

Performance of on-duty risk model by officer race/ethnicity				
	All Officers	White, Non-Hispanic Officers	Black Officers	Hispanic Officers
Base rate of outcome	1.8%	1.6%	2.1%	2%
Flagging Rate	2%	2.1%	1.5%	2.2%
Rate of future misconduct among flagged officers	10.8%	10.5%	13.8%	10.4%

Table 7b: Performance of off-duty risk model by officer race/ethnicity

Performance of on-duty risk model by officer race/ethnicity				
	All Officers	White, Non-Hispanic Officers	Black Officers	Hispanic Officers
Base rate of outcome	4.8%	4%	6.8%	4.7%
Flagging Rate	2%	1.4%	3.6%	1.7%
Rate of future misconduct among flagged officers	25.0%	24.4%	25.4%	26.5%

Notes: Flagging rates and predictive performance across officer race/ethnicity for the on-duty model (top table) and off-duty model (bottom table). Each table shows the misconduct rates by officer race/ethnicity, the flagging rates across race/ethnicity when flagging the top 2% of officers by predicted risk, and the rate of future misconduct among flagged officers by race/ethnicity. The on-duty risk model performs nearly identically across race/ethnicity, with similar flagging rates across and similar rates of future misconduct among flagged officers.

The off-duty model is more likely to flag Black officers, but those flags are as accurate in the sense that flagged Black officers are equally likely to have future off-duty misconduct relative to other officers. Whether these observed differences reflect true differences in rates of off-duty misconduct or instead reflect biased data cannot easily be determined. One possibility is that the bar for making an off-duty complaint against a Black officer is lower than for other officers. We don't find evidence of this concern (recognizing that we cannot test this theory perfectly given the limits of the data); off-duty complaints received by both flagged Black and flagged White officers are sustained at a rate of 28.3% and 28.9%.

Table 8: Overlap of on-duty and off-duty risk

		Level of risk for future off-duty misconduct (Rows sum to 100%)			
	Rate of future off-duty misconduct	Low (Less than average; <5% chance of future off-duty misconduct)	Average (1-2x average; 5-10% chance of future off-duty misconduct)	Elevated (2-3x average; 10-15% chance of future off-duty misconduct)	High (More than 3x average; > 15% chance of future off-duty misconduct)
Officers in the top 2% of on-duty risk	17.6%	1%	30%	28%	41%
All Officers	4.8%	68%	25%	4%	3%

Notes: Distribution of off-duty misconduct risk among officers in the top 2% of on-duty risk. This analysis shows that officers with the highest on-duty risk also have greatly elevated rates of off-duty misconduct risk, with nearly 70% of high on-duty risk officers having either an elevated or high level of risk of off-duty misconduct. This finding has two important implications. First, these results suggest that some officers face multiple dimensions of risk simultaneously and efforts to improve officer wellness (such as addressing PTSD, stress, or substance abuse) could be a useful part of the efforts to improve risk and misconduct management. This would seem to be an important open question for future research. The second implication is that effective interventions may not be a 'one-size-fits-all' solution. Officers who are high-risk for both on-duty and off-duty misconduct may require a different approach than officers who only have high on-duty risk. To effectively address these different challenges, departments will both need a suite of interventions to address different root causes, and a process for determining which interventions are appropriate for each officer.

Table 9: Government cost-savings estimates from intervention targeting

	Savings (complaint + litigation) from targeting versus targeting-at-random baseline			
	Machine Learning Estimate Cost = 5,000,000		Rank by Complaints (RBC) Estimated Cost = 500,000	
Flagging Rate	Annual savings	5 year savings	Annual savings	5 year savings
1% 127 flags per year	$414,101	$2,070,505	$336,688	$1,683,440
2% 253 flags per year	$638,105	$3,190,525	$551,641	$2,758,205
5% 632 flags per year	$1,035,853	$5,179,265	$882,773	$4,413,865

Notes: Estimated government cost savings from targeting a preventative intervention using a machine learning or the rank-by-complaints policy instead of targeting at random. We estimated the cost of misconduct to the government based on the expenses associated with conducting complaint investigations, the expenses of representing the city in lawsuits against CPD, and the payouts made to plaintiffs in lawsuits (see the appendix for more details on these cost estimates). For the sake of this simulation, we assume that the targeted intervention reduces the misconduct rate of officers that receive the intervention by 20% and hence lowers misconduct costs by 20%. The table above shows the additional government savings from reduced misconduct costs when allocating the preventive intervention using a machine learning model (the first two columns) or the rank-by-complaints policy instead of random targeting (see the appendix for a comparison against an alternative baseline policy of targeting interventions to officers based on prior sustained complaints).

The savings from the machine learning model are higher than the rank-by-complaints policy due to the increased accuracy of the ML model but it's notable that they are close. For instance, at a 2% flagging rate, the estimated 5-year savings from the ML model is 3.2 million while its nearly 2.8 million for RBC. Given the large difference in estimated costs of building the two systems, the RBC model is more likely to generate a high (or infinite) marginal value of public funds.

A. Methods and robustness checks

A1: Dataset descriptions and feature construction

All of the data used for this research was exported from CPD's administrative data systems. Below we list the databases that were used in this work and the type of features (covariates) that we created from each dataset.

Complaint data includes records of all formal complaints lodged against CPD officers. Complaints can be filed by the public as well as internally by other members of CPD. Complaints are investigated either internally by CPD's Bureau of Internal Affairs or by Chicago's Civilian Officer of Police Accountability (COPA); the nature of the complaint determines which body conducts the investigation. A single complaint can name multiple officers and each officer can have multiple misconduct allegations. Complaint and allegation categories fall into 15-20 high-level categories[42] (e.g., "Verbal Abuse", "Operational Violations", "Excessive Force", etc) and approximately 80-100 specific categories. The most common category is excessive force, followed by 'operational/personnel violation'. If a complaint is sustained (meaning that the allegations are proven to have occurred and the officer's actions were not in line with CPD's policy), the officer will receive a punishment that ranges from 'violation noted' all the way to suspensions or termination. We created features from prior complaints that measured the total count of received complaints, the count by high-level categories, counts by detailed categories, counts by investigatory outcomes (e.g. how many sustained findings, how many 'exonerated' findings, etc), counts by penalty types (eg how many 'violations noted', how many suspensions, count of suspension days, etc), and finally counts by category and finding ("number of sustained excessive force complaints", etc). We also created features measuring the penalties associated with the disciplinary outcome, which range from "violation noted" through suspensions and terminations.

*SPAR*s (Summary Punishment Action Report) are lower-level internal transgressions recorded by a supervisor after observing a policy violation. The most common SPAR types are failure to appear in court and a low-level vehicle crash. Unlike complaints, SPARs are not investigated and are assumed to be true by default, but officers can appeal if they feel the SPAR was unwarranted. The possible outcome of a SPAR ranges from a 'violation noted' to a suspension of up to 3 days. Similar to the scheme we used for complaints, we created features based on total counts of SPARs, counts of SPARs by type, and counts of SPARs by penalty type.

Use of force reports are records filled out by police officers any time they use force on a subject, where force can range from tactics like 'emergency handcuffing' all the way to the use of a service weapon. Each report includes a series of checkboxes to indicate the actions of the officer

[42] https://directives.chicagopolice.org/forms/CPD-44.248.pdf

as well as the actions of the subject, as well as other information about the nature of the associated arrest (if any), the time of day, weather conditions, etc. Use of force reports are classified into three possible levels based on the severity of force used. We created features from prior complaints based on the total count of use of force reports, the count by force level, counts by specific force tactics (e.g. “number of use of force reports with a taser use”, etc), and counts by the nature of the arresting charges (eg “number of use force reports where the arresting charge was a misdemeanor” or “number of UoF reports where the arresting charge was ‘resisting arrest’’).

Attendance records are records of every day an officer has spent employed by CPD, with an indicator for whether the officer was present or absent each day, whether the absence was an unexcused absence (i.e., not appearing for work without approval and without a medical reason), and the reason for the absence (if excused). We created features based on the number of days worked in a certain time period, number of excused absences, number of unexcused absences, and number of absences.

Overtime records include the number of overtime hours worked in a given period. We created features based on the total number of overtime hours worked in a given period, as well as counts by different types of overtime (e.g., whether the overtime hours were from working additional hours at the end of a typical shift or whether they were for coming in on a day that the officer would normally be off).

Arrest records record each arrest an officer was involved with, the charges associated with the arrest, and the role of the officer on the arrest (first arresting officer, secondary arresting officer, and assisting arresting officer). We created features for the number of arrests each officer made by charge type (misdemeanor vs felony, and then counts by each specific charge types), by arresting role, and by whether the charge types were ‘discretionary arrest charges (likes ‘resisting arrest’ or ‘disturbing the peace’), which prior research have identified as potential risk factors.

Activity records include the number of traffic stops, investigatory stops, warrants issued, and awards received by an officer. We created features based on the counts of each activity type over a given time period.

Unit and assignment records list the unit that each officer was assigned to (e.g. a geographic unit like “District 1” or a specialized unit like “Narcotics enforcement”) and their role (e.g. “Police officer”, “Sergeant”, etc). We created a feature based on what unit and role each officer was assigned to at the time of prediction.

Lawsuit records from CPD were excluded from this analysis because they were not consistently tracked in CPD’s administrative data systems over the research period.

Using the datasets listed in above, we constructed a large set of features (covariates) for the models to use as potential predictors of future misconduct. We constructed a set of features for different time horizons - events that occurred within the past year, events that occurred in the past two years, and events that occurred in the past five years. These time windows are overlapping by construction - all of the events counted by 'past year' features are also counted in the past two and five year features. The use of multiple time horizons allows the models trade off between completeness and recency, as well as allowing the models to handle observations where the officer has less than five full years of data.

The table below summarizes the features that we created from each data source. Each data source has a set of 'simple' features that capture basic counts of prior events. For data sources that are more complex, like complaints and use of force, we created features that measure prior events by 'type'. Features in the 'intermediate' category use types are fairly broad (like "prior excessive force") while features in the 'complex' category use very fine-grained types ('prior sustained allegations of excessive force with a weapon').

	Feature Complexity		
All datasets ~1000 features per time horizon	**Simple** ~30 features per time horizon	**Intermediate** ~150 features per time horizon	**Complex** ~800 features per time horizon
Complaints ~700 features per time horizon	Count of complaints, Count of penalties associated with complaints	Counts by high-level complaint category *Example:* *Prior complaints alleging excessive force,* *Prior complaints alleging conduct unbecoming of an officer*	Counts by detailed categories, counts by category and investigation *Prior complaints alleging excessive force with use of a weapon,* *Prior sustained complaints alleging failure to activate body-worn camera,* *Open complaints alleging conduct unbecoming of an officer*
SPARs ~50 features per time horizon	Count of SPARs, Count of penalties associated with SPARs	Counts by high-level category *Example:* *Prior SPARs for missing court*	
Use of force ~100 features per time horizon	Count of use of force	Counts by CPD force categorization *Prior level 1 use of force*	Counts by specific tactics *Prior use of force with open-handed strike*
Attendance records ~60 features per time horizon	Count of days worked, Count of days absent	Counts by reason for absence	

Overtime records ~5 features per time horizon	Count of overtime hours worked	Counts by overtime hours worked by type	
Arrest records ~45 features per time horizon	Count of arrests	Counts by arrest type, Counts by role (primary arresting officer or second arresting officer), Counts of discretionary arrests	Counts by arrest type and
Activity records ~6 features per time horizon	Count of investigatory stops, traffic stops, warrants served, and awards received		
Unit and assignment ~2 features per time horizon	Most recent unit of assignment, most recent role		

A2. Choice of prediction outcome

In this section, we provide further detail on different prediction outcomes. In Section 2, we defined four possible misconduct outcomes based on whether a complaint was an on-duty (e.g. excessive force) or off-duty (e.g. altercations while off-duty), and whether the complaint was sustained or not. We first show that predicting 'all off-duty complaints' is better than predicting 'sustained off-duty complaints' because the 'all off-duty model' ends up being a better predictor of both outcomes. We trained two models - one in which the outcome variable was 'all off-duty complaints' and one in which the outcome variable was 'sustained off-duty complaint' - and measured how well those models predict each outcome. The results of this exercise, shown in the table below, show that the 'all off-duty model' is a better predictor of 'all off duty' complaints and a nearly equal predictor of 'sustained off duty' complaints. In other words, regardless of which outcome is the preferred one, the 'all off-duty model' is a weakly better predictor of that outcome.

	'All off-duty' outcome	'Sustained off-duty' outcome
Model that predicts all off-duty	Recall@5%: 19.1% AUC: .682	Recall@5%: 21.3% AUC: .682
Model that predicts sustained off-duty	Recall@5%: 18.2% AUC: .671	Recall@5%: 22.3% AUC: .682

On the other hand, the choice of whether to predict ‘all on-duty complaints’ or ‘sustained on-duty complaints’ is less clear. The table below shows that the sustained on-duty model is a slightly better predictor of the ‘sustained on-duty’ outcome while the ‘all on-duty outcome’ is a better predictor of the ‘all on-duty outcome’. Hence the conclusion of which model is a better predictor depends on the relative value of correctly flagging a few more officers who have a sustained on-duty complaint in the outcome period versus flagging many more officers who have a non-sustained complaint in the outcome period.

	‘All on-duty’ outcome	‘Sustained on-duty’ outcome
Model that predicts all on-duty	Recall@5%: 17.5% AUC: .788	Recall@5%: 20.8% AUC: .741
Model that predicts sustained on-duty	Recall@5%: 15.5% AUC: .761	Recall@5%: 22.2% AUC: .752

The other dimension that we can compare the risk models on is how correlated the risk estimates are with policing activity and assignment (this potential cost is one of the most commonly-cited objections to the use of early intervention systems). Both the sustained off-duty and all off-duty models have the same low degree of correlation with activity and assignment. The all on-duty risk scores have a significantly higher correlation with activity and assignment relative to the sustained on-duty risk scores (shown in the table below). Hence the sustained on-duty model potentially reduces one important dimension of the ‘cost’ of using risk prediction in the policing context. See Section 5b for more detail on this analysis.

	R^2 from regressing on assignment and all activity measures
All Off-duty risk score	.18
Sustained Off-duty risk score	.09
All On-duty risk score	.67
Sustained On-duty risk score	.41

A3: Length of the outcome period

The set-up of the misconduct prediction problem - given what you know at time T, predict whether an officer will engage in misconduct in the near future - requires choosing the length of the outcome period. Shorter outcome periods have the advantage that features measured at time T will be more relevant for the near future relative to the distant future (e.g. events from 2010 are presumably more predictive of events in 2011 than the events in 2015). On the other hand, shorter outcome periods have less time to observe outcomes and hence an overall lower base rate.

We examined this choice empirically by building and evaluating models with different outcome periods, and found that the resulting models perform essentially the same regardless of the length of the outcome period. The tables below show that all models perform essentially the same when holding fixed the outcome period during evaluation. The first column in each table shows that all models have a recall@5% between 22.8% and 23.7% when evaluated on how well they predict 1-year outcomes. Similarly, columns 2 and 3 show that all models have a recall@5% of 21.8%-22.3% for 2-year outcomes and 19.6%-20.4% for 4-year outcomes. We also see evidence that recall decreases as the length of the outcome period increases, which is consistent with our original theory that features at time T will be less predictive of events that happen in the far future relative to the near future.

Recall @ 5% - On-duty Misconduct			
	Evaluation on 1 year outcomes	Evaluation on 2 year outcomes	Evaluation on 4 year outcomes
Train on 1 year outcomes	23.7%	22.2%	19.6%
Train on 2 year outcomes	23.6%	22.3%	20.1%
Train on 4 year outcomes	22.8%	21.8%	20.4%

Recall @ 5% - Off-duty Misconduct			
	Evaluation on 1 year outcomes	Evaluation on 2 year outcomes	Evaluation on 4 year outcomes
Train on 1 year outcomes	21.7%	18.9%	16.2%
Train on 2 year outcomes	21.7%	19.1%	16.4%
Train on 4 year outcomes	20.5%	18.2%	16.1%

A4: Machine learning model selection

We tested three different machine learning models - gradient boosted trees, random forests, and regularized logistic regression. For each of these models, we tuned hyperparameters using grid search and 3-fold cross-validation. We opted for 3-fold cross-validation, as opposed to the typical 5 fold, to reduce the computational burden of cross validation search because hyperparameter tuning is nested inside our cross-fitting procedure (the cross-fitting procedure itself is described in Section 2) and hence is called many times. All cross-validation folds are split by officer ID to ensure that the selected models generalize well to officers not in the training folds.

For gradient boosting, we tuned the maximum depth of each tree over a grid of [1, 3, 5], and tuned the learning rate over a grid of [.01, .1, .5]. We set the number of trees to be 500 and used early stopping to prevent overfitting. For random forests, we set the number of trees in the random forest at 200,[43] and tuned the fraction of features tested at each node over a grid of [.1, .5] and the depth of each tree over a grid of [3, 5, no maximum depth]. For elastic net regression, we tuned the weight on the penalty term over a grid of [.01, 1,10, 1000, 10000] and the L1 ratio at [.1, .5, and .9]. All parameters not mentioned were kept at the default values as set by scikit-learn.

The table below compares the performance of gradient boosting, random forests, and elastic net regression, as well as an ensemble of all three models (the prediction for the ensemble model is simply the average of the predictions from all three models). While all models perform similarly, gradient boosting performs the best.

[43] In principle, setting the number of trees to a higher number could yield more accurate models but practical studies have shown that the gain in relative accuracy from hundreds of trees to thousands of trees tends to be quite small and might not outweigh the cost of additional computation (Probst & Boulesteix 2017).

Comparison of machine learning methods		
	Predicting on-duty misconduct	
	Recall @ 5%	AUC
Gradient Boosting	**22.2%** [20%, 24.1%]	**.752** [.740, .768]
Random Forest	19.9% [17.4%, 22.4%]	.726 [.714, .741]
Elastic Net (regularized logistic regression)	20% [17.7%, 22.8%]	.727 [.715, .740]
Ensemble	21.9% [19.2%, 22.4%]	.744 [.732, .759]
	Predicting off-duty misconduct	
	Recall @ 5%	AUC
Gradient Boosting	**19.1%** [17.6%, 20.3%]	**.682** [.671, .688]
Random Forest	18% [16.5%, 19.3%]	.666 [.656, .675]
Elastic Net (regularized logistic regression)	18.5% [16.9%, 19.7%]	.671 [.660, .681]
Ensemble	18.7% [17.3%, 19.9%]	.678 [.668, .686]

A5: Hyperparameter tuning

In this section we present a sensitivity test for whether the predictive accuracy of our machine learning models are sensitive to the choice of procedure used to tune hyperparameters. Specifically, we test whether using a grid search procedure over a limited set of hyperparameters results in models that are less accurate relative to a procedure that efficiently optimizes over a large set of hyperparameters. For the latter procedure, we use a recent technique known as FLAML (fast-and-lightweight automated machine learning; Wang, Wu, Weimer, and Zhu 2021) that searches the space of hyperparameter configurations through a weighted random sampling technique that proportionally samples and tests new configurations based on an estimate of "accuracy gain per computation time" – essentially decreasing the time required to find a good set of hyperparameters. The decreased time to find good hyperparameters means that we can optimize over all hyperparameters rather than just a subset.

For this experiment, we built competing models to predict on-duty misconduct (the results for off-duty are essentially the same). We allowed FLAML to search over two implementations of gradient-boosting – LightGBM (Ke et al 2017) and XGBoost (Chen and Guestrin 2016) - and

give it a time budget of 60 minutes to tune parameters. FLAML searches over 8-9 different hyperparameters for LightGBM and XGBoost, as compared to the 2 hyperparameters that we searched over during grid search (Section A4 describes the grid search procedure).
This experiment is run within the context of our iterated cross-fitting procedure. Recall that in each iteration, we partition the data into 3 sets – P1, P2, and P3. Partitions P1 and P2 are used to fit a model to make predictions for fold P3, P1 and P3 are used to fit a model to make predictions for P2, and P2 and P3 are used to fit a model to make predictions for P1. We repeat this procedure over 10 iterations with different random partitions each time. This training scheme allows us to assess the difference between FLAML and grid search in two ways – either evaluating the average predictive accuracy over the 10 iterations or evaluating the predictive accuracy of the averaged predictions across the 10 iterations. Although these methods seem similar, the crucial difference is that the aggregated predictions across the 10 iterations benefit from a variance reduction and hence we should expect them to be more accurate than the models produced over the 10 different iterations.

The table below shows the results of this experiment where the accuracy of each method is evaluated both by AUC and by recall@5% (the bolded values indicate the best performance within each column). For both methods, the aggregated predictions are moderately more accurate than the median accuracy over the 10 iterations - demonstrating the value of the variance reduction from aggregating the predictions. The FLAML model has a better AUC than the grid search model but the differences are fairly small. The FLAML model also has a better recall@5% when evaluating at the median iteration but the grid search model has a better recall@5% after aggregating predictions. Again, all differences are small.

Method	**Time per iteration**	**AUC - On duty misconduct**		**Recall @ 5% - On duty misconduct**	
		Median iteration	Aggregated predictions	Median Iteration	Aggregated predictions
FLAML	60	**.7497**	**.7538**	**21.83%**	21.9%
Grid Search	15	.7425	.7520	20.48%	**22.2%**

Overall, the results of this experiment show that a more principled hyperparameter tuning method can yield better results but the gains are quite small, particularly when aggregating predictions over the 10 iterations of cross-fitting. As a result of these experiments, we use the standard grid search method throughout the paper to save on computational budget.

A3: Robustness to population specification

In this work, we constructed our sample by pooling observations from all active officers in each year. One concern with this modeling choice is that using all officers creates a very heterogeneous population - e.g. officers with 25+ years of experience might have very different patterns than officers with 5 years of experience. Although our models can theoretically control for those patterns because years of experience is included as a feature in the risk models, there's no guarantee that simply including those features is enough to adequately control for population heterogeneity. Hence it is possible that we are understating predictability among key sub-populations.

We tested whether our results were sensitive to the choice of using all officers in the modeling sample by repeating the analysis only using data from officers with 5-15 years of experience at the time of prediction. Specifically, we created another machine learning model that was only trained on officers with 5-15 years of experience and checked whether that model more accurately predicted misconduct among officers with 5-15 years of experience than the full model did. We conducted this sensitivity analysis using both on-duty and off-duty misconduct.

The table below shows that the model built on the full population is a more accurate predictor of misconduct among officers with 5-15 years of experience than the model whose training data was limited to officers with that experience level for both on-duty and off-duty misconduct. This is likely due to the fact that it has much more data to learn from, that data from other subgroups are relevant to the target population, and that it is able to model any interactions between years experience and other features due to the fact that years experience is included in the model. In short, the heterogeneity in the full population does not cause the model to be less accurate than a model that is built on a more homogenous population

		Evaluated on officers with 5-15 years of experience	
	Number of observations in dataset	Recall @ 5% - On-duty misconduct	Recall @ 5% - Off-duty misconduct
Trained on full population	113,768	18.3%	17.2%
Trained on officers with 5-15 years of experience	37,947	16.9%	17.1%

A different way to test the sensitivity of our results is to compare the accuracy of the full model on the full population against the accuracy of the full model on a restricted sample, like officers

with 5-15 years of experience. The table below shows the accuracy of the on-duty risk model (the one built on the full sample) when evaluated on the full population and when evaluated only on officers with 5-15 years of experience (we omit the result for the off-duty model for brevity but the conclusions are the same). This sensitivity check shows that the performance is slightly more accurate when applied to the full population, implying that the model is able to find high risk officers outside of the group of officers with 5-15 years of experience.

	On-duty Misconduct				
Evaluation Sample	Base Rate	AUC	Recall @ 5%	Precision @ 5%	Lift (Misconduct rate of flagged officers divided by base rate)
Full population	1.8%	.752	22.2%	.083	4.5x
Officers with 5-15 years of experience	1.9%	.733	18.3%	.074	3.8x

We finally conducted one other sensitivity test where we removed supervisors (sergeants, lieutenants, etc) from the sample and tested how flagging the top 250 non-supervisors compared to flagging the top 250 officers from the full sample (i.e. including supervisors). Similarly, the table below shows that prediction accuracy on the full sample is higher than on the restricted population.

	On-duty Misconduct				
Evaluation Sample	Base Rate	AUC	Recall @ 5%	Precision @ 5%	Lift (Misconduct rate of flagged officers divided by base rate)
Full population	1.8%	.752	22.2%	.083	4.5x
Population w/supervisors removed	1.9%	.754	21.9%	.085	4.5x

In sum, our choice to use the full population of officers does not cause us to underestimate predictability. Although the population heterogeneity was a concern in principle, the machine learning model is able to accurately model subgroup effects when trained on the full population.

A3: Robustness to definition of on-duty misconduct

Our definition of on-duty misconduct includes allegations that fall into any of the following categories: excessive force, improper arrest or search, verbal abuse, coercion, search warrant incident, arrest and lockup incident, bribery or official corruption, and weapon discharge. The intent of this outcome was to capture a wider variety of potential harms when a CPD officer was carrying out some sort of enforcement action.[44] The potential downside of using this wider index is that it reduces the focus on what is arguably the most harmful outcome - sustained allegations of excessive force.

In this section we test whether our results are sensitive to using this wider index of on-duty misconduct versus solely focusing on sustained excessive force allegations. We conducted this sensitivity test by retraining a model that solely predicted sustained excessive force allegations and comparing whether that model better predicted future sustained excessive force allegations than the general on-duty misconduct model. The table below shows the on-duty misconduct model when evaluated for well it predicts on-duty misconduct (first row, as a baseline) and how well it specifically predicts sustained excessive force allegations (second row). The sustained excessive force outcome is more rare than the general on-duty misconduct outcome - the base rate drops from 1.8% to .5%, and as a result, the precision of the on-duty misconduct model drops when evaluated on sustained excessive force allegations due the narrow definition of the outcome variable. However the AUC, recall, and lift of the on-duty misconduct model are all higher when only evaluated on the sustained excessive force outcome.

The third row shows the performance of the model constructed to exclusively predict sustained excessive force allegations - that model is slightly more accurate at predicting sustained excessive force allegations relative to the general on-misconduct models when evaluated on precision or recall but less accurate when evaluated on AUC.

Training	Evaluation	Base	AUC	Recall @	Precision	Lift

[44] This definition excludes on-duty complaints like failure to take action, failure to file a report, activate a body camera, etc. We omitted these allegation categories from the definition so we could focus the models on the most acute harms.

outcome	outcome	Rate		5%	@ 5%	(Misconduct rate of flagged officers divided by base rate)
On-duty misconduct	On-duty misconduct	1.8%	.752	22.2%	.083	4.5x
On-duty misconduct	Sustained excessive force allegations	0.59%	.765	24.3%	.029	4.9x
Sustained excessive force allegations	Sustained excessive force allegations	0.59%	.758	25.8%	.031	5.2x

In summary, the decision to use the broader on-duty misconduct outcome doesn't really lead us to overstate the predictability of the most serious outcomes (sustained excessive force allegations), nor did it suppress the model's ability to predict these events.

B. Additional Results

B1. Additional measurement error results

In Section 3, we argued that a series of reforms in Chicago and the Chicago Police Department around 2016-2017 caused a decrease in the amount of measurement error in the on-duty misconduct outcome because true incidents of on-duty misconduct were more likely to result in a sustained on-duty complaint. We then showed that machine learning models trained on data from before those reforms (the 'early' data) were less accurate than models trained on data from after those reforms (the 'late' data). We now argue that the change in measurement error is the best explanation for the relative performance of the early and late model by ruling out two alternative explanations for why the late model is better than the early model.

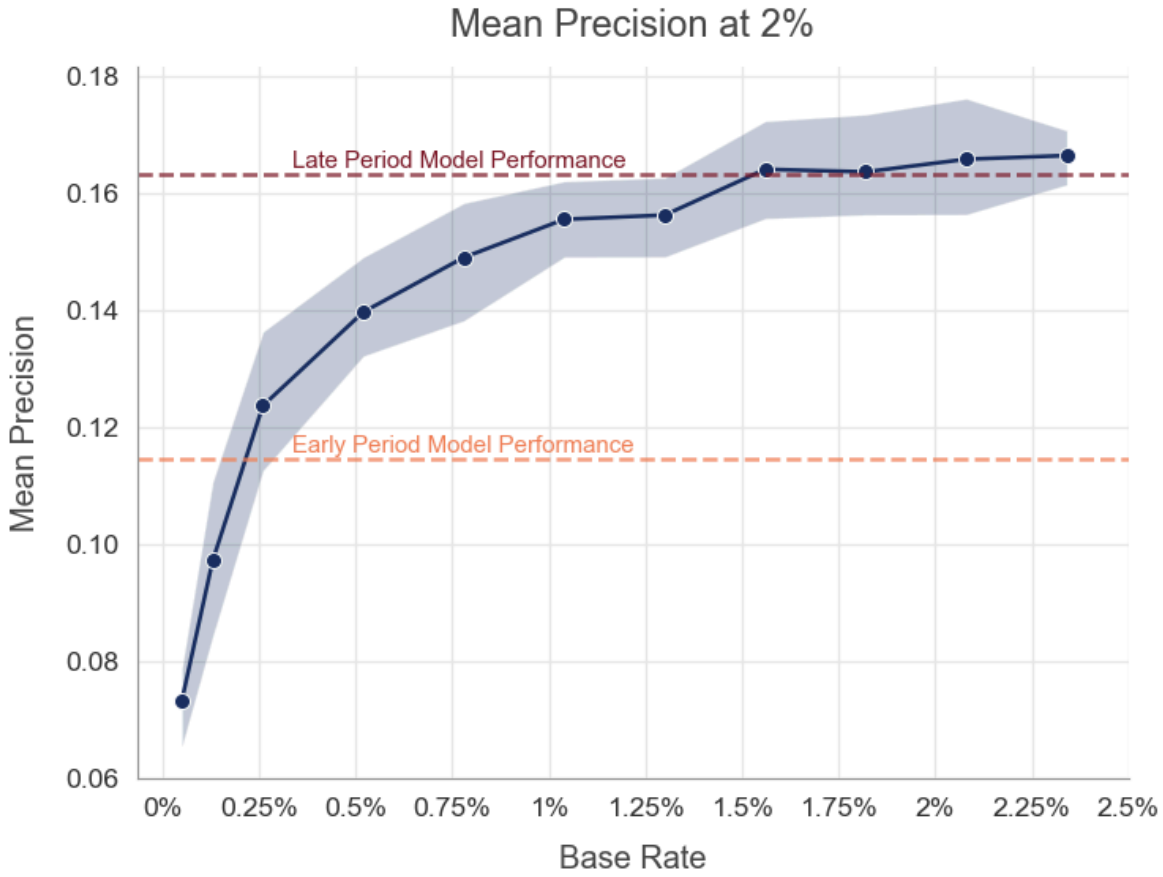
The first alternative explanation is data drift[45], a term from the computer science literature that refers to systematic differences between the data used to build a machine learning model and the data that the ML model is applied to in production. One of the most common reasons for data drift is the time difference between when a model is trained and when it is used. Recent work in machine learning has demonstrated that data drift can cause loss in accuracy. It's possible that data drift is the cause of our finding that the early model performs worse than the late model on data from the later period, rather than our theory about change in measurement error.

If data drift were the best explanation for the relatively weak performance of the early model, then we'd expect to see similar results when we repeat the 'early vs late' experiment for the off-duty outcome. The first row in the table below shows performance of the 'early' and 'late' models when they are trained to predict off-duty misconduct rather than on-duty misconduct and are evaluated on observations from 2017. While the late off-duty model is a better predictor than the early off-duty model, the difference in performance between the two models is much smaller than the difference in the early and late on-duty models (shown in the second row of the table). Hence the large difference in performance of the on-duty models is likely to be solely due to a generic explanation of data drift.

[45] This term is a catch-all that captures concepts like covariate shift, label shift, concept drift, etc. See Rabanser, Gunnemann, and Lipton (2019) (and references therein) for recent work on this topic.

	Rate of future misconduct (observations from 2017)	
Type of misconduct	Flagged by early model (2% flagging rate)	Flagged by late model (2% flagging rate)
Off-duty	21.9%	25.9%
On-duty	11.7%	23.5%

The second alternative explanation is that the late model is better because there are more y=1 instances in the late data to learn from. Indeed, the base rate of on-duty misconduct in the later data is 2.6% versus a base rate of 1.4% in the early data. We tested whether the difference in base rates is a better explanation for the superior accuracy of the late model by randomly flipping y=1 cases to 0 (hence lowering the base rate) and retraining models on the modified data. The plot below demonstrates that models trained on the modified late data are more accurate than the earlier data even as we artificially decrease the base rate from 2.6% (the original base rate) all the way down to .25%. These results show that the difference in base rate is not a good explanation of the superior performance of the late on-duty models.



B2. Evaluating on a broader ‘negative event’ outcome

Our primary evaluation of each risk model measures how well they predict the specific outcome that they were trained to predict, i.e. we evaluated the on-duty misconduct model based on how well it predicts on-duty misconduct. This approach is conservative because, for example, if an officer flagged for on-duty misconduct has an off-duty complaint (even one that is sustained or results in discipline), we would not count that as a successful prediction. This conservative approach potentially leads to understating the level of risk among flagged officers. For example, our finding that only 12% of officers in the top 2% of on-duty misconduct risk actually engage in on-duty misconduct might seem to imply that 88% of flagged officers have no negative outcomes, but that would be ignoring all other types of negative events those officers may have engaged in.

We addressed this issue by defining a broader outcome called ‘any negative event’ which measures the number of sustained complaints, off-duty complaints, or suspensions that the officer receives over the target period. This is a broader outcome than on-duty or off-duty misconduct in several ways because it counts both on and off-duty misconduct, includes sustained complaints for operational things like ‘failure to activate body worn camera’ that were previously count counted, and includes suspensions from low-level transgressions (known as SPARs) like missing a required court appearance. We also treat ‘any negative event’ as a count variable rather than a binary indicator because officers semi-frequently have more than one negative event over the target period.

The table below shows that officers flagged by the on-duty and off-duty risk models have a rate of future negative events that is significantly higher than just their rate of future on-duty or off-duty misconduct. For example, while 12.7% of officers flagged for on-duty misconduct have a future instance of on-duty misconduct, 27% of them have some negative event in the future. The far right column shows the rate of future negative events, which accounts for officers that have multiple negative events during the target period[46], for each group and shows that many officers flagged for on-duty or off-duty misconduct have multiple future negative events.

[46] We ensure that we don’t ‘double count’ any single event, e.g. if an officer had an off-duty complaint that was also sustained, we only count that as a single negative event.

	% of population with future on-duty misconduct	% of population with future off-duty misconduct	% of population with future negative event	Rate of future negative events (# of negative events per 100 officers)
All officers	1.9%	5%	9%	11
Top 2% by on-duty misconduct risk	12.7%	–	27.3%	49.2
Top 2% by off-duty misconduct risk	–	36.6%	41.5%	62.9

B3. Focus on patterns, not events

In this section, we expand upon the analysis in Section 4 that argued that a good heuristic for thinking about risk prediction is to focus on patterns, not events. Specifically, we argue that officers who have a "frequent" pattern of prior events (where frequent is defined as having more of that event than most officers) are more likely to be involved in a future negative outcome relative to officers who had a "serious" version of that event. While it might be natural for departments to focus on officers who had a proven or egregious prior serious event for the purposes of reactive risk management (e.g. ensuring there are appropriate sanctions or responses to proven instances of misconduct), we find that proactive risk management should instead focus on officers who have the most prior events, even if none of them are proven to be serious.

To make this concrete, we compared three different "frequent" and "serious" policies, as shown in the table below. In each case, the "serious" policy flags any officer who has a serious event in the recent past, while the "frequent" policy flags an equivalently-sized group of officers with the most prior events. For example, the first row in the table below contrasts the policy of flagging any officer with a sustained complaint in the past five years with the policy of flagging officers with the most complaints over the past five years. We then measure the rate at which each of those groups have a sustained complaint in the future.

This comparison shows that, across a variety of negative outcomes, officers flagged by the "frequent" policy are more likely to have a future serious event than officers with a prior serious event (see the table below). For instance, officers with the most prior lawsuits are roughly twice as likely to have a future expensive lawsuit as compared to officers with a prior expensive

lawsuit. The heuristic of "focus on patterns, not events" seems to be robust across a number of outcomes and settings.

Serious event type	"Frequent" event type	Rate of future serious outcome among all officers	Rate of future serious outcome among "serious" flagged	Rate of future serious outcome among "frequent" flagged
On-duty sustained complaint (CPD)	All complaints	1.9%	3.2%	4.9%
Sustained CCRB complaint (NYPD)	All CCRB Complaints	3.1%	7.5%	10.8%
Expensive lawsuit[47] (NYPD)	All lawsuits	1.6%	4.0%	8.1%

B4. Risk and and police activity

In Section 5b we showed that removing the correlation between risk scores and activity/assignment through a residualization procedure only had a minor effect on which officers were flagged as the highest risk. The implication of that result is that officers who are high risk in absolute sense (before any adjustment for activity and assignment) are also high risk relative to the other officers with similar assignment or activity.

In this section, we give more details on the regression used in the residualization procedure. We let $r_{i,t}$ denote the risk score for officer i and time t, $A_{i,t}$ denote a vector capturing an officer i's assignment (unit assigned and position) at time t, $W_{i,t}$ denote a vector capturing their activity in the two years leading up to time t, and Z_t be a year fixed effect. We then run the following regression:

$$r_{i,t} \sim A_{i,t} + W_{i,t} + Z_t$$

[47] An expensive lawsuit is defined as a lawsuit with a total cost of more than 50,000 thousand dollars. See Section C in the appendix for more details about this definition and the NYPD lawsuit data.

This regression enables us to do two things: measure the relationship between risk scores and activity/assignment, and compute residualized risk scores. We can also change the set of activity measures in the regression depending on what kind of correlations policymakers might be concerned about. For instance, some policymakers believe that many low-level arrests are indicative of a higher risk officer and hence wouldn't be concerned about correlation between risk scores and the number of low-level arrests.

The table below shows the correlation between activity/assignment and three different risk scores: on-duty misconduct risk, off-duty misconduct risk, and risk score from a model that was designed to predict any future on-duty complaint (regardless of whether it was sustained or not). We tested two different activity measures - the "all activity" measures that includes all arrests, traffic stops, investigatory stops, guns recovered and awards received, and "limited" activity set that only includes felony arrests, guns recovered, and awards received.

These results show that most of the variation in on-duty and off-duty misconduct risk are not explained by assignment and activity (i.e. the assignment/activity regressions explain less than 50% of the variation in risk scores for the on-duty and off-duty misconduct models). Furthermore, that explained variation reduces slightly when using the limited set of activity metrics. These results also show that a model that predicts risk of any future on-duty complaint has a significantly higher correlation with activity/assignment than the model that only tries to predict future sustained on-duty complaints.

	Correlation between risk and activity/assignment	
	Assignment + "all" activity	Assignment + "limited" activity
On-duty risk (Risk of a future sustained on-duty complaint)	.41	.36
Off-duty risk	.18	.15
Risk of any future on-duty complaint	.67	.61

We further explored these results by first measuring the correlation between risk scores and assignment, and then measuring the correlation between risk score and activity after accounting for assignment. We first measure the correlation of risk scores and assignment, and then activity and assignment, by regressing both risk scores and activity measures on assignment. We find that assignment by itself explains 24% of the variation in risk scores, and between 22-50% of the

variation in activity measures (shown in the table below). Put differently, most of the variation in both risk and activity seems to be within, not between, work assignments.

	R^2 from regressing on unit, position, and year
Off-duty risk score	.11
On-duty risk score	.24
Arrests	.50
Felony arrests	.43
Guns Recovered	.22
Investigatory stops	.37
Traffic stops	.25
Department awards	.32

We next measure the correlation between activity and risk by first residualizing the effect of assignment and role on both risk scores and activity measures, and then regressing residualized risk scores on residualized activity (results shown in the table below). We find that residualized activity explains 23% of the variation in residualized risk scores for the on-duty models and 9% of the variation for the off-duty model. The fact that activity only explains a quarter of the variation in risk scores (after controlling for position and unit) suggests that risk of on-duty misconduct is not purely a mechanical byproduct of active policing. The choice of activity measures can further reduce that correlation; for example, using only guns recovered, felony arrests, and awards reduces the correlation between residual activity and residual risk to 16%.

	R^2 from regressing on all activity measures	R^2 from regressing on limited activity measures
Off-duty risk score	.09	.05
On-duty risk score	.24	.16

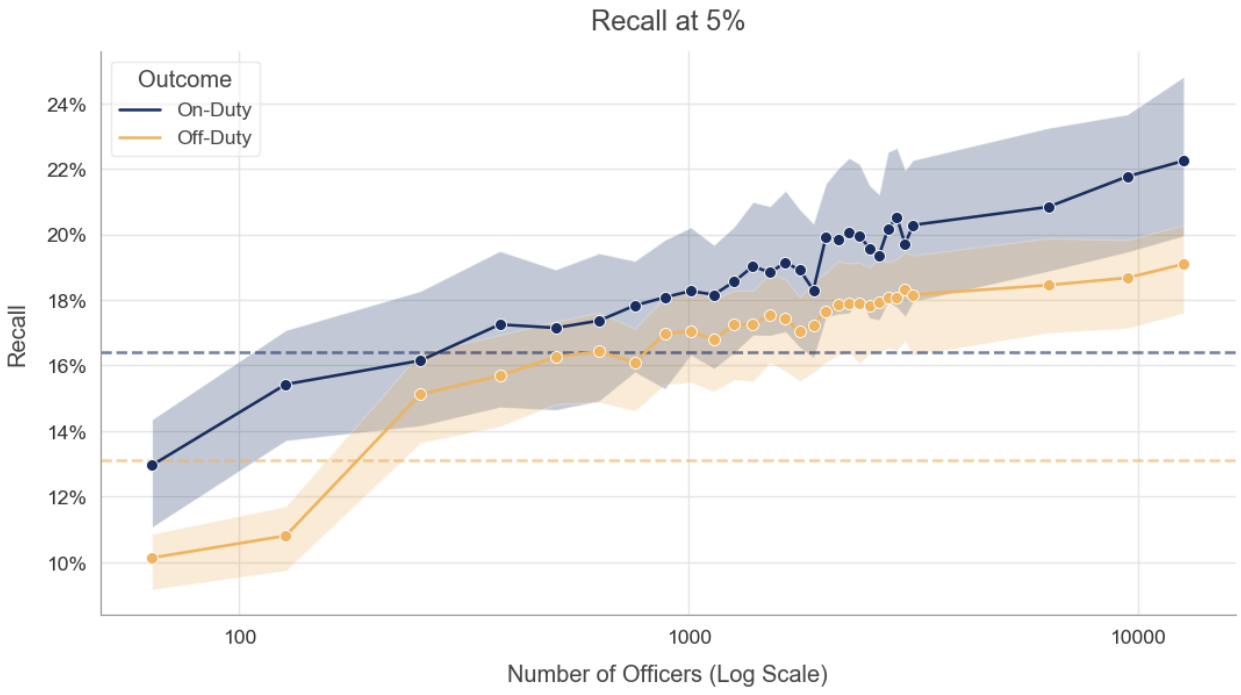
B5. The effect of sample size on model accuracy

In this section, we show how sample size affects the accuracy of machine learning models. The conventional wisdom in machine learning is that more data enables the creation of better models because the additional sample size allows the algorithms to discover subtle relationships between

the features and outcome. We estimated the effect of sample size on model performance by subsampling the Chicago data, training models on the sub-sampled data, and observing the change in accuracy.[48]

We subsampled the CPD data by selecting R officers at random, limiting the training data to only observations from those R officers, and training a model on that subsample. We then used that model to make predictions for the entire dataset.[49] We repeated this sample-and-train procedure 10 times and plotted the recall of the median-performing model (as well as the confidence intervals for that model) for each sampling rate. Results are shown in the plot below. The dashed horizontal lines show the recall of the rank-by-complaints policy for on-duty misconduct (blue line) and off-duty misconduct (yellow line).

This exercise shows that accuracy of the ML models (as measured by recall @ 5%) roughly improves linearly with every order of magnitude increase in sample size. At very small sample sizes (fewer than 200 officers), the machine learning models perform worse than the 'rank-by-complaints' baseline because there's simply not enough data to learn from. For on-duty misconduct, the performance of the ML models is only significantly better than RBC with a sample size of 1000 officers. For off-duty misconduct, the ML model is significantly better when the sample size has ~500 officers.



[48] We used recall @ 5% as our measure of accuracy but the results would be qualitatively similar with other flagging rates and performance metrics.
[49] We use cross-fitting to make out-of-sample predictions for the observations that were in the subsampled data. See Section 2 for details on the cross-fitting procedure.

The above results suggest that most police departments in the United States do not have a large enough sample size to build a machine learning model that is more accurate than a simple policy. Policing in the United States is famously decentralized - there are over 18,000 police departments with 90% of them having fewer than 100 officers. In this exercise, the machine learning model is only clearly better at predicting on-duty misconduct with at least 1000 officers, but only 80 departments in the United States have 1000 officers or more. Put another way, it seems that only large departments should consider building their own risk models - smaller departments may not have the data, nor the scale, to justify the efforts of building a data-driven risk model.

One alternative to building a custom risk model for each department is to use a risk model from another department (or set of departments). The accuracy of this approach depends on how similar the “training” and “customer” departments are, both in terms of true risk patterns and in terms of systems for recording data. While this is a crucial question for the scalability and MVPF of machine learning models for predicting misconduct risk, it is beyond the scope of this paper given the lack of data from a larger set of departments.

C : Replication on NYPD data

We replicated our analysis with public data from the New York Police Department (NYPD) to test if our general conclusions hold in another jurisdiction. Although the NYPD is more limited than the Chicago data in terms of the types of prior datasets available (for example, the public NYPD dataset does not have use of force data), we generally find the same conclusions regarding the levels of predictability, the similar performance between machine learning models and simple ranking policies, and the predictive value of prior non-sustained complaints.

C1: Data, features, and outcomes

The public data includes complaint records from New York City's Citizen Complaint Review Board (CCRB) (which investigates external complaints against NYPD officers), lawsuit data recorded by NYC's law department, and a roster of NYPD officers.[50] We constructed a similarly-structured panel dataset where the features for each observation includes everything known about an officer's behavior up to and including year and the outcomes include whether the officer was involved in any negative outcomes in year T and T+1. We focus on prediction years from 2015-2019.[51] The final dataset has between 33,000 - 36,000 observations per year, resulting in a total of 175,000 officer-year observations.

We constructed features (covariates) from an officer's history of CCRB complaints and lawsuits from the five years prior to the date of prediction. An officer's history of complaints is represented by the total number of complaints, total number of complaints by type, number of complaints by finding, and complaints by type and finding (eg "number of sustained excessive force complaints"). An officer's history of lawsuits is represented by the total number of lawsuits, the number of lawsuits by type,[52] and the total monetary payouts in their lawsuits.[53] Each set of features were measured over the prior year, the prior two years, and prior five years in order to allow the models to weight factors differently based on how recently they occurred. In total, there are 117 features per observation.

---

[50] All of this data is made publicly available by New York City's Citizen Complaint Review Board (CCRB) and NYC's law department. The data and code used to generate this analysis can be found with our replication materials https://github.com/uchicago-urbanlabs-crimelab/predicting-police-misconduct.

[51] Lawsuit data is not available prior to 2013. We selected 2015 as the beginning year to ensure we had at least two years of features with both lawsuit and complaint data. We selected 2019 as the last prediction year due to right truncation bias; many complaints and lawsuits that were filed in 2021 and later are still pending.

[52] The NYC law department classifies lawsuits by whether they involve an allegation of force, assault/battery, malicious prosecution, or false arrest/imprisonment.

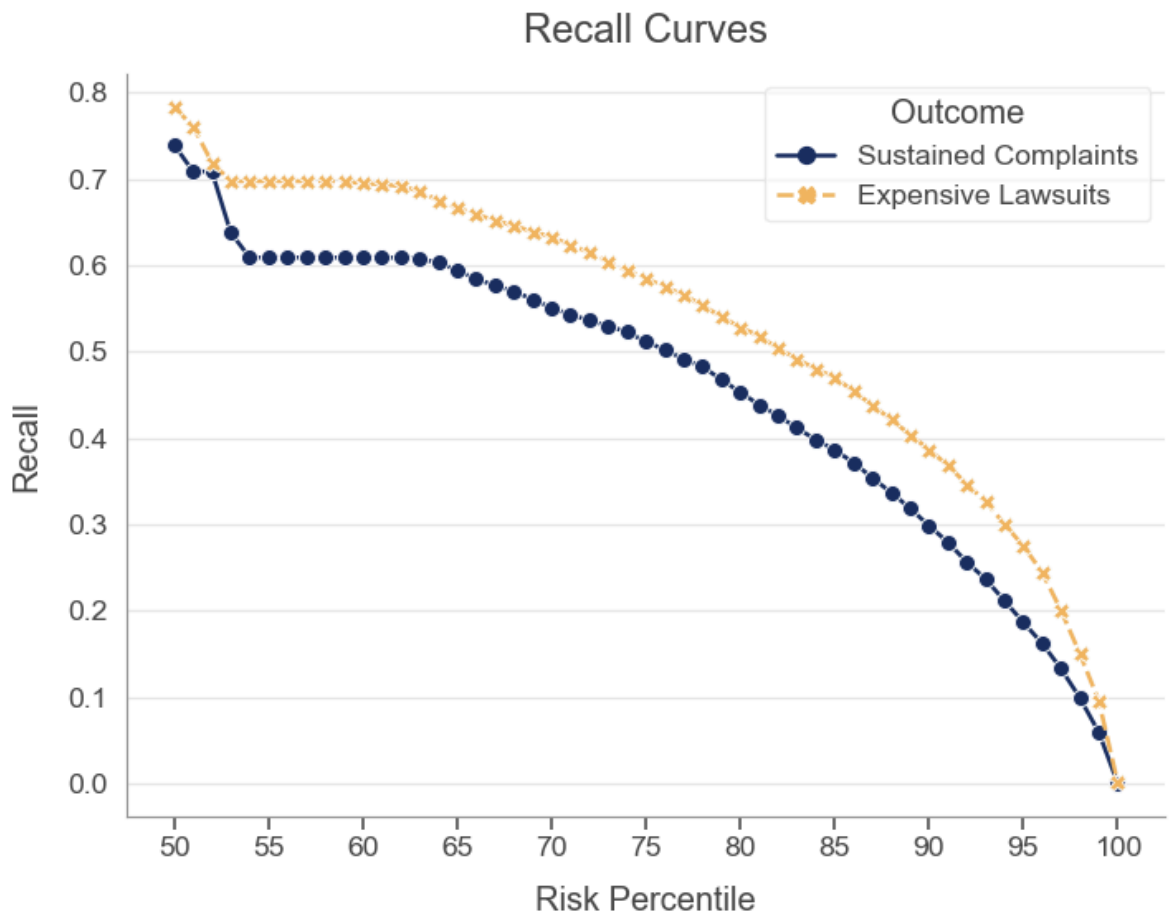
[53] In the case where multiple officers are named on a single lawsuit, we attribute the cost evenly to all officers on the suit. This is not a perfect scheme as the claims against some officers on the suit may have been dropped prior to disposition but the lawsuit data does not contain enough detail for more nuanced measurements.

We defined two outcomes in the NYPD data. The first outcome is whether an officer has a sustained CCRB complaint during the outcome period, which is analogous to the CPD on-duty misconduct outcome since the CCRB only investigates complaints related to use of force, abuse of authority, discourteous policing, and offensive language. This is a rare outcome - only 3% to 4.5% of officers have a sustained complaint over these two-year outcome periods. The second outcome is whether an officer was named in a lawsuit whose payout is $50,000 or greater[54] during the outcome period. For simplicity, we refer to these as *expensive lawsuits*. This outcome is also rare - only 1-2% of officers are named in an expensive lawsuit over these two-year outcome periods.

C2: Performance of machine learning models

We begin by documenting how well the machine learning models built on NYPD data predict future sustained complaints and future expensive lawsuits. The plots below show the recall curves of the machine learning models for each outcome. The sustained complaints model shows a similar level of predictability to the CPD models, with the top 5% of officers by predicted risk accounting for 18.6% of all officers who have a sustained complaint in the follow-up period. The ‘expensive lawsuit’ model shows a higher degree of predictability, with the top 5% of officers by predicted risk accounting for nearly 30% of officers named in an expensive lawsuit during the outcome period.

[54] Roughly 20% of non-pending lawsuits have a payout of $50,000 or more.



C3: Comparison of machine learning and simple models

We next compared simple policies like ranking officers either by the prior number of complaints or the prior number of lawsuits to machine learning models. The first column in the table below shows that rank-by-complaints (RBC) captures nearly as many future sustained complaints as the machine learning model that was trained to predict sustained complaints. The gap in performance between ML and RBC is smaller in NYC than Chicago, potentially due to the fact that there are fewer types of officer behavior data in the public NYPD data. The rank-by-lawsuits (RBL) policy, on the other hand, is significantly worse than both RBC and ML.

The second column shows the machine learning model trained to predict future expensive lawsuits is more accurate than either the RBC or RBL policies, but both policies successfully identify a high-risk group of officers. The top of 5% of officers by predicted lawsuit risk account for 27.5% of all officers named in a future lawsuit (a rate that's nearly 6x times higher than the

average officer) while the top 5% of officers identified by RBL or RBC account for 21 or 22% of officers named in an expensive lawsuit (which is roughly 4x higher than the average officer).

Comparison of risk models and rank-by-complaints		
	Recall and Annual True Positives @5%	
	Future Sustained Complaint	Future Expensive Lawsuit
ML model	18.6% Annual true positives = 204	27.5% Annual true positives = 187
Rank by complaints in past two years	17.6% Annual true positives = 193	21.8% Annual true positives = 150
Rank by lawsuits in past two years	11.2% Annual true positives = 120	20.3% Annual true positives = 138

C4: Predictive value of non-sustained complaints

We finally repeated the experiment of testing whether records of non-sustained complaints have predictive value. Specifically, we constructed two statistical models for each outcome. The “all complaints” model used information derived from all complaints filed against an officer in the five years prior to the date of prediction, while the “only sustained complaints” model only had access to information from complaints that were sustained.

This experiment shows that non-sustained complaints carry predictive signals about the risk, which echoes our findings on the CPD data. When flagging the top 5% of officers by predicted risk of a future sustained complaint, the ‘All complaints’ model correctly flags 18.6% of officers with a sustained complaint in the outcome period while the ‘Only sustained complaints’ model flags 11.4% of officers with a sustained complaint during the outcome period. This is a relative drop in accuracy of 38%, and translates to 80 fewer correctly flagged officers per year. Similarly, the machine learning model trained to predict future expensive lawsuits suffers a relative drop in accuracy of 38% when limiting prior features to only sustained complaints, which translates to 56 fewer correctly flagged officers per year.

	Recall@5%	
Risk Model	Future sustained complaint	Future expensive lawsuit
All complaints	18.6% Annual true positives: 203	24.2% Annual true positives: 168
Only sustained complaints	11.4% Annual true positives: 123	15.0% Annual true positives: 102

D: Marginal Value of Public Funds calculation

D1: Savings to the government

In this section, we estimate the potential savings to the government from using a risk algorithm to target an intervention. We assume that the department has an intervention that reduces negative events - lawsuits, complaints, etc - by 20% in the year following the intervention, and that, as a baseline policy, the department administers it to a random X% of officers per year. The question is how much additional misconduct could be avoided by targeting based on risk and how the cost of that additionally prevented misconduct compares to the cost of implementing a risk targeting algorithm.

We quantify the government savings from reduced misconduct based on reduced litigation costs (including both payouts to plaintiffs as well as costs of litigation) and reduced costs from fewer misconduct investigations. In Chicago, we estimated the cost of a single complaint investigation to be around $15,000 dollars based on the ratio of the budget (including the cost of fringe benefits) of the agency that investigates the most serious complaints (the Civilian Office of Police Accountability, or COPA) to the number of complaints it receives each year.[55] We can then compute the cost of complaints for each officer by multiplying the number of complaints an officer received by $15,000. However this calculation would 'double count' costs in instances where multiple officers are named on a complaint, so we instead calculate an officer's sum of 'weighted complaints' where each complaint is inversely weighted by the number of officers on the complaint (e.g. a complaint with two officers receives a weight of ½, a complaint with three officers receives a weight of ⅓, etc)

The table below shows the average number of weighted-complaints (where each complaint is divided by the number of officers named on a complaint to avoid double counting) received by flagged officers in the year following being flagged for different policies, as well as the average number of future complaints for all officers. Those policies are:

- **ML flagging**: Flag the top X% of officers with the highest predicted risk by the ML models
- **Rank-by-complaints** (RBC): Flag the top X% of officers with the most complaints over the prior two years (ties are broken randomly).
- **Rank-by-serious-event** (RBSE): Flag the top X% of officers with the most sustained complaints over the prior five years (ties are broken randomly).

[55] Specifically, we computed the cost-per-complaint for every year between 2018 and 2022 and took the average cost-per-complaint over those years. All source figures for these estimates can be found in Section D2. This is within the range of costs found by other studies; Ariel, Farrar, and Sutherland estimate a cost of $20,000 per complaint while Braga et al (2017) estimate a cost of $6,776 per complaint (prior to the implementation of body cameras).

For example, the table shows that the top 1% of officers flagged by the ML models have a future weighted-complaint rate of .419 whereas the top 1% of officers by prior number of complaints have a future weighted-complaint rate of .382.

	Average number of weighted future complaints (inversely weighted by # of officers named in a complaint)			
Flagging Rate	All officers	ML-flagged officers (On-duty model)	RBC	RBSE
1% 127 flags per year	.055	.419	.382	.149
2% 253 flags per year	.055	.349	.319	.155
5% 632 flags per year	.055	.273	.226	.114

Finally, we can compute the savings from implementing a smarter targeting policy (either ML or RBC) relative to a baseline (either flag at random or flag based on prior serious events). The annual savings is equal to the difference in the future weighted-complaint rate of the new and the baseline policy multiplied by the number of flags multiplied by the cost-per-complaint ($10,500)[56] multiplied by the intervention efficacy (.2). In terms of an equation that looks like:

$$Complaint\ savings\ =\ (\bar{Y}(k)_{new\ policy}\ -\ \bar{Y}_{baseline})\ *\ K\ *\ \$15,000\ *\ .2$$

The first two columns in the table below show savings from either using ML or RBC instead of flagging at random, while the last two columns show the gains relative to flagging officers with the most prior "serious events" (sustained complaints).

[56] The assumption that each complaint incurs the average cost likely understates the gain from targeting because complaints received by the riskiest officers have longer-than-average investigations, and hence incur more time from city employees.

	Annual complaint savings from targeting versus baseline			
	Baseline = Random		Baseline = Serious event	
Flagging Rate	ML	RBC	ML	RBC
1% 127 flags per year	$138,746	$124,888	$102,949	$89,092
2% 253 flags per year	$223,542	$201,082	$138,905	$116,445
5% 632 flags per year	$413,529	$324,411	$301,766	$212,648

We next compute the potential savings from lawsuit payouts based on extrapolating from NYPD data because we lack consistent lawsuit data from Chicago. The NYPD data links lawsuits and their payouts directly to the involved officers,[57] allowing us to compute the average lawsuit costs of flagged officers. The table below similarly compares the average future payouts of the same three policies we tested on the Chicago data: ML flagging (see Section C in the appendix for how we constructed the ML models on the NYPD data), ranking by complaints, and ranking by prior serious events.

	Average future litigation payouts			
Flagging Rate	All officers	ML-flagged officers (On-duty model)	RBC	RBSE
1% 127 flags per year	$1028	$8772	$6985	$4187
2% 253 flags per year	$1028	$6881	$5977	$3585
5% 632 flags per year	$1028	$4545	$4184	$2797

[57] Technically, the NYPD lawsuit data lists each officer that was named in the beginning of the case and the final payouts (if any). We evenly attribute the lawsuit payout to each officer named on the case (e.g. if there was a $10,000 settlement and 2 officers named on the case, we attribute a cost of $5,000 to each of them). We note that this even attribution is a simplification because the actions of some officers may have been deemed more serious than others but the dataset lacks the necessary detail to disaggregate further.

We use these NYPD estimates for our Chicago estimates with one adjustment. We first note that the lawsuit costs per officer are roughly the same in Chicago and NYC. CPD pays out roughly 1/3rd of the annual lawsuit payouts that NYPD does, but also has 1/3rd of the officers (36,000 vs 12,000). However we scale up the per-officer lawsuit costs from NYPD to account for the cost of litigation, i.e. the fees paid to outside counsel to represent Chicago in these cases. On average, Chicago's outside counsel fees are about 40% of the annual litigation payouts (for example, Chicago paid out 86 million in 2022 and incurred 25 million dollars in outside counsel fees), so we scale up costs by 40%. Then the savings from implementing some new targeting policy instead of the baseline policy can be calculated, and summarized, with the equation and table below.

$$Lawsuit\ savings\ =\ (C(\bar{K})_{policy}\ -\ \bar{C}_{baseline})\ *\ K\ *\ 1.4\ *\ .2$$

	Annual litigation savings from targeting versus baseline			
	Baseline = Random		Baseline = Serious event	
Flagging Rate	ML	RBC	ML	RBC
1% 127 flags per year	$275,355	$211,799	$163,054	$99,498
2% 253 flags per year	$414,562	$350,559	$233,486	$169,482
5% 632 flags per year	$622,324	$558,362	$309,292	$245,329

Finally, we combine the cost savings from complaint investigations and lawsuits together to yield the following savings estimates:

	Annual savings (complaint + litigation) from targeting versus baseline			
	Baseline = Random		Baseline = Serious event	
Flagging Rate	ML	RBC	ML	RBC
1% 127 flags per year	$414,101	$336,688	$266,004	$188,590
2% 253 flags per year	$638,105	$551,641	$372,391	$285,927
5% 632 flags per year	$1,035,853	$882,773	$611,058	$457,978

D2: Data for cost estimates

Table D2.1:

COPA budget per complaint			
Year	**COPA budget** (including fringe costs)	**# of complaints retained by COPA**	**Cost per complaint**
2019	$18,415,879	2089	$8,815
2020	$18,390,073	1740	$10,569
2021	$20,589,908	1021	$20,166
2022	$23,464,609	1106	$21,215
Average	$20,215,117	1432	$15,191

Notes: COPA receives roughly 5 times the number of complaints as listed above but only a subset of those complaints (roughly 20%) are retained to be investigated by COPA based on the nature of the allegations. Complaints not retained by COPA are sent to CPD for internal investigation. Source: Budget and number of complaints are taken from COPA's annual reports, with the exception of the 2019 budget estimate which comes from City of Chicago annual budget statements.

Table D2.2:

Comparison of CPD and NYPD lawsuit payouts			
Year	**Chicago litigation payouts**	**NYPD litigation payouts**	**Ratio of Chicago to NYPD payouts**
2019	$46,000,000	$237,000,000	0.19
2020	$40,000,000	$225,000,000	0.17
2021	$123,000,000	$206,000,000	0.59
2022	$86,000,000	$237,000,000	0.36
Average	$73,750,000	$226,250,000	0.33

Source: City of Chicago Annual Reports on CPD litigation, New York City Comptroller Annual Claims Reports

Table D2.3:

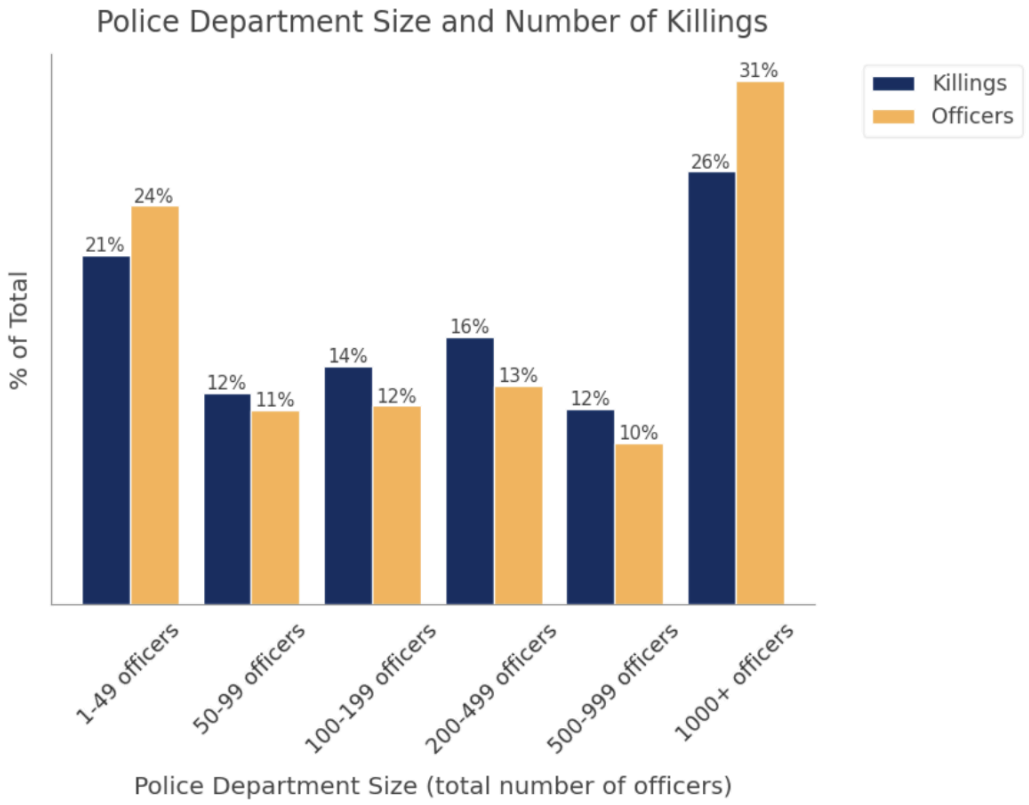
Costs of litigating complaints against CPD			
Year	**Chicago litigation payouts**	**Chicago outside counsel fees**	**Ratio of counsel fees to litigation payouts**
2019	$46,000,000	$25,000,000	0.54
2020	$40,000,000	$24,000,000	0.6
2021	$123,000,000	$25,000,000	0.20
2022	$86,000,000	$25,000,000	0.29
Average	$73,750,000	$24,750,000	.41

Source: City of Chicago Annual Reports on CPD litigation

E: Distribution of police killings by agency size

We merged data from the Mapping Police Violence project (https://mappingpoliceviolence.org/) with the 2016 Law Enforcement Management and Administrative Statistics[58] (LEMAS) survey to estimate the share of police killings that are committed by small versus mid-size versus large police departments. MPV constructs this dataset by monitoring a stream of news articles produced by Google News and then hand-verifying the details of a possible police killing. We merge this data to the LEMAS survey by agency name. In the cases where the MPV data lists multiple agencies as responsible, we attribute that event to the first listed agency (those events are rare and have no qualitative impact on these findings). We limit our study to MPV-collected killings that occurred between January 2013 and April 2023.

Departments with fewer than 500 officers account for 63% of killings in the MPV database, and departments with fewer than 200 officers account for 47% of killings. We note this is roughly proportional to the collective number of officers that work in those departments - suggesting a roughly equal rate of killings per officer across agency size.



[58] United States Department of Justice. Office of Justice Programs. Bureau of Justice Statistics. Law Enforcement Agency Roster (LEAR), 2016. Inter-university Consortium for Political and Social Research [distributor], 2017-04-05. https://doi.org/10.3886/ICPSR36697.v1