How Many Complaints Against Police Officers Can Be Abated by Incapacitating A Few "Bad Apples?"*

Aaron Chalfin[1] and Jacob Kaplan[1]

[1]Department of Criminology, University of Pennsylvania

August 14, 2020

Abstract

Research Summary: The notion that the unjustified use of force by police officers is concentrated amongst a few "bad apples" is a popular descriptor which has gained traction in scholarly research and achieved considerable influence among policymakers. But is removing the bad apples likely to have an appreciable effect on police misconduct? Leveraging a simple policy simulation and data from the Chicago Police Department, we estimate that removing the top 10 percent of officers identified based on *ex ante* risk and replacing them with officers drawn from the middle of the risk distribution would have led to only a 6 percent reduction in use of force incidents in Chicago over a ten-year period.

Policy Implications: Our analysis suggests that surgically removing predictably problematic police officers is unlikely to have a large impact on citizen complaints. By assembling some of the first empirical evidence on the likely magnitude of incapacitation effects, we provide critical support for the idea that early warning systems must be designed, above all, to deter problematic behavior and promote accountability.

Keywords: police use of force, early warning systems

*For helpful comments, we thank Richard Berk, Maria Cuellar, Zubin Jelveh, Greg Ridgeway and Sarah Tahamont. Correspondence: Aaron Chalfin, Department of Criminology, 558 McNeil Building, 3718 Locust Walk, University of Pennsylvania, Philadelphia, PA 19104. E-Mail: achalfin@sas.upenn.edu.

1 Introduction

The idea that a small number of "bad apples" are responsible for an outsize share of complaints against police officers has gained considerable traction over the course of the last four decades both in the scholarly literature (Berkow, 1996; Alpert and MacDonald, 2001; Walker et al., 2001; Rozema and Schanzenbach, 2019; Goncalves and Mello, 2020) and in government reports (Christopher, 1991; Mollen, 1994) and popular media accounts (Arthur, 2018; Invisible Institute, 2018; Wu, 2019; Ba and Rivera, 2020; Kelly and Nichols, 2020; MacDonald and Klick, 2020). Such a claim is inspired by numerous anecdotal descriptions of "rogue cops" (Greek, 2007; Sherman, 2020) as well as by the Pareto principle (also sometimes called the "80/20 rule"), the empirical regularity that, in many areas of human inquiry, approximately 80 percent of the effects accrue from 20 percent of the causes.[1] With respect to policing, appeals to the Pareto principle were instrumental in spurring the creation of the first "early warning systems" to identify problematic police officers in the 1970s (Walker et al., 2000) and has informed numerous police reform initiatives in the intervening years (Alpert and Walker, 2000; Walker et al., 2001; Hughes and Andre, 2007). Given the increasing availability of "big data" to inform police practice (Ridgeway, 2018), many observers have expressed optimism that large administrative datasets can potentially be put to use to identify problematic police officers and incapacitate them before they have the opportunity to do serious harm to members of the communities they serve (Sherman, 2018).

Empirically the claim that a small number of police officers account for an outsize share

[1]The Pareto Principle derives from the observation of 19th century economist, Vilfredo Pareto, that, in many societies, 80 percent of the wealth is concentrated in the hands of 20 percent of the population (Pareto et al., 1971). In mathematical statistics, the Pareto principle has stimulated the study of "power law distributions" which characterize an astonishing array of natural phenomena in physics, biology, earth and planetary sciences as well as in the social and computational sciences (Newman, 2005).

of serious misconduct rests on analyses of individually identified microdata on complaints against police officers. By collapsing the data at the officer level and sorting the officers in descending order with respect to the number of complaints they have generated, researchers can compute the share of complaints over a given time period that are accounted for by the top k percent of officers. Prior analyses from police departments across the United States suggest that a small share of officers indeed account for a large share of complaints against police. Indeed a common estimate is that the top 2 percent of officers account for approximately 50 percent of known misconduct by police officers (Walker et al., 2001). As the other 98 percent of officers are responsible for the remaining 50 percent of misconduct, the implication is that the top 2 percent of officers are, incredibly, 49 times more likely to commit misconduct than other officers.[2] With respect to public policy, such an analysis suggests that if only the small number of "bad apples" can be identified and successfully intervened upon, law enforcement agencies could make substantial progress in reducing police misconduct without making any other institutional changes to policy or practice. Indeed, a relative risk ratio of 49 naively suggests that nearly 50 percent of use of force complaints could potentially be abated by replacing the top 2 percent of officers with officers drawn from the remainder of the distribution.[3]

Unfortunately, such an analysis suffers from three problems. First, the above computation assumes that we can predict bad acts among police officers with perfect foresight. Second, the analysis makes no provision for the replacement of "bad apples" with other

[2]To see this, consider a municipal law enforcement agencies which employs 1,000 officers and experiences 100 misconduct complaints. In this agency, the top 20 officers account for 50 complaints and the remaining 980 officers account for 50 complaints. The relative risk ratio is given by: $\frac{50/20}{50/980} = 49$.

[3]Suppose a department has $N = 1{,}000$ officers and $m = 100$ use of force complaints over some time period. The top 20 officers are known to account for 50 percent of the 100 complaints and the remaining 980 officers account for the other 50 percent of complaints. Replacing the top 20 officers with 20 officers whose risk is equal to the remainder of officers would lead to a reduction of 100 - 1000 $\times \frac{50}{980} = 51$ complaints.

police officers, who while less likely to commit misconduct, will nevertheless continue to generate complaints. Finally, the computation suffers from a simple but, to date, seldom identified problem which we refer to as "data density bias." Put simply, when the number of complaints is relatively small compared to the number of police officers, it will be true, by definition, that a small share of the officers will account for a large share of the known incidents. To see this, consider the simplified but nevertheless instructive case in which there are only 5 serious complaints filed in a city which employs 1,000 police officers. It is easy to see that even if all 5 complaints implicate a different officer, 100 percent of the complaints would be accounted for by just $\frac{5}{1{,}000} = 0.5$ percent of officers. While the headline — 0.5 percent of officers account for all of the serious complaints — sounds impressive, it is merely a statistical artifact that is intrinsic to the analysis of sparse data.[4] As we show in Section 2.1, sparse data are common in studies which use complaint data. Accordingly, calculations such as the one above have the potential to distort the policy conversation to a considerable degree.

How can we correct for data density bias? The solution lies in identifying the correct benchmark against which to compare a conventional assessment of the degree to which complaints are concentrated among police officers. In particular, we need to know what share of a law enforcement agency's use of force complaints would be accounted for by the top k percent of officers in a world in which the use of force by police officers were completed unconcentrated. Happily, such a counterfactual is easy to both identify and to compute. By randomizing complaints with replacement to police officers, we can generate the share of complaints that would be accounted for by the top k percent of officers in the

[4] A similar argument has been made in the literature on crime concentration by Hipp and Kim (2017), Levin et al. (2017) and Chalfin et al. (2020) among others.

complete absence of concentration. Referring to the example above, were we to randomize 5 complaints among 1,000 officers a large number of times, in nearly all iterations, the 5 complaints would be randomly assigned to different officers. Since 0.5 percent of officers will have accounted for 100 percent of the complaints in both the real data and the simulated data, we would conclude that there is, in fact, no concentration in the use of serious force by police officers. Accordingly the Pareto principle would constructively fail to hold even though it would be supported by a naive analysis of the data. Put differently, some of the "bad apples" may have simply been "unlucky apples."[5]

To the extent that use of force is relatively unconcentrated, this narrows the scope for incapacitating problematic police officers to have a large effect on use of force complaints. However, a more salient policy question is how much force can be abated by incapacitating predictably problematic police officers, a question which hinges on the ability of analysts to make an *ex ante* prediction about which officers will receive future complaints. As we demonstrate in Section 2.3, even though there is reasonably strong persistence in use of force complaints throughout an officer's career, incapacitating the small number of officers who generate the greatest number of complaints early in their career is likely to lead to only a modest reduction in future use of force complaints. Drawing on a simple but realistic policy simulation, we estimate that replacing the 10 percent of officers who generated the largest number of use of force complaints early in their career with officers drawn from the middle of the distribution would have led to only a 6 percent reduction in use of force complaints against the Chicago Police Department over a ten-year period. The modesty of

[5]This conceptualization of risk is similar in spirit to an approach that is found Ridgeway and MacDonald (2009) who identify NYC police officers who are the most likely to engage in biased policing. In order to ensure that officers flagged by their statistical algorithm are, in fact, high-risk and not merely "unlucky," they motivate a statistical framework in which the risk threshold is raised until false discovery rates are tolerably low.

this impact is, in part, due to the difficulty of predicting future complaints and, in part, due to the extent to which data density bias has obfuscated the true degree of concentration in the use of force by police officers.

Our conclusion is that while incapacitating predictably problematic officers serves an important instrumental purpose, this practice is, in of itself, unlikely to lead to a large reduction in use of force complaints, absent appreciable deterrence effects or broader cultural change. As such, early warning systems should be designed to promote accountability among a broader set of officers, rather than to serve as a narrowly-tailored tool to surgically remove high-risk personnel. While the importance of accountability has long been a focal point in the scholarly literature on early warning systems (Alpert and Walker, 2000; Walker et al., 2001), references to the concentration of misconduct amongst a small number of "bad apples" are pervasive in popular media accounts and public commentary. By assembling some of the first empirical evidence on the likely magnitude of incapacitation effects, we provide critical support for the idea that early warning systems must be designed, above all, to deter problematic behavior and promote accountability.

2 Empirical Example

2.1 Data and Methods

We explore the extent to which use of force is concentrated amongst police officers using individually identified microdata on use of force complaints made public by the Chicago Police Department. These data come from the Citizens Police Data Project which is a collection of nearly 250,000 complaints against Chicago Police Department officers filed since 1988 (Ba and Rivera, 2020). The data were collected and released publicly by the

Invisible Institute and have been used in recent research on police use of force including that of Rozema and Schanzenbach (2019) and Ba and Rivera (2019) among others.[6] This dataset is ideal for our purposes as it includes information on all complaints as well as a full roster of Chicago police officers including, critically, those who have never been named in a complaint.

In this research, our focus will be on citizen complaints which implicate one or more Chicago police officers. Naturally, not every complaint will be justified and, in practice, many complaints will fail to be sustained upon detailed review. Likewise, some errant behavior, in particular criminal acts, may be committed by officers while they are off-duty (Fyfe, 1980; Kane and White, 2009). While we acknowledge that complaints are an imperfect proxy for official misconduct, we believe this exercise is both appropriate and useful for three reasons. First, research by Rozema and Schanzenbach (2019) has shown that complaints are a surprisingly good predictor of high-impact events such as lawsuit payouts by municipal officials. While legal settlements are rare and therefore extraordinarily difficult to predict, complaints are more common and, as such, have greater predictive signal. Second, by focusing on all complaints instead of sustained complaints, we use data that has not been filtered through the lens of what a law enforcement agency deems problematic and which therefore may better reflect community norms. Finally, by focusing on all complaints instead of sustained complaints, we generate a lower bound on the extent to which data density bias distorts the policy conversation. Since sustained complaints are a subset of all complaints, the degree of data density bias will be even greater in such an analysis.

To explore the concentration of complaints, we focus on complaints made against Chicago police officers during the five-year period between September 17, 2012 and Septem-

[6]The data was downloaded from their website https://beta.cpdp.co/

ber 17, 2017.[7]. For each complaint, we have information on the police officers involved in the complaint as well as the nature of the complaint. We focus on the 11,283 police officers who were employed by the Chicago Police Department as of September 17, 2017 — the last date for which we have information on complaints — and use the last five years of available data so that we have up to five years of data for each officer.[8] Among these 11,283 officers, between September 17, 2012 and September 17, 2017, there were 16,023 complaints of which 2,566 involved the use of force. Thus, over a five-year period, there were approximately 1.4 total complaints and 0.2 use of force complaints per police officer. While many use of force incidents are likely to go unreported given the difficulty of filing a complaint in Chicago and many other U.S. cities (Ba and Rivera, 2019), these data capture all complaints known to the authorities and therefore form the basis for any data-driven early warning system that could potentially be developed.

In a second analysis, we focus on the cohort of Chicago police officers hired between 2000 and 2007 and follow them prospectively over the following ten years. We rank officers with respect to the number of complaints they receive early in their careers and use this information in order to predict future complaint risk. This exercise forms the basis for our policy simulation, described in Section 2.3, in which we identify high-risk officers using data generated during their early career probationary period and simulate the replacement of these officers with officers in the middle of the risk distribution. We use this policy simulation to estimate the share of use of force complaints over a ten-year period that could have been abated solely by incapacitating the "bad apples", identified *ex ante*.

[7]September 17, 2017 is the last date where the data indicates that a police officer retired.

[8]We restrict the data to individuals employed at the rank of police officer, police officer-training officer, sergeant or police-officer-detective in order to focus on the subset of officers who routinely encounter citizens while on patrol.

2.2 How Concentrated are Citizen Complaints?

In **Figure 1**, we explore the concentration of overall complaints against Chicago police officers (Panel A) and use of force complaints specifically (Panel B). In each graph, there are three lines. The solid red line plots the cumulative distribution of complaints — that is, the share of complaints that are accounted for by the top k percent of officers. The dashed gray line plots the cumulative distribution function which arises from randomly assigning complaints to police officers, with replacement. The solid black line is a 45 degree line which represents uniformity in concentration — that is, the condition in which top k percent of officers account for k percent of complaints for all values of k. Naturally, uniformity can only hold prior to the saturation point at which 100 percent of the complaints are accounted for. The figure can be used to make a number of useful comparisons which reveal the extent to which complaints are concentrated amongst police officers. Graphically, the degree of data density bias is greatest when the simulated density function under randomization lies closer to the empirical density function than to the 45 degree line. Indeed when the empirical and simulated density functions lie on top of one another, complaints are, in fact, unconcentrated.

Referring to Panel A, we see that the top 20 percent of officers, ranked according to the number of complaints they have generated, account for approximately 65 percent of the complaints and the top half of officers account for nearly all of the complaints. However, the slope of the curve is steepest at the top of the distribution. Here, we observe that the top 2 percent of police officers account for approximately 14.4 percent of total complaints against the department. Put differently, the top 2 percent of officers are $\frac{\frac{14.4}{2}}{\frac{(100-14.4)}{(100-2)}} = 8.2$ times more likely to generate complaints than the remaining 98 percent of officers. This naive

computation suggests that complaints are concentrated to a large degree amongst a very small number of officers. However, this computation does not account for the obfuscating effect of data density bias. To see how important data density bias is empirically, we turn to our simulated data in which we randomly assigned complaints to police officers with replacement. Referring to the dashed gray line, under random assignment, the top 2 percent of officers generate 7.4 percent of the complaints. This suggests that, even under random assignment, the top 2 percent of officers are $\frac{\frac{7.0}{2}}{\frac{(100-7.0)}{(100-2)}} = 3.7$ times more likely to generate complaints than the remaining 98 percent of officers. As such, in the real-world data, the top 2 percent of officers are $\frac{8.2}{3.7} = 2.2$ times more likely to generate complaints than in a condition in which there is no concentration in use of force by construction. While this computation suggests that complaints are, in fact, concentrated, the naive comparison overstates the degree of concentration by a factor of nearly 4.

Next, we turn to use of force complaints which account for 16 percent of all complaints against Chicago Police officers during our five-year study period. Given that these types of complaints are less common, the degree to which data density bias obfuscates comparisons will be greater. In Panel B, we see that the top 10 percent of officers, ranked according to the number of complaints they have generated, account for 70 percent of the complaints and the top 16 percent of officers account for all of the use of force complaints. In other words, 84 percent of officers generated no use of force complaints during the sample period. At the top of the distribution, the top 2 percent of officers account for 26.2 percent of the use of force complaints. In other words, these officers are $\frac{\frac{26.2}{2}}{\frac{(100-26.2)}{(100-2)}} = 17.4$ times more likely to generate complaints than the remaining 98 percent of officers. This comparison suggests an extraordinary degree of concentration and accordingly that the Chicago Police

Department could appreciably reduce use of force complaints by removing a small number of "bad apples." However, the figure also shows a considerable degree of concentration even when complaints are randomized to officers. Indeed, even in the simulated data, the top 2 percent of officers account for 18.3 percent of use of force complaints. As such, even under randomization, these officers are $\frac{\frac{18.4}{2}}{\frac{(100-18.3)}{(100-2)}} = 11$ times as likely to generate force complaints than other officers. Thus, rather than use of force complaints being 19 times more common among the top 2 percent officers, they are, in fact, only $\frac{17.4}{11} = 1.6$ times as likely once data density bias is accounted for.

2.3 Policy Simulation

Although use of force complaints are not nearly as concentrated as they might have seemed at first blush, there still appear to be officers who are at an elevated risk to generate citizen complaints. We next assess the extent to which "bad apples" are *ex ante* predictable. We begin by considering the extent to which there is persistence in the use of force. Specifically, we assess the degree to which the officers who are most likely to generate complaints early in their careers are also the most likely to generate complaints later in their careers. Next, we motivate a simple but realistic policy simulation in which we estimate the share of use of force complaints which could be abated by removing a small number of officers who generate the greatest number of complaints early in their careers and replacing them with officers drawn from the middle (40th-60th percentile) of the use of early career use of force distribution.

In settings in which there are rich cross-sectional data — for example, detailed demographic data or pre-employment information — predictions about police officer risk are typically made using sophisticated machine learning-based algorithms (Ridgeway and Mac-

Donald, 2009; Carton et al., 2016; Chalfin et al., 2016; Helsby et al., 2018) or, at a minimum, logistic regression (Leinfelt, 2005; White, 2008).[9] The advantage of machine learning methods in this context is that the approach allows researchers to automate the detection of signal in the data, a task which is complicated considerably when the number of predictors is large and the relationships between variables are non-linear and conditional (Hastie et al., 2009).[10]

Given the longitudinal nature of complaint data, we focus instead on a simpler but, we argue, especially policy-relevant prediction exercise which captures the extent to which there is persistence in complaints among officers. We focus on the eight cohorts of Chicago police officers hired between 2000 and 2007 and who remain employed by the Chicago police department in 2017.[11] For each officer, we retain 11.5 years of data and divide the 11.5-year sample period into an 18-month pre-period and a ten year post-period. We choose a pre-period of 18 months as this is the standard probationary period for new police officers hired in Chicago — as a robustness check, we repeat this analysis using a five-year probationary period. The purpose of this exercise is to identify the police officers who generate the largest number of complaints during their probationary period and to see how many of them continue to generate an outsize number of complaints throughout the prime of their careers. As this exercise is intended to be illustrative, we do not condition on the officer's precinct and shift; nor are we able to observe the duties to which an officer is assigned.

[9]For example, leveraging a wealth of data on pre-employment characteristics, Chalfin et al. (2016) use stochastic gradient boosting to predict police misconduct among a sample of police officers in Philadelphia. Machine learning-based algorithms are indeed used in a variety of settings in the criminal justice system including to inform decisions about sentencing (Berk and Hyatt, 2015), parole (Berk, 2017) and arraignment (Berk et al., 2016; Kleinberg et al., 2018). For an excellent reference on the development of machine learning algorithms in the U.S. criminal justice system, we refer readers to Berk (2019).

[10]See Beutler et al. (1985), Leinfelt (2005) and Ridgeway (2020) for assessments of the predictors of police use of force and Lum (2016) and Ridgeway (2016) for assessments with respect to race.

[11]Naturally over a ten-year period, some officers will be terminated as a result of a use of force incident. However, termination is exceedingly rare — less than 0.2 percent of officers are terminated annually (Ba and Rivera, 2020).

However, these elements could be straightforwardly incorporated by a law enforcement agency which intends to use a prediction exercise to make personnel decisions. We likewise note that, to the extent that consistent officer assignments serve to generate persistence in complaints, failing to account for these will bias the quality of our predictions *upwards*. That is, our results are, if anything, optimistic with respect to the potential for incapacitating problematic officers to lead to meaningful reductions in the use of force.

Among the officers who are in the top k percent, ranked according to the number of complaints accrued during their probationary period, we identify the share who are in the top j percent of officers, ranked according to use of force complaints in the ten-year post-period. This estimand corresponds with the "positive predictive value" in the prediction literature.[12] The results of this exercise are presented in **Table 1** which reports the positive predictive value for $k, j = 2$, 5, 10 and 20 percent. Among the officers in the top 2 percent of complaints in the probationary period, 7.4 percent are also in the top 2 percent in the post-period and nearly one third are in the top 10 percent in the post-period. Among the officers in the top 5 percent of complaints in the probationary period, 25 percent are in the top 10 percent in the post-period. With respect to use of force complaints, 17.7 percent of officers who are in the top 2 percent of the pre-period distribution are also in the top 2 percent in the post-period distribution, 26.5 percent are in the top 10 percent in the post-period distribution and nearly half are in the top 20 percent in the post-period distribution.

Overall, these results suggest that there is considerable persistence in use of force over an officer's career generating optimism that use of force a predictable phenomenon. However, the critical policy question is how many use of force complaints could be abated by

[12]Formally, the positive predictive value is computed as the number of true positives (here, those who are in the top j percent of the use of force distribution in the post-probationary period) divided by the sum of the number of true positives and false positives.

terminating high-risk officers identified on the basis of their early career complaint activity and replacing them with less risky officers, who we proxy for using officers drawn from the middle of the use of force distribution. In **Table 2** we report the share of post-period complaints that would have been abated if the top k percent of officers identified at the end of the probationary period were replaced with an equivalent number of officers drawn at random from the middle 20 percent of the distribution. For instance, for $k = 2$ percent, we remove the 68 police officers who generated the greatest number of complaints during the 18-month probationary period and replace those officers with 68 officers drawn at random from the middle 20 percent of the probationary period distribution. Referring to the table, we estimate that removing the top 2 percent of officers — identified ex ante — from circulation would abate just 1.5 percent of total complaints and just 2 percent of use of force complaints. Even the replacement of the top 10 percent of the workforce — an enormously difficult task given the current rate of approximately 0.2 percent per year — with officers drawn from the middle of the distribution is estimated to reduce use of force complaints by just 6 percent.

Perhaps an 18-month probationary period is insufficient to be able to identify high-risk officers. In order to assess the sensitivity of our policy simulation to this parameter of the analysis, we repeat this exercise focusing on a longer probationary period. Estimates using a five-year probationary are presented in **Table 3**. While using a longer probationary period improves the quality of our predictions, there are two key drawbacks of such an approach. First, it is preferable to identify high-risk officers early in their careers before they have the opportunity to generate complaints. Second, it is generally more difficult to terminate or reassign officers after their official probationary period has ended. Referring

to Table 3, even when we use a five-year probationary period, the results do not suggest that terminating a small number of "bad apples" is likely to have a substantial impact on use of force complaints. We estimate that terminating the top 5 percent of officers, ranked according to use of force complaints in the five-year pre-period, would have abated slightly more than 1 in 10 use of force incidents during the remaining six and a half years.[13]

Given that an accounting of the raw data suggest that the top 2 percent of officers are 17 times more likely to generate use of force complaints than other officers, these estimates — which are fairly modest — may appear surprising. However, the estimates are, in fact, sensible given that future complaints cannot be predicted with perfect foresight and that complaints are not especially concentrated among a small number of officers. For example, officers in the 90th percentile of the probationary period distribution of use of force are not very different in the number of complaints generated (2.76 per officer during the ten-year post-period) than the median officer (1.67 per officer during the ten-year post-period). In the next section, we discuss the implication of these findings for public policy.

3 Policy Implications

In thinking about complaints, there are four possible ways of describing the ease with which problematic police officers can be identified and incapacitated. The most helpful scenario is when complaints are both highly concentrated and highly predictable. When this is the case, policymakers will find it easy to identify the "bad apples" and likewise achieve large reductions in complaints by incapacitating those officers. A second possibility is that complaints are predictable but not particularly concentrated. In this scenario, it is

[13] We can also focus on the most serious use of force complaints — those that involve the discharge of a firearm. When we do so, the estimates are extraordinarily similar to those reported in Tables 2 and 3.

possible to identify which officers will commit bad acts but, given that there are likely to be feasibility constraints with respect to the number of officers who can be incapacitated, the number of complaints that can be abated will be tempered by the lack of appreciable concentration in the data. A third possibility is that complaints are concentrated but not very predictable. Such a scenario might come to pass if, in a given year, a large share of the complaints accrue to a small share of officers but, in each year, the problematic officers are different. Such a scenario is unwelcome in the sense that a shifting policy environment makes it difficult to successfully intervene prior to the accrual of harms. A fourth possibility is that complaints are neither particularly predictable nor particularly concentrated. This possibility leaves little room for optimism that complaints can be meaningfully reduced solely through prediction and incapacitation.

The data suggest that the use of force by police officers is concentrated amongst a small number of problematic officers to a degree, albeit far less concentrated than naive calculations would suggest. This finding, in turn, suggests that the scope for incapacitating problematic police officers to have an appreciable effect on misconduct is narrower than the standard calculations imply. Such a claim is further underscored by the difficulty of predicting who the most problematic police officers will be at the time they are hired or early in an officer's career (Cuttler and Muchinsky, 2006; Fyfe and Kane, 2006; White, 2008; Chalfin et al., 2016). Consistent with the prior literature, when we use an officer's early career accumulation of complaints to predict an his or her subsequent career performance, the accumulation of complaints is somewhat predictable. However, the positive predictive value is modest — between 7 and 40 percent depending on the threshold used. Accordingly the number of false positives remains high, complicating the extent to which such a process

could be used to make personnel decisions such as terminating or even simply reassigning police officers (Goldman and Puro, 2001; White, 2008; Dharmapala et al., 2019). While the number of false positives can be reduced by raising the risk threshold used to flag problematic officers (Ridgeway and MacDonald, 2009), the cost of doing so is inevitably fewer abated complaints — a tradeoff that is informed by the impossibility of simultaneously minimizing both Type I and Type II errors.

With respect to public policy, the most direct implication of this analysis is that, absent appreciable deterrence effects or broader cultural change, early warning systems that are designed to identify problematic police officers and incapacitate them — either through termination or re-assignment — are unlikely to lead to large reductions in the use of force. Likewise, a surgical focus on "bad apples" may be less effective than broad-based measures to improve managerial practices and increase accountability (Sherman, 1978; Skolnick and Fyfe, 1993; Ivkovic, 2009; Mummolo, 2018). While pitting these two policy solutions against each other, in principle, presents a false choice, in practice, constraints on political capital may require policymakers to invest in a limited set of actions. With respect to the efficacy of broad-based police reform efforts, while there continues to be a dearth of high-quality evidence in this domain (Sherman, 2018; Engel et al., 2020), there is, at least, some evidence to support the efficacy of de-escalation training (Engel et al., 2020) and procedural justice training (Owens et al., 2018; Nagin and Telep, 2020; Wood et al., 2020), federal oversight of police agencies (Powell et al., 2017; Goh, 2020) as well as the use of and training in non-lethal weapons (MacDonald et al., 2009; Sousa et al., 2010). There is likewise support for the idea that reforms involving police unions may be effective (Dharmapala et al., 2019) especially if unions can be incentivized to "self-regulate" which might potentially be encouraged by

transferring the burden of liability insurance from municipalities to unions (Ramirez et al., 2018). Finally, as noted by Mummolo (2018), police officers tend to be highly responsive to managerial directives, leaving room for optimism that procedural reforms can dramatically alter officer behavior.

A second implication of this analysis is that it is critical for policymakers to incentivize better reporting and discovery of police misconduct (Long et al., 2013; Knox et al., 2019). Incomplete reporting of misconduct by citizens inevitably leads to noisy data which, in turn, leads to poor predictability (Ivkovic, 2009) and a diminishing of the ability of data-driven early warning systems to have maximum impact. Happily, there is evidence that more complete reporting can be achieved through actions that are available to many municipal policymakers. For example, Ba (2018) finds that when police departments make it easier for citizens to report complaints, the number of complaints increases. Likewise, as suggested by Rozema and Schanzenbach (2019), requiring civilians making allegations to swear out an affidavit before an investigation may proceed may have a chilling effect on reporting misconduct. Similarly, to the extent that there are other markers of complaints such as internal investigations, ad hoc performance assessments or pre-employment information, this information will be critical to deploy in order to further enhance the predictability of bad acts. One especially promising idea is the collection of customer service data arising from police-citizen encounters (Burn, 2010). While there are challenges to collecting data from individuals who are the recipients of police service, the richness of such data might well be a goldmine for prediction, especially over a short time window.

Finally, while our policy simulation suggests that identifying and surgically incapacitating the "bad apples" is unlikely to have a large and direct impact on use of force, early

warning systems, coupled with rigorous oversight and genuine accountability have the potential to have a far larger effect by generating deterrence effects or by holistically changing departmental culture. To the extent that identifying the officers who are genuinely problematic has appreciable spillover effects or serves as a deterrent to officers on the margin, these efforts may well be capable of producing marked changes in use of force well in excess of the estimates we report in this research. We therefore emphasize that our policy simulation, by design, does not identify the promise of early warning systems more generally. Indeed the net impact of data-driven efforts to identify "bad apples" will depend critically on the extent to which these efforts are coupled with initiatives that change behavior among police officers who are unlikely to be flagged as being high-risk.

References

Alpert, G. P. and J. M. MacDonald (2001). Police use of force: An analysis of organizational characteristics. *Justice Quarterly 18*(2), 393–409.

Alpert, G. P. and S. Walker (2000). Police accountability and early warning systems: Developing policies and programs. *Justice Research and Policy 2*(2), 59–72.

Arthur, R. (2018). 130 chicago officers account for 29 percent of police shootings. *The Intercept*.

Ba, B. (2018). Going the extra mile: The cost of complaint filing, accountability, and law enforcement outcomes in chicago. Technical report, Working paper.

Ba, B. and R. Rivera (2020). Police think they can get away with anything. that's because they usually do.

Ba, B. A. and R. Rivera (2019). The effect of police oversight on crime and allegations of misconduct: Evidence from chicago. *Faculty Scholarship at Penn Law*. (19-42).

Berk, R. (2017). An impact assessment of machine learning risk forecasts on parole board decisions and recidivism. *Journal of Experimental Criminology 13*(2), 193–216.

Berk, R. (2019). *Machine learning risk assessments in criminal justice settings*. Springer.

Berk, R. and J. Hyatt (2015). Machine learning forecasts of risk to inform sentencing decisions. *Federal Sentencing Reporter 27*(4), 222–228.

Berk, R. A., S. B. Sorenson, and G. Barnes (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies 13*(1), 94–115.

Berkow, M. (1996). Weeding out problem officers. *Police Chief 63*, 21–29.

Beutler, L. E., A. Storm, P. Kirkish, F. Scogin, and J. A. Gaines (1985). Parameters in the prediction of police officer performance. *Professional Psychology: Research and Practice 16*(2), 324.

Burn, C. (2010). The new south wales police force customer service programme. *Policing: A Journal of Policy and Practice 4*(3), 249–257.

Carton, S., J. Helsby, K. Joseph, A. Mahmud, Y. Park, J. Walsh, C. Cody, C. E. Patterson, L. Haynes, and R. Ghani (2016). Identifying police officers at risk of adverse events. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 67–76.

Chalfin, A., O. Danieli, A. Hillis, Z. Jelveh, M. Luca, J. Ludwig, and S. Mullainathan (2016). Productivity and selection of human capital with machine learning. *The American Economic Review 106*(5), 124–27.

Chalfin, A., J. Kaplan, and M. Cuellar (2020). Measuring marginal crime concentration: A new solution to an old problem.

Christopher, W. (1991). Independent commission on the los angeles police department.(1991) report of the independent commission on the los angeles police department. *Los Angeles, CA: The Commission*.

Cuttler, M. J. and P. M. Muchinsky (2006). Prediction of law enforcement training performance and dysfunctional job performance with general mental ability, personality, and life history variables. *Criminal Justice and Behavior 33*(1), 3–25.

Dharmapala, D., R. H. McAdams, and J. Rappaport (2019). Collective bargaining and police misconduct: Evidence from florida.

Engel, R. S., H. D. McManus, and T. D. Herold (2020). Does de-escalation training work? a systematic review and call for evidence in police use-of-force reform. *Criminology & Public Policy*.

Fyfe, J. J. (1980). Always prepared: Police off-duty guns. *The Annals of the American Academy of Political and Social Science 452*(1), 72–81.

Fyfe, J. J. and R. Kane (2006). *Bad cops: A study of career-ending misconduct among New York City police officers*. John Jay College of Criminal Justice.

Goh, L. S. (2020). Going local: Do consent decrees and other forms of federal intervention in municipal police departments reduce police killings? *Justice Quarterly*, 1–30.

Goldman, R. L. and S. Puro (2001). Revocation of police officer certification: A viable remedy for police misconduct. *Saint Louis University Law Journal 45*, 541.

Goncalves, F. and S. Mello (2020). A few bad apples?: Racial bias in policing. *The American Economic Review*.

Greek, C. (2007). The big city rogue cop as monster: Images of nypd and lapd. *Monsters in and among us: Toward a Gothic criminology*, 164–198.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Helsby, J., S. Carton, K. Joseph, A. Mahmud, Y. Park, A. Navarrete, K. Ackermann, J. Walsh, L. Haynes, C. Cody, et al. (2018). Early intervention systems: Predicting adverse interactions between police and the public. *Criminal Justice Policy Review 29*(2), 190–209.

Hipp, J. R. and Y.-A. Kim (2017). Measuring crime concentration across cities of varying sizes: Complications based on the spatial and temporal scale employed. *Journal of Quantitative Criminology 33*(3), 595–632.

Hughes, F. and L. Andre (2007). Problem officer variables and early-warning systems. *Police Chief 74*(10), 164.

Invisible Institute, T. (2018). The citizens police data project.

Ivkovic, S. K. (2009). Rotten apples, rotten branches, and rotten orchards: A cautionary tale of police misconduct. *Criminology & Public Policy 8*, 777.

Kane, R. J. and M. D. White (2009). Bad cops: A study of career-ending misconduct among new york city police officers. *Criminology & Public Policy 8*(4), 737–769.

Kelly, J. and M. Nichols (2020). We found 85,000 cops who've been investigated for misconduct. now you can read their records. *USA Today*.

Kleinberg, J., H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan (2018). Human decisions and machine predictions. *The Quarterly Journal of Economics 133*(1), 237–293.

Knox, D., W. Lowe, and J. Mummolo (2019). Administrative records mask racially biased policing. *The American Political Science Review*, 1–19.

Leinfelt, F. H. (2005). Predicting use of non-lethal force in a mid-size police department: A longitudinal analysis of the influence of subject and situational variables. *The Police Journal 78*(4), 285–300.

Levin, A., R. Rosenfeld, and M. Deckard (2017). The law of crime concentration: An application and recommendations for future research. *Journal of Quantitative Criminology 33*(3), 635–647.

Long, M. A., J. E. Cross, T. O. Shelley, and S. Kutnjak Ivković (2013). The normative order of reporting police misconduct: Examining the roles of offense seriousness, legitimacy, and fairness. *Social Psychology Quarterly 76*(3), 242–267.

Lum, C. (2016). Murky research waters: The influence of race and ethnicity on police use of force. *Criminology & Public Policy 15*, 453.

MacDonald, J. and J. Klick (2020). Hire more cops to reduce crime. *City Journal*.

MacDonald, J. M., R. J. Kaminski, and M. R. Smith (2009). The effect of less-lethal weapons on injuries in police use-of-force events. *American Journal of Public Health 99*(12), 2268–2274.

Mollen, M. (1994). *Commission report*. The Commission.

Mummolo, J. (2018). Modern police tactics, police-citizen interactions, and the prospects for reform. *The Journal of Politics 80*(1), 1–15.

Nagin, D. S. and C. W. Telep (2020). Procedural justice and legal compliance: A revisionist perspective. *Criminology & Public Policy*.

Newman, M. E. (2005). Power laws, pareto distributions and zipf's law. *Contemporary Physics 46*(5), 323–351.

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

Pareto, V. et al. (1971). Manual of political economy.

Powell, Z. A., M. B. Meitl, and J. L. Worrall (2017). Police consent decrees and section 1983 civil rights litigation. *Criminology & Public Policy 16*(2), 575–605.

Ramirez, D., M. Wraight, L. Kilmister, and C. Perkins (2018). Policing the police: Could mandatory professional liability insurance for officers provide a new accountability model. *American Journal of Criminal Law 45*, 407.

Ridgeway, G. (2016). Officer risk factors associated with police shootings: a matched case–control study. *Statistics and Public Policy 3*(1), 1–6.

Ridgeway, G. (2018). Policing in the era of big data.

Ridgeway, G. (2020). The role of individual officer characteristics in police shootings. *The ANNALS of the American Academy of Political and Social Science 687*(1), 58–66.

Ridgeway, G. and J. M. MacDonald (2009). Doubly robust internal benchmarking and false discovery rates for detecting racial bias in police stops. *Journal of the American Statistical Association 104*(486), 661–668.

Rozema, K. and M. Schanzenbach (2019). Good cop, bad cop: Using civilian allegations to predict police misconduct. *American Economic Journal: Economic Policy 11*(2), 225–68.

Sherman, L. W. (1978). *Scandal and reform: Controlling police corruption*. University of California Press.

Sherman, L. W. (2018). Reducing fatal police shootings as system crashes: Research, theory, and practice.

Sherman, L. W. (2020). Targeting american policing: Rogue cops or rogue cultures?

Skolnick, J. H. and J. J. Fyfe (1993). *Above the law: Police and the excessive use of force*. Free Press New York.

Sousa, W., J. Ready, and M. Ault (2010). The impact of tasers on police use-of-force decisions: Findings from a randomized field-training experiment. *Journal of Experimental Criminology 6*(1), 35–55.

Walker, S., G. P. Alpert, and D. J. Kenney (2000). Early warning systems for police: Concept, history, and issues. *Police Quarterly 3*(2), 132–152.

Walker, S., G. P. Alpert, and D. J. Kenney (2001). *Early warning systems: Responding to the problem police officer*. US Department of Justice, Office of Justice Programs, National Institute of Justice.

White, M. D. (2008). Identifying good cops early: Predicting recruit performance in the academy. *Police Quarterly 11*(1), 27–49.

Wood, G., T. R. Tyler, and A. V. Papachristos (2020). Procedural justice training reduces police use of force and complaints against officers. *Proceedings of the National Academy of Sciences 117*(18), 9815–9821.

Wu, K. J. (2019). Study finds misconduct spreads among police officers like contagion. *PBS*.

Figure 1: Actual Versus Simulated Concentration of Complaints Against Chicago Police Officers

A: All Complaints

B: Use of Force Complaints

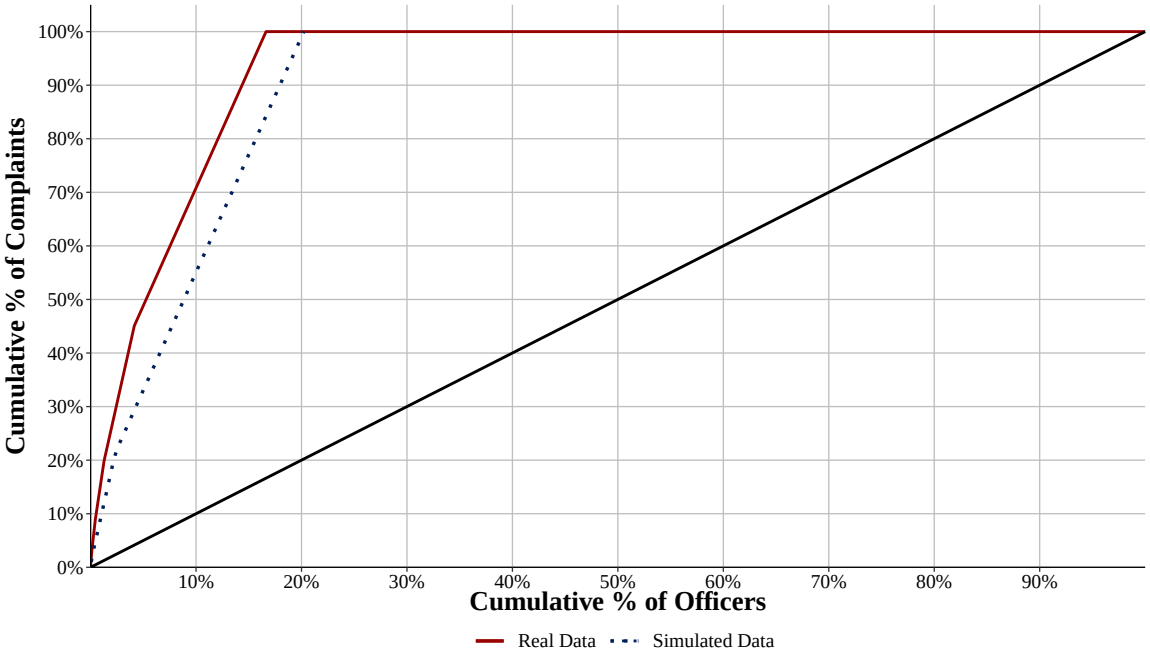

Table 1: Persistence in Use of Force

		Ten-Year Post Probationary Period			
		Top 2%	Top 5%	Top 10%	Top 20%
Probationary Period	Top 2%	7.35%	13.24%	32.35%	51.47%
	Top 5%	5.29%	11.18%	25.29%	41.76%
	Top 10%	4.41%	10.29%	21.76%	39.12%
	Top 20%	3.24%	8.24%	16.62%	30.00%

Panel A: All Complaints

		Ten-Year Post Probationary Period			
		Top 2%	Top 5%	Top 10%	Top 20%
Probationary Period	Top 2%	7.35%	17.65%	26.47%	47.06%
	Top 5%	6.47%	14.71%	24.71%	42.35%
	Top 10%	5.00%	11.18%	18.82%	33.53%
	Top 20%	3.68%	9.26%	17.65%	31.18%

Panel B: Use of Force Complaints

Note: For both total complaints and use of force complaints, we estimate the percent of people who are in the top k% of complaints in their probationary period (the first 18 months after they are hired) that are also in the top k% of complaints in the 10 year follow-up period.

Table 2: Policy Simulation: Estimated Percent Change in Complaints When Replacing the Top k Percent of Officers With Officers in the Middle 20 Percent of the Risk Distribution [18-Month Probationary Period]

	All Complaints	Use of Force Complaints
Top 1%	-0.78%	-1.01%
Top 2%	-1.46%	-2.00%
Top 5%	-2.62%	-4.21%
Top 10%	-4.72%	-5.98%

Note: For both total complaints and use of force complaints, we estimate the reduction in the number of complaints during a ten-year period that would have accrued if a given share, k, of Chicago police officers, ranked according to the number of complaints they accrued during their probationary period (the first 18 months after they are hired), had been terminated at the end of their probationary period and replaced with officers drawn from the middle 20 percent of the distribution.

Table 3: Policy Simulation: Estimated Percent Change in Complaints When Replacing the Top k Percent of Officers With Officers in the Middle 20 Percent of the Risk Distribution [5-Year Probationary Period]

	All Complaints	Use of Force Complaints
Top 1%	-2.00%	-3.51%
Top 2%	-3.99%	-5.79%
Top 5%	-7.55%	-11.62%
Top 10%	-13.21%	-16.47%

Note: For both total complaints and use of force complaints, we estimate the reduction in the number of complaints during a ten and a half-year period that would have accrued if a given share, k, of Chicago police officers, ranked according to the number of complaints they accrued during the first five years after they are hired, had been terminated at the end of their probationary period and replaced with officers drawn from the middle 20 percent of the distribution.