Cambridge Journal of Evidence-Based Policing (2024) 8:4
https://doi.org/10.1007/s41887-023-00094-1

RESEARCH

Forecasting for Police Officer Safety: A Demonstration of Concept

Brittany Cunningham[1] · James Coldren[1] · Benjamin Carleton[1] · Richard Berk[2] · Vincent Bauer[1]

Received: 21 August 2023 / Revised: 30 August 2023 / Accepted: 12 December 2023

Abstract

Purpose Police officers in the USA are often put in harm's way when responding to calls for service. This paper provides a demonstration of concept for how machine learning procedures combined with conformal prediction inference can be properly used to forecast the amount of risk associated with each dispatch. Accurate forecasts of risk can help improve officer safety.

Methods The unit of analysis is each of 1928 911 calls involving weapons offenses. Using data from the calls and other information, we develop a machine learning algorithm to forecast the risk that responding officers will face. Uncertainty in those forecasts is captured by nested conformal prediction sets.

Results For approximately a quarter of a holdout sample of 100 calls, a forecast of high risk was correct with the odds of at least 3 to 1. For approximately another quarter of the holdout sample, a forecast of low risk was correct with an odds of at least 3 to 1. For remaining cases, insufficiently reliable forecasts were identified. A result of "can't tell" is an appropriate assessment when the data are deficient.

Conclusions Compared to current practice at the study site, we are able to forecast with a useful level of accuracy the risk for police officers responding to calls for service. With better data, such forecasts could be substantially improved. We provide examples.

Keywords Police officer risks · Calls for service · 911 professionals · Forecasting · Machine learning · Conformal prediction sets

✉ Richard Berk
berkr@sas.upenn.edu

[1] CNA, Arlington, USA

[2] University of Pennsylvania, Philadelphia, USA

Published online: 05 March 2024

Introduction

In the USA, police officers sometimes face the most unpredictable, traumatic, and violent circumstances of any profession (White et al., 2019). Although much of an officer's workday is comprised of brief, unremarkable events, some calls for service can escalate into life-threatening situations. For officers to adequately mitigate the risks, they must be well informed about the types of risks they might face, settings that are especially dangerous, and tactics that can enhance their safety.

Prior literature has focused heavily on how technical advances such as body armor and patrol car design can improve officer safety in the field, especially when coupled with better training (Cunningham et al., 2021). Less attention has been given to the role that 911 professionals (i.e., call-takers and dispatchers) play in enhancing officer safety. 911 professionals are responsible for extracting critical details from 911 callers, assessing the risks, and transmitting such information clearly to responding police units (Gillooly, 2020; 2022; Neusteter et al., 2019; https://www.911.gov/). Consequently, they have a huge role in shaping officers' perceptions of risk when responding to calls (Gillooly, 2022; Taylor, 2020).

Call-takers and dispatchers are not immune to fatigue, overload, and questionable judgments that can influence their assessments of risk (Gillooly, 2022; Manning, 1988). In addition, callers will sometimes provide incorrect and confusing information that typically is transmitted in a very short time; there is little opportunity to confirm and double-check key information. Despite these and other challenges, the information conveyed to dispatched police officers will in practice become the raw material for an operational *forecast* of what might happen at the crime scene when the police arrive. It is the forecast, not a set of loosely connected and incomplete facts, that helps frame how police officers respond. Yet, this forecasting framing has received almost no attention when dispatching is studied.

Forecasts have long been at least implicit in a wide variety of police planning and decision making. The allocation of police assets to different neighborhoods is one common example. Over the past several decades, the forecasting has in many law enforcement settings become more data driven, more statistical, and more concerned with statistical and causal inference. COMPSTAT was an early illustration, but currently "AI" in the form of predictive policing and offender risk assessment have made explicit forecasting almost commonplace (Berk, 2021).[1]

The pages ahead draw on these criminal justice forecasting experiences. We offer a demonstration of concept illustrating the possibilities from using machine learning to forecast the risks that police officers face when dispatched in response to a call for service.[2] Our forecasting algorithm is developed from information provided by 911 calls, officers' completed offense forms, and a variety of particulars from other

[1] These days, it has become fashionable to include a heterogeneous mix of statistical procedures under the rubric artificial intelligence or "AI." Yet, artificial intelligence has no consensual definition nor a definitive list of included procedures (Berk, 2021). To avoid confusion, we will not use the term.

[2] As used here, a "demonstration of concept" is meant to illustrate in a credible fashion that a proposed method or set of procedures has initial face validity, feasibility and promise. In contrast, a "proof of concept" should make compelling case that the method or procedures work as intended, even if just within a somewhat artificial or simplified setting (i.e., a "test bed").

sources. We quantify the uncertainty in our forecasts using nested conformal prediction sets, which are a relatively recent statistical development (Gupta et al., 2022).

We hold that useful forecasts of officer risk are feasible with only modest improvements in our approach. Some might argue that our current work already could be useful in practice. We make no claim that our data or procedures are the last word.

Past Research

Risks for Police Officers

Policing in the USA is widely considered a high-risk occupation (Bierie et al., 2016; Bierie, 2017; Crifasi et al., 2016; Mumford et al., 2021; Ricciardelli, 2018). As Brandl and Stroshine (2012: 268) note, police are "responsible for intervening in situations where they may not be invited and where they may be dealing with hostile citizens and suspects." Policing differs from other occupations "in that injury and death come not just from accidents, but from job performance" (Moskos, 2009: 1). The characterization of policing as a high-risk occupation is also supported by the conclusion that seven of the ten National Institute for Occupational Safety and Health's risk factors for workplace violence are central to police work (Crifasi et al., 2016; Fridell et al., 2009).

The numbers support most broad claims (Bierie et al., 2016; Bierie, 2017; Brandl & Stroshine, 2003; Hine et al., 2018; Sierra-Arevalo et al., 2022; White et al., 2019). For example, Maguire and colleagues (2002) found that the fatality rate of police officers is nearly three times that of the average US worker. Sierra-Arévalo et al. (2022) find that the rate of gun homicide of police is 1.6 times larger than the US rate, and for non-fatal firearm assaults, the rate is 2.7 times larger. Peek-Asa and colleagues (1997), in their study of nonfatal workplace assault injuries, report that police were 73.1 times more likely to be assaulted while at work compared to the overall average over other occupational settings. In addition, approximately 10% of police officers are assaulted each year (Bierie, 2017). According to the FBI's Law Enforcement Officers Killed and Assaulted database, 43,649 officers were assaulted in 2021 and 15,368 of the police officers sustained injuries.[3] Lastly, Sierra-Arévalo et al. (2023) found that the murder of George Floyd in 2020 was associated with a 3-week spike in firearm assaults on police.

But such statistics do not convey a full story. Officer injury data surely are imperfect. Uchida and King (2002) highlight that some agencies may not keep precise counts of officer injuries, which, in turn, compromise the precision of national estimates. Furthermore, injuries are not always reported. For example, injuries sustained during higher status calls for service are more likely to be reported than those sustained during lower status calls for service. There is little to be gained from reporting a bruised hip caused by slipping on an icy sidewalk while ticketing a parking violation.

[3] https://leb.fbi.gov/bulletin-highlights/additional-highlights/crime-data-law-enforcement-officers-assaulted-in-2021

There are also concerns that the reality of police work can be misconstrued when operationalized primarily by killings and assaults of officers in the line of duty. Such incidents are very rare (Brandl & Stroshine, 2003; Bierie et al., 2016; Hine et al., 2018; Sierra-Arevalo and Nix, 2020), and the very low risk probabilities offer an incomplete accounting. White et al. (2019) point out that, nevertheless, the large volume of police and citizen encounters makes violence against police "a daily event" widely noted in the interpersonal networks of sworn officers. Recruiting, retention, and morale can be adversely affected (Fridell et al., 2009; Kaminski & Sorensen, 1995).

Looking more closely complicates appearances. Although being shot was the leading cause of law enforcement deaths between 2012 and 2022, deaths from COVID-19[4] and motor vehicle-related incidents, including being struck by a vehicle or being involved in a crash, have consistently been among the leading causes of line-of-duty deaths of law enforcement officers.[5] Moreover, nonfatal assaults represent the most common type of violence directed at police (Sierra-Arevalo & Nix, 2020), and accidental injuries are the most common job-related hazard (Brandl & Stroshine, 2003). Accidental injuries also can have undesirable consequences. The raw number of police officers on work-related disability can be quite large and create a drain on department staffing and budgets (Brandl & Stroshine, 2002; Bierie, 2017; Fridell et al., 2009; Kaminski & Sorensen, 1995). In short, although shootings of police garner the most media attention, there are a wide variety of other risks that are far more numerous and can have adverse consequences for police officers and the departments in which they serve.

The Importance of 911 Professionals

In the USA and elsewhere, law enforcement agencies have made efforts to improve officer safety. These efforts include investments in body armor (e.g., bulletproof vests, shields, helmets), technology (e.g., electronic control weapons and conducted energy devices), and training (e.g., tactical preparedness training), among others (Cunningham et al., 2021). One domain that has not received the same attention is dispatch. Dispatch allocates police units to calls for service and transmits information about the incident to the responding units (Gillooly, 2020). "When dispatched to a distal call, an officer's initial understanding of the incident will be formed almost entirely by the information received from dispatch" (Taylor, 2020: 315).[6]

Some scholars characterize 911 professionals as gatekeepers (Lum et al., 2020), in part because call-takers and dispatchers filter out calls that may not require a police response. Lum and colleagues (2020) observed that approximately 50% of calls were resolved without a police response. Gillooly (2020) appends the characterization of "risk appraisers" to the job description. As a practical matter, 911 professionals determine the priority level of each call and the number of units that

[4] https://nleomf.org/memorial/facts-figures/officer-fatality-data/causes-of-law-enforcement-deaths/

[5] https://www.cdc.gov/niosh/topics/leo/default.html

[6] In some jurisdictions, the 911 professional answering calls (i.e., the call-taker) differs from the individual allocating units and transmitting information to the responding units (Gillooly, 2020; Neusteter et al., 2019). In other jurisdictions, a 911 professional fills both roles.

will initially respond (Gillooly, 2020). Their summary of risk then informs how quickly a unit arrives at the scene, whether the unit should wait for back-up, and how the unit should prepare tactically, among other considerations.

Gillooly (2022: 766) also stresses that 911 professionals are not "neutral conduits" that simply transfer information to responding units. Call-takers and dispatchers interpret information provided by callers, which inherently introduces personal judgments (Gillooly, 2022; Manning, 1988). In her study of how call-takers appraise risk and classify calls, Gillooly (2022) found that different call-takers will often classify similar calls differently. Accuracy can suffer as well. Call-takers and dispatchers tend to overestimate the severity of calls, with some call centers having between 20 and 40% of all crime calls answered by call-takers downgraded by once police are at the scene (Gillooly, 2020; 2022)

And it matters. Call-takers can exert substantial influence over police perceptions of the calls to which they are responding (Gillooly, 2022: 780). For mental health and public assault calls, police officers were much more likely to classify the incidents as high-priority when the call-taker initially classified the incident as high priority. In his study examining dispatch priming and the decision to use deadly force, Taylor (2020) found that when officers were told earlier that a potential perpetrator appeared to be talking on a cell phone and that individual subsequently produced the cellphone during the encounter, 6% of officers made a shooting error. Conversely, when officers were told earlier that a potential perpetrator might be holding a gun, and subsequently produced a cellphone during the encounter, 62% of the officers made a shooting error. Yet, in a replication of Taylor's (2020) study, Potts et al. (2022) used a realistic virtual reality scenario to test the effects of dispatch priming and found no overall effect of dispatch priming on the responder's likelihood of firing a gun. In short, the precise nature and size of call-taker influence on police perceptions are unresolved.

In summary, 911 professionals provide consequential links connecting calls for service to police responses. Some communication errors are inevitable. Some overestimates and underestimates of risk are inevitable as well. Thus, Bierie (2017) has emphasized the potential in "risk assessment tools for police."

We agree. To set the stage, we see this exercise as a demonstration of concept. Could dispatch data be employed to anticipate especially risky situations for police officers when they respond to 911 calls? We use the Camden County, New Jersey Police Department as a study site. Insofar as our forecasting tools work well with the data available, we hope that at other sites, better data might be collected, 911 professionals might be trained to improve risk forecasts, and the dangerous work undertaken by police officers can be made more safe.

Data Collection and Preparation

Our research team partnered with the Camden County Police Department (CCPD) to evaluate the promise of machine learning forecasts of police officer risks for improving the information extracted from 911 calls and the quality of subsequent dispatches. The CCPD is the primary provider of law enforcement services to the

City of Camden, New Jersey. Camden is located in southern New Jersey directly across the Delaware River from Philadelphia, Pennsylvania, and has an estimated population of 73,562. Camden has historically had one of the highest homicide rates in the country, with 87 murders per 100,000 residents in 2012. In recent years, the homicide rate has fallen significantly, a change that is often attributed to the adoption of "community policing" in 2013.

CCPD's primary response mechanism is its neighborhood response team mobilized units. According to CCPD's website, these units "serve as the primary tiered responders to emergency calls for service and perform neighborhood directed patrols in alignment with the daily resource deployment plan."[7] The department also uses a call prioritization system in which the closest unit is automatically dispatched to emergency calls. Units are dispatched through the Camden County Communications Center's Police Central, which is responsible for the dispatching of police from 27 municipalities and has designated dispatchers for the city of Camden.[8]

The CCPD provided us with data on *every* 911 call for service from January 1, 2015 until, December 31, 2019. During this time, the CCPD averaged 105,000 calls for service every year or about 290 calls for service every day.[9] However, any interest in forecasting the risks for police officers anticipated by these calls was complicated by the *relative* lack of calls from incidents that actually placed police officers in harm's way. Armed robberies in progress, for example, are known to be dangerous for police officers, but are typically quite rare. Most calls for service do not encode such risks. Nevertheless, progress can be made, as we hope to show below.

The unit of analysis for this forecasting enterprise is the call for service[10] We view these calls as exchangeable. The calls are seen as generated by the social, psychological, and other factors responsible for 911 calls such that the order in which the calls are received does not affect the probability distribution of those calls. This will suffice to justify the nested conformal prediction sets we later construct to convey forecasting uncertainty, and in any case is probably a realistic assumption.[11]

[7] (https://camdencountypd.org/operations-bureau/).

[8] (https://www.camdencounty.com/service/public-safety/communications-center/).

[9] There were several meetings with members of the CCPD and many follow-up phone calls and emails. Cooperation was excellent. We were provided with all of the relevant data that were collected and stored. There were data that might well have been useful that were not systematically collected. Going forward, improvements in data collection and management could be beneficial.

[10] We use the term "call" because that is the term used in the literature and by the CCPD. Multiple calls for the same incident are collapsed into a single "call."

[11] Consider a very simple illustration. Suppose for a given week there were 1000 911 calls received for three kinds of crimes. Thirty percent were for crime type A, 50% were for crime type B, and 20% were for crime type C. Now, for the same social, psychological, and other forces functioning in the same manner, suppose that nature randomly shuffled the order in which those same calls were received. The distribution of crime types would still be approximately 30% As, 50% Bs, and 20%, Cs. Time order does not affect the marginal distribution of the crime types as it would if the data were a conventional times series. However, like for *all forecasting*, we assume that the processes that generated the data on hand are reasonably stable going forward; the future is sufficiently like the past. For some kinds of longitudinal data, this can more formally imply "strict stationarity" (Box et al., 2017: 506-508); the joint distribution of the data does not change over time. Temporal stability can be addressed empirically as new data become available, from which statistical adjustments and improvements should be introduced as needed. "Updating" forecasting procedures is an important part of recommended practice.

Our binary response variable is whether a dispatch leads to a high compared to a low risk encounter——both defined in some detail shortly. For our CCPD data, the proportion of incidents overall putting police officers in substantial danger is less than 1% because the base number of calls overall is so large (i.e., 309,490). Statistically, the several thousand calls conveying significant risk were *relatively* rare events. It follows that our binary response variable is very badly unbalanced.[12]

For some preliminary analyses, the forecasts of risk for police officers using all the available data were very accurate and demonstrably pointless. Using no predictors at all, a forecast of low risk would be correct over 99% of the time employing information solely from the Bernoulli distribution of the response variable. Very high accuracy from the marginal distribution of a binary response variable by itself typically makes moot information available from promising predictors. Simply put, the vast majority of dispatches from 911 calls in our data posed virtually no risk to responding officers, and forecasting low risk automatically for all new calls as they were received would almost certainly be right most of the time. Nevertheless, the relatively small number of dispatches that posed a danger to responding officers remained a legitimate concern. Police officer injuries and deaths could result. And even for incidents in which the dangers were safely managed, emotional stress would likely be significantly elevated. These were for police officers low probability, high cost incidents. We were tasked with searching for the proverbial needle in a haystack.[13]

It is routine practice when using the statistical tools that we describe shortly to subset the data into random disjoint subsets, usually labeled training data and testing (or test) data (Berk, 2018). The training data are used to construct a forecasting algorithm. The testing data are used to evaluate the algorithm, in an "honest" manner, uncontaminated by the how the forecasting algorithm was built and how well the forecasting algorithm performs. There were 232,118 calls in the training data and 77,372 calls in the testing data representing 75% and 25% of the data respectively.[14]

Addressing the highly unbalanced response variable led to additional data splits. We further subdivided the training and testing data by broad categories for different kinds of calls for service and then examined each subset separately. We were seeking crime types with less unbalanced response variables. In effect, we sought to make the haystack smaller. For example, one such subset was calls stating or implying the use or presence of firearms. In the training data, there were 1928 such

[12] Information on officer injury types was available as short free-text descriptions but was not systematically organized. There also was no scale of injury seriousness, and we did not impose any threshold. Future research probably would benefit from systematic data permitting an empirical consideration of seriousness. No doubt seriousness helps shape which injuries are reported.

[13] Low-probability, high-cost events are increasingly being studied in a wide variety of fields. Climate change is perhaps the most salient setting. Rare and extreme heat waves are one example. They are highly localized in time and space but with devastating consequences. On any given day, the chances of such an event anywhere is close to 0.0, but over the course of a year, several sites will beat the odds. Extreme, rare heat waves need to be far better understood and forecasted, but because the events are rare, the statistical challenges are substantial (McKinnon & Simpson, 2022).

[14] There is little formal statistical guidance on what splitting proportions to use (Dobbin & Simon, 2011; Xu & Goodacre, 2018). Given the large sample sizes, any of the usual splitting proportions (e.g., 50% and 50%, 60% and 40%, 70% and 25%) would not materially affect the results.

calls or about 4.5 calls per day. Of these, 278 resulted in weapons-related criminal charges. Although knowing that these calls resulted in weapons-related charges does not indicate that the specific calls were truly high-risk when they occurred, we can reasonably assume that at least a portion of these 278 calls required police officers to apprehend an armed offender at some substantial risk to themselves.[15]

The resulting response variable was far less unbalanced. 86% of the calls, not 99%, were low risk. An illustrative very low risk incident had the alleged offender leaving the scene before the police arrived. In short, the response variable was still unbalanced but within a range that often allows for performance gains from available predictors.

We generated 81 predictor variables *conceivably known to a dispatcher*, from a call for service and other information that could be rapidly and routinely accessed. We were careful to exclude predictors that could only be known after officers arrived at the scene of the 911 call because such information could not be known when a dispatch was made. These 81 variables fell into eight categories as follows.

- Crime type included in the dispatch
- Date and time: We included information about the hour, day of week, month, and year in which the incident took place, which allows the algorithm to identify some important temporal patterns
- Initiation type: Whether a community member or an officer initiated the call for service
- Weather conditions: We included several variables related to the weather conditions on the day that the incident took place, including falling snow, rain, temperature, and the presence of fog
- Local trends: Local patterns of activity may be important predictors of how a particular incident will unfold, so we included counts of arrests and calls for service for any reason within the past 30, 90, and 180 days in the same police sector (a sector is a spatial unit used by the CCPD roughly comparable to a neighborhood). All counts are generated one day prior to the incident in question, which means that the counts include only past events for any given call for service[16]
- Local timing details: We included for each police sector the number of days since the last injury, arrest, or call for service for any reason at the address because repeated calls to the same address may be a risk factor. All counts were gen-

[15] Such charges typically are filed long after police have left the crime scene. Nevertheless, the charges are an important indication that the apprehended offender likely had access to a firearm, which in principle can put responding officers at greater risk. The same reasoning applies to the timing of officer offense forms. Information from neither source was used to construct predictors. Such information was used to help construct the response variable only.

[16] Consider counts for the past 30 days from a given CCPD sector. Calls for service on, say, day 10 would have past incident counts downwardly biased because incidents eleven or more days earlier would not be known. An overall solution would be to drop from later forecasting analyses all 911 calls received 30 days or less from the earliest date in the data. The same logic applies to all 90- and 180-day variables. In the forecasting analyses to follow, we dropped the incidents from the first 6 months to remove any downward measurement bias from the incomplete data. Thanks to our large sample size, the loss of the earliest year of data did not seem to cause any analysis problems.

erated one day prior to the incident in question, which means that the counts include only past events for any given call for service (see also footnote 16)

- Census tract information: We included aggregated information on housing vacancies, employment, population density, and race for the census tract in which the incident took place, using data from the 2010 Census. Neighborhood characteristics raise important questions about fairness but may also be important predictors of call outcomes. We do not know, however, how census tracts overlap with the spatial units used by the Camden Police Department (CCPD), which makes them additionally problematic for our analyses.

The binary response variable was constructed using several indicators of risk to police officers, some taken from the dispatch information and some from officer-provided offense forms.[17] Risk included officer injuries, incidents in which the suspect(s) eluded or resisted arrest, in which the suspect(s) possessed a weapon(s), and all crimes "in progress" when the call for service was received. Incidents in which the suspect(s) eluded arrest were defined as dangerous because they often involved a chase on foot. A crime in progress was defined as dangerous because police officers likely would drive at high speeds to reach the incident location and because perpetrators were more likely to still be actively engaged in their criminal behavior. "High risk" was coded "1," and "low risk" was coded "0."

It was apparent that many of the predictors were highly correlated and challenging to interpret in a "held constant" framework. To illustrate, for a given police sector, what would one make of the relationship between the risk for police officers and the number of past calls for service over the preceding 180 days holding constant the number of past calls over the preceding 90 days and also the preceding 30 days? Using methods described in the next section, we were able to reduce the number of predictors to seven with no meaningful reduction in forecasting performance. The seven predictors were (1) Fridays, (2) Saturdays, (3) evening hours, (4) the month June, (5) the month of July, (6) the month of August, and (7) the number of past 911 calls from the particular police sector over the past 30 days prior to a given call. The locale was the same as the spatial origination of the call. Our variable selection was undertaken by applying stochastic gradient boosting to the training data as described immediately below and by removing predictors that were not "important" for the fit. Importance was measured by contributions to reductions in the deviance. Given the bias likely introduced by variable selection (Berk et al., 2013), all results reported below are computed as needed from the testing data, not the training data. Post model-selection bias from empirically chosen predictors is not carried forward in testing data (Berk et al., 2013).

[17] The response variable required for algorithmic training necessarily includes information that cannot be known for *new* incidents when a forecast of risk is needed. If that information were available, there would be no need for a forecast. This is a standard forecasting protocol.

Statistical Methods for Forecasting Risk

We begin this section with the most relevant meta-issues. In mathematical statistics and common statistical applications, a model is meant to represent literally how the data were generated (Freedman, 2009). Conventional linear regression is a popular example. Developing a model for this paper was ruled out a priori because there is now ample evidence and formal mathematics demonstrating that machine learning algorithms in the social and biomedical sciences typically will forecast at least as well as models, and usually better (Berk & Bleich, 2013; Berk, 2018). Moreover, models typically require, before the analysis begins, far greater subject-matter knowledge than algorithms require. In particular, models depend on a pre-specified structure (e.,g., linear and additive in the predictors), whereas most forecasting algorithms are non-parametric and can respond adaptively to linear as well as nonlinear associations found in the data (Berk et al., 2023). As Kearns and Roth emphasize in their book *The Ethical Algorithm* (2020: 4) "An algorithm is nothing more than a very precisely specified series of instructions for performing some concrete task." Consequently, misspecification is not even defined. This can serve forecasting skill very well, but usually makes explanation a secondary consideration.

A valid data analysis requires "training data" to develop a forecasting procedure and separate "testing data" to properly evaluate its performance. The data from which forecasts subsequently are constructed must be generated in the same manner as the data used to build and assess the forecasting tool. All such observations should be realized independently and identically from the same distribution (i.e., i.i.d.). Exchangeable observations can be a proper fallback position. Note that testing data are well known to be essential in algorithmic forecasting for obtaining valid measures of forecasting uncertainty, protecting against potential cherry-picked results, and precluding overfitting (Hastie et al., 2016).

If the outcome to be forecasted is categorical, the forecasting procedure should be a classifier. We applied stochastic gradient boosting (Friedman, 2001) used as a classifier because it can be easily tuned to capture relatively rare events.[18] Stochastic gradient boosting is a form of supervised machine learning.

We applied an excellent implementation of stochastic gradient boosting, *gbm,* available in the R programming language. Stochastic gradient boosting is among the best performing classifiers readily available and will yield asymptotically unbiased forecasts conditional on the predictors available and the values of tuning parameters. Interaction depth was set at 5, because deep classification trees were needed to find the rare, high risk events. Minimum node size was fixed at 1 consistent with the recommendations of Wyner et al. (2015), which helps our classifier perform like an interpolator that, in turn, can formally determine how superior performance may be

[18] When used as a classifier, stochastic gradient boosting can be seen as a form of non-linear, non-parametric, logistic regression. Because the specification is non-parametric, there are no regression coefficients to interpret. We could almost as easily have used random forests or deep neural networks with no important differences in the results (Berk, 2020).

achieved (Liang & Recht, 2021). Because the number of predictors was small, the fit stopped improving in a meaningful way at about 50 iterations.[19]

Finally, the consequences of false negatives and false positives are not the same in our forecasting application. Their differential costs were incorporated into the analysis by weighting the data. In particular, the costs of having responding police officers surprised by unexpected risk were seen by stakeholders as substantially worse than having officers over-prepared.[20] Provisionally, the costs were set at 10 to 1; being under-prepared was seen as about 10 times worse than being over-prepared. With a forecast of high risk labeled a "positive," and low risk labeled a "negative," the weighting meant that the boosting algorithm would work substantially harder to avoid false negatives than to avoid false positives. This was precisely the intent.[21] A cost ratio is a subject-matter decision, not a tuning parameter. Different cost ratios might be appropriate depending on the settings and stakeholders.[22]

Results

Stochastic gradient boosting was applied to each dataset for each broad crime category whose response variable imbalance was not an insurmountable obstacle. In the interest of space, we focus on weapons-related dispatches and consider other kinds of dispatches only in passing.

Table 1 is a standard confusion table constructed using the testing data. It is a cross-tabulation of the outcome class labels in the testing data and the outcome classes determined by trained, risk algorithm applied to the testing data. One can see that the target cost ratio of 10 to 1 is very well approximated (524/53 = 9.88) despite being a product of the testing data.[23]

An apparent problem is that there are nearly 5 times more false positives than true positives (i.e., 524/109 = 4.81). However, the large number of false positives (i.e., 524) follows directly from the 10 to 1 cost ratio. When a classifier is working especially hard to avoid false negatives, the mathematical tradeoff

[19] As its name implies, algorithm has stochastic features. Consequently, different fitting attempts can vary a bit in the results and the number of iterations used. But the results are always asymptotically unbiased, at least in principle.

[20] For this preliminary study, the main stakeholders were representatives of the CCPD and Bureau of Justice Assistance grant monitors.

[21] The labels "positive" and "negative" are somewhat arbitrary but convey the intent of the forecasting task. A dangerous encounter was a positive because it was the kind of incident motivating the forecasting enterprise to begin with. It was the sought-after needle in a haystack.

[22] The parties included as stakeholders will vary by setting. Broadly speaking, stakeholders have skin in the game and potential influence. Obvious candidates on the matter of officer safety are top police administrators, police union representatives, unions representing dispatchers, citizen groups authorized to provide oversight of the police department, community groups concerned about policing practices, and politicians concerned about the monetary costs such as paid leaves while an officer recovers from injuries in the line of duty. Some stakeholders have an intermittent role triggered by particular safety issues. One example is representatives of firms that sell police body armor or body worn cameras.

[23] The training data have the same structural properties as the testing data because the disjoint split was applied at random, but estimates from random disjoint samples can differ a bit.

Table 1 Confusion table for weapons dispatches (labeled outcome classes high risk or low risk refer to outcome classes in the data; predicted outcome classes high risk or low risk are outcome classes determined by the algorithm)

	Predict low risk	Predict high risk	Classification error
Labeled low risk	286	524	.64
Labeled high risk	53	109	.33
Forecasting error	.16	.82	–

encourages false positives, which are far less costly. Should stakeholders choose to make the cost ratio more balanced, there would be fewer false positives and more false negatives. This tradeoff depends on a policy choice and can easily be changed.

The cost ratio leads as well to rather different misclassification rates for classes labeled in the data as low risk or high risk calls. Sixty-four percent of the calls labeled low risk in the data are misclassified as high risk calls. Thirty-three percent of the calls labeled as high risk in the data are misclassified for low risk calls. The good news is that about 2/3rds of the calls labeled high risk were correctly classified. More good news, also a product of the 10 to 1 cost ratio, is that although the confusion table forecasting error for high risk class is 82% (i.e., 525/633), the confusion table forecasting error for low risk dispatches is only 16% (i.e., 53/339).[24] With the algorithm working so hard to avoid false negatives, it should be no surprise that there are only 53 of them. Forecasts of low risk from the confusion table can from this analysis be quite credible.

Note that the term “forecasting error” is with respect to outcomes classes as labeled *in the data on hand*, not to outcome classes that are at this point unknown. Those unknown outcome classes to be forecasted are addressed with conformal prediction sets provided shortly.

Also of interest is which predictors are driving the risk algorithm’s fitted values. This information can have important policy implications. For example, even very good forecasting results can be questioned if the most important predictors strongly contradict existing research and/or widely accepted subject-matter assumptions.[25]

Figure 1 shows the *relative* contribution of each predictor to the classifier’s fit of the training data (i.e., the contributions sum to 100%). About 80% of the deviance reduction can be attributed to the number of 911 calls from particular locations in the past 30 days, even though most are not repeat calls. This replicates the well

[24] By definition, classification error conditions on the outcome class labels in the data, and forecasting error conditions on the outcome class predicted by the machine learning algorithm. For example, the .64 in the upper right hand corner of Table 1 has the total number of outcomes labeled in the data as low risk in the denominator (i.e., 286 + 524) and the number of those cases claimed by the algorithm as high risk (i.e., 525) in the numerator. The .16 in the lower left hand corner of Table 1 has the total number of outcomes claimed by the algorithm to be low risk in the denominator (i.e., 286 + 53) and the number of those cases labeled in the data as high risk (i.e., 53) in the numerator.

[25] The issues can be subtle because as noted above, an algorithm is not a model.

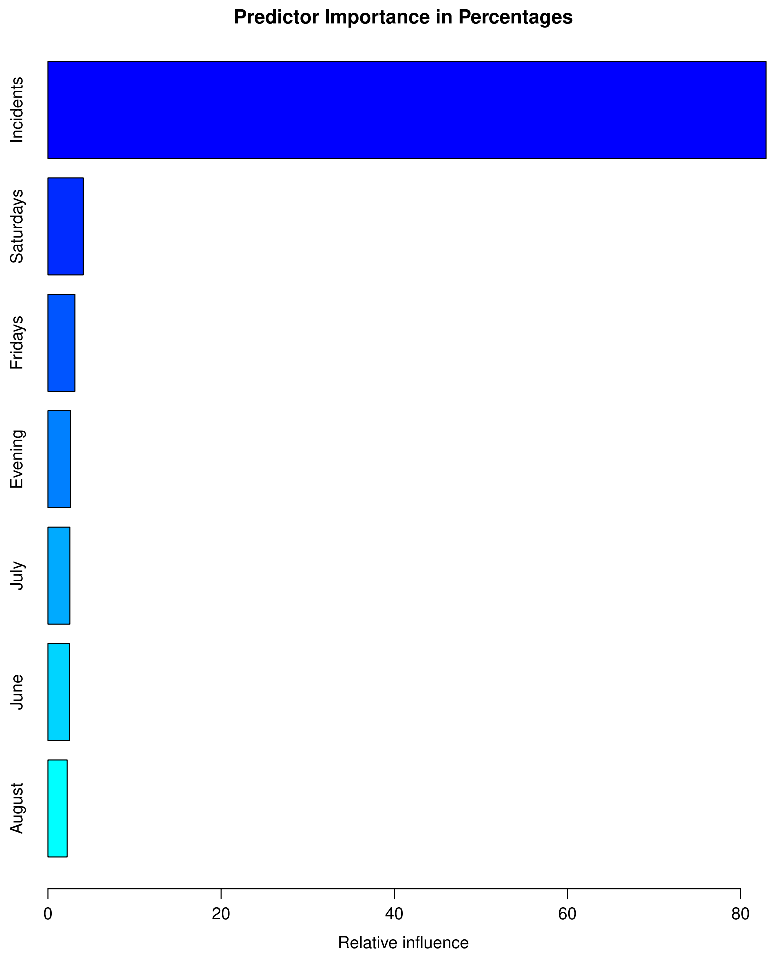

Fig. 1 Weapons dispatches: relative variable importance for the boosting fit

known finding that crime and calls for service are spatially concentrated. All of the other predictors matter as well, but far less. For example, there are greater risks to police officers on weekends, evenings and during the summer months. None of these is a surprise.

However, all of the predictors are empirically related such that, for example, a substantial fraction of the 80% relative deviance reduction attributed to the number of past 911 calls is actually shared with other predictors because of statistical interaction effects. To provide more details about these interaction effects can be somewhat involved (Molnar, 2022) and beyond the scope of this paper. We would proceed with such a discussion were we building a causal model.[26]

Further information can be extracted from partial dependence plots (Friedman, 2001). These show the relationship between a given predictor and the response, holding all other predictors constant in a novel manner.[27] For example, Fig. 2 displays how the number of incidents leading to calls for service in the preceding 30 days from a given police sector is related to the probability of a high risk dispatch. The rug plot at the bottom indicates that there are relatively few calls from police sectors with more than 1200 calls for service. The plot to the right of 1200 should probably be ignored. It rests on very sparse data, and the smoothed values in black are distorted downward. For the remaining cases, one essentially has a step function, shown in red.[28]

The relationship in Fig. 2 is highly nonlinear. A credible conclusion is that the probability of a high risk dispatch increases sharply from around .20 to about .60 (with a maximum of more than .80) as the number of prior incidents increases from about 100 to about 800 (with considerable local variation) and then levels off. The slight decline in risk from about 800 calls to about 1200 calls cannot be distinguished from noise. Within the range of calls with sufficient data, a greater density of 911 calls is associated in a complex fashion with a higher risk for police officers, all other included predictors held constant.[29]

One must be careful about attributing a *causal* effect to the past number of calls for service. It is unlikely that the number of past calls for service in a given locale directly increases risk. More likely, the number of past calls for service is an indicator of the causal social factors responsible for high crime rates, such as social disorganization (Sampson & Groves, 1989). These, in turn, affect the risks experienced by police officers. Also, if our boosting formulation were (inappropriately) treated as a causal model, it would no doubt be badly misspecified. That is acceptable when forecasting accuracy is the priority. Recall that the forecasts are asymptotically unbiased given the tuning and predictors included. But that does not preclude greater accuracy with a better set of predictors.

Prediction Accuracy Through Nested Conformal Prediction Sets

One of the major problems with interpretations of confusion tables is that the reliability of the predictions is not considered. The outcome class predicted by the classifier is the outcome class with the largest estimated probability. For example, if for a given case the estimated probability of high risk is .80, and the estimated probability

[26] An example of a simple interaction effect is that evening dispatches are made especially high risk during the summer months. The association with risk is greater than the association with the risks of summer months and evenings added together.

[27] The familiar covariance adjustments used in regression models is not being used. A clever form of matching is employed instead (Friedman, 2001).

[28] The smoothed values were fit with the procedure loess in R. The step function is just an eye-balled summary for ease of interpretation.

[29] This complexity would be fundamentally misrepresented by a conventional logistic regression model.

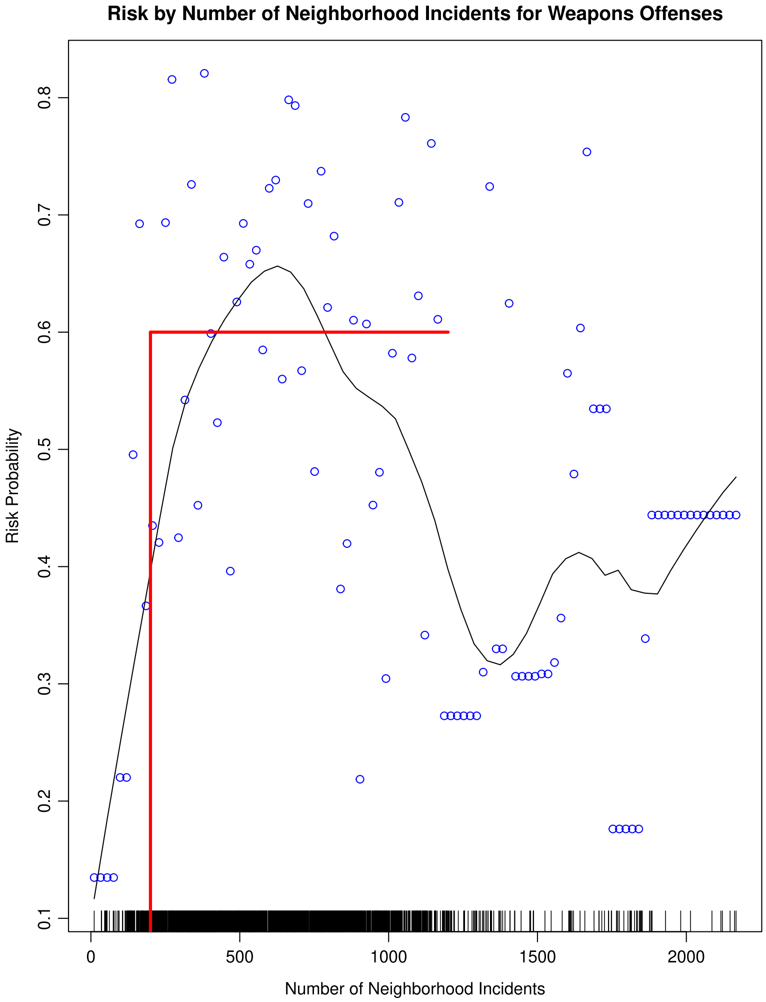

Fig. 2 Weapons dispatches: smoothed dependence plot for the number incidents in the past 30 days in particular police reporting districts

of low risk is .20, most would consider the forecast of high risk quite reliable. But if the two probabilities are .53 and .47 respectively (or even closer), the forecast is being made by a procedure that is just a little better than a 50-50 coin flip. The reliability is low and likely provides poor guidance for 911 professionals to pass along to the responding officers.

When the two outcome probabilities are near one another, a proper policy conclusion is that the algorithm is unable to make a definitive decision about the most likely outcome class. These and other difficulties with confusion tables are discussed in Kuchibhotla and Berk (2023). We need to do better forecasting police officer risk.

Conformal prediction sets provide a better, principled solution (Angelopoulos & Bates, 2022). They have some of the look and feel of confidence intervals. But whereas a confidence interval is an estimated region in which population parameter estimates (e.g., an estimate of a population proportion) will fall with a certain probability, conformal prediction sets estimate the *future outcome class (or classes)* that will be right with a certain probability. The former conveys the "certainty" associated with, say, the estimated proportion of high risk incidents over many past dispatches to armed robberies in progress. The latter conveys the "certainty" associated with *forecasts* of high risk for many *new* dispatches in response to *new* armed robberies in progress.[30]

Consider now the two outcome classes we have been using: high risk and low risk. With conformal prediction sets and two outcome classes coded "1" for high risk and "0" for low risk, there are four prediction sets logically possible for each case needing a forecast: $\{1\}$, $\{0\}$, $\{1, 0\}$, $\{\varnothing\}$. The first set is a forecast of high risk, the second set is a forecast of low risk, and the third set is a "can't tell" result because the classifier is unable to make a reliable distinction between high risk and low risk. The fourth is an empty set indicating that no forecast at all is made because the given case likely is an outlier that is very different from the data on which the classifier was trained.

We held out 100 cases at random from calibration data and treated them as if we did not know the outcome class, just as if each was a dispatch for which a forecast was needed. We constructed a *nested* conformal prediction set for each dispatch based on the pseudocode provided by Kuchibhotla and Berk (2023).[31] The coverage probability was fixed in advance at .75.

Of those 100 randomly selected holdout cases, the prediction set included only the high risk outcome class for 26% of the cases, meaning that for these cases, high risk was the forecast. The prediction set included only the low risk outcome class for 24% of the cases, meaning that for these cases, low risk was the forecast. Each of these two prediction sets contain the true future outcome with a probability of at least .75. In practice, this means that about 75 out of 100 such conformal forecasts will be correct.

For 50% of the 100 cases, both outcome classes were included in the prediction set, meaning that the classifier could not make reliable distinction between the

[30] The validity of both approaches in practice depends on how the data were generated. Confidence intervals are usually computed assuming that the data were generated by probability sampling or natural, close approximations thereof. Valid conformal prediction intervals and prediction sets are usually computed assuming that the data are exchangeable. With exchangeable data, the new cases needing forecasts are from the same distribution as the training and testing data. They just happen to be the most recent observations, which under exchangeability does not matter. Consequently, one already knows from the existing data the probability that any particular forecasted outcomes falls within a particular range. For technical reasons, however, one does not work with outcome classes, but nonconformal scores constructed in part for those outcomes.

[31] Nested conformal prediction sets are the most precise prediction sets possible (i.e., with the fewest outcome classes), given exchangeable data, conditional on the classifier used, the predictors included, and the specified coverage probability. No other assumptions are needed (Arun Kumar Kuchibhotla, personal communication).

two. The 50% figure is an "honest" reliability evaluation for the forecasting exercise undertaken, but is otherwise unsatisfactory. For real dispatches, having half of the risk assessments too unreliable to provide dispatch guidance is a policy disappointment. The large faction of “can’t tells” likely results substantially from the large imbalance in the binary outcome coupled with the need for better predictors. The implications will be further addressed shortly.

There were no empty prediction sets, which is a necessary result from employing nested conformal scores (Kuchibhotla & Berk, 2023). Each nested conformal prediction set will always include at least one outcome class. Consequently, the issue of outlier cases requiring forecasts does not arise as it can for other conformal prediction set methods.

There is nothing special about the coverage probability of .75. It represents an odds of 3 to 1 (.75/.25); the odds are at least 3 to 1 that the true outcome class is included in each prediction set. If one is prepared to accept a lower coverage probability (e.g., .70), the fraction of “can’t tell” prediction sets could be reduced. If one prefers a higher coverage probability (e.g., .80) the “can’t tell” fraction could increase. This is a call to be made by stakeholders, although work in progress may make the choice moot. In any case, there is no formal rationale for automatically employing a common default probability such as .95 or .99.

Results for Other Crime Categories

We applied stochastic gradient boosting to four other broad crime categories. The results for assaults and the results for robberies were qualitatively very similar to the results for weapons offenses. For domestic violence and general disturbances, the number of high risk dispatches was roughly the same as for weapons offenses. However the number of calls for service was about five times larger. This made the lack of balance far more dramatic, and we could not produce any useful results.[32]

Thinking about steps toward implementation, careful consideration must be given to how different sets of crime categories should be determined. That will depend on local stakeholder views and on the local proportions of dispatches that are high risk for police officers. In addition, a lot will rest on making improvements in forecasting accuracy.

The broad crime categories chosen have more than statistical consequences. One can imagine a 911 professional having access to several sets of forecasting algorithms, each for a different broad crime category. With the crime category chosen, the relevant forecasting algorithm could be selected by dispatchers (and others) in an informed manner. A forecast could then be produced in a few seconds.

[32] The claims by some that domestic violence (DV) cases are especially dangerous for police officers may rest on the raw number of DV cases, not their risk per case. In many jurisdictions, more police are injured responding to DV cases than any other broad category of crime. But that might result from a substantially larger number of DV incidents to begin with. The risks case by case might be relatively small.

Discussion

Implementation

Using information that is routinely available, we have shown that interpretable and technically defensible forecasts of risk can be produced for police officers responding to a dispatch. But we make no claim that our procedures are ready for implementation. To begin, in most applications, "can't tell" prediction sets can occur at least once in a while. They indicate that for a given case when a forecast is sought, the classifier cannot make a reliable distinction between different outcome classes. Stakeholders might decide to default such prediction sets to either a high risk or a low risk outcome, depending on whether false alarms are more or less costly than the absence of an alarm when needed. The choice will probably vary across different police departments. It may also be possible to introduce outside information to help inform an otherwise ambiguous forecast. For example, the caller's tone of voice may convey imminent danger. In short, careful thought must still be given to possible responses to equivocal forecasts.

Either for accuracy or policy reasons, the outcome risk variable might be usefully defined in other ways, depending substantially on data availability and quality. For example, risk might be more narrowly defined as an actual injury or fatality if either are, unfortunately, sufficiently common. In other jurisdictions, injuries and fatalities might be so rare that they can perhaps be ignored in the risk definition. In addition, the risk outcomes do not have to be binary. In principle, one might construct a scale of risk with fatalities having the high score, threats of violence having the lowest score, and non-fatal injuries in the middle. None of these options were a practical choice for the data from the CCPD.

There are likely to be useful predictors that were not available from the CCPD dispatches but perhaps available elsewhere. For example, it might be very instructive if the alleged perpetrator's gender, approximate age, and relationship to the caller/victim were elicited. Young males can be especially dangerous (Berk et al., 2009). Were drugs or alcohol involved? Romantic couples in the process of separating can place officers in situations that are emotionally charged (Berk et al., 2016). Articulated threats of violence might be predictive as well. And, there also might be useful predictors that vary by locale such as whether gang hostilities are involved. In short, a richer set of predictors is needed and surely considerable progress can be made.[33]

We appreciate that it can be difficult to obtain fruitful information from some calls for service, and that key information must be obtained quickly. But, we believe that in many police departments it is possible to improve on current practice. For example, when forms are filled out as each call is taken (either by hand or by data entry), those forms might be improved by having boxes to be filled in for the alleged perpetrator's age and gender. Relying solely on an incident narrative that might

[33] There are more speculative possibilities in the not too distant future. For example, Large Language Models now used for the analysis of speech (https://openai.com/blog/chatgpt) might be applied to extract useful predictors from the spoken words of 911 callers. Certain phrases (or perhaps even tone of voice) implying that a caller is fearful of being overheard, for instance, could be a violence precursor.

include such information will typically produce less complete results. Good advice for designing effective data entry forms is readily available (Wiggins et al., 2011).

Finally, there are a host of operational details to be worked out site by site.

1. A feasibility assessment must be mounted.
2. If the estimated feasibility conveys promise, support (or at least not opposition) from the relevant stakeholders must be obtained.
3. Resources must be provided for staff time, IT support, dedicated computer hardware and software, and perhaps outside consultants. The computation costs themselves should be modest. On a modern laptop, the elapsed time to train a risk forecasting algorithm should be less than a minute, and subsequently, forecasts can be produced almost instantaneously.
4. Methods must be provided to transmit relevant information from each call for service to a risk algorithm along with existing information such as the number of past calls in the recent past from the same neighborhood.
5. The forecasting algorithm must be made operational, which should be done with foresight and technical skill in a site-specific manner. Note that there is no need to re-train an algorithm with each 911 call. All that is required is a module from the trained algorithm that does the forecasting and some code to input predictor values and output the forecasts. This form of AI should not be confused with current Large Language Models that power ChatGPT and related software. Currently, these algorithms are so expensive to train and use that even deep-pocket firms like Google and Microsoft are having serious difficulties finding a sustainable business model for them (Oremus, 2023).[34]
6. Some retraining of dispatchers will likely be needed.
7. Ample time should be provided to thoroughly test each step in the implementation.

These requirements may seem daunting, but similar challenges have been successfully overcome in other criminal justice settings (Berk, 2017).

There is also an important methodological message going forward. The use of conformal prediction sets underscores that the reliability of forecasted response variable classes from a confusion table can be very low. For our analysis of police officer risks, the large fraction of "can't tell" prediction sets means that about half of the forecasted outcome classes from a classifier's confusion table could be properly treated as unreliable and difficult to justify for real dispatches. Part of the reason is the very limited dataset given to the classifier we used. But a lot also depends on how the conformal enterprise is tuned. Kuchibhotla and Berk (2023) show that even strong performance by a classifier will produce a large proportion of legitimate "can't tell" results when a very high coverage probability is required. In short, relying solely on a confusion table to forecast outcome classes can be ill advised, or at least insufficient, and conformal prediction tuning should not be perfunctory.

[34] Very recent work indicates that size reductions of various kinds are available potentially making Large Language Models cost-effective and relatively easy to train and apply (Schick & Schütz, 2021). Often there is little loss in accuracy.

The Role of Race

Because of the recent concerns about artificial intelligence and the particular claim of racial bias (Berk et al., 2023), use of almost any risk algorithm by a police department could become very controversial. Yet, our proposed algorithm does not target individuals based on their race. Neither the race of the caller nor the race of the alleged perpetrator are predictors, and we are not proposing that they should be.

The role of neighborhoods is more subtle because of residential segregation based on income, nationality, and race. There can then be, in particular, concerns about “over-policing” (Boehme et al., 2022), sometimes characterized as racial profiling in certain areas (Grogger & Ridgeway, 2006). Within such scenarios, police choose to concentrate their assets in disadvantaged neighborhoods as a top-down process, even if well-intentioned.

People in disadvantaged communities are inordinately victims of crime, and we capitalize on a bottom-up process initiated by concerned neighborhood residents themselves. To some observers, this may look like over-policing. But it is a response to requests for help that arguably are justified by a Hobbesian-like social contract. One consequence of that contract is that police who respond can be put in harm’s way. Better informed dispatches may help to reduce that risk.

It also is important to appreciate that our algorithm only provides a forecast of risk. What responding police officers do with that information can be consequential but should be determined by department policy and police officer training. The risk algorithm is not responsible for either. One might add to the implementation steps listed above that any police department adopting a risk algorithm for police officers should integrate and train for the best police practices that depend on the amount of risk. Also, for some calls, the best response may be deferring to responders who are not police officers. In Brooklyn, New York, for example, the police have from time to time stepped aside and let community members of the Brownsville Safety Alliance respond to 911 calls that do not involve a violent crime in progress (Cramer, 2023). These low risk incidents can account for a large fraction of calls for service. Informal assessments look promising.

Some may wonder why a similar effort could not be made for predicting when citizens are at risk from unnecessary force and/or racially motivated practices by police officers. There are a large number of studies about these matters, such as the path breaking work by Grogger and Ridgeway (2006) and by Ridgeway and MacDonald (2014), but none to our knowledge have been framed explicitly as a forecasting problem. Moreover, there are operational challenges. For example, it is very difficult except in extreme cases to determine when a police apprehension is racially tainted. It is also hard to tell, again, except in extreme cases, when a use of force is unnecessary. There would likely be resistance from police officer organizations as well.

Conclusions

Police officer safety has long been a policy concern. If high risk incidents could be better anticipated, there is the possibility of introducing more effective and more timely safety measures. We have shown that improved forecasting of the risks when responding to calls for service is possible, although our work is a somewhat circumscribed first attempt. The requisite statistical tools are readily available and can be implemented and maintained in police departments with sufficient IT expertise. IT expertise can also be purchased. Perhaps the major obstacle is appropriate, readily available data, in part because calls for service are a very small window of data collection opportunity. Still, the skills of call-takers and dispatchers can be improved along with the ways 911 call data are organized and stored. And some very simple improvements have the promise to move this demonstration of concept toward at least a provisional implemetation.

Acknowledgements Preliminary summaries of the work were presented at following conferences: IACP Officer Health and Wellness Symposium 2021, IACP Technology Conference 2022, American Society of Criminology 2022, The American Society of Evidenced-based Policing in 2021 and 2022, and The National Homeland Security Conference 2023. We received very helpful comments and suggestions at each. Funding for the project was provided by the U.S. Bureau of Justice Assistance.

Authors' contributions B.C. handled day to day administration and contributed in essential ways to all phases of the research. J.C. initiated the project and provided essential oversight for all phases of the research. B.C. provided essential input to all phases of the project, especially through his review of the literature. R.B. provided technical expertise for the data analysis, prepared the figures, and wrote the initial draft of the paper. V.B. was responsible for data management and initial statistical analyses, and provided essential input to all phases of the research.

Funding This project was supported by Cooperative Agreement Number BJA-2018-DP-BX-K015 awarded by the Bureau of Justice Assistance, U.S. Department of Justice.

Availability of data and materials The data belong to the Camden (New Jersey, USA) Police Department, and there are major privacy issues. We are not allowed to share the data provided to us.

Declarations

Competing interests The authors declare no competing interests.

References

Angelopoulos, A.N., & Bates, S. (2022). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.

Berk, R. A. (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. A. (2018). *Machine learning forecasts of risk in criminal justice settings*. New York: Springer.

Berk, R.A. (2020). *Statistical Learning from a regression perspective* (3rd ed.). Springer.

Berk, R. A. (2021). Artificial intelligence, predictive policing, and risk assessment for law enforcement. *Annual Review of Criminology, 4*, 209–237.

Berk, R.A., Sherman, L., Barnes, G., Kurtz, E., & Ahlman, L. (2009). Forecasting murder within a population of probationers and parolees: A high stakes application of statistical learning. *Journal of the Royal Statistical Society (Series A), 172*(part 1), 191–211.

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behavior: A comparative assessment. *Criminology and Public Policy, 12*(3), 515–544.

Berk, R. A., Brown, L., Buja, A., George, E., Zhang, K., & Zhao, L. (2013). Valid post-selection inference. *Annals of Statistics, 41*(2), 802–837.

Berk, R. A., Sorenson, S. B., & Barnes, G. (2016). Forecasting domestic violence: A machine learning approach to help inform arraignment decisions. *Journal of Empirical Legal Studies, 31*(1), 94–115.

Berk, R. A., Kuchibhotla, A. K., & Tchetgen Tchetgen, E. (2023). Fair risk algorithms. *Annual Review of Statistics and Its Applications, 10*, 165–187.

Bierie, D. M. (2017). Assault of police. *Crime NS Delinquency, 63*(8), 899–925.

Bierie, D. M., Detar, P. J., & Craun, S. W. (2016). Firearm violence directed at police. *Crime & Delinquency, 62*(4), 501–524.

Boehme, H., Cann, D., & Isom, D. A. (2022). Citizen perceptions of over- and under-policing: A look at race, ethnicity, and community characteristics. *Crime & Delinquency, 68*(1), 1213–154.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., & Ljung, G.M. (2017). *Time series analysis: Forecasting and control*. Wiley.

Brandl, S. G., & Stroshine, M. S. (2012). The physical hazards of police work revisited. *Police Quarterly, 15*(3), 262–282.

Brandl, S. G., & Stroshine, M. S. (2003). Toward an understanding of the physical hazards of police work. *Police Quarterly, 6*(2), 172–191.

Cramer, M. (2023). What happened when a Brooklyn neighborhood policed itself for five days. *New York Times*. June 4, 2023 (accessed online).

Crifasi, C. K., Pollack, K. M., & Webster, D. W. (2016). Assaults against US law enforcement officers in the line-of-duty: Situational context and predictors of lethality. *Injury Epidemiology, 3*(1), 1–6.

Cunningham, B., Dockstader, J., & Thorkildsen, Z. (2021). *Law enforcement officer safety: Risks,* recommendations, and examples from the field. Arlington: CNA.

Dobbin, K.K., & Simon, R.M. (2011). Optimally splitting cases for training and testing high dimensional classifiers. *BMC Medical Genomics, 4*(31). Published online April 8, 2011.

Freedman, D.A. (2009). *Statistical models: Theory and* practice (revised edition). Cambridge Press.

Fridell, L., Faggiani, D., Taylor, B., Brito, C. S., & Kubu, B. (2009). The impact of agency context, policies, and practices on violence against police. *Journal of Criminal Justice, 37*(6), 542–552.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Gillooly, J. W. (2020). How 911 callers and call-takers impact police encounters with the public: The case of the Henry Louis Gates Jr. arrest. *Criminology and Public Policy, 19*(3), 787–804.

Gillooly, J. W. (2022). 'Lights And Sirens': Variation in 911 call-taker risk appraisal and its effects on police officer perceptions at the scene. *Journal of Policy Analysis and Management, 41*(3), 762–786.

Grogger, J., & Ridgeway, G. (2006). Testing for racial profiling in traffic stops from behind a veil of darkness. *Journal of the American Statistical Association, 101*(475), 878–887.

Gupta, C., Kuchibhotla, A. K., & Ramdas, A. K. (2022). Nested conformal prediction and quantile out-of-bag ensemble methods. *Pattern Recognition, 127*(July 2022), 108496.

Hastie, T., Tibshirani, R., & Friedman, J. (2016). *The elements of statistical learning* (2nd ed.). Springer.

Hine, K. A., Porter, L. E., Westera, N. J., & Alpert, G. P. (2018). The understated ugly side of police-citizen encounters: Situation, suspect, officer, decision-making, and force predictors of officer injuries. *Policing and Society, 28*(6), 665–683.

Kaminski, R. J., & Sorensen, D. W. (1995). A multivariate analysis of individual, situational and environmental factors associated with police assault injuries. *American Journal of Police, 14,* 3–48.

Kearns, M., & Roth, A. (2020). *The ethical algorithm.* Oxford Press.

Kuchibhotla, A. K., & Berk, R. A. (2023). Nested conformal prediction sets for classification with applications to probation data. *Annals of Applied Statistics, 17*(1), 761–785.

Liang, T., & Recht, B. (2021). Interpolating classifiers make few mistakes. arXiv:2101.11815v2 [stat.ML].

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

Maguire, B. J., Hunting, K. L., Smith, G. S., & Levick, N. R. (2002). Occupational fatalities in emergency medical services: a hidden crisis. *Annals of emergency medicine, 40*(6), 625–632.

Manning, P. K. (1988). *Symbolic communication: Signifying calls and the police response* (Vol. 9). MIT press.

McKinnon, K. A., & Simpson, I. R. (2022). How unexpected was the 2021 Pacific Northwest heatwave? *Geophysical Research Letters, 49*, 1–9.

Molnar, C. (2022). *Interpretable machine learning* (2nd ed., licensed under Creative Commons). https://christophm.github.io/interpretable-ml-book/.

Moskos, P. (2009). *Cop in the hood: My year policing Baltimore's* eastern district. Princeton: Princeton University Press.

Mumford, E. A., Liu, W., & Taylor, B. G. (2021). Profiles of US law enforcement officers' physical, psychological, and behavioral health: Results from a nationally representative survey of officers. *Police Quarterly, 24*(3), 357–381.

Neusteter, S. R., Mapolski, M., Khogali, M., & O'Toole, M. (2019). T*he 911 Call Processing System: A Review of the Literature as it Relates to Policing*. Vera Institute of Justice.

Oremus, W. (2023) AI chatbots lose money every time you use them. That is a problem. Washington Post, June 5, 2023.

Peek-Asa, C., Howard, J., Vargas, L., & Kraus, J. F. (1997). Incidence of non-fatal workplace assault injuries determined from employer's reports in California. *Journal of Occupational and Environmental Medicine*, 44–50.

Potts, J., Hawken, A., Hillhouse, M., & Farabee, D. (2022). Virtual reality for law enforcement training: A demonstration and implication for dispatch priming. *Police Practice and Research, 23*(5), 623–632.

Ricciardelli, R. (2018). 'Risk It Out, Risk It Out': Occupational and organizational stresses in rural policing. *Police Quarterly, 21*(4), 415–439.

Ridgeway, G., & MacDonald, J. (2014). Methods for assessing racially biased policing. *Crime and Deliquency, 60*(1), 145–162.

Sampson, R. J., & Groves, W. B. (1989). Community structure and crime: Testing social-disorganization theory. *American Journal of Sociology, 94*(4), 744–802.

Schick, T., & Schütz, H. (2021). It's not just size that matters: Small language models are also few-shot learners. arXiv:2009.07118

Sierra-Arévalo, M., & Nix, J. (2020). Gun victimization in the line of duty: Fatal and nonfatal firearm assaults on police officers in the United States, 2014–2019. *Criminology and Public Policy, 19*(3), 1041–1066.

Sierra-Arévalo, M., Nix, J., & O'Guinn, B. (2022). A national analysis of trauma care proximity and firearm assault survival among US police. *Police Practice and Research, 23*(3), 388–396.

Sierra-Arévalo, M., Nix, J., & Mourtgos, S. M. (2023). The "War on Cops,'' retaliatory violence, and the murder of George Floyd. *Criminology, 61*(3), 385–675.

Taylor, P. L. (2020). Dispatch priming and the police decision to use deadly force. *Police Quarterly, 23*(3), 311–332.

Uchida, C. D., & King, W. R. (2002). Police employee data: Elements and validity. *Justice Research and Policy, 4*(1–2), 11–19.

White, M. D., Dario, L. M., & Shjarback, J. A. (2019). Assessing dangerousness in policing: An analysis of officer deaths in the United States, 1970–2016. *Criminology and Public Policy, 18*(1), 11–35.

Wiggins, A., Newman, G., Stevenson, R.D., & Crowston, K. (2011). Mechanisms for data quality and validation in citizen science. In *2011 IEEE Seventh International Conference on e-Science Workshops*. Stockholm. pp. 14 –19.

Wyner, A. J., Olson, M., Bleich, J., & Mease, D. (2015). Explaining the success of AdaBoost and random forests as interpolating classifiers. *Journal of Machine Learning Research, 18*(1), 1–33.

Xu, Y., & Goodacre, R. (2018). On splitting training and validation set: A comparative study of cross-validation, bootstrap and systematic sampling for estimating the generalization performance of supervised learning. *Journal of Analysis and Testing, 2*, 2249–262.