Journal of Experimental Criminology (2021) 17:15–41
https://doi.org/10.1007/s11292-019-09400-2

The Philadelphia predictive policing experiment

Jerry H. Ratcliffe[1] · Ralph B. Taylor[1] · Amber Perenzin Askey[2] · Kevin Thomas[3] · John Grasso[3] · Kevin J. Bethel[4] · Ryan Fisher[1] · Josh Koehnlein[1]

Published online: 9 January 2020

Abstract

Objectives This place-based, randomized experiment explored the impact of different patrol strategies on violent and property crime in microscale predicted crime areas. The experiment aimed to learn whether different but operationally realistic police responses to crime forecasts, estimated by a predictive policing software program, could reduce crime.
Methods Twenty Philadelphia city districts were randomized to three interventions and one control condition. The three interventions comprised awareness districts (where officers were made aware of predicted areas on roll-call), marked car districts (where a marked patrol police car was dedicated to treatment areas), and unmarked car districts (a plain-clothes vehicle was dedicated to treatment areas). A business-as-usual approach represented the control condition in districts where staff had no access to the predictive software program. Two distinct 3-month phases examined crime outcomes for property and violent crime, respectively.
Results The marked car treatment showed substantial benefits for property crime (31% reduction in expected crime count), as well as temporal diffusion of benefits to the subsequent 8-h period (40% reduction in expected crime count). No other intervention demonstrated meaningful crime reduction. These reductions were probably not substantial enough to impact city or district-wide property crime. Some violent crime results ran contrary to expectations, but this happened in a context of extremely low crime counts in predicted areas. The small grid size areas hampered achieving statistical power.
Conclusions The experiment found reductions in property crime resulting from the marked car focused patrols. It also demonstrated the real-world challenges of estimating and preventing crime in small areas.

Keywords Predictive policing · Randomized controlled trial · Police patrol · Philadelphia · Police

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11292-019-09400-2) contains supplementary material, which is available to authorized users.

✉ Jerry H. Ratcliffe
jhr@temple.edu

Extended author information available on the last page of the article

Introduction

Background

Predictive policing refers to analytical techniques identifying "likely targets for police intervention" (Perry et al. 2013, p. xiii). The National Institute of Justice's first predictive policing symposium in Los Angeles in 2009 identified numerous potential applications of predictive policing, the primary use being the "time and location of future incidence in a crime pattern or series."[1] Another definition more germane to street policing is therefore "the use of historical data to create a spatiotemporal forecast of areas of criminality or crime hot spots that will be the basis for police resource allocation decisions with the expectation that having officers at the proposed place and time will deter or detect criminal activity" (Ratcliffe 2014, p. 4).

The current study uses predictive policing to explore what geographic patrol activities should be pursued in these priority areas. This is not just of relevance to many police departments. It pertains as well to advancing our theoretical understanding of predictive policing and a closely related concept, hot spots policing.[2] Hot spots policing is relevant because it continues to be the case that "focused police resources on crime hot spots provided the strongest collective evidence of police effectiveness that is now available" (National Research Council 2004, p. 250).

Unfortunately, extant research is less clear on what specific activities police officers should be doing in hot spots. Concepts such as the Koper Curve (Koper 1995) or strategies such as focused foot patrol (Ratcliffe et al. 2011) may only be effective for long-term, chronic hot spots and prove ineffectual at addressing dynamic day-to-day crime projections, like those often identified by modern algorithms. If we consider predictive policing strategies are related to hot spots policing, then two recent observations from Weisburd and Telep are relevant: (1) there are numerous strategies that have not yet been rigorously tested, and (2) much more needs to be learned about the impact of new technology on policing effectiveness (Weisburd and Telep 2014).

In consideration of these observations, the current study reports on a randomized, controlled field experiment in America's sixth largest city, Philadelphia, Pennsylvania (PA). After a predictive policing software algorithm (called HunchLab) generated mission grids, the Philadelphia Police Department (PPD) tested three different patrol responses over the course of two different experimental periods, each lasting 3 months. The first experimental period applied the algorithm and patrol responses to property crime, and the second to violence.

Our study reflects the practical reality of many police departments. The policing intervention did not involve swamping predicted areas with massive amounts of resources (saturation patrol) because such an option is unsustainable on an ongoing

[1] http://www.nij.gov/topics/law-enforcement/strategies/predictive-policing/symposium/Pages/technical-breakout.aspx

[2] "Predictive policing…is hot-spot policing, significantly enhanced by technology" *[Bill Bratton, NPR interview (Around the Nation), November 26, 2011].*

basis for all but the largest police departments. Instead, we examined whether crime can be reduced in predicted areas either through greater patrol awareness or with the application of modest additional resources—a dedicated police car.

At present, "there is little experimental evidence from the field demonstrating whether implementing an advanced analytics predictive model, along with a prevention strategy…works to reduce crime" (Saunders et al. 2016, p. 348). The motivating questions we explore are thus, (1) "can we predict crime at the microgeographical level?" and, (2) "if we can predict crime, can we prevent crime with these predictions?"

Where items are applicable, our article follows a CONSORT framework for the reporting of randomized trials (Moher et al. 2010) and is organized as follows. This introduction continues with the theoretical foundation and objectives of the experiment. We then discuss the methods, including the trial design, study setting, outcomes, and randomization process. Then we explain the statistical methods used to determine the intervention impact, followed by a description of the most pertinent experimental results. Inevitably, with an experiment of this size and scope, we lack sufficient space to describe every aspect of the work; therefore, a fuller description of the experimental activity is in an online technical Appendix. A discussion of the intervention's limitations, generalizability, and implications for policing completes the article.

Theoretical foundation

Our study stands at the intersection of two ideas relevant to modern policing: first, to what extent the spatial distribution of crime is predictable; and second, whether policing strategies can act on those predictions and deter criminal activity.

A substantial canon demonstrates that crime is unevenly distributed among places and victims (Felson 1987; Sherman et al. 1989; Weisburd and Eck 2004), and a number of spatial theories of crime can explain short-term changes in crime risk for small areas (Chainey and Ratcliffe 2005). As Perry et al. (2013) point out, the theoretical foundations allowing for crime prediction are ably supported by routine activity theory (Cohen and Felson 1979), the rational choice perspective (Cornish and Clarke 1986), and crime pattern theory (Brantingham and Brantingham 1981–1982)—collectively known as opportunity theories.[3] A pertinent empirical question, however, remains. How fluid are these patterns over the short term?

The underlying theoretical platform for the HunchLab software, like so many spatial prediction algorithms, is the reduced randomness in short-term crime event distributions. This is founded on evidence that offenders "forage" in nearby areas (Johnson et al. 2009b), and this acts in conjunction with a generalized risk heterogeneity model that highlights increased risk in certain areas (Johnson 2010; Tseloni and Pease 2003). Gorr and Lee (2015) have suggested that short-term, temporary hot spots, which may be identified by predictive policing methods, can be a more effective focus for crime prevention than large chronic hot spots and provide for greater equity of crime

[3] Perry et al. (2013) consolidated these approaches into what they called *blended theory*, but environmental criminologists usually refer to these as the *opportunity theories*.

reduction resources. Previous randomized studies examining policing in crime hot spots, however, have relied on a year of crime prior to the experiment (Sherman and Weisburd 1995) or even longer (Ratcliffe et al. 2011; Taylor et al. 2011) to identify hot spot locations. Predictive policing approaches such as risk terrain modeling (Caplan et al. 2011; Kennedy et al. 2011; Moreto et al. 2014) and techniques using short-term event patterns (Johnson et al. 2009a; Mohler et al. 2011) have emerged over the last decade or two with the promise of identifying short-term crime hot spots.

In the current study, we also employ a software program that generates a risk surface for a forthcoming period. We stress, however, that estimating the efficacy of the software program is not the primary goal of this study. Although we report the capacity of the software to predict crime in the control sites, thereby allowing readers to assess the context of the overall study, this is only a subsidiary goal. Instead, the starting assumptions for the study were that (1) predictive software exists, (2) it is capable of performing better than chance or against a uniform spatial distribution of crime, and (3) police departments are as yet unclear as to the most viable strategy to police predicted crime areas. We therefore focus on the policing aspect, and in doing so, our research is a response to numerous calls for more research in this area. For example, Sherman and colleagues' comment of being "unaware of any field experiments conducted with patrol dosage randomly assigned to predictive policing hot spots" (Sherman et al. 2014, p. 108) and "Currently, the policing aspect is mostly overlooked in evaluations of predictive policing" (Rummens et al. 2017, p. 7).

A Campbell Collaboration review reported that 80% of police interventions focused on crime hot spots "reported noteworthy crime and disorder reductions" (Braga et al. 2012, p. 6). Many required significant resource commitments, such as dedicated foot patrols of multiple officers for 16 h a day (Ratcliffe et al. 2011), or sufficient directed patrol resources to "saturate" a crime hot spot (Taylor et al. 2011). This question of "dosage" appears important for crime prevention. Durlauf and Nagin (2011, p. 17) show that certainty and severity of punishment is central to deterrence theory and that "the empirical support for the deterrent effect of certainty is far stronger than for severity." Offender decision-making is likely driven by a subjective assessment of the risk of apprehension, reinforced either by the presence of visible police resources or specific exposure to arrest. Arrest can reinforce "the capacity of the police to deter crime through the threat of apprehension" (Lum and Nagin 2017, p. 344). Focusing police resources in a small area, as can occur with predictive policing, is also relevant. Johnson et al. (2009a, p. 177) determined that a spatial scale between meso and micro is appropriate for patrolling choices and that at this scale "such patterns are overwhelmingly a reflection of the activity of individual offenders."

Choice of policing strategy ties not only to a theoretical foundation but also to operational capacity. Crime analysis in Philadelphia has shown that while there are chains of connected violent crime events, they have a short half-life. This requires any police department to be extremely responsive if a predictive intervention is to be successful (Haberman and Ratcliffe 2012; Ratcliffe and Rengert 2008; Wyant et al. 2012). Haberman and Ratcliffe (2012) concluded that any organization seeking to invest in predictive policing needs a proficient surveillance and analysis mechanism to identify crime patterns, as well as a capable decision-making framework that supports operational flexibility. The latter theme appears in other work (Johnson et al. 2009a). The current study explores this confluence of deterrence and operational response. The application of patrol resources to predicted crime areas during the current study was grounded in a more sustainable and realistic appreciation of the operational resource constraints confronting police chiefs on a day-to-day basis.

Objectives

Can an operationally sustainable police response to predicted crime areas reduce crime? The goal of our study is to investigate the crime reduction link between certain tactics associated with predictive policing, guided by a prediction algorithm, and related concepts of crime control grounded in theories of deterrence. We examine the effects of policing on two different types of high-volume street offenses: violence and property crime. Previous research has shown that street crime (Chainey et al. 2008) and property crime patterns (Gorr et al. 2003) are more predictable than other types of crime.

The experiment tested a number of hypotheses related to the operational implementation of predictive policing.

Hypothesis 1: Greater awareness among general duties patrol officers of the predicted crime areas will measurably deter crime.

It may be that simply making the general duties patrol officers in a police district aware of the predicted crime areas during roll-call will be sufficient to increase the amount of time they spend in the areas, leading to potential offenders being deterred from committing crime during that tour of duty, thereby resulting in measurable crime reduction. During this experiment, the relevant shift officers were made aware of the predicted crime areas during roll-call at the start of their tour of duty.

Hypothesis 2: A dedicated uniform patrol attendance in predicted crime areas will increase visible police presence sufficiently in the local area to deter crime.

The main difference between this and the first hypothesis is in the dosage, focus, and quality of the patrol intervention. In Hypothesis 1, any additional patrol in the predicted areas stems from patrol officers assigned to response policing in the district paying additional attention to the predicted crime areas during the time when they are not assigned to response policing activities. In this hypothesis (2), the awareness model is supplemented by reassigning one marked patrol vehicle to focus specifically on the predicted crime areas to the exclusion of most other police work, increasing treatment dosage.

Hypothesis 3: Dedicated plain-clothes units in unmarked patrol vehicles in predicted crime areas will increase police presence sufficiently in the local area to reduce crime.

Uniform marked police cars are a deliberately obvious police presence linked to a general deterrence mechanism. Rather than deter crime through visible presence (Hypothesis 2), it may be possible to reduce crime through focused surveillance. The use of a (plain-clothes) unmarked police vehicle provides officers with greater opportunity to observe criminal activity surreptitiously, effect an intervention, and theoretically reduce crime through a specific deterrence mechanism.

Hypothesis 4: The previously described interventions will cause temporal crime displacement.

For the current experiment, each predicted crime estimate is valid for an 8-h shift, which leaves up to 16 h before the next experimental treatment period. Because each treatment district reverts to a business-as-usual state after eight treatment hours, it is possible to establish if there is a detectable increase in offenses during a subsequent nonexperimental time block, which would suggest temporal crime displacement (Bowers and Johnson 2003). Alternatively, there might be a reduction in crime, suggestive of a temporal diffusion of benefits from the intervention (Weisburd and Green 1995).

Study setting

With 1.5 million residents, Philadelphia is the sixth largest city in the country. Census population data[4] as of July 2016 estimate that 45% of the city is white, 44% black, and 14% report being Hispanic or Latino. One quarter of the city has a bachelor's degree or higher, and a quarter live in poverty. In the year the study commenced (2015), the city had a homicide rate of 17.8 and a violent crime rate of 1029 per 100,000 population, compared to national rates of 4.9 and 383, respectively.[5]

Research design

Experimental design and participants

The Philadelphia Police Department comprises 22 geographic police districts. One of these districts is the city's international airport. In each iteration (the property crime phase and the violent crime phase), the airport and the lowest crime district for that phase were dropped from the experiment (district 7 for the property crime phase and district 5 for the violent crime phase). The remaining 20 were randomly assigned to one of four experimental conditions using block randomization with a 1:1:1:1 allocation ratio. These conditions were (1) a control condition, with a business-as-usual patrol strategy; (2) an *awareness* condition where officers were made aware of the treatment areas at roll-call and asked to concentrate there when able, congruent with Hypothesis 1; (3) an awareness model treatment enhanced with a dedicated *marked* patrol car and uniformed officers to exclusively patrol the treatment areas consistent with Hypothesis 2; and (4) an awareness model treatment supplemented with dedicated officers and an *unmarked* vehicle to exclusively patrol the treatment areas, as per Hypothesis 3. Areas were patrolled for 8-h-a-day, for two 3-month periods (one for property crime and one for violence).

Prediction algorithm

Although HunchLab was sold to ShotSpotter® in 2018 (and rebranded as ShotSpotter® Missions™), it originated as a web-based predictive policing system

[4] Source: https://www.census.gov/quickfacts/fact/table/philadelphiacountypennsylvania/PST045216
[5] Source: FBI Crime in the US, Table 8 (Offenses known to Law Enforcement by City) and Table 2 ("Crime in the United States, by Community Type, 2015).

designed by Philadelphia-based Azavea, Inc. This section draws extensively on documentation received from Azavea and conversations with Azavea employees.

The crime prediction algorithm (Azavea 2014) uses machine learning techniques that incorporate components of risk terrain modeling (Caplan et al. 2011), near repeats (Ratcliffe and Rengert 2008), collective efficacy (Sampson 2012), and self-exciting point processes (see Mohler et al. 2011). The software can accept a variety of available data sources, based on data availability. During the Philadelphia experiment, data came from the 2013 release of the American Community Survey data, Philadelphia's Open Street Map, the Philadelphia Zoning Authority, and the Philadelphia Police Department. HunchLab can make use of the following variables to generate predicted crime grid cells[6]:

1. Baseline crime levels (historic crime levels measured over the preceding 28, 56, 84, 112, 168, and 364 days).
2. Near repeat patterns (crime trends over the last 3-, 7-, and 14-day periods).
3. Risk terrain modeling (elevation; zoning and water area coverage; density of hospitals, fire departments, transportation points such as bus stops and subway entrances, and schools; distance to the closest bicycle network, hospital, fire department, road, water areas, zoning, and schools).
4. Sociodemographic indicators (median household income, percent households with no income, mean household size, housing density, population density, median rent, percent rented houses, percent vacant houses, density of vehicles, median population age).
5. Guardianship was measured using proximity to and density of nearby police stations.
6. Target availability was reflected in various indicators including housing density, population density, and density of vehicles.
7. Temporal cycles (day of the week, day of the month, week of year, month of the year, police shift).
8. Recurring temporal events (no recurring events such as sporting events or holidays were used during the experiment).
9. Weather (weather variables such as temperature or precipitation were not used during the experiment).

Like a number of other predictive systems, the likely crime locations in HunchLab are defined by grid cells that are 500 ft × 500 ft in size, laid over the city in a fishnet pattern. The algorithm first uses a gradient boosting machine predicting whether a crime will occur in each grid cell. This process relies on several years of historical crime data, which are separated into training versus testing datasets. A portion of the training dataset is fed into a decision tree making a prediction about whether or not a crime will occur in a particular location (grid cell). Next, the data feed another decision

[6] While this list includes the variables that were used as inputs to the models, it is important to note that HunchLab used a different model for each crime type. Log files that identified which variables were relevant in predicting each crime type were no longer available at the time this manuscript was drafted due to the transfer of ownership from Azavea to ShotSpotter; therefore, it is unknown how important each of these variables were in generating the forecast for each crime type.

tree that corrects incorrect predictions from the first model. This process continues several hundred times until the model has "learned" how to correct its prior mistakes.

Cross-validation methods prevent the gradient boosting machine process from overfitting the training data. Once the appropriate number of training iterations is selected, smoothing functions available through generalized additive models calibrate the scores that are created from the gradient boosting machines. The predicted outcomes are converted into expected crime counts for each 500 ft × 500 ft cell. HunchLab then selects the cells with the highest expected counts and displays them as likely crime grid cells on a map, called "mission grids."

In its normal operation mode, HunchLab will identify the mission grids within a jurisdiction; however, for the experiment, we constrained the software in two ways, both of which reduced its accuracy. First, we required the software to forecast three 500 ft × 500 ft mission grids for each day for each police district—regardless of the distribution of the crime problem. This meant that it had to forecast mission grids in low crime police districts, and it could not overpopulate higher crime districts with more than three mission grids. Second, the software introduced a slight randomization component to reduce the possibility that the same mission grids were selected for several days in a row. This was because the HunchLab team had implemented the software in another city and received feedback that continually attending the same location had become boring for officers. To try to reduce the boredom factor and possibly suffer a concomitant loss of treatment fidelity, we agreed to a minor randomization of the highest prediction mission grids within districts.

From mission grids to treatment areas as the unit of analysis

Our original plan was to use the three mission grids in each district as the spatial unit of analysis; however, once we conducted fieldwork (described later in this article and further in the online), it became clear that this did not reflect the policing occurring on the ground. During the experiment, researchers accompanied officers on just over 100 dedicated car assignment patrols (marked and unmarked cars). It was clear from these observations that it was either impossible or unrealistic for officers to patrol the mission grids exclusively. A 500 ft × 500 ft mission grid constitutes not much more than a city block in Philadelphia, given the mean street block length is more than 400 ft. Many of the city's blocks are one-way streets. As a result, officers had to drive through the streets outside of the mission grids in order to police each mission grid. Furthermore, some of the mission grids were contiguous or closely located to other mission grids. For example, in Fig. 1, two mission grids (shaded) are close to one another meaning officers would likely traverse between the mission grids frequently.

The maps presented to the officers showed only the three mission grids. However, because the officers had to navigate the surrounding streets to return to the mission grids, we estimated that any intervention effect would likely also affect the grid cells immediately surrounding the (HunchLab-selected) mission grids. These contiguous cells—first-order queen contiguous cells adjacent to the mission grids—were therefore included in our analysis of the experimental outcomes. These form a more realistic evaluation focus than just the three mission grids per district each day. Because the mission grids were sometimes contiguous or near each other, the contiguous cells sometimes overlapped. In other words, a contiguous cell for one mission grid could

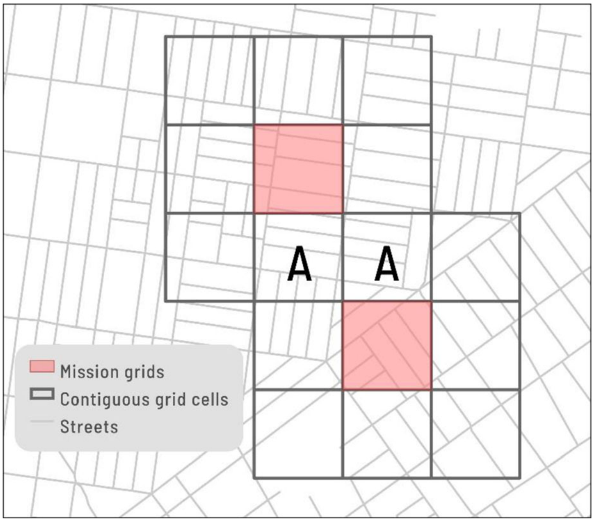

Fig. 1 Example street layout with two mission grids and contiguous grid cells surrounding each mission grid. Cells contiguous to both mission grids are labeled A

also be a contiguous cell for another mission grid. This can be seen in Fig. 1 where two contiguous cells (labeled A) are adjacent to *both* mission grids in the picture.

Because of this inability to distinguish contiguous areas that were unique to specific mission grids, the unit of analysis in this experiment is the combination of all three mission grids in each district on each day, as well as their respective contiguous cells. We term this combined spatial unit the "treatment area." Because each day the experiment operated for one 8-h shift, we can explain the unit of analysis thus: Each day, within each district assigned to a treatment condition within the experiment, the district's "treatment area" comprises the three mission grids as well as the contiguous cells adjacent to the mission grids for the relevant 8-h period. There is therefore one treatment area per intervention district per day. Given low crime counts and associated concerns, outlined in the online Appendix, we aggregated up to weekly crime counts for each district. Crime counts are therefore organized into treatment area by week within districts. With 14 weeks and 20 districts, we have 280 "district–weeks" per experimental phase as the unit of analysis used in this study.

Algorithm accuracy

Although (as we stated earlier) estimating the efficacy of the software program is not the primary goal of this study, we are able to report some indication of the software's capacity to identify appropriate treatment areas in the districts. We report the prediction accuracy index (PAI), a ratio measure comprising a numerator of the proportion of crime events falling within the mission grids for each district and a denominator of the percentage of the relevant police district covered by the mission grids (Chainey et al. 2008; Eck et al. 2017; Flaxman et al. 2019; Tompson and Townsley 2010). Note the following two features about our deployment of the PAI. First, while Drawve (2016) demonstrated that the PAI is more suited to comparisons between techniques or parameter choices, we report PAI mean and standard deviations for the five control districts across the property crime phases. We do not report these values for treatment districts due to possible contamination resulting from operational activities. Second, although we use treatment areas for much of the analysis in this article, we report the PAI values for mission grids only. This is a more accurate indication of the predictive algorithm's abilities and more reflective of the task it was asked to perform.

As per Chainey et al. (2008), the PAI is calculated thus:

$$\mathrm{PAI} = \frac{\left(\frac{n}{N}\right)*100}{\left(\frac{a}{A}\right)*100} \tag{1}$$

Where n is the number of crimes that occurred in the mission grids, N is the total number of crimes in the district on each given day, a is the area of the mission grids, and A is the total area of the police district under examination. We also report the prediction efficiency index (PEI) which indicates how well a forecasting algorithm performs compared to the best it could have done, given the available crime and distribution (Lee and Eck 2017), as shown:

$$\mathrm{PEI} = \frac{\frac{n}{N} / \frac{a}{A}}{\frac{n^*}{N} / \frac{a}{A}} \tag{2}$$

Where n^* is the maximum number of crimes within the district that are forecastable within the area predicted ("a," or in our case, three mission grids), and other variables are as per Eq. (1).

Interventions

In "awareness" districts, officers were informed at roll-call of the mission grids for that shift and were asked to pay attention to the areas when they were able, within the constraints of their operational duties (consistent with Hypothesis 1). Perry et al. (2013) note with specific regard to predictive policing that "*situational awareness* among officers and staff is a critical part of any intervention plan" (p. xviii, emphasis in original). We did not impose any requirement to pay additional attention to the mission grids, nor did we dictate any order of attention or other constraints.

In "marked" car districts, in addition to having officers at roll-call aware of the mission grids, two (though occasionally one) officers were assigned to a single marked vehicle creating a dedicated patrol for the mission grids (Hypothesis 2). This patrol was not assigned calls for service or other operational roles. In "unmarked" car districts, the operational implementation was the same as in marked car districts with the exception that in these districts the officers always used an unmarked vehicle and were often dressed in plain clothes as well (Hypothesis 3).

For the marked and unmarked car districts, officers were instructed to remain in the three mission grids as much as possible and to move between them as necessary during the tour of duty. The instructions to officers did not dictate any particular order of attendance to the three mission grids or specify for how long or how frequently officers attended each mission grid. The officers were instructed to concentrate on the mission grids, to patrol each of the mission grids throughout the shift, and to respond to calls for service in their mission grids. They were permitted to leave this patrol pattern to backup other officers in the district if they asked for assistance. Because the city's police cars were not fitted with automatic vehicle location (AVL) software or devices at the time of the experiment, we are unable to measure the treatment fidelity precisely; however, as explained in a following section, we are able to provide an estimate of fidelity based on field observations.

In the five "control" districts, personnel did not have access to the crime prediction software and continued policing as usual.

The property crime phase ran for 90 days from June 1, 2015, through August 25, 2015, 8 a.m. to 4 p.m. daily. A break was scheduled between the property and violent crime phases of the experiment for two reasons: (1) to allow the police department to prepare for, and recover from, the visit of Pope Francis to Philadelphia in late September 2015; and (2) to allow crime patterns to return to normalcy for the districts. This was estimated to be more than necessary given the short half-life of successes associated with hot spot policing interventions (for example, see Sorg et al. 2013). The study resumed on November 1, 2015, for the violent crime phase of the experiment which ran for 92 days through January 31, 2016, 6 p.m. to 2 a.m. daily. District assignments were re-randomized for the violence phase. Both phases spanned 14 weeks.

We disseminated information regarding the experiment as follows: Deputy Police Commissioner Kevin Bethel (an author on this article and responsible for patrol policing at the time) held a command briefing (that also featured the lead academic researcher and primary author of this article) for all available mid-level and senior district-level leaders. Instructional information also was sent to each district. Two short instructional videos were made available to both crime analysts responsible for accessing the crime predictions and personnel assigned to marked and unmarked cars. Finally, the research and analysis section at police headquarters assigned a senior analyst to the project, and he spent considerable time on outreach and project education.

Outcomes and data collection

The research design incorporates various strands. As stated earlier, we report on metrics examining the software predictive efficacy. As a generalized indication of treatment fidelity, we also chart aggregate time usage of the dedicated police cars based on trained observers conducting 101 ride-alongs in the back of cars assigned to either the marked or unmarked car condition. Details of the training and data collection for the field observations are included in the online Appendix. Finally, we report outcome results for property crime and violent crime.

All crime data came from the police department's geolocated incident database. For the property crime phase, the following UCR part 1 property crime counts were included in the property crime analysis: residential and commercial burglary, theft from, auto and stolen vehicles. For the violent crime phase, the following UCR part 1 crimes were included: homicide, rape, robbery, and aggravated assault.

Randomization

Blocked experimental designs have been around for roughly a century and are generally credited to statistician R. A. Fisher (Box 1980; Fisher 1935). Criminal justice researchers using spatial units of analysis—like police districts, drug markets, or hot spots—have been drawn to block randomization for two reasons. First, it elevates the likelihood that random assignment, even with small numbers of spatial units, will achieve initial probabilistic equivalence across treatment groups, or across treatment and control groups. Second, block randomized designs may improve statistical power (Gill and Weisburd 2013). We blocked on a weighted count of three variables: SES index, percent population white non-Hispanic, and crime gravity score.

We created the SES index from four census variables reflecting households reporting income less than US$20,000 (reversed), households reporting income greater than US$50,000, median house value, and median household income. The online Appendix has more information about this process.

Crime gravity scores are a measure of crime harm. Constructing crime harm levels involves, in part, weighting crime occurrences by seriousness (Ignatans and Pease 2016; Ratcliffe 2015a, b; Sherman 2007; Sherman et al. 2016). In our case, every crime incident over a 3-year period was weighted by a gravity score derived from the Pennsylvania Sentencing Guidelines. We weighted recent crime more heavily than older crime events. To create one stratification variable, each district's ranks on the relevant variables (SES index [reversed], percent population white non-Hispanic [reversed], crime gravity) were summed, counting the crime gravity rank twice for greater prominence. Districts were then ordered by this (reversed) summed rank variable. This was conducted independently for property crime and violent crime.

Along with the airport district, the district for each experiment phase with the highest summed rank (on a reversed seriousness scale) was excluded. The remaining ordered 20 districts were separated into five blocks of four districts each. This resulted in one randomly selected awareness, marked, unmarked, and control district within each block of four districts (again, see the online Appendix for more details).

Concealment and blinding

The prediction software is web-based and required a login and password. This permitted electronically blocking the control districts in each experimental phase from accessing the software platform. We were not able to conceal the existence of the experiment across the police department. Indeed, to achieve buy-in, it had been widely discussed; however, only a few individuals had access to the logins for each district and there is no evidence of control district personnel attempting to access crime predictions for their district.

Field observations

A total of eight graduate research assistants and primary investigators were deployed as field researchers to collect observations during selected ride-alongs with officers to monitor treatment integrity. The observation shifts were determined via random assignment for both the property crime and violent crime phases of the experiment. Data captured by field researchers helped to produce a more holistic view of the implementation challenges faced by the department (Ratcliffe et al. in press), echoing Greene's (2014) call to explore the process concerns and contextual dynamics of innovation in policing. Field observers completed a one-page field observation form during each ride-along, noting where the officers spent the majority of their time for 32 × 15-min blocks during the 8-h experimental treatment phases. For each 15-min period, the observer noted the dominant activity of the assigned vehicle from one of eight possible conditions: at station, car outside grids, car inside grids, in grid talkingwith citizens, on break, incident within the grid, incident responded to outside the grid, and other (see the Appendix for the form used). Prior to the beginning of the experiment, observers discussed potential questions and themes relevant to the ride-alongs at a training session. We conducted 79 observations for the property crime phase and 22 for the violent crime phase. The reason for this imbalance is explained in the online Appendix.

Describing daily crime patterns

Before moving to statistical models of district–week crime occurrence in treatment areas, we describe daily district crime patterns in each of the control districts.

As explained earlier, we examined outcomes for treatment areas per district per day (a reminder that a treatment area is the three mission grids plus their contiguous cells). Even when we enhance the mission grids to include the contiguous cells into a combined treatment area for a police district, we are still dealing with small geographic areas in an 8-h shift. As such, there were many days with either no crimes reported in the treatment area during the shift, or no crimes reported in the entire district during a shift. This constraint particularly hampered the violent crime phase.

As can be seen in Table 1 for the property crime phase, a control district only had a property crime during the 8-h experimental time period about 60% of the time. About 40% of the time, there were no offenses for the software to predict, even though the experimental design required the software to predict three mission grids per district per day. On average, the treatment area predicted was about 7% of the total area of the district, though with considerable variation. Note that in the district labeled D, the treatment area averaged about 13% of the district, but in district E, the treatment area was on average less than 3%. Overall, the software—which was constrained to meet the terms of the experiment—predicted about 14% of the possible crimes, using an area of about 7% of the districts. Note that in the block randomization process, districts with

Table 1 Crime during the 8-h treatment shift for control districts during the property crime phase

District	Days with crime in the district (max. possible = 90)	Total crimes in district	Crimes within treatment area	Mean area of district for treatment area (%)
A	61 (67.7%)	127	11 (8.6%)	5.1
B	46 (51.1%)	64	12 (18.7%)	6.1
C	74 (82.2%)	157	22 (14%)	5.7
D	59 (65.5%)	92	15 (16.3%)	12.9
E	34 (37.7%)	49	8 (16.3%)	2.6
Total	274 (60.1%)	489	68 (13.9%)	6.8

Table 2 Crime during the 8-h treatment shift for control districts during violent crime phase

District	Days with crime in the district (max. possible = 92)	Total crimes in district	Crimes within treatment area	Mean area of district for treatment area (%)
F	49 (53.2%)	68	7 (10.3%)	5.8
G	32 (24.4%)	40	4 (10%)	5.9
H	19 (20.6%)	20	4 (20%)	7.4
I	27 (29.3%)	37	8 (21.6%)	2.6
J	10 (10.8%)	11	1 (9.1%)	1.4
Total	137 (29.8%)	176	24 (13.6%)	5.1

generally more crime are A in the property phase and F in the violence phase, and lower crime districts are E and J.

For the violent crime phase, the counts are all lower (Table 2). A district could only expect to experience a violent crime during the experimental 8-h window about 30% of the time. The software predicted about 14% of these, using an area of about 5% of a district.

Intention to treat

Because crimes reported during the entire 8-h shift are included in the treatment period, even though officers were not always actively patrolling during the entire time (see Fig. 2 and Fig. 3), the approach here is tantamount to an "intention-to-treat" analysis. This is the best way to "obtain an unbiased estimate of the effect of selecting one treatment over another" (Detry and Lewis 2014, p. 85) in a randomized control trial. Should a police department at some point wish to instigate a predictive policing approach, it is more reflective of the operational realities they are likely to encounter.

Statistical method

We used mixed effects models, with weeks (n = 14) nested within districts (n = 20), estimating full Bayesian models with Markov chain Monte Carlo (MCMC) estimation (Gelman et al. 2003; Gilks et al. 1996). The case for using full Bayesian estimation is based on working with population data and our interest in what happened with these specific data. We are not interested in treating the data as a sample and making inferences back to unobserved population data. Confidence intervals around parameters obtained from a non-Bayesian statistical model would be making that type of inference; such resultant confidence intervals would not apply here given the data examined. The Bayesian assumptions, by contrast, capture exactly what we seek. With these data and these assumed prior distributions, these models reveal the best guess for these specific impacts (b weights and the 2.5th percentile and 97.5th percentile of those estimates). To put this more technically: "A primary motivation for believing Bayesian thinking important is that it facilitates a common-sense interpretation of statistical conclusions. For instance, a Bayesian (probability) interval for an unknown quantity of interest can

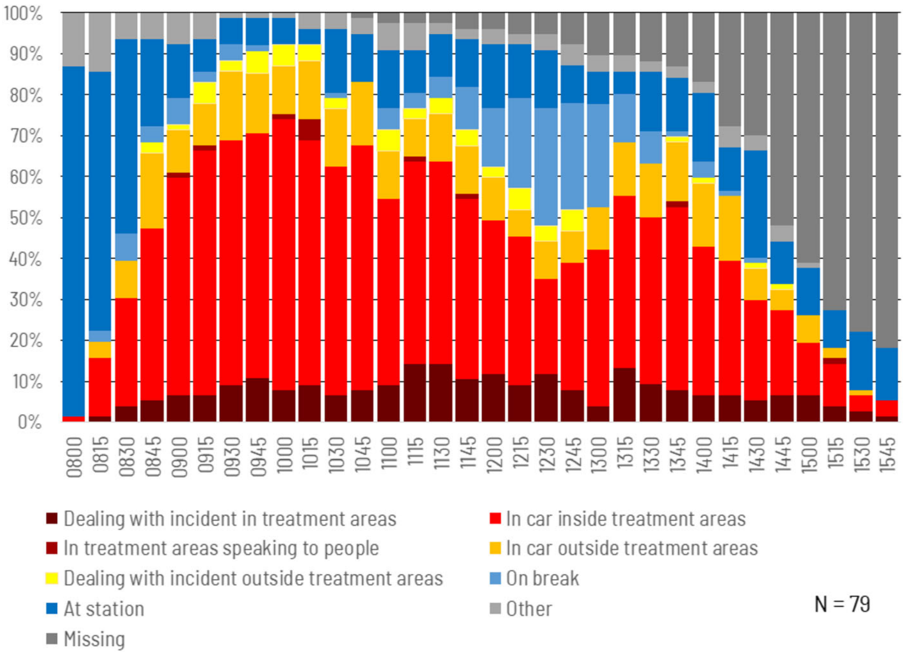

Fig. 2 Main car activity for 15-min blocks for marked and unmarked cars in the property crime experimental phase (8 a.m. to 4 p.m., *n* = 79)

be directly regarded as having a high probability of containing the unknown quantity, in contrast to a frequentist (confidence) interval, which may strictly be interpreted only in relation to a sequence of similar inferences that might be made in repeated practice" (Gelman et al. 2003, pp. 3–4).

Within this statistical approach, models were estimated with MLwiN (Browne 2012). We ran logit models for a *crime presence* variable (crime presence vs. absence of crime) and negative binomial models for a *crime counts* variable. For all models, MCMC estimation was conducted with chains ranging from 50,000 to 200,000. We only report results from estimation with the largest number of chains estimated, but stability was observed throughout.

The exposure variable was the total number of grid cells in all of the treatment areas across the entire week. For the property crime experiment, they ranged from 40 to 189 (mean 133.33). For the violent crime experiment, they ranged from 39 to 204 (mean 134.05). Exposure was modeled in various ways with comparable results, so we report the results that are easiest to understand with the mean-centered grid cell count as a predictor. We first estimated a null model, with random intercepts at the district level and the exposure variable. A second model followed with the dummy treatment predictors added to the model with the exposure variable. These three separate binary dummy variables represented each of the three treatment conditions—awareness, unmarked patrol, and marked patrol. District–weeks in the control districts represented the reference string. We examined various diagnostics, reported in the online Appendix.

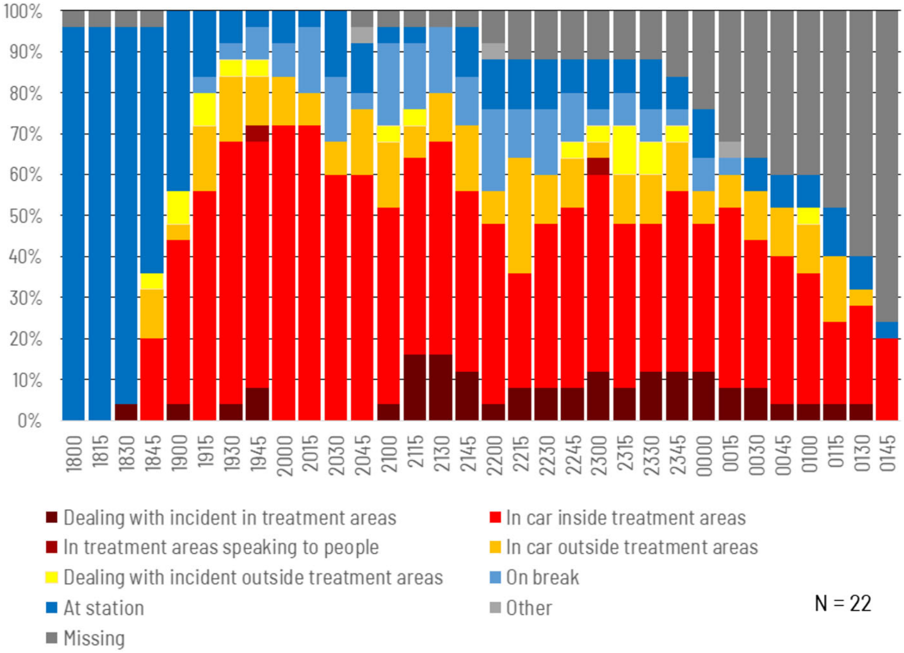

Fig. 3 Main car activity for 15-min blocks for marked and unmarked cars in the violent crime experimental phase (6 p.m. to 2 a.m., $n = 22$)

Results

Software efficacy

Given the operational requirement to predict into an 8-h window, even in low crime districts, it is unsurprising that many control districts did not experience a crime event during the experimental shift (8 a.m. to 4 p.m. for property crime and 6 p.m. to 2 a.m. for violence) as shown in Tables 1 and 2. Since the PAI uses the total crime during the shift as a denominator, we were unable to calculate the PAI in these zero-crime instances. For the calculations, we therefore exclude days when a district did not experience a crime. Furthermore, the district-level violent crime counts were too low to provide reliable measures across all of the districts (as seen in Table 2, and recall that the table shows values for entire treatment areas, not just the mission grids). We therefore report the PAI and PEI mean and standard deviation for the 90-day property crime phase for the five control district mission grids[7] at the mission grid level in Table 3.

Implementation fidelity

In total, there were 101 individual field note entries completed through both phases of the experiment, with 79 conducted with the property crime phase, and the remaining 22

[7] At the request of the police department, we have anonymized the district location information. Therefore, A–E may, or may not, include districts also identified as F–J.

Table 3 PAI mean and standard deviation for the control district mission grids during the property crime phase

District	Mean PAI	PAI std. dev.	Mean PEI	PEI std. dev.
A	12.01	30.11	0.18	0.43
B	17.98	34.12	0.25	0.43
C	18.30	35.50	0.33	0.63
D	8.19	16.40	0.27	0.50
E	25.21	51.37	0.24	0.43

with the violent crime phase (more information in the online Appendix). The following aggregate charts represent all observed shifts for both marked and unmarked cars. As can be seen in Fig. 2 and Fig. 3, officers patrolled the treatment areas to varying levels throughout the shift, with officers getting to the treatment areas earlier for the 8 a.m. to 4 p.m. property crime phases, but officers patrolling treatment areas more extensively later in the shift for the violent crime phase (6 p.m. to 2 a.m.). At least 50% saturation of treatment areas was achieved for 3.5 and 3.75 h, respectively, for property and violent crime phases.

Property crime phase results

Prior to adding predictors, we found that the treatment areas in a typical district in a typical week would experience one or more property crimes about 60% of the time (Table 1 and see the online Appendix). Table 4 shows the impacts of the three treatment conditions in contrast with the control condition.

The marked patrol car condition noticeably lowered the chances of a property crime occurring during the treatment shift. In a marked condition district with a typically-sized treatment area, the expected odds of one or more reported property crimes occurring during a shift versus not occurring was 55% lower (mean OR = 0.45) compared to the control district–weeks. This translated to an expected 49.5% of district–weeks with a typical count of cells in the treatment area having one or more reported property crimes in this condition[8]. This represents an expected 27.9% reduction in the presence of property crime during treatment shifts in marked condition districts relative to control districts with typically-sized treatment areas. Despite this sizable benefit, the coefficient was not statistically significant (two-tailed $p = 0.17$).

The two other treatment conditions had smaller effects. In the typical awareness district (compared to the control condition), the expected odds of any property crime occurring were about 7% lower relative to typical control condition district–weeks. In the unmarked condition, the corresponding expected odds were about 2% higher.

These "crime versus not crime" (crime presence) results were mimicked when we explored the effect of the treatment on specific crime counts. Table 5 shows

[8] The expected odds = exp((log odds for control district (constant)) + (coefficient for marked condition)), then calculating the expected proportion. The expected proportion is derived as [odds ratio / (1 + odds ratio)]

Table 4 Property crime presence/absence during treatment shift: treatment condition logit model MCMC results

					Mean *b* weight		Mean OR	
Fixed portion of model	Mean OR	Mean *b*	SD of *b*	*p*	95% LCL	95% UCL	95% LCL	95% UCL
Constant (control)	2.185	0.782	0.614	0.093	− 0.427	2.008	0.652	7.449
N of cells (centered)	1.017	0.017	0.005	< 0.001	0.008	0.027	1.008	1.027
Awareness	0.928	− 0.075	0.864	0.459	− 1.798	1.666	0.166	5.292
Unmarked	1.025	0.025	0.892	0.489	− 1.759	1.796	0.172	6.024
Marked	0.448	− 0.802	0.862	0.166	− 2.558	0.887	0.077	2.429
Random portion of model		Mean	SD		95% LCL	95% UCL		
District level variation		1.430	0.871		0.382	3.632		

N = 280 district–weeks

the property–crime count model with the treatment conditions included. In the marked districts as compared to the control districts, expected property crime counts in the treatment areas were 31% lower (OR = 0.69). Although sizable, this treatment benefit did not attain conventional levels of statistical significance (two-tailed $p = 0.208$).

Violent crime phase results

Compared to property crime in Philadelphia, violent crime proves an even rarer event at the microgrid scale. Prior to adding predictors, we found that the treatment area in a typical district in a typical week would experience one or more violent crimes only about 29% of the time (Table 2 and see the online Appendix). Table 6 shows the impacts of the three treatment conditions in contrast with the control condition.

Table 5 Property crime count during treatment shift: treatment condition MCMC results

					Mean b weight		Mean IRR	
Fixed portion of model	Mean IRR	Mean *b*	SD of *b*	*p*	95% LCL	95% UCL	95% LCL	95% UCL
Constant (control)	0.933	− 0.069	0.331	0.411	− 0.731	0.603	0.482	1.828
N of cells (centered)	1.010	0.010	0.002	< 0.001	0.006	0.015	1.006	1.015
Awareness	1.112	0.106	0.461	0.398	− 0.838	1.012	0.432	2.752
Unmarked	1.128	0.121	0.476	0.388	− 0.854	1.044	0.426	2.841
Marked	0.694	− 0.366	0.470	0.208	− 1.303	0.574	0.272	1.775
Random portion of model		Mean	SD					
District level variation		0.460	0.246		0.161	1.088		
Overdispersion		0.036	0.040		0.002	0.148		

N = 280 district–weeks

Table 6 Violent crime presence/absence during treatment shift: treatment condition MCMC results

					Mean b		Mean OR	
Fixed portion of model	Mean OR	Mean *b*	SD of *b*	*p*	95% LCL	95% UCL	95% LCL	95% UCL
Constant (control)	0.327	−1.118	0.427	0.005	−2.005	−0.316	0.135	0.729
N of cells (centered)	1.008	0.008	0.005	0.048	−0.001	0.018	0.999	1.019
Awareness	1.624	0.485	0.591	0.191	−0.651	1.701	0.522	5.481
Unmarked	0.952	−0.050	0.609	0.468	−1.284	1.139	0.277	3.124
Marked	1.404	0.339	0.589	0.267	−0.839	1.508	0.432	4.519
Random portion of model		Mean	SD					
District level variation		0.467	0.444		0.003	1.598		

N = 280 district–weeks

Relative to the control districts, the chances of violent crime in a treatment area were somewhat higher in the marked districts (expected 31.5% of district weeks with one or more violent crimes, $p = 0.267$) and in the awareness districts (expected 34.7% of district weeks with one or more violent crimes, $p = 0.191$). The chances were slightly lower in the unmarked districts (expected 23.7% with one or more violent crimes, $p = 0.468$).

We see similar patterns when examining the crime count models rather than the "crime present versus no crime present" model. Relative to control districts, expected counts were nonsignificantly higher in the awareness districts (IRR = 1.56; two-tailed $p = 0.15$) and in the marked districts (IRR equals 1.24; two-tailed $p = 0.31$). Expected counts in unmarked districts were closely comparable to the control districts (IRR = 1.04). The full violent crime count models results appear in the online Appendix.

Posttreatment shift effects

Per Hypothesis 4, we did examine the potential temporal displacement and temporal diffusion of benefits in the 8-h period after a treatment shift. To manage the size of this article, we report these results fully in the online Appendix and provide only a summary here. For property crime, we did find a temporally lagged benefit of the marked car treatment condition, resulting in an expected 39.7% reduction in posttreatment shifts with one or more reported property crimes. Again, however, this benefit did not achieve conventional levels of statistical significance (two-tailed $p = 0.072$). Other treatment conditions saw smaller effects. We caution that property crime was generally lower in the 8 h (4 p.m.–midnight) after the treatment shift (8 a.m.–4 p.m.).

Violent crime was scarce during the 8 h in the shift after the treatment window (from 2 a.m. to 10 a.m.). The expected violent crime count in treatment areas was 0.17 per district week across all 20 districts. Expected counts increased nonsignificantly in the awareness and marked car districts by upwards of 70% and in the unmarked car districts by 23%. Nevertheless, while these percentage changes can seem impressive, they related to minute amounts of crime, so we therefore caution that any differences can appear exaggerated (see the online Appendix for more details).

Discussion

Predicting crime to an 8-h window in a few small mission grids is inherently challenging, especially when focusing on violent crime. At this point, given a grid cell is little more than a single city block, we may be operating below the current limit of analytical capacity, data accuracy, data availability, and human stochastics. Over the course of a week, the average district would have a 63% chance of there being one or more property crimes in the treatment area during treatment shifts, but only about a 29% chance of one or more violent offenses in those locations for those shifts. We caution that at these volumes, any differences tend to appear exaggerated when expressed as percentage changes (as we do in this article). The floor effects (see the online Appendix) tend to explain the absence of statistically significant results generally.

With these caveats in mind, the substantive (not statistical) results showed a 31% reduction in expected reported crime counts when marked cars were deployed to property crime districts (28% reduction in presence when examining crime presence of crime versus absence). In the 8 h *after* the property phase shift, we found a 40% reduction in expected property crime occurrence (46% with count models) from 4 p.m. to midnight. The other experimental conditions did not demonstrate any remarkable results.

For property crime, we therefore do not find evidence to support Hypothesis 1: *Greater awareness among general duties patrol officers of the predicted crime areas will measurably deter crime*. However, we did find some modest evidence to support Hypothesis 2: *A dedicated uniform patrol attendance in predictive areas will increase visible police presence sufficiently in the local area to deter crime.* We stress *modest* in terms of evidence, because a combination of either floor effects or weak experimental effects meant this change was not statistically significant.

Because we did not find a meaningful effect for the unmarked condition, we reject, for the property phase, Hypothesis 3: *Dedicated plain-clothes units in unmarked patrol vehicles will increase police presence sufficiently to reduce crime*. Finally, Hypothesis 4 predicted that *the previously described interventions will cause temporal displacement*. In the 8 h after the property treatment, the marked car districts were associated with a diffusion of benefits, such that expected crime counts were 40–46% lower depending on whether the logit or count models were considered.

Violent crime findings were hampered by low volumes overall and a general scarcity of violence in the treatment areas during the study frame (which may come as a welcome surprise to city residents). Descriptively speaking, relative to control districts, increased crime counts were experienced in awareness districts (56%) and marked car districts (24%) and little change in unmarked car districts (4% increase). In the 8 h posttreatment, we find evidence of nonsignificant increased crime counts in all three treatment conditions (see the online Appendix). We do not find support for any of the first three hypotheses, and there is no apparent theoretical explanation for these increases. Again, these changes should be interpreted with respect to the extremely low underlying raw counts and as such should be interpreted with considerable caution.

Limitations

Predicting crime and the tactics used in the subsequent predicted areas are arguably two distinct considerations. This experiment was not designed to explicitly evaluate the first consideration; however, we do report a number of metrics for HunchLab as a crime prediction tool. Different results would likely be observed using different software or using HunchLab with different parameters. Furthermore, it is almost certainly the case that the software will perform more effectively when released from the constraints imposed by our experiment.

The intervention is evaluated here "as intended" and we present intention-to-treat analyses. Inevitably, operational constraints prevented the police department from achieving a complete implementation of the experimental treatments. Sometimes in the car districts, officers made arrests and the police district was unable to spare personnel to replace the operational crew for the remainder of the shift. Sometimes, a vehicle broke down and a replacement was not available. If this is seen as a constraint, we would argue that it is a constraint that reflects the reality of the policing environment the world over.

We also note that property crime suffers from reporting inaccuracies since victims are usually not present during the crime event. While techniques such as aoristic analysis (Ratcliffe 2000; Ratcliffe and McCullagh 1998) can help, limitations with Philadelphia Police data reporting precluded using this approach here. In lieu of this, we were limited to using the time that the crime was reported to police.

Discussions with police personnel during field observations did raise the possibility that some of the unmarked police cars got "burned." That is, the community became aware that unmarked cars were police vehicles. Police officers suspected this was the case due to the frequency of the patrols in the treatment areas as well as the make and model of the cars commonly used as unmarked vehicles by the police department. If sufficient people became aware that unmarked vehicles policed the treatment areas, this information would likely transmit to potential offenders. As a result, this would negate any *specific* deterrence effects; however, it would have a concomitant effect of increasing *general* deterrence. The extent of this potential (though unanticipated) treatment condition change is difficult to estimate. Nevertheless, in light of the experimental constraint of driving around a small number of streets for extended periods of time, it certainly cannot be ruled out. Given the notably different property crime impacts for the marked as compared to the unmarked condition, two possible conclusions exist. Either the cars were not "burned" and the specific deterrence mechanism of unmarked cars was insufficient to affect the crime rate, or the subsequent general deterrence mechanism from burned cars was ineffective. Finally, a combination of burned and not burned may have occurred throughout the experiment lessening any overall impact.

Perhaps one of the biggest limitations arises from crime infrequency at extremely small spatial and temporal scales. The attendant constraints are sketched in the online Appendix and elaborated elsewhere (Taylor and Ratcliffe 2019). Whether this adverse statistical power impact of the geographic fine-tuning accompanying current predictive policing efforts is inevitable, or can be counterbalanced in other ways, proves an interesting question (Taylor and Ratcliffe 2019). Researchers advocating microscale interventions would do well to attend to these concerns.

Costs

Most vehicle-bound police personnel involved in the experiment were released from general duties police units, as were the marked cars themselves. The unmarked vehicles were normally assigned to covert units such as drug squad officers. As a result, the police department incurred no direct financial costs. This does however represent a realignment of one or two officers within each district during the specific 8-h shifts of the experimental phases and a repurposing of the unmarked cars. As a result, some reallocation of workload within the district will have been experienced. This may have manifested as increased response times to calls for service outside of treatment areas. It may have also affected crime in other areas of the district, for example due to the removal of a vehicle from covert narcotics work. Future work may be able to explore this issue further. The police department also assigned a centralized analyst (one of the authors) to the experiment for much of his assigned work time. This will have affected other crime and intelligence analysis activities.

Generalizability

As Weisburd and colleagues note, “police and scholars must ask not just ‘what the strategy is’ that will be effective but also ‘how we gain enough dosage’ for such strategies to have large-scale crime prevention impacts across the city” (Weisburd et al. 2015, p. 384). This experiment was designed in close collaboration with the PPD so that the intervention conditions reflected realistic treatments that the police department could sustain indefinitely. While external validity is ultimately an empirical question (Taylor 1994, p. 164), our ad hoc discussions with police personnel from other departments suggest that our experimental conditions reflect the likely manner by which other police departments would implement a long-term predictive policing strategy.

Interpretation

If predictive policing is, in essence, an application of microlevel hot spots policing, then there is a need to consider explicitly the strategic translation of a place-specific microlevel policing tactic up to a district-wide or city-level implementation (Sherman et al. 2014). Reliably predicting crime for small grids comprising little more than a single street block remains a monumental challenge for the nascent discipline of predictive crime analysis; further, the vagaries of stochastic and dynamic microlocations can play havoc with even the best predictions. As a result, we chose to expand the analytical areas to more closely reflect the realities on the ground. Even when expanding to treatment areas, we experienced floor effects, especially in terms of violent crime counts. As such, we caution against interpreting the results purely in terms of patterns of statistical significance.

One reviewer reasonably asked why we bother reporting the violent crime results at all. Although the low violence counts likely negated any chance of discovering meaningful results, the policing field remains focused on tackling violent crime. We felt there was value in exploring the violence question, even if we are only able to highlight the significant challenges in both predicting and affecting this societal issue. The odds ratio changes with regard to the property crime phase and the marked car units are a little more substantial, again with the caveat that the numbers remain relatively small.

Overall, the quantitative results suggest that a dedicated and focused marked police car working in a small area may reduce property crime by a substantial fraction; however, participants and the community need to manage expectations about the volume of crime prevented. Small microgrids will never generate a massive volume of crime in a short time frame like an 8-h window; therefore, the raw numbers will remain small. Police commanders should interpret these results carefully before deciding if any costs in terms of using dedicated resources are offset by the potentially limited gains to be earned from the sort of predictive policing strategy employed in this experiment. Freeing any software product from the constraints we imposed will improve matters, but it is unclear by how much and what further operation implications there might be.

Recent related qualitative work suggests the value of visible deterrence for the positive marked car results. When Summers and Rossmo (2019) provided over 130 offenders with a vignette featuring two police cars driving by in a short space of time, 63% said they would displace their offending to another time or place, and 23% said they would give up for the day. The desistance/displacement results for marked patrol cars were larger than any other security measure they tested.

Until now, there has been little available evidence regarding how best to use patrol police officers beyond hot spots policing. Thus, while there are numerous caveats expressed in this article, the property crime results are among the first positive robust empirical findings derived from a randomized experiment of a tactic tied to a predictive policing implementation. This experiment provides some operational insights for patrol commanders.

This remains the result of just one experiment. Not only is there a vital need for more research with the highest standards of scientific validity, such investigations also should test deployments that are democratically acceptable, realistic, and operationally viable. If the current enthusiasm for predictive policing is to develop into a mainstream strategy, there is no point testing resource-intensive law enforcement strategies that police departments cannot sustain on an ongoing basis. Moreover, questions about what police should do in predicted crime areas remain central to progress. These will have to be discussed and studied in an open and transparent manner. Predictive policing may help law enforcement to be more targeted and precise, but for long-term effectiveness, analytical solutions to crime problems must be developed alongside transparent and democratically viable prevention mechanisms.

Funding information This research was made possible by award number 2014-R2-CX-0002 from the National Institute of Justice.

Compliance with ethical standards

Conflict of interest Any views or opinions expressed herein do not necessarily reflect the official policies of the Department of Justice, the National Institute of Justice, the Philadelphia Police Department, or the City of

Philadelphia. Any mention of trade names, commercial practices, or organizations does not imply endorsement by the Department of Justice, the National Institute of Justice, the Philadelphia Police Department, or the City of Philadelphia. None of the authors has any financial investment in Azavea, Inc. or ShotSpotter Inc., and none of the authors has received any financial or other incentive from either company related to the HunchLab product.

References

Azavea. (2014). *HunchLab: under the hood*. Philadelphia: Azavea.

Bowers, K. J., & Johnson, S. D. (2003). Measuring the geographical displacement and diffusion of benefit effects of crime prevention activity. *Journal of Quantitative Criminology, 19*(3), 275–301.

Box, J. F. (1980). R. A. Fisher and the design of experiments, 1922-1926. *The American Statistician, 34*(1), 1–7.

Braga, A. A., Papachristos, A., & Hureau, D. (2012). Hot spots policing effects on crime. Campbell Systematic Reviews, report no. 8. Oslo, Norway.

Brantingham, P. L., & Brantingham, P. J. (1981-1982). Mobility, notoriety, and crime: a study in the crime patterns of urban nodal points. *Journal of Environmental Systems*, 11(1), 89-99.

Browne, W. J. (2012). *MCMC estimation in MLwiN*. Bristol: University of Bristole Centre for Multilevel Modeling.

Caplan, J. M., Kennedy, L. W., & Miller, J. (2011). Risk terrain modeling: brokering criminological theory and GIS methods for crime forecasting. *Justice Quarterly, 28*(2), 381.

Chainey, S., & Ratcliffe, J. H. (2005). *GIS and crime mapping*. London: Wiley.

Chainey, S., Tompson, L., & Uhlig, S. (2008). The utility of hotspot mapping for predicting spatial patterns of crime. *Security Journal, 21*(1-2), 4–28.

Cohen, L. E., & Felson, M. (1979). Social change and crime rate trends: a routine activity approach. *American Sociological Review, 44*, 588–608.

Cornish, D., & Clarke, R. (1986). *The reasoning criminal: rational choice perspectives on offending*. New York: Springer.

Detry, M. A., & Lewis, R. J. (2014). The intention-to-treat principle: how to assess the true effect of choosing a medical treatment. *JAMA, 312*(1), 85–86.

Drawve, G. (2016). A metric comparison of predictive hot spot techniques and RTM. *Justice Quarterly, 33*(3), 369–397.

Durlauf, S. N., & Nagin, D. S. (2011). Imprisonment and crime: can both be reduced? *Criminology & Public Policy, 10*(1), 13–54.

Eck, J. E., Lee, Y. O. S., & Martinez, N. (2017). Compared to what? Estimating the relative concentration of crime at places using systematic and other reviews. *Crime Science, 6*(8), 1–17.

Felson, M. (1987). Routine activities and crime prevention in the developing metropolis. *Criminology, 25*(4), 911–932.

Fisher, R. A. (1935). *Design of experiments*. Edinburgh: Oliver and Boyd.

Flaxman, S., Chiricoz, M., Pereirax, P., & Loeffler, C. (2019). Scalable high-resolution forecasting of sparse spatiotemporal events with kernel methods: a winning solution to the NIJ "real-time crime forecasting challenge". *Annals of Applied Statistics*.

Gelman, A., Carlin, J. B., & Stern, H. S. (2003). *Bayesian data analysis* (2nd ed.). Boca Raton: Chapman & Hall/CRC.

Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1996). Introducing Markov chain Monte Carlo. In W. R. Gilks, S. Richardson, & D. J. Spiegelhalter (Eds.), *Markov Chain Monte Carlo in practice* (pp. 1–20). Boca Raton: Chapman & Hall/CRC.

Gill, C. E., & Weisburd, D. (2013). Increasing equivalence in small-sample, place-based experiments: taking advantage of block randomization methods. In B. C. Welsh, A. A. Braga, & G. J. N. Bruinsma (Eds.), *Experimental criminology: prospects for advancing science and public policy* (pp. 141–162). Cambridge: Cambridge University Press.

Gorr, W. L., & Lee, Y. (2015). Early warning system for temporary crime hot spots. *Journal of Quantitative Criminology, 31*(1), 25–47.

Gorr, W. L., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting, 19*(4), 594.

Greene, J. R. (2014). New directions in policing: balancing prediction and meaning in police research. *Justice Quarterly, 31*(2), 193–228.

Haberman, C. P., & Ratcliffe, J. H. (2012). The predictive policing challenges of near repeat armed street robberies. *Policing: A Journal of Policy and Practice, 6*(2), 151–166.

Ignatans, D., & Pease, K. (2016). Taking crime seriously: playing the weighting game. *Policing: A Journal of Policy and Practice, 10*(3), 184–193.

Johnson, S. D. (2010). A brief history of the analysis of crime concentration. *European Journal of Applied Mathematics, 21*(4/5), 349–370.

Johnson, S. D., Bowers, K. J., Birks, D., & Pease, K. (2009a). Predictive mapping of crime by ProMap: accuracy, units of analysis and the environmental backcloth. In D. Weisburd, W. Bernasco, & G. J. N. Bruinsma (Eds.), *Putting crime in its place: units of analysis in geographic criminology* (pp. 171–198). New York: Springer.

Johnson, S. D., Summers, L., & Pease, K. (2009b). Offender as forager? A direct test of the boost account of victimization. *Journal of Quantitative Criminology, 25*(2), 181–200.

Kennedy, L. W., Caplan, J. M., & Piza, E. (2011). Risk clusters, hotspots, and spatial intelligence: risk terrain modeling as an algorithm for police resource allocation strategies. *Journal of Quantitative Criminology, 27*(3), 362.

Koper, C. S. (1995). Just enough police presence: reducing crime and disorderly behavior by optimizing patrol time in crime hot spots. *Justice Quarterly, 12*(4), 649–672.

Lee, Y. O. S., & Eck, J. E. (2017). *A theory-driven algorithm for real-time crime hot spot forecasting*. National Institute of Justice: Final report. Washington DC.

Lum, C., & Nagin, D. S. (2017). Reinventing American policing. *Crime and Justice, 46*(1), 339–393.

Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J., et al. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *British Medical Journal, 340*(c869), 1–28.

Mohler, G. O., Short, M. B., Brantingham, P. J., Schoenberg, F. P., & Tita, G. E. (2011). Self-exciting point process modeling of crime. *Journal of the American Statistical Association, 106*(493), 108.

Moreto, W. D., Piza, E. L., & Caplan, J. M. (2014). "A plague on both your houses?": risks, repeats and reconsiderations of urban residential burglary. *Justice Quarterly, 31*(6), 1102–1126.

National Research Council. (2004). *Fairness and effectiveness in policing: the evidence (Editors: Wesley Skogan and Kathleen Frydl). Committee to Law and Justice*. Washington: Division of Behavioral and Social Sciences and Education.

Perry, W. L., McInnis, B., Price, C. C., Smith, S. C., & Hollywood, J. S. (2013). *Predictive policing: the role of crime forecasting in law enforcement operations*. Washington: Rand Corporation.

Ratcliffe, J. H. (2000). Aoristic analysis: The spatial interpretation of unspecific temporal events. *International Journal of Geographical Information Science, 14*(7), 669–679.

Ratcliffe, J. H. (2014). What is the future... of predictive policing? *Translational Criminology, 2014*(Spring), 4–5.

Ratcliffe, J. H. (2015a). *Harm-focused policing*. In *Ideas in American Policing (19)*. Washington: Police Foundation.

Ratcliffe, J. H. (2015b). Towards an index for harm-focused policing. *Policing: A Journal of Policy and Practice, 9*(2), 164–182.

Ratcliffe, J. H., & McCullagh, M. J. (1998). Aoristic crime analysis. *International Journal of Geographical Information Science, 12*(7), 751–764.

Ratcliffe, J. H., & Rengert, G. F. (2008). Near repeat patterns in Philadelphia shootings. *Security Journal, 21*(1-2), 58–76.

Ratcliffe, J. H., Taniguchi, T., Groff, E. R., & Wood, J. D. (2011). The Philadelphia foot patrol experiment: a randomized controlled trial of police patrol effectiveness in violent crime hotspots. *Criminology, 49*(3), 795–831.

Ratcliffe, J. H., Taylor, R. B., & Fisher, R. (in press). Conflicts and congruencies between predictive policing and the patrol officer's craft. *Policing and Society.*

Rummens, A., Hardyns, W., & Pauwels, L. (2017). The use of predictive analysis in spatiotemporal crime forecasting: building and testing a model in an urban context. *Applied Geography, 86*, 255–261.

Saunders, J., Hunt, P., & Hollywood, J. S. (2016). Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot. *Journal of Experimental Criminology, 12*(3), 347–371.

Sherman, L. W. (2007). The power few: experimental criminology and the reduction of harm. *Journal of Experimental Criminology, 3*(4), 299–321.

Sherman, L. W., & Weisburd, D. (1995). General deterrent effects of police patrol in crime "hot spots": a randomized, controlled trial. *Justice Quarterly, 12*(4), 625–648.

Sherman, L. W., Gartin, P., & Buerger, M. E. (1989). Hot spots of predatory crime: routine activities and the criminology of place. *Criminology, 27*(1), 27–55.

Sherman, L. W., Williams, S., Ariel, B., Strang, L. R., Wain, N., Slothower, M., & Norton, A. (2014). An integrated theory of hot spots patrol strategy: implementing prevention by scaling up and feeding back. *Journal of Contemporary Criminal Justice, 30*(2), 95–112.

Sherman, L. W., Neyroud, P. W., & Neyroud, E. (2016). The Cambridge crime harm index: measuring total harm from crime based on sentencing guidelines. *Policing: A Journal of Policy and Practice, 10*(3), 171–183.

Sorg, E. T., Haberman, C. P., Ratcliffe, J. H., & Groff, E. R. (2013). Foot patrol in violent crime hot spots: longitudinal impacts of deterrence and post-treatment effects of displacement. *Criminology, 51*(1), 65–101.

Summers, L., & Rossmo, D. K. (2019). Offender interviews: Implications for intelligence-led policing. *Policing: An International Journal, 42*(1), 31–42.

Taylor, R. B. (1994). *Research methods in criminal justice*. New York: McGraw-Hill.

Taylor, R. B., & Ratcliffe, J. H. (2019). *Is the Pope's visit to blame, or something else?: statistical power, the lack thereof, and nano-scale predictive policing randomized control trials. Paper presented at the annual meetings of the American Society of Criminology.* San Francisco: November.

Taylor, B., Koper, C. S., & Woods, D. J. (2011). A randomized controlled trial of different policing strategies at hot spots of violent crime. *Journal of Experimental Criminology, 7*(2), 149–181.

Tompson, L., & Townsley, M. (2010). (Looking) Back to the future: using space–time patterns to better predict the location of street crime. *International Journal of Police Science and Management, 12*(1), 23–40.

Tseloni, A., & Pease, K. (2003). Repeat personal victimization: 'boosts' or 'flags'? *British Journal of Criminology, 43*(1), 196–212.

Weisburd, D., & Eck, J. (2004). What can police do to reduce crime, disorder, and fear? *The Annals of the American Academy of Political and Social Science, 593*(1), 43–65.

Weisburd, D., & Green, L. (1995). Measuring immediate spatial displacement: methodological issues and problems. In J. E. Eck & D. Weisburd (Eds.), Crime and place (Vol. Volume 4, pp. 349-361). Monsey: Criminal Justice Press.

Weisburd, D., & Telep, C. (2014). Hot spots policing: what we know and what we need to know. *Journal of Contemporary Criminal Justice, 30*(2), 200–220.

Weisburd, D., Groff, E. R., Jones, G., Cave, B., Amendola, K. L., Yang, S.-M., & Emison, R. F. (2015). The Dallas patrol management experiment: can AVL technologies be used to harness unallocated patrol time for crime prevention? *Journal of Experimental Criminology, 11*(3), 367–391.

Wyant, B. R., Taylor, R. B., Ratcliffe, J. H., & Wood, J. (2012). Deterrence, firearm arrests, and subsequent shootings: a micro-level spatio-temporal analysis. *Justice Quarterly, 29*(4), 524–545.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Jerry H. Ratcliffe is a professor in the Department of Criminal Justice at Temple University. A former police officer, his research areas include spatial criminology and evidence-based policing, and he is the host of the Reducing Crime podcast.

Ralph B. Taylor researches, teaches undergraduates and graduate students, and supervises or contributes to dissertations at Temple University in the Department of Criminal Justice. Current interests center on community criminology, and questions of disproportionality within police/community dynamics.

Amber Perenzin Askey is a data scientist at Deloitte Consulting LLP. Amber received her PhD from Temple University where she studied gangs, policing, and crime analysis.

Kevin Thomas is the Director of Research and Analysis for the Philadelphia Police Department, overseeing research, crime analysis, intelligence analysis, and analytics/business intelligence. With a multi-disciplinary background including urban planning, engineering, and design, he supports the Department in the development of data-driven, evidence-based crime-fighting strategies.

John Grasso is the Intelligence Manager for the Philadelphia Police Department Intelligence Bureau managing both sworn and civilian intelligence analysts. Prior to this, John was a Federal law enforcement officer, assigned to the Violent Crimes Task Force and to the Counter-Terrorism Task-Force as an intelligence officer. He was awarded his Juris Doctor in 2006.

Kevin J. Bethel is a retired Deputy Police Commissioner in the Philadelphia Police Department where he commanded patrol operations for the city. He also served as a Stoneleigh Foundation Fellow. Kevin is a senior advisor and Chief of School Safety for the Philadelphia School District.

Ryan Fisher is a doctoral candidate in the Department of Criminal Justice at Temple University. His main research interests include environmental criminology, spatial crime analysis, and policing.

Josh Koehnlein is a PhD student in the Department of Criminal Justice at Temple University. His research interests include environmental criminology and policing.

Affiliations

Jerry H. Ratcliffe[1] · Ralph B. Taylor[1] · Amber Perenzin Askey[2] · Kevin Thomas[3] · John Grasso[3] · Kevin J. Bethel[4] · Ryan Fisher[1] · Josh Koehnlein[1]

1 Department of Criminal Justice, Temple University, Philadelphia, USA

2 Deloitte Consulting LLP, New York, USA

3 Philadelphia Police Department, Philadelphia, USA

4 Philadelphia School District, Philadelphia, USA