Machine Learning Can Predict Shooting Victimization Well Enough to Help Prevent It

Sara B. Heller, Benjamin Jakubowski, Zubin Jelveh & Max Kapustin

July 27, 2024

Using arrest and victimization records for almost 644,000 people from the Chicago Police Department, we train a machine learning model to predict the risk of being shot in the next 18 months. Out-of-sample accuracy is strikingly high: of the 500 people with the highest predicted risk, almost 17 percent are shot within 18 months, a rate 106 times higher than the average Chicagoan. A central concern with using police data is that predictions will "bake in" bias, overestimating risk for groups likelier to interact with police conditional on behavior. We show that Black male victims more often have enough police contact to generate predictions. But those predictions are not, on average, inflated; the demographic composition of predicted and actual shooting victims is almost identical. There are legal, ethical, and practical barriers to using these predictions to target law enforcement. But using them to target social services could increase the potential for prevention programs to reduce shootings: predictive accuracy among the top 500 people justifies spending up to $190,900 per person for an intervention that could cut the probability of being shot by half.

Heller: University of Michigan & NBER (sbheller@umich.edu). Jakubowski: New York University (benjamin.u.jakubowski@gmail.com). Jelveh: University of Maryland (zjelveh@umd.edu). Kapustin: Cornell University & NBER (kapustin@cornell.edu). Contribution statement: Heller, Jelveh, and Kapustin shared responsibility for conceptualization, analysis, and visualization. Jelveh was responsible for data curation, methodology design, and implementation, assisted by Jakubowski. Heller and Kapustin led writing and revision. We are grateful to the Chicago Police Department for making available the data upon which this analysis is based. We thank Jalon Arthur, Phil Cook, Jen Doleac, Leif Elsmo, Jens Ludwig, Doug Miller,

1 Introduction

Gun violence in the U.S. causes widespread harm—to its direct victims and to the children, families, and communities around them (Sharkey, 2018)—generating social costs of at least $100 billion annually (Cook and Ludwig, 2000). Because addressing this problem with aggressive policing can generate its own significant social costs (e.g., Ang, 2021; Geller et al., 2014; Jones, 2014; Chalfin et al., 2022), local policymakers are spending millions of dollars to prevent gun violence with social services rather than law enforcement.[1] How much these services can reduce shootings is shaped by how well program operators can anticipate participants' *ex ante* risk; even a very effective intervention will prevent few shootings if few participants would be victims or offenders in its absence. Unfortunately, we know little about whether a person's risk of future shooting involvement can be predicted accurately, a prerequisite for individually-targeted interventions to make a cost-effective difference.

In settings from child welfare to opioid misuse, machine learning algorithms help solve this kind of prediction problem by forecasting future behavior accurately, consistently, and at scale (e.g., Obermeyer and Emanuel, 2016; Chouldechova et al., 2018; Kleinberg et al., 2018a; Hastings et al., 2020). But using algorithms to predict shootings faces two key challenges. First, predicting outcomes as rare as shootings has been a major challenge across many disciplines involving human behavior or other complex systems (Lo-Ciganic et al., 2019; Qi and Majda, 2020; Japkowicz, 2000; Salganik et al., 2020).[2] Achieving

Emily Nix, Andy Papachristos, Greg Ridgeway, Mark Saint, Pat Sharkey, Ravi Shroff, and Megan Stevenson for their feedback. We thank Xander Beberman, Melissa McNeill, and Gargi Sundaram for outstanding research assistance. This paper builds on a predictive model the authors developed to identify men for referral into READI Chicago, an experimental preventive social service intervention. The larger READI research effort had support from the philanthropic community, including the Partnership for Safe and Peaceful Communities, JPMorgan Chase, and the Chicago Sports Alliance. All opinions and any errors are our own and do not necessarily reflect those of our funders or of the Chicago Police Department.

[1] See, e.g., City of Chicago (2020), City of Baltimore (2021), City of Philadelphia (2021), Washington, D.C. (2022), and City of Oakland (2024).

[2] Even in our Chicago setting, where gun violence rates are high (though far from the highest in the U.S.), shootings injure or kill about 0.1 percent of residents annually.

adequate predictive performance may be particularly hard given the noise and distortions in crime data.[3]

Second, most input data likely come from the criminal legal system, and algorithms trained on crime data may "bake in" that system's biases. For example, harsher treatment of non-White individuals, and especially of Black men, is well documented (e.g., Antonovics and Knight, 2009; Arnold et al., 2018; Eberhardt et al., 2004; Goncalves and Mello, 2021; Hoekstra and Sloan, 2022; Rehavi and Starr, 2014). If Black men are likelier to be arrested conditional on their behavior, then even "accurate" predictions of an outcome like shooting arrest may nevertheless overestimate their true risk of committing a shooting relative to people in other groups (Mayson, 2019; Starr, 2014). When such an algorithm is used to target legal interventions that curtail civil liberties, the burden of its false positive mistakes will be borne by the same groups that have historically been treated unfairly by the criminal legal system (e.g., Angwin et al., 2016; Lum and Isaac, 2016; Richardson et al., 2019; Mehrabi et al., 2021). This concern recently led some mathematicians to abandon predictive policing because it is "simply too easy to create a 'scientific' veneer for racism."[4]

This paper shows that even the biased information in police data can predict shootings with enough accuracy to guide prevention efforts and without distorting average risk across demographic groups. The key is to not predict shooting *arrest*. In our setting, shootings often do not result in an arrest. The likelihood that police arrest a shooter may vary, including due to bias in police decisions about whom to investigate and arrest. An algorithm trained to predict this kind of biased proxy may yield predictions that incorporate or "bake in" these biases, a scenario known as "target variable bias" (Fogliato et al., 2020). This can lead to getting average group differences in risk dramatically wrong (Mullainathan and Obermeyer, 2021; Obermeyer et al., 2019), such as by systematically

[3] These distortions include, among others: arrests or convictions of innocent people, under-reporting of crimes, and crimes that do not result in any arrests.

[4] `https://www.popularmechanics.com/science/math/a32957375/mathematicians-boycott-predictive-policing/`

overestimating shooting offense risk for members of a group who are disproportionately likely to be arrested conditional on their behavior. Crucially, because true offending is unobserved, there is no way to assess how accurately, or with what bias, predictions of shooting arrests reflect actual shooting offenses.

Instead, we predict shooting *victimization*. Intervening with people likely to be shot to keep them safe is a plausible alternative to intervening with people likely to shoot for reducing gun violence (Cooper et al., 2006; Zun et al., 2006; Cheng et al., 2008; Green et al., 2017). And as we argue below, shooting victimization is likely to be measured consistently across demographic groups in our setting. Theoretical work suggests that predicting this kind of well-measured outcome will recover accurate estimates of risk at the group level even if the predictors used to do so are measured differently across groups (Kleinberg et al., 2018b). If these predictions are accurate enough, then they could help to cost-effectively reduce gun violence, especially when paired with interventions that reduce victimization risk without imposing significant costs when mis-targeted (e.g., preventive services). Currently, there is very little data about the predictability of shooting victimization, overall or by demographic group.

To fill this gap, we build a gradient boosting machine (GBM) model (Friedman, 2001) to predict shooting victimization. This decision tree-based approach is trained and tested using 20 years of records for 643,975 people from the Chicago Police Department (CPD), including over 1,400 predictive features that capture a person's demographic information, arrest and victimization histories, and the arrest and victimization histories of people who were co-involved in prior criminal incidents. We evaluate predictive performance on an out-of-sample 18-month period.

We have three main sets of results. First, the model successfully identifies a small group of people at extraordinarily high risk of being shooting victims. Of the 500 people at highest predicted risk, almost 17 percent are shot during the following 18 months—a rate 19 times higher than everyone in our prediction sample of people with recent police

contact (0.9 percent across 327,181 people) and 106 times higher than everyone in Chicago (0.2 percent across 2.7 million people). An intervention that could cut by half the risk of being shot for these 500 people would generate an estimated social cost savings of $96 million from the victimization reduction alone (Cook and Ludwig, 2000; Ludwig and Cook, 2001). If the intervention cost less than $190,900 per person, it would pay for itself. Our analysis unpacks what information the model uses to achieve this predictive performance.

Second, the predictions do not misrepresent average victimization risk across demographic groups. We show that Black male shooting victims are likelier to *have* a predicted risk, because they are likelier to have prior police contact.[5] This finding highlights how using only police data limits an algorithm's ability to identify future victims with little or no prior police contact. But importantly, the algorithm accurately recovers group-level victimization risk regardless of race, age, or gender. As a result, the demographic composition of predicted shooting victims matches almost exactly that of actual shooting victims. In other words, the over-representation of Black men in police data does not yield predictions that overestimate the average victimization risk of Black men; the predictions are well-calibrated (right on average) within groups and across the risk distribution.

Third, while a flexible machine learning approach achieves the greatest predictive performance, simpler prediction methods also perform well. For example, a GBM model identifies about 10 percent more victims among the 4,244 people with the highest predicted risk than does an ordinary least squares (OLS) model, when both use either the full set of over 1,400 features or only the 10 most predictive features.[6] Providing OLS with greater flexibility by adding two-way interactions of the 10 most predictive features closes half of the performance gap with GBM. In absolute terms, even these modest performance

[5] Black men make up 71 percent of all shooting victims during the outcome period studied here. The model generates predictions for 78 percent of them, compared to half or fewer of the victims from other groups.

[6] We use 4,244 as a rank cut-off example throughout the paper because there were 4,244 shooting victims in Chicago during the out-of-sample 18-month outcome period.

gaps can yield large differences: for the full feature set, GBM identifies 45 more shooting victims among the 4,244 people with the highest predicted risk than OLS. Whether this performance advantage of GBM outweighs the relative simplicity and feasibility of various OLS models likely varies across settings.

Two points about these findings are important for interpretation. First, we predict shooting victimization risk under the status quo amount of intervention, incarceration, and mortality, or $Y(0)$ in a potential outcomes framework. To minimize shootings, interventions should ideally target the people for whom treatment effects, or $Y(1) - Y(0)$, are largest.[7] However, in contexts with base rates as low as those for shootings, predicting $Y(0)$ is crucial to generating the evidence about $Y(1) - Y(0)$ that is required for such targeting to be possible. Choosing a study sample where $\bar{Y}(0)$ is too low leaves little room for an intervention to reduce shootings, making it hard to detect a treatment effect. And even at high risk levels, statistical power is sensitive to small changes in $\bar{Y}(0)$. For example, the sample size required to detect a 50 percent reduction in shooting victimization is 61 percent larger when the sample is drawn from the 98^{th}-99^{th} percentile of the full model's predicted risk distribution, relative to when it is drawn from above the 99^{th} percentile, where $\bar{Y}(0)$ is higher.[8] With an outcome as rare as being shot, accurately anticipating who will have a high $Y(0)$ is a necessary, if not sufficient, condition for testing preventive interventions and identifying optimal targeting strategies.

The second point important for interpretation is that getting predictions right on average across demographic groups is not the same as the algorithm being "fair" or "unbiased." As we discuss in section 4.2, even algorithms that get group averages right across the risk distribution can still mis-rank people, both within and across demographic groups. How much mis-ranked predictions matter for "fairness" depends on the decision rule one adopts—whether to serve everyone above a global or group-specific threshold

[7] In practice, big *changes* in Y are sometimes (e.g., Heller, 2022), but by no means always (e.g., Ascarza, 2018; Haushofer et al., 2022), correlated with high *levels* of $Y(0)$.

[8] See section 5 for additional details.

of predicted probability, whether to apply geographic or age restrictions as often done in practice, and so on—and what kind of fairness one chooses to prioritize.[9] And importantly, whether an algorithmic decision rule improves fairness under a given definition depends on the counterfactual decision-making process. Because we have no systematic data on how violence prevention services are currently allocated, we cannot say whether a given algorithmic decision rule would be fairer than the status quo in our setting.

As a result, we leave it as a central task for policymakers to map predictions onto service decisions in a way that satisfies normative preferences about fairness in a given setting, and to weigh whether using an algorithm increases or decreases bias relative to the alternative decision-making process. Our goal in this paper is not to evaluate particular fictional use cases. Rather, it is to demonstrate that predicting a well-measured outcome can get average demographic group risk right even when biased input data over-represent some groups, and that shootings are predictable enough with these data to make cost-effective individual interventions—and better research on those interventions—a plausible reality.

To be clear, predicting shootings with police data is by no means a complete solution to gun violence, and predictions should be used with care. It is important to attend to whom this kind of algorithm misses, and to the dangers and limitations of using such predictions to target law enforcement rather than social services (see section 5). Still, this paper establishes that shooting victimization is predictable enough for algorithmic screening to help preventive social services reach the people who need them, in the same way that algorithms have been proposed to screen for risk of depression, opioid abuse, and suicide in broader populations to prevent future harm (Garza et al., 2021; Eichstaedt et al., 2018; Hastings et al., 2020). Of course, actually preventing shootings requires effective interventions. Understanding what kind of preventive services can reduce shootings for different parts of the risk distribution should be a priority for future research.

[9] It is well known that different definitions of fairness conflict with each other and not all can be simultaneously satisfied (e.g., Chouldechova, 2017; Kleinberg et al., 2017).

2 Related literature

The literature on using machine learning-based predictions to guide decision-making (e.g., Kleinberg et al., 2018a; Athey, 2017; Chouldechova et al., 2018; Hastings et al., 2020; Obermeyer and Emanuel, 2016) does not engage with the risk of shootings specifically. But it does provide two key priorities for evaluating predictive models generally (also see the discussion in Berk, 2008). First, predictions must be a true forecast, relying only on information available to the decision-maker at the time they are made. Second, model performance must be assessed out-of-sample, i.e., using data separate from those with which the model is trained. Since in-sample predictive performance overstates how well observable features can predict future behavior due to over-fitting, it does not demonstrate the predictability of future violence. There is another key priority specific to our context: generating predictions that capture true differences in risk across demographic groups, rather than differences in behavior by actors in the legal system embedded within the predicted outcome.

For these reasons, the large literatures in psychology and criminology on risk assessment instruments to predict different types of violent offending (see reviews in Otto and Douglas, 2010; Hanson, 2005; Singh et al., 2011), as well as the risk factors correlated with violence more generally (Hawkins et al., 1998; Farrington et al., 2017), do not speak to the predictability of gun violence.[10] These studies typically either collect information from an interview to assess a known person's risk (e.g., a detainee or parolee) or examine in-sample correlations to identify potential risk factors. As such, they are not designed to establish how predictable shootings are, forecast and rank the risk of future shooting victimization across a large population, or assess whether those predictions distort true differences in risk due to underlying biases in the data generating process.

We build on several peer-reviewed papers and technical reports that have made im-

[10] For an overview of this literature, see Wheeler et al. (2019).

portant progress toward testing whether shootings are predictable. Berk et al. (2009) use machine learning to generate true forecasts and carefully explore predictive performance. But partly because the predictions were intended to target probation and parole services, they predict homicide charges rather than victimization. Using an outcome partially determined by legal system actors makes it hard to assess whether the algorithm is predicting a person's risk of homicide offending or potentially biased police and prosecutor decision-making about whom to arrest or charge. And relatively low homicide arrest rates make it impossible to assess which offenders are being missed. Wernick (2018) and Wheeler et al. (2019) both study algorithms that predict a combination of shooting offending and victimization, and are therefore both subject to the same concern as Berk et al. (2009).[11] Chandler et al. (2011) predict the out-of-sample risk of being a shooting victim using ordinary least squares with Chicago Public Schools data. But their analysis is limited to high school students (a small minority of shooting victims) and does not report performance by group.

A large and influential body of work by Andrew Papachristos and coauthors (e.g., Green et al., 2017; Papachristos et al., 2012; Papachristos and Wildeman, 2014; Papachristos et al., 2015a; Papachristos et al., 2015b; Papachristos and Bastomski, 2018; Wood and Papachristos, 2019) documents the concentration of gun violence within social networks and explores the role these networks play in determining one's own risk of being shot. Green et al. (2017) provide a seminal insight about the role of social network measures in predicting the risk of shootings for prevention purposes, which directly influenced our feature selection below. But their prediction model relies on measures of co-arrest ties that do not appear in the data until after when an intervention would be delivered, making it infeasible for use as a pure forecasting method. They also fit and assess the performance of their model using the same data, making it difficult to determine how accurately the model predicts out-of-sample behavior.

[11] Wheeler et al. (2019) also do not explore performance by demographic group.

This study improves upon and extends the prior literature by providing the three types of assessment needed to know if it is possible to predict who will be shot within a population without distorting risk across demographic groups. We perform a pure forecasting exercise using only data available at the time of prediction. We assess performance on data not used in the model-building process, including information about shooting victims who are not in the prediction data at all. And by predicting an outcome that is consistently measured across demographic groups, we document how many shootings a given use of data-driven predictions would capture or miss and for whom, what shapes those predictions, and how performance varies across race, gender, and age groups.

3 Method

We build a model that predicts a person's likelihood of being reported in Chicago police data as injured or killed by gunfire (reported shooting victimization) in the next 18 months.[12] The key modeling decision we make is to predict reported shooting victimization rather than arrest. Most shooting offenses in Chicago do not result in an arrest, nor do all shooting arrests mean the arrestee committed a shooting.[13] Both the likelihood of a shooting resulting in an arrest and the likelihood of a mistaken or wrongful arrest may vary across groups due to, among other factors, differences in police behavior.

In contrast, police records likely capture the vast majority of shooting victimizations in Chicago. As a result, reported shooting victimizations are likely to measure actual victimizations consistently across demographic groups (and certainly more consistently than shooting arrests are to measure offenses). This is true of shooting injuries that are

[12] This excludes suicides and shootings by police officers; these incidents are not in the data and the interventions to reduce them likely differ substantively from those designed to reduce the gun assaults that are our focus. We use an 18-month outcome period because a closely related model was used to identify participants for an 18-month intervention (Bhatt et al., 2024). Appendix B.1 shows that our results are consistent across alternative evaluation and training durations.

[13] In the first half of the 2010s, under half of homicides and fewer than 10 percent of non-fatal shootings in Chicago resulted in an arrest (Kapustin et al., 2017).

immediately fatal and to which police are among the first to respond. But it is also true of non-fatal shootings, as most victims receive medical care (Barber et al., 2022) and healthcare providers in Illinois are mandated to report firearm injuries to law enforcement (20 ILCS 2630/3.2). There may still be a concern about selective under-reporting of non-fatal shooting injuries by victims who avoid receiving medical care due to mistrust or fear of the police (e.g., Liebschutz et al., 2010). However, surveys of jail inmates across the U.S. (May et al., 2002), and specifically of Chicago men incarcerated for gun crimes (White et al., 2021), consistently find that almost all sought medical care after they were shot. We provide additional empirical evidence of the low rate of shooting victimization under-reporting in Chicago police data in Appendix A.1.3.

To build our model, we start with Chicago Police Department (CPD) data on 12.7 million event-level records from January 1999 through October 2019. These records contain information on demographics, arrests, and reported victimizations in Chicago for youth and adults. We then use a probabilistic record linkage algorithm (Tahamont et al., 2021) to group together records belonging to the same people. Finally, we construct detailed, retrospective predictive features about each person in our sample and use these features to train a model to predict their risk of reported shooting victimization in the next 18 months. We describe key aspects of this process below; for additional details, see Appendix A.

To predict a person's risk at a point in time, we first require that they meet a sample inclusion criterion: having at least one arrest or two reported victimizations in the 50 months before the prediction date.[14] We do this mainly to remove from the analysis people at very low risk of being shot: those who do not meet the inclusion criterion have a rate of shooting victimization 70 times lower than that of people who meet it. Furthermore, excluding people with a single reported victimization reduces the influence of record-linkage error caused by missing date of birth information in some victimization

[14] Importantly, we observe all reported shooting victimizations in the CPD data, even for people who do not meet the inclusion criterion.

records (see Appendix A.2).

We then train and test a gradient boosting machine (GBM) model (Friedman, 2001). We choose this approach because GBM models have been shown to generally outperform other machine learning methods on tabular data (Caruana et al., 2008; Shwartz-Ziv and Armon, 2022). Mimicking how such a model might be used in practice, we train it to predict a person's risk of a reported shooting victimization in the 18 months after a given prediction date. To increase the amount of data available for training, validation, and testing, we divide the data into four calendar time cohorts (Appendix Figure A.1). Each cohort has a prediction date, preceded by a 50-month sample inclusion period and followed by an 18-month outcome period. The four cohorts' outcome periods do not overlap, a point we return to below. There are 643,975 people who meet the inclusion criterion across one or more of these cohorts and contribute to the modeling process.

To predict reported shooting victimization, we construct 1,411 features for each person-cohort using only data from before the cohort's prediction date (see Appendix A.3 for details on feature construction). These features fall into four categories. Demographic features include age, gender, race and ethnicity,[15] and police beats associated with home and incident addresses.[16] Arrest and victimization features include time-windowed, cumulative counts of these prior incidents, separately by the type of crime involved (e.g., robbery, shooting, vandalism).[17] For example, one feature counts the number of arrests for robberies involving a firearm within the past two years, while another counts the number of shooting victimizations within the past 90 days.

[15] It is worth noting that, in practice, there are many legal issues with the inclusion of race and ethnicity in algorithms (Yang and Dobbie, 2020). We include it in our full model because it may help the algorithm make more accurate predictions by racial group when predictors are recorded differently by race (Kleinberg et al., 2018b). But as we show in Appendix B.5.2, excluding race from the model leaves our results basically unchanged.

[16] There are 277 total police beats in Chicago, compared to 866 census tracts and 77 community areas (neighborhoods).

[17] The time windows are within 30, 60, 90, 180, 365, 730, and 1825 days before the prediction date, as well as the time since January 1999.

Finally, network features include time-windowed, cumulative counts of prior arrests and victimizations among people to whom the focal person is connected through co-involvement in prior criminal incidents ("neighbors"), defined as either being arrested for the same incident or being a victim-arrestee pair in the same incident. Crucially, network links are created only using data on incidents that occurred prior to the cohort's prediction date; two people who are first co-arrested after the prediction date, for example, are not (yet) neighbors. Examples of network features include counts of the number of gun possession arrests in the past 180 days among neighbors to whom the focal person is connected directly ("first-degree neighbors"), or counts of the number of robbery victimizations in the past 90 days among neighbors one degree further removed ("second-degree neighbors"). Also included are features describing the local structure of the network graphs themselves, such as the focal person's centrality and number of neighbors.

An important consideration for how we train the model is that, when people are connected, the outcomes of one person may contain information about those of another person (Chouldechova et al., 2018). For example, if persons i and j commit crimes together, then their arrests in a given period may be correlated. This poses an issue for traditional model training approaches that randomly split data into training, validation, and test sets: if i and j are in the training and test sets, respectively, then their being connected may cause information leakage between these sets through their correlated outcomes. The more people there are in the validation and test sets whose outcomes are correlated with those of people in the training set, the more information leakage may result, potentially inflating performance estimates and, ultimately, leading to poor model selection and subpar real-world performance.[18] To avoid this issue, we depart from traditional model training approaches and define these sets using our cohorts: the first two cohorts are the training set, the next is the validation set used for hyperparameter tuning, and the last is

[18] Note that removing network features from the model does not address this concern, as the contemporaneous outcomes of connected people may still be correlated.

the test set. This ensures that no information from outcomes in the validation or test sets leaks into the training set, as the outcome periods for the validation and test sets occur after those of the training set. All the results we report speak to the model's performance at predicting reported shooting victimization for the 327,181 people in the test set, for the out-of-sample 18-month outcome period starting April 1, 2018.

A different concern that arises from how we define and use cohorts to train the model is potential correlation in outcomes over time within the same person. When a person appears across cohorts—as those with frequent police contact often do—their time-windowed features and outcomes are defined relative to each cohort's prediction period. If a person's outcomes across cohorts are not independent, and if the person appears in the training set and in either the validation or test sets, then there may be information leakage that results in overly optimistic performance estimates. In Appendix B.2, we show that our results are robust to alternative approaches for training a GBM model that account for this potential correlation within a person over time.

4 Results

Our main results describe how well the algorithm can predict future gun violence to help target prevention services, with particular transparency around two major concerns with using algorithms in practice: racial disparities and what influences predictions. We first assess the model's overall performance, focusing on its ability to identify the relatively small number of people at high predicted risk of being shot.[19] We then show how predictions vary by demographic group, unpacking who is identified and missed with this kind of approach. Finally, we describe how changing the type of information

[19] Given how rare shootings are in the overall population, we avoid two common performance metrics: accuracy and area under the curve (AUC). Accuracy, defined as the share of all predictions made correctly, will be mostly driven by correctly classifying non-victims. AUC describes performance across the entire risk distribution and may not effectively assess a model's ability to identify people at the highest risk. Instead, we focus on performance measures at or above approximately the top 1 percent of the predicted risk distribution, the segment most relevant for directing preventive services.

available to the algorithm or the modelling method affects performance.

4.1 Performance overall

The top left panel of Figure 1 shows the overall distribution of the model's predictions compared to realized rates of shooting victimization for the sample. The x-axis is the average predicted risk for each percentile of the risk distribution, with each point containing 1 percent of the test sample, or 3,272 people. The y-axis is the rate of shooting victimization in the 18-month outcome period for the 3,272 people in each bin. The bootstrapped confidence intervals capture sampling variation in the outcome. Computing capacity prevents us from re-building the model within each bootstrap sample, so the intervals do not capture uncertainty from the model-building process.[20]

Three features about the overall predictions are apparent. First, on average, the model's risk predictions are accurate (well-calibrated): their slope is close to the 45-degree line. Second, the vast majority of people in the sample are predicted to have a shooting victimization risk close to zero, as indicated by the mass of points in the bottom left of the graph. Finally, the predicted risk distribution is highly positively skewed, with points in the upper right of the graph corresponding to a small group of people in the long right tail whose predicted risk of being shot in the 18-month outcome period is very high. We discuss the other panels of Figure 1 in the next section.

Figure 2 reports two measures of model performance across the predicted risk distribution. For each panel, the x-axis ranks everyone in the prediction sample by their predicted risk of victimization, with highest predicted risk on the left. Figure 2a shows Precision_k on the y-axis, or the share of people who are shot during the 18-month outcome period among the k people with the highest predicted risk: $\text{Precision}_k = \sum_{i=1}^{k} \mathbb{1}(\text{Shooting victim}_i = 1)/k$. Figure 2b shows Recall_k on the y-axis, or the share of shooting victims during the 18-month outcome period who are among the k people with highest predicted risk:

[20] For additional details, see Appendix A.5.2.

$\text{Recall}_k = \sum_{i=1}^{k} \mathbb{1}(\text{Shooting victim}_i = 1)/(\text{Total shooting victims})$.[21]

We show two versions of recall in Figure 2b. The first, labeled recall, uses the total number of shooting victims in the prediction sample as the denominator, or 2,827. The second, labeled total recall, uses the total number of shooting victims in the entire city during the outcome period as the denominator, or 4,244. The difference highlights a point we return to in the following section about whom predictions based on police data miss: one-third of eventual shooting victims are not in our prediction sample and are therefore not assigned a predicted risk by the model. Though it is more common when evaluating the performance of a predictive algorithm to report recall, total recall helps to assess the ability of algorithmic prediction to identify shooting victims city-wide, regardless of whether they have enough prior police contact to be included in the prediction sample.

The share of people shot in the 18-month outcome period is startlingly high among those in the right tail of the distribution (Figure 2a). Among the $k = 500$ people with highest predicted risk, 16.6 percent, or 83 people, are shot. This is 19 times higher than the victimization rate for the whole prediction sample (327,181 people) of 0.9 percent, and 106 times the victimization rate for the entire city (2.7 million people) of 0.2 percent. Among the $k = 4,244$ people with highest predicted risk—corresponding to the actual number of shooting victims during the outcome period—11.5 percent are shot. Those at higher predicted risk for shooting victimization are also at elevated risk for other adverse outcomes, like shooting arrest and other violent victimization (Appendix Table B.3).

The recall rates confirm that those in the right tail of the distribution account for an outsized share of all shooting victims (Figure 2b). Despite representing just under 0.02 percent of the city's population, the $k = 500$ people with highest predicted risk include 2.0 percent of the 4,244 total victims during the 18-month outcome period.[22] The $k = 4,244$

[21] In the public health literature, precision is commonly referred to as positive predictive value, and recall is commonly referred to as sensitivity or the true positive rate.

[22] Considering only the 2,827 shooting victims in the prediction sample rather than all 4,244 victims, recall at this threshold is 2.9 percent.

people with highest predicted risk—just over 0.1 percent of the city's population—include 11.5 percent of total victims.

Still, the recall rates make clear that not all shootings are easily predicted using observable factors derived from police data. Future victims are missed in two ways. First, by construction, the algorithm misses the 33.4 percent of victims who are not included in the prediction sample. This can be seen by the gap between the recall and total recall curves at $k = 327,181$ in Figure 2b. Second, most eventual victims are assigned a low predicted risk that leaves them outside the top $k = 500$ or $k = 4,244$. This may be partly because being shot is inherently difficult to predict: it is the product of both a complex social phenomenon (i.e., engaging in high-risk behavior) and of randomness (being hit when fired at). But it may also be because the model can better distinguish risk among people about whom it has more, and more recent, information. For example, eventual victims among the $k = 4,244$ people with highest predicted risk have almost 11 times as many arrests in the prior year as eventual victims not among the $k = 4,244$ (2.1 vs. 0.2).

4.2 Performance by group

A major concern about using police data for prediction is that differences in those data across demographic groups may not reflect differences in the group members' behavior, but rather differences in police officers' behavior toward those groups. As a result, though the model's predictions match realized rates of shooting victimization overall (top left panel of Figure 1), it may still over- or under-predict risk—or fail to predict it altogether—more for members of some groups than others due to differences in how or whether they appear in police data. For example, Black individuals appear in the prediction sample four times as often as White individuals and almost three times as often as Hispanic individuals, despite each group making up roughly a third of the city's population. (Note that throughout the paper, we refer to individuals of any race as Hispanic if this is their indicated ethnicity; those to whom we refer as White or Black include only those who are

non-Hispanic.[23]) A key concern is that if some of this over-representation is because Black residents are more likely to come into contact with the police due to over-policing of Black neighborhoods, or due to a greater propensity among officers to stop, search, or arrest them conditional on their behavior, then police data will systematically misrepresent Black individuals' behavior in a way that could generate inflated predictions of the shooting victimization risk they face. Similarly, under-policing of other demographic groups may lead the model to under-predict their victimization risk.

The three remaining panels of Figure 1, which report calibration separately by race or ethnicity, show this is not the case on average, across the distribution of predicted risk. Each point is a bin containing one percent of people of the indicated race or ethnicity in the prediction sample. Relative to other groups, the distribution of predicted risk is wider—extends further to the right—for Black individuals, and their average predicted risk is 2.2 times higher. Yet importantly, the slope of the line shows that the higher predicted risks of shooting victimization for Black individuals are not inflated: they are, on average, accurate probability estimates, falling close to the 45-degree line across the risk distribution. If anything, it is the Hispanic individuals predicted to be in the right tail of shooting victimization risk for whom the predictions may slightly overestimate risk, as indicated by the points below the 45-degree line (see Appendix B.4 for further quantification and discussion). But on average, the predictions are accurate about the risk of shooting victimization across racial groups and across the risk distribution.

We next provide a fuller accounting of how using police data shapes the demographic composition of the predictions relative to the demographic composition of those who are shot. As shown in Figure 2b, of the 4,244 shooting victims in the outcome period, two-thirds, or 2,827, have enough prior police contact to have a prediction. Not all shooting victims are equally likely to have a prediction: three-fourths of all Black male victims,

[23] Race and ethnicity information contained in the data likely reflect the views of officers rather than the subjects themselves.

but only half or fewer of the victims from other demographic groups, have a prediction (Appendix Figure B.1). This pattern is consistent with the over-representation of Black men in police data more generally. But in this case, over-representation may help predict shootings since Black men comprise the largest share of all victims (71 percent). A key implication is the need for other methods and data sources to help identify and prioritize for prevention people at high risk of victimization who would be missed by an algorithm trained solely using police data.

Figure 3 provides further evidence that the predictions are successfully matching the true demographic composition of shooting victims, as well as the demographic implications of one particular use of the predictions. The first two rows break down who is included or missed in the sample by comparing the demographic composition of all 4,244 shooting victims city-wide (first row) to the demographic composition of the 2,827 victims in the prediction sample (second row). Comparing the first and second rows shows that Black male victims, both those above and below the median shooting victim age of 23, are slightly over-represented in the data relative to the other groups. The third row shows the demographic composition of "predicted victims" in the sample, calculated by averaging across all 327,181 people in the prediction sample while weighting each person by their predicted risk of victimization (see Appendix A.5.1 for details). Comparing the second and third rows again shows that the calibration of the model's predicted probabilities does not vary systematically by demographic group; the demographic shares of predicted victims are quite close to those of actual victims in the prediction sample, with predictions just barely under-stating the proportion of younger Black male victims and over-stating the proportion of older Hispanic male victims. Additional details on predictive performance by demographics are in Appendix B.4.

If we predicted an outcome like arrest, we would be unable to determine whether differences in predictions across demographic groups are due to true differences in behavioral risk across them, or whether they are due to differences in police decision-making

about whom to arrest from each group. In our case, however, we predict an outcome (reported shooting victimization) that captures the true behavior of interest (actual shooting victimization) with little differential error across groups, and the model is relatively well-calibrated by race. So we can conclude that even if our arrest predictors represent a distorted picture of differences in offending across groups, the resulting predictions of our outcome—whether someone is shot—are not, on average, systematically biased across the race, age, and gender groups in the data. This is broadly consistent with theoretical work finding that an algorithm with access to information that allows it to reconstruct race can "learn" accurate race-specific rankings of risk (Kleinberg et al., 2018b).

It is important to note, however, that calibration within demographic groups does not imply that the algorithm removes all potential influence of differential policing (across or within groups). For example, suppose Black neighborhoods are over-policed relative to non-Black neighborhoods. The resulting differential measurement error in the predictors can still affect how well the algorithm can rank across groups, even when getting group averages right (Corbett-Davies et al., 2017). And if differences in how the predictors are measured are driven by unobservables (e.g., if there is unobserved heterogeneity in over-policing within Black neighborhoods), then the algorithm may be unable to learn the differential relationship between the predictors and the outcome. In that case, there could still be mis-ranking within groups (Kleinberg et al., 2018b).

As discussed in the introduction, the implications of this kind of mis-ranking for measures of fairness or bias hinge on the decision rule that maps predictions onto service decisions, and how that compares to the counterfactual non-algorithmic decision rule. The costs of each decision rule also rest on whether the services provided are helpful or harmful to the individual and society.[24] These issues are crucial for using algorithms in practice. But in the absence of specific use cases, we do not have enough information

[24] Costs may also be a function of the sources of inequities. For example, all else equal, stakeholders may place a higher cost on imbalance in fairness measures if the cause is bias from policing as opposed to if the cause was from true underlying risk differences.

to make broad claims about fairness or bias. Instead, we offer one simple example to highlight how decision rules shape who would be served.

As suggested by the dramatically different risk distributions by race/ethnicity in Figure 1, any decision rule that offers prevention services to everyone above some high threshold of predicted risk in this setting will end up serving a disproportionately Black population, as well as a small number of Hispanic and White individuals. The fourth row of Figure 3 shows the demographic implications of one such threshold rule as a stylized example: serving the 4,244 people at highest predicted risk.

Compared to actual and predicted victims, this group overwhelmingly comprises Black men, and particularly young Black men. It includes almost no women. And older Black men are under-represented despite making up the plurality of actual victims. Importantly, the concentration of young Black men at the top of the predicted risk distribution does not indicate falsely inflated risk; even within this above-threshold group, young Black men have the highest realized risk (see Appendix Table B.6 for performance measures by group under this decision rule). The model is most calibrated for Black men overall while systematically over-predicting risk for Hispanic men (consistent with the right side of the Hispanic panel in Figure 1). This pattern likely reflects both the higher true risk of shooting victimization among some young Black men and the model's ability to identify those individuals. Further examination about why some groups of victims are more easily identified by the model, and whether this informs what kind of services would be most useful to them, is warranted.[25]

There are, of course, normative fairness questions involved with any way of allocating scarce services, including an algorithmic threshold rule (see section 5 for discussion). The descriptive result here, which may help inform those normative discussions, is that a

[25] For example, if domestic violence shootings are harder to predict using police data, and if a larger share of female victims are shot in such incidents, then this could explain female victims' low predicted risk. Conversely, if shootings with young Black male victims are easier to predict using police data, then this could explain these victims' higher predicted risk. We cannot explore these issues because our data lack information on the nature of shooting incidents, but this would be a useful avenue for future work.

threshold rule would allocate services disproportionately to young Black men in a case where the algorithm is, on average, getting the demographic distribution of shooting victims quite close to correct. And depending on where it is drawn, this kind of threshold allocation rule may miss almost all female victims, likely including many victims of domestic gun violence. Such a rule violates many fairness measures, including ones that require equal representation across groups. We therefore make no claim that successful group calibration, even across the risk distribution, means that a given use-case of the algorithm, such as the threshold allocation rule described here, would be "fair."

4.3 What matters for performance?

4.3.1 Feature sets

We are interested not only in whether the model can identify people at high risk of being shooting victims, but also what information allows it to do so. A common strategy for answering this question in machine learning applications is to report the "importance" of individual features. One way to do this is by assessing how much a given feature affects the predictions of a model that has already been built, such as by permuting the feature's values and measuring the impact on prediction errors using the same model (Breiman, 2001). However, this approach can be easily misinterpreted, especially when closely correlated features exist (Tolo̧si and Lengauer, 2011). For example, if a model loads heavily on one feature and not its correlated counterpart, then the former feature may be "important" in terms of affecting predictions within a given model, but unimportant in terms of not materially changing model performance when that feature is left out entirely.

An alternative approach that better answers the importance question is to retrain the model leaving out the feature in question (Lei et al., 2018). By allowing the remaining features to substitute for the missing information, this approach determines which features capture information that is substantively important for predictive performance and cannot be found in other features. While ideal, it is often impractical to leave out one feature at

a time and rerun the computationally expensive model-building process. Therefore, to implement this in practice and aid interpretation, we focus on removing *sets* of features grouped by substantive type and retraining the model each time.

Figure 4 reports precision for the full model and three other models that each exclude certain feature sets.[26] For the k people with the highest predicted risk on the x-axis, the y-axis reports the share actually victimized during the outcome period. Because noise in our precision measure increases as k, the number of people above a predicted risk threshold, decreases, we start the graph at $k = 500$. Bootstrapped 95 percent confidence intervals are plotted around each model.

Two feature sets of particular interest are those containing information about a person's own arrest history and those containing information about the arrest and victimization histories of people in a person's "network." As others have noted (e.g., Richardson et al., 2019; Lum and Isaac, 2016; Luh, 2022), arrest data contain errors, may be subject to manipulation, and are shaped in part by officer behavior. In the extreme, if arrest data provide little signal about individual behavior, then even if individual behavior plays a large role in a person's risk of shooting victimization, arrest data would provide little predictive power. Separately, Green et al. (2017) show that network information may be useful in predicting shooting victimization, particularly if gun violence propagates through a social network as people co-engage in risky behavior with their peers.

As Figure 4 shows, features related to a person's own arrests matter substantially for performance, as excluding them reduces precision by between one and three percentage points relative to the full model, depending on the rank k. To make this concrete, at $k = 4,244$, the 1.3 percentage point higher precision of the full model relative to one that excludes own arrest information, a statistically significant difference, means an additional 54 victims identified (486 versus 432). Setting aside the confidence intervals, the full model

[26] Performance measures for additional models excluding different feature sets are reported in Appendix B.5.

would identify almost 13 percent more victims than one omitting own arrest information.

The story is less clear for network features. Over most of the values of k shown in Figure 4, precision for the model that excludes network features is statistically indistinguishable from precision for the full model. For groups larger than approximately the $k = 2,500$ at highest predicted risk, the precision of both models nearly converges. Excluding network features on their own does not appear to substantially affect performance. But excluding network features *in addition to* own arrest features lowers precision by about three to four percentage points relative to the full model (equivalent to 103 more victims identified at $k = 4,244$, a 27 percent increase).[27] This pattern suggests that both a person's own arrest history *and* the arrest and victimization histories of their network neighbors contain valuable signal for predicting their shooting victimization risk. But while much of the signal contained in the network features is likely also captured by a person's own arrest history, a person's own arrest history contains substantial additional signal that is not captured by their network features.[28]

4.3.2 Modeling approach

Given the rarity of our outcome, *ex ante* we expected machine learning models to have better predictive performance than simpler linear approaches like ordinary least squares (OLS). But the difficulty of explaining what information machine learning models are using, as well as the logistical challenge of implementing them, may make linear models with fewer features a more practical choice in some settings. To help policymakers quantify the predictive power loss that comes from using simpler approaches, Table 1 reports performance differences between our full GBM model and several simpler variants. We

[27] At $k = 500$, we cannot distinguish between the performance of the "no own arrests" and "no own arrests, no network information" models. However, at $k = 4,244$, the performance differences between the two models are statistically significant.

[28] This may be partly because a person's network features are constructed using information from their own arrest history. For additional analyses exploring the sensitivity of the model's performance to feature count and modeling complexity, see Appendix B.5.3.

lack the statistical power to differentiate across models, especially for the $k = 500$ people with the highest predicted risk. But for the $k = 4,244$ people, the estimates are suggestive of the performance loss from simplifying the modeling approach, so we focus on those.

To assess how much the flexibility and non-linearity of a machine learning approach matters, the first two rows report precision and recall for our full GBM model and an OLS model using the same 1,411 features. GBM's precision is 1.1 percentage points higher at $k = 4,244$ (0.115 versus 0.104), which translates to an additional 43 victims identified.

Another way to simplify the model is to use fewer features. The next two rows in Table 1 report results for both GBM and OLS using only the 10 most predictive features (see Appendix B.5.3 for details on choosing these features). GBM's flexibility still helps here relative to OLS: the 10-feature GBM identifies 464 victims among the $k = 4,244$ people with the highest predicted risk, similar to OLS with the full feature set (443 victims), whereas OLS with only 10 features identifies only 419 victims. We can recover much (but not all) of the predictive power by making the OLS model less restrictive; when we add all two-way interactions, the 10-feature OLS model identifies 439 victims.

The results confirm that both the richness of information entering the model and the flexibility of machine learning improves performance, typically on the order of identifying about 5-10 percent more victims (at least at $k = 4,244$). This translates to a larger number of identified victims in absolute terms and so may be worth the complexity in some settings. But it is notable that OLS is able to identify about 90 percent of the victims identified by the machine learning approaches, with improved performance when using more features or two-way interactions. In settings where more complicated approaches are infeasible or undesirable, a more easily explained and implemented linear model can still help identify a population at high risk of shooting victimization.

5 Discussion

This paper demonstrates that re-purposing police data allows us to identify a small group of people at outsized risk of being shot. The immense social cost of gun violence—to victims, their families, and their communities—justifies spending a lot to reduce this risk. For example, the 500 people with the highest predicted risk represent just 0.02 percent of Chicago's population but 2.0 percent of its shooting victims over an 18-month period. This amount of gun victimization generates an estimated social cost of just over $191 million (Cook and Ludwig, 2000; Ludwig and Cook, 2001).[29] If an intervention could cut this group's risk by half, it would save $190,900 in social costs for each of the 500. The algorithm could also help target larger interventions: the 4,200 people with highest predicted risk are under 0.2 percent of Chicago's population but account for almost 12 percent of its shooting victims during the outcome period. At an estimated social cost of $1.1 billion, reducing this risk by half would save $133,071 for each of the 4,200. Even with the uncertainty inherent in estimates of gun violence's social costs, the magnitudes involved are likely to be staggering. The fact that it is possible to anticipate who so many shooting victims will be, given the huge social costs involved, is a strong argument for spending more to prevent their victimization.

Predicting shooting victimization can also be important for research aimed at identifying effective interventions. While interventions should target the people whom treatment would most benefit—those with large negative values of $Y(1) - Y(0)$—reaching participants with lower $Y(0)$ may reduce statistical power in a given sample size (or conversely, require a larger sample size to detect a given effect). For example, suppose a study sample had the same average shooting victimization risk as people above the 99$^{\text{th}}$ percentile of the full model's predicted risk distribution ($\bar{Y}(0)$ = 12.3 percent). Detecting a 50 percent

[29] These studies estimate the social cost of a gunshot injury to be $1.2 million in 1998, or $2.3 million in inflation-adjusted 2024 dollars, using a nationally representative contingent valuation survey of adults in the U.S.

reduction in shooting victimization would require an experiment with a sample size of 680.[30] In contrast, if the sample had the same average shooting victimization risk as people in the 98th-99th percentile ($\bar{Y}(0)$ = 7.9 percent), then detecting a 50 percent reduction in shooting victimization would require an experiment with a sample size of 1,092, a 61 percent increase. In a world of limited resources, better prediction of $Y(0)$ can be an important input into identifying $Y(1) - Y(0)$.

Of course, identifying those in need of prevention is only the first step. Preventing shooting victimization also requires interventions that can address why a person is at high risk of it or change something external to reduce that risk. If using a threshold rule, such as serving a small group at the highest predicted risk, consideration must also be given to that group's demographic composition when designing interventions. Research about social service interventions' effectiveness at reducing gun violence for this population is relatively limited.[31] Generating evidence about who is responsive to which kinds of prevention efforts, and how that varies across the risk distribution, is a prerequisite for any prediction method to effectively target interventions.

It is also crucial to acknowledge that even a model capable of identifying a group of people at very high risk of being shooting victims will get that prediction wrong for many—in our case, most—people in the group. And as discussed above, getting group averages right still leaves room for differential error rates by demographic group, depending on how predictions are used. Given these realities, the costs of misdirecting an intervention can vary significantly. For example, providing a slot in a social program to someone whose actual risk is much lower than predicted incurs an important opportunity cost but

[30] The calculation uses a two-sided difference of proportions test with 80 percent power and 5 percent chance of Type I error.

[31] The most well-studied model, Cure Violence, has mixed evidence of success (Butts et al., 2015; Buggs et al., 2020). Other programs providing mentorship and life coaching to those at high risk of gun violence in the community (Corburn and Fukutome-Lopez, 2020) or who are hospitalized (Cheng et al., 2008; Cooper et al., 2006; Zun et al., 2006) are being studied non-experimentally or at small scale. A preventive intervention delivered by police in Chicago to men identified by a predictive model was not found to reduce victimization (Saunders et al., 2016).

is unlikely to harm the recipient.[32] Targeting proactive policing efforts that could infringe on someone's civil liberties or perpetuate racially-discriminatory police practices in their community, on the other hand, may impose unacceptably high costs on the recipient (Stevenson and Mayson, 2022) and those around them.

There are other reasons not to use shooting victimization predictions to target proactive policing efforts. In addition to the potential legal barriers posed by using any algorithmic predictions for such targeting, these proactive policing efforts are usually designed to intervene with (and restrain) future *offenders*, not the future *victims* we seek to predict.[33] Our results provide no basis for concluding that the risks of shooting victimization and offending are interchangeable. Without a measure of true offending, we cannot assess how well predicting victimization does at predicting offending, nor whether it is more or less accurate in identifying future offenders than the status quo policing methods. This uncertainty points to an ethical challenge: it is difficult to justify targeting policing efforts—which often create large negative externalities—on the basis of shooting victimization risk, given the unclear marginal benefits and high potential costs of doing so.

Importantly, however, the results in this paper suggest that ignoring the ability of police data to predict shooting victimization altogether is not a solution; the counterfactual of *not* using information that might improve the targeting of gun violence prevention efforts carries its own cost. Current resource allocation mechanisms often rely on the staff of community violence prevention organizations, sometimes in partnership with law enforcement or hospital staff. Such individuals' social networks and expert judgment likely capture risk factors that police data miss. But they also introduce their own potential for bias and may miss high-need people whom the relevant staff do not know. Additionally,

[32] Even when a model's high predicted risk of victimization is correct, offers of preventive services made on the basis of algorithmic predictions need to be implemented carefully to avoid stigmatizing or even potentially further endangering the recipients.

[33] Using predictions based on prior police actions to justify future police actions that infringe on civil liberties may not meet the necessary legal standards. For example, an algorithmic prediction alone may not satisfy the "reasonable suspicion" standard needed for a traffic stop or the "probable cause" standard needed for an arrest, especially if police actions can deliberately elevate a person's predicted risk.

local organizations have good reason to target those who are easiest to find and least costly to serve. If the people at highest risk are also the hardest to identify and serve, then algorithms may be an effective way to direct potentially life-saving services toward those who might not otherwise receive them.

One example of how algorithmic prediction can be used to direct gun violence prevention services is READI Chicago (Bhatt et al., 2024). In that setting, a predictive model closely related to the one studied here identified men at very high risk of involvement in future gun violence. Publicly available information about them was provided to community violence prevention organizations, who offered the men a chance to voluntarily participate in an intervention designed to reduce their risk. No information about them was shared with law enforcement. Other men who could benefit from the intervention were identified by the outreach workers at the community organizations themselves, or by jail, prison, and parole staff. Crucially, this approach was developed in consultation with people who live in the affected communities.

Ex post, outreach referrals were more responsive to the intervention than algorithm referrals. But the evidence does not suggest that this heterogeneity is because outreach workers focused on $Y(1) - Y(0)$. Rather, they seemed to select participants based largely on the likelihood of taking up services. In fact, the bulk of outreach referrals were not responsive to READI; only the subset who were *also* predicted to be at high risk by the algorithm drove the decline in serious violence. In this way, the model ended up being a complement to, rather than a substitute for, human expertise. It helped find people who might benefit from programming but who would not otherwise be found, and when combined with the unobservables humans used to make referrals, it helped identify program responsiveness.

The key insight of this paper is that an algorithm using police data—which are readily available in most cities—to predict a well-measured outcome can be a useful tool for aiding efforts to prevent morbidity and mortality from gun violence. Training the algo-

rithm to predict shooting victimization rather than arrest makes it likelier to predict the outcome of interest, rather than whom police decide to arrest (Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021). We note that the algorithm's ability to predict shooting victimization occurs in a social context where law enforcement is often the primary state institution enmeshed in the lives of Black men at high risk of gun violence. In a different context, where other government agencies and non-profit organizations more extensively engage with people facing such risks, there will likely be other information available to help target preventive services. Shifting toward this context could have a number of benefits, including reducing the social costs of excessive police contact (e.g., Pager, 2003; Harris, 2016; Mello, 2021; Agan and Starr, 2017). Until then, a small group of people face an extraordinarily high risk of being shot, with few systematic ways to identify them available. We demonstrate that it is currently possible for an algorithm to predict shooting victimization well enough to help direct and test services that might save lives.

References

Agan, Amanda and Sonja B. Starr (2017). "The effect of criminal records on access to employment." *American Economic Association: Papers & Proceedings* 107.5, pp. 560–564.

Ang, Desmond (2021). "The Effects of Police Violence on Inner-City Students." *Quarterly Journal of Economics* 136.1, pp. 115–168.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016). "Machine Bias." *ProPublica*. URL: `https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing`.

Antonovics, Kate and Brian G. Knight (2009). "A new Look at racial profiling: Evidence from the Boston Police Department." *Review of Economics and Statistics* 91.1, pp. 163–177.

Arnold, David, Will Dobbie, and Crystal S. Yang (2018). "Racial Bias in Bail Decisions." *The Quarterly Journal of Economics* 133.4, pp. 1885–1932.

Ascarza, Eva (2018). "Retention futility: Targeting high-risk customers might be ineffective." *Journal of Marketing Research* 55.1, pp. 80–98.

Athey, Susan (2017). "Beyond prediction: Using big data for policy problems." *Science* 355, pp. 483–485.

Barber, Catherine, Philip J. Cook, and Susan T. Parker (2022). "The emerging infrastructure of US firearms injury data." *Preventive Medicine* 165, p. 107129.

Berk, Richard (2008). "Forecasting Methods in Crime and Justice." *Annual Review of Law and Social Science* 4, pp. 219–238.

Berk, Richard, Lawrence Sherman, Geoffrey Barnes, Ellen Kurtz, and Lindsay Ahlman (2009). "Forecasting murder within a population of probationers and parolees: a high stakes application of statistical learning." *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 172.1, pp. 191–211.

Bhatt, Monica P., Sara B. Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman (2024). "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago." *The Quarterly Journal of Economics* 139 (1) (1), pp. 1–56.

Breiman, Leo (2001). "Random forests." *Machine Learning* 45, pp. 5–32.

Buggs, Shani A., Daniel W. Webster, and Cassandra K. Crifasi (2020). "Using synthetic control methodology to estimate effects of a Cure Violence intervention in Baltimore, Maryland." *Injury Prevention*, pp. 1–7.

Butts, Jeffrey A., Caterina Gouvis Roman, Lindsay Bostwick, and Jeremy R. Porter (2015). "Cure Violence: A Public Health Model to Reduce Gun Violence." *Annual Review of Public Health* 36, pp. 39–53.

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina (2008). "An empirical evaluation of supervised learning in high dimensions." *Proceedings of the 25th International Conference on Machine Learning*, pp. 96–103.

Chalfin, Aaron, Benjamin Hansen, Emily K. Weisburst, Morgan C. Williams, et al. (2022). "Police Force Size and Civilian Race." *American Economic Review: Insights*.

Chandler, Dana, Steven D. Levitt, and John A. List (2011). "Predicting and Preventing Shootings among At-Risk Youth." *American Economic Review: Papers & Proceedings* 101.3, pp. 288–292.

Cheng, Tina L., Denise Haynie, Ruth Brenner, Joseph L. Wright, Shang En Chung, and Bruce Simons-Morton (2008). "Effectiveness of a mentor-implemented, violence prevention intervention for assault-injured youths presenting to the emergency department: Results of a randomized trial." *Pediatrics* 122.5, pp. 938–946.

Chouldechova, Alexandra (2017). "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments." *Big Data* 5.2, pp. 153–163.

Chouldechova, Alexandra, Diana Benavides-Prado, Oleksandr Fialko, and Rhema Vaithianathan (2018). "A case study of algorithm-assisted decision making in child maltreatment hotline screening decisions." *Conference on Fairness, Accountability and Transparency*. PMLR, pp. 134–148.

Lo-Ciganic, Wei-Hsuan, James L. Huang, Hao H. Zhang, Jeremy C. Weiss, Yonghui Wu, C. Kent Kwoh, Julie M. Donohue, Gerald Cochran, Adam J. Gordon, Daniel C. Malone, et al. (2019). "Evaluation of machine-learning algorithms for predicting opioid overdose risk among medicare beneficiaries with opioid prescriptions." *JAMA Network Open* 2.3.

City of Baltimore (2021). *Violence Prevention Framework and Plan*. URL: www.monse.baltimorecity.gov/sites/default/files/MayorBMS_Draft_ViolenceReductionFrameworkPlan.pdf.

City of Chicago (2020). *Our City, Our Safety: A Comprehensive Plan to Reduce Violence in Chicago*. URL: https://www.chicago.gov/city/en/depts/mayor/press_room/press_releases/2020/september/ComprehensiveViolenceReductionPlan.html.

City of Oakland (2024). *Oakland's Ceasefire Strategy*. URL: https://www.oaklandca.gov/topics/oaklands-ceasefire-strategy.

City of Philadelphia (2021). *Philadelphia Roadmap to Safer Communities: Spring 2021 Update*. URL: www.phila.gov/media/20210414123750/RoadmapToSaferCommunitiesSpring2021.pdf.

Cook, Philip J. and Jens Ludwig (2000). *Gun Violence: The Real Costs*. New York: Oxford University Press.

Cooper, Carnell, Dawn M. Eslinger, and Paul D. Stolley (2006). "Hospital-based violence intervention programs work." *Journal of Trauma - Injury, Infection and Critical Care* 61.3, pp. 534–537.

Corbett-Davies, Sam, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq (2017). "Algorithmic decision making and the cost of fairness." *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pp. 797–806.

Corburn, Jason and Amanda Fukutome-Lopez (2020). *City of Sacramento/Advance Peace Sacramento Youth Peacemaker Fellowship Program CalVIP, BSCC Final Local Evaluation Report*. Tech. rep.

Eberhardt, Jennifer L., Phillip Atiba Goff, Valerie J. Purdie, and Paul G. Davies (2004). "Seeing black: race, crime, and visual processing." *Journal of Personality and Social Psychology* 87.6, p. 876.

Eichstaedt, Johannes C, Robert J Smith, Raina M Merchant, Lyle H Ungar, Patrick Crutchley, Daniel Preoţiuc-Pietro, David A Asch, and H Andrew Schwartz (2018). "Facebook language predicts depression in medical records." *Proceedings of the National Academy of Sciences* 115.44, pp. 11203–11208.

Farrington, David P., Hannah Gaffney, and Maria M. Ttofi (2017). "Systematic reviews of explanatory risk factors for violence, offending, and delinquency." *Aggression and Violent Behavior* 33, pp. 24–36.

Fogliato, Riccardo, Alexandra Chouldechova, and Max G'Sell (2020). "Fairness evaluation in presence of biased noisy labels." *International conference on artificial intelligence and statistics*. PMLR, pp. 2325–2336.

Friedman, Jerome H. (2001). "Greedy function approximation: a gradient boosting machine." *Annals of statistics*, pp. 1189–1232.

Garza, Ángel Garcıa de la, Carlos Blanco, Mark Olfson, and Melanie M Wall (2021). "Identification of suicide attempt risk factors in a national US survey using machine learning." *JAMA Psychiatry* 78.4, pp. 398–406.

Geller, Amanda, Jeffrey Fagan, Tom Tyler, and Bruce G. Link (2014). "Aggressive policing and the mental health of young urban men." *American Journal of Public Health* 104.12, pp. 2321–2327.

Goncalves, Felipe and Steven Mello (2021). "A Few Bad Apples? Racial Bias in Policing." *American Economic Review* 111.5, pp. 1406–1441.

Green, Ben, Thibaut Horel, and Andrew V. Papachristos (2017). "Modeling contagion through social networks to explain and predict gunshot violence in Chicago, 2006 to 2014." *JAMA Internal Medicine* 177.3, pp. 326–333.

Hanson, R. Karl (2005). "Twenty years of progress in violence risk assessment." *Journal of Interpersonal Violence* 20.2, pp. 212–217.

Harris, Alexes (2016). *A Pound of Flesh: Monetary Sanctions as Punishment for the Poor*. American Sociological Association's Rose Series. Russell Sage Foundation.

Hastings, Justine S., Mark Howison, and Sarah E. Inman (2020). "Predicting high-risk opioid prescriptions before they are given." *Proceedings of the National Academy of Sciences of the United States of America* 117.4, pp. 1917–1923.

Haushofer, Johannes, Paul Niehaus, Carlos Paramo, Edward Miguel, and Michael W. Walker (2022). "Targeting impact versus deprivation." *NBER Working Paper No. 30138*.

Hawkins, J. David, Todd Herrenkohl, David P. Farrington, Devon Brewer, Richard F. Catalano, and Tracy W. Harachi (1998). "A review of predictors of youth violence."

Heller, Sara B. (2022). "When scale and replication work: Learning from summer youth employment experiments." *Journal of Public Economics* 209, p. 104617.

Hoekstra, Mark and CarlyWill Sloan (2022). "Does race matter for police use of force? Evidence from 911 calls." *American Economic Review* 112.3, pp. 827–60.

Japkowicz, Nathalie (2000). "The Class Imbalance Problem: Significance and Strategies." *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, pp. 111–117.

Jones, Nikki (2014). ""The Regular Routine": Proactive Policing and Adolescent Development Among Young, Poor Black Men." *New Directions for Child and Adolescent Development* 143, pp. 33–54.

Kapustin, Max, Jens Ludwig, Marc Punkay, Kimberley Smith, Lauren Speigel, and David Welgus (2017). "Gun violence in Chicago, 2016." *University of Chicago Crime Lab*.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan (2018a). "Human Decisions and Machine Predictions." *The Quarterly Journal of Economics* 133.January, pp. 237–293.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan (2018b). "Algorithmic Fairness." *American Economic Review: Papers & Proceedings* 108, pp. 22–27.

Kleinberg, Jon, Sendhil Mullainathan, and Manish Raghavan (2017). "Inherent Trade-Offs in the Fair Determination of Risk Scores." *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lei, Jing, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman (2018). "Distribution-Free Predictive Inference for Regression." *Journal of the American Statistical Association* 113.523, pp. 1094–1111.

Liebschutz, Jane, Sonia Schwartz, Joel Hoyte, Lauren Conoscenti, Anthony B. Christian Sr., Leroy Muhammad, Derrick Harper, and Thea James (2010). "A chasm between injury and care: experiences of black male victims of violence." *The Journal of Trauma* 69.6, p. 1372.

Ludwig, Jens and Philip J. Cook (2001). "The Benefits of Reducing Gun Violence: Evidence from Contingent-Valuation Survey Data." *The Journal of Risk and Uncertainty* 22.3, pp. 207–226.

Luh, Elizabeth (2022). "Not So Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports." URL: https://papers.ssrn.com/sol3/Papers.cfm?abstract_id=3357063.

Lum, Kristian and William Isaac (2016). "To predict and serve?" *Significance* 13.5, pp. 14–19.

May, John P., David Hemenway, and Alicia Hall (2002). "Do criminals go to the hospital when they are shot?" *Injury Prevention* 8.3, pp. 236–238.

Mayson, Sandra G. (2019). "Bias In, Bias Out." *Yale Law Journal* 128.8, pp. 2122–2473.

Mehrabi, Ninareh, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan (2021). "A Survey on Bias and Fairness in Machine Learning." *ACM Computing Surveys* 54.6, pp. 1–35.

Mello, Steven (2021). "Fines and Financial Wellbeing." URL: https://mello.github.io/files/fines.pdf.

Mullainathan, Sendhil and Ziad Obermeyer (2021). "On the Inequity of Predicting A While Hoping for B." *American Economic Association: Papers & Proceedings* 111, pp. 37–42.

Obermeyer, Ziad and Ezekiel J. Emanuel (2016). "Predicting the Future - Big Data, Machine Learning, and Clinical Medicine." *New England Journal of Medicine* 375.13, pp. 1212–1216.

Obermeyer, Ziad, Brian Powers, Christine Vogeli, and Sendhil Mullainathan (2019). "Dissecting racial bias in an algorithm used to manage the health of populations." *Science* 366.6464, pp. 447–453.

Otto, Randy K. and Kevin S. Douglas (2010). *Handbook of Violence Risk Assessment*. International perspectives on forensic mental health. Routledge. ISBN: 9780415962148.

Pager, Devah (2003). "The Mark of a Criminal Record." *American Journal of Sociology* 108.5, pp. 937–975.

Papachristos, Andrew V. and Sara Bastomski (2018). "Connected in crime: the enduring effect of neighborhood networks on the spatial patterning of violence." *American Journal of Sociology* 124.2, pp. 517–568.

Papachristos, Andrew V., Anthony A. Braga, and David M. Hureau (2012). "Social networks and the risk of gunshot injury." *Journal of Urban Health* 89.6, pp. 992–1003.

Papachristos, Andrew V., Anthony A. Braga, Eric Piza, and Leigh S. Grossman (2015a). "The Company You Keep? The Spillover Effects of Gang Membership on Individual Gunshot Victimization." *Criminology* 53.4, pp. 624–649.

Papachristos, Andrew V. and Christopher Wildeman (2014). "Network exposure and homicide victimization in an African American community." *American Journal of Public Health* 104.1, pp. 143–150.

Papachristos, Andrew V., Christopher Wildeman, and Elizabeth Roberto (2015b). "Tragic, but not random: The social contagion of nonfatal gunshot injuries." *Social Science and Medicine* 125, pp. 139–150.

Qi, Di and Andrew J. Majda (2020). "Using machine learning to predict extreme events in complex systems." *Proceedings of the National Academy of Sciences* 117.1, pp. 52–59.

Rehavi, M. Marit and Sonja B. Starr (2014). "Racial disparity in federal criminal sentences." *Journal of Political Economy* 122.6, pp. 1320–1354.

Richardson, Rashida, Jason M. Schultz, and Kate Crawford (2019). "Dirty Data, Bad Predictions: How Civil Rights Violations Impact Police Data, Predictive Policing systems, and Justice." *New York University Law Review* 94.2, pp. 192–233.

Salganik, Matthew J. et al. (2020). "Measuring the predictability of life outcomes with a scientific mass collaboration." *Proceedings of the National Academy of Sciences* 117.15, pp. 8398–8403.

Saunders, Jessica, Priscillia Hunt, and John S. Hollywood (2016). "Predictions put into practice: a quasi-experimental evaluation of Chicago's predictive policing pilot." *Journal of Experimental Criminology* 12.3, pp. 347–371.

Sharkey, Patrick (2018). "The long reach of violence: A broader perspective on data, theory, and evidence on the prevalence and consequences of exposure to violence." *Annual Review of Criminology* 1, pp. 85–102.

Shwartz-Ziv, Ravid and Amitai Armon (2022). "Tabular data: Deep learning is not all you need." *Information Fusion* 81, pp. 84–90.

Singh, Jay P., Martin Grann, and Seena Fazel (2011). "A comparative study of violence risk assessment tools: A systematic review and metaregression analysis of 68 studies involving 25,980 participants." *Clinical Psychology Review* 31.3, pp. 499–513.

Starr, Sonja B. (2014). "Evidence-based sentencing and the scientific rationalization of discrimination." *Stanford Law Review* 66, p. 803.

Stevenson, Megan T and Sandra G Mayson (2022). "Pretrial detention and the value of liberty." *Virginia Law Review* 108.3, pp. 709–782.

Tahamont, Sarah, Zubin Jelveh, Aaron Chalfin, Shi Yan, and Benjamin Hansen (2021). "Dude, where's my treatment effect? Errors in administrative data linking and the destruction of statistical power in randomized experiments." *Journal of Quantitative Criminology* 37, pp. 715–749.

Toloşi, Laura and Thomas Lengauer (2011). "Classification with correlated features: unreliability of feature ranking and solutions." *Bioinformatics* 27.14, pp. 1986–1994.

Washington, D.C. (2022). *People of Promise*. URL: https://onse.dc.gov/service/people-promise.

Wernick, Miles N. (2018). "A Data-Driven Crime Prevention Program."

Wheeler, Andrew P., Robert E. Worden, and Jasmine R. Silver (2019). "The accuracy of the violent offender identification directive tool to predict future gun violence." *Criminal Justice and Behavior* 46.5, pp. 770–788.

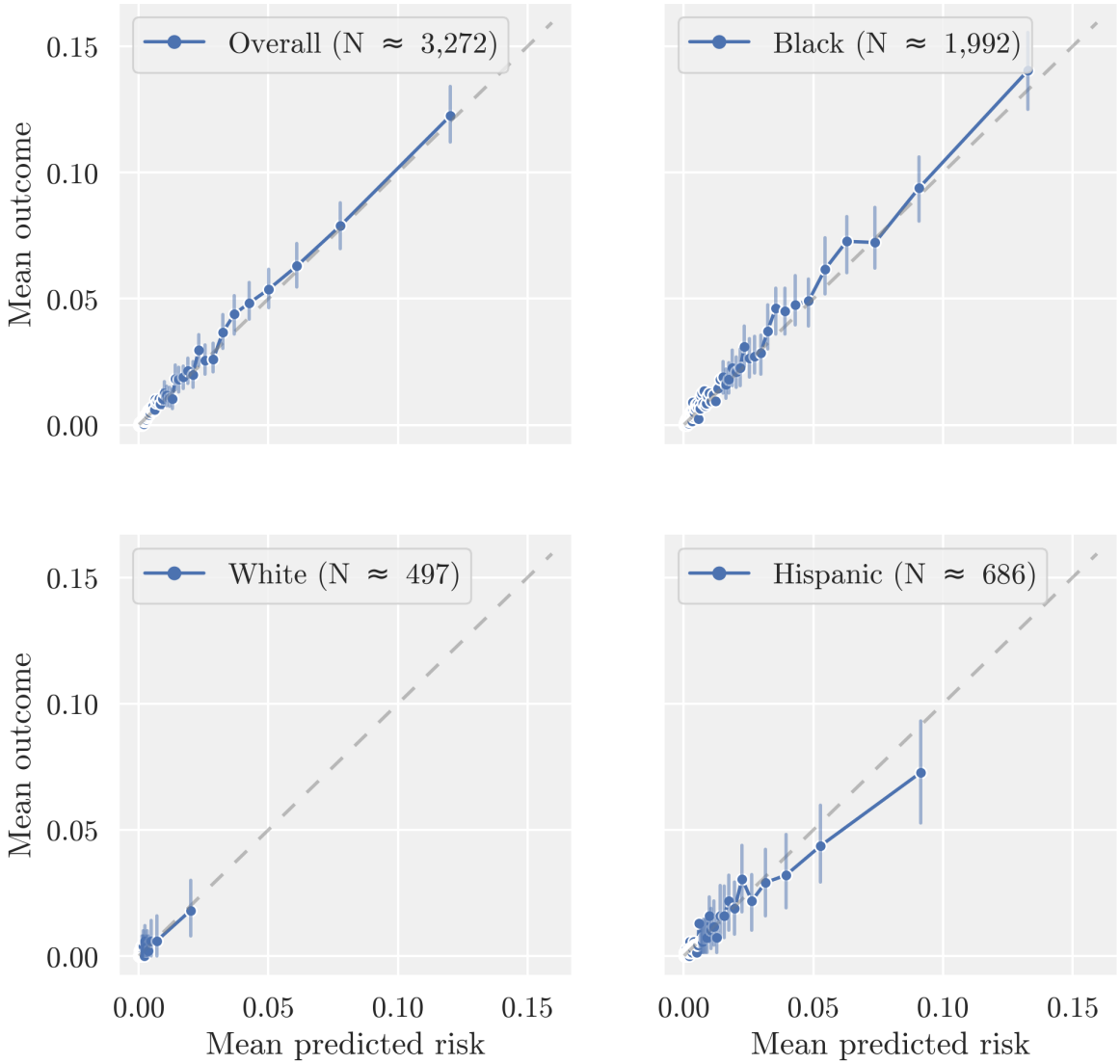
White, Kailey, Philip J. Cook, and Harold A. Pollack (2021). "Gunshot-victim cooperation with police investigations: results from the Chicago inmate survey." *Preventive Medicine* 143, p. 106381.

Wood, George and Andrew V. Papachristos (2019). "Reducing gunshot victimization in high-risk social networks through direct and spillover effects." *Nature Human Behaviour* 3.11, pp. 1164–1170.

Yang, Crystal S. and Will Dobbie (2020). "Equal protection under algorithms: A new statistical and legal framework." *Michigan Law Review* 119, p. 291.

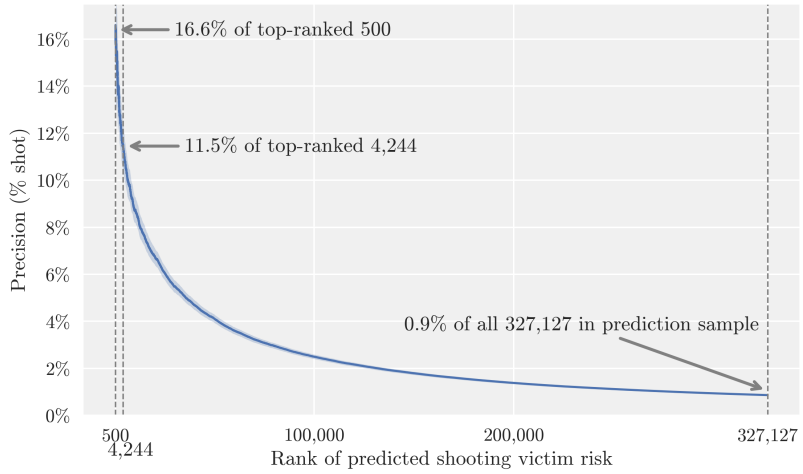
Zun, Leslie S., La Vonne Downey, and Jodi Rosen (2006). "The effectiveness of an ED-based violence prevention program." *American Journal of Emergency Medicine* 24.1, pp. 8–13.

Figure 1: Predicted versus actual risk of shooting victimization by bin (calibration), overall and by race/ethnicity

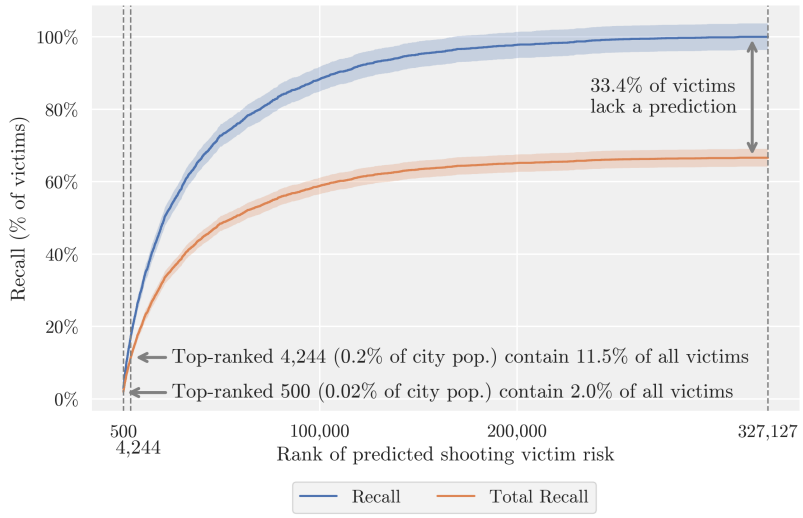


Note: Figure shows mean predicted shooting victimization risk and shooting victimization rate within each percentile of the overall (top left panel) and race/ethnicity-specific (remaining panels) predicted risk distributions. Race/ethnicity categories are mutually exclusive: non-Hispanic White, non-Hispanic Black, and Hispanic of any race. Bootstrapped 95 percent confidence intervals shown (see Appendix A.5.2 for details).

Figure 2: Predictive performance for shooting victimization



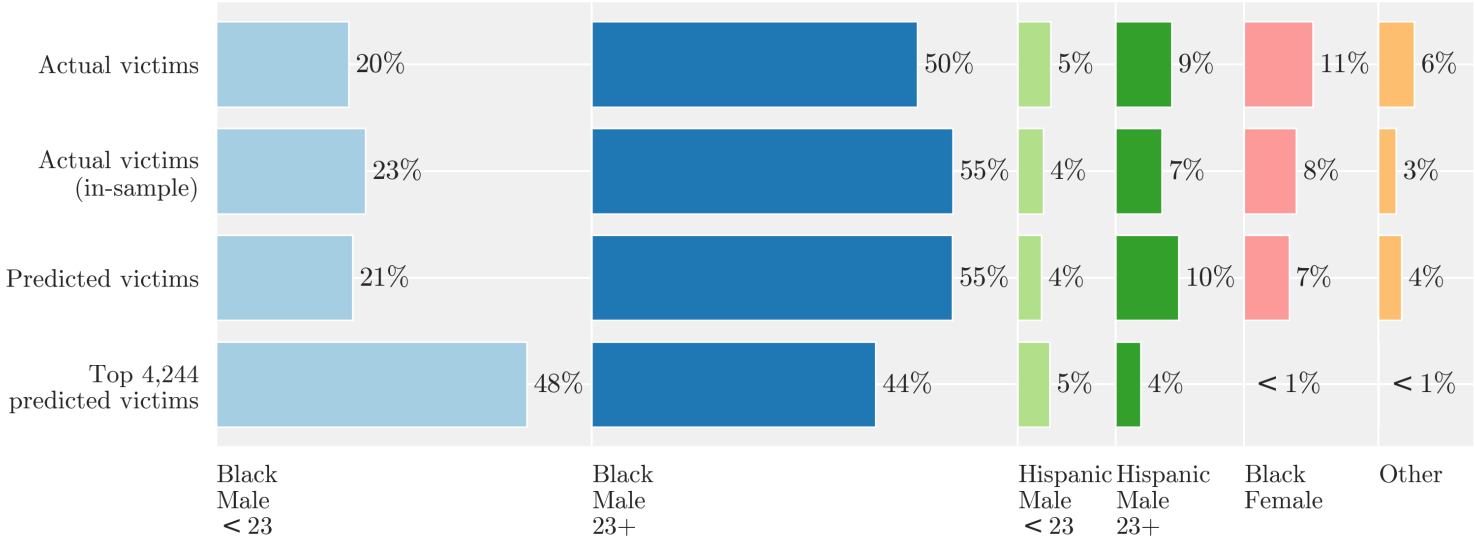
(a) Precision



(b) Recall

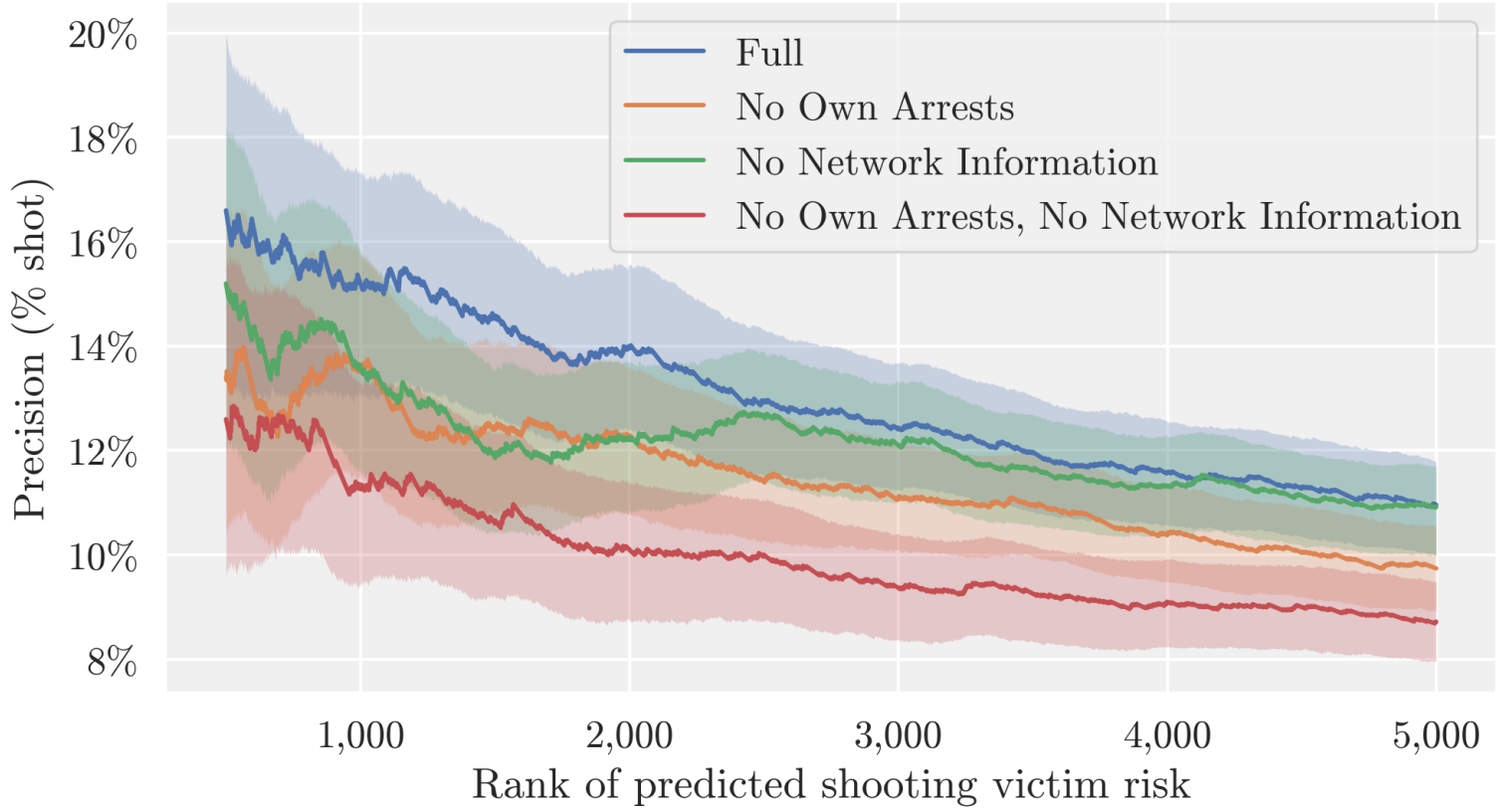
Note: Performance of the full model during the 18-month outcome period starting April 1, 2018. Precision shows the share of the k people with the highest predicted risk of shooting victimization who are shot during the outcome period. Recall shows the share of all 2,827 shooting victims in the prediction sample during the outcome period who are among the k people with highest predicted risk. Total recall shows the share of all 4,244 shooting victims in Chicago during the outcome period who are among the k people with highest predicted risk. Bootstrapped 95 percent confidence shown (see Appendix A.5.2 for details).

Figure 3: Demographic composition across victim groups



Note: Figure reports the proportion of each row in the indicated demographic category, with rows showing all actual shooting victims, those in the prediction sample, predicted shooting victims, and the 4,244 people with the highest predicted risk of victimization. To reduce visual clutter, demographic groups accounting for very small shares of actual and predicted victims—Hispanic women, White men, White women, people with missing race/ethnicity or gender information, and Black or Hispanic men with missing age information—are combined in the "Other" category. The demographic shares for predicted shooting victims (third row) are based on the 327,181 people in the prediction sample reweighted by their predicted risk of victimization (see Appendix A.5.1 for details).

Figure 4: Precision across models with different feature sets



Note: Figure shows precision, or the share of the $k \leq 5,000$ people with the highest predicted risk of shooting victimization who are shot during the 18-month outcome period, for models trained with different feature sets. Due to noise in precision at low values of k, we start the graph at $k = 500$. Bootstrapped 95 percent confidence intervals shown (see Appendix A.5.2 for details).

Table 1: Comparison across modeling approaches

Model	Top 500				Top 4,244			
	True Positives	Precision	Recall	Total Recall	True Positives	Precision	Recall	Total Recall
Full GBM	83	0.166	0.029	0.020	486	0.115	0.172	0.115
	(67, 100)	(0.134, 0.200)	(0.024, 0.035)	(0.016, 0.024)	(444, 527)	(0.105, 0.124)	(0.157, 0.186)	(0.105, 0.124)
Full OLS	74	0.148	0.026	0.017	443	0.104	0.157	0.104
	(58, 90)	(0.116, 0.180)	(0.021, 0.032)	(0.014, 0.021)	(404, 484)	(0.095, 0.114)	(0.143, 0.171)	(0.095, 0.114)
Simple GBM	83	0.166	0.029	0.020	464	0.109	0.164	0.109
	(65, 97)	(0.130, 0.194)	(0.023, 0.034)	(0.015, 0.023)	(424, 504)	(0.100, 0.119)	(0.150, 0.178)	(0.100, 0.119)
Simple OLS	63	0.126	0.022	0.015	419	0.099	0.148	0.099
	(48, 77)	(0.096, 0.154)	(0.017, 0.027)	(0.011, 0.018)	(381, 459)	(0.090, 0.108)	(0.135, 0.162)	(0.090, 0.108)
Simple OLS	71	0.142	0.025	0.017	439	0.103	0.155	0.103
Interactions	(57, 87)	(0.114, 0.174)	(0.020, 0.031)	(0.013, 0.020)	(399, 478)	(0.094, 0.113)	(0.141, 0.169)	(0.094, 0.113)

Note: GBM refers to our preferred gradient boosting machine model; OLS refers to an ordinary least squares model. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details). Full models use all 1,411 features. Simple models use the 10 most independently predictive features (see Appendix B.5.3 for details on how these features are selected). OLS with interactions includes all two-way interactions of the 10 most independently predictive features.

Online Appendix for
"Machine Learning Can Predict Shooting Victimization
Well Enough to Help Prevent It"

Sara B. Heller, Benjamin Jakubowski, Zubin Jelveh & Max Kapustin

A Methods

This section provides additional details regarding our modeling process. First, we describe the raw Chicago Police Department (CPD) data, as well as the record linkage algorithm used to identify unique people across records. Next, we discuss the features (predictors) generated from these records. Finally, we discuss model training, parameter selection, and performance measures.

A.1 Data

Our model predicts a person's likelihood of being reported in the CPD data as being injured or killed by gunfire (reported shooting victimization) in the 18 months following the prediction date. To do this, we use information from 12.7 million CPD records that broadly fall into two categories: suspected offender and victim. These records are used to identify unique people in the record linkage process (Appendix A.2), generate predictive features (Appendix A.3), and construct outcomes. The following sections describe the relevant attributes of each record type and how they are used.

A.1.1 Suspected offender records

The suspected offender records contain information about people suspected of having committed criminal offenses and details about those offenses.[34] The main set of suspected offender records we use are arrest records between January 1999 and October 2019. The arrest records contain several types of identifying information about the person arrested that we use in the record linkage process, including name, date of birth, and home address. Most important for record linkage, the arrest records contain a unique person identifier called an Illinois Record (IR) number, which is based on a fingerprint scan. We use the IR number to construct a person's entire CPD arrest history.

We also extensively use the arrest records to generate predictive features. The first and most direct way that we do this is by generating features using the information contained in

[34] The CPD data do not contain information on the final dispositions of these cases, so we do not know which people were subsequently convicted.

these records about arresting charges,[35] including charge descriptions and Uniform Crime Reporting (UCR) codes; the location and time of the incident and the arrest; demographics of the arrestee; and information about whether the arrest was gang-related (and, if so, the arrestee's CPD-identified gang affiliation). The second way that we use the arrest records to generate predictive features is by using the unique incident identifier they contain to link together the arrestees and victims associated with a single incident, allowing us to identify a person's network connections (see Appendix A.3.4).

In addition to the arrest records, we use a smaller set of "homicide offender" records that contain similar identifying information (IR number, name, date of birth, home address) about people arrested for, or suspected of having committed (but who have not been arrested for), homicide. These records are not used to generate features, but they provide additional information that helps us refine the record linkage process.

A.1.2 Victim records

The victim records contain information about people reported as victims of a crime to the police. These records are generated either through victim self-reporting, third-party reporting (i.e., by healthcare providers for non-fatal shooting victims), or through police discovery (i.e., homicide victims).

The main set of general victim records we use spans January 1999 to October 2019. These records contain information about victimization incidents used to generate predictive features, such as detail about the incident type (including a description and UCR code), the location and time when the incident occurred, and the incident's unique identifier.

There are two limitations of the general victim records. First, relative to the arrest records, the general victim records contain a more limited set of information that can be used to identify the victim. Most notably, the general victim records do not include the unique person identifier included in the arrest records (IR number). This means we cannot construct a person's entire reported victimization history with the same degree of accuracy as we can construct a person's entire arrest history. Instead, the general victim records contain fields like name, home address, and date of birth, which support probabilistic record linkage (Appendix A.2). However, the reliability of the information recorded in these fields is uncertain and some fields have a high rate of missingness, though this varies over time. For example, date of birth information is missing for 73 percent of victim records since January 1999 and 47 percent of records since January 2011. We discuss the

[35] Arrests are associated with one or more arresting charge, and our arrest features consider the full set of charges on the arrest.

implications for probabilistic matching below.

Second, the information about incident type contained in the general victim records is insufficient to reliably identify shooting victimizations. Consider non-fatal shootings, which are most likely to be identified as cases of "aggravated battery with a firearm." Using this definition can lead to both false negative and false positive classifications of non-fatal shootings. For example, an incident in which a person sustained a gunshot injury during a robbery might be recorded as an armed robbery (false negative), while an incident in which a person was physically injured by an offender who was wielding (but did not fire) a gun might be recorded as an aggravated battery with a firearm (false positive).

As a result, for our shooting outcomes we rely on a separate dataset that CPD maintains containing records of both fatal and non-fatal shoot victimizations ("shooting victim records"). These shooting victim records allow us to overcome the second limitation of the general victim records by identifying victimizations in which a person was injured or killed by gunfire. The shooting victim records also contain name and date of birth information with almost no missingness, supporting more reliable probabilistic matching. We use these shooting victim records to construct our main outcome: reported shooting victimization. One limitation of the shooting victim records is that they only start in January 2011.

Finally, CPD also maintains a separate dataset containing homicide victim records. Relative to the general and shooting victim records, a key advantage of the homicide victim records is that approximately 80 percent of them contain an IR number, the same unique person identifier included in the arrest records. We use the homicide victim records to refine the record linkage process. In addition, we use the general, shooting, and homicide victim records to construct the secondary "Violent Crime Victim" outcome that we report in Appendix B.3.

A.1.3 Reporting of shooting victimizations

A key argument we make is that police records likely capture the vast majority of shooting victimizations in Chicago. As a result, reported shooting victimizations are likely to measure actual victimizations consistently across demographic groups (and certainly more consistently than shooting arrests are to measure offenses). Predicting an outcome that is consistently measured across demographic groups in this way avoids "target variable bias" (Fogliato et al., 2020) and allows us to accurately recover estimates of risk at the group level.

This argument could be undermined if there is a substantial gap between the actual

Table A.1: Comparison of shooting victim counts by data source

	Shooting Victims					
	Gun Violence Archive			Chicago Police Department		
Year	Total	Fatal	Non-Fatal	Total	Fatal	Non-Fatal
2014	2,291	416	1,875	2,558	365	2,193
2015	2,824	454	2,370	2,922	432	2,490
2016	3,597	599	2,998	4,273	695	3,578
2017	3,357	590	2,767	3,384	604	2,780
2018	2,906	478	2,428	2,876	481	2,395
2019	2,641	473	2,168	2,639	446	2,193

Note: Annual counts of fatal and non-fatal shooting victims. Data from the Gun Violence Archive are available from 2014 onward. Data from the Chicago Police Department are from the City of Chicago's Violence Reduction Dashboard: `https://www.chicago.gov/city/en/sites/vrd/home.html`.

number of shooting victimizations and the number captured in police records. If such a gap exists, one might worry that the likelihood of a shooting victimization being captured in police records could vary systematically with the victim's demographic group. For example, non-fatal shooting victims in some demographic groups might be less likely to contact the police or seek medical care that would lead to third-party reporting. Even when a non-fatal shooting victim seeks medical care, if they do so outside of Chicago—which may be likelier if they live near the city's border—then their injury may not be reported to the CPD.[36]

Two pieces of empirical evidence suggest that there is actually relatively little under-reporting of shooting victimizations in the CPD data. First, we compare counts of fatal and non-fatal shooting victims in the CPD data to those in data from the Gun Violence Archive (GVA), an independent group that collects information on shootings from both law enforcement and news sources. Appendix Table A.1 below reports these counts annually, from 2014 (the earliest year of available GVA data) to 2019 (the last year of data used in our analysis). Before 2018, the number of shooting victims was actually greater in the CPD data than in the GVA data. In 2018 and 2019, this pattern reversed, with a slightly greater number of shooting victims in the GVA data than in the CPD data (but never more than a 1 percent difference). To the extent that the pattern since 2018 is due to under-reporting, the magnitude of this under-reporting is minimal.

Second, the rate of any under-reporting of non-fatal shooting victimizations appears, if anything, to be smaller in Chicago than in other cities. We can see this in the case-

[36] Based on our conversations with practitioners, and consistent with the empirical evidence below, we think the magnitude of such selective under-reporting is likely to be quite small.

fatality rate, or the ratio of fatal to all reported shooting victimizations. Assuming a similar lethality of gun assault injuries across cities (and that basically all homicides are known to the police), the higher a city's case-fatality rate is, the more its non-fatal shooting victimizations are being under-reported. The national case-fatality rate, derived from vital statistics and hospital data, is 22 percent (Cook et al., 2017). Chicago's case-fatality rate in recent years, as calculated from the data reported in Appendix Table A.1, ranges from 14 to 17 percent. This suggests that Chicago may have a lower rate of under-reporting of non-fatal shooting victimizations than the national average, possibly due to the mandatory reporting of firearm injuries to the police in Illinois (20 ILCS 2630/3.2).

A.2 Record linkage

As mentioned above, while CPD arrest records include a unique person identifier (IR number), with the exception of homicide victim records, most victim records do not. As such, we use a probabilistic record linkage algorithm to associate unique individuals with all of their records across the CPD data. For details on the algorithm itself, see McNeill and Jelveh (2021). In this section, we describe the basics of the linking procedure.

To link CPD records that refer to the same person, we take the IR number from 2010 onward as ground truth, allowing us to identify the set of unique individuals arrested during the study period and to associate these individuals with their arrest records.[37] Since records are already linked within the arrest data, probabilistic record linkage primarily allows us to address two remaining data challenges: associating arrested individuals with their victim records, and identifying unique individuals across the victim records who did not experience a CPD arrest during the study period.

Our record linkage algorithm produces a collection of records referring to the same person, which we call a *cluster*. In assigning records to clusters, the algorithm follows researcher-specified rules based on the context of the data. For our linkage, we specify the following constraints. First, a cluster can have at most one IR number from 2010 onward. Second, a homicide victim record cannot link to another record if the homicide record's event date came before the other record's event date. Third, while virtually all shooting victim records have date of birth information, 73 percent of the general victim records do not have date of birth information—an important predictor of true positive links—which can lead to a large number of false positive links. To reduce the chance of false positives, we introduce a constraint that if at least one record in a record pair is missing date of

[37] The consistency of IR numbers is somewhat spotty at the beginning of the records but improves considerably over time. As such, we do not treat IR numbers prior to 2010 as ground truth.

birth information, then enforce that the age field (if not missing) in the two records is within 3 years. We also enforce that if at least one record in a potential cluster is missing date of birth information, then all other records in the cluster not missing date of birth information must have similar dates of birth.[38]

The record linkage procedure identifies 3,263,111 people (clusters) across the two decades of our data. We filter the set of clusters to exclude two sets of people: those with no CPD records from the past 50 months relative to a given prediction date, and those with only a single victimization in the past 50 months. We exclude these people for two reasons. First, they have much lower baseline risk: 0.05 percent of them were shot during the follow-up period, compared to 0.9 percent of people in our test set. Second, because many victim records do not have date of birth information and are therefore more likely to incorrectly link to another record, dropping clusters with just a single victimization reduces the influence of record-linkage error caused by missing data. As such, our sample inclusion criterion reduces data integrity issues while still capturing most of the identifiable population with elevated risk. This filtering removes 1,804,430 people with no arrests or victimizations during the 50 months prior to a prediction date, and 813,601 people with a single victimization in that period. We further drop 1,105 people who, for the earliest cohort in which they meet the inclusion criterion, were homicide victims during the 50 months prior to the prediction date. This leaves us with 643,975 people whose records we use to train and test our model.

A.3 Feature generation

Record linkage identifies the set of unique people represented in the CPD data, and associates each person with their CPD arrest, homicide offender, victim, shooting victim, and homicide victim records. To predict a person's risk of being shot as of a given prediction date, we aggregate over these associated records to construct features at the person-prediction date level.[39] We construct four broad types of features: demographic, arrest, victimization, and network features. Appendix Table A.2 provides a summary of this final feature set, described by type below.

When a person has no data in either the arrest or the victim records, we assign a count of 0 to each relevant set of features. For the time-since features, which are not counts, we assign a missing value to the relevant features rather than a 0, and program the LightGBM

[38] We operationalize this by enforcing that these dates of birth be within two character edits of each other.

[39] When generating features for a given person-prediction date, we restrict to records available prior to the prediction date.

package to include those instances and count their features as missing. Similarly, when a categorical feature is missing (e.g., police beat or gender), we assign a special category which is treated as missing. If a person is missing network features due to having no co-arrests or co-victimizations, we assign 0s for those features and include an indicator that the set of those features is missing (i.e., the person is not part of the network map).

Table A.2: Feature counts by type and subtype

Feature Type	Feature Subtype	Count
Demographics	Age	4
Demographics	Race	3
Demographics	Gender	3
Demographics	Police Beat	3
Arrest	Indexed	102
Arrest	Fine-Grained	387
Arrest	Gang	3
Victimization	Indexed	82
Victimization	Fine-Grained	224
Network	1^{st} and 2^{nd} Degree	592
Network	Centrality, degree	8
Total		1,411

Note: Indexed features include time since first incident (arrest or reported victimization), most recent incident, and cumulative incident counts within different time windows prior to the prediction date, classified by the broad crime type categories described below. Fine-grained features include similar cumulative counts but classified by the more granular crime type categories defined by unique UCR code and charge. Network features include counts of the number of incidents involving a focal person's direct network neighbors (first degree) and those one degree further removed (second degree), and counts of the number of first and second degree neighbors involved in incidents, both by broad crime type category. Network features also include measures describing the local structure of the network graphs such as a focal person's centrality.

A.3.1 Demographic features

We construct 13 demographic features from information on a person's age, race, gender, and home address.[40] As with most administrative data, police records are often noisy, with different values of theoretically invariant characteristics appearing across multiple records for the same person. We represent age and race using the modal value across a person's record set. When exact date of birth is missing, we treat the age feature as missing and construct a missing indicator; this occurs only for 10,766 people with only victim records (i.e., people who have never been arrested). However, most of these records include an approximate age, which we use to construct an additional approximate age feature for

[40] For discussion regarding the inclusion of race in the model, see Appendix B.5.2.

each person, as well as a similar missing indicator for approximate age information.[41] We represent gender using three separate features: a person's (1) most recently recorded gender, (2) modal gender, and (3) the number of distinct genders with which they are associated. We summarize a person's home address and race using these same three types of features for their police beats.

A.3.2 Arrest features

We construct 492 features summarizing a person's arrest history prior to that cohort's prediction date. These arrest features fall into three broad types: indexed arrest features, fine-grained arrest features, and gang features.

To compute indexed arrest features, we bucket the charges associated with an arrest into several broad, overlapping categories: domestic incidents, drug crime, drug dealing, gun assault or battery, gun battery, gun robbery, property crime, violent crime, Part I violent crime, and all types of crimes. Then, we summarize individual arrest histories within each index using three types of time-aware features:

1. Time since first indexed arrest;
2. Time since most recent indexed arrest;
3. Cumulative counts of the number of indexed arrests within the following time windows: the previous 30, 60, 90, 180, 270, 365, or 730 days, and over the individual's entire previous CPD arrest history (beginning in January 1999).

While these indexed arrest features provide a rich summary of a person's arrest history, they could still potentially mask heterogeneity in the predictive value of different sorts of incidents collapsed into each index. As such, we augment our representation of prior arrests with 387 fine-grained arrest features that count how many arrests a person has, within each time window, by unique UCR code and charge.

Finally, in addition to indexed and fine-grained prior arrest features, we compute three measures of a person's prior CPD-identified gang affiliation. These measures include (i) an indicator of whether the person has any prior gang-affiliated arrests, (ii) the number of unique gangs with which a person has previously been associated, and (iii) the most recent gang with which a person is associated.

[41] We combine true and approximate age information to classify people as over- or under-23 when reporting performance metrics.

A.3.3 Victimization features

We construct 306 features summarizing a person's history of victimization prior to that cohort's prediction date. Paralleling our treatment of prior arrests, we compute both indexed and fine-grained measures of prior victimization, using the same cumulative time windows and indices.

A.3.4 Network features

Since CPD arrest, homicide offender, victim, shooting victim, and homicide victim records all share an event identifier, we construct a network using information on events within the five years prior to that cohort's prediction date that includes two types of links: (i) links between co-arrestees, and (ii) links between arrestees and victims.[42] After constructing this network, we generate two types of features summarizing a person's position within it.

First, we compute aggregate statistics describing a person's network connections (whom we refer to as neighbors). We compute two types of aggregate statistics. The first counts incidents, while the second counts people. Specifically, the first type of aggregate counts the number of incidents involving a person's neighbors, by incident type and time window. For example, we count the number of property crime incidents that occurred in the last 365 days and resulted in the arrest of a neighbor. The second type of aggregate counts the number of neighbors involved in incidents, again by incident type and time window. For example, we count the number of neighbors arrested for property crime incidents within the last 365 days. We compute these two types of aggregates separately for a person's first- and second-degree neighbors.

Second, we compute features describing the underlying network structure, including a person's degree and eigenvector centrality, as well as the maximum degree and eigenvector centrality of their first- and second-degree neighbors.

A.4 Model training

To maximize flexibility, especially in the right tail of the risk distribution, we train and test a gradient boosting machine (GBM) model (Friedman, 2001) using the LightGBM implementation of gradient boosting in the Python programming language (Ke et al., 2017). Though deep learning methods are generally more accurate than tree-based methods like GBM when using "unstructured" input data like audio, video, images, or text, they have

[42] Note this corresponds to the bipartite projection of the bipartite $person \leftrightarrow incident$ graph.

been shown to be outperformed by tree-based methods, including GBM, on structured data like the kind we use here (Caruana et al., 2008; Shwartz-Ziv and Armon, 2022).

As described in section 3, we depart from traditional machine learning applications in how we generate our hold-out test set. The typical approach is to subsample observations and hold out a set of people from the training process, so that tests of predictive power are performed on entirely unseen data. However, given the social nature of some criminal activity (e.g., Billings et al., 2019), it is possible that holding out individuals would not be enough to prevent group-level information from leaking from the test set into the training set. That could result in estimates of the model's performance that are more optimistic than those that would attain in real world use.

In our setting, this is of particular concern due to the inclusion of network features: even if person i is removed from the training set, the predictors we define for person j include arrest and victimization histories for those co-involved in incidents, which could include i. Information about i could therefore still appear in the training data through j's network features, even if i is in a randomly subsampled hold-out test set. Typical subsampling would therefore not adequately address the risk that information about people in the test set could be leaked to the training set, particularly given our inclusion of network features.

To avoid this kind of information leakage between the training and test sets, we do not subsample observations. Instead, we adopt an approach that involves dividing the data into four calendar time cohorts: the first two serve as the training set, the next as the validation set for hyperparameter tuning, and the last as the test set. We describe this approach in greater detail below.

A.4.1 Defining cohorts

To define cohorts, we first establish four non-overlapping 18-month outcome periods, the beginning of which we consider a "prediction date." Then, we construct cohorts of people who meet the inclusion criterion as of that date: having at least one arrest or two reported victimizations during the 50 months before each cohort's 18-month outcome period (Appendix Figure A.1).

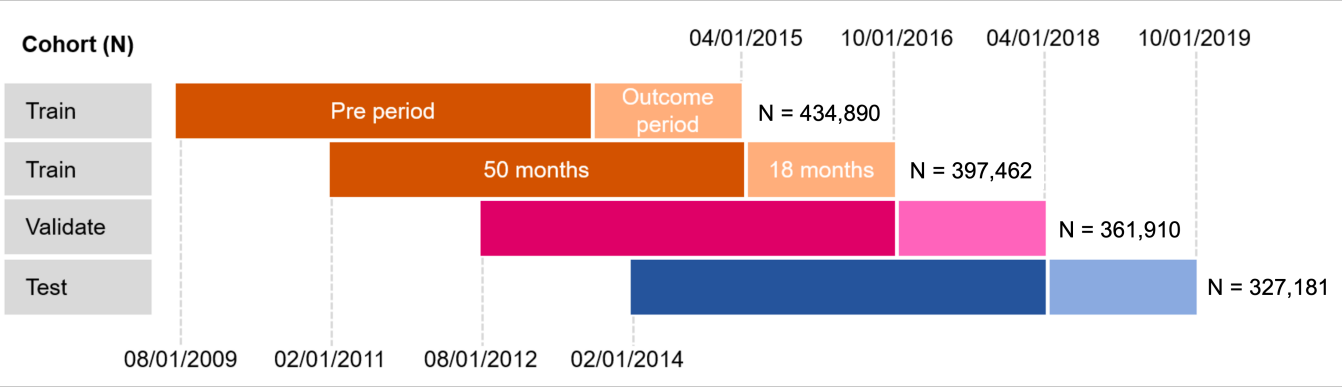
We use the first two cohorts to train the model. We split the third cohort into a 50 percent validation set for hyperparameter tuning, a 25 percent set for calibrating the predictions from the model, and a 25 percent set to optimize the number of trees used by GBM via "early stopping" (Raskutti et al., 2011).[43] The final cohort is our test set, where

[43] We found that model calibration was not meaningfully improved when we applied our calibration proce-

we predict shooting victimization risk for each person i ($\hat{p}_i$) among the 327,181 people in the test set, for the out-of-sample 18-month outcome period starting on April 1, 2018.

Note that a person can appear across multiple cohorts if they meet the inclusion criterion for them. However, a person's time-windowed features and outcomes for a given cohort are defined relative to that cohort's prediction date. As a result, even when a person appears in multiple cohorts, their features and outcomes are defined over different time periods. Because some people appear in both the training and test sets, this creates the possibility of information leakage between the two that could theoretically result in over-fitting. In Appendix B.2, we show that our results are robust to alternative model training approaches that guard against this type of information leakage.

Figure A.1: Model cohort structure



A.4.2 Hyperparameter tuning

We optimize the performance of our GBM model using random search over the following hyperparameters: number of leaves, minimum number of observations in each leaf, learning rate, and the fraction of data instances and features to use in building each tree. Our random search procedure is as follows:

1. We randomly sample $N = 100$ hyperparameter configurations from this search space.

2. For each hyperparameter configuration, we fit a GBM model over the two cohorts in the training set, using early stopping (based on minimizing log loss on a partition of the validation cohort) to optimize the number of rounds of boosting (i.e., the number of decision trees in the ensemble).

dure, and therefore we only report the results for the raw predictions in this paper.

3. From this set of $N = 100$ random hyperparameter configurations, we select the configuration that maximizes precision evaluated at the rank that equals the number of shooting victims in the validation set.

4. Finally, we refit the model, using the selected hyperparameters, over the combined training and validation sets.

A.5 Model evaluation

We evaluate the performance of our model on the test set (prediction sample). While our primary evaluation metrics are described in the paper, this section provides additional detail on (i) construction and interpretation of the $\hat{p}$-weighted prediction sample (Figure 3) and (ii) construction of bootstrap confidence intervals for precision and recall at k.

A.5.1 $\hat{p}$-weighted prediction sample

Figure 3 includes a "Predicted victims" series that shows the demographic composition of a weighted sample, where people in the prediction sample are weighted based on their predicted shooting risk $\hat{p}_i$. Specifically, for all people i belonging to a given demographic subgroup G, this series shows

$$\% \text{ in demographic group } G = \frac{\sum_{i \in G} \hat{p}_i}{\sum_i \hat{p}_i}$$

where $\hat{p}_i$ is the predicted risk for person i. If the model generated perfect predictions, then the demographic composition of predicted victims would be the same as the demographic composition of actual victims in the prediction sample. As such, differences between the second and third horizontal bars in Figure 3 indicate misprediction.

A.5.2 Bootstrap confidence intervals

For many of the estimates in this paper, we report 95 percent confidence intervals at different k (Figures 1, 2, and 4; Table 1; Appendix Figures B.3, B.4, and B.5; and Appendix Tables B.3, B.6, B.7, and B.12). These are constructed from 1,000 bootstrap samples, where each bootstrap sample is generated by:

1. Bootstrap resampling the prediction sample (i.e., drawing $N_{prediction} = 327,181$ instances from the test set, with replacement).

2. Within each bootstrap sample, computing Precision_k and Recall_k at different k (e.g., $k = 1, 2, \ldots, 5000$).

The 95 percent confidence intervals report the 2.5^{th} and 97.5^{th} percentiles from this bootstrap distribution.

While this bootstrap procedure characterizes prediction set sample variance, it does not account for other sources of variation in our procedure (e.g., training set sample variance, explicit randomness in the gradient boosting algorithm, etc.).

B Additional results

B.1 Sensitivity to length of evaluation and training duration

Our preferred model specification is characterized by two types of outcome duration. The first is the duration, starting April 1, 2018, over which we evaluate the model's performance in the test set ("evaluation duration"), which we chose to be 18 months. The second is the duration we use in the model training process to construct cohorts with non-overlapping outcome periods ("training duration"), which we also chose to be 18 months. Our choice of 18 months for both the evaluation and training durations corresponds to the length of an intervention for which a closely related predictive model was used to identify participants (Bhatt et al., 2024).

To investigate the sensitivity of our results to different choices of evaluation and training durations, we consider four candidate time periods: 6 months, 12 months, 18 months, and 24 months. For each candidate evaluation duration, we evaluate the model's performance at predicting whether a person in the test set has a reported shooting victimization within the specified duration starting April 1, 2018. For each candidate training duration, we redefine the training and validation cohorts described in Appendix A.4.1 to have non-overlapping outcome periods of the specified duration.

We then re-estimate our main model for each combination of evaluation and training duration, and assess the impact on the model's predictive performance. Appendix Table B.1 presents the results of this sensitivity analysis, reporting the number of true positives, precision, and recall for the $k = 4,244$ people with the highest predicted risk for each combination of evaluation and training duration.

Table B.1: Predictive performance by evaluation and training duration

	Evaluation Duration			
Training Duration	6mo	12mo	18mo	24mo
True Positives				
6mo	237 (208, 266)	330 (296, 367)	474 (434, 515)	496 (455, 539)
12mo	236 (205, 263)	333 (296, 367)	470 (429, 511)	491 (447, 531)
18mo	237 (208, 267)	336 (302, 371)	486 (444, 527)	509 (467, 551)
24mo	233 (204, 262)	331 (294, 365)	477 (434, 516)	500 (457, 541)
Precision				
6mo	0.056 (0.049, 0.063)	0.078 (0.070, 0.086)	0.112 (0.102, 0.121)	0.117 (0.107, 0.127)
12mo	0.056 (0.048, 0.062)	0.078 (0.070, 0.086)	0.111 (0.101, 0.120)	0.116 (0.105, 0.125)
18mo	0.056 (0.049, 0.063)	0.079 (0.071, 0.087)	0.115 (0.105, 0.124)	0.120 (0.110, 0.130)
24mo	0.055 (0.048, 0.062)	0.078 (0.069, 0.086)	0.112 (0.102, 0.122)	0.118 (0.108, 0.127)
Recall				
6mo	0.199 (0.175, 0.224)	0.179 (0.160, 0.199)	0.168 (0.154, 0.182)	0.168 (0.154, 0.183)
12mo	0.198 (0.172, 0.221)	0.180 (0.160, 0.199)	0.166 (0.152, 0.181)	0.167 (0.152, 0.180)
18mo	0.199 (0.175, 0.224)	0.182 (0.163, 0.201)	0.172 (0.157, 0.186)	0.173 (0.159, 0.187)
24mo	0.196 (0.171, 0.220)	0.179 (0.159, 0.198)	0.169 (0.154, 0.183)	0.170 (0.155, 0.184)

Note: Precision and recall for the $k = 4,244$ people with the highest predicted risk of shooting victimization. In each cell, results are reported from the full model predicting shooting victimization during the outcome period of the specified evaluation duration starting April 1, 2018, trained using cohorts with non-overlapping outcome periods of the specified training duration. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details).

We start by focusing on the 18-month evaluation duration used in our main results. Across rows representing different training durations, we see that model performance is always highest with the 18-month training duration used in our main results. However, performance differences across different training durations are substantively small and not statistically significant.

Comparing performance across evaluation durations is less straightforward. Holding training duration fixed, as the evaluation duration grows longer, there is mechanically more time for someone among the $k = 4,244$ people with the highest predicted risk to be shot, which should weakly increase the number of true positives and precision. Comparing results across the columns of Appendix Table B.1 demonstrates this pattern, with the number of true positives and precision increasing with evaluation duration.[44] Recall for shooting victims in the prediction sample, on the other hand, declines modestly with evaluation duration.

Overall, the results in Appendix Table B.1 confirm that our main conclusion—an algorithm using police data to predict shooting victimization could help guide prevention

[44] We would not expect the increase in true positives and precision to be linear in evaluation duration for three reasons. First, due to seasonality: because the evaluation period always begins April 1, 2018, months 1-6 and 13-18 correspond to the months of April through September and encompass the summer (when rates of gun violence are usually highest), while months 7-12 and 19-24 correspond to October through March and encompass the winter (when rates of gun violence are usually lowest). Second, due to secular trends in gun violence: there were almost 10 percent fewer shooting victims in 2019 than in 2018. And third, due to possible risk decay: the risk of being shot may be highest at the start of the outcome period and decline further out.

efforts—is not limited to an 18-month outcome period. For outcome periods as short as 6 months and as long as 24 months, the model is able to identify a small group of people who experience high rates of shooting victimization and who encompass a large share of all shooting victims in the prediction sample over the period.

B.2 Sensitivity to alternative GBM model training approaches

In this section, we explore the sensitivity of our performance measures to alternative approaches for training a GBM model. We designed our preferred cohort-based approach, described in section 3 and above in Appendix A.4, in part to avoid the kind of information leakage across social networks that might stem from a traditional hold-out approach (i.e., where information about a person's neighbors could be included in the training data even if the person was held out for the test set). At least in theory, this leaves another possible source of information leakage that could result in over-fitting: within-person correlation over time across cohorts. While we did not expect this kind of information leakage to be particularly problematic, the alternatives reported in this section provide empirical confirmation.

To elaborate: the preferred approach we use in the main text to avoid over-fitting involves splitting our data into four cohorts. A person appears in any cohort for which they meet the inclusion criterion: having at least one arrest or two reported victimizations during the 50 months before the cohort's 18-month outcome period. As a result of our cohort definitions, a person can—and those with frequent police contact often do—appear in multiple cohorts, including in both cohorts used for testing and training. When a person appears in multiple cohorts, their time-windowed predictors and outcomes are defined over different time periods. Still, it is possible that observations across cohorts for the same person are not entirely independent, potentially resulting in information leakage between the training and test sets, and in overly optimistic performance estimates.

To confirm that our main approach is not introducing significant over-fitting via this kind of leakage, we investigate three alternative training approaches, described with their associated trade-offs below. In each of these three alternative approaches, the model used to generate predictions for people in the test set is never trained on data that includes observations from those same people. To further assess whether information leakage arising from the training and test sets containing people who are co-involved in criminal activity is causing over-fitting, we implement versions of each of these approaches that also drop the network features.

We show each of these three approaches to be transparent about which, if any, changes in the training procedure matter for our results. As explained below, however, we think

approach (3) is the most reasonable alternative to our main cohort-based approach for avoiding over-fitting in our setting.

1. **Hold out all current test set individuals from the training set.** Here we retain the cohort structure described in Appendix A.4.1, but drop all people in the test set from earlier cohorts. We then train the model using the cohort-based approach described in Appendix A.4.

   Relative to our main approach, this approach has two major limitations. First, by excluding people in the test set from earlier cohorts, we may be discarding potentially informative data that could improve the model's performance. The information lost is disproportionately from people with frequent police contact—those who meet the inclusion criterion for the last cohort and earlier ones as well—who are among those at the highest risk of being shot.

   Second, this approach may lead to a mismatch in the covariate distribution between the training and test sets, as people who appear in the test set may have different characteristics than those who only appear in earlier cohorts. For example, people in the test set have an average of 5.1 arrests in their histories, compared to an average of 3.0 for people in the other three cohorts. So while this is a clean way to avoid having a person in the test set appear in earlier cohorts, it also fundamentally changes the group contributing to model training in a way that makes it less like the population in the test set.

2. **Nested cross-validation.** Here we stack all person-cohort observations together into one dataset. We randomly divide people (and all their associated cohort-specific observations) into five folds. We start by performing an "outer" loop over the folds: In each iteration of the outer loop, we hold out one fold and combine the remaining four. We then perform an "inner" loop over the combined folds, training a gradient boosting model using five-fold cross-validation and using this model to generate out-of-sample predictions for the held-out fold from the outer loop. We repeat this process until all people have received an out-of-sample prediction.

   Unlike approach (1), this approach uses all available data for each person, regardless of whether they appear in multiple cohorts or not. Stratifying at the person level also helps to address the mismatch in covariate distributions between the training and test sets that arises when excluding people in the test set from earlier cohorts. By including all instances of a person meeting the inclusion criterion, we can ensure that the model is trained and tested on a more representative sample of the population,

with a similar distribution of arrest histories and other relevant covariates. For example, the average number of prior arrests is similar among people in the training and validation cohorts and people in the test cohort (an average of 4.63 versus 5.1, respectively).

However, a key limitation of this approach is that, unlike with the cohort-based approaches, we train and test on observations from the same time periods. In a real-world application of such a model, we would not have access to future information.

3. **Nested cross-validation with cohort-based inner loop.** Here we stack all person-cohort observations together into one dataset. We randomly divide people (and all their associated cohort-specific observations) into five folds. We start by performing an "outer" loop over the folds: In each iteration of the outer loop, we hold out one fold and combine the remaining four. We then perform an "inner" loop over the combined folds, training a gradient boosting model with the cohort-based approach in Appendix A.4, then using this model to generate out-of-sample predictions for the held-out fold from the outer loop. We repeat this process until all people have received an out-of-sample prediction.

   This approach is similar to (2), with one crucial difference: the out-of-sample predictions for people in the test cohort are generated using a model trained on data only from earlier cohorts, and not data from other people in the same contemporaneous test cohort. This avoids training and testing on observations from the same time period and maintains the forward-in-time aspect of model training and validation that characterizes our main approach.

Appendix Table B.2 presents the results of this sensitivity analysis. It reports the number of true positives, precision, and recall for the $k = 4,244$ people with the highest predicted risk under different modeling approaches. The results in the top panel are from models trained on all 1,411 features, while the results in the bottom panel are from models trained excluding network features. The bottom panel effectively limits both potential sources of over-fitting at once.

The first row in each panel of Appendix Table B.2 reports the performance of our main model training approach, while the next three rows report the performance of the alternative model training approaches described above. Within each panel, performance is very similar across all of the approaches. Though nested cross-validation (approach 2) appears to slightly outperform our main approach, the differences are not statistically significant. Looking across panels, the exclusion of network features reduces performance very slightly, but the differences are not statistically significant, either.

Table B.2: Predictive performance by model training approach

	Top 4,244			
Model Training Method	True Positives	Precision	Recall	Total Recall
All features				
Main approach	486 (444, 527)	0.115 (0.105, 0.124)	0.172 (0.157, 0.186)	0.115 (0.105, 0.124)
Exclude test cohort individuals from earlier cohorts	476 (435, 517)	0.112 (0.102, 0.122)	0.168 (0.154, 0.183)	0.112 (0.102, 0.122)
Nested cross-validation	501 (460, 544)	0.118 (0.108, 0.128)	0.177 (0.163, 0.192)	0.118 (0.108, 0.128)
Nested cross-validation with cohorts	481 (441, 519)	0.113 (0.104, 0.122)	0.170 (0.156, 0.184)	0.113 (0.104, 0.122)
No network information				
Main approach	483 (441, 521)	0.114 (0.104, 0.123)	0.171 (0.156, 0.184)	0.114 (0.104, 0.123)
Exclude test cohort individuals from earlier cohorts	448 (408, 490)	0.106 (0.096, 0.115)	0.158 (0.144, 0.173)	0.106 (0.096, 0.115)
Nested cross-validation	488 (449, 528)	0.115 (0.106, 0.124)	0.173 (0.159, 0.187)	0.115 (0.106, 0.124)
Nested cross-validation with cohorts	469 (430, 508)	0.111 (0.101, 0.120)	0.166 (0.152, 0.180)	0.111 (0.101, 0.120)

Note: Precision and recall for the $k = 4,244$ people with the highest predicted risk of shooting victimization. Results are reported for a model trained using the indicated model training approach and evaluated during the 18-month outcome period starting April 1, 2018. In the top panel, models are trained using the full set of 1,411 features. In the bottom panel, models are trained excluding network features. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details).

If information was being leaked between the training and test sets due to the presence of the same people (but in different cohorts) in our main cohort-based model training approach, we would expect the corresponding performance metrics to be inflated relative to those from approaches that limit the possibility of this type of information leakage. However, we see no evidence of that, giving us confidence that the performance estimates we report in the main text are not overly optimistic due to over-fitting.

B.3 Prevalence of other outcomes among those with high predicted risk of shooting victimization

The main text reports predictive performance for the primary outcome of interest, reported shooting victimization, when ranking people by their predicted risk of that outcome. Because the risk of being shot is likely correlated with the risk of other socially costly outcomes, efforts to reduce the risk of shooting victimization among this group may reduce the risk of these other outcomes as well. We do not focus on quantifying the benefits of reducing the risk of these other outcomes, since they are less reliable measures of the underlying behavior of interest (i.e., the relationship between arrest for violent crime and true violent offending is likely to be noisier and to differ by racial group, relative to the relationship between reported shooting victimization and actual shooting victimization).

Nonetheless, because efforts to prevent shooting victimization among this group may

produce other large benefits, this section reports on the prevalence of other measures of violence among those predicted to be at high risk of being shot. Note that we are not training a model to predict these other outcomes, since that would likely confound police behavior or willingness to report violence to the police with true individual risk. Rather, we are reporting on the prevalence of different violence measures among groups defined by their ranking in the shooting victimization predictions.

Appendix Table B.3 below reports our standard measures of model performance, precision and recall, for the full shooting victimization model evaluated on four different outcomes: shooting victimization, shooting arrest, violent crime victimization, and violent crime arrest.

Table B.3: Predictive performance of shooting victimization predictions for other outcomes

k	True Positives	Precision	Recall	Total Recall
Shooting Victim				
500	83 (67, 100)	0.166 (0.134, 0.200)	0.029 (0.024, 0.035)	0.020 (0.016, 0.024)
4,244	486 (444, 527)	0.115 (0.105, 0.124)	0.172 (0.157, 0.186)	0.115 (0.105, 0.124)
327,181	2,827 (2726, 2933)	0.009 (0.008, 0.009)	1.000 (0.964, 1.038)	0.666 (0.642, 0.691)
Shooting Arrest				
500	32 (21, 43)	0.064 (0.042, 0.086)	0.054 (0.035, 0.072)	0.042 (0.027, 0.056)
4,244	162 (138, 188)	0.038 (0.033, 0.044)	0.272 (0.232, 0.315)	0.210 (0.179, 0.244)
327,181	596 (546, 642)	0.002 (0.002, 0.002)	1.000 (0.916, 1.077)	0.773 (0.708, 0.833)
Violent Crime Victim				
500	112 (94, 131)	0.224 (0.188, 0.262)	0.007 (0.006, 0.008)	0.003 (0.002, 0.003)
4,244	717 (666, 767)	0.169 (0.157, 0.181)	0.044 (0.040, 0.047)	0.017 (0.016, 0.018)
327,181	16,475 (16245, 16713)	0.050 (0.050, 0.051)	1.000 (0.986, 1.014)	0.397 (0.391, 0.403)
Violent Crime Arrest				
500	100 (82, 117)	0.200 (0.164, 0.234)	0.020 (0.017, 0.024)	0.013 (0.011, 0.016)
4,244	578 (533, 620)	0.136 (0.126, 0.146)	0.117 (0.108, 0.126)	0.077 (0.071, 0.083)
327,181	4,940 (4814, 5077)	0.015 (0.015, 0.016)	1.000 (0.974, 1.028)	0.662 (0.645, 0.680)

Note: Precision and recall from the full model trained to predict shooting victimization during the 18-month outcome period starting April 1, 2018. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details). Model performance is evaluated on the four outcomes shown, for the k people with the highest predicted risk of shooting victimization. Violent crimes refer to the Part I violent index offenses: aggravated assault, aggravated battery, forcible rape, murder, and robbery. Prediction sample size is 327,181.

The people whom the model predicts to be at higher risk of shooting victimization are indeed at higher risk for these other adverse outcomes during the 18-month outcome period as well. For example, among the 500 people at highest predicted risk of shooting victimization, 6.4 percent are arrested on suspicion of carrying out a shooting (32 times the base rate in the whole test set of 0.2 percent); 22.4 percent are reported as victims

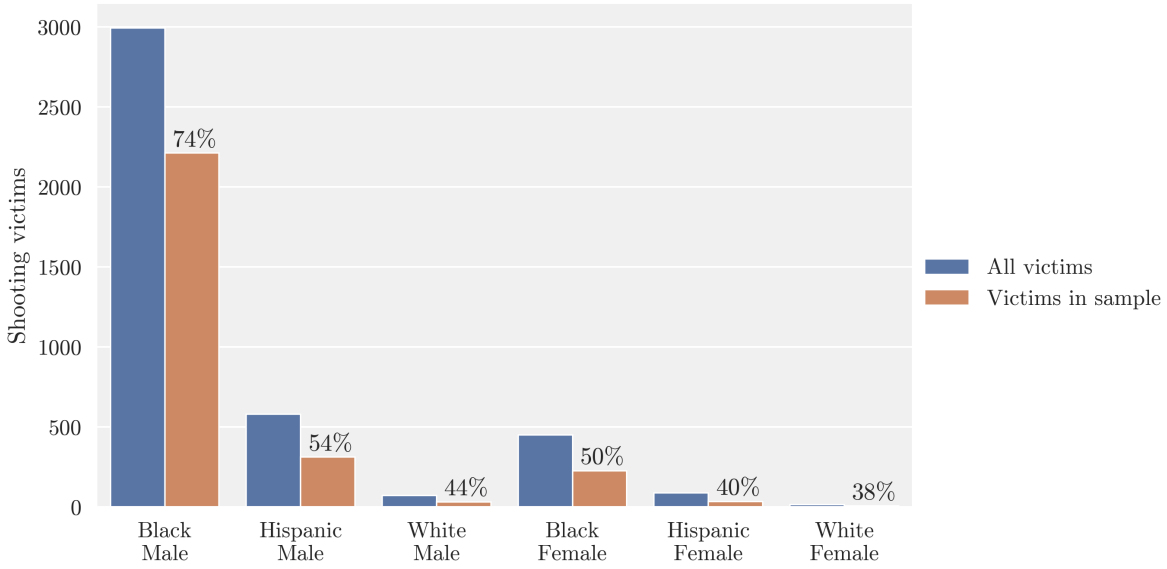
of a violent offense (4.5 times the base rate); and 20 percent are arrested on suspicion of carrying out a violent offense (13.3 times the base rate).

B.4 Victim counts and performance by demographic group

Figure 3 in the main text shows the proportion of shooting victims that fall into different demographic groups. This section adds some additional information to the summaries in the main text.

Appendix Figure B.1 compares the race/ethnicity and gender composition of all 4,244 shooting victims to the 2,827 victims with enough data to receive a prediction. The blue bars show the number of all shooting victims in each group and the orange bars the number in our prediction sample, with the label reporting the share of all victims in that group who are in our sample. Three-fourths of all Black male victims are in our sample and therefore receive predictions, compared to roughly half or fewer of the victims from other demographic groups.

Figure B.1: Demographic composition of all victims and those in the prediction sample



Note: Figure reports counts of shooting victims separately by race/ethnicity and gender, among all 4,244 shooting victims during the 18-month outcome period and the 2,827 victims in the prediction sample. Percentages above the in-sample bars report the share of all shooting victims in that demographic group (each blue bar) who appear in the prediction sample.

To be transparent about the underlying size of each group in Appendix Figure B.1, Appendix Table B.4 below reports the counts across demographic categories of four groups: all shooting victims, shooting victims in the prediction sample, predicted victims (see discussion above in Appendix A.5.1), and the $k = 4,244$ people with the highest predicted

risk.

Table B.4: Demographic composition of actual and predicted shooting victims

Race	Gender	Age	Actual Victims	Actual Victims (Not In Sample)	Actual Victims (In Sample)	Predicted Victims	Top 4,244 Predicted Victims
Black	Male	<23	1,097	264	833	717	2,453
		23+	1,885	507	1,378	1,308	1,415
	Female	<23	188	109	79	58	1
		23+	261	113	148	126	0
Hispanic	Male	<23	254	119	135	129	239
		23+	323	146	177	225	131
	Female	<23	30	15	15	10	0
		23+	58	38	20	22	0
White	Male	<23	14	7	7	8	2
		23+	58	33	25	33	3
	Female		16	10	6	12	0
Other/Missing			60	56	4	11	0
Total			4,244	1,417	2,827	2,658	4,244

Note: Counts for White females are not disaggregated by age due to small cell sizes.

Figure 1 in the main text shows that the predictions are well-calibrated overall and by racial group, with some overestimation among those predicted to be at the very highest risk within the distributions for White and Hispanic individuals. Appendix Table B.5 sheds additional light on calibration by contrasting the base shooting victimization rate within the prediction sample and the average prediction, both by race/ethnicity (as in Figure 1) and further broken down by age and gender (as in Figure 3).

Consistent with the calibration plots in the main text (Figure 1), average predictions are generally quite similar to observed rates of shooting victimization, even within race/ethnicity-age-gender groups. Where the model's predictions deviate the most from base shooting victimization rates are for young Black and Hispanic men, for whom the model under-predicts risk by 0.021.

Table B.5: Base rate and average predicted risk by race, gender, and age for prediction sample

Race	Gender	Age	N	Base Rate By Group	Mean Predicted Risk
Black	All	All	200,186	0.017	0.011
	Female	All	77,607	0.006	0.002
		<23	9,073	0.016	0.005
		23+	68,534	0.004	0.002
	Male	All	122,579	0.024	0.017
		<23	14,297	0.060	0.039
		23+	108,282	0.020	0.014
Hispanic	All	All	68,916	0.010	0.006
	Female	All	20,385	0.004	0.002
		<23	2,459	0.009	0.003
		23+	17,926	0.004	0.001
	Male	All	48,531	0.012	0.007
		<23	5,432	0.039	0.018
		23+	43,099	0.008	0.006
White	All	All	49,717	0.002	0.001
	Female	All	18,452	<0.001	<0.001
		<23	860	0.003	0.001
		23+	17,592	<0.001	<0.001
	Male	All	31,265	0.002	0.001
		<23	1,448	0.008	0.004
		23+	29,817	0.002	0.001
Other Race/Gender			7,302	0.002	0.001
Missing Race/Gender/Age			2,477	0.017	0.001

Note: Table shows the base rate, or the proportion of each group that becomes a shooting victim during the outcome period, along with the average predicted risk within each group. Note that the "All" age rows include individuals of that race/ethnicity and gender who are missing age information; as a result, the number of observations in the under- and over-23 rows do not exactly sum up to those for the "All" row. The final row groups together everyone with missing race/ethnicity, gender, or age information.

Of course, average predictions being similar to base rates at a group level does not mean each person's prediction is accurate. To assess accuracy at the individual level, one must establish a decision rule that translates predicted risk levels into classifications of "positive" (predicted to be shot) and "negative" (predicted not to be shot) for each person. There are many different classification rules one could use. Given the uneven demographic distribution of individuals across the risk distribution, different decision rules could have different implications for who is correctly and incorrectly classified.

Since a natural kind of decision rule is a threshold rule, where policymakers would consider everyone above some global risk threshold as a positive prediction and everyone

below as a negative prediction, we show the implications of one such threshold (the same that is shown in Figure 3): serving the $k = 4,244$ people with the highest predicted risk (motivated by the fact that there are 4,244 actual victims in the outcome period). Appendix Table B.6 shows precision and average predicted risk within race/ethnicity and age groups for the subset of men among the $k = 4,244$. We omit women and the age breakdown for White men in this table because there are so few of these individuals in this top-ranked group.

Table B.6: Precision and average predicted risk by race and age for men among the top 4,244

Race	Gender	Age	N	Precision	Mean Predicted Risk
Black	Male	All	3,868	0.117 (0.108, 0.127)	0.113 (0.112, 0.114)
		<23	2,021	0.125 (0.111, 0.138)	0.119 (0.118, 0.121)
		23+	1,847	0.109 (0.096, 0.124)	0.105 (0.104, 0.106)
Hispanic	Male	All	370	0.086 (0.059, 0.116)	0.108 (0.106, 0.111)
		<23	207	0.092 (0.053, 0.135)	0.109 (0.106, 0.113)
		23+	163	0.080 (0.043, 0.123)	0.107 (0.104, 0.112)
White	Male	All	5	0	0.120 (0.099, 0.138)

Note: Table reports statistics for White males of all ages together and omits 6 individuals belonging to other demographic groups due to small cell sizes. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details).

Comparing the two columns gives a sense for subgroup calibration for this subsample, and precision shows the proportion of true positives (such that 1 − Precision is the false discovery rate). Again we emphasize that this not reflective of performance across the whole sample, but rather provides additional information on the fairness implications of a "top 4,244" decision rule.

Comparing the mean predicted risk with the realized risk (precision) in Appendix Table B.6 shows several key patterns. First, consistent with the subgroup calibration panels in Figure 1, predicted risk among this right tail is quite close to the realized risk for Black men, but slightly overstates the realized risk for Hispanic and White men on average, albeith with overlapping confidence intervals.

In terms of classification among the top 4,244, the model has the highest true positive rate (and thus lowest false discovery rate) for Black men, of whom 11.7 percent are correctly classified, i.e., become shooting victims in the outcome period. In contrast, among Hispanic men—a much smaller group of 370 compared to 3,868 Black men—only 8.6 percent are correctly classified. This is consistent with argument in the main text that the over-representation of Black men in the right tail of the risk distribution is not because

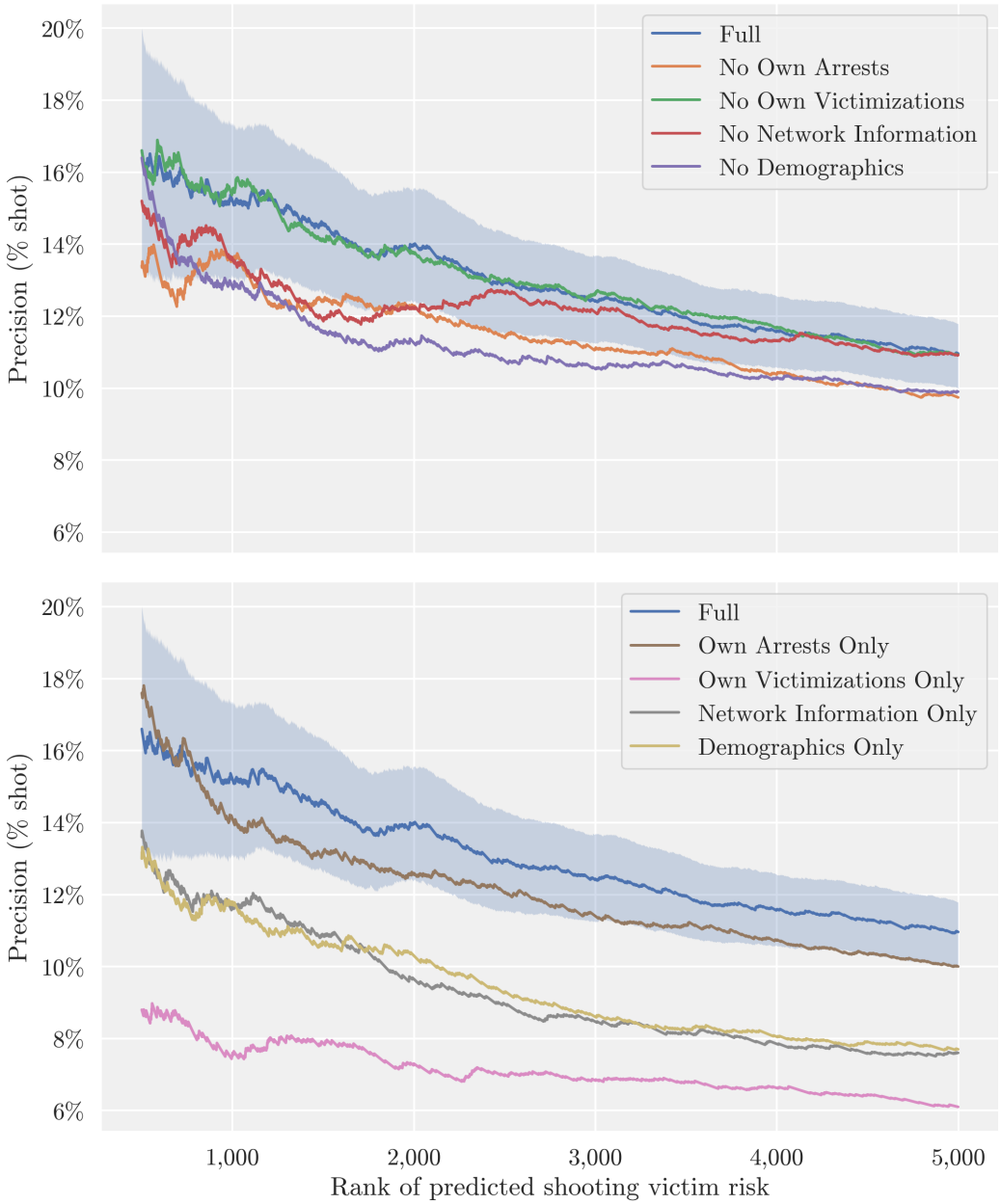
estimates of their risk are inflated, but rather because the model does a better job at identifying Black men who face genuinely higher risk of victimization. These true positive rates are extremely high from a substantive standpoint, identifying over 450 Black men and 30 Hispanic men for whom preventive services might have kept them from serious injury or death. Nonetheless, the fact that almost 90 percent of Black men and 91 percent of Hispanic men above this threshold are not shot during the outcome period again emphasizes how costly it would be target any intervention that reduced people's civil liberties based on these predictions.

B.5 Further detail on what matters for prediction

B.5.1 Performance by groups of features

The main text presents predictive performance leaving out three sets of features: own arrests, peer information (networks), and both (Figure 4). We expand this exercise below for additional combinations of features. Appendix Figure B.2 reports precision for the full model and different models that each exclude certain feature sets. To ensure the lines are not all on top of each other, we limit the scale to the top 5,000 ranked individuals in each model. Past 5,000, most of the differences in performance tend to be quite small. Appendix Table B.7 quantifies the precision differences and 95 percent bootstrapped confidence intervals at $k = 500$ and $k = 4,244$, as well as reporting recall and total recall.

Figure B.2: Precision across models with different feature sets



Note: Figure shows precision, or the share of the $k \leq 5,000$ people with the highest predicted risk of shooting victimization who are shot during the 18-month outcome period, for models trained with different feature sets. Due to noise in precision at low values of k, we start the graph at $k = 500$. Bootstrapped 95 percent confidence interval for the full model shown (see Appendix A.5.2 for details).

Table B.7: Predictive performance by feature set

Feature Set	Top 500				Top 4,244			
	True Positives	Precision	Recall	Total Recall	True Positives	Precision	Recall	Total Recall
Full	83 (67, 100)	0.166 (0.134, 0.200)	0.029 (0.024, 0.035)	0.020 (0.016, 0.024)	486 (444, 527)	0.115 (0.105, 0.124)	0.172 (0.157, 0.186)	0.115 (0.105, 0.124)
No Network Information	76 (60, 91)	0.152 (0.120, 0.182)	0.027 (0.021, 0.032)	0.018 (0.014, 0.021)	483 (441, 521)	0.114 (0.104, 0.123)	0.171 (0.156, 0.184)	0.114 (0.104, 0.123)
No Own Victimizations	83 (66, 98)	0.166 (0.132, 0.196)	0.029 (0.023, 0.035)	0.020 (0.016, 0.023)	484 (445, 527)	0.114 (0.105, 0.124)	0.171 (0.157, 0.186)	0.114 (0.105, 0.124)
No Race	83 (68, 100)	0.166 (0.136, 0.200)	0.029 (0.024, 0.035)	0.020 (0.016, 0.024)	501 (457, 541)	0.118 (0.108, 0.127)	0.177 (0.162, 0.191)	0.118 (0.108, 0.127)
No Own Arrests or Victimizations	76 (60, 91)	0.152 (0.120, 0.182)	0.027 (0.021, 0.032)	0.018 (0.014, 0.021)	424 (386, 461)	0.100 (0.091, 0.109)	0.150 (0.137, 0.163)	0.100 (0.091, 0.109)
No Demographics	82 (66, 98)	0.164 (0.132, 0.196)	0.029 (0.023, 0.035)	0.019 (0.016, 0.023)	436 (396, 477)	0.103 (0.093, 0.112)	0.154 (0.140, 0.169)	0.103 (0.093, 0.112)
Own Arrests Only	88 (71, 104)	0.176 (0.142, 0.208)	0.031 (0.025, 0.037)	0.021 (0.017, 0.025)	446 (407, 485)	0.105 (0.096, 0.114)	0.158 (0.144, 0.172)	0.105 (0.096, 0.114)
Own Arrests and Network Information Only	76 (60, 91)	0.152 (0.120, 0.182)	0.027 (0.021, 0.032)	0.018 (0.014, 0.021)	436 (396, 477)	0.103 (0.093, 0.112)	0.154 (0.140, 0.169)	0.103 (0.093, 0.112)
No Own Arrests	67 (52, 83)	0.134 (0.104, 0.166)	0.024 (0.018, 0.029)	0.016 (0.012, 0.020)	432 (394, 470)	0.102 (0.093, 0.111)	0.153 (0.139, 0.166)	0.102 (0.093, 0.111)
Network Information Only	68 (53, 82)	0.136 (0.106, 0.164)	0.024 (0.019, 0.029)	0.016 (0.012, 0.019)	330 (295, 365)	0.078 (0.070, 0.086)	0.117 (0.104, 0.129)	0.078 (0.070, 0.086)
No Own Arrests, No Network Information	63 (48, 78)	0.126 (0.096, 0.156)	0.022 (0.017, 0.028)	0.015 (0.011, 0.018)	383 (349, 416)	0.090 (0.082, 0.098)	0.135 (0.123, 0.147)	0.090 (0.082, 0.098)
Demographics Only	65 (52, 82)	0.130 (0.104, 0.164)	0.023 (0.018, 0.029)	0.015 (0.012, 0.019)	338 (305, 374)	0.080 (0.072, 0.088)	0.120 (0.108, 0.132)	0.080 (0.072, 0.088)
Own Victimizations Only	44 (32, 55)	0.088 (0.064, 0.110)	0.016 (0.011, 0.019)	0.010 (0.008, 0.013)	273 (243, 308)	0.064 (0.057, 0.073)	0.097 (0.086, 0.109)	0.064 (0.057, 0.073)

Note: Models differ based on the feature sets available to them during training (see text below). Model performance is evaluated on shooting victimization during the outcome period for the $k = 500$ and $k = 4,244$ people with the highest predicted risk of shooting victimization. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details). Models are sorted by Total Recall for the top 4,244.

The definitions of the models that leave out particular feature sets are as follows:

Top panel

1. Full: The main model reported in the text with all available features
2. No Own Arrests: Excludes all arrest features for the focal person (but includes them for first- and second-degree peers), including gang-related features
3. No Own Victimizations: Excludes all victimization features for the focal person (but includes them for first- and second-degree peers)
4. No Network Information: Excludes all features about the focal person's first- and second-degree peers, and about the local structure of the network graphs themselves, such as the focal person's centrality and number of neighbors
5. No Demographics: Excludes race, gender, age, and location information

Bottom panel

1. Full: Same as above
2. Own Arrests Only: Uses only arrest features for the focal person, excluding all victimization, demographic, and network features
3. Own Victimizations Only: Uses only victimization features for the focal person, excluding all arrest, demographic, and network features
4. Network Information Only: Uses only features about the focal person's first- and second-degree peers, and about the local structure of the network graphs themselves, such as the focal person's centrality and number of neighbors, excluding all arrest, victimization, and demographic features for the focal person
5. Demographics Only: Uses only information on demographics, excluding arrests and victimization information

As the top panel shows, the feature sets that reduce precision the most when excluded are a person's own arrest history and their demographics. As shown in Appendix Table B.7, for the $k = 4,244$ people with the highest predicted risk, the performance measures of models that exclude either of these two feature sets fall below the confidence intervals of those performance measures for the full model.[45] In contrast, and as described in

[45] For the $k = 500$ people with the highest predicted risk, the "no demographics" model performs similarly to the full model. In contrast, the "no own arrests" model performs much worse, identifying 67 victims compared to 83 for the full model, though the performance differences are just shy of being statistically significant.

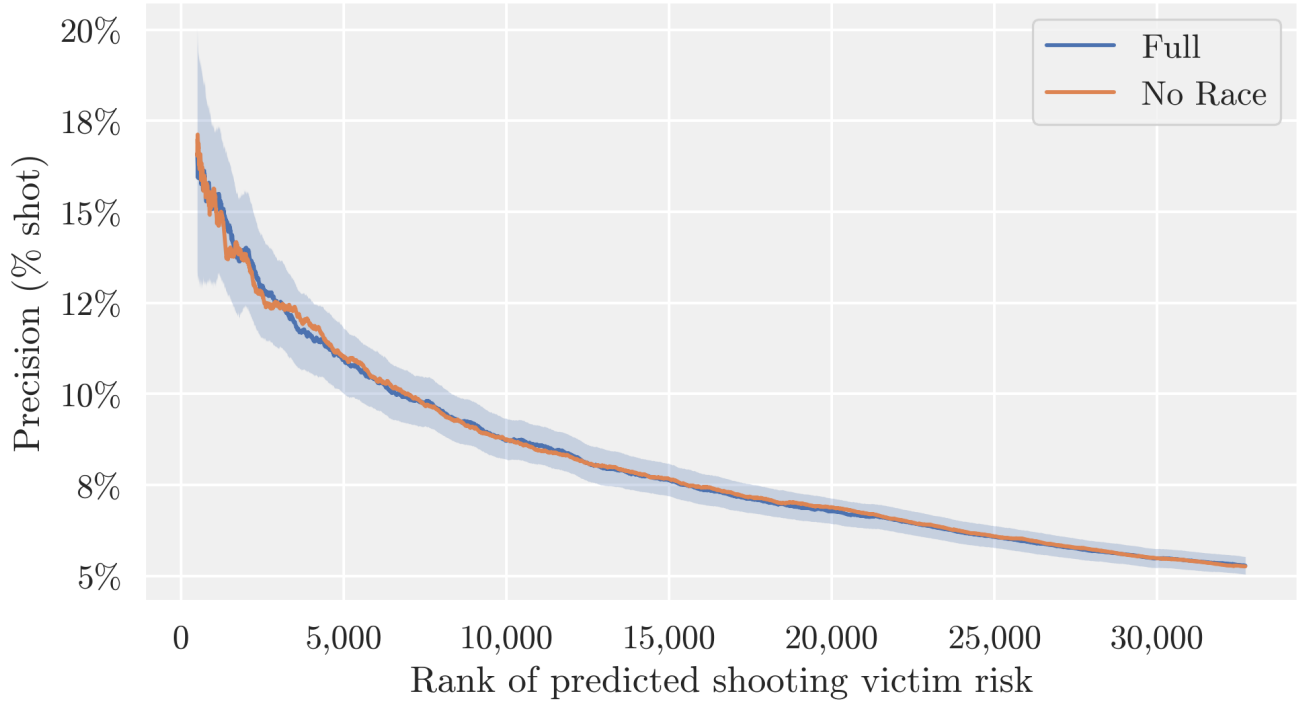
section 4.3.1, over most values of k—outside of approximately $k = 1,000$ to $k = 2,000$—the precision of the model that excludes network features is statistically indistinguishable from precision for the full model, as confirmed in Appendix Table B.7 for both $k = 500$ and $k = 4,244$.

As the bottom panel shows, the biggest loss of information comes from using only victimization records when building the model. Using just demographics or just network features does slightly better than using victimizations alone, but both still fall short of the full model. For example, compared to the number of shooting victims identified by the full model among the $k = 4,244$ people with the highest predicted risk, the model trained using only demographic information identifies 30 percent fewer shooting victims (338 vs. 486), while the model trained using only network information identifies 32 percent fewer shooting victims (330 vs. 486). In contrast, the performance of the model trained using just own arrest features is statistically indistinguishable from that of the full model at $k = 500$ and $k = 4,244$ (Appendix Table B.7). This pattern echoes the finding in the main text that while other features contain some of the same information as own arrests and some independent information, the details of a person's own arrests are particularly valuable in predicting their risk of being a shooting victim.

B.5.2 Prediction without race

Our main results come from a model that includes race in the model-building process. Many legal scholars believe that including race as an algorithmic input is likely unconstitutional, though the debate around this question is not completely settled (e.g., Yang and Dobbie, 2020). Importantly, as shown in Appendix Figure B.3, the inclusion of race has a trivial effect on predictive performance.

Figure B.3: Precision across models with and without race indicators

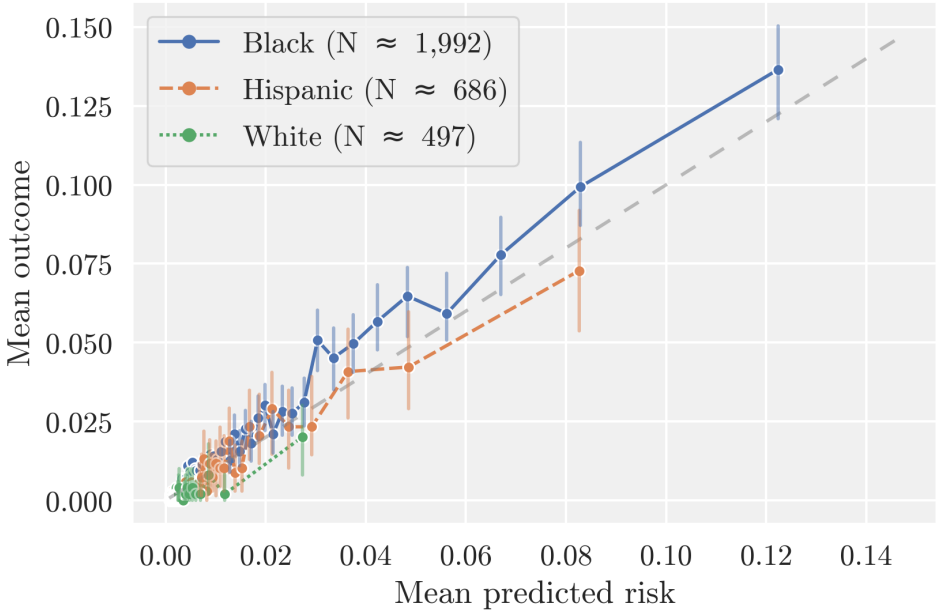


Note: Figure shows precision, or the share of the k people with the highest predicted risk of shooting victimization who are shot during the 18-month outcome period, for models trained with and without race indicators. Due to noise in precision at low values of k, we start the graph at $k = 500$. Bootstrapped 95 percent confidence interval for the full model (with race indicators) shown (see Appendix A.5.2 for details).

Appendix Figure B.4 also shows little change in calibration within race/ethnicity groups relative to the full model shown in Figure 1. So although we show the main results from a model including race, the arguments contained in the paper are equally applicable for settings that require the model to exclude race.[46]

[46] When a somewhat different version of this prediction model was used for social service referrals in practice (Bhatt et al., 2024), we excluded race; see, e.g., `https://osf.io/ap8fj/`.

Figure B.4: Calibration for model built without race indicators



Note: Figure shows mean predicted shooting victimization risk and shooting victimization rate within each percentile of the race/ethnicity-specific predicted risk distributions, from a model trained without race indicators. Race/ethnicity categories are mutually exclusive: non-Hispanic White, non-Hispanic Black, and Hispanic of any race. Bootstrapped 95 percent confidence intervals shown (see Appendix A.5.2 for details).

B.5.3 Performance by number of features & model complexity

A different way to ask what information matters is not to focus on sets of features grouped by theme, but on the number of features available and the complexity of the algorithm used to predict with them. Black box models may not be appropriate in all high-stakes settings (Rudin, 2019). A simpler model with only a few features may aid in interpretability, trust, and uptake (Ustun and Rudin, 2019). Multiple researchers have identified settings where complex models with more features provide minimal performance improvements over simple models with fewer features (e.g., Dressel and Farid, 2018; Jung et al., 2017; Angelino et al., 2018; Stevenson and Slobogin, 2018; Stevenson and Mayson, 2022). Thus, for use in these contexts, it is important to understand how much of the predictive accuracy of the full model can be captured by a drastically smaller set of features and simpler, more transparent modeling techniques.

We explore these questions in our setting by first creating a rank-ordered set of the 50 features that are most correlated with the outcome from the full set of all 1,411 features. To generate this smaller set of 50 features, we use a simple stepwise residualization procedure. First, we select the single feature that is most highly correlated with shooting victimization in the first two cohorts. Then we remove the correlation between all other

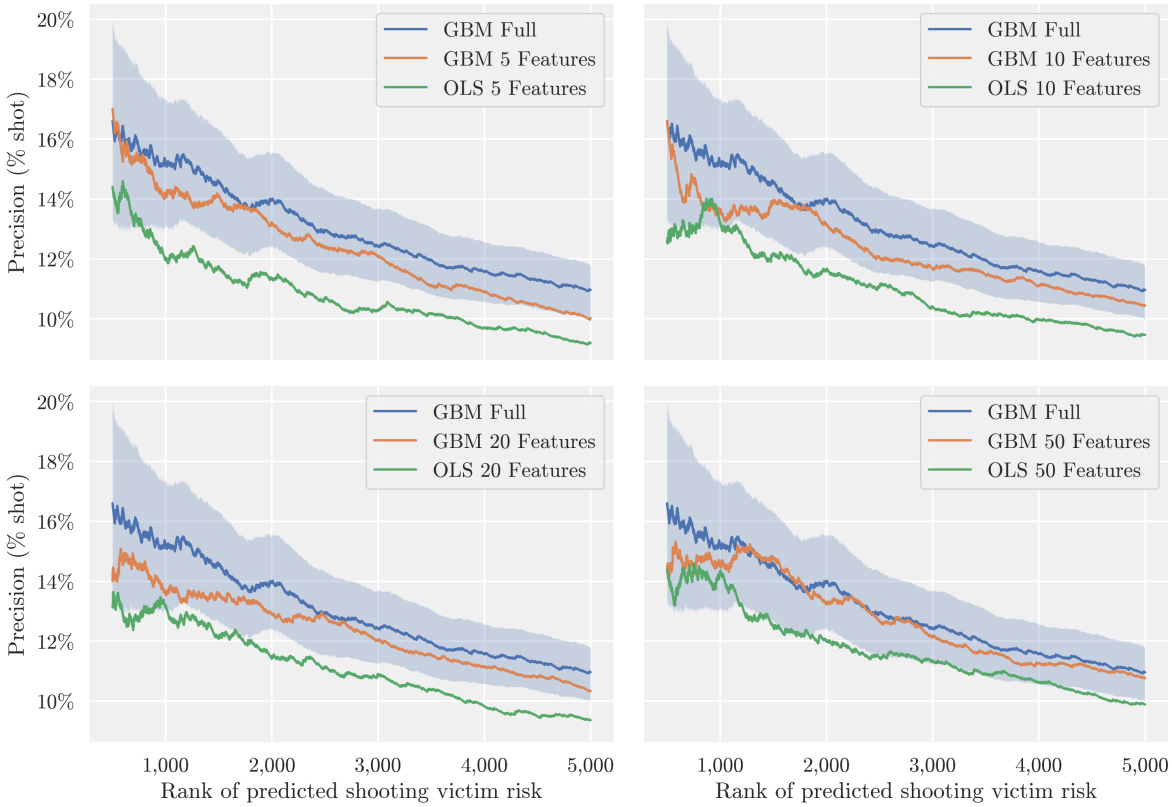
features and the selected feature. To do so, we replace the value of each unselected feature with the residual from a linear regression of each unselected feature onto the feature with the highest correlation. We then repeat the process, searching each time for the feature with the highest remaining correlation with the outcome after removing the correlation with already-selected features. Given a particular collection of features to start with, this approach produces a rank ordering of the features in that collection with the highest linearly independent relationship with the outcome.

Finally, we build models using both GBM and ordinary least squares (OLS) using only the $n \in \{5, 10, 20, 50\}$ highest-ranked features, comparing their performance to that of the full model built using GBM with 1,411 features.

Appendix Table B.8 reports the set of 50 features chosen by this process. The first column shows the set to which each feature belongs; the second column provides a description of the feature, where text in parentheses indicate a subtype of the feature; the third column shows, where appropriate, the time window over which the feature was measured, where "Total" indicates features that look back to the beginning of the data (January 1999 for all features except those drawn from the shooting victim records, which start in January 2011); the fourth column shows the correlation between the residualized version of the feature and the outcome; and the fifth column shows the correlation between the unresidualized version of the feature and the outcome.

Appendix Figure B.5 reports the same precision plot as Figure 4, with separate panels for different numbers of the top n features reported in Appendix Table B.8. Each panel shows the precision for the full model, and for models using GBM and OLS with only the indicated top n features. Across the panels, the parsimonious GBM models perform similarly to the full model, with its precision usually falling within the full model's confidence interval. In contrast, the parsimonious OLS models, while appearing to improve slightly in performance with the number of features available to them, still perform more poorly: even with $n = 50$ features, the precision of the OLS model is between one and two percentage points lower than that of the full model, although these differences are not always statistically significant when accounting for sampling variation. This pattern of results suggests that it may be possible to achieve similar performance to the full model using a relatively small set of features using a flexible modeling technique like GBM or, as noted in section 4.3.2, OLS with interaction terms.

Figure B.5: Precision across models with different model types and number of features



Note: Figure shows precision, or the share of the $k \leq 5,000$ people with the highest predicted risk of shooting victimization who are shot during the 18-month outcome period, for the full model, a gradient boosting machine (GBM) model with a limited set of features, and an ordinary least squares (OLS) model with the same limited set of features (Appendix Table B.8). Due to noise in precision at low values of k, we start the graphs at $k = 500$. Bootstrapped 95 percent confidence interval for the full model shown (see Appendix A.5.2 for details).

To provide further insight into which predictors provide the most independent information, and to emphasize how the information in features can be substitutable, we repeat the stepwise residualization procedure described above for the other feature sets shown in Figure 4. Appendix Tables B.9, B.10, and B.11 respectively show the list of the top 50 features identified by this process for the following three feature sets: no network features, no own arrest features, and the combination of no network and no own arrest features.

We then reran our modeling process, but only gave the algorithm access to the 50 most correlated variables for each feature set, identified in our stepwise residualization procedure. The results, reported in Appendix Table B.12, are consistent with those shown in Appendix Figure B.5, confirming that it is generally possible to achieve comparable performance to the full model in the tail (at $k = 500$ and $k = 4,244$) with a limited number of features, even when restricting the sets from which those features are drawn. The exception, consistent with the results in Figure 4 in the main text, is a model that

removes all arrest information on both focal individuals and their neighbors (last row). The overall similarity in performance emphasizes the point in the main text that standard "importance" measures within a single model do not capture which variables are uniquely important to prediction; other correlated variables can often capture similar information when the "important" variables are removed. For this reason, it would be a mistake to assign any kind of causal interpretation to the importance of individual features at the top of Appendix Tables B.8, B.9, B.10, and B.11. To get a clearer understanding of which kinds of features truly matter, in the sense that their removal would harm predictive performance, we must compare predictive performance in models trained without particular variables, as in the main text.

Of course, it is typically impossible to know *a priori* which small set of features will achieve performance as close as possible to a model with access to the full set of features. The process of solving this constrained optimization problem is itself a machine learning challenge (Rudin, 2019). In practice, settings that require smaller numbers of features could engage in this process.[47]

[47] See Luminosity and York (2020) for a real-world example of developing a risk assessment for pretrial arraignment decisions in New York City.

Table B.8: Top 50 features from the stepwise residualization procedure when given access to all feature sets

Feature Set	Description	Time Window	Correlations Residualized	Original
arrests	# of own-arrests (any)	730 days	0.118	0.118
arrests	Affiliated gang (most recent)		0.069	0.111
networks	# of 1st degree neighbors victimized (shooting)	Total	0.052	0.098
arrests	# of own-arrests (larceny)	730 days	-0.034	0.014
demographics	Age (modal)		-0.031	-0.056
victims	# of victimizations (shooting)	Total	0.028	0.068
arrests	# of own-arrests (warrant arrests)	730 days	-0.025	0.049
arrests	# of own-arrests (weapons violation)	Total	0.024	0.079
arrests	Ever gang-affiliated		-0.022	0.105
demographics	Police beat (modal)		0.023	0.048
arrests	# of own-arrests (public alcohol consumption)	730 days	-0.020	0.015
arrests	# of own-arrests (reckless conduct)	Total	0.020	0.079
arrests	# of own-arrests (misc. municipal code violation)	730 days	-0.018	0.047
arrests	# of own-arrests (gambling)	Total	0.019	0.076
arrests	# of own-arrests (possession controlled substance)	730 days	-0.016	0.026
arrests	# of own-arrests (robbery)	Total	0.016	0.061
networks	# of victimizations (shooting) of 1st degree neighbors	Total	-0.016	0.097
networks	# of arrests (property crime) of 1st degree neighbors	365 days	0.019	0.080
arrests	# of own-arrests (simple battery)	730 days	-0.017	0.030
demographics	Missing date of birth		-0.014	-0.016
demographics	# of unique police beats		-0.015	0.031
arrests	# of own-arrests (soliciting)	730 days	0.012	0.057
arrests	# of own-arrests (criminal trespass-land)	730 days	-0.012	0.043
arrests	# of days since arrest (first property crime)		-0.012	-0.039
arrests	# of days since arrest (first)		0.012	-0.016
arrests	# of own-arrests (shooting)	730 days	-0.011	0.019
victims	# of victimizations (gun assault or battery)	270 days	0.011	0.036
demographics	Sex (most recent)		0.011	0.052
arrests	# of own-arrests (gun robbery)	730 days	-0.011	0.013
arrests	# of own-arrests (traffic violation)	730 days	-0.010	0.027
arrests	# of own-arrests (reckless conduct)	365 days	0.010	0.057
networks	# of victimizations (domestic) of 2nd degree neighbors	90 days	-0.010	0.057
arrests	# of own-arrests (aggravated battery)	Total	0.010	0.048
arrests	# of own-arrests (obstructing identification)	Total	0.009	0.034
arrests	# of own-arrests (drug paraphenelia possession)	730 days	-0.009	0.001
arrests	# of own-arrests (aggravated assault school employee)	Total	0.008	0.029
arrests	# of own-arrests (public alcohol consumption)	Total	-0.008	0.011
arrests	# of own-arrests (criminal trespass-vehicles)	270 days	0.008	0.036
networks	# of 1st degree neighbors arrested (gun battery)	60 days	-0.008	0.006
networks	# of 1st degree neighbors arrested (violent crime)	90 days	0.008	0.052
victims	# of victimizations (drug abuse)	Total	-0.008	-0.003
arrests	# of own-arrests (violent crime)	270 days	-0.008	0.028
arrests	# of own-arrests (fbi code 04a)	365 days	0.008	0.025
arrests	# of own-arrests (criminal trespass-real property)	730 days	-0.007	0.024
arrests	# of own-arrests (battery cause bodily harm)	Total	0.007	0.044
arrests	# of own-arrests (bail bond violation)	Total	-0.008	0.040
arrests	# of own-arrests (soliciting)	Total	0.008	0.066
arrests	# of own-arrests (heroin posession)	Total	-0.008	0.010
arrests	# of own-arrests (weapons violation)	Total	-0.007	0.008
arrests	# of own-arrests (criminal trespass-land)	180 days	0.007	0.029

Note: Features are listed in descending order of residualized correlation, except for the first feature. The first column shows the set to which each feature belongs. The second column provides a description of the feature. Text in parentheses indicate a subtype of the feature. The third column shows, where appropriate, the time window over which the feature was measured. Time windows listed as "Total" indicate features that look back to the beginning of our data (January 1999 for all features except those drawn from the shooting victim records, which start in January 2011). The fourth and fifth columns show the correlation between the residualized and unresidualized version of the feature and the outcome, respectively.

Table B.9: Top 50 features from the stepwise residualization procedure when not given access to network features

Feature Set	Description	Time Window	Correlations Residualized	Original
arrests	# of own-arrests (any)	730 days	0.118	0.118
arrests	Affiliated gang (most recent)		0.069	0.111
victims	# of victimizations (shooting)	Total	0.043	0.068
demographics	Age (modal)		-0.034	-0.056
arrests	# of own-arrests (larceny)	730 days	-0.034	0.014
arrests	# of own-arrests (weapons violation)	Total	0.027	0.079
arrests	# of own-arrests (warrant arrests)	730 days	-0.024	0.049
arrests	# of own-arrests (gambling)	Total	0.024	0.076
arrests	# of own-arrests (public alcohol consumption)	Total	-0.023	0.011
arrests	# of own-arrests (reckless conduct)	Total	0.023	0.079
arrests	Ever gang-affiliated		-0.023	0.105
demographics	Police beat (modal)		0.024	0.048
arrests	# of own-arrests (robbery)	Total	0.019	0.061
arrests	# of own-arrests (drug abuse)	730 days	-0.018	0.079
arrests	# of own-arrests (motor vehicle theft)	270 days	0.016	0.041
arrests	# of own-arrests (simple battery)	730 days	-0.015	0.030
demographics	Missing date of birth		-0.014	-0.016
arrests	# of own-arrests (panhandling)	Total	-0.014	-0.002
arrests	# of own-arrests (reckless conduct)	730 days	0.014	0.074
arrests	# of own-arrests (traffic violation)	730 days	-0.013	0.027
demographics	Sex (most recent)		0.011	0.052
arrests	# of days since arrest (first)		0.012	-0.016
arrests	# of days since arrest (first property crime)		-0.012	-0.039
demographics	# of unique police beats		-0.011	0.031
arrests	# of own-arrests (soliciting)	Total	0.011	0.066
arrests	# of own-arrests (misc. municipal code violation)	730 days	-0.010	0.047
arrests	# of own-arrests (gun robbery)	730 days	-0.010	0.013
arrests	# of own-arrests (soliciting)	365 days	0.010	0.045
arrests	# of own-arrests (heroin posession)	Total	-0.010	0.010
arrests	# of own-arrests (aggravated battery w/o firearm)	Total	0.010	0.022
victims	# of victimizations (gun assault or battery)	270 days	0.009	0.036
victims	# of days since victimization (last shooting)		0.009	-0.064
arrests	# of own-arrests (aggravated assault school employee)	Total	0.009	0.029
arrests	# of own-arrests (obstructing identification)	Total	0.009	0.034
arrests	# of own-arrests (gun assault or battery)	Total	0.008	0.036
arrests	# of own-arrests (gun battery)	730 days	-0.009	0.011
arrests	# of own-arrests (criminal trespass-land)	730 days	-0.008	0.043
arrests	# of own-arrests (burglary)	Total	0.008	0.039
arrests	# of own-arrests (mfg del heroin sch pub hs pk)	Total	0.008	0.030
arrests	# of own-arrests (weapons violation)	Total	-0.008	0.008
arrests	# of own-arrests (battery cause bodily harm)	Total	0.007	0.044
arrests	# of own-arrests (bail bond violation)	Total	-0.007	0.040
arrests	# of own-arrests (gang loitering)	Total	0.007	0.045
victims	# of victimizations (shooting)	180 days	0.007	0.036
arrests	# of own-arrests (firearm possession)	730 days	-0.007	0.017
arrests	# of own-arrests (weapons violation)	730 days	0.008	0.058
arrests	# of own-arrests (simple assault)	730 days	-0.007	0.024
arrests	# of own-arrests (criminal trespass-land)	180 days	0.007	0.029
arrests	# of own-arrests (drug paraphenelia possession)	730 days	-0.007	0.001
arrests	# of own-arrests (obstructing traffic)	Total	0.006	0.027

Note: See bottom of Table B.8 for column definitions.

Table B.10: Top 50 features from the stepwise residualization procedure when not given access to own-arrest features

Feature Set	Description	Time Window	Correlations Residualized	Original
networks	# of 1st degree neighbors ever gang affiliated	365 days	0.112	0.112
victims	# of victimizations (shooting)	Total	0.051	0.068
victims	# of victimizations (drug abuse)	Total	-0.042	-0.003
demographics	Age (modal)		-0.039	-0.056
demographics	Sex (modal)		0.039	0.052
demographics	Police beat (modal)		0.039	0.048
demographics	Missing date of birth		-0.021	-0.016
networks	# of 1st degree neighbors victimized (any)	365 days	0.016	0.088
victims	# of victimizations (any)	730 days	-0.019	-0.015
demographics	# of unique police beats		0.014	0.031
networks	# of 1st degree neighbors ever gang affiliated	180 days	-0.014	0.104
networks	# of arrests (property crime) of 1st degree neighbors	365 days	0.017	0.080
victims	# of victimizations (reckless conduct)	Total	-0.012	-0.002
victims	# of victimizations (weapons violation)	Total	0.013	-0.001
victims	# of victimizations (gun battery)	Total	0.010	0.062
victims	# of days since victimization (last shooting)		0.011	-0.064
victims	# of victimizations (gun assault or battery)	270 days	0.011	0.036
demographics	Approximate age (modal)		-0.009	-0.053
networks	# of 1st degree neighbors arrested (gun battery)	60 days	-0.009	0.006
networks	# of 1st degree neighbors victimized (gun battery)	270 days	0.009	0.070
networks	# of 1st degree neighbors ever gang affiliated	270 days	-0.011	0.109
networks	# of arrests (drug deal) of 1st degree neighbors	730 days	0.011	0.070
networks	# of victimizations (domestic) of 2nd degree neighbors	90 days	-0.008	0.057
networks	# of victimizations (shooting) of 2nd degree neighbors	Total	0.010	0.092
networks	# of 2nd degree neighbors arrested (gun robbery)	60 days	-0.009	0.034
demographics	Race (most recent)		0.008	0.044
networks	# of 1st degree neighbors arrested (violent crime)	90 days	0.007	0.052
networks	# of 2nd degree neighbors victimized (property crime)	60 days	-0.007	0.051
demographics	Sex (most recent)		-0.006	0.052
victims	# of victimizations (property crime)	Total	-0.007	-0.034
networks	# of arrests (domestic) of 2nd degree neighbors	30 days	0.006	0.050
networks	# of 2nd degree neighbors		-0.007	0.016
networks	# of 2nd degree neighbors victimized (shooting)	Total	0.006	0.093
networks	# of 1st degree neighbors ever gang affiliated	730 days	-0.008	0.111
networks	# of 1st degree neighbors arrested (gun assault or battery)	Total	0.006	0.070
victims	# of days since victimization (first shooting)		0.006	-0.062
victims	# of victimizations (shooting)	180 days	0.007	0.036
networks	# of 1st degree neighbors victimized (shooting)	60 days	-0.006	0.027
victims	# of days since victimization (first)		-0.006	0.001
networks	# of 1st degree neighbors arrested (any)	30 days	0.005	0.084
networks	# of 1st degree neighbors ever gang affiliated	60 days	-0.006	0.090
networks	# of 1st degree neighbors victimized (gun assault or battery)	365 days	0.006	0.081
victims	# of victimizations (to land)	730 days	-0.005	-0.003
victims	# of victimizations (robbery)	Total	-0.005	0.004
networks	# of 2nd degree neighbors victimized (gun robbery)	30 days	0.005	0.029
networks	# of arrests (domestic) of 1st degree neighbors	Total	-0.005	0.052
victims	# of victimizations (marijuana posession)	Total	0.006	-0.002
victims	# of victimizations (reckless conduct)	Total	-0.006	-0.003
victims	# of victimizations (aggravated-handgun)	Total	0.004	0.051
victims	# of days since victimization (last gun battery)		0.005	-0.064

Note: See bottom of Table B.8 for column definitions.

Table B.11: Top 50 features from the stepwise residualization procedure when not given access to network or own-arrest features

Feature Set	Description	Time Window	Correlations Residualized	Original
victims	# of victimizations (shooting)	Total	0.068	0.068
demographics	Age (modal)		-0.052	-0.056
demographics	Sex (modal)		0.050	0.052
demographics	Police beat (modal)		0.048	0.048
demographics	Missing date of birth		-0.028	-0.016
demographics	# of unique police beats		0.020	0.031
victims	# of victimizations (property crime)	Total	-0.017	-0.034
victims	# of victimizations (aggravated-handgun)	Total	0.014	0.051
victims	# of days since victimization (last shooting)		0.012	-0.064
demographics	Approximate age (modal)		-0.012	-0.053
victims	# of victimizations (gun battery)	270 days	0.012	0.041
demographics	Race (most recent)		0.011	0.044
victims	# of victimizations (robbery)	Total	-0.008	0.004
victims	# of days since victimization (first shooting)		0.008	-0.062
victims	# of victimizations (simple domestic battery)	Total	-0.008	-0.015
victims	# of victimizations (gun battery)	Total	0.007	0.062
demographics	Sex (most recent)		-0.006	0.052
victims	# of victimizations (violent crime)	270 days	0.006	0.023
victims	# of days since victimization (last domestic)		0.005	0.013
demographics	Police beat (most recent)		0.005	0.047
victims	# of victimizations (larceny)	Total	0.004	-0.029
victims	# of victimizations (child abduction)	Total	0.004	0.007
victims	# of victimizations (fbi code 04a)	Total	-0.004	0.008
victims	# of victimizations (gun assault or battery)	180 days	0.004	0.031
victims	# of days since victimization (last gun battery)		0.004	-0.064
victims	# of victimizations (aggravated battery)	730 days	0.004	0.038
victims	# of victimizations (drug abuse)	Total	-0.004	-0.003
victims	# of victimizations (simple assault)	730 days	-0.004	-0.009
victims	# of victimizations (fbi code 04a)	730 days	0.003	0.010
victims	# of victimizations (credit card fraud)	Total	0.003	-0.009
victims	# of victimizations (child endangerment)	Total	0.003	0.011
victims	# of victimizations (gun battery)	60 days	-0.003	0.017
victims	# of victimizations (shooting)	180 days	0.003	0.036
victims	# of victimizations (shooting)	730 days	-0.003	0.056
victims	# of victimizations (agg po hands no min injury)	Total	-0.003	-0.005
victims	# of victimizations (aggravated–other dangerous weapon)	Total	0.003	0.010
victims	# of victimizations (fbi code 2)	Total	-0.003	-0.006
victims	# of victimizations (fbi code 04a)	270 days	-0.003	0.007
victims	# of victimizations (gun assault or battery)	730 days	0.003	0.048
victims	# of days since victimization (first part one violent crime)		0.003	-0.028
victims	# of days since victimization (first)		-0.003	0.001
victims	# of days since victimization (first property crime)		0.004	0.028
victims	# of victimizations (pocket picking)	730 days	0.003	-0.003
victims	# of victimizations (attempt armed handgun)	Total	0.003	0.006
victims	# of victimizations (aggravated domestic battery)	Total	0.003	0.008
victims	# of victimizations (to property)	Total	0.002	-0.019
victims	# of victimizations (telephone threat)	Total	-0.003	-0.012
victims	# of victimizations (violate order of protection)	Total	0.002	-0.007
victims	# of victimizations (aggravated)	Total	-0.002	-0.002
victims	# of victimizations (to vehicle)	Total	-0.002	-0.017

Note: See bottom of Table B.8 for column definitions.

Table B.12: Predictive performance for limited feature sets chosen by the stepwise residualization procedure

Feature Set	Top 500				Top 4,244			
	True Positives	Precision	Recall	Total Recall	True Positives	Precision	Recall	Total Recall
All Feature Sets	83 (67, 100)	0.166 (0.134, 0.200)	0.029 (0.024, 0.035)	0.020 (0.016, 0.024)	486 (444, 527)	0.115 (0.105, 0.124)	0.172 (0.157, 0.186)	0.115 (0.105, 0.124)
All Feature Sets - Top 50	72 (58, 87)	0.144 (0.116, 0.174)	0.025 (0.021, 0.031)	0.017 (0.014, 0.020)	477 (435, 517)	0.112 (0.102, 0.122)	0.169 (0.154, 0.183)	0.112 (0.102, 0.122)
No Network Information - Top 50	74 (59, 89)	0.148 (0.118, 0.178)	0.026 (0.021, 0.031)	0.017 (0.014, 0.021)	435 (395, 473)	0.102 (0.093, 0.111)	0.154 (0.140, 0.167)	0.102 (0.093, 0.111)
No Own Arrests - Top 50	76 (60, 90)	0.152 (0.120, 0.180)	0.027 (0.021, 0.032)	0.018 (0.014, 0.021)	427 (389, 466)	0.101 (0.092, 0.110)	0.151 (0.138, 0.165)	0.101 (0.092, 0.110)
No Own Arrests, No Network Information - Top 50	51 (39, 65)	0.102 (0.078, 0.130)	0.018 (0.014, 0.023)	0.012 (0.009, 0.015)	381 (342, 415)	0.090 (0.081, 0.098)	0.135 (0.121, 0.147)	0.090 (0.081, 0.098)

Note: Performance and recall from models trained to predict shooting victimization during the 18-month outcome period starting April 1, 2018. Models differ based on the feature sets available to them during training. Model performance is evaluated on shooting victimization during the outcome period, for the $k = 500$ and $k = 4,244$ people with the highest predicted risk of shooting victimization. Bootstrapped 95 percent confidence intervals are in parentheses (see Appendix A.5.2 for details).

References

Angelino, Elaine, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin (2018). "Learning certifiably optimal rule lists for categorical data." *Journal of Machine Learning Research* 18, pp. 1–78.

Bhatt, Monica P., Sara B. Heller, Max Kapustin, Marianne Bertrand, and Christopher Blattman (2024). "Predicting and Preventing Gun Violence: An Experimental Evaluation of READI Chicago." *The Quarterly Journal of Economics* 139 (1) (1), pp. 1–56.

Billings, Stephen B., David J. Deming, and Stephen L. Ross (2019). "Partners in crime." *American Economic Journal: Applied Economics* 11.1, pp. 126–150.

Caruana, Rich, Nikos Karampatziakis, and Ainur Yessenalina (2008). "An empirical evaluation of supervised learning in high dimensions." *Proceedings of the 25th International Conference on Machine Learning*, pp. 96–103.

Cook, Philip J., Ariadne E. Rivera-Aguirre, Magdalena Cerdá, and Garen Wintemute (2017). "Constant lethality of gunshot injuries from firearm assault: United States, 2003-2012." *American Journal of Public Health* 107.8, pp. 1324–1328.

Dressel, Julia and Hany Farid (2018). "The accuracy, fairness, and limits of predicting recidivism." *Science Advances* 4.1.

Fogliato, Riccardo, Alexandra Chouldechova, and Max G'Sell (2020). "Fairness evaluation in presence of biased noisy labels." *International conference on artificial intelligence and statistics*. PMLR, pp. 2325–2336.

Friedman, Jerome H. (2001). "Greedy function approximation: a gradient boosting machine." *Annals of statistics*, pp. 1189–1232.

Jung, Jongbin, Connor Concannon, Ravi Shroff, Sharad Goel, and Daniel G. Goldstein (2017). "Simple Rules for Complex Decisions."

Ke, Guolin, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu (2017). "Lightgbm: A highly efficient gradient boosting decision tree." *Advances in Neural Information Processing Systems* 30, pp. 3146–3154.

Luminosity and Crime Lab New York (2020). "Updating the New York City Criminal Justice Agency Release Assessment." URL: `https://www.nycja.org/assets/downloads/Updating-the-NYC-Criminal-Justice-Agency-Release-Assessment-Final-Report-June-2020.pdf`.

McNeill, Melissa and Zubin Jelveh (2021). *Manual for Name Match*. URL: `https://github.com/urban-labs/namematch`.

Raskutti, Garvesh, Martin J. Wainwright, and Bin Yu (2011). "Early stopping for non-parametric regression: An optimal data-dependent stopping rule." *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1318–1325.

Rudin, Cynthia (2019). "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead." *Nature Machine Intelligence* 1.5, pp. 206–215.

Shwartz-Ziv, Ravid and Amitai Armon (2022). "Tabular data: Deep learning is not all you need." *Information Fusion* 81, pp. 84–90.

Stevenson, Megan T. and Christopher Slobogin (2018). "Algorithmic risk assessments and the double-edged sword of youth." *Behavioral Sciences & the Law* 36.5, pp. 638–656.

Stevenson, Megan T and Sandra G Mayson (2022). "Pretrial detention and the value of liberty." *Virginia Law Review* 108.3, pp. 709–782.

Ustun, Berk and Cynthia Rudin (2019). "Learning Optimized Risk Scores." *J. Mach. Learn. Res.* 20, pp. 150–1.

Yang, Crystal S. and Will Dobbie (2020). "Equal protection under algorithms: A new statistical and legal framework." *Michigan Law Review* 119, p. 291.