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

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

June 14, 2022

Abstract

This paper shows that shootings are predictable enough to be preventable. 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. We address central concerns about police data and algorithmic bias by predicting shooting victimization rather than arrest, which we show accurately captures risk differences across demographic groups despite bias in the predictors. Out-of-sample accuracy is strikingly high: of the 500 people with the highest predicted risk, 13 percent are shot within 18 months, a rate 130 times higher than the average Chicagoan. Although Black male victims more often have enough police contact to generate predictions, 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 have enormous preventive benefits: predictive accuracy among the top 500 people justifies spending up to $123,500 per person for an intervention that could cut their risk of being shot in half.

Heller: University of Michigan & NBER. Jakubowski: New York University. Jelveh: University of Maryland. Kapustin: Cornell University. 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, Emily Nix, Andy Papachristos, Greg Ridgeway, Mark Saint, Pat Sharkey, Ravi Shroff, and Megan Stevenson for their input and feedback. We thank Xander Beberman for outstanding research assistance. This research builds on a predictive model the authors developed to identify men at high risk of future gun violence involvement for referral into READI Chicago, an experimental preventive social service intervention. The larger research effort around READI Chicago was made possible with 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 Introduction

Gun violence in the U.S. causes widespread harm—to its direct victims and to the children, families, and communities surrounding them (Sharkey, 2018)—and generates social costs on the order of $100 billion annually (Cook and Ludwig, 2000). Because addressing this problem with more policing can generate its own significant social costs (e.g., Ang, 2021; Geller et al., 2014; Jones, 2014; Chalfin et al., forthcoming), local policymakers are spending millions of dollars to prevent gun violence with social services rather than law enforcement.[1] Whether these targeted preventive services can reduce shootings hinges on whether they reach people at high enough risk; even an exceptionally effective intervention will not stop gun violence if too few participants would be a victim or offender in its absence. Yet we know surprisingly little about whether we can predict someone's future shooting involvement accurately enough to make individual interventions a plausible solution to gun violence.

In other settings, 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, when most input data likely come from the criminal legal system, faces two key challenges. First, predicting outcomes as rare as shootings[2] 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; Martin et al., 2016; Salganik et al., 2020). Prediction may be particularly challenging given the noise and distortions in crime data.[3] Second, due to the criminal legal system's

[1] See, e.g., https://www.chicago.gov/content/dam/city/sites/public-safety-and-violenc-reduction/pdfs/OurCityOurSafety.pdf, https://www.phila.gov/2021-04-14-how-the-city-is-addressing-gun-violence-2021-update-to-the-roadmap-to-safer-communities/, https://www.oaklandca.gov/topics/oaklands-ceasefire-strategy, and https://monse.baltimorecity.gov/sites/default/files/MayorBMS_Draft_ViolenceReductionFrameworkPlan.pdf.

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

[3] Arrests or convictions of innocent people mean criminal legal records can overstate actual crim-

discriminatory treatment of minority groups, especially Black men (e.g., Antonovics and Knight, 2009; Eberhardt et al., 2004; Goncalves and Mello, 2021; Hoekstra and Sloan, 2022; Arnold et al., 2018; Abrams et al., 2012; Rehavi and Starr, 2014), an algorithm trained using data from this system may "bake in" bias—that is, predictions may differ across groups because of disparate treatment rather than (or in addition to) differences in behavior (e.g., Starr, 2014; Angwin et al., 2016; Lum and Isaac, 2016; Chouldechova, 2017; Kleinberg et al., 2017; Richardson et al., 2019; Mayson, 2019; Mehrabi et al., 2021). The risk of confounding discrimination or selective measurement with actual behavioral differences has led some academics to abandon crime prediction exercises altogether, because it is "simply too easy to create a 'scientific' veneer for racism."[4]

This paper demonstrates that even the biased information in police data can predict shootings with enough accuracy to save lives, and without distorting risk across demographic groups. The key is not to predict shooting *arrest*, which may capture shooting offending but also potentially biased police decisions about whom to arrest. Training algorithms on an outcome that measures the behavior of interest differentially by group is precisely the problem that yields predictions with "baked in" bias, which can get group differences in behavioral risk dramatically wrong (Mullainathan and Obermeyer, 2021; Obermeyer et al., 2019). And in the case of shootings, where we lack any unbiased measure of actual offending, we cannot even assess the extent of this bias (nor overall predictive performance).

Instead, we predict shooting *victimization*. Intervening with people at high risk of being victimized to keep them safe is a plausible alternative for reducing shootings to intervening with potential offenders (Cooper et al., 2006; Zun et al., 2006; Cheng et al., 2008; Green

inal behavior; low reporting rates for crimes like domestic violence mean not every offense is brought to the attention of law enforcement; and low clearance rates mean not everyone who commits a reported crime is arrested. And some cases are alleged to be outright data falsification; see, e.g., https://www.nydailynews.com/news/crime/fabricated-drug-charges-innocent-people-meet-arrest-quotas-detective-testifies-article-1.963021.

[4] Hundreds of mathematicians recently boycotted all efforts to work with police departments to predict crime for this reason. https://www.popularmechanics.com/science/math/a32957375/mathematicians-boycott-predictive-policing/

et al., 2017; Chalfin et al., forthcoming). And as we argue below, shooting victimization is likely to be measured consistently across demographic groups in our setting. Theoretical work suggests that this kind of well-measured outcome can avoid biased predictions even when the predictors themselves are biased (Kleinberg et al., 2018b). But in practice, there is very little data about whether shooting victimization is predictable, overall or by group.

To test this question, we build a model to predict shooting victimization in Chicago over an 18-month period. The model uses arrest and victimization records for 643,914 people from the Chicago Police Department (CPD), including over 1,400 predictors that capture a person's demographic information, arrest and victimization histories, and the arrest and victimization histories of peers who were co-involved in prior criminal incidents.

We have two 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, 13 percent are actually shot during the following 18 months—a rate almost 19 times higher than everyone in our prediction sample of people with recent police contact (0.7 percent across 327,127 people) and 130 times higher than everyone in Chicago (0.1 percent). An intervention that could cut the risk of being shot in half for these 500 people would generate an estimated social cost savings of $62 million from the victimization reduction alone (Cook and Ludwig, 2000; Ludwig and Cook, 2001). If the intervention cost less than $123,500 per person, it would pay for itself. Our analysis unpacks what information the model is using to achieve this predictive performance.

Second, the predictions do not misrepresent victimization risk across demographic groups. We show that Black male shooting victims are more likely to *have* a predicted risk, because they are more likely to have prior police contact.[5] This finding highlights how using police data limits an algorithm's ability to identify future victims with little or no prior police contact. But importantly, the *accuracy* of predictions is similar, on average, across race, age, and gender groups. The demographic composition of predicted

[5] Black men make up 69 percent of all shooting victims during the outcome period studied here. The model generates predictions for 75 percent of them, compared to half or fewer of the victims from other groups.

shooting victims matches almost exactly that of actual shooting victims. In other words, the predictions do not disproportionately inflate the victimization risk of Black men.

The impact of any predictive tool ultimately depends on how policymakers decide to use it and how it differs from non-algorithmic targeting. We show that one potential use—providing services to everyone above a high predicted risk threshold—will serve almost entirely Black men, and disproportionately younger Black men, in our setting. This likely reflects this group's higher true risk of shooting victimization and large number of total and recent police contacts, which could allow the algorithm to better distinguish risk within this group. But crucially, the fact that police data can successfully identify individual Black men at high risk of being shot means that re-purposing those data to target prevention efforts could help save Black lives.

Predicting shootings with police data is by no means a permanent solution to gun violence, and predictions should be used with care. It is particularly important to attend to whom this kind of algorithm misses and the dangers and limitations of using these kinds of 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 ensure that 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). A key remaining question is what kind of preventive services can reduce shootings for different parts of the risk distribution, which should be a priority for future research.

2 Related literature

The literature on using machine learning-based predictions to guide decision-making (e.g., Kleinberg et al., 2015, 2018a; Glaeser et al., 2016; Athey, 2017; Chouldechova et al.,

2018; Hastings et al., 2020; Obermeyer and Emanuel, 2016; Obermeyer et al., 2019) does not engage with the risk of shootings. But it does provide two key priorities for evaluating predictive models (also see discussion in Berk, 2008). First, predictions must be a true forecast, relying only on information available to the analyst at the time they are made. Second, a predictive model's 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. A third priority that is especially relevant in our context, as discussed above, is generating risk predictions that capture true differences in risk across demographic groups, rather than differences reflecting discrimination 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 used to predict the likelihood of 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.[6] 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 generation of criminal legal data.

We build on a handful of existing efforts that have made important progress toward testing whether shootings are predictable. Berk et al. (2009) use a machine learning strategy to generate true forecasts and carefully explore predictive performance. But partly

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

because the predictions were intended to target probation and parole services, they predict homicide charges rather than victimization. The use of 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 police and prosecutor decision-making about whom to arrest or charge, and relatively low clearance 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, but Wernick (2018) does not explore predictive performance overall, and neither explores performance by group. Chandler et al. (2011) predict the out-of-sample risk of being a shooting victim using Chicago Public Schools data and ordinary least squares. 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,b; 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 the time prevention 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.

This study improves upon and extends the prior literature by providing the three types of assessment needed to understand whether it is possible to predict who will be shot within a population without distorting risk across demographic groups. First, we

perform a pure forecasting exercise using only data available at the time of prediction. Second, we assess performance on data not used in the model-building process, including information about shooting victims who are not included in the prediction data at all. And third, by predicting an outcome much likelier to be 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 risk of being injured or killed by gunfire (shooting victimization) in the next 18 months using CPD data.[7] The key modeling decision we make is to predict reported shooting victimization rather than arrest. Reported victimization is much likelier to measure actual victimization consistently across demographic groups than arrest is to measure offending. Most shooting offenses do not result in an arrest in Chicago,[8] and the likelihood of a shooting resulting in an arrest may vary across groups due to, among other factors, differences in police behavior. In contrast, the Chicago data are consistent with most shooting victimizations, including non-fatal ones, being known to police,[9] leaving little scope for reporting to vary significantly across groups. This may

[7] This excludes suicides and incidents where people are shot by police officers. We focus on an 18-month outcome period because we are interested in determining who is at high risk over a period of time, rather than at one point in time. The former may be more appropriate for determining how to allocate preventive programming, which for a population at extremely high risk of gun violence can span months or years. For example, we used a related model to target a social service intervention in Chicago whose primary component is 18 months long; see https://osf.io/ap8fj/.

[8] The best available data suggest that, in the first half of the 2010s, under half of homicides and fewer than ten percent of non-fatal shootings in Chicago resulted in an arrest (Kapustin et al., 2017).

[9] We can see this by comparing the ratio of fatal to all shooting victims from non-police gun assaults in Chicago (16-18 percent in recent years) and from national estimates (22-26 percent) (Cook et al., 2017; Kaufman et al., 2021). If non-fatal shooting victimizations are being undererported in Chicago, then we would expect its ratio to be higher than one based on national estimates, which is not the case. Note that this comparison relies the assumption that fatal victimizations from non-police gun assaults are mostly free of underreporting, a widely-held view among researchers (Loftin and McDowall, 2010; Carr and Doleac, 2016).

be due, in part, to the requirement that healthcare providers in Illinois are mandated to report firearm injuries to law enforcement,[10] and the very high likelihood that shooting victims obtain medical care.[11]

We start with CPD data on 12.7 million event-level records between August 1999 and October 2019, containing information on demographics, arrests, and reported victimizations in Chicago for both youth and adults. We then limit our sample and use a probabilistic record linkage algorithm (Tahamont et al., 2019) to uniquely identify 643,914 people for our modeling process, construct detailed predictive features about each person in the sample, and train a model to predict their risk of shooting victimization in the following 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 require that they have at least one arrest or two reported victimizations in the 50 months prior (though importantly, we can observe all shooting victimizations even for individuals not in this prediction sample).[12] We then construct four categories of features for each person meeting this inclusion criterion. Demographic features include age, gender, race and ethnicity,[13] and police beats associated with home and incident addresses.[14] Arrest and victimization features include

[10] 20 ILCS 2630/3.2

[11] This high likelihood is widely reported across violence prevention, medical, and law enforcement practitioners in Chicago. There may still be some selective under-reporting of non-fatal shooting injuries by victims who self-treat or who live near the city border and seek care from providers outside CPD's jurisdiction, both of which in theory could vary by race. However, based on our conversations with practitioners, we think the magnitude of such selective under-reporting is likely to be quite small. Another potential source of measurement error that may be correlated with demographics, but also likely to be quite small in practice, is our procedure for linking records that lack a common identifier (see Appendix A).

[12] We exclude people with only one reported victimization due to their very low risk of future shootings and to reduce the influence of record-linkage error caused by poor data quality.

[13] It is worth noting that, in practice, there are many legal issues surrounding 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). As we show in Appendix B.3.2, however, there is enough information in the non-race features that excluding it leaves predictive power and calibration by race basically unchanged.

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

time-windowed counts of incidents, separately by incident type (e.g., robbery, shooting, vandalism).[15] 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.

Finally, network features include time-windowed counts of arrests and victimizations among people to whom the focal person is connected through co-involvement in prior criminal incidents ("neighbors"). For example, one feature counts the number of gun possession arrests in the past 180 days among neighbors to whom the focal person is connected directly ("first-degree neighbors"), while another counts 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. The full model includes 1,406 features in total.

We train and test a gradient-boosted decision tree model (Friedman, 2002). A key point of departure from traditional machine learning applications is the way we generate our hold-out test set. Because the predictors we define for person i include information about i's peers—arrest and victimization histories for those co-involved in incidents with i—information about i could still appear in the training data through i's peers, even if i is in a randomly subsampled 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 given our inclusion of network features—a situation which could lead to optimistic performance estimates. Instead, we divide the data into four calendar time cohorts (see Appendix Figure A.1). Each cohort has the same structure: a 50-month sample inclusion period followed by an 18-month outcome period. We use the first two cohorts to train the model and the third as a validation cohort for hyperparameter tuning. The final cohort is our test set, where we predict shooting risk for the 327,127 individuals in the

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

out-of-sample 18-month outcome period starting April 1, 2018.[16] All results speak to the predictive performance for this group during this period.

4 Results

Our main results describe how effectively the algorithm can predict future gun violence to help target prevention services, with particular transparency surrounding 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.[17] 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 available to the algorithm affects performance.

4.1 Performance overall

The top left panel of Figure 1 shows the overall distribution of the model's predictions and how they compare to realized rates of shooting victimization for the prediction sample. The x-axis is the average predicted risk for each percentile of the risk distribution, with each point containing 1 percent of the sample, or 3,271 people. The y-axis is the actual rate of shooting victimization in the 18-month outcome period for the 3,271 people in each bin.

[16] People can appear in more than one cohort, so observations are not entirely independent across cohorts. However, the time-windowed predictors and outcomes are defined relative to each cohort's prediction date. As a result, even when a person appears in multiple cohorts, their predictors and outcomes are defined over different time periods. Most importantly, the cohorts' outcome periods do not overlap, ensuring that the outcomes in the test set are never included during model training.

[17] Given how rare shootings are in the overall population, a common performance metric like accuracy, defined as the share of all predictions made correctly, will mostly be driven by correctly classifying people who are not shot. Another common metric, the area under the Receiver Operating Characteristic curve, or AUC, also characterizes performance across the entire risk distribution. In this paper, we are most interested in performance at the top of the risk distribution, since that will determine how effective any selection process for preventive services will be. As a result, we focus on performance measures at or above approximately the top 1 percent, recognizing that differences in AUC across models may not reflect differences in accuracy among the highest-ranked predictions.

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, albeit with some under-prediction for people in the right tail and some over-prediction for people in the highest-risk bin. 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.

Figure 2 reports two measures of model performance across the predicted risk distribution. Figure 2a shows Precision$_k$, or the share of people who are actually shot during the 18-month outcome period among the k people with the highest predicted risk:

$$\text{Precision}_k = \frac{\sum_{i=1}^{k} \mathbb{1}[\text{Shooting victim}_i = 1]}{k}$$

Figure 2b shows Recall$_k$, or the share of actual shooting victims during the 18-month outcome period who are among the k people with highest predicted risk:[18]

$$\text{Recall}_k = \frac{\sum_{i=1}^{k} \mathbb{1}[\text{Shooting victim}_i = 1]}{\text{Total shooting victims}}$$

We show two versions of recall in Figure 2b. The first, simply labeled recall, uses the total number of shooting victims in the prediction sample as the denominator, or 2,253. The second, labeled total recall, uses the total number of shooting victims in the entire city during the outcome period as the denominator, or 3,381. The difference between these two 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 therefore not assigned a predicted risk by the model. Though it is more

[18] 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.

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 during 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, 13 percent, or 65 people, are shot. This is almost 19 times higher than the base victimization rate for the prediction sample (327,127 people) of 0.7 percent, and 130 times the city-wide victimization rate (2.7 million people) of 0.1 percent. Among the $k = 3,381$ people with highest predicted risk—corresponding to the actual number of shooting victims during the 18-month outcome period—almost 9 percent are shot. Those at higher predicted risk for shooting victimization are also at significantly elevated risk for other adverse outcomes, like shooting arrest and violent victimization (Appendix Table B.1).

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 almost 2 percent of the 3,381 total victims during the 18-month outcome period.[19] The $k = 3,381$ people with highest predicted risk—just over 0.1 percent of the city's population—include almost 9 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 $k = 327,127$ in Figure 2b. Second, some eventual victims are assigned a low predicted risk that leaves them outside the top $k = 500$ or even $k = 3,381$. This may be partly

[19] Considering only the 2,253 shooting victims in the prediction sample rather than all 3,381 victims, recall at this threshold is 2.9 percent.

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 $k = 3,381$ with highest predicted risk have almost four times as many arrests in the prior year as eventual victims not among the $k = 3,381$ (2.7 vs. 0.7).

A major concern about using police data for prediction is that racial differences in those data may not arise from true differences in behavioral risk, but rather from differences in police behavior toward racial groups. Taking advantage of the fact that our relatively well-measured outcome allows us to compare predicted and actual risk for the behavior of interest, we next turn to showing how much these differ by race, gender, and age.

4.2 Performance by group

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.[20]) 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 a greater propensity among officers to stop, search, or arrest them, then police data will systematically misrepresent Black

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

individuals' behavior in a way that could generate inaccurate predictions of the shooting victimization risk they face.

The three remaining panels of Figure 1, which report calibration separately by race or ethnicity, show this is not the case. Each point is a bin containing 1 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 2.8 times higher. Yet importantly, the slopes of each line show 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. If anything, the predictions for Black individuals may slightly *underestimate* risk among many of those in the top decile, as indicated by all but the last point being above the 45-degree line. In contrast, among the White and Hispanic individuals predicted to be at the highest risk of shooting victimization, these predictions may *overestimate* the risk they face, as indicated by the points below the 45-degree line (see Appendix B.1 for further quantification and discussion). In other words, these data yield predictions that are accurate on average about the risk of shooting victimization across racial groups.

Figures 3 and 4 provide a fuller accounting of how the use of police data shapes the demographic composition of the predictions relative to the demographic composition of those who actually end up being shot.[21] As shown in Figure 2b, two-thirds of shooting victims have enough prior police contact to appear in our prediction sample. To show who is missing from the prediction sample, Figure 3 compares the race/ethnicity and gender composition of all 3,381 shooting victims to the 2,253 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

[21] Additional detail about the demographic compositions of the samples presented in these figures can be found in Appendix Table B.2.

fewer of the victims from other demographic groups. This pattern is consistent with the over-representation of Black men in police data more generally, though in this case it may aid predictions about shootings since Black men comprise the largest share of all victims (69 percent). A key implication of Figure 3 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 4 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 provide an additional breakdown of who is included or missed in the sample by showing the percentage of victims in each race, gender, and age group for all 3,381 shooting victims city-wide (first row) compared to the demographics of the 2,253 victims in the prediction sample (second row). Further dividing the data from Figure 3 by the median age of shooting victims, 23, does not change the picture of sample selection; comparing the second row to the first shows that Black male victims in both age categories 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 327,127 people in the prediction sample while weighting each person by his or her predicted risk of victimization (see Appendix A.5.1 for details). Comparing the second and third rows demonstrates that the accuracy of the model's predictions do 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 Black male victims in both age groups. Additional detail on predictive performance by demographics is in Appendix B.1.

If we were predicting an outcome like arrest, we would be unable to determine whether the demographic distribution of predictions was due to true differential behavioral risk or to differential police decision-making about whom to arrest from each group. In our

case, however, we predict an outcome that captures the true behavior of interest (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). What we additionally observe in our context is that, when the outcome is well-measured across groups, the algorithm can further produce probabilities that are accurate on average across them as well.

So far, we have discussed predictive performance and the demographic composition of predicted victims across the entire sample. As suggested by the dramatically different risk distributions by race/ethnicity in Figure 1, however, any decision rule that offers prevention services to everyone above some threshold of predicted risk in this setting will end up serving a disproportionately Black population, as well as likely over-predict risk for Hispanic and White individuals. The fourth row of Figure 4 shows the demographic implications of one such threshold rule as a stylized example: serving the 3,381 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 at all. 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 risk distribution does not indicate falsely inflated risk. In fact, even within this above-threshold group, the model is the most accurate for young Black men, while (consistent with the right side of each panel in Figure 1) over-predicting for most other groups (see Appendix Table B.4 for performance measures by group under this decision rule). 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.[22]

There are, of course, normative fairness questions involved with any way of allocating a limited amount of 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 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. Depending on where it is drawn, this kind of threshold allocation rule would likely miss almost all female victims and have higher false discovery rates for White and Hispanic men.

4.3 What matters for performance?

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 (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, while at the same time not materially chang-

[22] For example, if the risk of domestic violence shooting victimization is harder to predict using information contained in police data, and if a larger share of female shooting victims are injured or killed in such incidents, then this may help explain why the model assigns low predicted risk to these victims. In contrast, if young Black men are disproportionately victimized in shooting incidents that are easier to predict using information contained in police data, then this may help explain why the model assigns higher predicted risk to these victims. We cannot explore these issues directly since we have no information about the nature of the incident in which someone was shot, but this would be a useful avenue for future work.

ing 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 *groups* of features by substantive type and retraining the model each time.

Figure 5 reports precision for the full model and three other models that each exclude certain feature sets.[23] The x-axis ranks everyone in the prediction sample by their predicted risk of victimization, with highest predicted risk on the left. For each rank k, the y-axis reports the precision, or share actually victimized during the outcome period, of the people with that predicted risk or higher. For example, for the full model, among the $k = 1,000$ people with the highest predicted risk, just under 11 percent are shot during the outcome period; among the top $k = 4,000$ people, approximately 9 percent are shot. Because noise in our precision measure increases as k, the number of people above a predicted risk threshold, shrinks, we start the graph at $k = 500$. A bootstrap 95 percent confidence interval is plotted around each model.[24]

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, 2019), arrest data contain errors, may be subject to intentional manipulation, and are shaped in significant part by officer behavior. In the extreme, if arrest data provide little signal about individual behavior, then even if individual behav-

[23] Performance measures for additional models excluding different feature sets are reported in Appendix B.3.

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

ior plays a large role in a person's risk of shooting victimization, arrest data would provide little predictive power. Separately, Green et al. (2017) shows that network information may be useful in predicting shooting victimization, particularly if, as they argue, gun violence risk propagates through a social network as people co-engage in risky behavior with their peers.

As Figure 5 shows, excluding either of these feature sets does not appear to significantly degrade performance in the tail, while excluding both degrades it substantially.[25] One hypothesis that might explain this finding is that both feature sets contain similar information about a person's engagement in risky behavior. When one feature set is excluded, the other may substitute for it.[26] Yet when both are excluded, the model no longer has access to this signal about a person's behavior that is valuable for predicting their shooting victimization risk.[27]

5 Discussion

This paper demonstrates that re-purposing police data allows us to identify small groups 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.

[25] The precision of the "no own arrests" model appears to be *above* that of the full model from approximately $k = 1,000$ to $k = 4,500$, as is the "no network information" model from approximately $k = 2,000$ to $k = 4,500$. This result seems counterintuitive, suggesting that model performance improves with *less* information. It is worth noting that the "no own arrests" and "no network information" models fall within the 95 percent confidence interval of the full model, and their performance is not statistically distinguishable from that of the full model. We note that our models are trained to optimize performance across the full distribution, not only in the upper tail shown here. However, the AUC of the full model (0.872) is not appreciably higher than the AUCs of the "no network information" (0.873) and "no own arrests" (0.866) models. These patterns suggest that making more features available to the model can increase the signal available to it but also its complexity, leaving the net effect on performance ambiguous. It may be possible, for example, to obtain marginal improvements in performance by employing techniques, such as stacking (Wolpert, 1992), to further reduce variance beyond our current hyperparameter tuning procedure.

[26] While the performance of the full, "no own arrests," and "no network information" models are similar in the tail, the groups of people they identify as being at high predicted risk only partially overlap. For example, among the top 3,381, the overlap between the two models in those predicted at highest risk is 60.0 percent and increases to 69.5 percent for the top 10,000.

[27] For additional analyses exploring the sensitivity of the model's performance to the number of features and modeling complexity, see Appendix B.3.3.

For example, the 500 people with the highest predicted risk represent just 0.02 percent of Chicago's population but almost 2 percent of its shooting victims over an 18-month period. Using estimates of the social cost of a shooting (Cook and Ludwig, 2000; Ludwig and Cook, 2001), this amount of gun victimization generates a social cost of just over $123 million. Cutting this group's risk in half justifies spending $123,500 for each of the 500. The algorithm could also help target larger interventions: the 3,400 people with highest predicted risk are 0.1 percent of Chicago's population but account for about 9 percent of its shooting victims during the outcome period. At a social cost of $562 million, reducing this risk by half would justify spending $82,650 for each of the 3,400. The fact that it is possible to anticipate who so many shooting victims will be is a strong argument for trying to prevent their victimization, given its staggering cost.

Of course, identifying those in need of preventive services is only the first step. Preventing shooting victimization also requires understanding why a person is at high risk of it and what might reduce that risk. A person at high risk may not be highly responsive to any given intervention. And research about social service interventions' effectiveness at reducing gun violence for this population is relatively limited.[28] Generating evidence about who is responsive to which kinds of prevention efforts, and how that varies across the risk distribution, should be a high priority.

It is also crucial to acknowledge that even a model capable of identifying a group of people at very high risk of being involved in gun violence will get that prediction wrong for many—in our case, most—people in the group. When that happens, the costs of misdirecting an intervention can vary significantly. Providing a slot in a social program to someone whose actual risk is much lower than predicted, for example, incurs an important

[28] 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).

opportunity cost but is unlikely to harm the recipient.[29] 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 high costs on the recipient (Stevenson and Mayson, 2021) and those around them.

There are other reasons to be cautious about using predictions of shooting victimization risk to target proactive policing efforts. In addition to the potential legal barriers posed by using any algorithmic predictions for such targeting,[30] these proactive policing efforts are designed to intervene with (and prevent the actions of) future *offenders*, not the future *victims* we seek to predict. 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 methods used by police. 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.

But the results in this paper suggest that the solution is not to ignore the ability of police data to predict shooting victimization altogether; the counterfactual of *not* using information that could improve the effectiveness 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

[29] 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.

[30] Predictions based partly on prior police behavior may not meet the requisite standards for performing certain police actions that can infringe on a person's civil liberties. For example, it is unclear whether an algorithmic prediction would be sufficient to meet the "reasonable suspicion" standard necessary to effect a traffic stop or the "probable cause" standard necessary to effect an arrest, particularly if police could justify future action against someone by interacting more with them today in a way that raises their predicted 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, 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.

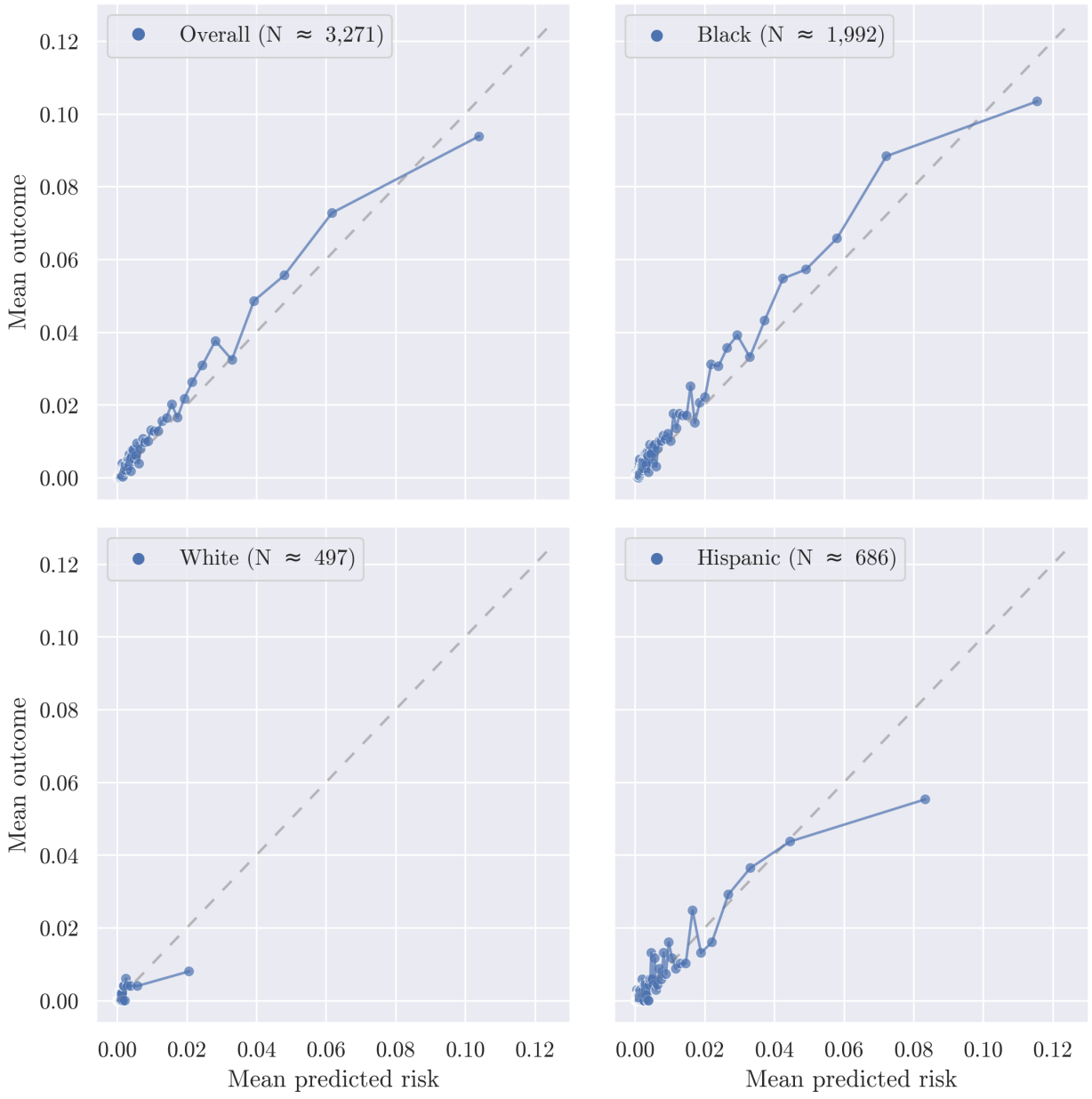
One example of how algorithmic prediction can be used to direct gun violence prevention services is an experiment currently underway in Chicago.[31] In that setting, a predictive model closely related to the one studied here identifies men at very high risk of involvement in future gun violence. Publicly available information about them is provided to community violence prevention organizations, who offer the men a chance to voluntarily participate in an intervention designed to reduce their risk. No information about them is shared with law enforcement. Other men who could benefit from the intervention are identified by the community organizations themselves, or by jail, prison, and parole staff. In this way, the model is a complement to, rather than a substitute for, human expertise; it helps find people who could benefit from programming but who might not otherwise be found. Crucially, this approach was developed in consultation with people who live in the affected communities.

The key insight of this paper is that an algorithm trained on police data—which is readily available in most cities—to predict a well-measured outcome can be a useful tool for preventing morbidity and mortality from gun violence. Training the algorithm on shooting victimization rather than arrest ensures that it is predicting the outcome of interest, rather than whom police decide to arrest (Obermeyer et al., 2019; Mullainathan and Obermeyer, 2021). This approach is best able to identify shooting victims who are Black men, a group constituting almost 70 percent of victims and that is over-represented in police data relative to victims from other groups. We note that the algorithm's ability

[31] For details, see https://osf.io/ap8fj/.

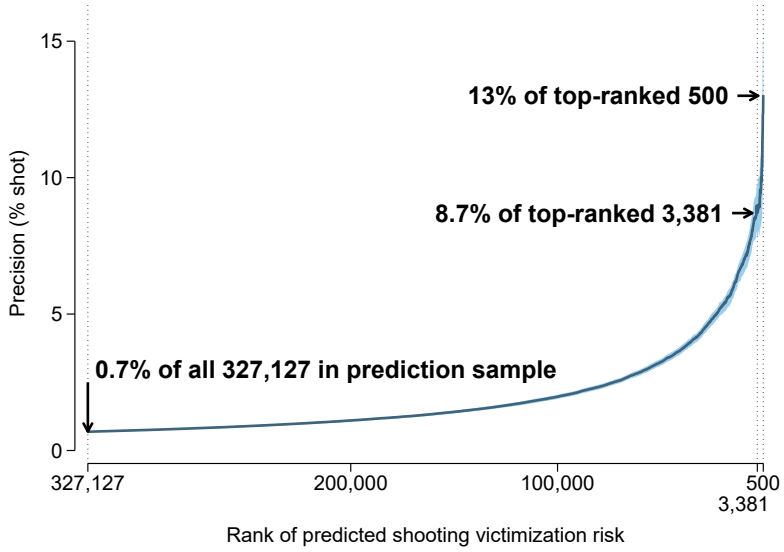
to predict gun victimization occurs in a social context where law enforcement is often the primary state institution enmeshed in the lives of Black men at high risk of gun violence. In a different context, where other government agencies and non-profit organizations more extensively engage with people 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 direct services that can help save lives.

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

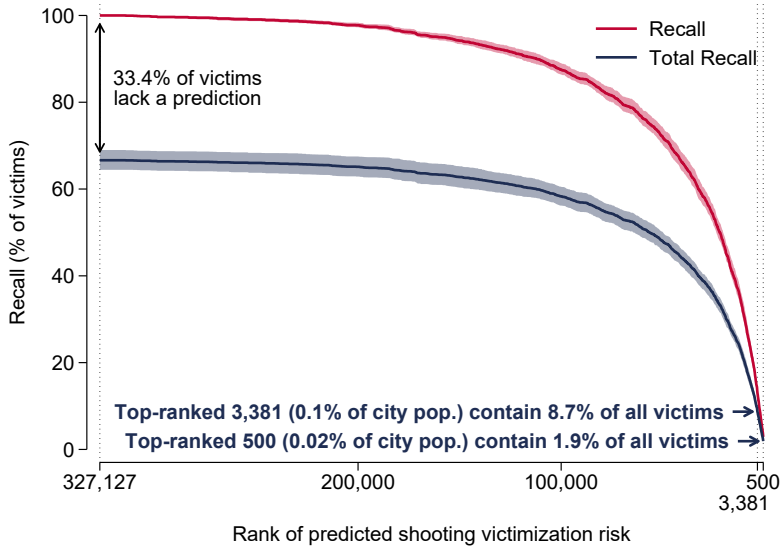


Note: Figure shows mean predicted risk and shooting victimization rate within each percentile of the overall (top left panel) and race/ethnicity-specific (remaining panels) predicted risk distributions. Race/ethnicity categories are mutually exclusive: non-Hispanic White, non-Hispanic Black, and Hispanic of any race.

Figure 2: Predictive performance for shooting victimization



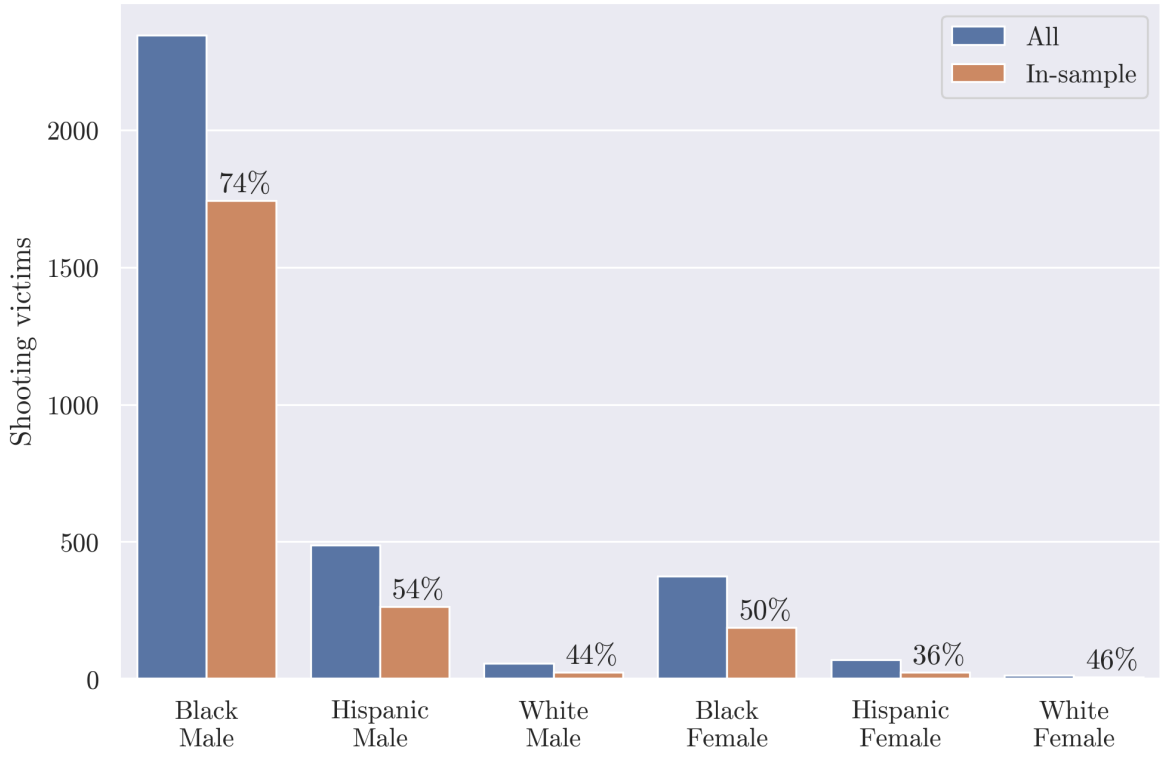
(a) Precision



(b) Recall

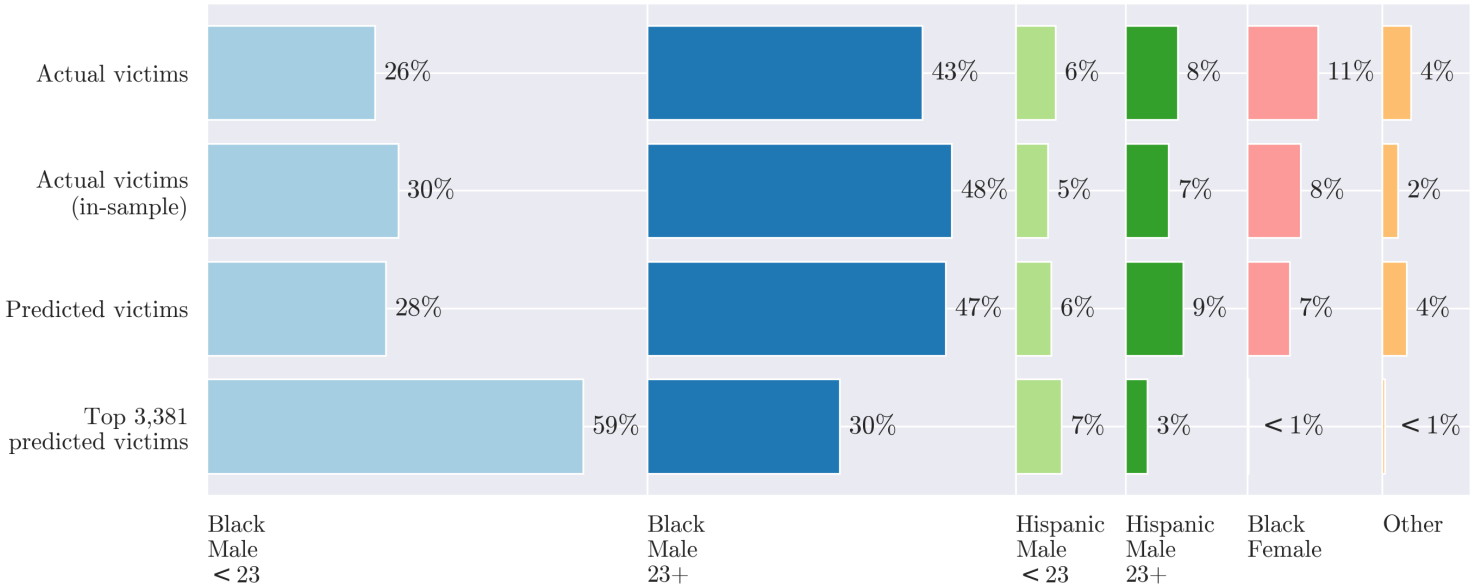
Note: Performance during the 18-month outcome period starting April 1, 2018 of the full model trained to predict shooting victimization. Precision shows share of the k people with the highest predicted risk who are actually shot during the 18-month outcome period. Recall shows the proportion of the 2,253 actual shooting victims in the prediction sample who are among the k people with highest predicted risk. Total recall shows the share of all 3,381 actual shooting victims in the city 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 of all victims and those in prediction sample



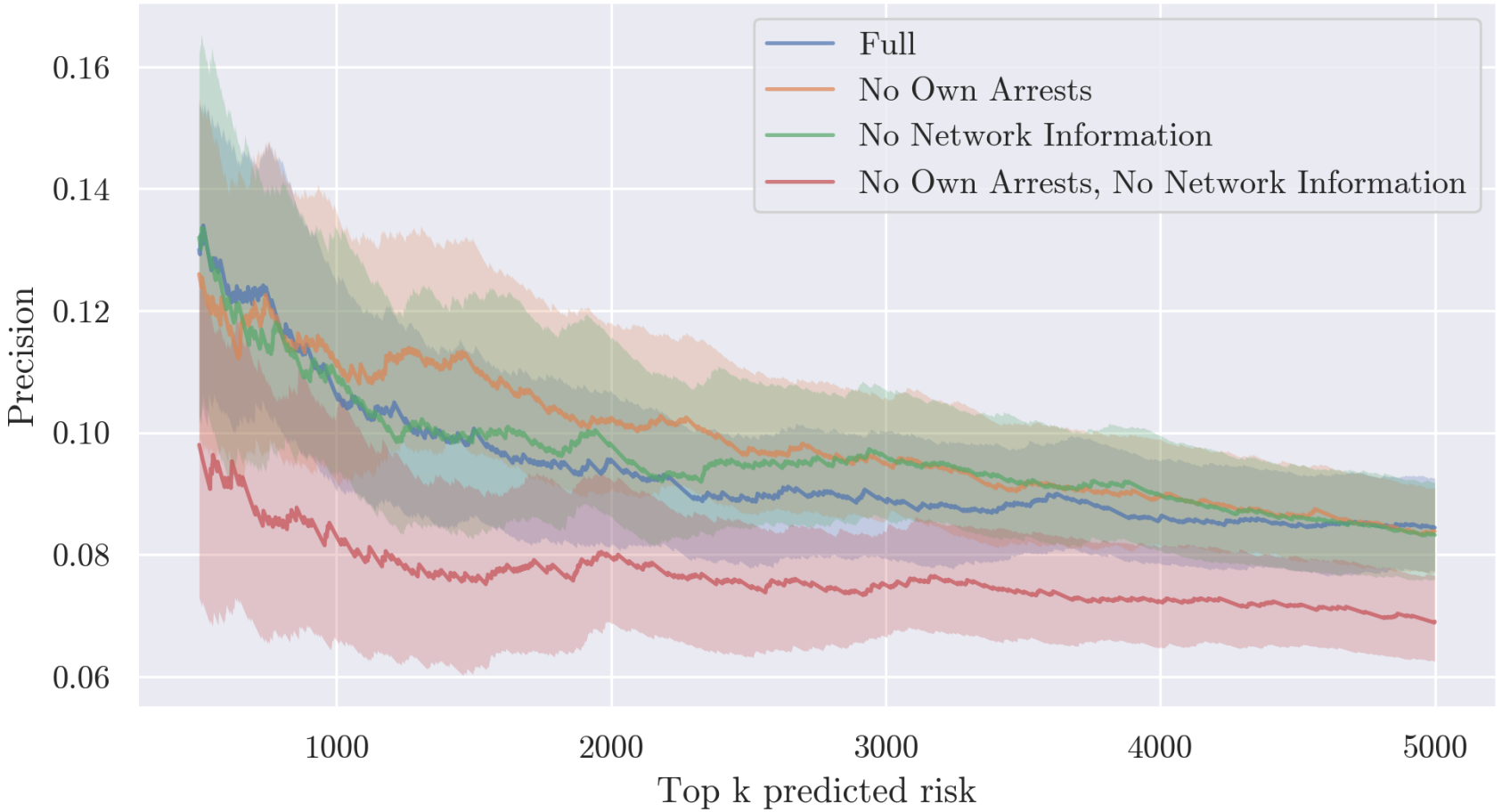
Note: Figure reports counts of shooting victims separately by race/ethnicity and gender, among all 3,381 shooting victims during the 18-month outcome period and the 2,253 victims in the prediction sample. Percentages above the in-sample bars report the share of all shooting victims in that demographic group (each blue bar) who appear in the prediction sample.

Figure 4: 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 3,381 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, individuals 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 horizontal bar) are based on the 327,127 people in the prediction sample reweighted by their predicted risk of victimization (see Appendix A.5.1 for details).

Figure 5: Precision across models with different feature sets



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

References

Abrams, David S., Marianne Bertrand, and Sendhil Mullainathan (2012) "Do judges vary in their treatment of race?," *Journal of Legal Studies*, 41 (2), 347–384.

Agan, Amanda and Sonja Starr (2017) "The effect of criminal records on access to employment," *American Economic Review*, 107 (5), 560–564.

Ang, Desmond (2021) "The Effects of Police Violence on Inner-City Students," *Quarterly Journal of Economics*, 136 (1), 115–168.

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, 1–78.

Angwin, Julia, Jeff Larson, Surya Mattu, and Lauren Kirchner (2016) "Machine Bias," 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), 163–177.

Arnold, David, Will Dobbie, and Crystal S. Yang (2018) "Racial Bias in Bail Decisions," *Quarterly Journal of Economics*, 133 (4), 1885–1932.

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

Berk, Richard (2008) "Forecasting Methods in Crime and Justice," *Annual Review of Law and Social Science*, 4, 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), 191–211.

Breiman, Leo (2001) "Random forests," *Machine Learning*, 45, 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*, 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, 39–53.

Carr, Jillian and Jennifer L Doleac (2016) "The geography, incidence, and underreporting of gun violence: new evidence using ShotSpotter data," *Incidence, and Underreporting of Gun Violence: New Evidence Using Shotspotter Data (April 26, 2016)*.

Chalfin, Aaron, Benjamin Hansen, and Rachel Ryley (forthcoming) "The minimum legal drinking age and crime victimization," *Journal of Human Resources*.

Chalfin, Aaron, Benjamin Hansen, Emily K Weisburst, Morgan C Williams et al. (forthcoming) "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), 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 preven-

tion intervention for assault-injured youths presenting to the emergency department: Results of a randomized trial," *Pediatrics*, 122 (5), 938–946.

Chouldechova, Alexandra (2017) "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments," *Big data*, 5 (2), 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," in *Conference on Fairness, Accountability and Transparency*, 134–148, PMLR.

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

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), 1324–1328.

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), 534–537.

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

Dressel, Julia and Hany Farid (2018) "The accuracy, fairness, and limits of predicting recidivism," *Science advances*, 4 (1).

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), 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), 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, 24–36.

Friedman, Jerome H (2002) "Stochastic gradient boosting," *Computational statistics & data analysis*, 38 (4), 367–378.

de la Garza, Ángel García, 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), 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), 2321–2327.

Glaeser, Edward L., Andrew Hillis, Scott Duke Kominers, and Michael Luca (2016) "Crowdsourcing city government: Using tournaments to improve inspection accuracy," *American Economic Review: Papers Proceedings*, 106 (5), 114–118.

Goncalves, Felipe and Steven Mello (2021) "A Few Bad Apples? Racial Bias in Policing," *American Economic Review*, 111 (5), 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), 326–333.

Hanson, R. Karl (2005) "Twenty years of progress in violence risk assessment," *Journal of Interpersonal Violence*, 20 (2), 212–217.

Harris, A. (2016) *A Pound of Flesh: Monetary Sanctions as Punishment for the Poor*, American Sociological Association's Rose Series: Russell Sage Foundation, https://books.google.com/books?id=c3MADAAAQBAJ.

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), 1917–1923.

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.."

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

Japkowicz, Nathalie (2000) "The Class Imbalance Problem: Significance and Strategies," in *In Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI)*, 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), 33–54.

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

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

Kaufman, Elinore J, Douglas J Wiebe, Ruiying Aria Xiong, Christopher N Morrison, Mark J Seamon, and M Kit Delgado (2021) "Epidemiologic trends in fatal and nonfatal firearm injuries in the US, 2009-2017," *JAMA internal medicine*, 181 (2), 237–244.

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

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer (2015) "Prediction policy problems," *American Economic Review: Papers & Proceedings*, 105 (5), 491–495.

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

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

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), 1094–1111.

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).

Loftin, Colin and David McDowall (2010) "The use of official records to measure crime and delinquency," *Journal of quantitative criminology*, 26 (4), 527–532.

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), 207–226.

Luh, Elizabeth (2019) "Not So Black and White: Uncovering Racial Bias from Systematically Misreported Trooper Reports," *Available at SSRN 3357063*.

Lum, Kristian and William Isaac (2016) "To predict and serve?," *Significance*, 13 (5), 14–19.

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

Martin, Travis, Jake M Hofman, Amit Sharma, Ashton Anderson, and Duncan J. Watts (2016) "Exploring limits to prediction in complex social systems," in *International World Wide Web Conference*.

Mayson, Sandra G. (2019) "Bias In, Bias Out," *Yale Law Journal*, 128 (8), 2122–2473.

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

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), 1–35.

Mello, Steven (2021) "Fines and Financial Wellbeing."

Mullainathan, Sendhil and Ziad Obermeyer (2021) "On the Inequity of Predicting A While Hoping for B," *American Economic Review: Papers & Proceedings*, 111, 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), 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), 447–453.

Otto, R.K. and K.S. Douglas (2010) *Handbook of Violence Risk Assessment*, International perspectives on forensic mental health: Routledge, https://books.google.com/books?id=p1JoYbAAN7QC.

Pager, Devah (2003) "The Mark of a Criminal Record," *American Journal of Sociology*, 108 (5), 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), 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), 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), 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), 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, 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), 52–59.
Raskutti, Garvesh, Martin J. Wainwright, and Bin Yu (2011) "Early stopping for non-parametric regression: An optimal data-dependent stopping rule," in *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 1318–1325.
Rehavi, M Marit and Sonja B Starr (2014) "Racial disparity in federal criminal sentences," *Journal of Political Economy*, 122 (6), 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), 192–233.
Rudin, Cynthia (2019) "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, 1 (5), 206–215.
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), 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), 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, 85–102.
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), 499–513.
Starr, Sonja B (2014) "Evidence-based sentencing and the scientific rationalization of discrimination," *Stan. L. Rev.*, 66, 803.
Stevenson, Megan T. and Sandra G. Mayson (2021) "Pretrial Detention and the Value of Liberty," https://papers.ssrn.com/abstract=3787018.
Stevenson, Megan T and Christopher Slobogin (2018) "Algorithmic risk assessments and the double-edged sword of youth," *Behavioral sciences & the law*, 36 (5), 638–656.
Tahamont, Sarah, Zubin Jelveh, Aaron Chalfin, Shi Yan, and Benjamin Hansen (2019) "Administrative Data Linking and Statistical Power Problems in Randomized Experiments," (25657).
Toloşi, Laura and Thomas Lengauer (2011) "Classification with correlated features: unreliability of feature ranking and solutions," *Bioinformatics*, 27 (14), 1986–1994.
Ustun, Berk and Cynthia Rudin (2019) "Learning Optimized Risk Scores.," *J. Mach. Learn. Res.*, 20, 150–1.
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), 770–788.
Wolpert, David H (1992) "Stacked generalization," *Neural networks*, 5 (2), 241–259.
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), 1164–1170.

Yang, Crystal S and Will Dobbie (2020) "Equal protection under algorithms: A new statistical and legal framework," *Mich. L. Rev.*, 119, 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), 8–13.

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 appendix provides additional details regarding our modeling process. First, we describe the raw 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 how we construct a set of cohorts and use them for model training and evaluation.

A.1 Data

Our model predicts a person's probability of becoming a shooting victim in the 18 months following the date of prediction, using information from 12.7 million CPD records. Available records describe 3,783,724 arrest, 8,911,412 victimization, 23,859 shooting, and 11,598 homicide events that occurred between August 1999 and October 2019 in Chicago.[32] We proceed to describe the relevant attributes of each of type of event record.

A.1.1 Arrest records

CPD arrest records include a unique person identifier (an *Illinois Record (IR) number*), based on a fingerprint scan, that allows us to construct a person's entire CPD arrest history. In addition to this person identifier, CPD arrest records include an incident identifier that can be used to link together the arrestees and victims associated with a single incident. The arrest data also contain police-recorded information about: arresting charges,[33] charge descriptions, and 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).

A.1.2 Victim records

While victimization records include an incident identifier, they do not include a unique person identifier. As such, the raw data do not allow us to construct a person's entire victimization history. However, the records provide each victim's identifying information (including name, home address, and date of birth), and therefore support probabilistic

[32] The shooting data start in 2010.

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

record linkage, described below. The victimization records also contain information about the type of victimization incident, including a description and UCR code.

A.1.3 Shooting and homicide victimization records

Similar to victimization records, shooting and homicide victimization records include an incident identifier but no unique person identifier. These records also include demographic information, facilitating record linkage. Shooting and homicide records are used to construct our outcome.

A.2 Record linkage

While CPD arrest records include a unique person identifier, victimization, shooting, and homicide 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 same person, we take the post-2010 IR number (the person identifier associated with arrest records) 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.[34] 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 non-arrest records—the various victimization incidents—and identifying the unique individuals represented across the victimization, shooting, and homicide data 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 post-2010 IR number. Second, a homicide record cannot link to another record if the homicide record's event date came before the other record's event date. Third, 73.2 percent of victimization 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 these false positives, we introduce a constraint that if at least one record in a record pair is missing date-of-birth information, 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, all other records in the cluster not missing date-of-birth information must have similar dates of birth.[35]

Our record linkage procedure identifies 5,426,703 individuals across the three decades of our data. We filter the set of clusters before training and evaluating our predictive models (see details in Appendix A.4.1).

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

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

A.3 Feature generation

Record linkage identifies the set of unique individuals represented in the CPD data, and associates each individual with their CPD arrest, victimization, shooting, and homicide record set. To predict an individual's risk of being shot as of a given prediction date, we aggregate over these associated records to construct *(person, prediction date)*-level features.[36] We construct four broad types of features: demographic, arrest, victimization, and network features. Table A.1 provides a summary of this final feature set, described by type below.

When an individual has no data in either the arrest or the victimization 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 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 an individual 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.1: 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	104
Arrest	Fine-Grained	365
Arrest	Gang	3
Victimization	Indexed	84
Victimization	Fine-Grained	233
Network	1^{st} and 2^{nd} Degree	598
Network	Centrality, degree	10
Total		1,406

A.3.1 Demographic features

We construct 13 demographic features from information on an individual's age, race, gender, and home address.[37] As with most administrative data, police records are often noisy, with different values of theoretically invariant characteristics appearing across multiple records for the same individual. We represent age and race using the modal value across

[36] When generating features for a given *(person, prediction date)*, we restrict to records available prior to the prediction date.

[37] For discussion regarding the inclusion of race in the model, see Appendix B.3.2.

an individual'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 who are only in the victimization records (i.e., have never been arrested). However, most of these records include an approximate age, which we use to construct an additional approximate age feature for each individual, as well as a similar missing indicator for approximate age information. [38] We represent gender using three separate features: an individual'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.

A.3.2 Arrest features

We construct 472 features summarizing an individual's prior arrest history. 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. Total 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 CPD arrest history (beginning in August 1999).

While these indexed arrest features provide a rich summary of an individual'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 365 fine-grained arrest features that count how many arrests an individual 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 an individual's prior CPD-identified gang affiliation. These measures include (i) an indicator of whether the individual has any prior gang-affiliated arrests, (ii) the number of unique gangs with which an individuals has been associated, and (iii) the most recent gang with which an individual is associated.

A.3.3 Victimization features

We construct 317 features summarizing an individual's history of victimization. Paralleling our treatment of prior arrests, we compute both indexed and fine-grained measures of prior victimization, using the same time windows and indices.

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

A.3.4 Network features

Since CPD arrest, victimization, shooting, and homicide records all share an event identifier, we construct a network using information on events within five years of the prediction date that includes two types of links: (i) links between co-arrestees, and (ii) links between arrestees and victims.[39] After constructing this network, we generate two types of features summarizing an individual's position within it.

First, we compute aggregate statistics describing an individual's network connections (whom we also 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 an individual'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 an individual's first- and second-degree network connections.

Second, we compute features describing the underlying network structure, including an individual'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 top tail of the risk distribution, we train and test a gradient-boosted decision tree model (Friedman, 2002). Because we include network features in the model, we cannot use traditional sub-sampling to generate a hold-out test set. Even if individual i were part of a randomly sub-sampled test set, information about i's risk could still be used in model-building to the extent he has peers in the training data and appears in their features. To avoid this kind of overfitting, we divide the data into calendar time cohorts as follows.

A.4.1 Defining cohorts

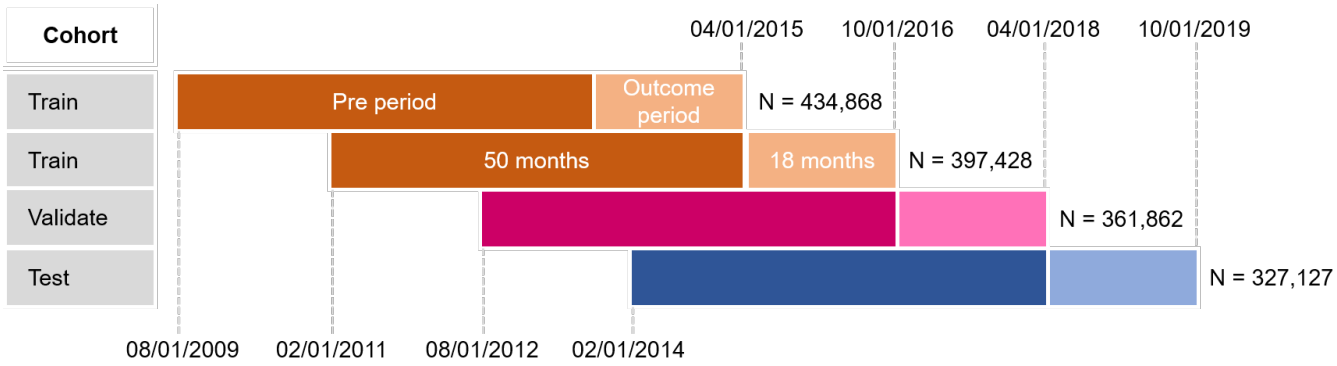
To define cohorts, we first establish four non-overlapping 18-month outcome periods. Then, we identify cohorts of individuals who have had either an arrest or two victimizations during the 50 months preceding each 18-month outcome period (see Figure A.1). This cohort definition excludes two sets of people: those with no CPD records from the past 50 months, and those who only had a single victimization in the past 50 months. We exclude these individuals for two reasons. First, they have much lower baseline risk: 0.01 percent of these individuals were shot during the follow-up period, compared to 0.7 percent of people in our cohort. Second, and as we describe above, we are more conservative in allowing links for victimization records without date-of-birth information. This leads to large number of singleton victimization records which remain unlinked. Therefore, dropping single victimizations implicitly reduces the influence of record-linkage error

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

caused by poor data quality. As such, our sample selection criteria reduce data integrity issues, while still capturing most of the identifiable population with elevated risk.

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 in the gradient boosting model via "early stopping" (Raskutti et al., 2011).[40] The final cohort is our test set, where we predict shooting risk ($\hat{p}$) for the out-of-sample 18-month outcome period starting on April 1, 2018.

Figure A.1: Model cohort structure



A.4.2 Hyperparameter tuning

We optimize the performance of our gradient-boosted decision tree model using random search over the following hyperparameters: number of leaves, minimum number of data 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 gradient-boosted decision tree model over the two training cohorts, 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).

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 cohorts.

[40] In this paper, we only report the results for the raw predictions, and not the predictions that are rescaled by the calibration model.

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 main 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 at k (Figure 4 and Appendix Figure B.4).

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 individuals in the prediction sample are weighted based on their predicted shooting risk $\hat{p}_i$. Specifically, for 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 predicted risk for the i^{th} individual. 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

Tables B.1 and B.5, as well as Figures 2, 4, and Appendix Figure B.4, include 95 percent bootstrap confidence intervals for several statistics at different k. These are constructed from 1,000 bootstrap samples, where each bootstrap sample is generated by:

1. Bootstrap resampling the prediction sample (i.e., drawing $N_{prediction} = 327,127$ 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, or explicit randomness in the gradient boosting algorithm).

B Additional Results

B.1 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, 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 shooting victimization in the police data and true 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 shootings. 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.1 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.1: Predictive performance of shooting victimization predictions for other outcomes

k	Precision	Recall	Total Recall
Shooting Victim			
500	0.130 (0.100, 0.160)	0.029 (0.022, 0.036)	0.019 (0.015, 0.024)
3,381	0.087 (0.078, 0.097)	0.130 (0.117, 0.146)	0.087 (0.078, 0.097)
327,127	0.007 (0.007, 0.007)	1.000 (0.958, 1.041)	0.666 (0.638, 0.694)
Shooting Arrest			
500	0.034 (0.018, 0.050)	0.047 (0.025, 0.070)	0.037 (0.019, 0.054)
3,381	0.020 (0.016, 0.025)	0.192 (0.148, 0.234)	0.148 (0.114, 0.181)
327,127	0.001 (0.001, 0.001)	1.000 (0.900, 1.097)	0.772 (0.695, 0.847)
Violent Crime Victim			
500	0.214 (0.178, 0.250)	0.007 (0.005, 0.008)	0.003 (0.002, 0.003)
3,381	0.170 (0.158, 0.182)	0.035 (0.032, 0.038)	0.014 (0.013, 0.015)
327,127	0.050 (0.049, 0.051)	1.000 (0.985, 1.015)	0.397 (0.391, 0.402)
Violent Crime Arrest			
500	0.178 (0.148, 0.216)	0.018 (0.015, 0.022)	0.012 (0.010, 0.014)
3,381	0.151 (0.138, 0.162)	0.103 (0.095, 0.111)	0.068 (0.062, 0.073)
327,127	0.015 (0.015, 0.016)	1.000 (0.974, 1.028)	0.658 (0.641, 0.676)

Note: Performance and recall from the full model trained to predict shooting victimization during the 18-month outcome period starting April 1, 2018. Model performance is evaluated on the four outcomes shown, for the k people with the highest predicted risk of shooting victimization. Violent crime arrest refers to the Part I violent index offenses: aggravated assault, aggravated battery, forcible rape, murder, and robbery. Prediction sample size is 327,127.

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, 3.4 percent are arrested on suspicion of carrying out a shooting (34 times the base rate in the whole test set of 0.1 percent); 21.4 percent are reported as victims of a violent offense (4.3 times the base rate); and 17.8 percent are arrested on suspicion of carrying out a violent offense (11.9 times the base rate).

B.2 Victim counts and performance by demographic group

Figures 2 and 3 in the main text show the proportion of shooting victims that fall into different demographic groups. This section adds some additional information to the summaries in the main text. To be transparent about the underlying size of each group, Appendix Table B.2 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 $k = 3,381$ people with the highest predicted

risk.

Table B.2: Demographic composition of actual and predicted shooting victims

Race	Gender	Age	Actual victims (N=3381)	Actual victims (in sample) (N=2253)	Predicted victims (rounded)	Top 3,381
Black	Male	<23	885	672	579	1976
		23+	1456	1072	966	1031
	Female	<23	159	65	39	1
		23+	214	123	98	4
Hispanic	Male	<23	210	112	114	236
		23+	276	152	188	121
	Female	<23	22	10	9	0
		23+	47	15	19	0
White	Male	<23	10	6	8	4
		23+	46	19	29	8
	Female		13	6	11	0
Other/Missing			43	1	9	0

Note: Counts for White females of all ages reported 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 under-prediction in the high-risk tail of the distribution for White and Hispanic individuals but over-prediction in the last bin. Appendix Table B.3 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. The model slightly under-predicts risk for Black men and women (by anywhere from 0.001 for older Black men and women to 0.004 for younger Black men), while it slightly over-predicts risk for older Hispanic men (by 0.001) and older White women.

Table B.3: 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	199192	0.01	0.008
	Female	All	77371	0.002	0.002
		<23	12141	0.005	0.003
		23+	65230	0.002	0.001
	Male	All	121782	0.014	0.013
		<23	18939	0.035	0.031
		23+	102842	0.01	0.009
Hispanic	All	All	68613	0.004	0.005
	Female	All	20333	0.001	0.001
		<23	3442	0.003	0.003
		23+	16890	0.001	0.001
	Male	All	48260	0.005	0.006
		<23	7639	0.015	0.015
		23+	40619	0.004	0.005
White	All	All	49710	0.001	0.001
	Female	All	18443	<0.001	0.001
		<23	1258	0.001	0.001
		23+	17183	<0.001	0.001
	Male	All	31235	0.001	0.001
		<23	2281	0.003	0.003
		23+	28943	0.001	0.001
Other Race/Gender		All	7287	<0.001	0.001
Missing Race/Gender/Age		All	2434	<0.001	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 total to N for the "All" row. The final row groups everyone with missing race/ethnicity, gender, and/or age information together.

Of course, average predictions being similar to base rates at a group level does not mean each individual'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 3,381 individuals with the highest predicted risk (motivated by the fact that there are 3,381 actual victims in the outcome period). Appendix Table B.4 shows precision and average predicted risk within race/ethnicity and age groups for the subset of men among the top 3,381 highest predictions. 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.4: Precision and average predicted risk by race and age for men among the top 3,381

Race	Gender	Age	N	Precision	Mean predicted risk
Black	Male	All	3007	0.092	0.102
		<23	1976	0.104	0.106
		23+	1031	0.069	0.096
Hispanic	Male	All	357	0.048	0.104
		<23	236	0.055	0.103
		23+	121	0.033	0.105
White	Male	All	12	0.0	0.198

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.

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 3,381" decision rule.[41]

Comparing the mean predicted risk with the realized risk (precision) in Appendix Table B.4 shows several key patterns. First, consistent with the subgroup calibration panels in Figure 1, predicted risk among this top tail is quite close to the realized risk for Black men, but overstates the realized risk for Hispanic men on average. The age breakdown further shows that the predictions are best calibrated for younger minority men, who tend to have fewer but more recent police contacts. The older minority men chosen with this decision rule tend to have elevated predictions relative to their realized risk.

In terms of classification among the top 3,381, the model has the highest true positive rate (and thus lowest false discovery rate) for Black men, of whom 9.2 percent are correctly classified, i.e., become shooting victims in the outcome period. In contrast, among Hispanic men—a much smaller group of 357 compared to 3,007 Black men—only 4.8 percent are correctly classified. This is consistent with argument in the main text that the over-representation of Black men in the top tail of the risk distribution is not because estimates of their risk are distorted (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 276 Black men and 17

[41] The reported numbers for White men have huge implicit confidence intervals since there are only 15 of them; while it is notable that none of these 15 is shot during the outcome period, we hesitate to over-interpret a pattern from so few individuals.

Hispanic men for whom preventive services might have kept them from serious injury or death. Nonetheless, the fact that 91 percent of Black men and 95 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.3 Further detail on what matters for prediction

B.3.1 Performance by groups of features

The main text presents predictive performance leaving out 3 sets of features: own arrests, peer information (networks), and both. We perform a similar exercise, dropping sets of features and retraining a new model, for additional combinations of features. Appendix Figure B.1 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. We do not show confidence intervals to make the figure more readable, but note that there is a fair amount of noise from sampling variation at any given k. Appendix Table B.5 quantifies the precision differences at $k = 500$ and $k = 3,381$, as well as reporting recall and total recall.

Figure B.1: Precision across models with different feature sets

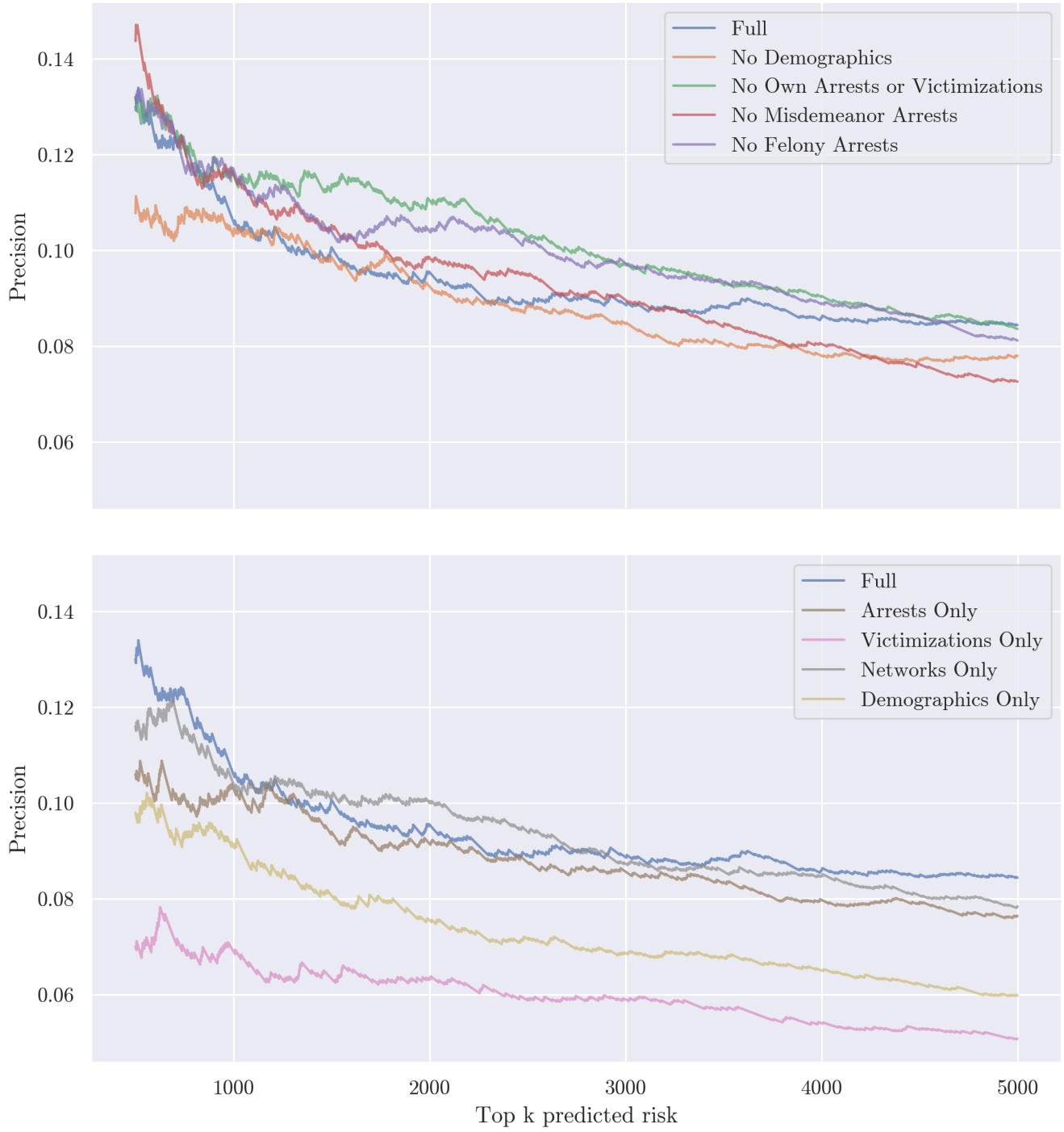


Table B.5: Predictive performance by feature set

	Top 500			Top 3,381		
Feature Set	Precision	Recall	Total Recall	Precision	Recall	Total Recall
Full	0.130 (0.102, 0.154)	0.029 (0.023, 0.034)	0.019 (0.015, 0.023)	0.087 (0.079, 0.098)	0.130 (0.118, 0.147)	0.087 (0.079, 0.098)
No Network Information	0.132 (0.103, 0.162)	0.029 (0.023, 0.036)	0.020 (0.015, 0.024)	0.093 (0.083, 0.103)	0.139 (0.124, 0.154)	0.093 (0.083, 0.103)
No Own Arrests	0.126 (0.099, 0.155)	0.028 (0.022, 0.034)	0.019 (0.015, 0.023)	0.091 (0.082, 0.102)	0.137 (0.122, 0.153)	0.091 (0.082, 0.102)
No Own Arrests, No Network Information	0.098 (0.073, 0.124)	0.022 (0.016, 0.028)	0.014 (0.011, 0.018)	0.075 (0.066, 0.084)	0.112 (0.099, 0.126)	0.075 (0.066, 0.084)
No Race	0.114 (0.090, 0.141)	0.025 (0.020, 0.031)	0.017 (0.013, 0.021)	0.093 (0.085, 0.104)	0.140 (0.127, 0.157)	0.093 (0.085, 0.104)
No Demographics	0.108 (0.086, 0.136)	0.024 (0.019, 0.030)	0.016 (0.013, 0.020)	0.081 (0.073, 0.089)	0.122 (0.110, 0.134)	0.081 (0.073, 0.089)
No Own Arrests or Victimizations	0.132 (0.109, 0.173)	0.029 (0.024, 0.038)	0.020 (0.016, 0.026)	0.095 (0.086, 0.106)	0.142 (0.129, 0.158)	0.095 (0.086, 0.106)
Arrests Only	0.106 (0.079, 0.137)	0.024 (0.017, 0.030)	0.016 (0.012, 0.020)	0.085 (0.076, 0.094)	0.127 (0.114, 0.140)	0.085 (0.076, 0.094)
Victimizations Only	0.070 (0.053, 0.089)	0.016 (0.012, 0.020)	0.010 (0.008, 0.013)	0.057 (0.051, 0.068)	0.086 (0.077, 0.102)	0.057 (0.051, 0.068)
Demographics Only	0.098 (0.075, 0.126)	0.022 (0.017, 0.028)	0.014 (0.011, 0.019)	0.069 (0.060, 0.077)	0.103 (0.090, 0.116)	0.069 (0.060, 0.077)
Networks Only	0.116 (0.097, 0.148)	0.026 (0.021, 0.033)	0.017 (0.014, 0.022)	0.087 (0.078, 0.098)	0.130 (0.118, 0.147)	0.087 (0.078, 0.098)
Arrests + Networks Only	0.124 (0.095, 0.155)	0.028 (0.021, 0.034)	0.018 (0.014, 0.023)	0.091 (0.083, 0.101)	0.137 (0.124, 0.152)	0.091 (0.083, 0.101)
No Own Victimizations	0.116 (0.088, 0.144)	0.026 (0.020, 0.032)	0.017 (0.013, 0.021)	0.093 (0.085, 0.105)	0.140 (0.128, 0.158)	0.093 (0.085, 0.105)
No Misdemeanor Arrests	0.144 (0.117, 0.180)	0.046 (0.037, 0.057)	0.021 (0.017, 0.027)	0.086 (0.076, 0.095)	0.184 (0.164, 0.203)	0.086 (0.076, 0.095)
No Felony Arrests	0.132 (0.109, 0.168)	0.035 (0.029, 0.045)	0.020 (0.016, 0.025)	0.094 (0.086, 0.103)	0.171 (0.156, 0.187)	0.094 (0.086, 0.103)

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 $k = 3,381$ people with the highest predicted risk of shooting victimization.

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 Victimizations: Excludes all victimization records for the focal person (but includes them for first- and second-degree peers)
3. No Demographics: Excludes race, gender, age, and location information
4. No Own Arrests or Victimizations: Excludes both arrest and victimization features for the focal person (but includes them for first- and second-degree peers)
5. No Misdemeanor Arrests: Excludes all misdemeanor arrests when constructing features, altering counts for focal person as well as first- and second-degree peers
6. No Felony Arrests: As above but excludes felony arrests and includes misdemeanor arrests

Bottom panel

1. Full: Same as above
2. Arrests Only: Uses only information on arrests, excluding victimization and demographic information
3. Victimizations Only: Uses only information on victimizations, excluding arrests and demographic information
4. Networks Only: Uses only information on first- and second-degree peers, excluding all information on the focal person
5. Demographics Only: Uses only information on demographics, excluding arrests and victimization information

Across models that include some version of arrest information (top panel), it appears that different features are correlated enough that the information lost when one set of features is excluded can be replaced by the information in the remaining sets of features. Although some of the prediction gaps between these models could be substantively important in practice—on the order of 2 percentage points, with felony arrests containing the most information—we cannot distinguish the differences from statistical noise.[42] Even the imprecise differences are concentrated at the very top of the risk distribution; after about $k = 3,000$, precisions across most models are basically on top of each other.

As the bottom panel shows, the biggest loss of information comes from using only victimization records when building the model. Using just demographics or a combination of demographics and victimization records does slightly better than victimizations alone, but not as well as other models. This pattern echoes the finding in the main text that the information contained in arrests records (both of one's self and one's peers) is particularly valuable in estimating the risk of being a shooting victim.

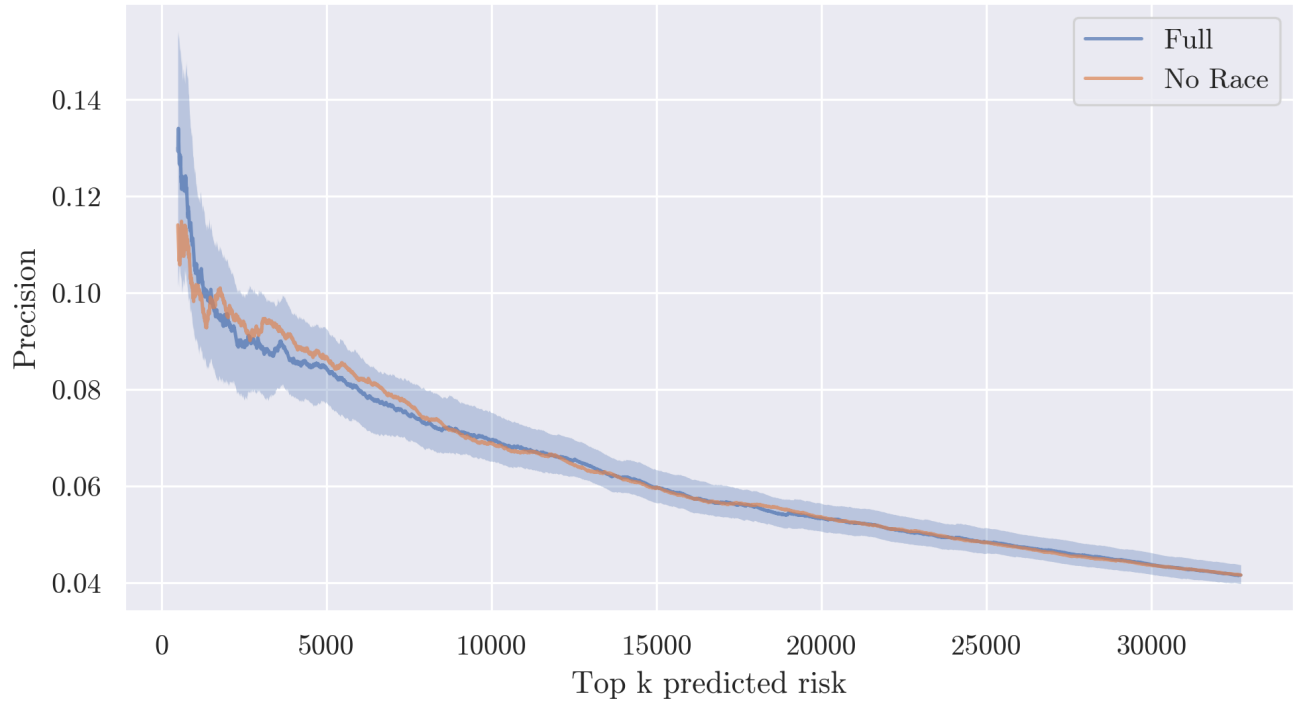
[42] Noise helps to explain why there is a small gap above $k = 1,500$ where the "no own arrest" model appears to outperform the full model.

B.3.2 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).

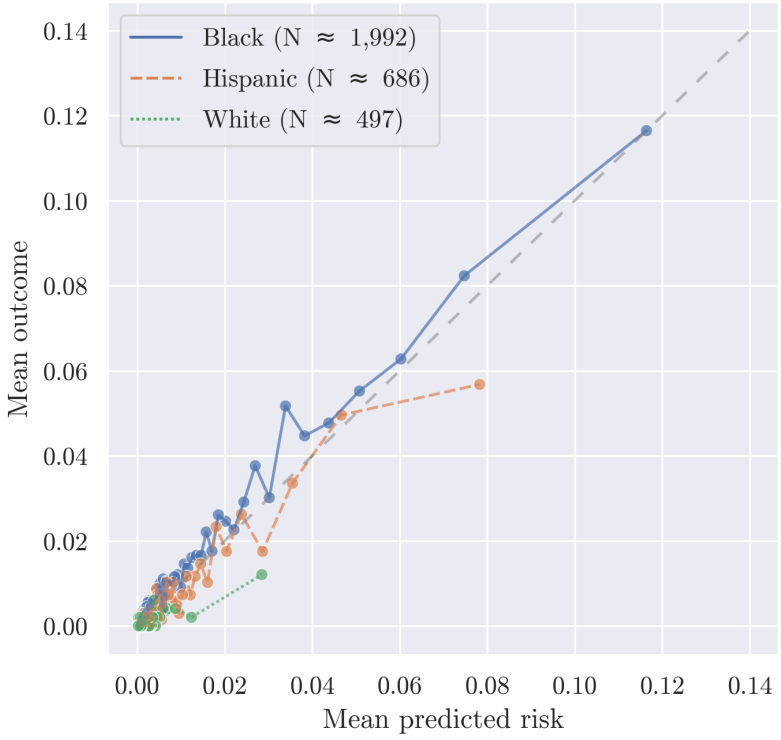
Importantly, as shown in Appendix Figure B.2, the inclusion of race has a relatively trivial effect on predictive performance.[43]

Figure B.2: Precision across models with and without race indicators



[43] We show performance for the top 10 percent of the sample, since showing the full sample makes the scale too small to see the very small differences at the top.

Figure B.3: Calibration for model built without race indicators



There is perhaps some small loss of precision at the extreme top of the risk distribution, but it is not statistically distinguishable from the full model that uses race. This is not unusual in settings where many other features in the prediction are correlated with race (Starr, 2014). Appendix Figure B.3 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.[44]

B.3.3 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, 2021). Thus, for use in these contexts, it is important to understand how much of the predictive accuracy of the

[44] When a somewhat different version of this prediction model was used for social service referrals in practice, we excluded race; see, e.g., https://osf.io/ap8fj/.

full model can be captured by a drastically smaller set of features and simpler modeling techniques.

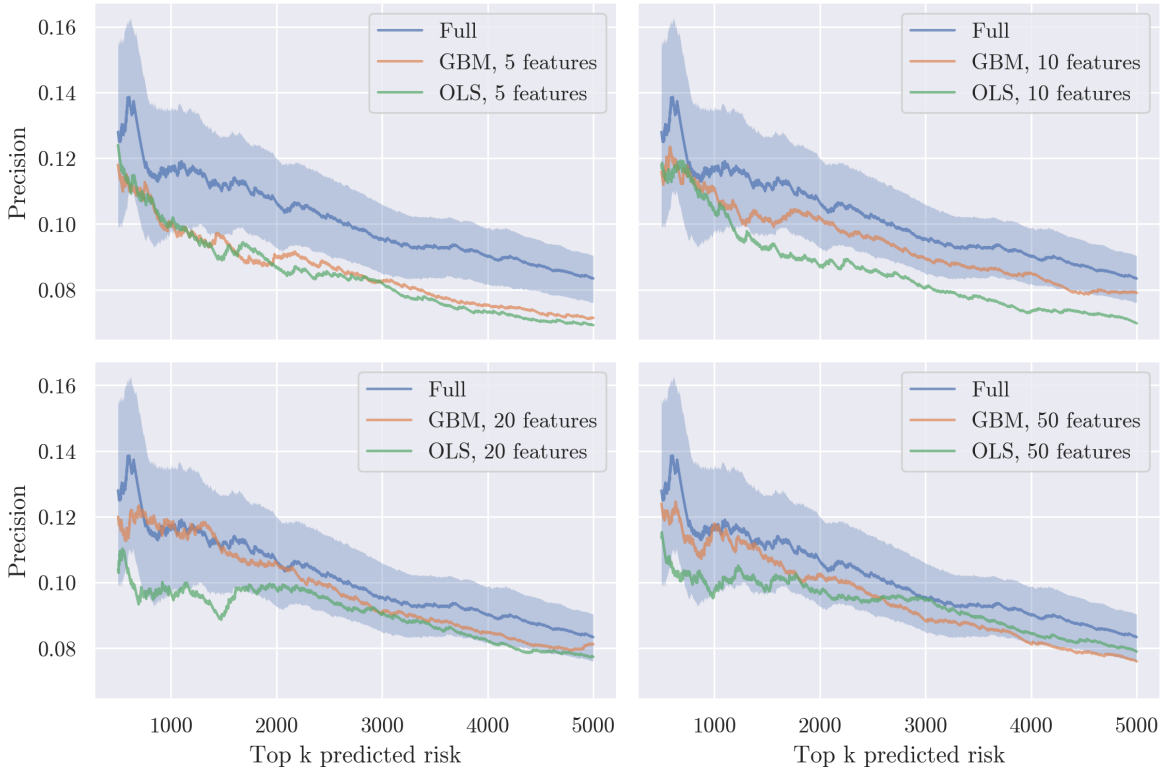
We explore these questions in our setting by first creating a rank-ordered set of 50 features from the full set of all 1,406 features, where the first feature has the highest correlation with shooting victimization, the second has the highest correlation after residualizing out the first feature, and so on. Then we build models using both gradient-boosted decision trees (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 1,406 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. Finally, we 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.

Appendix Table B.6 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 (August 1999); 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.4 reports the same precision plot as Figure 4, with separate panels for different numbers of the top n features reported in Appendix Table B.6. Each panel shows the precision for the full model, and for models using GBM and OLS with only the indicated top n features. As the upper left panel of Appendix Figure B.4 shows, though the differences are noisy, models using only the top $n = 5$ features attain somewhat lower precision: 10 percent of the $k = 1,000$ people with highest predicted risk identified by these very parsimonious models are actually shot in the outcome period, compared to almost 12 percent of those identified by the full model. However, as the number of features available to them rises, the performance of both the GBM and OLS models improve, substantially closing the gap with the full model. This gap closes somewhat faster for the GBM model than for the OLS model. Although the differences in precision are not always statistically significant when accounting for sampling variation, the pattern of results suggests that the additional flexibility of GBM could materially improve performance in situations where the number of available features is more limited. When enough information is available, however, less computationally-intensive prediction models perform as well as more flexible algorithms.

Figure B.4: Precision across models with different model types and number of features



Note: Figure shows precision, or the share actually victimized during the outcome period, of the $k \leq 5,000$ people with the highest predicted risk of shooting victimization, for the full model, a gradient-boosted model with a limited set of features, and an ordinary least squares model with the same limited set of features (Appendix Table B.6). Due to noise in precision at low values of k, we start the graphs at $k = 500$. 95 percent bootstrap confidence interval for full model shown (see Appendix A.5.2 for details).

In addition to Appendix Table B.6, we repeat the stepwise residualization procedure described above for the other feature sets shown in Figure 4. Appendix Tables B.7, B.8, and B.9 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.

Appendix Table B.10 reports the main performance metrics for these different models: the full model and each leave-out-a-feature-set model. To generate these metrics, we reran our modeling process but only gave the algorithm access to the 50 most correlated variables as per our stepwise residualization procedure. As the feature lists show, the top 50 predictors change quite a bit as different feature sets are removed. But both precision and recall at $k = 500$ and $k = 3,381$ are quite similar across all models that use 50 features. This 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. 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.

Appendix Table B.10 and Appendix Figure B.4 highlight that it is generally possible to achieve comparable performance to the full model in the tail. The biggest loss in performance is from removing all arrest information on both focal individuals and their neighbors. 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.[45]

[45] 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.6: Top 50 features from the stepwise residualization procedure when given access to all feature sets

Feature Set	Description	Time Window	Correlations Residualized	Original
networks	# of 1st degree neighbors arrested (any)	365 days	0.120	0.120
arrests	Ever gang-affiliated		0.062	0.095
victims	# of victimizations (shootings)	Total	0.037	0.064
arrests	# of own-arrests (any)	730 days	0.031	0.109
demographics	Age (modal)		0.029	-0.052
arrests	# of own-arrests (gambling)	Total	0.024	0.068
arrests	# of own-arrests (violent)	Total	0.021	0.074
networks	# of 2nd degree neighbors arrested (drug)	30 days	0.021	0.025
arrests	# of own-arrests (reckless conduct)	730 days	0.020	0.067
arrests	# of own-arrests (criminal trespass)	270 days	0.016	0.036
networks	# of 1st degree neighbors arrested (drug)	730 days	0.016	0.102
demographics	Sex (most recent)		0.015	-0.017
demographics	Missing date of birth		0.014	-0.014
arrests	# of own-arrests (solicitation)	730 days	0.014	0.055
arrests	# of own-arrests (drug)	Total	0.014	0.049
victims	# of victimizations (shootings)	270 days	0.013	0.040
networks	# of 2nd degree neighbors victimized (gun assault or battery)	60 days	0.012	0.013
arrests	# of own-arrests (solicitation)	Total	0.012	0.060
networks	# of arrests (any) by 1st degree neighbors	Total	0.012	0.097
arrests	# of own-arrests (firearm possession)	Total	0.012	0.035
demographics	# of unique police beats		0.012	0.027
arrests	# of own-arrests (reckless conduct)	Total	0.012	0.072
networks	# of victimizations (gun battery) by 2nd degree neighbors	60 days	0.012	0.006
networks	# of arrests (drug) by 2nd degree neighbors	730 days	0.012	0.057
networks	# of arrests (drug) by 2nd degree neighbors	30 days	0.011	0.032
arrests	# of own-arrests (robbery)	Total	0.011	0.056
networks	# of victimizations (gun battery) by 1st degree neighbors	Total	0.011	0.071
arrests	# of own-arrests (gun assault or battery)	Total	0.011	0.034
networks	# of 1st degree neighbors arrested (property)	Total	0.010	0.105
networks	# of 1st degree neighbors arrested (drug)	Total	0.010	0.114
networks	# of 2nd degree neighbors victimized (gun battery)	60 days	0.010	0.005
arrests	# of own-arrests (property)	730 days	0.010	0.036
arrests	# of own-arrests (obstructing identification)	Total	0.010	0.033
arrests	# of own-arrests (gang loitering)	Total	0.009	0.043
arrests	# of own-arrests (replica firearm)	Total	0.009	0.030
networks	# of arrests (gun battery) by 2nd degree neighbors	730 days	0.009	0.006
arrests	# of own-arrests (burglary)	Total	0.009	0.036
arrests	# of own-arrests (dealing marijuana public school)	Total	0.008	0.035
networks	# of arrests (gun robberies) by 1st degree neighbors	730 days	0.008	0.018
victims	# of victimizations (any victimization)	Total	0.008	-0.017
arrests	# of own-arrests (drug)	730 days	0.008	0.040
arrests	# of own-arrests (gang name indicator)		0.008	-0.014
networks	# of 1st degree neighbors arrested (gun robberies)	Total	0.008	0.044
arrests	# of own-arrests (aggravated assault school employee)	Total	0.008	0.027
arrests	# of own-arrests (weapons possession)	Total	0.008	0.031
arrests	# of own-arrests (gun robberies)	365 days	0.008	0.010
arrests	# of own-arrests (drug deals)		0.007	-0.016
networks	# of arrests (property) by 2nd degree neighbors	30 days	0.007	0.009
networks	# of 1st degree neighbors victimized (domestic incident)	730 days	0.007	0.006
arrests	# of own-arrests (resist/obstruct officer)	Total	0.007	0.022

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 (August 1999). The fourth and fifth columns show the correlation between the residualized and unresidualized version of the feature and the outcome, respectively.

Table B.7: Top 50 features from the stepwise residualization procedure when not given access to network features

Feature Set	Description	Time Window	Correlations Residualized	Correlations Original
arrests	# of own-arrests (any)	730 days	0.109	0.109
arrests	Ever gang-affiliated		0.060	0.095
victims	# of victimizations (shootings)	Total	0.043	0.064
demographics	Age (modal)		-0.034	-0.052
arrests	# of own-arrests (gambling)	Total	0.032	0.068
arrests	# of own-arrests (reckless conduct)	730 days	0.030	0.067
arrests	# of own-arrests (violent)	Total	0.024	0.074
arrests	# of own-arrests (criminal trespass)	270 days	0.020	0.036
arrests	# of own-arrests (solicitation)	730 days	0.019	0.055
demographics	Sex (most recent)		-0.015	-0.017
demographics	Missing date of birth		-0.015	-0.014
arrests	# of own-arrests (gang loitering)	Total	0.015	0.043
arrests	# of own-arrests (drug)	Total	0.014	0.049
victims	# of victimizations (shootings)	270 days	0.014	0.040
demographics	# of unique police beats		-0.013	0.027
arrests	# of own-arrests (robbery)	Total	0.013	0.056
arrests	# of own-arrests (reckless conduct)	Total	0.013	0.072
arrests	# of own-arrests (public alcohol consumption)	Total	-0.011	0.010
arrests	# of own-arrests (firearm possession)	Total	0.011	0.035
arrests	# of own-arrests (any)	365 days	0.011	0.095
arrests	# of own-arrests (drug deals)		0.011	-0.016
arrests	# of own-arrests (burglary)	Total	0.011	0.036
arrests	# of own-arrests (drug)	730 days	0.010	0.040
arrests	# of own-arrests (replica firearm)	Total	0.010	0.030
arrests	# of own-arrests (property)	365 days	-0.010	0.030
arrests	# of own-arrests (chicago municipal code)	730 days	-0.010	0.041
victims	# of victimizations (gun assault or battery)	Total	0.009	0.046
arrests	# of own-arrests (obstructing identification)	Total	0.009	0.033
arrests	# of own-arrests (fbi code wrt)	270 days	-0.009	0.030
arrests	# of own-arrests (solicitation)	Total	0.009	0.060
arrests	# of own-arrests (gun assault or battery)	Total	0.009	0.034
arrests	# of own-arrests (weapons possession)	Total	0.008	0.031
arrests	# of own-arrests (aggravated assault school employee)	Total	0.008	0.027
arrests	# of own-arrests (drug paraphenelia possession)	730 days	-0.008	0.001
arrests	# of own-arrests (gang name indicator)		-0.008	-0.014
arrests	# of own-arrests (domestic)	365 days	-0.008	0.010
arrests	# of own-arrests (disorderly conduct)	Total	0.008	0.024
arrests	# of own-arrests (heroin possesion)	Total	-0.008	0.008
arrests	# of own-arrests (resist/obstruct officer)	Total	0.007	0.022
arrests	# of own-arrests (aggravated robbery)	Total	0.007	0.028
arrests	# of own-arrests (gun assault or battery)	365 days	0.007	0.016
arrests	# of own-arrests (gun battery)	730 days	-0.007	0.011
arrests	# of own-arrests (manufacturing/dealing public school)	Total	0.007	0.027
arrests	# of own-arrests (possession controlled substance)	365 days	-0.007	0.019
arrests	# of own-arrests (robbery)	Total	-0.007	0.004
arrests	# of own-arrests (public alcohol consumption)	730 days	-0.007	0.013
arrests	# of own-arrests (aggravated battery)	Total	0.007	0.019
arrests	# of own-arrests (stolen property)	365 days	-0.007	-0.002
arrests	# of own-arrests (probation violation)	Total	0.006	0.022
arrests	# of own-arrests (violent)	270 days	-0.006	0.027

Note: See bottom of Table B.6 for column definitions.

Table B.8: Top 50 features from the stepwise residualization procedure when not given access to own-arrest features

Feature Set	Description	Time Window	Correlations Residualized	Correlations Original
networks	# of 1st degree neighbors ever gang-affiliated	365 days	0.120	0.120
victims	# of victimizations (shootings)	Total	0.044	0.064
networks	# of 1st degree neighbors arrested (drug)	Total	0.039	0.114
demographics	Age (modal)		-0.032	-0.052
demographics	Missing date of birth		-0.023	-0.014
networks	# of 2nd degree neighbors arrested (drug)	30 days	-0.020	0.025
victims	# of victimizations (gun battery)	Total	0.017	0.057
networks	# of arrests (violent) by 1st degree neighbors	Total	0.016	0.089
demographics	Sex (most recent)		-0.016	-0.017
networks	# of arrests (drug) by 2nd degree neighbors	30 days	0.015	0.032
victims	# of victimizations (shootings)	270 days	0.013	0.040
networks	# of 1st degree neighbors victimized (domestic incident)	Total	-0.013	0.018
networks	# of 1st degree neighbors arrested (property)	Total	0.012	0.105
networks	# of 1st degree neighbors arrested (drug)	730 days	0.012	0.102
networks	# of 1st degree neighbors arrested (drug deals)	Total	0.012	0.077
networks	# of 2nd degree neighbors victimized (gun assault or battery)	60 days	0.012	0.013
networks	# of victimizations (gun battery) by 2nd degree neighbors	60 days	0.012	0.006
networks	# of 2nd degree neighbors		0.012	0.075
networks	# of arrests (any) by 1st degree neighbors	Total	-0.011	0.097
victims	# of victimizations (days since first victimization)		-0.011	-0.020
networks	# of arrests (any) by 1st degree neighbors	365 days	0.011	0.102
networks	# of arrests (gun battery) by 2nd degree neighbors	730 days	-0.011	0.006
networks	# of arrests (violent) by 2nd degree neighbors	270 days	-0.011	0.028
networks	# of 2nd degree neighbors victimized (gun battery)	60 days	-0.010	0.005
networks	# of 1st degree neighbors arrested (gun robberies)	Total	0.010	0.044
networks	# of arrests (any) by 1st degree neighbors	60 days	-0.010	0.062
networks	# of 2nd degree neighbors ever gang-affiliated	730 days	0.009	0.090
networks	# of 2nd degree neighbors arrested (any)	730 days	-0.009	0.090
networks	# of arrests (gun robberies) by 1st degree neighbors	Total	-0.009	0.038
networks	# of 2nd degree neighbors		-0.009	0.016
networks	# of victimizations (shootings) by 1st degree neighbors		0.008	0.033
networks	# of arrests (drug) by 2nd degree neighbors	730 days	-0.008	0.057
networks	# of 1st degree neighbors arrested (gun assault or battery)	Total	0.007	0.035
networks	# of arrests (gun robberies) by 1st degree neighbors	730 days	-0.007	0.018
networks	# of arrests (property) by 2nd degree neighbors	30 days	-0.007	0.009
demographics	# of unique police beats		0.007	0.027
networks	# of 2nd degree neighbors arrested (drug)	60 days	0.007	0.034
networks	# of arrests (any) by 1st degree neighbors	90 days	0.007	0.074
victims	# of victimizations (any victimization)	730 days	-0.007	-0.014
networks	# of arrests (drug deals) by 2nd degree neighbors	90 days	-0.007	0.014
networks	# of 1st degree neighbors arrested (domestic)	60 days	0.006	0.012
networks	# of arrests (domestic) by 1st degree neighbors	Total	-0.006	0.033
networks	# of arrests (drug deals) by 1st degree neighbors		0.006	0.036
networks	# of arrests (property) by 2nd degree neighbors	Total	-0.006	0.055
networks	# of arrests (domestic) by 1st degree neighbors	365 days	-0.006	0.022
networks	# of arrests (domestic) by 1st degree neighbors	30 days	-0.006	0.004
demographics	Number of unique races recorded		-0.006	-0.006
networks	# of 1st degree neighbors arrested (domestic)	730 days	0.006	0.035
networks	# of 1st degree neighbors victimized (shootings)	365 days	-0.006	0.038
networks	# of 1st degree neighbors arrested (any)	30 days	0.006	0.060

Note: See bottom of Table B.6 for column definitions.

Table B.9: 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	Correlations Original
victims	# of victimizations (shootings)	Total	0.064	0.064
demographics	Age (modal)		-0.048	-0.052
demographics	Missing date of birth		-0.036	-0.014
demographics	# of unique police beats		0.027	0.027
victims	# of victimizations (any victimization)	Total	-0.023	-0.017
victims	# of victimizations (gun battery)	Total	0.019	0.057
demographics	Sex (most recent)		-0.017	-0.017
victims	# of victimizations (shootings)	270 days	0.015	0.040
victims	# of victimizations (days since first domestic incident)		-0.013	-0.007
victims	# of victimizations (property)		-0.011	-0.007
victims	# of victimizations (gun assault or battery)	730 days	0.011	0.044
victims	# of victimizations (any victimization)	730 days	-0.007	-0.014
victims	# of victimizations (gun assault or battery)	90 days	0.007	0.022
victims	# of victimizations (aggravated battery)	Total	0.007	0.035
demographics	Number of unique races recorded		-0.006	-0.006
victims	# of victimizations (simple domestic battery)	Total	-0.006	-0.014
victims	# of victimizations (property)	Total	-0.006	-0.031
victims	# of victimizations (to property)	Total	0.006	-0.017
victims	# of victimizations (gun battery)	60 days	-0.006	0.013
victims	# of victimizations (gun assault or battery)		0.006	-0.039
victims	# of victimizations (days since first victimization)		-0.006	-0.020
victims	# of victimizations (gun battery)	270 days	-0.005	0.037
victims	# of victimizations (child abduction)	Total	0.005	0.007
victims	# of victimizations (days since last victimization)		0.005	0.006
victims	# of victimizations (aggravated battery)	730 days	0.005	0.033
victims	# of victimizations (criminal sexual assault)	Total	-0.004	-0.005
victims	# of victimizations (property)		-0.004	0.002
victims	# of victimizations (days since first shooting victimization)		-0.004	-0.028
victims	# of victimizations (shootings)	365 days	-0.004	0.041
demographics	Modal race indicator		-0.004	-0.003
victims	# of victimizations (aggravated handgun)	Total	0.003	0.046
victims	# of victimizations (gun battery)		0.003	-0.037
victims	# of victimizations (gun assault or battery)		0.003	-0.029
victims	# of victimizations (child endangerment)	Total	0.003	0.010
victims	# of victimizations (gun robberies)		0.003	-0.004
victims	# of victimizations (aggravated domestic battery)	Total	0.003	0.007
victims	# of victimizations (gun robberies)	60 days	0.003	0.004
victims	# of victimizations (attempted strongram)	Total	-0.003	-0.003
victims	# of victimizations (telephone threat)	Total	-0.003	-0.011
victims	# of victimizations (armed knife)	Total	-0.003	-0.003
victims	# of victimizations (gun assault or battery)	Total	-0.003	0.046
demographics	Modal police beat		0.003	0.001
victims	# of victimizations (gun robberies)		-0.003	-0.009
victims	# of victimizations (shootings)		0.003	0.034
victims	# of victimizations (aggravated)	Total	-0.003	-0.003
victims	# of victimizations (gun battery)		-0.003	-0.030
victims	# of victimizations (gun battery)	90 days	-0.003	0.020
victims	# of victimizations (gun battery)	365 days	0.003	0.039
victims	# of victimizations (aggravated (other dangerous weapon))	Total	0.002	0.009
victims	# of victimizations (retail theft)	Total	0.002	-0.002

Note: See bottom of Table B.6 for column definitions.

Table B.10: Predictive performance for limited feature sets chosen by the stepwise residualization procedure

	Top 500			Top 3,381		
Feature Set	Precision	Recall	Total Recall	Precision	Recall	Total Recall
Full	0.130	0.029	0.019	0.087	0.130	0.087
Full - Top 50	0.118	0.026	0.017	0.089	0.133	0.089
No Own Arrests - Top 50	0.120	0.027	0.018	0.088	0.132	0.088
No Networks - Top 50	0.114	0.025	0.017	0.085	0.127	0.085
No Own Arrests or Networks - Top 50	0.066	0.015	0.010	0.060	0.090	0.060

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 $k = 3,381$ people with the highest predicted risk of shooting victimization.