NBER WORKING PAPER SERIES

COMPARING CONVENTIONAL AND MACHINE-LEARNING APPROACHES TO RISK ASSESSMENT IN DOMESTIC ABUSE CASES

Jeffrey Grogger
Sean Gupta
Ria Ivandic
Tom Kirchmaier

Working Paper 28293
http://www.nber.org/papers/w28293

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
December 2020

We thank Terence Chau for excellent research assistance. Special thanks also to Ian Hopkins, Rob Potts, Chris Sykes, as well as Gwyn Dodd, Emily Higham, Peter Langmead-Jones, Duncan Stokes and many others at the Greater Manchester Police for making this project possible. All findings, interpretations, and conclusions herein represent the views of the authors and not those of Greater Manchester Police, its leadership, its members, or the National Bureau of Economic Research. We thank Richard Berk, Ian Wiggett, and three anonymous referees for helpful comments. No financial support was received for this project.

Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases

Jeffrey Grogger, Sean Gupta, Ria Ivandic, and Tom Kirchmaier

NBER Working Paper No. 28293
December 2020
JEL No. K14,K36

ABSTRACT

We compare predictions from a conventional protocol-based approach to risk assessment with those based on a machine-learning approach. We first show that the conventional predictions are less accurate than, and have similar rates of negative prediction error as, a simple Bayes classifier that makes use only of the base failure rate. Machine learning algorithms based on the underlying risk assessment questionnaire do better under the assumption that negative prediction errors are more costly than positive prediction errors. Machine learning models based on two-year criminal histories do even better. Indeed, adding the protocol-based features to the criminal histories adds little to the predictive adequacy of the model. We suggest using the predictions based on criminal histories to prioritize incoming calls for service, and devising a more sensitive instrument to distinguish true from false positives that result from this initial screening.

Jeffrey Grogger
University of Chicago
Harris School of Public Policy
1307 E. 60th Street
Chicago, IL. 60637
and NBER
jgrogger@uchicago.edu

Sean Gupta
Center for Economic Performance
London School of Economics and
Political Science
Houghton Street
London WC2A 2AE
United Kingdom
arijitsean.gupta1@googlemail.com

Ria Ivandic
Houghton Street
London WC2A 2AE
United Kingdom
r.ivandic@lse.ac.uk

Tom Kirchmaier
Houghton Street
London WC2A 2AE
United Kingdom
t.kirchmaier@lse.ac.uk

Comparing Conventional and Machine-Learning Approaches to Risk Assessment in Domestic Abuse Cases *

Jeffrey Grogger[†] *Sean Gupta*[‡] *Ria Ivandic*[§] *Tom Kirchmaier*[¶]

October 1, 2020

Abstract

We compare predictions from a conventional protocol-based approach to risk assessment with those based on a machine-learning approach. We first show that the conventional predictions are less accurate than, and have similar rates of negative prediction error as, a simple Bayes classifier that makes use only of the base failure rate. Machine learning algorithms based on the underlying risk assessment questionnaire do better under the assumption that negative prediction errors are more costly than positive prediction errors. Machine learning models based on two-year criminal histories do even

*We thank Terence Chau for excellent research assistance. Special thanks also to Ian Hopkins, Rob Potts, Chris Sykes, as well as Gwyn Dodd, Emily Higham, Peter Langmead-Jones, Duncan Stokes and many others at the Greater Manchester Police for making this project possible. All findings, interpretations, and conclusions herein represent the views of the authors and not those of Greater Manchester Police, its leadership, or its members. We thank Richard Berk, Ian Wiggett, and three anonymous referees for helpful comments. No financial support was received for this project.

[†]University of Chicago, IZA, and NBER

[‡]London School of Economics and Political Science

[§]London School of Economics and Political Science

[¶]London School of Economics and Copenhagen Business School

better. Indeed, adding the protocol-based features to the criminal histories adds little to the predictive adequacy of the model. We suggest using the predictions based on criminal histories to prioritize incoming calls for service, and devising a more sensitive instrument to distinguish true from false positives that result from this initial screening.

I. Introduction

Domestic abuse is a global problem. Worldwide, nearly one-third of women in a relationship report having experienced physical or sexual violence at the hands of an intimate partner (Buzawa and Buzawa (2017)). In the United States, roughly 1 in 4 women experience serious intimate partner violence over their lifetimes (National Coalition Against Domestic Violence (2014)). In England, one-third of all assaults involving injury stem from domestic abuse-related crimes (Her Majesty's Inspectorate of Constabulary (2014)).

Since the mid-1990s, an important part of the response to domestic abuse, as to other aspects of policing, has been risk assessment (Dutton and Kropp (2000); Campbell, Webster and Glass (2009); Ericson (1997)). Although details vary across assessment protocols, many share basic features. They typically involve a questionnaire, which can be administered to the victim or the perpetrator, or be answered on the basis of an official records search (Messing and Thaller (2013)). Some involve a scoring rule, whereas others take a structured-judgment approach, asking the person administering the questionnaire to use the responses, together with their professional judgement, to classify the risk presented by the case (Kropp (2004)). Risk assessments can be carried out by police, probation offices, or victim support organizations for the purposes of criminal-justice decision making, victim safety planning, and the prevention of future violence (Kropp (2004); Her Majesty's In-

spectorate of Constabulary (2014); Robinson et al. (2016)). They are conducted routinely in several European countries and in many parts of Canada and the US (Kropp (2004); Berk, He and Sorenson (2005); Roehl et al. (2005); Turner, Medina and Brown (2019)).

In England and Wales, most police forces carry out risk assessment using the DASH (Domestic Abuse, Stalking, and Harassment and Honor-Based Violence) risk identification and assessment model (Robinson et al. (2016)). DASH was developed by police, victim advocates, and academics with the goal of identifying victims at high risk and facilitating investigations, planning, and interventions so as to reduce the likelihood of violence (Robinson (2011)). The items included on the questionnaire were based on prior research on violent recidivism and included a number of items identified as particularly predictive of serious violence and murder (Richards (2009)).

DASH shares features with risk assessment protocols in use elsewhere. Its first component is a questionnaire, consisting of 27 items, administered to the victim by a police officer responding to a domestic abuse call. Its second component is the officer's risk grade, which amounts to a prediction of future risk. The officer grades the case as standard-, medium-, or high-risk, where high risk implies that "[t]here are identifiable indicators of risk of serious harm. The potential event could happen at any time and the impact would be serious" (Richards (2009)). Victims in high-risk cases are offered services designed to keep them safe and provide them with time to consider their options (Her Majesty's Inspectorate of Constabulary (2014)). The risk grade involves structured judgement: the officer's risk classification may be informed by the victim's responses to the questionnaire, but the officer is instructed to use his/her professional judgement in making it (Robinson et al. (2016)).

Such risk assessment protocols have been criticized on a number of grounds.

These include low predictive power, which has been observed among predictive methods used in a variety of criminal justice contexts (Farrington and Tarling (1985); Campbell, Webster and Mahoney (2005)). Similarly, methods based on informal scoring or vague decision rules have been criticized for inconsistency (Grove and Meehl (1996); Gottfredson and Moriarty (2006)).

Inconsistency is clearly a problem for DASH, since variation in its use has resulted in striking differences across police forces in the share of domestic abuse calls graded as high-risk. Her Majesty's Inspectorate of Constabulary (2014) reports that 10 police forces, out of 28 for which data were available, classified fewer than 10 percent of domestic abuse cases as high-risk. At the other end of the spectrum, three forces designated over 80 percent as high-risk. Such wide variation casts doubt on the reliability of the predictions. It is consistent with findings of unreliable professional judgements in a wide variety of settings (Gottfredson and Moriarty (2006); Kahneman et al. (2016)).

In this paper we develop a method for predicting violent recidivism in domestic abuse cases that outperforms DASH. Our work makes several contributions. First, we analyze predictions based on the DASH risk grade. We show that these predictions are only trivially more accurate than those based on the simple Bayes classifier, which makes no use of DASH whatsoever. We then apply machine learning methods to the DASH data. We show that the resulting predictions do no worse than those based on the risk grade, and under plausible conditions regarding the relative costs of different types of forecasting errors, can do considerably better.

We next apply the same machine learning methods to predictors derived from the victim's and perpetrator's criminal histories. We show that these predictions are better than those based on the DASH data. Finally, we show that adding the DASH data to the criminal histories generates predictions that improve only slightly

on those based on the criminal histories alone.

The prior literature includes several examples of machine learning methods being applied to forecast recidivism in domestic abuse cases. Berk, He and Sorenson (2005) build models to predict recidivism on the basis of limited information available at the scene. Berk, Sorenson and Barnes (2016) consider a pre-trial detention setting, where more information may be available. Turner, Medina and Brown (2019) train models to predict violent recidivism based on DASH. Relative to these prior studies, our contributions are: i) to compare machine learning models based on a conventional risk-assessment protocol to those based on criminal histories, and ii) to show that adding information from the protocol to the models based on criminal histories leads to little improvement. In the next section of the paper we discuss our data. In the following section we discuss our approach and present results. In the final section, we discuss other risk assessment protocols, limitations of both the DASH protocol and our machine-learning procedure, and how our approach could be put into day-to-day practice.

II. Data

A. Calls for domestic abuse, violent recidivism, and the analysis sample

Our data on calls for service for domestic abuse (DA) are drawn from the command and control database of Greater Manchester Police (GMP), which includes a record for each call for service. Details about calls for service that result in crime reports are held in GMP's crime database. Not all DA calls involve criminal offenses, but all DA calls are retained by the system, whether the underlying incidents ultimately prove to be crimes or not. DASH data come from a separate file. We merge

these files together, retaining records that pertain to DA calls.

In England and Wales, domestic abuse is broadly defined to include incidents between individuals age 16 years or older who are or have been intimate partners or family members (Crown Prosecution Service (2020)). This can include incidents between siblings, incidents between adult children and parents, or intimate-partner incidents involving current or past spouses or romantic partners. Since the great majority of domestic abuse incidents involve heterosexual intimate partners, we restrict attention to those calls[1].

The objective of our analysis is to forecast domestic abuse recidivism involving violence with injury or a sex offense that is reported to police. Hereafter, we refer to this as violent recidivism so as to economize on language. We define violent recidivism at the level of the dyad, that is, the victim-perpetrator pair. We code violent recidivism as any reported DA incident involving violence with injury or a sex offense that occurs within one year of a preceding call from the same dyad, ignoring multiple calls on the same day[2]. Of course, many DA incidents are not reported to police; we are unable to forecast those incidents[3].

In principle, recidivism could be defined differently, for example based on arrests rather than calls for service. Other analysts, such as Berk, Sorenson and Barnes (2016) and Turner, Medina and Brown (2019), have taken this approach. We selected our measure of violent recidivism with the goal of capturing a reasonable notion of serious harm, given the offense categories recorded in the GMP data. Violence with injury consists primarily of two classes of offenses: Wounding with Intent to do Grievous Bodily Harm (GBH) and Assault Occasioning Actual Bodily Harm

[1]Intimate partners include ex-partners, partners, wives, girlfriends, ex-wives, husbands, boyfriends, ex-husbands, and civil partners. Turner, Medina and Brown (2019) show that DASH predicts repeat intimate partner violence better than non-intimate partner domestic violence.

[2]Dyads are defined to involve the same two people, but not necessarily in the same roles as victim and perpetrator.

[3]Since our data come from GMP's command and control database, we may also miss any incidents that take place in another police jurisdiction.

(ABH). GBH includes injuries resulting in permanent disability; permanent, visible disfigurement; compound fractures; substantial loss of blood; lengthy treatment, or psychiatric injury. GBH clearly accords with any notion of serious harm; the maximum sentence for GBH is life imprisonment. ABH also includes injuries involving considerable harm, such as broken noses, broken digits, loss of teeth, and shock, as well as lesser injuries such as grazes, swelling, and black eyes. It carries a maximum sentence of five years. It is important to note that violence with injury does not include Common Assault, which may entail slaps, punches, or other attacks that leave no visible mark or injury, and which carries a maximum sentence of six months (Home Office (2020)).

As a measure of serious harm, ABH may seem too inclusive, whereas GBH seems too restrictive. A final statistical consideration led us to focus on violence with injury: GBH accounts for only about 1 percent of the domestic abuse cases in our sample, making it essentially unforecastable. Below we see that our violent recidivism measure, which includes violence with injury plus sex offenses, is already rare enough (in a statistical sense) to pose considerable challenges[4].

The variables that we use in our analysis, plus the time periods over which the various components of our data are available, define our sample period. The call-for-service and crime data are available from April 2008 to July 2019. These are used to construct our outcome measure as well as the predictors that are based on two-year histories of DA calls and criminal records. The DASH data are available from July 2013 to July 2019. However, we only include those dyads whose first call for service took place after April 2014, since information linking victims and perpetrators is only available beginning in April 2012. To ensure that we have a full year to measure violent recidivism for all DA calls, we only include calls that

[4] Turner, Medina and Brown (2019) seemingly worked with an earlier version of the Home Office classification system, which allowed them to distinguish finer categories of offenses.

occurred before July 2018. Figure 1 illustrates this timeline and the Data Appendix provides further details about the sample.

Table 1 displays the frequency distribution of calls by the sequence number of the call. It also shows the probability of at least one repeat call and the probability of violent recidivism by call number. Here as throughout the analysis, the unit of observation is a call for service, which we also refer to as an incident. The first row shows that 72,972 calls, or 44.2 percent of the sample, are first calls. Of those, 42.5 percent are followed by at least one more call. The third column shows that the probability of a repeat call rises with the number of calls. Bland and Ariel (2015) report a lower rate of repeat calls in rural Suffolk, but similarly report an increase in the likelihood of a repeat call as the number of calls rises.

The third column of the table shows that the likelihood of violent recidivism also rises with the number of calls. Although only 6.7 percent of first calls result in violent recidivism, that share rises to 10.7 at the second call, and continues to increase thereafter. Nevertheless, because the sample disproportionately consists of first calls, the overall rate of violent recidivism is 11.8 percent. Viewed from a strictly statistical perspective, this means that violent recidivism is a relatively rare event. This will have important implications for our analysis below.

B. Responses from DASH questionnaires

As mentioned above, responding officers are instructed to complete a DASH form for each DA call. Answers come from the victim[5]. The standard DASH questionnaire contains 27 questions; GMP adds a 28th question, asking whether the officer gathered any other relevant information about the case. The officer then

[5]The primary victim provides the answers to the DASH questionnaire. Standard procedure is to separate the parties before administering the protocol. In a small number of cases there are multiple victims, who are most often underage children. Our original datasets contain only data on the primary victim as recorded by the police. We do not have data on secondary victims.

grades the case as standard, medium, or high risk. For dyads with multiple calls, both the responses to the questions and the officer's risk grade may vary from call to call, reflecting the dynamics of the situation.

Table 2 displays short versions of the questions and tabulates their responses. For each question, there were three possible responses: yes, no, and omitted. We treat each question as a three-level factor, with the omitted category as the third level. Approximately 10 percent of the DASH questionnaires consisted entirely of omitted responses. We include these cases in the analysis below, although dropping them from the analysis sample had little effect on the results.

Panel (a) of Table 2 shows wide variation in the responses to the DASH questions. Abusers are unlikely to be reported to threaten or harm children, for example, whereas nearly half are reported to be in trouble with the police. On the average DASH report, 4.9 features were checked yes. On a question-by-question basis, the share of omitted responses runs from about 12 to 20 percent. The average number of omissions was 5.2.

Panel (b) reports the distribution of officer risk grades. The vast majority of cases are assessed as standard- or medium-risk. However, 9.0 percent are assessed as high-risk, meaning that the officer believes that the victim could be seriously harmed at any time. We discuss the accuracy of these assessments in the next section.

C. Criminal history variables

Table 3 lists the criminal history variables that we use to predict violent domestic recidivism. One set pertains to the dyad, another to the perpetrator, and another to the victim. All history variables extend back two years from the date of the focal DA call.

Looking first at DA calls, we see that the perpetrator is the male in 83.4 percent of incidents. The typical dyad averages 2.42 calls over the past two years. Roughly 0.79 of those incidents were classified as crimes, on average, and 0.23 of them involved violence. Over the prior two years, the average perpetrator had been involved in 0.864 crimes, compared to only 0.262 crimes for the average victim[6]. The perpetrator was also roughly three times more likely than the victim to have been involved in violence, either with or without injury.

III. Prediction models and results

A. Predictions based on the DASH risk assessment

Although our focus is on forecasts derived from the machine learning methods, we begin by discussing predictions based on the DASH risk grade. In the DASH protocol, an officer's grade of high risk is a prediction of serious harm. We dichotomize the three-level risk assessment, combining standard and medium risk into a single category, which we denote as "lesser" risk. In Table 4 we present a cross-tabulation of this dichotomous assessment and our outcome variable, which equals one if the dyad was involved in a DA incident involving violent recidivism within a year of the focal call. For comparability with the results to follow, we tabulate results based on the test sample.

We present results in the form of a confusion matrix, as we do with the machine learning models to follow. The first two columns classify DA calls according to the risk grade, that is, the predicted outcome. The first two rows classify them according to the actual outcome, that is, whether or not they resulted in violent recidivism. The third column presents the row shares, also known as base rates,

[6]None of the victim or perpetrator offenses need involve the other party to the dyad.

showing that 11.8 percent of the calls in the test sample resulted in failure, that is, violent recidivism. The third row of the table presents the column shares, showing that 8.5 percent of the calls in the test sample were graded as high-risk, that is, predicted to fail.

The final column of the table presents rates of classification error. Classification error conditions on the actual outcome of the incident. The first entry shows that 8.2 percent of the successes, which we define as cases that did not result in violent recidivism within a year, were classified as failures[7]. Equivalently, 91.8 percent were classified as successes. The second entry shows that 88.8 percent of the failures were classified as successes.

The final row of the table presents rates of prediction error. Prediction error conditions on the predicted, rather than actual, outcome of the incident. The first entry shows that 11.5 percent of the incidents that were predicted to succeed actually failed. The second entry shows that 84.4 percent of the cases that were predicted to fail actually succeeded. The bottom right-most element in the table is the overall error rate. It is a weighted sum of the classification errors, where the weights are given by the row shares. The overall error rate is 17.7 percent. Put equivalently, the accuracy rate of the forecasts based on the DASH risk assessment, equal to one minus the overall error rate, is 82.3 percent.

The bottom element on the left is the area under the receiver operating characteristic (ROC) curve, or AUC. This is a common measure of classifier performance, which equals the probability that a randomly selected failure will have a greater predicted likelihood of failure than a randomly selected success. A value of one indicates perfect prediction, whereas a value of 0.5 is equivalent to random guessing (Fawcett (2006)). A value of 0.515 indicates that the DASH risk grade performs

[7]No value judgement is implied here. We are simply using the term success to refer to the complement of failure, as is common in the statistical literature.

little better than random guessing[8].

Table 4 also makes several other important points. The first is that that it is difficult to predict a rare outcome in a manner that beats the simple Bayes classifier. The simple Bayes classifier just predicts the majority class for each case. Since the base failure rate is 11.8 percent, this would amount to predicting that no incident would result in violent recidivism. This forecast would have a 11.8 percent prediction error. Equivalently, it would be 88.2 percent accurate, beating the DASH risk grade. At the same time, the simple Bayes approach would be unacceptable, since every failure would amount to a false negative[9].

This argument leads to another point, which is that conventional accuracy may not be the correct objective for the prediction exercise. There are two possible types of prediction errors, negative prediction errors and positive prediction errors. Negative prediction errors occur when the forecasting method predicts no recidivism, but recidivism in fact occurs. Positive prediction errors occur when a case that is predicted to recidivate does not do so.

In this setting, the costs of negative and positive prediction errors are likely to be asymmetric, a point that has been made forcefully by Berk and his co-authors (Berk, He and Sorenson (2005), Berk (2008), Berk (2012), Berk and Bleich (2013)). The cost of a positive prediction error may involve protective measures that were undertaken unnecessarily. The cost of a negative prediction error may be a violent attack on a victim who was provided with no such protection. Negative prediction errors are thus more costly, that is, more dangerous.

This leads to our third point: when negative prediction errors are more costly than positive prediction errors, reducing the rate of negative prediction errors should

[8] Turner, Medina and Brown (2019) report a similar finding. To calculate the AUC in Table 4, we used a univariate logistic regression of violent recidivism on the dichotomized DASH risk grade to compute the ROC curve, then integrated the area beneath it.

[9] Rare outcomes are common in DA prediction settings. See Berk and Sorenson (2020) for an alternative to the approach we take below.

take precedence over increasing accuracy. The negative prediction error rate is the number of false negatives divided by the number of cases predicted to succeed. Since the number of false negatives in the test sample is 1702, and the total number of cases predicted not to fail is 14,823 (=1702 + 13,121), the negative prediction error rate of the forecast based on the DASH assessment is 11.5 percent. This is only trivially better than 11.8 percent negative prediction error rate from the simple Bayes classifier. Put differently, despite all the effort devoted to the DASH system, it generates a less accurate forecast, with a negative prediction error rate that is only slightly lower, than a forecast which makes no use of the DASH information whatsoever. The approach we take below lets us reduce negative prediction error substantially.

B. Machine learning methods

Our main analyses are based on machine learning algorithms trained to various sets of predictors. We consider two algorithms, logistic regression and random forest. In each case, we train one set of models to the responses from the DASH protocol, one to the criminal history variables, and one to both sets of features combined. In each case, the goal is to predict violent recidivism.

Logistic regression is a parametric procedure in which the log-odds of failure is modelled as a linear function of the predictors. An advantage of logistic regression is that it is relatively easy to interpret. The procedure produces a coefficient for each predictor, which can be interpreted as the partial effect of the predictor, that is, as the mean change in the log-odds of failure due to a unit change in the predictor, holding the other predictors constant. The coefficient can be used to assess the direction and magnitude of the effect of the predictor on the outcome and serve as a basis for inference. A disadvantage of logistic regression is the assumption of

linearity between the predictors and the log-odds of failure, since there is no reason to expect linearity to hold in practice.

The random forest is a non-parametric procedure. It is constructed from a predetermined number of classification trees, where each tree is grown from a different random sample of the data, drawn with replacement, and a different random subsample of predictors. Each tree classifies each observation as a predicted failure or success on the basis of its predictors[10]. Once all the trees have been grown, each observation is classified as a failure or success based on a majority vote across the trees[11].

An advantage of random forests is that they impose no assumptions on the functional relationship between predictors and the log-odds of failure. If the relationship is linear, the random forest will learn that, but the splitting algorithm used to grow the trees can uncover non-linearities as well as interactions among the predictors (Berk (2008); Hastie, Tibshirani and Friedman (2009)). A disadvantage is that random forests are more difficult to interpret than logistic regression, since they produce no coefficients.

Instead, insights into the workings of the model can be gained from predictor importance measures and partial dependence plots. The relative importance of a predictor can be assessed by randomly shuffling the values of the predictor, assessing the classification accuracy of the predicted outcomes based on the shuffled predictor, and comparing it to the classification accuracy based on the actual predictor. Likewise, the direction and magnitude of the effect of a predictor can be assessed by means of a partial dependence plot, which is essentially a simulation of how the outcome changes as the predictor changes, holding constant the other predictors (Hastie, Tibshirani and Friedman (2009), p. 369).

[10] We use the randomForest package in R to train our models.

[11] Excellent detailed discussions of random forests can be found in Breiman (2001), Hastie, Tibshirani and Friedman (2009), and Berk (2012).

For estimation and prediction, we follow standard practice of splitting our full data set into a training sample and a test sample. Since some dyads are involved in multiple incidents, those involving the same dyad may be statistically dependent. To ensure that the training and test samples are independent, we assign all incidents involving the same dyad to either the training sample or the test sample. The training sample consists of a randomly selected 90 percent of the dyads, which are used to train the prediction models. The remaining 10 percent test sample is used to judge the adequacy of the models.

By default, both logistic regression and random forests assume symmetric costs of positive and negative prediction errors, and will tend to maximize classification accuracy as a result. In order to account for asymmetric costs, as we advocated above, we target different cost ratios by downsampling the majority class, that is, the successes. Given the differences between logistic regression and random forests, downsampling is carried out slightly differently for the two algorithms. For logistic regression, we sample with certainty all the failures in the training sample. We then sample that number of successes which most nearly achieves the desired cost ratio. For the random forests, downsampling involves stratified sampling with replacement at the construction of each tree. From the failures, we draw a sample with replacement of size equal to the number of failures in the training set. From the successes, we draw a sample with replacement of that size which most nearly achieves the desired cost ratio. The desired cost ratio generally cannot be achieved exactly, due to the inherent randomness of the procedure. However, we typically come close, as will be seen below.

C. *Logistic regressions based on DASH features*

Before discussing the predictions from our models, we present estimation results that may help the reader understand the relationship between the DASH features and violent recidivism. Column (1) of Table 5 presents coefficients from a logistic regression model estimated from the training sample[12]. In each case, the dependent variable is the binary indicator for violent recidivism. As predictors, the model in column (1) includes a set of two dummy variables for the responses to each of the 28 DASH questions, plus a dummy variable equal to one if the incident was graded as high-risk by the responding officer. For each of the DASH questions, one of the two dummies is coded as one if the response was yes and zero otherwise, and the other dummy is coded as one if the response was omitted and zero otherwise. The omitted category in each case is a negative response.

Twenty-three of the 28 coefficients corresponding to affirmative answers are statistically significant at conventional levels. The high-risk coefficient is also significant, as are several of the coefficients on the omitted-response dummies. The magnitudes of the coefficients can be compared to each other, since each can be interpreted as the mean change in the log-odds of failure from changing the corresponding predictor from zero to one, holding constant all other predictors. In general, the magnitudes vary quite a bit across coefficients, meaning that the features have different partial effects on violent recidivism. Two of the strongest positive factors are affirmative indicators that the abuser has alcohol problems and that the abuser has previously been in trouble with police. Other important positive predictors include whether the current incident resulted in injury, whether the victim feels isolated, and whether the abuse is happening more often. Omitted responses to whether the abuser ever strangled or choked the victim, or hurt anyone else, are

[12] Standard errors, presented in parentheses below the coefficients, have been clustered at the level of the dyad to account for possible dependence among incidents involving the same participants.

also highly predictive.

At the same time, we see that several of the coefficients are negative and significant, including indicators of whether the abuser ever hurt children, whether the victim was very frightened, and whether the abuser stalked the victim. To understand these negative coefficients, we estimated a series of logistic regressions in which we included the responses to the individual DASH questions one-at-a-time. In results not reported here, we found that affirmative responses to each of the questions are positively correlated with violent recidivism.[13] Therefore, what the negative coefficients show is that the partial effects of several of the DASH predictors are negative, once one accounts for the other predictors in the questionnaire. To the extent that responding officers weigh all the questions in forming their risk assessment, understanding and keeping track of which predictors have negative partial effects must greatly complicate their prediction problem.[14]

D. Random forests based on DASH features

We estimated a random forest from the same training data. Again we use the 28 DASH questions and the dichotomized DASH risk grade to predict violent recidivism. Figure 2 plots shuffle importance measures for the predictors. Measured by loss of fit when the variable's influence on the model is effectively eliminated by shuffling, the most important predictors are indicators of whether the abuse is happening more often, the victim feels isolated, the abuser hurt anyone else, or the abuser ever strangled or choked the victim. These variables were also identified as important by the logistic regression.[15]

[13] Turner, Medina and Brown (2019) report a similar finding.

[14] We also analyzed the responses for collinearity by estimating the tolerance of each feature with respect to the others. For the most part, tolerances ranged from 0.80 to 1.0, suggesting that collinearity was not the main explanation for the negative coefficients.

[15] To judge the stability of the importance measures, we trained 100 random forests, changing the randomization sequence each time. We calculated the share of times each predictor received an importance ranking within one rank of that shown in Figure 2. For the top four features, the share was 100 percent.

Figure 3 presents partial dependence plots for those four predictors. Because the DASH features are categorical variables, the partial dependence plots are bar charts. They show how the log-odds of failure differ, on average, between yes, no, and omitted responses. The figures show that victims who indicated that the abuse was happening more often, or for whom this response was omitted, were surprisingly less likely to experience violent recidivism than those whose response to this question was negative. The opposite pattern, which is more consistent with the logistic regression coefficients in Table 5, holds for whether the victim felt isolated. Omitted responses to whether the abuser had hurt anyone else, or ever choked or strangled the victim, were more predictive of violent recidivism, all else equal, than either yes or no responses.

E. Predictions of violent recidivism based on DASH features

We now consider the predictions of violent recidivism that are generated by these models, which are summarized in the confusion matrices that appear in Table 6. These matrices apply the logistic regression and random forest from the training sample to data from the independent test sample. They cross-tabulate the actual class (the row variable, as in Table 4) by the predicted class (the column variable). The confusion matrices for the logistic regression appear on the left, and those for the random forest appears on the right.

The predictions summarized in Panel (a) are based on the assumption that negative and positive prediction rates have symmetric costs. This implicit assumption is the default for both algorithms, and results in models that maximize accuracy. When failure is a statistically rare event, as in our case, accuracy is maximized largely by minimizing classification error within the majority class (i.e., the suc-

For the remainder of the top 10 features, it exceeded 70 percent.

cesses). This is evident in the confusion matrices in Panel (a). The logistic regression literally returns the Bayes classifier. The random forest does little better. For both models, the rate of negative prediction error is 11.8 percent.

To allow for asymmetric costs, we retrained the algorithms using the down-sampling approach described above. Since the true cost of negative and positive prediction errors is unknown, we selected three different relative cost ratios: 5:1, 10:1, and 15:1. We present confusion matrices based on the 10:1 cost ratio in Panel (b) of Table 6; those for the other cost ratios appear in Appendix Table A.1.[16] In presenting these results, we do not mean to take a stand on what the correct cost ratio is; that would require one to know the relative costs of possible outcomes in each case[17]. Our point here is to illustrate the importance of allowing for asymmetric costs.

Raising the relative cost of negative prediction errors reduces the classification error among the failures, as one would expect. At the same time, it raises classification error among the successes, raising the overall error rate. When positive prediction errors are relatively less costly than negative prediction errors, more cases are predicted to fail, raising the number of false positives and the overall error rate.

More importantly, the negative prediction error rate falls. At the 10:1 cost ratio, the rate of negative prediction error from the logistic regression is 7.7 percent. The random forest does a bit worse, yielding a negative prediction error rate of 7.9 percent.

The absolute number of false negative cases falls from 1702 based on the DASH risk grade to 634 based on the random forest with a 10:1 cost ratio. If we project these numbers onto our full sample, the reduction in false negatives comes to roughly

[16] As noted above, it is impossible to achieve the targeted cost ratio exactly due to the randomness in the procedure. However, models in Panel (b) perform fairly well achieving cost ratios of 10.52 (logistic regression) and 10.86 (random forest).

[17] Fortunately, only relative costs matter. It is not necessary to specify the absolute costs of either false negative or false positive errors. See Berk (2012).

10,680 ($= 10 \times (1702 - 634)$). Annualizing over our sample period of four years and four months, this amounts to a reduction of 2,465 false negative cases per year. If negative prediction errors are more costly than positive prediction errors, as seems reasonable in this case, this is a substantial improvement.

In part, this reduction is due to the fact that the machine learning algorithms make many fewer predictions of success, and many more predictions of failure, than do the police based on the DASH risk grade. Although this is precisely what one would want in a setting where negative prediction errors are more costly, it is nonetheless interesting to ask how the algorithms would do if they generated roughly the same share of predicted failures as the police. To address this question, we used the downsampling technique described above to target not the relative cost of prediction errors, but rather the share of predicted failures, targeting an 8.5 percent predicted failure rate to accord with the results of the DASH risk grade in Table 4. Results are shown in 6.[18] Both the logistic regression and the random forest yield a negative prediction error of about 11 percent, slightly lower than the 11.5 percent achieved by the DASH risk assessment. Thus the advantage of the machine learning algorithms over the DASH risk assessment stems from two factors. First, they achieve a lower rate of negative prediction error, holding constant the share of failures that are predicted. Second, they allow one to adjust the share of predicted failures according to the relative cost of negative vs. positive prediction errors.

Finally, with one exception, the AUC statistics in Table 6 are all notably higher than those in Table 4. This further indicates that the predictions based on the machine learning models outperform the predictions based on the DASH risk grade.[19]

[18] Again, the randomness of the procedure makes it impossible to hit the target exactly.

[19] The exception involves the random forest trained at the default cost ratio. For this model, half of the observations were predicted to be successes by unanimous vote. As a result, the ROC curve coincides with the 45-degree line for much of its upper reach.

However, the important point from Table 6 is that using the DASH items to train a machine learning algorithm can generate many fewer negative prediction errors than the predictions based on the DASH risk grade.

F. Logistic regressions based on criminal history features

We now turn to the results of models trained to criminal histories rather than the DASH features. As above, we begin with estimation results in order to help the reader get a sense of the workings of the model.

Table 7 presents estimated coefficients from a logistic regression of the violent recidivism indicator on the criminal history features. Many of the variables are significant at conventional levels. The first coefficient indicates that incidents involving a male perpetrator are less likely to result in violent recidivism than the much less common incidents involving a female perpetrator.

Since each criminal history variable represents the number of events in the past two years, their coefficients can be interpreted as the partial effect of a one-unit change in the corresponding variable on the log odds of failure, holding the other variables constant. Thus we can infer that, among the variables that positively predict violent recidivism, the most important include the number of prior incidents involving DA violence within the dyad, and the number of incidents in which the victim was implicated that involved violence, both with and without injury.

As with the models based on the DASH features above, several of the coefficients are negative. However, the interpretation of these coefficients is different from that of the negative coefficients from the DASH model. The reason is that many of the variables are highly correlated. For example, if the number of DA violence incidents within a dyad rises by one, then the number of DA violence incidents involving the perpetrator rises by one and the number of DA violence incidents involving the

victim rises by one. The reason the three terms are not perfectly collinear is that the perpetrator and victim may be involved in incidents of DA violence involving someone other than the other member of the dyad. When predictors are highly correlated, it can be difficult to distinguish their partial effects, even when they remain jointly predictive. [20]

G. Random forests based on criminal history features

We also trained a random forest on the same data. Figure 4 presents a shuffle importance plot. In the random forest, DA-related calls prove more important than the DA-related crimes and violence that were important in the logistic regressions.[21] The reason for this difference could involve interactions. Shuffle importance measures capture the influence of a feature in all its facets, either as a main effect or (possibly high-order) interaction. If DA-related calls interact with other features, this would be reflected in the shuffle importance plots, but such interactions would not be captured by the logistic regression due to the linearity assumption.

Figure 5 presents partial dependence plots for four important predictors, including DA calls by the victim, the dyad, and the perpetrator, as well as DA crimes by the dyad. These plots show how the log-odds of failure change, on average, as the predictor increases. Since the predictors are quantitative, they appear as line graphs rather than the bar charts above. The graphs for each of the three types of DA calls are fairly similar. Initially, an increase in the number of calls leads to a sharp increase in the log-odds of failure, after which the effect flattens out. This effect is highly non-linear, although it is important to keep in mind that the

[20] Collinearity in a logistic regression can make the coefficients difficult to interpret, but it does not necessarily affect the predictive power of the model. Collinearity is less of an issue for random forests, since each tree is built from a randomly chosen subset of predictors, and correlated predictors will often appear in different subsets rather than together.

[21] These importance rankings were quite stable, where stability is defined as it was for the models based on the DASH features.

vast majority of incidents involve a small number of prior calls.[22] Thus over the range of most of the data, an increase in DA calls is associated with an increase in violent recidivism. The effect of DA crimes for the dyad is also highly non-linear, first rising, then falling, before rising and falling again over a region where data is sparse.

H. Predictions of violent recidivism based on criminal history features

We now consider the predictions of violent recidivism that are generated by these models, which are summarized in the confusion matrices that appear in Table 8. As above, these matrices apply the logistic regressions and random forests from the training sample to data from the independent test sample. The confusion matrices for the logistic regression appears on the left, and those for the random forest appears on the right.

Panel (a) shows that the models trained on the criminal history features at the default cost ratio perform slightly better than those trained on the DASH features at the same cost ratio. Whereas the DASH-based predictions were equivalent to the simple Bayes classifier, the models based on criminal histories at least predict a few dozen incidents of violent recidivism. As a result, the rates of negative prediction error for both models are slightly lower for the criminal history-based models than for the DASH-based models. The AUC statistics are higher as well.

A more important comparison involves the predictions from models trained under asymmetric costs. These are obtained by downsampling the training data, as above. Panel (b) of Table 8 presents predictions from models trained to a 10:1 cost ratio. Confusion matrices for models trained to 5:1 and 15:1 cost ratios ap-

[22] The 90th percentile for prior DA calls for the dyad is 6. For the victim it is also 6 and for the perpetrator it is 7.

pear in Appendix Table A.2. Moving from the DASH features to the criminal history features reduces the rate of negative prediction error. Whereas the logistic regression-based predictions from Panel (b) of Table 6 yielded a negative prediction error rate of 0.077, the corresponding error rate from Table 8 is 0.069. The rate of negative prediction error from the random forest is lower still, at 0.066.[23] The difference between the performance of the logistic regression and the random forest may be due to the fact that the random forest can capture the non-linearities in the effects of the predictors that were evident in Figure 5.

The absolute number of false negative cases within the test sample, based on the random forest trained to the criminal history features at the 10:1 cost ratio, is 587. This compares to 634 for the random forest trained to the DASH features. Projecting and annualizing as above, this amounts to 108 fewer false negatives per year. Put differently, as compared to the predictions from the random forest based on the DASH data, there are roughly 10 fewer cases per month that are predicted to succeed, but which nevertheless result in an incident of violent recidivism.

Panel (c) presents predictions from models trained on subsets of the training data that were subsampled so as to target a predicted failure rate of 8.5 percent, similar to the failure rate achieved by the DASH risk grade. Holding constant the predicted failure rate, both machine learning models achieve a lower rate of negative prediction error than the DASH risk grade in Table 4. They also achieve a lower rate of negative prediction error than the machine learning algorithms trained on the DASH features from Table 6 (c).

[23] This is an example of a lower-AUC classifier performing better in a specific region of the ROC curve than a higher-AUC classifier (Fawcett (2006)).

I. *Machine learning models based on both DASH and criminal history features*

Finally, we ask whether adding the DASH features to the criminal history features improves predictions. Logistic regression coefficients appear in Appendix Table A.3. These include all of the predictors that appear in Tables 5 and 7. Many of the coefficients on both sets of variables are similar to their counterparts from the simpler models above.

We trained a random forest to the same set of features. Feature importances appear in Appendix Figure A.1 and partial dependence plots appear in Appendix Figure A.2. The most important variables come from the criminal history features; their partial dependence plots are generally similar to those from the model based on the criminal history features alone.

J. *Predictions of violent recidivism based on both DASH and criminal history features*

Confusion matrices based on these models appear in Table 9. The layout is the same as that for Table 6 and 8. As above, these matrices apply the logistic regressions and random forests from the training sample to data from the independent test sample.

Panel (a) shows that, at the default cost ratio, the models based on all the features perform the same in terms of negative prediction error as those based on only the criminal history features. At the 10:1 cost ratio, adding the DASH features improves performance by slightly more, reducing the negative prediction error rate of the logistic regression from 0.069 to 0.065 and that of the random forest from 0.066 to 0.061 (see Panel (b)). The AUC statistics from the combined models are a

bit higher than those based on the criminal history features alone. Nonetheless, the practical consequence of the improvement is fairly small. Focusing on the random forests, which perform better than the logistic regressions, it amounts to reducing the number of prediction errors from 587 to 561 in the test sample. Annualizing as above, this comes to about 60 fewer negative prediction errors per year, or about five per month. Adding the DASH features has little effect on the rate of negative prediction error when we target a predicted failure rate of 8.5 percent, as seen in Panel (c).

IV. Robustness

In this section, we report the results from two additional training and forecasting exercises designed to help evaluate the robustness of our results. In the first, we re-train the models after deleting from the sample all observations for which all of the DASH questions had been left blank. As reported above, this amounted to about 10 percent of the observations in the sample.

Appendix Table A.5 presents confusion matrices for the models trained to this smaller data set. Most of the models have a slightly lower rate of negative prediction error than the corresponding models trained to the full sample (see Panel (c) in Tables 6-9). However, in all three cases, the change is small.

In the second exercise, we dropped from the sample all the cases that received a DASH assessment of high risk. We do this for the following reason. A grade of high risk is supposed to trigger an intervention, generally involving a referral to a multi-agency council and specialist services to protect the victim. If that intervention is effective, it means that violent recidivism is averted, at least for some share of such cases. In that situation, a case that would have resulted in a failure results in success

instead. This means that the failure rate for these cases is lower than it would have been, albeit by an unknown amount. Dropping the high-risk cases reduces the influence of protective services on our models. It does not fully eliminate it, since some lower-risk cases may also receive some services. However, such services tend to be of much lower intensity than those provided to high-risk cases (Stanley et al. (2010)). Dropping high-risk cases removes the influence of high-intensity protective interventions and allows us to gauge the extent to which such interventions affect the performance of our models.

We note first that removing these observations from the training sample had little effect on the coefficients in our logistic regressions. The second columns in Tables 5, 7, and Appendix Table A.3. report these estimates. In most cases, the sign of the coefficient in the smaller sample is the same as its counterpart in the full sample. For the most part, the magnitudes are similar as well.

Confusion matrices for both the logistic regressions and random forests, trained at the 10:1 cost ratio, are presented in Appendix Table A.6. These models generally perform similarly to their full-sample counterparts. Apparently, the protective services offered to high-risk victims do not unduly influence our predictions.

In other results that we do not report here, we experimented with different approaches to feature selection. When some features are essentially noise variables, providing little predictive variation beyond that which is available from other features, forecasting performance can suffer. We experimented with dropping variables with low shuffle importance and with forward-selection and backward-elimination approaches. Backward elimination improved performance, but only slightly; the other approaches generally resulted in models that performed similarly or worse. We also experimented with truncating the criminal history features. Our concern was that random forests may overuse long-tailed features in constructing the underlying

classification trees, which may result in overfitting and exaggerate the importance of such variables (Strobl et al. (2007)). However, when we truncated long-tailed features at the 95th percentile, which typically involved substantial truncation, neither model performance nor feature importance were much affected.

V. Discussion and conclusion

One important conclusion from our findings is that a machine learning approach can provide better forecasts of violent recidivism than a structured-judgement approach based on data from the same assessment protocol. A second important finding is that machine learning methods applied to criminal history data provide better forecasts of violent recidivism than the same methods applied to the data from the assessment protocol. A third finding is that adding data from the assessment protocol does little to improve the performance of machine learning forecasts based on criminal history data.

One might wonder why the DASH risk assessment does not perform better. A reasonable first question would be to ask how DASH fares compared to similar instruments. Messing and Thaller (2013) recently evaluated several protocols, including the Ontario Domestic Assault Risk Assessment (ODARA), the Danger Assessment (DA), the Spousal Assault Risk Assessment (SARA), the Domestic Violence Screening Inventory (DVSI) and the Kingston Screening Instrument for Domestic Violence (K-SID). These protocols vary in their data collection methods and in the precise measure of repeat domestic violence that they seek to predict, but like the DASH risk assessment protocol, do not use machine learning to generate their predictions. With one exception, the average AUC for these protocols ranged from roughly 0.58 to 0.67. Thus with the exception of the K-SID, which had an

average AUC of 0.537, the DASH risk-grading system falls well below its peers.

We suspect that several features of DASH may contribute to its low predictive power. First, it consists of many items; many officers are likely to find it difficult to keep track of all 27 questions and the weight that each should receive as the officer formulates her risk grade. This task is made particularly difficult since several of the questions actually have negative, rather than positive, partial effects on the likelihood of violent recidivism, as we noted above. It is further complicated by the fact that many officers handle relatively few DA calls; even the largest number handled by any officer in our data was 282. To top it off, officers may not receive regular updates on the accuracy of their past risk grades, precluding them from learning from their own track record. In contrast, machine learning methods have the benefit of learning from tens of thousands of incidents, for which they see the outcome in each case. On the basis of that information, they learn weights that are optimized to forecast failure, taking into account the relative cost of false negatives to false positives.

These issues are exacerbated by the environment in which the DASH questionnaires are administered. The DASH interview involves an interaction between the victim and a police officer, typically at a moment of considerable stress. The willingness of the victim to provide information may depend on a number of factors, including the officer's sex and ethnicity, as well as other circumstances, such as the situation of the victim at the time (Kirkwood (1993); Jose Medina Ariza, Robinson and Myhill (2016)). Another issue involves the adequacy of training and officers' motivation in conducting the interview (Robinson (2011)). All these issues affect the quality of the DASH data. As one example, Turner, Medina and Brown (2019) note that negative responses to the question about whether the perpetrator had been in trouble with police were wrong about half the time. Of course, these difficulties are

not unique to DASH; Messing and Thaller (2013) note issues in administering the instruments for many of the risk assessment protocols they studied.

Data quality issues may affect all forecasts based on such data. This includes not just those based on the DASH risk grade, but also those based on the machine learning algorithms. We speculate that this may help explain why the machine learning predictions based on the criminal history features perform better than the models based on the DASH features. The criminal histories come from administrative data bases, rather than lengthy questionnaires administered under challenging circumstances.

Of course, the machine learning forecasts have limitations of their own. One is that we are only able to predict violent recidivism that is reported to police. Although more serious incidents of domestic abuse are more likely to be reported (Barrett et al. (2017)), we are unable to observe, and therefore to predict, violent domestic incidents that are not reported.

Another limitation is that we cannot predict other types of serious harm that may result from domestic abuse. There is evidence that non-violent aspects of abuse, including controlling and coercive behavior on the part of the perpetrator, can have serious consequences for the victim (Stark (2007); Johnson (2010)). Violence with injury represents but one form of serious harm facing victims of domestic abuse. Unfortunately, large-scale data on other forms of serious harm are generally unavailable.

We close by discussing how our forecasting procedure might be used to improve police response to DA calls. Considering that our model is most useful for predicting who is unlikely to violently recidivate, we posit that it could be most useful at the time when a DA call first arrives at the call center.

When calls for service arrive, call handlers take basic information and assign

a three-level priority score. The highest priority calls then get the most rapid responses; lower priority calls may take more time. With the right software and digital information system, a model such as ours could be used to prioritize DA calls for service. This would require that call handlers have a computer dashboard with access to a criminal history database, a coded-up version of the forecasting model, and an interface to pass data from the database to the model. When the call came in, the call handler would take enough information to identify the victim and perpetrator in the database. With their identities established, their criminal histories would be digitally retrieved and passed to the model, which would output a prediction. If the prediction were for no violent recidivism, the case would be given a low priority score. If the prediction were for violent recidivism, the call handler would gather other information to determine whether to assign the call a high or intermediate priority. Automating the first step would mean that the call handler could deal with more calls in a given amount of time, and at the same time, have more time to ask questions designed to distinguish true from false positives.[24]

A remaining question is how to deal with the high rate of positive prediction error that is generated by our approach. Here we envision a two-part screening procedure, analogous to medical testing for certain conditions. In testing for breast cancer, for example, the initial screening typically involves a self-examination or a physical examination by a healthcare provider. Problems detected at this initial stage are often referred for mammography, a costlier but more discriminating test. At this point, many initial referrals are determined to be false positives, while predicted positives based on mammography are referred for costlier, but even more

[24] Berk (2012) and Berk and Bleich (2013) propose a similar process in somewhat different settings. Berk (2012) also makes the important point that automated priority rankings should only be overridden on the basis of information not available to the model. Since the model is optimized to make predictions on the basis of the data used to train it, overriding the model's predictions on the basis of such information will generally worsen the quality of the resulting prediction. In contrast, outside information, such as a credible threat to the victim's safety, may provide a strong basis for overriding the priority ranking from the model.

accurate, testing.

In the risk assessment setting, the first screen would be made by the forecasting model applied to the criminal history information. The second screening would be applied to the cases that were predicted to fail, which include a large number of false positives. The idea for the second screen would be to develop an instrument with greater ability to distinguish true from false positives. Although the DASH items were not useful in this role, some of the instruments discussed above may better distinguish the highest- from lower-risk cases.[25] We suspect that such a two-part procedure would do better than the DASH protocol both in prioritizing calls for service and in providing protective resources to victims with the greatest need for them.

[25] In work not reported above, we experimented with a two-stage procedure. The first stage predicted violent recidivism using the criminal histories. The second stage used the DASH items to distinguish true from false positives among the predicted positives from the first stage. The second-stage model predicted very few positives, regardless of the cost ratio.

References

Barrett, Betty Jo, Amy Peirone, Chi Ho Cheung and Nazim Habibov. 2017. "Pathways to police contact for spousal violence survivors: The role of individual and neighborhood factors in survivors' reporting behaviors." *Journal of interpersonal violence* .

Berk, Richard. 2012. *Criminal justice forecasts of risk: A machine learning approach*. Springer Science & Business Media.

Berk, Richard A. 2008. *Statistical learning from a regression perspective*. Vol. 14 Springer.

Berk, Richard A and Justin Bleich. 2013. "Statistical procedures for forecasting criminal behavior: A comparative assessment." *Criminology & Pub. Pol'y* 12:513.

Berk, Richard A and Susan B Sorenson. 2020. "Algorithmic approach to forecasting rare violent events: An illustration based in intimate partner violence perpetration." *Criminology & Public Policy* 19(1):213–233.

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

Berk, Richard A, Yan He and Susan B Sorenson. 2005. "Developing a practical forecasting screener for domestic violence incidents." *Evaluation Review* 29(4):358–383.

Bland, Matthew and Barak Ariel. 2015. "Targeting escalation in reported domestic abuse: Evidence from 36,000 callouts." *International criminal justice review* 25(1):30–53.

Breiman, Leo. 2001. "Random forests." *Machine learning* 45(1):5–32.

Buzawa, Eve S and Carl G Buzawa. 2017. *Global Responses to Domestic Violence*. Springer.

Campbell, Jacquelyn C, D Webster and P Mahoney. 2005. "Intimate Partner Violence Risk Assessment Validation Study. Final Report.".

Campbell, Jacquelyn C, Daniel W Webster and Nancy Glass. 2009. "The danger assessment: Validation of a lethality risk assessment instrument for intimate partner femicide." *Journal of interpersonal violence* 24(4):653–674.

Crown Prosecution Service. 2020. "Domestic abuse." http://www.cps.gov.uk/domestic-abuse. Accessed: 2020-01-20.

Dutton, Donald G and P Randall Kropp. 2000. "A review of domestic violence risk instruments." *Trauma, violence, & abuse* 1(2):171–181.

Ericson, Richard V., 1948. 1997. *Policing the risk society*. Toronto, Ont.: University of Toronto Press.
URL: *http://search.ebscohost.com/login.aspx?direct=truescope=sitedb=e000xnaAN=683064*

Farrington, David P and Roger Tarling. 1985. "Criminological prediction: An introduction." *Prediction in criminology* pp. 2–33.

Fawcett, Tom. 2006. "An introduction to ROC analysis." *Pattern recognition letters* 27(8):861–874.

Gottfredson, Stephen D and Laura J Moriarty. 2006. "Statistical risk assessment: Old problems and new applications." *Crime & Delinquency* 52(1):178–200.

Grove, William M and Paul E Meehl. 1996. "Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy." *Psychology, public policy, and law* 2(2):293.

Hastie, Trevor, Robert Tibshirani and Jerome Friedman. 2009. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media.

Her Majesty's Inspectorate of Constabulary. 2014. "Everyone's business: Improving the police response to domestic abuse." *Report, HMIC, UK* .

Home Office. 2020. "Counting Rules for Violence Against the Person." *Report, Home Office, UK* .
URL: *https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment$_d$a violence – apr – 2020.pdf*

Johnson, Michael P. 2010. *A typology of domestic violence: Intimate terrorism, violent resistance, and situational couple violence*. Upne.

Jose Medina Ariza, Juan, Amanda Robinson and Andy Myhill. 2016. "Cheaper, faster, better: Expectations and achievements in police risk assessment of domestic abuse." *Policing: a journal of policy and practice* 10(4):341–350.

Kahneman, Daniel, Andrew M Rosenfield, Linnea Gandhi and Tom Blaser. 2016. "Noise: How to overcome the high, hidden cost of inconsistent decision making." *Harvard business review* 94(10):38–46.

Kirkwood, Catherine. 1993. *Leaving abusive partners: From the scars of survival to the wisdom for change*. Sage.

Kropp, P Randall. 2004. "Some questions regarding spousal assault risk assessment." *Violence against women* 10(6):676–697.

Messing, Jill Theresa and Jonel Thaller. 2013. "The average predictive validity of intimate partner violence risk assessment instruments." *Journal of interpersonal violence* 28(7):1537–1558.

National Coalition Against Domestic Violence. 2014. "Domestic Violence Facts." *Crisis* 402:826–2332.

Richards, Laura. 2009. "Domestic abuse, stalking and harassment and honour based violence (DASH, 2009) risk identification and assessment and management model.".

Robinson, Amanda L. 2011. "Risk and intimate partner violence.".

Robinson, Amanda L, Andy Myhill, Julia Wire, Jo Roberts and Nick Tilley. 2016. "Risk-led policing of domestic abuse and the DASH risk model." *What Works: Crime Reduction Research. Cardiff & London: Cardiff University, College of Policing and UCL Department of Security and Crime Science* .

Roehl, Janice, Chris Sullivan, Daniel Webster and Jacquelyn Campbell. 2005. "Intimate Partner Violence Risk Assessment Validation Study, Final Report." *Assessment* .

Stanley, Nicky, Pam Miller, Helen Richardson Foster and Gillian Thomson. 2010. "Children and families experiencing domestic violence: police and childen's social services' responses.".

Stark, Evan. 2007. *Coercive control: the entrapment of women in personal life*. Oxford University Press.

Strobl, Carolin, Anne-Laure Boulesteix, Achim Zeileis and Torsten Hothorn. 2007. "Bias in random forest variable importance measures: Illustrations, sources and a solution." *BMC bioinformatics* 8(1):25.

Turner, Emily, Juanjo Medina and Gavin Brown. 2019. "Dashing Hopes? The Predictive Accuracy of Domestic Abuse Risk Assessment by Police." *The British Journal of Criminology* .

Figure 1: Timeline showing availability of data and sample period

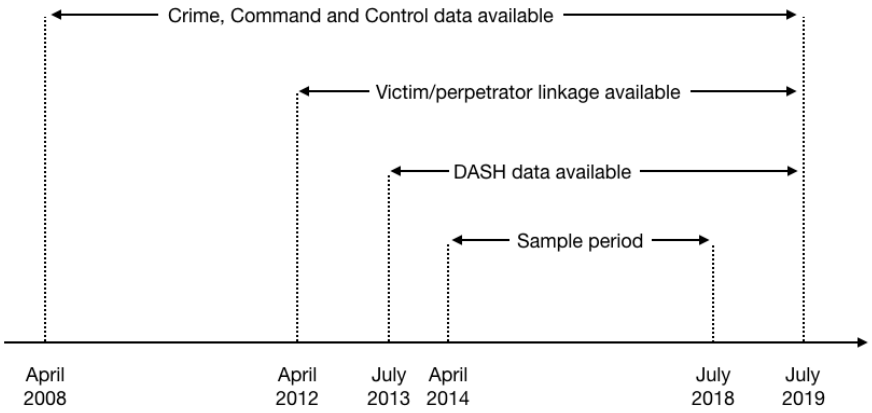


Figure 2: Predictor importance for random forest based on DASH features

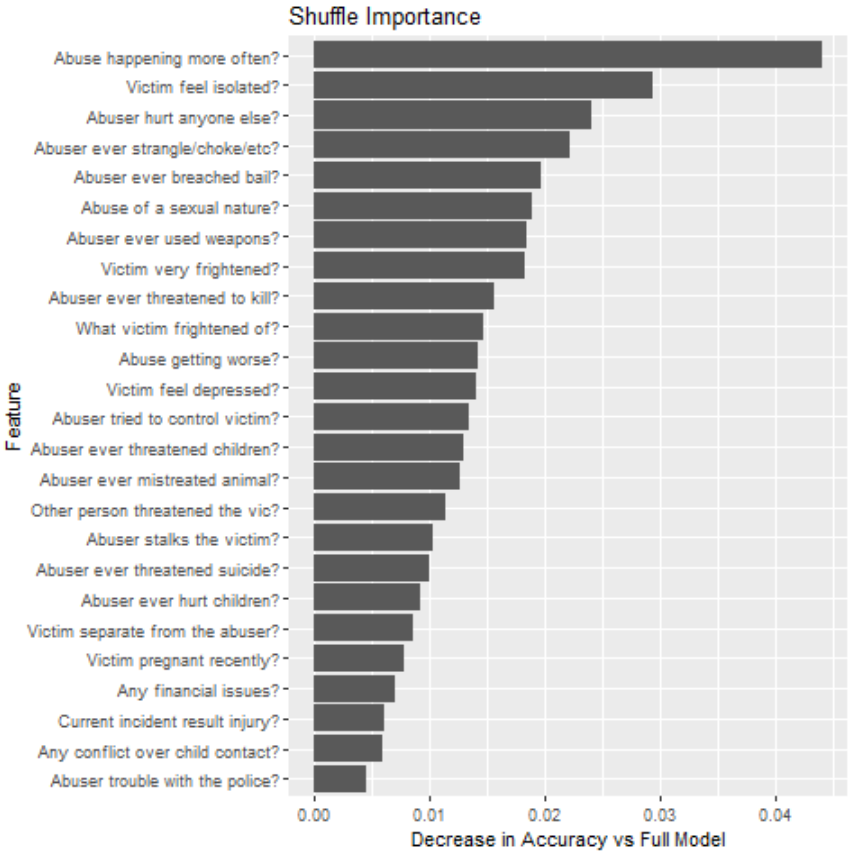
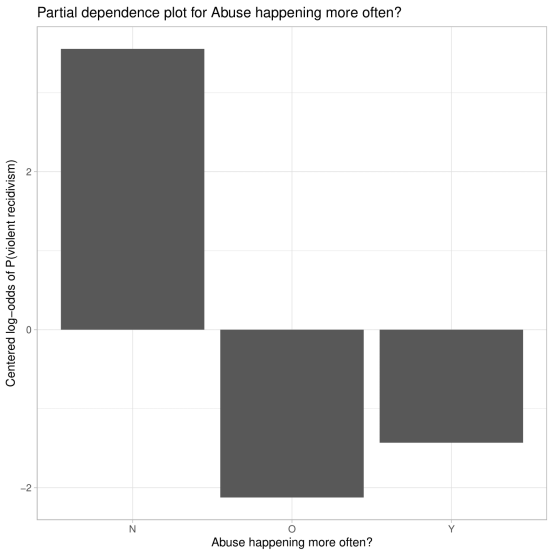
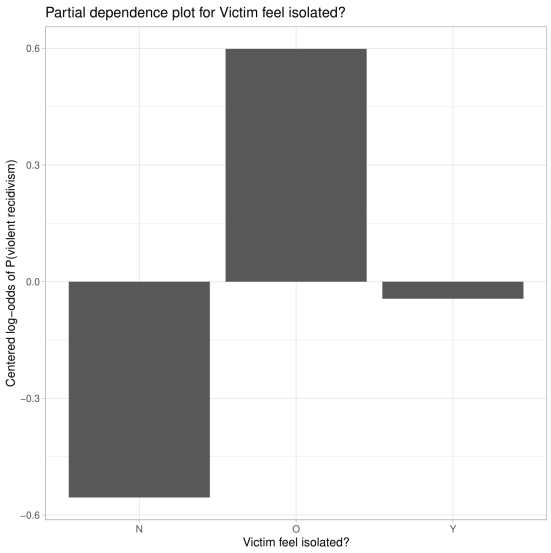


Figure 3: Partial dependence plots for selected features from random forest based on DASH features

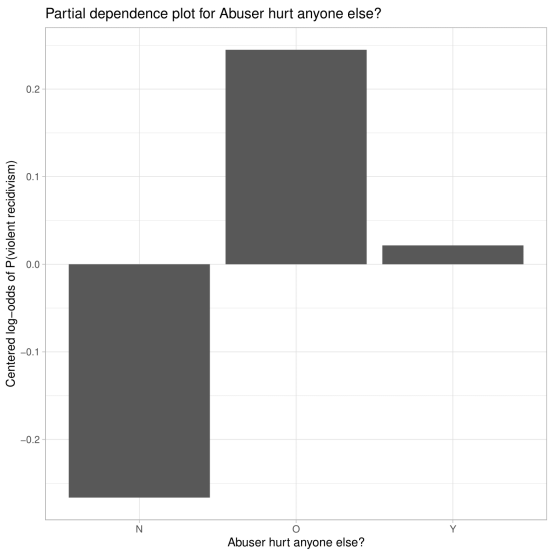
(a)



(b)



(c)



(d)

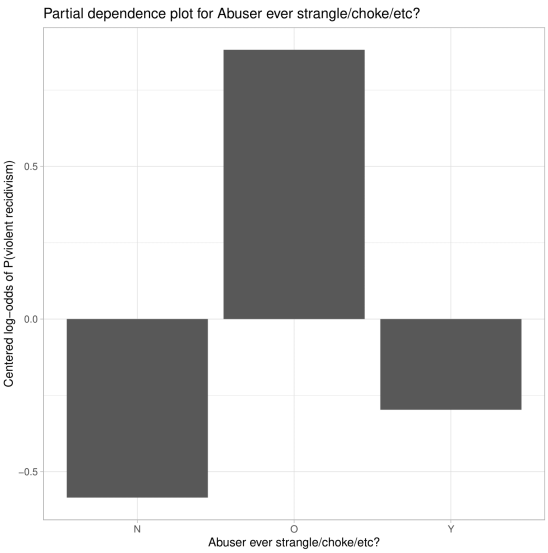


Figure 4: Predictor importance for random forest based on criminal history features

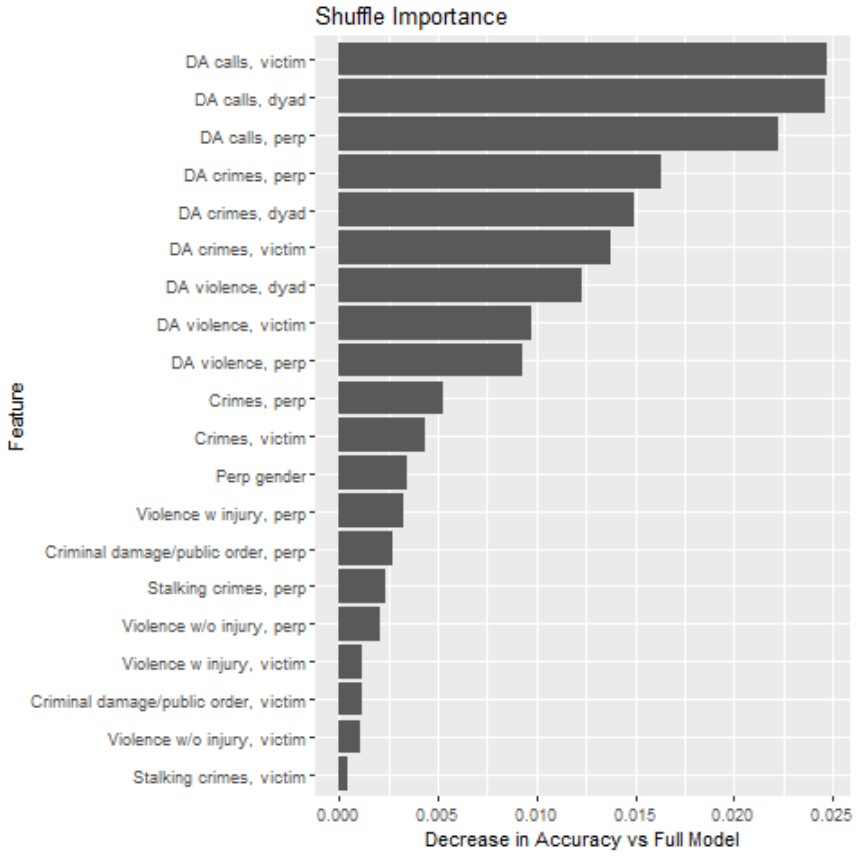
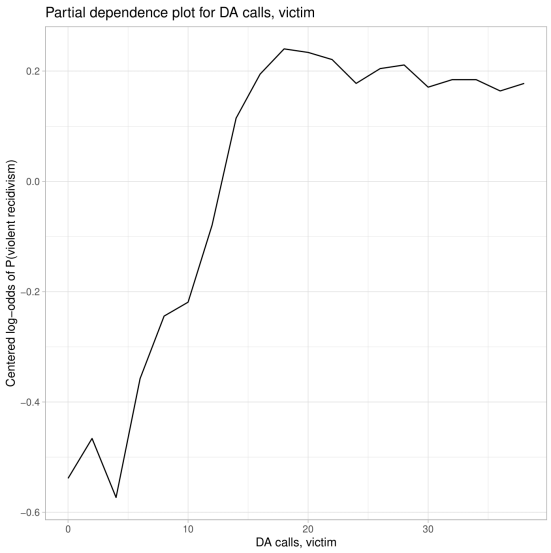
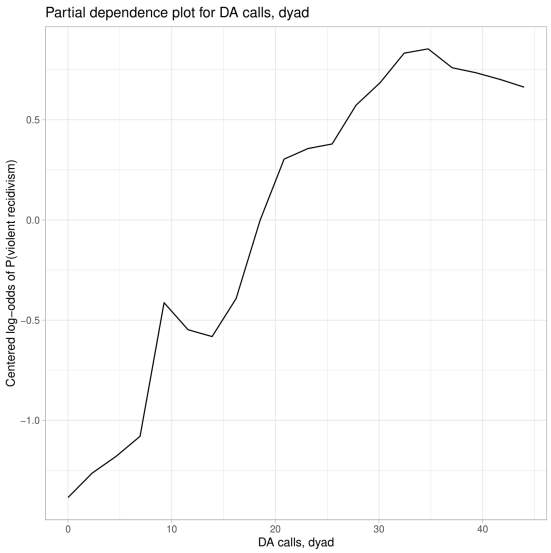


Figure 5: Partial dependence plots for selected features from random forest based on criminal history features

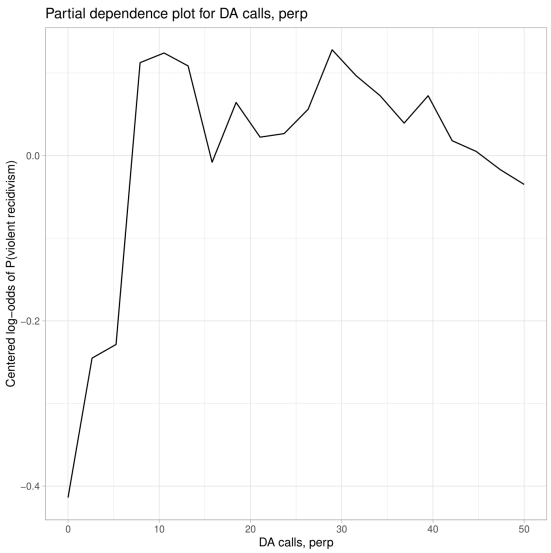
(a)



(b)



(c)



(d)

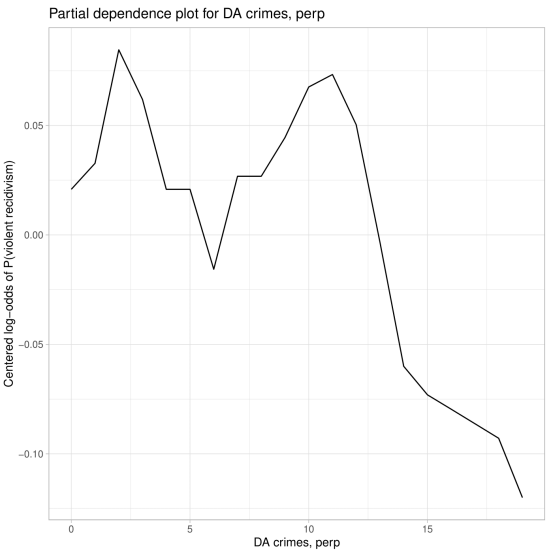


Table 1: Frequency Distribution of Calls for Service for Domestic Abuse, by Dyad, Probability of Repeat Call, and Probability of Violent Recidivism

Call number	Frequency	Relative frequency	Probability of repeat call	Probability of violent recidivism
1	72,972	0.442	0.425	0.067
2	31,014	0.188	0.574	0.107
3	17,812	0.108	0.643	0.134
4	11,462	0.069	0.683	0.154
5	7,831	0.047	0.700	0.163
6	5,484	0.033	0.726	0.181
7	3,979	0.024	0.736	0.204
8+	14,510	0.088	0.798	0.288
Overall	165,064	1	0.558	0.118

Table 2: DASH Questions, Response Frequencies, and Distribution of Risk Assessments

(a) DASH questions and Response Frequencies

	Y	N	O
Current incident result injury?	0.136	0.737	0.127
Victim very frightened?	0.283	0.547	0.170
What victim frightened of?	0.299	0.519	0.182
Victim feel isolated?	0.126	0.689	0.185
Victim feel depressed?	0.174	0.637	0.190
Victim separate from the abuser?	0.451	0.369	0.180
Any conflict over child contact?	0.151	0.676	0.173
Abuser stalks the victim?	0.190	0.622	0.188
Victim pregnant recently?	0.162	0.665	0.172
Any children not the abuser's?	0.154	0.673	0.174
Abuser ever hurt children?	0.028	0.785	0.187
Abuser ever threatened children?	0.025	0.786	0.189
Abuse happening more often?	0.205	0.600	0.194
Abuse getting worse?	0.198	0.607	0.195
Abuser tried to control victim?	0.254	0.548	0.198
Abuser ever used weapons?	0.093	0.707	0.200
Abuser ever threatened to kill?	0.103	0.695	0.202
Abuser ever strangle/choke/etc?	0.132	0.665	0.202
Abuse of a sexual nature?	0.074	0.721	0.206
Other person threatened the vic?	0.034	0.764	0.201
Abuser hurt anyone else?	0.116	0.683	0.202
Abuser ever mistreated animal?	0.036	0.763	0.201
Any financial issues?	0.166	0.637	0.197
Abuser had problems alcohol,etc?	0.373	0.435	0.192
Abuser ever threatened suicide?	0.171	0.626	0.203
Abuser ever breached bail?	0.109	0.691	0.199
Abuser trouble with the police?	0.448	0.365	0.188
Other relevant information?	0.186	0.674	0.140

(b) Risk Assessment

	Standard	Medium	High
Risk Assessment	0.602	0.308	0.090

NOTE: Sample size is 165,064.

Table 3: Means and Standard Deviations of Predictor Variables Derived from Two-year Criminal and Domestic Abuse Histories

	Mean	Std.Dev.
Perp is male	0.834	0.372
DA calls, dyad	2.416	4.283
DA crimes, dyad	0.787	1.549
DA violence, dyad	0.226	0.623
DA calls, perp	2.457	4.075
DA crimes, perp	0.834	1.580
DA violence, perp	0.217	0.588
DA calls, victim	2.307	3.873
DA crimes, victim	0.780	1.506
DA violence, victim	0.214	0.594
Violence w injury, perp	0.150	0.470
Violence w/o injury, perp	0.149	0.490
Criminal damage/public order, perp	0.182	0.681
Crimes, perp	0.864	1.957
Stalking crimes, perp	0.100	0.524
Violence w injury, victim	0.047	0.256
Violence w/o injury, victim	0.049	0.295
Criminal damage/public order, victim	0.051	0.332
Crimes, victim	0.262	0.987
Stalking crimes, victim	0.015	0.170

NOTE: Sample size is 165,064. Criminal and DA histories calculated from April 2012 to June 2018. Perp stands for perpetrator.

Table 4: Violent Recidivism by DASH Risk Assessment

	Lesser Risk	High Risk	Row Share	Class Error
Actual No	13121	1165	0.882	0.082
Actual Yes	1702	215	0.118	0.888
Col Share	0.915	0.085	1	
Pred Error	0.115	0.844		0.177

NOTE: AUC=0.515

Table 5: Logistic Regression of Violent Recidivism on Responses to DASH Questions

	Dependent variable:	
	Violent recidivism	
	(1)	(2)
Current incident result injury? O	−0.003	−0.005
	(0.041)	(0.043)
Current incident result injury? Y	0.206	0.184
	(0.027)	(0.030)
Victim very frightened? O	−0.027	−0.040
	(0.070)	(0.074)
Victim very frightened? Y	−0.232	−0.240
	(0.039)	(0.041)
What victim frightened of? O	0.095	0.105
	(0.074)	(0.078)
What victim frightened of? Y	−0.117	−0.082
	(0.039)	(0.041)
Victim feel isolated? O	0.010	−0.047
	(0.084)	(0.089)
Victim feel isolated? Y	0.182	0.207
	(0.033)	(0.036)
Victim feel depressed? O	−0.104	−0.090
	(0.084)	(0.090)
Victim feel depressed? Y	0.111	0.104
	(0.028)	(0.030)
Victim separate from the abuser? O	0.263	0.299
	(0.071)	(0.076)
Victim separate from the abuser? Y	0.090	0.095
	(0.024)	(0.025)
Any conflict over child contact? O	−0.284	−0.256
	(0.069)	(0.074)
Any conflict over child contact? Y	−0.669	−0.647
	(0.041)	(0.043)
Abuser stalks the victim? O	0.249	0.223
	(0.086)	(0.094)
Abuser stalks the victim? Y	−0.251	−0.279
	(0.030)	(0.032)
Victim pregnant recently? O	0.011	0.028
	(0.067)	(0.071)
Victim pregnant recently? Y	0.137	0.149
	(0.035)	(0.037)

Any children not the abuser's? O	0.054	0.052
	(0.075)	(0.079)
Any children not the abuser's? Y	−0.208	−0.174
	(0.035)	(0.036)
Abuser ever hurt children? O	−0.223	−0.222
	(0.112)	(0.121)
Abuser ever hurt children? Y	−0.524	−0.503
	(0.076)	(0.086)
Abuser ever threatened children? O	−0.212	−0.208
	(0.111)	(0.121)
Abuser ever threatened children? Y	−0.194	−0.163
	(0.073)	(0.086)
Abuse happening more often? O	0.060	0.083
	(0.119)	(0.133)
Abuse happening more often? Y	0.208	0.203
	(0.028)	(0.030)
Abuse getting worse? O	0.070	0.074
	(0.123)	(0.137)
Abuse getting worse? Y	−0.172	−0.148
	(0.030)	(0.033)
Abuser tried to control victim? O	−0.153	−0.098
	(0.100)	(0.109)
Abuser tried to control victim? Y	−0.001	0.007
	(0.028)	(0.029)
Abuser ever used weapons? O	0.292	0.236
	(0.106)	(0.120)
Abuser ever used weapons? Y	0.112	0.161
	(0.039)	(0.044)
Abuser ever threatened to kill? O	−0.091	−0.149
	(0.129)	(0.142)
Abuser ever threatened to kill? Y	−0.167	−0.126
	(0.038)	(0.044)
Abuser ever strangle/choke/etc? O	0.238	0.319
	(0.108)	(0.122)
Abuser ever strangle/choke/etc? Y	0.089	0.097
	(0.034)	(0.037)
Abuse of a sexual nature? O	−0.211	−0.187
	(0.110)	(0.127)
Abuse of a sexual nature? Y	−0.082	−0.074
	(0.043)	(0.047)
Other person threatened the vic? O	0.069	0.067
	(0.136)	(0.153)

Other person threatened the vic? Y	−0.092	−0.131
	(0.054)	(0.065)
Abuser hurt anyone else? O	0.192	0.181
	(0.110)	(0.123)
Abuser hurt anyone else? Y	0.004	−0.016
	(0.033)	(0.037)
Abuser ever mistreated animal? O	0.087	0.035
	(0.115)	(0.125)
Abuser ever mistreated animal? Y	−0.203	−0.180
	(0.062)	(0.072)
Any financial issues? O	0.080	0.080
	(0.081)	(0.086)
Any financial issues? Y	−0.039	−0.056
	(0.029)	(0.031)
Abuser had problems alcohol,etc? O	−0.007	−0.013
	(0.066)	(0.070)
Abuser had problems alcohol,etc? Y	0.375	0.357
	(0.029)	(0.030)
Abuser ever threatened suicide? O	−0.062	−0.085
	(0.095)	(0.105)
Abuser ever threatened suicide? Y	−0.129	−0.109
	(0.032)	(0.034)
Abuser ever breached bail? O	0.245	0.229
	(0.075)	(0.081)
Abuser ever breached bail? Y	0.190	0.202
	(0.041)	(0.044)
Abuser trouble with the police? O	0.178	0.174
	(0.059)	(0.063)
Abuser trouble with the police? Y	0.358	0.349
	(0.025)	(0.026)
Other relevant information? O	0.012	0.035
	(0.039)	(0.041)
Other relevant information? Y	0.118	0.129
	(0.023)	(0.025)
High risk	0.289	
	(0.041)	
Constant	−2.441	−2.451
	(0.027)	(0.028)
Observations	148,861	135,356

Note: Standard errors, in parentheses, have been clustered by dyad.

Table 6: Confusion Matrices from Logistic Regressions and Random Forests Based on DASH Features

Panel (a)

Default, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14286	0	0.882	0
Actual Yes	1917	0	0.118	1
Col Share	1	0	1	
Pred Error	0.118	NaN		0.118

AUC=0.636

Default, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14279	7	0.882	0
Actual Yes	1914	3	0.118	0.998
Col Share	0.999	0.001	1	
Pred Error	0.118	0.7		0.119

AUC= 0.518

Panel (b)

10:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7603	6683	0.882	0.468
Actual Yes	635	1282	0.118	0.331
Col Share	0.508	0.492	1	
Pred Error	0.077	0.839		0.452

AUC=0.636

10:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7403	6883	0.882	0.482
Actual Yes	634	1283	0.118	0.331
Col Share	0.496	0.504	1	
Pred Error	0.079	0.843		0.464

AUC=0.611

Panel (c)

Targeting 8.5% predicted failures, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13345	941	0.882	0.066
Actual Yes	1620	297	0.118	0.845
Col Share	0.924	0.076	1	
Pred Error	0.108	0.76		0.158

AUC=0.635

Targeting 8.5% predicted failures, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13489	797	0.882	0.056
Actual Yes	1674	243	0.118	0.873
Col Share	0.936	0.064	1	
Pred Error	0.11	0.766		0.153

AUC=0.568

Table 7: Logistic Regression of Violent Recidivism on Criminal History Variables

	Dependent variable:	
	Violent recidivism	
	(1)	(2)
Perp is male	−0.281	−0.264
	(0.032)	(0.032)
DA calls, dyad	0.053	0.056
	(0.015)	(0.017)
DA crimes, dyad	−0.111	−0.114
	(0.037)	(0.039)
DA violence, dyad	0.334	0.355
	(0.054)	(0.055)
DA calls, perp	0.008	0.003
	(0.014)	(0.014)
DA crimes, perp	0.060	0.072
	(0.034)	(0.035)
DA violence, perp	−0.130	−0.141
	(0.061)	(0.063)
DA calls, victim	0.001	0.002
	(0.015)	(0.015)
DA crimes, victim	0.080	0.089
	(0.036)	(0.036)
DA violence, victim	0.173	0.160
	(0.061)	(0.061)
Violence w injury, perp	0.057	0.089
	(0.038)	(0.041)
Violence w/o injury, perp	0.095	0.113
	(0.034)	(0.037)
Criminal damage/public order, perp	0.032	0.027
	(0.028)	(0.031)
Crimes, perp	0.028	0.032
	(0.011)	(0.012)
Stalking crimes, perp	−0.160	−0.180
	(0.047)	(0.048)
Violence w injury, victim	0.204	0.214
	(0.052)	(0.054)
Violence w/o injury, victim	0.207	0.214
	(0.043)	(0.044)
Criminal damage/public order, victim	0.099	0.095
	(0.049)	(0.050)

Crimes, victim	0.078	0.071
	(0.016)	(0.017)
Stalking crimes, victim	0.063	0.057
	(0.077)	(0.079)
Constant	−2.230	−2.259
	(0.033)	(0.034)
Observations	148,861	135,356

Note: Standard errors, in parentheses, have been clustered by dyad.

Table 8: Confusion Matrices from Logistic Regressions and Random Forests Based on Criminal History Features

Panel (a)

Default, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14212	74	0.882	0.005
Actual Yes	1809	108	0.118	0.944
Col Share	0.989	0.011	1	
Pred Error	0.113	0.407		0.116

AUC=0.680

Default, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14190	96	0.882	0.007
Actual Yes	1823	94	0.118	0.951
Col Share	0.988	0.012	1	
Pred Error	0.114	0.505		0.118

AUC=0.665

Panel (b)

10:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8537	5749	0.882	0.402
Actual Yes	629	1288	0.118	0.328
Col Share	0.566	0.434	1	
Pred Error	0.069	0.817		0.394

AUC=0.684

10:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8302	5984	0.882	0.419
Actual Yes	587	1330	0.118	0.306
Col Share	0.549	0.451	1	
Pred Error	0.066	0.818		0.406

AUC=0.665

Panel (c)

Targeting 8.5% predicted failures, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13405	881	0.882	0.062
Actual Yes	1467	450	0.118	0.765
Col Share	0.918	0.082	1	
Pred Error	0.099	0.662		0.145

AUC=0.683

Targeting 8.5% predicted failures, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13354	932	0.882	0.065
Actual Yes	1461	456	0.118	0.762
Col Share	0.914	0.086	1	
Pred Error	0.099	0.671		0.148

AUC=0.687

Table 9: Confusion Matrices from Logistic Regressions and Random Forests Based on All Features

Panel (a)

Default, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14199	87	0.882	0.006
Actual Yes	1805	112	0.118	0.942
Col Share	0.988	0.012	1	
Pred Error	0.113	0.437		0.117

AUC=0.700

Default, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	14203	83	0.882	0.006
Actual Yes	1819	98	0.118	0.949
Col Share	0.989	0.011	1	
Pred Error	0.114	0.459		0.117

AUC=0.687

Panel (b)

10:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8626	5660	0.882	0.396
Actual Yes	596	1321	0.118	0.311
Col Share	0.569	0.431	1	
Pred Error	0.065	0.811		0.386

AUC=0.706

10:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8566	5720	0.882	0.4
Actual Yes	561	1356	0.118	0.293
Col Share	0.563	0.437	1	
Pred Error	0.061	0.808		0.388

AUC=0.696

Panel (c)

Targeting 8.5% predicted failures, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13430	856	0.882	0.06
Actual Yes	1434	483	0.118	0.748
Col Share	0.917	0.083	1	
Pred Error	0.096	0.639		0.141

AUC=0.702

Targeting 8.5% predicted failures, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	13395	891	0.882	0.062
Actual Yes	1467	450	0.118	0.765
Col Share	0.917	0.083	1	
Pred Error	0.099	0.664		0.146

AUC=0.700

Appendix

Data Appendix

Our starting point is the Greater Manchester Police Public Protection Investigation database. There is one record in this database for each offense stemming from each call for service, although all DA calls generate at least one record, even if no charges are filed. We started with 275,954 DA records generated between April 2014 and July 2018. Although data are collected only for the primary victim of each incident, some incidents involve multiple perpetrators. Constructing one record for each victim-perpetrator dyad gave us 284,252 records. Out of these 284,252 records there are 257,145 unique calls to service. From these we dropped 11,832 records for which either the victim's or perpetrator's identity was unknown, since we could not link these calls to criminal histories.

For records where multiple crimes are recorded, we keep the most severe crime (ranging from Homicide, Rape, Sexual offences, Violence with injury, Violence without injury, down to Public order offences, and Fraud at the bottom). This resulted in dropping 18,707 records corresponding to the lesser offenses. Dropping multiple calls on the same day led to a further reduction of 3,114 records.

Of the 250,599 records remaining, 177,957 were classified as intimate partners, as defined in the text. Of those, 165,241 involved male-on-female or female-on-male incidents. A further 177 calls had no risk grade or no DASH questionnaire (as distinct from all responses being checked as omitted). Dropping these left us with our sample of 165,064 records.

Figure A.1: Predictor importance for random forest, based on all features

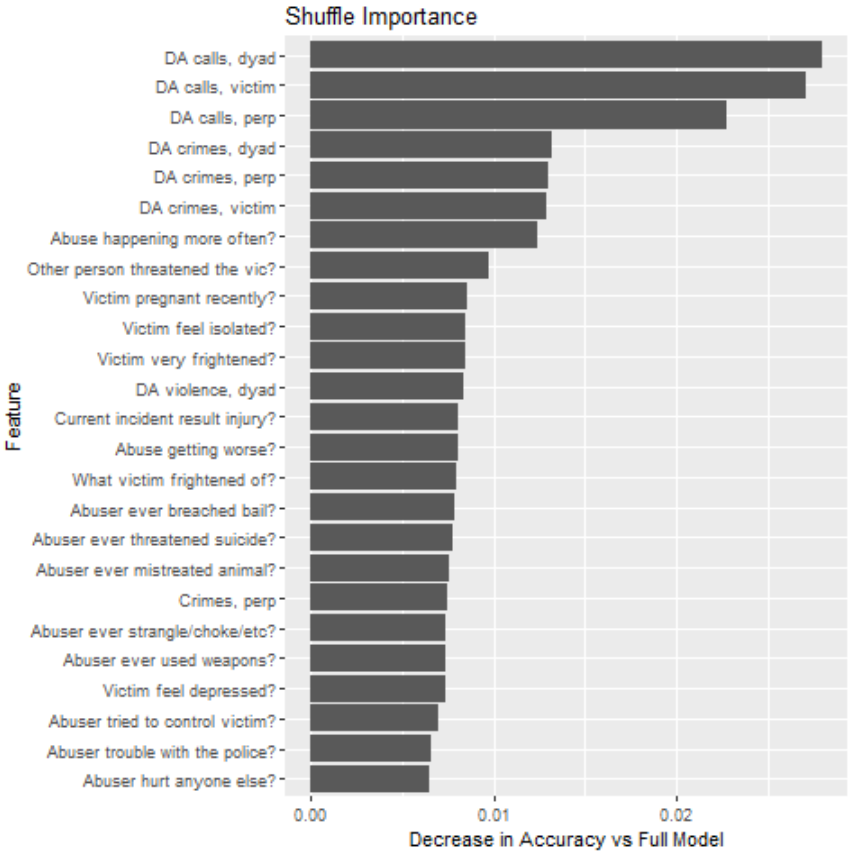
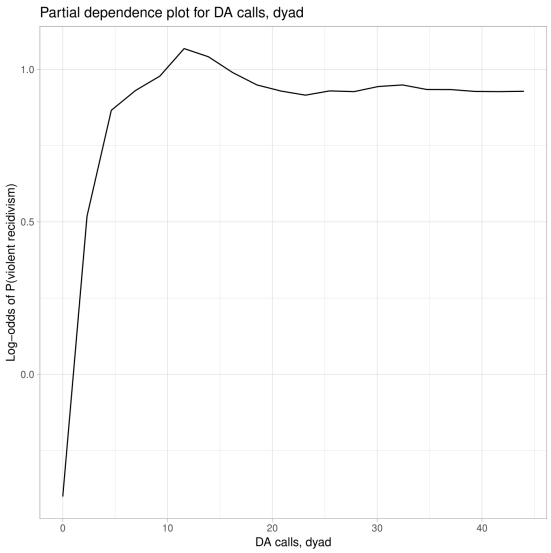
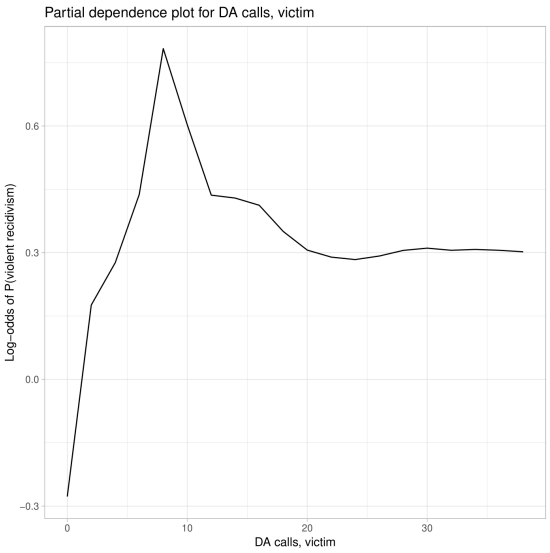


Figure A.2: Partial dependence plots for selected features from random forest based on all features

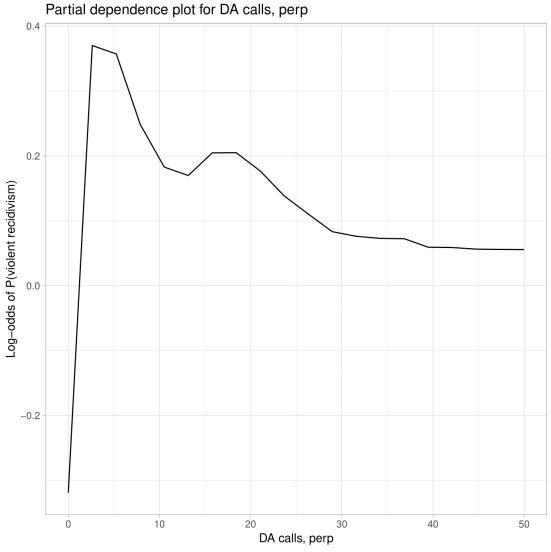
(a)



(b)



(c)



(d)

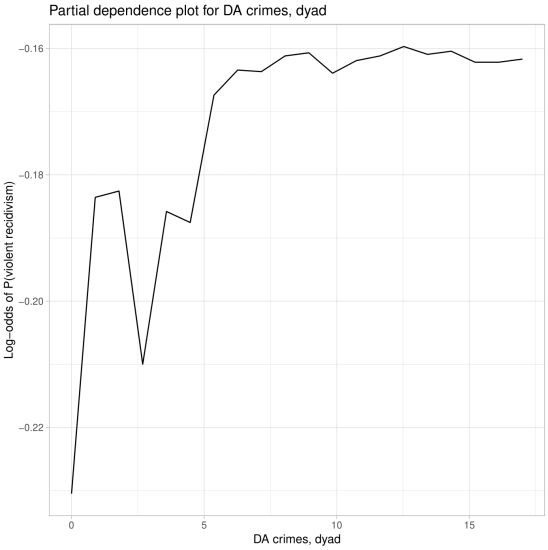


Table A.1: Confusion Matrices from Logistic Regressions and Random Forests Based on DASH Features, Alternative Cost Ratios

Panel (a)

5:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	9503	4783	0.882	0.335
Actual Yes	900	1017	0.118	0.469
Col Share	0.642	0.358	1	
Pred Error	0.087	0.825		0.351

AUC=0.637

5:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	9543	4743	0.882	0.332
Actual Yes	949	968	0.118	0.495
Col Share	0.648	0.352	1	
Pred Error	0.09	0.831		0.351

AUC=0.611

Panel (b)

15:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	6519	7767	0.882	0.544
Actual Yes	502	1415	0.118	0.262
Col Share	0.433	0.567	1	
Pred Error	0.071	0.846		0.51

AUC=0.634

15:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	6681	7605	0.882	0.532
Actual Yes	549	1368	0.118	0.286
Col Share	0.446	0.554	1	
Pred Error	0.076	0.848		0.503

AUC=0.614

Table A.2: Confusion Matrices from Logistic Regressions and Random Forests Based on Criminal Histories, Alternative Cost Ratios

Panel (a)

5:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	10165	4121	0.882	0.288
Actual Yes	837	1080	0.118	0.437
Col Share	0.679	0.321	1	
Pred Error	0.076	0.792		0.306

AUC=0.685

5:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	10082	4204	0.882	0.294
Actual Yes	839	1078	0.118	0.438
Col Share	0.674	0.326	1	
Pred Error	0.077	0.796		0.311

AUC=0.677

Panel (b)

15:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7877	6409	0.882	0.449
Actual Yes	569	1348	0.118	0.297
Col Share	0.521	0.479	1	
Pred Error	0.067	0.826		0.431

AUC=0.686

15:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7284	7002	0.882	0.49
Actual Yes	466	1451	0.118	0.243
Col Share	0.478	0.522	1	
Pred Error	0.06	0.828		0.461

AUC=0.648

Table A.3: Logistic Regression of Violent Recidivism on All Variables

	Dependent variable:	
	Violent recidivism	
	(1)	(2)
Current incident result injury? O	−0.013	−0.019
	(0.043)	(0.046)
Current incident result injury? Y	0.231	0.210
	(0.027)	(0.030)
Victim very frightened? O	−0.082	−0.098
	(0.074)	(0.078)
Victim very frightened? Y	−0.182	−0.191
	(0.040)	(0.043)
What victim frightened of? O	0.100	0.116
	(0.077)	(0.081)
What victim frightened of? Y	−0.080	−0.045
	(0.041)	(0.043)
Victim feel isolated? O	−0.003	−0.050
	(0.089)	(0.093)
Victim feel isolated? Y	0.195	0.214
	(0.033)	(0.037)
Victim feel depressed? O	−0.032	−0.010
	(0.087)	(0.093)
Victim feel depressed? Y	0.071	0.064
	(0.028)	(0.031)
Victim separate from the abuser? O	0.142	0.168
	(0.074)	(0.078)
Victim separate from the abuser? Y	0.018	0.024
	(0.025)	(0.025)
Any conflict over child contact? O	−0.257	−0.222
	(0.074)	(0.078)
Any conflict over child contact? Y	−0.579	−0.565
	(0.040)	(0.043)
Abuser stalks the victim? O	0.258	0.218
	(0.089)	(0.096)
Abuser stalks the victim? Y	−0.246	−0.274
	(0.030)	(0.033)
Victim pregnant recently? O	−0.007	0.020
	(0.072)	(0.076)
Victim pregnant recently? Y	0.226	0.236
	(0.034)	(0.036)

Any children not the abuser's? O	0.061	0.073
	(0.080)	(0.084)
Any children not the abuser's? Y	−0.065	−0.038
	(0.033)	(0.035)
Abuser ever hurt children? O	−0.261	−0.239
	(0.121)	(0.130)
Abuser ever hurt children? Y	−0.422	−0.443
	(0.077)	(0.086)
Abuser ever threatened children? O	−0.105	−0.127
	(0.119)	(0.128)
Abuser ever threatened children? Y	−0.173	−0.163
	(0.073)	(0.087)
Abuse happening more often? O	0.113	0.106
	(0.123)	(0.134)
Abuse happening more often? Y	0.207	0.206
	(0.029)	(0.031)
Abuse getting worse? O	0.067	0.095
	(0.130)	(0.142)
Abuse getting worse? Y	−0.063	−0.051
	(0.031)	(0.033)
Abuser tried to control victim? O	−0.098	−0.020
	(0.105)	(0.114)
Abuser tried to control victim? Y	0.065	0.073
	(0.028)	(0.029)
Abuser ever used weapons? O	0.140	0.032
	(0.108)	(0.120)
Abuser ever used weapons? Y	−0.017	0.018
	(0.038)	(0.043)
Abuser ever threatened to kill? O	−0.099	−0.137
	(0.132)	(0.144)
Abuser ever threatened to kill? Y	−0.153	−0.156
	(0.039)	(0.045)
Abuser ever strangle/choke/etc? O	0.114	0.180
	(0.112)	(0.125)
Abuser ever strangle/choke/etc? Y	0.084	0.083
	(0.034)	(0.038)
Abuse of a sexual nature? O	−0.183	−0.179
	(0.112)	(0.129)
Abuse of a sexual nature? Y	−0.061	−0.065
	(0.043)	(0.049)
Other person threatened the vic? O	0.090	0.102
	(0.142)	(0.154)

Other person threatened the vic? Y	−0.117	−0.178
	(0.056)	(0.067)
Abuser hurt anyone else? O	0.139	0.154
	(0.115)	(0.128)
Abuser hurt anyone else? Y	0.061	0.031
	(0.034)	(0.038)
Abuser ever mistreated animal? O	0.097	0.044
	(0.121)	(0.132)
Abuser ever mistreated animal? Y	−0.203	−0.169
	(0.064)	(0.074)
Any financial issues? O	0.040	0.031
	(0.084)	(0.090)
Any financial issues? Y	−0.007	−0.019
	(0.029)	(0.031)
Abuser had problems alcohol,etc? O	0.052	0.048
	(0.069)	(0.072)
Abuser had problems alcohol,etc? Y	0.222	0.203
	(0.027)	(0.028)
Abuser ever threatened suicide? O	−0.111	−0.114
	(0.093)	(0.102)
Abuser ever threatened suicide? Y	−0.103	−0.082
	(0.032)	(0.034)
Abuser ever breached bail? O	0.183	0.162
	(0.076)	(0.083)
Abuser ever breached bail? Y	−0.251	−0.245
	(0.041)	(0.046)
Abuser trouble with the police? O	0.105	0.099
	(0.063)	(0.066)
Abuser trouble with the police? Y	0.212	0.202
	(0.026)	(0.027)
Other relevant information? O	0.037	0.052
	(0.041)	(0.043)
Other relevant information? Y	0.091	0.087
	(0.024)	(0.026)
High risk	−0.010	
	(0.040)	
Perp is male	−0.276	−0.272
	(0.032)	(0.033)
DA calls, dyad	0.046	0.050
	(0.015)	(0.017)
DA crimes, dyad	−0.073	−0.081
	(0.035)	(0.037)

DA violence, dyad	0.284	0.310
	(0.053)	(0.054)
DA calls, perp	0.005	0.001
	(0.013)	(0.013)
DA crimes, perp	0.063	0.072
	(0.032)	(0.034)
DA violence, perp	−0.124	−0.130
	(0.058)	(0.061)
DA calls, victim	0.003	0.003
	(0.014)	(0.015)
DA crimes, victim	0.070	0.080
	(0.035)	(0.035)
DA violence, victim	0.152	0.139
	(0.059)	(0.059)
Violence w injury, perp	0.082	0.105
	(0.037)	(0.040)
Violence w/o injury, perp	0.100	0.116
	(0.034)	(0.037)
Criminal damage/public order, perp	0.041	0.034
	(0.028)	(0.030)
Crimes, perp	0.018	0.022
	(0.012)	(0.012)
Stalking crimes, perp	−0.106	−0.121
	(0.044)	(0.047)
Violence w injury, victim	0.197	0.209
	(0.051)	(0.054)
Violence w/o injury, victim	0.200	0.208
	(0.042)	(0.043)
Criminal damage/public order, victim	0.094	0.092
	(0.047)	(0.048)
Crimes, victim	0.053	0.049
	(0.016)	(0.016)
Stalking crimes, victim	0.073	0.068
	(0.074)	(0.076)
Constant	−2.434	−2.453
	(0.036)	(0.037)
Observations	148,861	135,356

Table A.4: Confusion Matrices from Logistic Regressions and Random Forests Based on All Features, Alternative Cost Ratios

Panel (a)

5:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	10184	4102	0.882	0.287
Actual Yes	817	1100	0.118	0.426
Col Share	0.679	0.321	1	
Pred Error	0.074	0.789		0.304

AUC= 0.704

5:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	10258	4028	0.882	0.282
Actual Yes	800	1117	0.118	0.417
Col Share	0.682	0.318	1	
Pred Error	0.072	0.783		0.298

AUC=0.701

Panel (b)

15:1 cost ratio, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7606	6680	0.882	0.468
Actual Yes	479	1438	0.118	0.25
Col Share	0.499	0.501	1	
Pred Error	0.059	0.823		0.442

AUC=0.703

15:1 cost ratio, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7316	6970	0.882	0.488
Actual Yes	452	1465	0.118	0.236
Col Share	0.479	0.521	1	
Pred Error	0.058	0.826		0.458

AUC=0.692

Table A.5: Confusion Matrices for Models Trained without Observations for Which all DASH Items were Left Blank. Cost Ratio 10:1

Panel (a)

DASH Items, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7122	5792	0.886	0.449
Actual Yes	575	1083	0.114	0.347
Col Share	0.528	0.472	1	
Pred Error	0.075	0.842		0.437

AUC=0.639

DASH Items, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	6809	6105	0.886	0.473
Actual Yes	562	1096	0.114	0.339
Col Share	0.506	0.494	1	
Pred Error	0.076	0.848		0.458

AUC=0.618

Panel (b)

Criminal Histories, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	6965	5949	0.886	0.461
Actual Yes	523	1135	0.114	0.315
Col Share	0.514	0.486	1	
Pred Error	0.07	0.84		0.444

AUC=0.679

Criminal Histories, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7725	5189	0.886	0.402
Actual Yes	546	1112	0.114	0.329
Col Share	0.568	0.432	1	
Pred Error	0.066	0.824		0.394

AUC=0.662

Panel (c)

DASH Items and Criminal Histories, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7818	5096	0.886	0.395
Actual Yes	528	1130	0.114	0.318
Col Share	0.573	0.427	1	
Pred Error	0.063	0.819		0.386

AUC=0.701

DASH Items and Criminal Histories, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7794	5120	0.886	0.396
Actual Yes	527	1131	0.114	0.318
Col Share	0.571	0.429	1	
Pred Error	0.063	0.819		0.388

AUC=0.695

NOTE: Test set size is 14,572. 1631 test set observations had missing values for all DASH items and were therefore excluded from the calculations.

Table A.6: Confusion Matrices for Models Trained without Observations Assessed as High-risk. Cost Ratio 10:1.

Panel (a)

DASH Items, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7676	6610	0.882	0.463
Actual Yes	637	1280	0.118	0.332
Col Share	0.513	0.487	1	
Pred Error	0.077	0.838		0.447

AUC=0.635

DASH Items, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7570	6716	0.882	0.47
Actual Yes	638	1279	0.118	0.333
Col Share	0.507	0.493	1	
Pred Error	0.078	0.84		0.454

AUC=0.611

Panel (b)

Criminal Histories, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8312	5974	0.882	0.418
Actual Yes	609	1308	0.118	0.318
Col Share	0.551	0.449	1	
Pred Error	0.068	0.82		0.406

AUC=0.684

Criminal Histories, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8254	6032	0.882	0.422
Actual Yes	578	1339	0.118	0.302
Col Share	0.545	0.455	1	
Pred Error	0.065	0.818		0.408

AUC=0.678

Panel (c)

DASH Items and Criminal Histories, Logit

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8438	5848	0.882	0.409
Actual Yes	566	1351	0.118	0.295
Col Share	0.556	0.444	1	
Pred Error	0.063	0.812		0.396

AUC=0.705

DASH Items and Criminal Histories, Random Forest

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8520	5766	0.882	0.404
Actual Yes	569	1348	0.118	0.297
Col Share	0.561	0.439	1	
Pred Error	0.063	0.811		0.391

AUC=0.693

NOTE: Test set size is 16,203.