



CEP Discussion Paper No 1676

February 2020

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

Jeffrey Grogger

Ria Ivandic

Tom Kirchmaier

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. A random forest based on the underlying risk assessment questionnaire does better under the assumption that negative prediction errors are more costly than positive prediction errors. A random forest based on two-year criminal histories does better still. Indeed, adding the protocol-based features to the criminal histories adds almost nothing 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.

Key words: domestic abuse, risk assessment, machine learning

JEL Codes: K42

This paper was produced as part of the Centre's Communities Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We thank Terence Chau and Sean Gupta 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 and Ian Wiggett for helpful comments. No financial support was received for this project.

Jeffrey Grogger, University of Chicago, IZA, NBER and Centre for Economic Performance, London School of Economics. Ria Ivandic, Centre for Economic Performance, London School of Economics. Tom Kirchmaier, Copenhagen Business School and Centre for Economic Performance, London School of Economics.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means without the prior permission in writing of the publisher nor be issued to the public or circulated in any form other than that in which it is published.

Requests for permission to reproduce any article or part of the Working Paper should be sent to the editor at the above address.

© J. Grogger, S. Ivandic and T. Kirchmaier, submitted 2020.

1 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 altercations ([Her Majesty's Inspectorate of Constabulary \(2014\)](#)).

Since the mid-1990s, an important part of the response to domestic abuse has been risk assessment ([Dutton and Kropp \(2000\)](#); [Campbell, Webster and Glass \(2009\)](#)). 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 Inspectorate 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, risk assessment is carried out using the DASH (Domestic Abuse, Stalking, and Harassment and Honor-Based Violence) risk identification and assessment model ([Robinson et al. \(2016\)](#)). DASH shares features with risk assessments in use elsewhere. Its first component is a 27-item protocol administered

to the victim by a police officer responding to a domestic abuse call. The same protocol is used by nearly all police forces nationwide, with only minor variations. Its second component is the officer's risk assessment, which amounts to a prediction of future risk. The officer classifies 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 officer's assessment 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 its use has resulted in striking differences across police forces in the share of domestic abuse calls assessed 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 assessment. We show that these predictions are only trivially more accurate than those based on the base recidivism rate, which make 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 assessment, 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 are only scant better than 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 model based on criminal histories leads to little if any 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 how machine-learning predictions based on criminal histories could be used much earlier in the process of responding to domestic abuse calls. Somewhat more speculatively,

we suggest that a two-stage procedure, of which our forecasting approach would compose the first stage, could improve both initial law enforcement response and the provision of protective resources.

2 Data

2.1 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. However, 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 DASH questions are most relevant for intimate partner incidents, we restrict attention to those calls¹.

The objective of our analysis is to forecast reported violent recidivism. We define violent recidivism at the level of the dyad, that is, the victim-perpetrator pair. We

¹Intimate partners include Ex Partners, Partners, Wives, Girlfriends, Ex Wives, Husbands, Boyfriends, Ex Husbands, and Civil Partners.

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 day². Of course, many DA incidents are not reported to police; we are unable to forecast those incidents.

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 either taken this approach or analyzed multiple definitions of recidivism. Here we focus on violent recidivism because DASH is supposed to predict, and help prevent, serious harm.

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 predictors based on two-year histories of DA calls and criminal records. The DASH data are available from July 2013 to July 2019. However, information linking victims and perpetrators is only available beginning in April 2012. To ensure that we have complete two-year histories for each dyad, we only include those dyads whose first call for service took place after April 2014. To ensure that we have a full year to measure violent recidivism for all DA calls, we only include calls that occurred before July 2018. Our resulting sample consists of 165,064 calls for service over the period April 2014 to July 2018³. [Figure 1](#) illustrates this timeline.

[Table 1](#) displays the frequency distribution of calls by dyad. The first column of the table shows that nearly 57.5 percent of the calls are singletons. Put differently, 42.5 percent of the dyads in our sample make more than one DA call to police. The

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

³We exclude calls for which there is no time stamp. We also exclude calls for which either the victim or the perpetrator is not recorded. If there are multiple crime reports for a call, we keep only the most serious one.

second column shows that the probability of a repeat call rises with the number of calls. It rises from 42.5 percent among dyads with at least one call, to 57 percent among dyads with at least two calls, to over 70 percent among dyads with at least five 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 rises with the number of calls as well. Although only 6.7 percent of first-time calls result in violent recidivism, that share rises to 10.7 at the second call, and continues to increase thereafter. Nevertheless, because so many dyads are involved in a single call, 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.

2.2 Responses from DASH questionnaires

As mentioned above, responding officers are instructed to complete a DASH form for each DA call. Answers come from the victim⁴. 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 assesses the case as standard, medium, or high risk. As indicated above, an assessment of high risk implies that an incident causing serious harm could take place at any time and triggers resources designed to promote the safety of the victim.

Table 2 displays the questions and tabulates their responses. For each question, there were three possible responses: yes, no, and omitted. We treat omitted as a

⁴The 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.

third level of the factor for each of these questions. 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 a question-by-question basis, the share of omitted responses runs from about 12 to 20 percent.

Panel B reports the distribution of officer risk assessments. 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.

2.3 Criminal history variables

Table 3 lists the criminal history variables that we use to predict violent domestic recidivism. One set pertains to the victim, another to the perpetrator, and another to the dyad. 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 0.834 crimes, compared to only 0.226 crimes for the average victim⁵. The perpetrator was also roughly three times more likely than the victim to have been involved in violence, either with or without injury.

⁵None of the victim or perpetrator offenses need involve the other party to the dyad.

3 Prediction models and results

3.1 Approach

Our main approach to prediction is to train a sequence of random forests in which the unit of observation is the DA call for service. The outcome is a dichotomous indicator of failure, that is, of violent recidivism during the following 12 months. We follow standard practice in the field of splitting our full sample into a training sample and a test sample. Since some dyads are involved in multiple incidents, we sample at the level of the dyad to ensure that the training and test samples are independent. 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.

Although a complete discussion of the random forest algorithm is beyond the scope of this paper, it is useful to provide a general description⁶. A random forest is constructed from a predetermined number of classification trees, where each tree is grown from a different random sample of the training set and a different random sample of predictors. Each tree classifies each observation as a predicted failure or predicted non-failure on the basis of its predictors.⁷

In random forests, one can use the observations that are not used to grow the tree, the so-called out-of-bag observations, to construct what is effectively an out-of-sample measure of predictive performance (Breiman (2001)). Like any other statistical procedure, machine learning models tend to do a better job predicting within the training sample than in independent samples. Since the purpose of a forecasting model is precisely to predict out-of-sample cases, it is important to evaluate its adequacy against data that were not used in training.

⁶Details can be found in Breiman (2001), Hastie, Tibshirani and Friedman (2009), Berk (2012).

⁷We use the *randomForest* package in R to train our models.

Although out-of-bag performance measures are usually adequate for this purpose, in this paper we go one step further, calculating such measures on our test sample which is independent of the training sample that was used to build the random forests. We take this additional step because our data contain multiple incidents for some dyads, and we were concerned that the usual out-of-bag performance measures might be overly optimistic as a result. In practice, the out-of-bag and test-sample measures were nearly the same.

We assess the adequacy of our models by means of confusion matrices, which compare predicted to actual outcomes within the test set. For each observation in the test sample, we obtained the predicted class (failure v. not failure) from each tree. We then obtained the prediction from the random forest as the majority class across all trees. Cross-classifying the actual class against the predicted class provides a useful measure of predictive performance.

3.2 Predictions based on the DASH risk assessment

Although our focus is on forecasts derived from machine learning methods, we begin by discussing predictions based on the DASH risk assessment. In the DASH protocol, an assessment of high risk is tantamount to a prediction of violent recidivism. 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 an DA incident involving violent recidivism within a year of a previous 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

assessed risk. 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, showing that 11.8 percent of the calls in the test sample resulted in failure. The third row of the table presents the column shares, showing that 8.5 percent of the calls in the test sample were classified as high-risk, and thus predicted to fail.

The final column of the table presents rates of classification error. It shows that 8.2 percent of the cases that did not result in violent recidivism were classified incorrectly. Equivalently, 91.8 percent were classified correctly. It also shows that 88.8 percent of the failures were classified incorrectly. The final row of the table presents rates of prediction error. The first column shows that 88.5 ($= 100 - 11.5$) percent of incidents that were predicted not to fail were predicted correctly. In contrast, 84.4 percent of the cases that were predicted to fail were predicted erroneously.

The final element in the table is the overall error rate. It is a weighted sum of the prediction errors, where the weights are given by the column shares. The overall error rate is 17.7 percent. Put equivalently, the accuracy rate of the forecasts based on the DASH risk assessment is 82.3 ($= 100 - 17.7$) percent.

Table 4 makes several important points. The first is that it is difficult to predict a rare outcome in a manner that beats the univariate, or 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 an 11.8 percent prediction error. Equivalently, it would be 88.2 percent accurate, beating the DASH risk assessment. At the same time, the simple Bayes approach would be unacceptable, since every failure would amount to a false negative.

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 fail does not. Of the two, negative prediction errors are probably more dangerous, or more costly. They represent cases where the perpetrator was classified as safe, but violently recidivated nonetheless. The cost of a positive prediction error may involve protectionary 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 measures.

This leads to our third point: when negative prediction errors are more costly than positive prediction errors, reducing negative prediction error should take precedence over increasing accuracy. The negative prediction error is the number of false negatives divided by the number of cases predicted not to fail. The negative prediction error of the forecast based on the DASH assessment is 11.5 percent. This is only trivially better than the 11.8 percent negative prediction error 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 that is only inconsequentially lower, than a forecast that requires no more data than the base failure rate. Fortunately, the approach we take below lets us reduce the negative prediction error substantially.

3.3 Predictions based on random forests using DASH items as predictors

Our first step is to ask whether the DASH data themselves can be used to form better forecasts. For this step we train a random forest using the data in the

training set. The outcome is violent recidivism. The predictors include all of the DASH questions, a variable that sums the number of DASH questions answered affirmatively (called "Totalyes"), and the three-level risk assessment. The output of the random forest is a predicted class for each case. We summarize the results from four random forests in the confusion matrices that appear in Table 5. These matrices apply the random forests 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).

A key step in training the random forests is to specify the relative cost of false negative and false positive errors, a point that has been made forcefully by Berk and his co-authors ([Berk, He and Sorenson \(2005\)](#), [Berk \(2008\)](#), [Berk \(2012\)](#) and [Berk and Bleich \(2013\)](#)). By default, the random forest algorithm will tend to maximize accuracy. When the failure is a statistically rare event, as in our case, it does this largely by minimizing the classification error within the majority class (i.e., the non-failures). This is evident in the confusion matrix in Panel A of Table 5. The classification error rate for the majority class is essentially zero. Even though the classification error rate for the failures is 0.998, the overall error rate is only 11.9 percent. Because the model predicts almost no failures, the negative prediction error is likewise 11.8 percent. Essentially, without specifying the relative cost of different types of errors, the random forest in this case performs no better than the simple Bayes classifier.

The remaining panels of the table present confusion matrices that were generated under different assumptions about the cost of false negative relative to false positive prediction errors. Panel B reports results for a 5:1 cost ratio, whereas Panels C and D report results for cost ratios of 10:1 and 15:1, respectively. In presenting these results, we do not mean to take a stand on what the correct cost ratio is; that would

require inputs on the relative costs of possible outcomes in each case⁸. Our point here is to illustrate how and to what extent different cost ratios affect the results.

We generate different relative cost ratios by down-sampling the majority class. That is, we make use of all the failures from the training sample, then sample that share of non-failures which balances the costs of error to most nearly achieve the desired cost ratio.⁹ Because of the inherent randomness of the random forest procedure, the desired cost ratio cannot be achieved exactly, but for the most part the approach approximates it reasonably well. To illustrate, in Panel B the ratio of positive to negative prediction errors is $4920/890 = 5.53$. In panels C and D, the achieved cost ratios are 10.52 and 13.79.

Raising the relative cost of negative prediction errors raises 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. At the same time, the negative error rate falls. At the 5:1 cost ratio, the rate of negative prediction error is 8.7 percent. At 10:1 it is 7.7 percent and at 15:1 it is 7.3 percent. The absolute number of false negative cases falls from 1702 based on the DASH risk assessment to 633 based on the random forest with a 10:1 cost ratio.

Using the DASH items to train a random forest can generate fewer negative prediction errors than those based on the DASH risk assessment alone. If negative prediction errors are more costly than positive prediction errors, as seems reasonable in this case, reducing false negatives is an important objective. In the next section we consider the performance of predictions based on criminal histories of the perpetrator and victim, rather than the DASH items.

⁸Fortunately, only relative costs matter. It is not necessary to specify the absolute costs of either false negative or false positive errors. See Berk (2012).

⁹Operationally, we used the *sampsiz*e option of the *randomForest* package to carry out the down-sampling. Finding the share of non-failures that achieves the desired cost ratio is generally an iterative exercise.

3.4 Predictions based on random forests using criminal histories as predictors

This exercise follows the approach above, except that we substitute the criminal history features for the DASH items in training our random forest models. Table 6 presents confusion matrices for the test sample. We again consider four cost ratios: the default, 5:1, 10:1, and 15:1.

Comparing Tables 5 and 6, we see that the models trained on the criminal histories outperform the models trained on the DASH items at each cost ratio. Although the improvements at the default cost ratio are negligible, those at higher cost ratios seem meaningful. At 10:1, the negative prediction error falls from 7.7 to 6.5 percent, leading to 50 fewer false negative predictions. At 15:1, the number falls by 72.

3.5 Predictions based on random forests using criminal histories and DASH items

Finally, we ask whether we can do better still by training models that use both the criminal histories and the DASH items as predictors. As above, we train the models on the training sample and present results based on the test sample. Table 7 reports the resulting confusion matrices.

At the default cost ratio, the model based on all the features performs the same in terms of negative prediction error as that based on only the criminal history features. This is the result of including what are essentially noise features at the training stage, that is, features that have little predictive power over and above those that are already included. At other cost ratios, the combined models perform slightly better. However, in all cases, the performance difference between the two

sets of models is negligible. To a close approximation, adding the DASH features to the criminal history features fails to improve the performance of the prediction models.

3.6 Predictor importance

The above results indicate that the machine learning approach to risk assessment outperforms the traditional approach based on the DASH risk assessment measure. It also shows that random forests based on criminal histories generally do better than those based on the DASH items. At the same time, the random forests are something of a black box. To shed some light on them, we construct importance measures of the predictors in the various models.

A typical measure of the importance of a feature is obtained by comparing the model's accuracy when that feature is used in forecasting to the model's accuracy when that feature's influence is removed from the forecast. To remove the feature's influence, the feature is randomly shuffled across observations, and predictions are obtained from the shuffled data. Importance is defined as the difference between the accuracy of the model based on the actual feature and the accuracy of the model based on the shuffled feature. Given our emphasis on negative prediction error, we modify the usual shuffle importance measure. Rather than reporting how shuffling the feature affects the model's accuracy, we report how it affects the model's rate of negative prediction error.

These modified shuffle importance measures are depicted in Figures 2 to 4. Figure 2 depicts importance plots for the model based on the DASH items; Figure 3 for the model based on criminal histories; and Figure 4 for the combined model. For illustrative purposes we present importance figures only for the models trained at a cost ratio of 10:1. In each panel, we restrict attention to the 20 most important

features.

Perhaps the most noteworthy feature of the Figures is that none of the features individually contributes very much to the performance of the model. The random forest is an ensemble forecasting procedure which is particularly useful in picking up important interactions among a large set of features. As a result, individual features in themselves often count for relatively little of the model's performance.

At the cost ratio 10:1, the rate of negative prediction error for the model based on the DASH items was 0.077. The most important individual predictor in the model is the indicator for whether the abuser is reported to be in trouble with the police; shuffling its values raises the prediction error by 0.009. The next two most important features are the risk assessment and the indicator for conflicts over child contact. These features each have an importance of about 0.007. After these features, importance falls off fairly quickly.

In the model trained on the criminal histories, the most important feature is the number of previous crimes in which the victim has been involved during the previous two years. This feature has lower importance than the most important feature in the DASH model, stressing the ensemble nature of the model. Some of the features have negative importance, suggesting that they predict worse than random numbers. Despite the low individual importance values, the criminal history features as a group forecast better than the DASH features.

Finally, Panel C reports modified importance measures for the features from the combined model. The most important feature is the indicator for conflict over child contact. The next most important is the number of previous crimes in which the victim has been involved. As with the criminal history model in panel B, none of the features has high individual importance.

3.7 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 8 presents confusion matrices for the models trained to this smaller data set. The DASH and the combined model have a slightly lower rate of negative prediction error than the corresponding models trained to the full sample (see Panel C in Tables 5-7). However, in all three cases, the change in the negative prediction error is very small.

In the second exercise, we dropped from the sample all the cases that received a DASH assessment of high risk. The reason is this. A high-risk assessment is supposed to trigger a protective response. If that response 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 instead results in a non-failure. This means that our failure rate for these cases is lower than it would have been, albeit by an unknown amount. Removing such cases removes the influence of any protective intervention and allows us to gauge the extent to which such interventions affect the performance of our models.

Confusion matrices are presented in Appendix Table 9. Panel A shows that the model trained on this sample using only the DASH items has a slightly lower rate of negative prediction error than the corresponding model trained to the full sample, whereas the other models have unchanged rates.

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.

4 Discussion and Conclusions

One important conclusion from our work is that a machine learning approach to risk assessment can perform better than a conventional approach based on an assessment protocol. In some sense, this might have been expected. The DASH questionnaire involves 27 questions. With such a large number of features to keep track of, one might expect salience effects and other cognitive biases to play a particularly important role. In contrast, under the assumptions that negative prediction errors are more costly than positive prediction errors, the machine learning models process the same information in a manner that substantially improves forecasting performance.

A second important finding is that machine learning methods applied to criminal history information provide better forecasts than the same methods applied to the DASH data. Apparently, the longer-run behavioral tendencies reflected in the

criminal histories are better predictors than details about the immediate incident, which are a primary focus of the DASH instrument. Whatever the reason, this finding could have important implications for the handling of domestic abuse calls.

To see this, it is necessary to understand how domestic abuse calls are handled. An incoming call is first handled by a call handler, who asks questions and assigns the incident a priority score. The urgency of the call is determined by the priority score; officers are supposed to respond to the highest-priority calls in a matter of minutes, whereas lower-priority calls may take longer. Only after an officer responds are the DASH data collected and the risk assessment made.

Whereas the DASH data are available only after an officer has appeared on the scene, electronic criminal history information is potentially available as soon as the call comes in and the call handler identifies the parties involved. This means that an initial prediction of violent recidivism could be made while the caller is on the line. Indeed, it could be used to set the priority score of the call, which we suspect would improve the allocation of calls across priority categories. In an exercise not reported above, we added the actual priority score to the criminal history variables, and found that it did not meaningfully improve predictions. Using predictions from a random forest model based on criminal history features would better distinguish cases at risk of violent recidivism. Of course, the main value of such an approach would be to identify calls that were not in need of urgent response. Nevertheless, the call handler would be left to prioritize a smaller number of calls based on other information obtained from the caller.

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 sensitive 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 sensitive, testing.

In the risk assessment setting, the first screen would be made by the random forest applied to the criminal history information. The second screening would be applied to the cases that were predicted to fail, which includes a large number of false positives. The idea for the second screen would be to develop an instrument with greater sensitivity for distinguishing true from false positives.¹⁰ Although the DASH items were not useful in this role, other instruments exist which may better distinguish the highest- from lower-risk cases ([Dutton and Kropp \(2000\)](#); [Campbell, Webster and Mahoney \(2005\)](#); [Messing and Thaller \(2013\)](#))¹¹. We suspect that such a two-part procedure would do better than the DASH risk assessment both in prioritizing calls for service and in providing protective resources to victims with the greatest need for them.

¹⁰[Berk, Sorenson and Barnes \(2016\)](#) propose a similar idea in a pre-trial detention setting.

¹¹In 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

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

- 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, 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* .
- 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

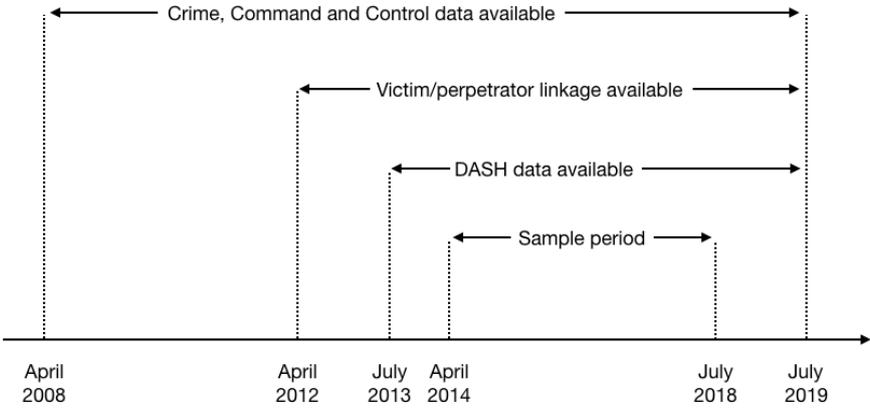
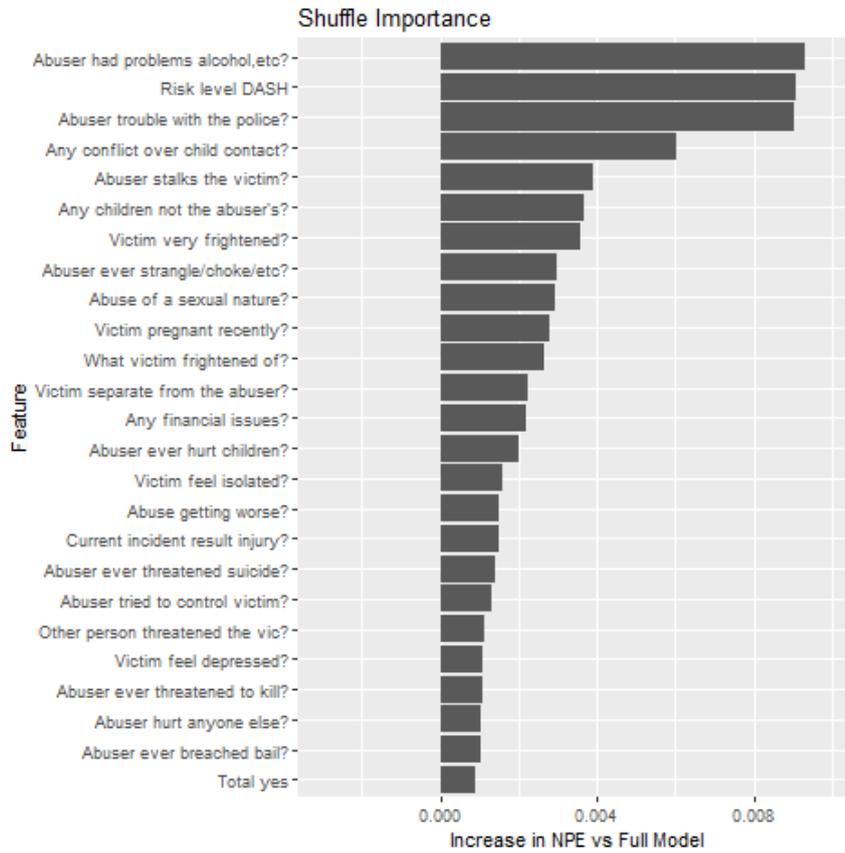
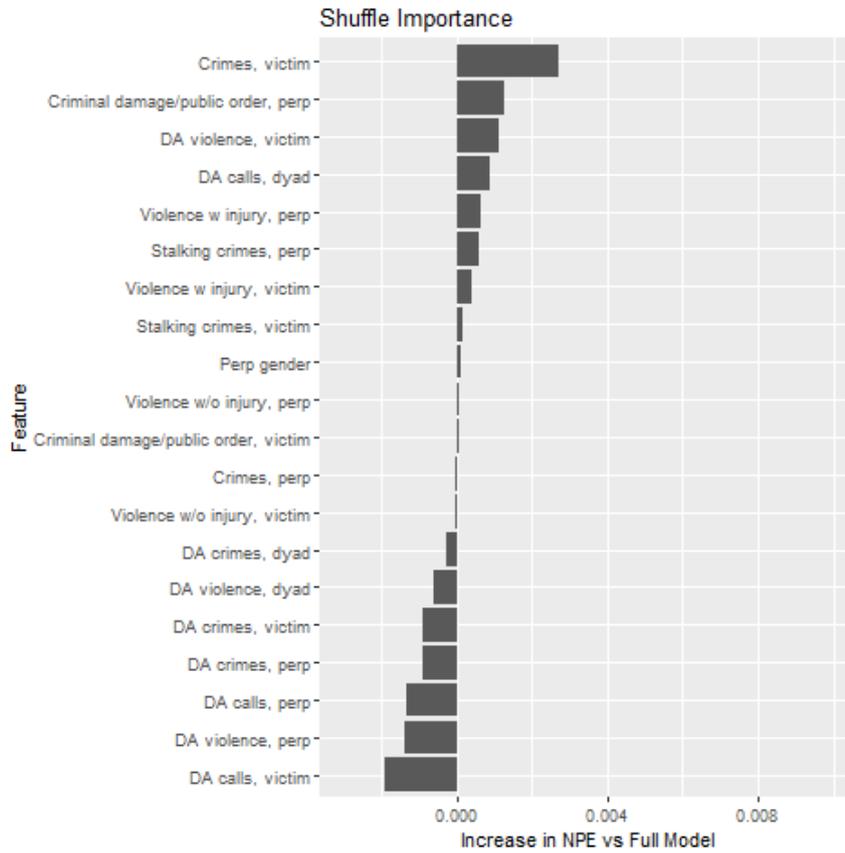


Figure 2: Modified predictor importance for random forests based on 10:1 cost ratio: DASH items



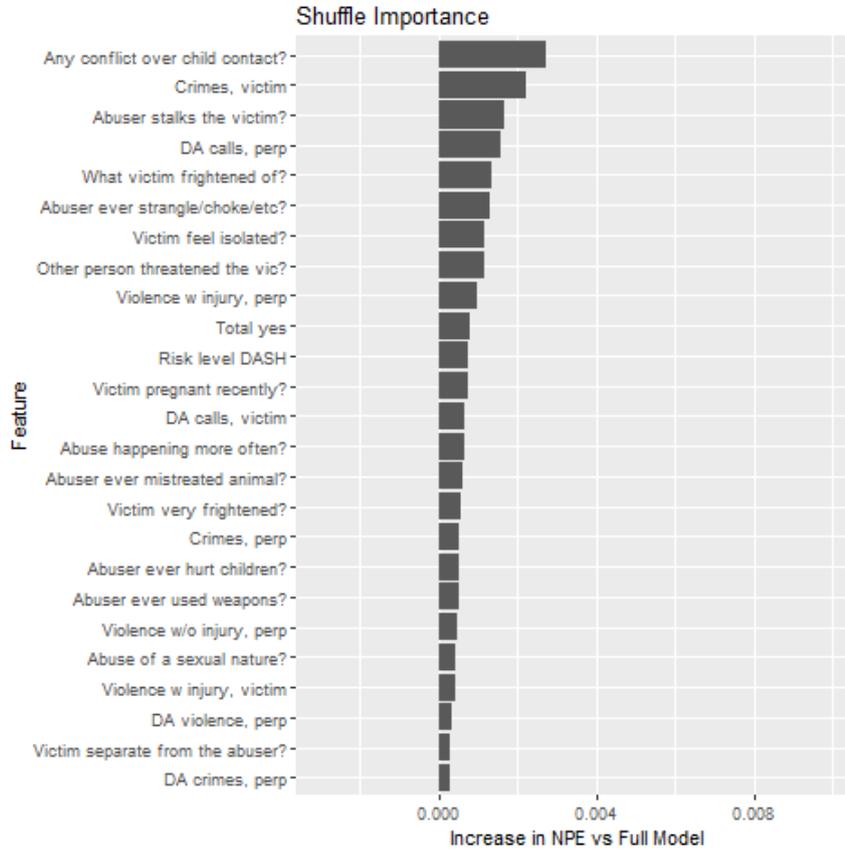
Note: NPE stands for negative prediction error; vic for victim; perp for perpetrator.

Figure 3: Modified predictor importance for random forests based on 10:1 cost ratio: Criminal Histories



Note: NPE stands for negative prediction error; vic for victim; perp for perpetrator.

Figure 4: Modified predictor importance for random forests based on 10:1 cost ratio: DASH items and Criminal Histories



Note: NPE stands for negative prediction error; vic for victim; perp for perpetrator.

Table 1: Frequency distribution of calls for service for domestic abuse, by dyad, probability of repeat call, and probability of violent recidivism

Call number	Relative frequency	Probability of repeat call	Probability of violent recidivism
1	0.575	0.425	0.067
2	0.181	0.574	0.107
3	0.087	0.643	0.134
4	0.050	0.683	0.154
5	0.032	0.700	0.163
6	0.021	0.726	0.181
7	0.014	0.736	0.204
8+	0.040	0.798	0.288
Overall		0.558	0.118

Note: Sample size is 165,064.

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	Classification Error
Actual No	13121	1165	0.882	0.082
Actual Yes	1702	215	0.118	0.888
Column Share	0.915	0.085	1	
Prediction Error	0.115	0.844		0.177

Note: Test set size is 16,203. Lesser risk includes standard and medium risk.

Table 5: Confusion matrices from random forests based on DASH, by cost ratio

(a) Default

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	14279	7	0.882	0
Actual Yes	1914	3	0.118	0.998
Column Share	0.999	0.001	1	
Prediction Error	0.118	0.7		0.119

(b) 5:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	9366	4920	0.882	0.344
Actual Yes	890	1027	0.118	0.464
Column Share	0.633	0.367	1	
Prediction Error	0.087	0.827		0.359

(c) 10:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	7623	6663	0.882	0.466
Actual Yes	633	1284	0.118	0.33
Column Share	0.51	0.49	1	
Prediction Error	0.077	0.838		0.45

(d) 15:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	6838	7448	0.882	0.521
Actual Yes	540	1377	0.118	0.282
Column Share	0.455	0.545	1	
Prediction Error	0.073	0.844		0.493

Note: Test set size is 16,203.

Table 6: Confusion matrices from random forests based on criminal history variables, by cost ratio

(a) Default

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	14179	107	0.882	0.007
Actual Yes	1819	98	0.118	0.949
Column Share	0.987	0.013	1	
Prediction Error	0.114	0.522		0.119

(b) 5:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	10130	4156	0.882	0.291
Actual Yes	836	1081	0.118	0.436
Column Share	0.677	0.323	1	
Prediction Error	0.076	0.794		0.308

(c) 10:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	8321	5965	0.882	0.418
Actual Yes	583	1334	0.118	0.304
Column Share	0.55	0.45	1	
Prediction Error	0.065	0.817		0.404

(d) 15:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	7251	7035	0.882	0.492
Actual Yes	468	1449	0.118	0.244
Column Share	0.476	0.524	1	
Prediction Error	0.061	0.829		0.463

Note: Test set size is 16,203.

Table 7: Confusion matrices from random forests based on criminal histories and DASH questions and risk assessment, by cost ratio

(a) Default

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	14201	85	0.882	0.006
Actual Yes	1823	94	0.118	0.951
Column Share	0.989	0.011	1	
Prediction Error	0.114	0.475		0.118

(b) 5:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	10282	4004	0.882	0.28
Actual Yes	806	1111	0.118	0.42
Column Share	0.684	0.316	1	
Prediction Error	0.073	0.783		0.297

(c) 10:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	8521	5765	0.882	0.404
Actual Yes	575	1342	0.118	0.3
Column Share	0.561	0.439	1	
Prediction Error	0.063	0.811		0.391

(d) 15:1 Cost

	Predicted No	Predicted Yes	Row Share	Classification Error
Actual No	7438	6848	0.882	0.479
Actual Yes	453	1464	0.118	0.236
Column Share	0.487	0.513	1	
Prediction Error	0.057	0.824		0.451

Note: Test set size is 16,203.

Appendix

Table 8: Confusion matrices for model trained without observations for which all DASH items were left blank. Cost ratio 10:1.

(a) DASH Items

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	6983	5931	0.886	0.459
Actual Yes	566	1092	0.114	0.341
Col Share	0.518	0.482	1	
Pred Error	0.075	0.845		0.446

(b) Criminal Histories

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7723	5191	0.886	0.402
Actual Yes	548	1110	0.114	0.331
Col Share	0.568	0.432	1	
Pred Error	0.066	0.824		0.394

(c) DASH Items and Criminal Histories

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7821	5093	0.886	0.394
Actual Yes	520	1138	0.114	0.314
Col Share	0.572	0.428	1	
Pred Error	0.062	0.817		0.385

Note: Test set size is 14,572. 1,631 test set observations had missing values for all DASH items and were therefore excluded from the calculations.

Table 9: Confusion matrices for models trained without observations assessed as high-risk. Cost ratio 10:1.

(a) DASH Items

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	7421	6865	0.882	0.481
Actual Yes	600	1317	0.118	0.313
Col Share	0.495	0.505	1	
Pred Error	0.075	0.839		0.461

(b) Criminal Histories

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8225	6061	0.882	0.424
Actual Yes	574	1343	0.118	0.299
Col Share	0.543	0.457	1	
Pred Error	0.065	0.819		0.409

(c) DASH Items and Criminal Histories

	Predicted No	Predicted Yes	Row Share	Class Error
Actual No	8462	5824	0.882	0.408
Actual Yes	570	1347	0.118	0.297
Col Share	0.557	0.443	1	
Pred Error	0.063	0.812		0.395

Note: Test set size is 16,203.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1675	Mary Amiti Stephen J. Redding David E. Weinstein	Who's Paying for the U.S. Tariffs? A Longer-Term Perspective
1674	Cheng Chen Claudia Steinwender	Import Competition, Heterogeneous Preferences of Managers and Productivity
1673	Stefano Bolatto Alireza Naghavi Gianmarco Ottaviano Katja Zajc Kejzar	Intellectual Property and the Organization of the Global Value Chain
1672	Timothy Besley Isabel Roland John Van Reenen	The Aggregate Consequences of Default Risk: Evidence from Firm-Level Data
1671	Jan-Emmanuel De Neve Daisy Fancourt Christian Krekel Richard Layard	A Local Community Course That Raises Mental Wellbeing and Pro-Sociality
1670	Tommaso Sonno	Globalization and Conflicts: The Good, the Bad and the Ugly of Corporations in Africa
1669	Michael Amior	Immigration, Local Crowd-Out and Undercoverage Bias
1668	Antoine Berthou John Jong-Hyun Chung Kalina Manova Charlotte Sandoz Dit Bragard	Trade, Productivity and (Mis)allocation
1667	Holger Breinlich Elsa Leromain Dennis Novy Thomas Sampson	Exchange Rates and Consumer Prices: Evidence from Brexit

1666	Fabrice Defever Michele Imbruno Richard Kneller	Trade Liberalization, Input Intermediaries and Firm Productivity: Evidence from China
1665	Philippe Aghion Antonin Bergeaud Richard Blundell Rachel Griffith	The Innovation Premium to Soft Skills in Low-Skilled Occupations
1664	Filip Gesiarz Jan-Emmanuel De Neve Tali Sharot	The Motivational Cost of Inequality: Pay Gaps Reduce the Willingness to Pursue Rewards
1663	Felix Koenig	Technical Change and Superstar Effects: Evidence From the Roll-Out of Television
1662	Enrico Moretti Claudia Steinwender John Van Reenen	The Intellectual Spoils of War? Defense R&D, Productivity and International Spillovers
1661	Decio Coviello Andrea Ichino Nicola Persico	Measuring the Gains from Labor Specialization
1660	Nicolás González-Pampillón	Spillover Effects from New Housing Supply
1659	Walter Bossert Andrew E. Clark Conchita D'Ambrosio Anthony Lepinteur	Economic Insecurity and the Rise of the Right
1658	Paul Frijters Andrew E. Clark Christian Krekel Richard Layard	A Happy Choice: Wellbeing as the Goal of Government

The Centre for Economic Performance Publications Unit

Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk

Website: <http://cep.lse.ac.uk> Twitter: @CEP_LSE