Predicting domestic abuse (fairly) and police risk assessment

Emily Turner, Gavin Brown\textsuperscript{1}, Juanjo Medina\textsuperscript{2}

\textsuperscript{1}School of Computer Science,
\textsuperscript{2}Department of Criminal Law and Crime Sciences, University of Seville

Published on: Feb 25, 2021
License: Creative Commons Attribution 4.0 International License (CC-BY 4.0)
Keywords: domestic abuse, risk assessment, machine learning, fairness, police

1 Abstract

Domestic abuse victim risk assessment is crucial for providing victims with the correct level of support. However, it has been shown that the approach currently taken by most UK police forces, the DASH risk assessment, is not identifying the most vulnerable victims. Instead, we propose a predictive model that incorporates information readily available in police databases, and census-area-level statistics. The model made significant improvement upon the predictive capacity of DASH, both for intimate partner violence (IPV) and other forms of domestic abuse (non-IPV). We show that the DASH questions contributed almost nothing to the predictive performance. We also provide an overview of model performance for ethnic and socioeconomic subgroups of the data sample.

2 Introduction

Risk assessment in domestic abuse cases has become increasingly relevant in criminal justice and public health practice. This is part of a broader trend towards risk assessment in criminal justice (Harcourt, 2006), no longer confined just within the penological context but also extending to sentencing and policing applications. A number of domestic abuse risk assessment tools have been developed for use in different settings, by different kinds of professionals, and for slightly different purposes (Messing and Thaller, 2013). Some take a structured professional judgement approach, whereas others (typically when deployed by front line personnel with limited training and more rushed contexts) rely on actuarial methods. Given the repetitive nature of domestic abuse, most of the tools and applications within a policing context have been oriented to identify victims that may exhibit a high risk of revictimization with a view to identifying protective actions that could be put in place to reduce this risk. Based on limited available evidence, it has been suggested by some that policing that is informed by this type of process may contribute to enhance victim safety (Messing et al. 2004).

Although research and practice in domestic abuse risk assessment first developed in North America, it has been in Europe where these practices have more quickly become national policy -although initiatives such as the Domestic Violence Homicide Prevention Initiative funded by NIJ is exploring the value of these measures. In 2004, the European Parliament called on Member States to implement measures against gender violence, including the development of ‘adequate risk assessments’. This was followed by the EU Victims’ Right Directive (2012/29) encouraging Member States to guarantee that ‘victims receive a timely and individual assessment, in accordance with national procedures, to identify specific protection needs’ (article 22). At present several European countries have national systems or sets of policies that require police officers to carry out some form of risk assessment in domestic abuse cases. This has led to an increased research focus on issues around the implementation...
of such systems (Robinson et al 2016) but also aiming to evaluate the quality of the classifications that result from using these risk assessment tools at large scale in the European context (Lopez-Ossorio et al. 2016; Svalin, 2018; Turner et al., 2019).

With the exception of ODARA, that was first developed in Canada, most risk assessment tools used by the police rely on short questionnaires or inventories that are completed after interviewing primarily the victim and typically (and in theory) examining police records on criminal history. This is certainly the case in the European countries that have developed these policies. Swedish police introduced B-Safer in the 1990s, a short version of SARA not requiring clinical training; Spain developed a nationally centralized system for dynamic ongoing assessment in 2007 called VIOGEN; Portugal recently adapted the Spanish model; and police forces in the United Kingdom have been using a standardized tool called DASH since 2009.

This paper aims to examine whether there is room to improve the existing systems to identify high risk victims. We will do so by developing machine learning models that not only use the information gathered by these tools used by front line police officers, but also other administrative data the police have. Given the ongoing debate about the ethics of predictive policing applications we will also perform an assessment of whether the resulting models meet basic requirements of fairness.

3 Literature review

3.1 Domestic abuse risk assessment in the UK

In this paper we focus on discussing the British case. From 2009 onwards most police forces in the UK started to use DASH when assessing domestic abuse. The Home Office defines domestic abuse as 'Any incident or pattern of incidents of controlling, coercive or threatening behaviour, violence or abuse between those aged 16 or over who are or have been intimate partners or family members regardless of gender or sexuality.' The bulk of the cases dealt with by the police under this definition are intimate partner violence incidents, but there is also a non-trivial proportion of cases of 'child'-to-parent and sibling violence. For example, in the data we use, in 2016 24% of victims were not of the intimate partner category. Of these, 51% was child-to-parent violence, 21% was sibling violence, 13% was parent-to-child violence (the remaining 15% included in-laws, and members of the extended family). DASH is also used by many other victim support organisations. DASH aims to identify domestic abuse victims at high risk of serious harm (CAADA, 2012) for referral to MultiAgency Risk Assessment Conferences (MARAC) - multiagency panels producing coordinated action plans to increase victims' safety and to manage perpetrator behaviour. DASH is a structured professional judgment scale. Where a victim answers 'yes' to 14 or more questions, police are advised to classify the case as high risk (CAADA, 2012), but the final judgment calls upon the professional judgment of the officer.
Although it is widely recognised that the adoption of DASH was a pivotal moment for encouraging positive change towards the policing of domestic abuse in the UK, the implementation of DASH has not been entirely uncontroversial. There are known issues regarding the quality of data gathering by officers in what often are rushed circumstances and may be highly emotional contexts (Robinson et al. 2016). There have been valid criticisms of the wording, phrasing and design of the tool. Police managers under the pressure of reducing budgets have sometimes come to view the process as time consuming and costly in terms of human resources. And in some practitioners quarters the tool has been seen as unnecessary red tape. Those promoting evidence-based policing were unimpressed with the large scale deployment of a tool that had not been subject to proper evaluation.

Indeed, a recent study demonstrated the tool performance was very poor in identifying victims and offenders at high risk of, respectively, revictimization and reoffending (Turner et al. 2019). The study determined that there was very little information in each DASH question response that could help us understand whether another incident was likely to happen (and come to the attention of the police). A model to predict serious subsequent abuse using the DASH questions as independent variables made improvement upon police professional judgment, but was only weakly predictive and not useful in practice (Turner et al., 2019).

Police auditing authorities in a landmark report published in 2014 recognised some of the problems of these risk assessment practices (HMIC, 2014) and tasked the College of Policing, the professional body of people working for the police in England and Wales, to review the system. At present, the College is piloting a new tool to be used by front line officers. The tool aims to reduce the burden on officers and victims by reducing the number of items, rethinking what may be appropriate questions in the context in which these encounters take place, and rephrasing them to enhance measurement. The new tool also slightly shifts the focus (from prediction) in that it places a greater emphasis in identifying (and properly evaluating the severity of) past unreported abuse and controlling behaviour. In some ways it could be said this new tool is less about risk assessment and more about harm detection, though ultimately also classifies victims according to the priority that they should be given. The pilots evaluating this new tool are ongoing.

4. Using data driven policing ethically

The police already has a large volume of information that may provide additional information to make assessments about the level of risk a victim experiences without placing an additional burden on front line officers (Pease et al. 2014). Although the judgments about victim safety continue to be primarily dependent on victim DASH responses, the police often have many other pieces of information about the victim, perpetrator, and general circumstances that can provide an enriched understanding of a given incident. Inspired by the glitter of predictive policing and some research applications (Berk, 2005), some practitioners wonder whether we could develop better predictions applying machine
learning algorithms to these administrative data sets. Some organisations, however, are calling for caution regarding these type of developments given the ethical implications associated with these uses of machine learning (Kearns and Muir, 2019), rehearsing arguments also raised in the US (Fergusson, 2017) about potential for bias and over-policing of certain individuals and communities. Indeed, the UK Government has set up an independent advisory board, the Center for Data Ethics and Innovation, that has been tasked to conduct a review on algorithmic decision making in various fields, the first of which will be criminal justice.

It is only in recent years that the Machine Learning community as a whole has woken up to the social threats posed by predictive modelling. The first Fairness and Transparency in Machine Learning (FAT/ML) Workshop took place in 2014 at NIPS, the largest conference in Artificial Intelligence. The inaugural FAT Conference was in 2018. High profile reporting, such as the work by ProPublica (Angwin et al., 2016), has led to increased scrutiny in the public domain. There are now countless reports outlining the current state of affairs and providing guidance on the considerations involved in defining fairness, and the considerable challenges that remain [Berk et al., 2018], [Chouldechova and Roth, 2018], [Leslie, 2019], [Partnership on AI, 2019].

The dangers posed to human rights including liberty, privacy, freedom of expression [Couchman, 2019; Leslie, 2019; Wachter et al., 2018] are numerous. On the other hand, the potential to improve outcomes such as the identification of high-risk victims of crime means that dismissing these methods out of hand is morally questionable at best. As more voices join the debate, from academia, government, industry, and other bodies, the only consensus there appears to be is that we are not yet anywhere near a satisfactory solution to ensuring fairness in an algorithmic society.

At the point of predictive modelling, there are several types of fairness that can be protected. Group or statistical fairness compares subgroups by one or more metrics of model performance and requires that there be group-wise parity in this respect. However, it is not possible to achieve parity across all metrics of group fairness at the same time so that a trade-off is required [Chouldechova, 2017; Kleinberg et al., 2017], and the overall accuracy of the model will likely also degrade. A more fundamental issue is that this type of fairness only protects the ‘average’ members of the groups and does not provide meaningful fairness guarantees for individuals [Couchman, 2019; Chouldechova, 2018]. Alternatively, there is the notion of individual fairness [Dwork et al., 2012], which, in theory, treats similar individuals similarly. However, in practice, this requires a way of measuring the similarity between two individuals, and obtaining this is fraught with difficulty [Friedler et al., 2016], [Chouldechova and Roth, 2018]. Both group and individual fairness ensure fairness within the world as it is represented by the data. A third form is counterfactual fairness ([Kusner et al., 2017]), which can capture externalities to the data set, such as the social biases that impact what data is available. Even the decision about which of these three approaches to take entails a value judgment about what sort of
fairness is more or less important. And given considerable variance in the costs of these approaches, it is also a judgment about what it is worth.

The purpose of this study is to evaluate the potential for identifying victims at high risk of harm at the hand of their abuser, and no such analysis would be complete without a consideration of whether this goal can be achieved within the limits of what is deemed to be fair and ethical. Throughout this study we draw attention to the most fundamental and challenging issues that were encountered in our work, and provide a cursory overview of model performance on subgroups of the data. However, a full exploration of fairness is beyond the scope of this paper, and in any case it is not for us to decide what fairness should look like in this context and how it should be achieved.

3 Data

3.1 Data pipeline

The data was provided by a large metropolitan police force in the UK. Our data sharing agreement and University Ethics Committee approval prevent us from identifying the force. Suffice to say it is a large force responsible for a diverse metropolitan area and that it is not unusual in terms of police auditing authorities' evaluations of the quality of services it provides (HMIC PEEL assessments).

Between 2011 and 2016, the police force responded to approximately 350,000 domestic violence incidents. Of this number, we only examine those with complete data in key fields: abuser and victim identifiers, victim-to-abuser relationship type (intimate partner or other), and data linking fields that permitted identification of whether or not there were charges associated with an incident. We only retained cases where officer risk grading had been specified. In this respect, there was complete data for 84% of the incidents. Each row of the data set corresponds to a unique abuser-victim pair (dyad) and incident.

Missing data in any other fields was handled, but the above mentioned fields were considered integral to our analysis. There are many unknowns around the missingness mechanism for each of these fields. Was it random data corruption? Or were officers less attentive to the details if an incident didn't seem serious enough? Was the relationship descriptor left out because the perpetrator and victim disagreed about the relationship status? Based on answers to questions like these, we could decide whether and how to treat these fields, statistically, in the model build and at the point of prediction. But this was beyond the scope of the paper so that we merely can point it out as a caveat to our findings.

There was one primary victim per incident, but some incidents also listed one or more secondary victims. We focus on the primary victim at the index incident because they would have provided the answers to the DASH questionnaire, which is victim-focused, and these questions form part of the predictor variable set.
A small proportion of dyads (1%) were recorded as being involved in more than one incident in a day. We did not know the time at which incidents occurred, so that the order in which incidents occurred on a single day could not be determined. It is possible that some of these were duplicated records. Thus, where this occurred, only one incident was kept and the rest were excluded. We also excluded the tiny proportion of cases (0.01%) where either victim or abuser were dead or too ill.

We exclude incidents occurring in 2011 and 2012 so that we could create predictor variables that represented two years of domestic abuse history. The outcome is defined to capture subsequent incidents happening up to one year after the index incident, so 2016 data is also excluded. This further reduced the number of incidents by 46%.

Of the remaining abuser-victim pairs, 37% were involved in more than one incident. Where this occurred, we randomly selected one incident from the several that they are involved in, to represent the index incident. In this way, we created a data set that is representative of the variety of incidents that the police encounter on a daily basis: they may have been meeting the abuser and victim for the first time, or they may have already dealt with the pair several times in the past. This approach also allows an evaluation of the importance of domestic abuse history for predicting subsequent incidents. By using only one event per dyad, we ensured that the assumption of independence of observations was preserved, a requirement for logistic regression modelling.

Finally, the dataset was split into IPV and non-IPV cases. The IPV data set included current/ex spouse and partner, girlfriend, and boyfriend relationship types. It was formed of ~60,000 unique dyads. The non-IPV data set contained ~24,000 unique dyads, less than half the number of IPV dyads for that same time period.

3.2 Variables

The predictor variables are mostly drawn from police and census data. The 27 DASH questions are included, alongside the officer’s DASH risk grading, and 62 additional variables, fitting into the broad categories of domestic abuse history, personal demographics, geographical demographics, and history of crime and victimisation. They are outlined below. We refer to the 58 additional variables as the augmented data set.

- **DASH (27 + 1 variables):** The 27 DASH questions are answered by the victim when an officer is called to a domestic abuse incident. The answers are 'yes'/‘no’/not known. Based on these responses, a risk grading of 'standard'/‘medium’/high’ is assigned to the case. A grading of 'high' indicates belief that the victim is at risk of serious harm and it may happen at any time; ‘medium’ predicts that serious harm is unlikely unless circumstances change for the victim or perpetrator; and ‘standard’ predicts
that there is no evidence indicating the likelihood of serious harm. We include the 27 questions and officer risk grading in the predictor variable set.

- **Additional Index Incident Descriptors (12 variables):** These include binary indicators for whether injuries were sustained, or alcohol or drugs involved at the index incident. This information is typically gathered by the officers when filling out an incident report and independently of any risk assessment process. There are variables to indicate whether there were crime charges related to the index incident. A variable was also created to represent the event where the victim responded ‘no’ or ‘not known’ to question 27 regarding the criminal history of the abuser even though police records show that the abuser did have a criminal history. A similar variable identified when the victim responded ‘yes’ to question 27 even though there was no history on police records. Note that there may not have been a record because the crime occurred outside of the metropolitan area, so care is needed when interpreting this variable.

- **Domestic abuse history (20 variables):** Repeat victimisation is higher for crimes of domestic violence [Ybarra and Lohr, 2002], so that it was important to create variables representing prior involvement. For both abuser and victim, we created a count of the number of times each has been victimised, or abused, and also include a count of prior incidents for a given dyad. There were two variables to represent the number of days since the abuser was last abused and days since last victimisation.

  This category also includes history of charges made in the context of domestic abuse. For each crime, we could identify the perpetrator and victim. We counted the number of times an abuser has perpetrated a crime against the victim in the last year and ten years, and the average harm score over that period. For incidents preceding 2011 it is not possible to tell whether these were committed in the context of domestic abuse. However, for dyad crime involvement occurring within the time frame for which we also have domestic abuse incident information, 92% of these were made in the context of a domestic abuse incident for which a corresponding DASH form was available.

- **Personal demographics (5 variables):** These variables covered basic biographical details of victim and abuser which were included in the incident reports. They were age of abuser, age difference between abuser and victim, and gender. A more fine grained description of victim relationship to abuser was also applied (current and ex partner for IPV; and sibling-sibling, parent-child, child-parent, and other for non-IPV).

- **Geographical demographics (4 variables):** These variables are small area statistics at the level of LSOA (lower-layer super output area), a common UK census geography that describe areas with an average population of 1,500. The Index of Multiple Deprivation is a relative measure of deprivation across England that is based on seven domains: income deprivation, employment deprivation,
education, skills, and training deprivation, health deprivation and disability, crime, barriers to housing and services, and living environment deprivation (Smith et al., 2015). Three additional variables were workday population density from the 2011 census, average property prices, and count of domestic abuse incidents in the last two years.

• **Criminal and victimisation history (28 variables):** By far the most important category of variables, in terms of risk prediction in this data set, is the history of crime charges and victimisations. This category covers the criminal and victimisation history of both the perpetrator and victim. There are the counts of charges for all crimes and serious harm crimes, and mean crime severity score in the year and ten years preceding the index incident. To measure harm we use the Office for National Statistics Crime Severity Score. This is a score that tries to measure the harm of each offence pairing this evaluation with the typical sentence given to that category of offence (for more details see Ashby 2018). Serious harm was defined as any crime in the violence against the person or sexual offences category with a score greater than or equal to 184 on the (ONS score). This is the score for “assault with injury”. The mean severity score is also based on the ONS score. Also the number of days since the first and most recent offences in the last ten years. We created analogous variables covering prior victimisations. We found that taking a longer view of criminal history was more informative than a narrower time window of, say, two or five years.

3.3 Revictimisation Outcome

Our analysis aimed to evaluate the extent to which victims at high risk of serious domestic abuse incidents could be identified using predictive modelling. We created several definitions of future harm, but focused on serious harm revictimisation for most of the discussion in the paper. This is what officers are predicting when they grade a case as high risk with the DASH tool. Officers are guided to classify a case as high risk if there is identifiable serious harm risk, the event could happen at any time, and the impact would be serious [CAADA, 2012]. DASH is victim-focused so that officers are encouraged to predict revictimisation instead of recidivism, where the perpetrator is involved in another domestic violence offence with the same or a different victim. If the primary victim of the index incident was a primary or secondary victim at the subsequent incident, we defined this as revictimisation. We defined serious harm as any violence against the person or sexual offences crime with an ONS score greater than or equal to 184 (this is the harm score for assault with injury). The event can happen ‘at any time’, which we interpreted as ‘any time up to 365 days after the index event’. This represented the ‘ground truth’, where ground truth is defined as that which we observe in the data rather than infer from a predictive model. By this definition, the prevalence of serious harm revictimisation was 3.6% for IPV and 1.1% for non-IPV victims. Note that if we instead consider revictimisation as any new incident within the year of the index, regardless of whether there was a
The charge associated with it, the domestic abuse prevalence rises considerably, to 22.5% and 11.5% for IPV and non-IPV respectively.

Apart from predicting whether a victim is at risk of further victimisation, police forces are also interested in perpetrator recidivism, regardless of whether it is with the same victim or not. New tools such as the Priority Perpetrator Identification Tool for domestic abuse (Robinson and Clancy, 2015) aim to help practitioners to identify such perpetrators. Thus we also evaluated whether we could predict recidivism. This was called the recidivism outcome, and the prevalence was 4.7% for IPV, and 4.4% for non-IPV incidents. Were we to define recidivism as any subsequent incident within the year, the prevalence would be 27.4% for IPV and 25.8% for non-IPV.

4 Methods

There were small numbers of missing data in the data sets. Postcodes were missing for 5.9% of the file and, where this occurred, geographical demographics could not be identified. Also, there were small proportions (<6%) of missing data for the age and gender of the victim and abuser. We applied multivariate imputation by chained equations (MICE, [Van Buuren et al., 1999]) to impute missing values for these fields.

We compared predictive algorithms based on six different statistical and machine learning models: Logistic Regression, Naive Bayes, Tree-Augmented Naive Bayes, Random Forests, Gradient Boosting, and Weighted Subspace Random Forests. The less usual Weighted Subspace Random Forests [Xu et al., 2012] was included because it often outperforms Random Forests when there are many unimportant variables present in the predictor set. In a single tree of a Random Forest, the best predictor for the next split in a node is chosen from a randomly selected subset of predictor variables. Each variable has an equal chance of being selected into the subset. The weighted subspace approach assigns varying probabilities for subset selection to each variable, based on the strength of the relationship between variable and outcome. Thus, a variable with only a weak relation to the outcome is less likely to be selected for a given subset.

Numeric variables were discretized before applying the Naive Bayes and Tree-Augmented Naive Bayes methods. Following [García et al., 2013], we compared two methods of discretization on both algorithms, FUSINTER [Zighed et al., 2003], and proportional discretization (PD; [Yang and Webb, 2009]). FUSINTER is supervised in that it takes the dependent variable into account when choosing cut-points, whereas PD is unsupervised. PD is a heuristic based on the idea that the more cut-points there are, the lower the risk that an instance is classified using an interval that contains a decision boundary. It is a trade-off between bias and variance.
Feature selection was required for Logistic Regression, Naive Bayes, and Tree-Augmented Naive Bayes. Logistic Regression was paired with Elastic Net [Zou, Hui, Hastie, 2005], an embedded feature selection method. Forward feature selection was applied for Naive Bayes and Tree-Augmented Naive Bayes.

It was desirable to compare variables in terms of their relative influence in the model. For this purpose we reported standardised coefficients on the logistic regression models. The method of standardization was to subtract the mean and divide by two standard deviations [Gelman, 2008]. Because many of the variables had units that were meaningful, for example, age in years, we also provided the odds-ratios related to the unstandardised variables.

As we are primarily concerned with the ability of a model to rank different individuals in order of risk, we evaluate models using ROC curves and area under the ROC curve (AUC). A ROC curve with an AUC of 1 indicates that the high risk cases have been perfectly separated from the rest. An AUC of 0.5 indicates that the model is no better than a classifier that randomly labels cases as high risk. We estimated the performance of each algorithm using cross-validation. The models were built on training data and evaluated on separate test data that was unseen by the model at the training stage. We report the mean and standard deviation of AUCs on the cross-validation results. We also provide the rate-wise mean ROC curve with 95% confidence intervals across cross-validation runs. The algorithm with the highest mean AUC was selected as the candidate best model. Where an algorithm had hyperparameters that required tuning, this was achieved with a further, nested set of cross-validation, where the best hyperparameters were again deemed to be those associated with the highest (nested) cross-validated AUC.

In this setting, the true positive rate represents the rate of revictimisation detection and false positive rate represents the rate of false alarms. The AUC represents the probability that the classifier will rank a randomly chosen revictimisation cases above a randomly chosen non-revictimisation case. We also make reference to the positive predictive value, which is the proportion of revictimisation predictions that were correct.

For a preliminary view of potential issues of unfairness that arose in the modelling process, we described model performance for two types of population subgroupings. These were based on officer-defined ethnicity (ODE) of the victim, and Index of Multiple Deprivation (IMD) ranking. ODE was 66% complete for both IPV and non-IPV. Note that as this was officer-defined, this is a source of measurement error. However, it was the only marker of race available for this study. IMD served as a proxy for social demographics. Model calibration is compared across subgroups, as are within-subgroup revictimisation rate, true and false positive rates, and positive predictive value. The intention of this part of the analysis was only to reveal disparities between groups where they exist.
We do not propose to resolve these issues here, because a definition of what is fair is elusive, context-dependent, so that it falls to policy-makers to decide what it should be.

5 Findings

5.1 Can we identify high-risk victims?

In short, yes we can. A classification model based on the augmented data set is far better than the DASH tool at identifying victims at highest risk of serious harm. Logistic regression with Elastic Net regularization was the best performing model, with an area under the curve (AUC) of 0.748 on the intimate partner violence (IPV) sample and 0.763 on the sample concerning domestic abuse in other relationship contexts (non-IPV). AUC measures how well a model ranks cases in order of risk. An AUC of 0.75 indicates that there is a 75% chance that a randomly selected victim who did go on to experience serious harm revictimisation, would have had a higher risk score than a victim who did not suffer revictimisation. When modeling criminal risk prediction, most “good” risk assessment instruments achieve AUCs in the range of 0.67 and 0.73 [Brennan et al., 2017; Jung et al., 2016; Messing et al. 2013], indicating that our models surpassed expected performance.

The various machine learning algorithms are compared in Table 1. The standard deviation of AUC on the non-IPV data was greater than that of the IPV data in all cases. This is in part due to the non-IPV data having far fewer observations relative to the IPV data. As we are primarily concerned with improving the process for identifying at-risk victims, and not with comparing different machine learning models, for the rest of the paper we focus exclusively on the best-performing model, logistic regression with Elastic Net.

5.2 Is the DASH form necessary?

The DASH form contributes almost nothing to the model. To establish this we rebuilt the model using the augmented data only, and excluded the DASH questions, DASH risk grading, and the two variables that represented disparities between victim-reported abusers’ criminal history (DASH question 27) and police records. We then compared the models based on each of these data sets in terms of ROC curve and AUC, provided in Figure 1.

Figure 1 consists of two main parts: part A displays ROC curves, and part B, boxplots. The boxplots pertain to model AUC (with the variation coming from cross-validation results). The boxplots are in pairs and within each pair, IPV is on the left, non-IPV on the right.

Figure 1A shows the ROC curves for the three scenarios. Note that there is so little difference between the predictive capacities of the full data set and the one that includes everything except DASH, that their ROC curves are almost completely overlapping. And so there is also negligible difference between
the sets of boxplots i) and ii), which concern the models built on the full data set and the data set including everything except DASH. The difference in mean AUC between these two boxplots is tiny (0.0001 on IPV data and 0.0007 on non-IPV data), indicating that DASH contributes almost nothing to the model. These results, and all results that are yet to be presented, are based on 500-times 2-fold cross-validation.

The boxplot pairs iii) and iv) of Figure 1B relate to models built on the limited predictor data set of DASH risk grading, and the data set of the 27 DASH questions, respectively. The poor performance of officer risk grading, (Fig. 1B iii), shows that officers are not able to identify high risk victims. And this is at least partly explained by the fact that they are not working with an effective tool, (Fig. 1B iv), which echoes previous findings, [Robinson et al., 2016], [Turner et al., 2019].

The remaining five pairs of boxplots, Figure B v) to iX), show the predictive capacity of several other subsets of the predictor data set. They indicate that data already sitting in police databases is much more effective for the purpose of risk prediction. The predictor subsets are: index incident descriptors that are not DASH (Fig. B v), domestic abuse history (Fig. B vi), demographics (personal and geographic) (Fig. B vii), history of crimes and victimisation (Fig. B viii), and history of crimes, victimisations, and domestic abuse (Fig. B iX). By far the most important feature subset criminal and victimisation history combined with domestic abuse history (Fig. 1B.iX).

In order to better understand the difference that a predictive model can make to those facing the highest risk of domestic abuse, we classify victims with the highest model predicted probabilities as high risk and compare with victims that were identified as high risk via structured professional judgment and the DASH form. We set the priors for high-risk prevalence in accordance with the proportion of cases that were ranked as high risk by officers. This was approximately 4.2% of IPV cases and 1.5% of non-IPV cases in each training data set (standard deviation in officer-high-risk prediction across training data was 0.0008 for IPV and 0.0007 for non-IPV). Cases in, approximately, the 95.9 percentile or above for IPV, and 98.5 percentile or above for non-IPV, were classified as high risk. In this way, the same number of cases were predicted as high risk by both the officers and the model, and thereby we could make a more direct comparison between officer and model performance. We compared two predictive models with officer performance: a model built on the full data set, and one that was built on the variable set that excluded all DASH variables.

We focus here on the predictive model that is based on the full feature set, but note that the results are near-identical for the feature set that excludes DASH, see Table 2. The predictive model correctly identified 5.2 and 8.2 times the number of high risk victims that the officers identified using DASH, for IPV and non-IPV respectively. Thus, although there is seemingly little difference in terms of overall accuracy between officer risk grading and logistic regression models, the improvements in true positive rate and positive predictive value are striking. A 1% increase in true positive rate amounts to 11
more IPV and 1 more non-IPV victim being correctly identified as high risk. There were approximately 30,000 IPV and 13,000 non-IPV victims in 2016. The model may have identified 166 (30,000 * 0.036 * (0.191 - 0.037)) more IPV and 14 (13,000 * 0.011 * (0.115 - 0.014)) more non-IPV cases than officers did that year using the DASH tool.

5.3 Model calibration

We focus on the model that includes all variables for the rest of the discussion, but the results are similar for the model that excludes DASH. Overall, the model is well-calibrated, see Figure 2. The mean predicted probability of revictimisation in the test data sets was very close to observed prevalence, and the majority of cases were reliably predicted to have less than 4% probability of revictimisation: 72.6% of IPV cases and 99.2% of non-IPV cases. However, in the higher regions of the predicted probabilities there is a lot less data, so that, although the mean prevalence is closely aligned with expected prevalence, we are less confident in predictions from individual models. This is captured in the increasing vertical spread of the box plots in the higher percentiles. Thus, predictions based on an ensemble of logistic regression models would be more reliable in this respect. Also note that only 1.5% of the IPV sample were predicted to have probability greater than 16 %, and only 0.2% had probability greater than 30%. In the non-IPV sample, only 0.2% had probability greater than 8% and 0.01% were predicted as serious harm revictimisation with a probability in excess of 30%. The main drivers of scores in excess of 30% were the criminal history of the perpetrator and, to a lesser extent, criminal involvement of the victim.

5.4 Variables selected

A total of 80 and 17 variables were selected in the IPV and non-IPV models respectively. Part of the reason why so many more variables were selected for IPV is statistical: the data set was larger and there was a higher rate of revictimisation. As some of these variables were far more influential than others, and due to considerations of space, we do not present all 97 variables in tabular format, but only those with a standardised coefficient of a magnitude in excess of 0.1, see Table 3. Note that Elastic Net logistic regression does not return p-values or confidence intervals on the coefficients. To improve linearity in the logit, all variables pertaining to crime counts and mean ONS scores were log-transformed. Thus, care is required when interpreting odds-ratios for these variables. Almost every one of the most influential variables were static, and variables concerning abuser criminal history are most dominant in count and influence.

Four variables are common to both IPV and non-IPV top predictors. These are highlighted in bold in Table 3. Domestic abuse history of the abuser, abuser criminal history, and victim history of victimisations (not in context of domestic abuse) are important in both. The top predictor in both IPV and non-IPV data sets is the time since the abuser’s last domestic abuse incident (where they were the
abuser and regardless of how serious the incident was). There is elevated risk if the less time has passed since the previous incident. Similarly, in both IPV and non-IPV, the less time that has passed since the abuser was involved in crime (not including crimes committed in the context of domestic abuse), the higher the risk of serious harm. The most influential LSOA-level variable for IPV is the LSOA-level count of domestic abuse incidents over the past 2 years. No LSOA variables were included in the non-IPV model.

Note that two DASH questions appear in the top IPV predictors list, question 9: ‘Are you currently pregnant or have you recently had a baby in the past 18 months?’, followed by question 7 ‘Is there conflict over child contact?’. If a victim answered ‘Yes’ to question 9, this indicated an increased risk, whereas a ‘Yes’ to question 7 predicted lower risk which can be attributed to greater third-party intervention in cases of child conflict.

We caution against over-interpretation of influential variables. Firstly, because logistic regression modelling cannot establish causal relationships between predictors and the outcome. It merely identifies correlations [Prins and Reich, 2017]. Secondly, if a variable is not in the model, this does not necessarily mean it is uncorrelated with the outcome. Many variables were excluded because they were redundant in the presence of other variables that were retained in the final model.

5.5 Is the model ‘fair’?

Although we could not answer in full this question here, we could explore and describe differences in how the model was working for protected groups, in terms of true positive rate (TPR), false positive rate (FPR), and positive predictive value (PPV), and accuracy. These are metrics of statistical fairness [Chouldechova and Roth, 2018] that provide the starting point for a conversation about fairness, and how it relates to victim risk prediction and the consequences of such predictions for both victim and perpetrator. Unlike in other criminal justice settings in which risk assessments are used, the focus here is not the offender and the consequence of high risk classification is not a liberty reducing measure (e.g., pretrial detention, etc). Instead a positive prediction means that a victim will receive additional safeguarding support, MARAC. As such, the consequence of a false positive prediction is that an unnecessary burden is placed on already resource-strapped services. The consequence of a false negative is much more serious, indicating that a victim that went on to endure serious harm at the hand of their abuser could not have received support which may have prevented that harm.

The subgroup analysis presented here is limited to two subgroupings: officer-defined ethnicity (ODE) of the victim, and the index of multiple deprivation (IMD) decile, representing social demographics. ODE was not included in the model as a predictor variable but IMD score was. The intersectionality of these factors was not explored. We applied the model with best-performance on the full data sets,
logistic regression with Elastic Net regularisation. Results are based on 500-times 2-fold cross-validation.

Note that of the victim ODEs that were known, there were very small counts of Black and Asian victims, which affected our ability to draw strong statistical conclusions in the group-wise comparisons. This is further exacerbated by the fact that serious harm revictimisation is relatively rare (3.6% for IPV and 1.1% for non-IPV). Any improvements made at the point of data collection, in terms of reducing the rate of unknown ODEs and accuracy of inputs, could provide more information with which to evaluate model performance on the smaller subgroups.

There were differences in revictimisation rate amongst subgroups, which could reflect real differences in patterns of domestic abuse, but could also be related to how different communities interact with the police (Awan et al., 2019, Jackson et al., 2012). In another sense, it is also possible that these differences were driven by officer perception of how serious an incident was. Where ODE was unknown, revictimisation prevalence was significantly lower. It is possible that the quality of data collection was related to the perceived seriousness of the case, so that an officer was less inclined to complete the ODE field if they did not think the case was serious. This is corroborated in the IPV data by the fact that officers perceived these cases to be of a lower risk level, see Table 4 where there is a lower prevalence of high risk cases for the set defined by unknown victim ODEs. Thus, if officer perception of risk varies by demographics, then the propensity to leave ODE descriptors blank is also affected by demographic, which impacts our measurement of revictimisation rate by ethnic subgroup.

The TPR, FPR, PPV indicate the quality of predictions that the model is making on each subgroup, an aspect of statistical fairness. However, because there were differences in risk prevalence, it was not possible to achieve equal TPR, FPR, and PPV across subgroups [Chouldechova, 2017, Corbett-Davies and Goel, 2018]. With this in mind, we present an analysis of group-wise disparities.

An assessment of model fairness must include a comparison with current procedures where this is possible. We achieved this by following the same steps as in Subsection 5.2 to create a high risk classification from the predicted probabilities output by the model. The prior belief about serious harm was set to match the overall prevalence of high risk gradings assigned by officers to incidents, approximately 4.2% and 1.5% for IPV and non-IPV respectively. By using a single threshold on model score (within IPV and non-IPV groups), cases are essentially treated the same, regardless of subgroup membership.

All groups experienced better predictions in terms of TPR and PPV when the predictive model is used instead of DASH, see Figure 3. As we are primarily concerned with identifying true positives, TPR and PPV are arguably the most important metrics to compare by. For a more direct group-wise comparison between the model and DASH, we can set the prior based on group-wise high risk prevalence. For
example, for Asians in the context of IPV, 5.6% of cases were classified as high risk. So, for this subgroup, the 5.6% with the highest model score are predicted as revictimisation. When the model and DASH were compared using these new group-wise classification thresholds (based on prevalence of high risk grading within each group), the model outperformed DASH risk across all metrics: TPR, FPR, and PPV, for every ODE and IMD subgroup. Thus there may be differences in model performance between the groups but if we consider prediction accuracy as a dimension of the fairness debate [Berk and Bleich, 2013], [Kleinberg et. al 2019], it may preferable to apply the model as opposed to remaining with the current procedure.

6 Discussion and Conclusion

The purpose of this study was to evaluate the potential of predictive modelling, when applied to data readily available to the police, to make improvements upon the approach that is currently deployed for assessing risk to victims of domestic abuse. We also provided an overview of disparities amongst ethnic and demographic subgroups, so that the nature and extent of differences in model performance between subgroups could be understood.

6.1 Model Accuracy

We have shown that the predictive model provides a clear advantage over structured professional judgment and the DASH questionnaire. The AUC for the best predictive model, logistic regression with Elastic Net regularisation, is 0.748 for IPV, and 0.763 for non-IPV. Compared to DASH, the model identifies 5.2 and 8.2 times the number of high risk victims that the officers identified using DASH, for IPV and non-IPV respectively. Of the victims that were correctly identified as high risk by the model, 40.9% IPV and 49.7% non-IPV cases were classified as standard risk by officers using DASH, and 43.7% IPV and 40.2% non-IPV cases were classified as medium risk with DASH.

The most influential variables in the model were in the categories criminal history and domestic abuse history. The DASH form and officer risk grading provided almost no benefit when it came to predicting the revictimisation outcome we had created. When all DASH variables were excluded from the model build, the AUC fell by only 0.0001 on IPV data and 0.0007 on non-IPV data. The model is well-calibrated. However, it is unable to predict revictimisation with great certainty. Only 0.2% of the IPV sample, and 0.01% of the non-IPV sample had a predicted probability of revictimisation that was in excess of 30%.

These results may look as surprising, insofar as the DASH questionnaire is composed of risk factors considered as relevant in past literature (victim pregnancy, use of guns, strangulation, etc.). We suspect the poor performance of these risk factors are linked to problems with the way and the context in which DASH is administered in most of British policing. In similar modeling we are
developing with data from the Spanish police, for example, victim pregnancy was in fact the most useful variable in the model (and the police questionnaire performed much better than DASH). This suggests that, when considering the development of risk assessment models that rely on police interviewing victims, as important as the selection of the risk factors, is the design of a system that ensure appropriate investment in police training, but also the development of protocols for ensuring that the questioning is done in conditions that are conducive to establish rapport with and generate trust on the victims.

Our comparison of officer predictions made using DASH, and predictive models, is not perfect. Whether an officer assigns a risk of high, medium, or standard determines the level of support a victim receives, which will, to some extent, influence whether or not another incident will occur. Among the false positives in this study (cases where a high DASH risk was assigned but no new serious harm incident came to the attention of the police), there must be a mix of genuinely mislabelled cases where no new incident would have occurred regardless of subsequent MARAC treatment, and accurately labelled cases where engaging with MARAC averted further abuse. And so it could be argued that we have underestimated the power of the DASH tool. However, we suggest that this occurred only to a minimal extent. In an earlier study on a much smaller data set from a different police force it was possible to adjust for a variety of post-call risk management actions including MARAC referrals (Peralta, 2015). The inclusion of these features did not improve model performance. Thus the effectiveness of MARAC is questionable, a finding which is echoed elsewhere (Svalin, 2018; Ward-Lasher et al. 2018). Furthermore, the DASH form performs poorly at identifying high risk victims, with 96.3% mislabelled as either standard or medium risk (false negative rate). The inaccuracy of DASH risk grading, in combination with the questionable effectiveness of the MARAC process, indicate that the effect of intervention on the outcome is minimal.

Although information on the risk management actions was not available in this study, we did have a descriptor of disposals and charges for each incident. We evaluated this information in a simplified setting. A single-variable model predicting the outcome with DASH risk grading was compared with a model consisting of two predictors, DASH risk grading and disposals / charges. This additional variable did not improve the model, and neither did the inclusion of an interaction term between the two predictors.

6.2 Model fairness

Statistical and machine learning models have been applied in ways that have a direct and serious impact on people's lives, and we are now reckoning with the consequences of this (Science and Technology Committee, 2018) [Angwin et al., 2016], [Partnership on AI, 2019], [Couchman, 2019]. Consequently, more voices are joining the debate about the legal, social, and ethical dimensions of predictive modelling. While our work is necessarily situated within this debate, we could not propose
to resolve these issues here. To even begin to grapple with these challenges would require defining what is 'fair', yet fairness as a concept is highly context dependent, involving considerations that are extrinsic to the modelling process.

Instead, we provided an overview of several metrics of statistical fairness [Chouldechova and Roth, 2018], that compare differences between subgroups based on ethnicity and demography. These methods are simple to apply, and it is feasible that such an approach could be used to monitor decision making at a high level. We also highlighted other considerations of fairness that pertain to the data, such as the validity of the outcome and predictor variables.

6.2.1 Model bias

Algorithmic decision-making, if well-understood and well-regulated, can provide an unprecedented opportunity for quantifying and handling discrimination that occurs, either intentionally or unintentionally, when those decisions are left to humans [Kleinberg et al., 2019]. We have shown that there were disparities between ethnic and demographic subgroups. However, we also demonstrated that everyone benefited from the increased accuracy of model-based predictions when compared with officer risk predictions. An important limitation of our analysis is that we did not explore the intersectionality of subgroup characteristics.

6.2.2 Model validity

A model’s validity depends upon its fidelity to the world it is purporting to represent. There will be validity issues if the data used in a model is not sufficiently representative of our world. Thus a crucial shortcoming in the analysis is that a record of domestic violence only exists in our data if the police were informed of that incident. Evidence from the Crime Survey of England and Wales 2015 indicates that 79% of victims did not report domestic abuse (ONS, 2018a).

It is also difficult to estimate the effect of coercive control on reporting of incidents. It may be that victims are not aware that they are being controlled for some time, but there is evidence that they are more likely in the long term to seek help than victims of physical abuse [Myhill, 2015]. The tendency to report domestic abuse, or crimes in general, most likely varies within different wealth and ethnic demographics (Awan et al., 2019; ONS, 2018b; Jackson et al., 2012). The true prevalence and nature of domestic abuse may also differ between different communities, further hindering our ability to understand what the observed outcome really represents in general, and how it may differ for different subgroups. And of course, the propensity of individual victims to report an incident is also shaped by a myriad of other, non-demographic, factors too (Kwak et al., 2019; Xie & Baumer, 2019). This all serves to highlight the complexities involved in interpreting and modelling heterogeneous data. A
redeeming factor of the study is that the outcome was defined to capture serious harm revictimisation, and serious incidents are more likely to be reported (Barrett et al., 2017; Smith, 2017).

The most important predictors in the model pertained to charge data, which is a proxy for the true variables of interest, concerning criminal history. There are serious concerns that use of such data is an inadequate approximation to the actual criminal involvement of a person [Couchman, 2019]. Charges represent how police responded to a crime, and this process is not without bias [Home Affairs Committee, 2009]. Statistics for England and Wales indicate that Black people were 9.5 times more likely than White people to be stopped and searched by police [ONS, 2019]. It may be considered fairer to use charge data for crimes that are known to be less subject to bias such as serious crimes instead of all crimes. This would involve a trade-off with accuracy. There are far fewer perpetrators with a history of serious charges and the predictors that counted all charges were stronger predictors in the model.

A further issue with criminal history is that it only covers the metropolitan area for which we had data, so that information was inevitably missing for an unknown number of individuals who were charged elsewhere between 2013 and 2016. Similarly, where it appeared that no revictimisation occurred, it may be that revictimisation did occur, but the persons concerned had moved outside of the area.

6.3 So what? What next for police risk assessment of domestic abuse victims

Our findings suggest that in the British context the use of administrative data subject to modelling can provide valuable information to support the decision making of officers that evaluate the risk faced by victims of domestic abuse. At the moment, this information seems of better quality than that obtained through the use of police-administered questionnaires. Past work has proposed that the limitations of the DASH questionnaire items for predictive purposes may be linked to measurement error. For example, it was shown that there are large divergences between victim-reporting of abuser criminal history and police records of that history [Turner et al., 2019]. If there is noise in the measurement (whether this is due to poor training or the situational variables present in police/victim encounters when responding to calls for services) predictions using this data will be poor. Unsurprisingly, in the British context, police authorities are piloting tools that simplify data capture and are aimed at minimising measurement error. These new tools are still in a piloting stage.

There may be a temptation to infer from this that we should just replace efforts to connect with victims to elicit valuable information with cheaper and faster mechanised processes that rely on this administrative data. In a context of diminished resources and pressures for police time this is a real temptation. However, by no means do we suggest that a predictive model can replace police decision making. What we think our work suggests is that whatever the value of risk assessment questionnaires may be, the decision-making by officers involved in assessing the risk for victims of
domestic abuse can be supplemented by implementing systems that use data sitting in police computers to develop useful (notwithstanding their limitations) predictions of victimization. But there is information that any model will fail to capture, so that officers must still be trained to identify critical signs of abuse. Whether information goes into a future iteration of a model or not, the police still need to be able to understand domestic abuse in its various forms, to gather the information in the first place. Therefore, investment in human capital is as necessary as ever. Particularly, since most risk assessment systems still leave open considerable room for officer discretion to disregard the model predictions. In contexts in which there is greater investment in training, development of careful protocols for gathering information, and design of interviewing contexts that are more conducive to the development of rapport police-interviews obtain information about risk factors that can be of use (Lopez-Ossorio et al. 2016).

Furthermore, despite its limitations and challenges, the use of data-driven approaches to inform criminal justice decision making is probably going to stay with us. Our findings suggest that there may be value on that. Thus, a predictive algorithm should form an essential part of the approach to tackling domestic abuse, if we are to allocate scarce police resources in the most efficient way to help more victims. Beyond making more accurate predictions, a model synthesizes information from many sources to produce results that are consistent across individuals, and which can be audited for inconsistencies between sensitive groups. Thus, decision-making can be inspected in a way that is not possible with human deliberations.

But findings from others in the field also suggest that for these approaches to provide real value we need to continue thinking and exploring how the use of the information provided by models can enhance the quality of human’s decision making. As some authors have suggested, the important driver of real-world effects will be how humans use the risk classification resulting from these algorithms and the research frontier is how we implement them in a way that get us closer to achieving our societal goals (Stevenson and Doleac, 2019).

It must also be recognised that the performance of these risk assessment tools is still rather limited. The predictive metrics in absolute terms suggest that many cases will continue to be misclassified regardless of how we assess them. In our view, this suggest we need to be moving toward systems like those implemented by Spanish police that require recontact with victims to re-evaluate the risk within time windows determined by the initial risk level classification (e.g., shorter for those initially classified as higher risk).

A serious limitation of our approach is that the outcome only captures physical violence, whereas we know that coercive control is the most harmful form of abuse in terms of serious harm risk [Sharp-Jeffs and Kelly, 2016], [Monckton-Smith et al., 2017]. It is also the type of domestic abuse that mostly goes unrecognised by frontline officers, (Robinson et al., 2017). Efforts are underway to equip officers
with the training and tools required to better identify these dynamics (Wine & Myhill, 2018). However, until this occurs, we simply do not have the data to analyse. Therefore, this model could only form a part of any risk assessment, and the full procedure must include efforts to identify cases of controlling or coercive behaviour. It is possible that we have captured some indicators of coercive control in variables relating to previous DAs - these represent a pattern that can be overlooked by frontline officers (Stark, 2012), [Myhill and Hohl, 2016].

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine Bias.

Awan, I., Brookes, M., Powell, M., and Stanwell, S. (2019). Understanding the public perception and satisfaction of a UK police constabulary. Police Practice and Research, 20(2):172–184.

Barrett, B.J., Peirone, A., Cheung, C.H., and Habibov, N. (2017) 'Pathways to Police Contact for Spousal Violence Survivors', Journal of Interpersonal Violence, 1--31.

Berk, R., Heidari, H., Jabbari, S., Kearns, M., and Roth, A. (2018). Fairness in Criminal Justice Risk Assessments: The State of the Art. Sociological Methods & Research.

Berk, R. A. and Bleich, J. (2013). Statistical Procedures for Forecasting Criminal Behaviour. Criminology and Public Policy, 12(3):513–544.

Brennan, T, Dietrich, W., and Oliver, W. (2017), 'Risk Assessment', Oxford Research Encyclopedia of Criminology, 1--38, DOI: 10.1081/E-EIA-120046557

Chouldekhova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. pages 1–17.

Chouldekhova, A. and Roth, A. (2018). The Frontiers of Fairness in Machine Learning. arXiv pre-print 1808.00023, pages 1–13.

Corbett-Davies, S. and Goel, S. (2018). The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. (Ec).

Couchman, H. (2019). Policing by Machine, 1--86. Liberty Human Rights, Technical report.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2012). Fairness Through Awareness. In Proceedings of the 3rd Innovations in Theoretical Science Conference, pages 214–226, New York, NY, USA. ACM.

Ferguson, A. G. (2017), The Rise of Big Data Policing. NYU University Press.
Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im)possibility of fairness. arXiv pre-print 1609.07236, pages 1–16.

Gama, J., Žliobaitė, I., Bifet, A., Pechenizkiy, M., and Bouchachia, A. (2014). A Survey on Concept Drift Adaptation. ACM Computing Surveys, 46(4):1:37.

García, S., Luengo, J., Sáez, J. A., López, V., and Herrera, F. (2013). A survey of discretization techniques: Taxonomy and empirical analysis in supervised learning. IEEE Transactions on Knowledge and Data Engineering, 25(4):734–750.

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. Statistics in medicine, 27:2865–2873.

Harcourt, B. E. (2006). Against prediction. Chicago University Press.

Home Affairs Committee (2009). The Macpherson Report - Ten Years On. Technical Report July, House of Commons.

Jackson, J., Bradford, B., Stanko, E. A. & Hohl, K. 2012. A focus on a special population: young males from Black and Minority Ethnic groups. Just Authority?: Trust in the Police in England and Wales 128- -136. Routledge

Jung, S. and Buro, K. (2016). ‘Appraising Risk for Intimate Partner Violence in a Police Context’, Criminal Justice and Behavior 44(2): 240--260

Kearns, I and Muir, R. 2019. Data driven policing and public value. The Police Foundation.

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. R. (2019). ‘Discrimination in age of algorithms’, SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3329669

Kuhn, M. and Johnson, K. (2013). Applied Predictive Modeling. Springer-Verlag New York.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual Fairness. NIPS.

Kwak, H., Dierfenfeldt, R., and McNeeley, S. (2019) `The code of the street and cooperation with the police: Do codes of violence, procedural injustice, and police ineffectiveness discourage reporting violent victimization to the police?’, Journal of Criminal Justice, 60: 25--34

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. Technical report, The Alan Turing Institute.
Medina Ariza, J. J., Robinson, A., and Myhill, A. (2016). Cheaper, faster, better: Expectations and achievements in police risk assessment of domestic abuse. Policing (Oxford), 10(4):341–350.

Messing J and Thaller, J. (2013) 'The Average Predictive Validity of Intimate Partner Violence Risk Assessment Instruments', Journal of Interpersonal Violence 28(7): 1537--1558.

Monckton-Smith, J., Szymanska, K., and Haile, S. (2017). Exploring the Relationship between Stalking and Homicide. Technical report, Suzy Lamplugh Trust, UK.

Myhill, A. (2015). Measuring Coercive Control: What Can We Learn From National Population Surveys? Violence Against Women, 21(3):355–375.

Office for National Statistics (2018a). Domestic abuse: findings from the Crime Survey for England and Wales: Year ending March 2017. (March 2017):1−19.

Office for National Statistics (2018b). Crime Survey for England and Wales, year ending March 2019. https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/confidence-in-the-local-police/latest (Last Accessed: 17 Sept 2019)

Office for National Statistics (2019). Stop and Search. Police powers and procedures England and Wales statistics. https://www.ethnicity-facts-figures.service.gov.uk/crime-justice-and-the-law/policing/stop-and-search/latest (Last Accessed: 17 Sept 2019)

Partnership on AI (2019). Report on Algorithmic Risk Assessment Tools in the U . S . Criminal Justice System. Technical report, Partnership on AI.

Pease, K., Bowen, E. and Dixon, L. (2014). DASHed on the rocks. http://www.policeprofessional.com/news.aspx?channel=0&keywords=dashed%20on%20the%20rocks (accessed 4th October 2018)

Peralta, D. (2015), Data Mining for the Prediction of Domestic Violence, masters dissertation, University of Manchester.

Prins, S. J and Reich, A. (2017). Can we avoid reductionism in risk reduction? Theoretical Criminology, page 136248061770794.

Robinson, A., Myhill, A., Wire, J., Roberts, J., and Tilley, N. (2016). Risk-led policing of domestic abuse and the DASH risk model, College of Policing.

Commons Select Committee: Science and Technology Committee, Algorithms in decision-making, 15 May 2018, HC 351 2017-2019.
Sharp-Jeffs, N. and Kelly, L. (2016). Domestic Homicide Review (DHR) Case Analysis. Technical Report June, Standing Together/London Metropolitan University, London.

Smith, V. (2017), `An Exploration into the Factors Shaping Victim Reporting of Partner Abuse to the Police', Manchester Review of Law, Crime and Ethics, 6: 95--120.

Stark, E. (2009). Coercive control: how men entrap women in personal life. Oxford University Press, Oxford.

Stark, E. (2012). Looking Beyond Domestic Violence: Policing Coercive Control. Journal of Police Crisis Negotiations, 12(2):199–217.

Stevenson, M., and Doleac, J. (2019). Algorithm Risk Assessment in the Hands of of Humans. Discussion Paper Series. IZA Institute of Labor Economics.

Turner, E., Medina, J., and Brown, G. (2019). ‘Dashing Hopes? The Predictive Accuracy of Domestic Abuse Risk Assessment By Police’, The British Journal of Criminology, Volume 59, Issue 5, September 2019, Pages 1013–1034

Van Buuren, S., Boshuizen, H. C., and Knook, D. L. (1999). Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine.

Veale, M., Van Kleek, M., and Binns, R. (2018). Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making. pages 1-14.

Wachter, S., Mittelstadt, B., & Russell, C. (2018), ‘Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR’, Harvard Journal of Law and Technology.

Ward-Lasher, A., Messing, J., Cimino, A. and Campbell, J. (2018), ‘The Association Between Homicide Risk and Intimate Partner Violence Arrest’, Policing, doi: 10.1093/police/pay004.

Wire, J & Myhill, A. (2018), ‘Piloting a new approach to domestic abuse frontline risk assessment’, College of Policing.

Xie, M. & Baumer, E. P., (2019), ‘Crime victims’ decisions to call the police: past research and new directions’, Annual Review of Criminology 2: 217--240.

Yang, Y. and Webb, G. I. (2009). Discretization for naive-Bayes learning: Managing discretization bias and variance. Machine Learning, 74(1):39–74.

Ybarra, L. M. and Lohr, S. L. (2002). Estimates of repeat victimization using the National Crime Victimization Survey.
Zighed, D. A., Rabaséda, S., and Rakotomalala, R. (2003). FUSINTER: A Method for Discretization of Continuous Attributes. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 06(03):307–326.

Zou, Hui, Hastie, T. (2005). Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society, 67(2):301–320.

| Model                                           | IPV        | Non-IPV    |
|-------------------------------------------------|------------|------------|
| Logistic Regression with Elastic Net            | 0.748 (0.004) | 0.763 (0.017) |
| Naive Bayes (with Proportional Discretization) | 0.726 (0.004) | 0.743 (0.023) |
| Augmented Naive Bayes (with FUSINTER discretization) | 0.728 (0.005) | 0.708 (0.032) |
| Random Forest                                   | 0.722 (0.002) | 0.712 (0.020) |
| Gradient Boosting                               | 0.743 (0.004) | 0.760 (0.019) |
| Weighted Subspace Random Forest                 | 0.718 (0.003) | 0.707 (0.017) |

Table 1. Mean and standard deviation on AUCs for all six statistical and machine learning models that were benchmarked for performance on both the intimate-partner, and non-intimate-partner violence (IPV and non-IPV) data sets.

| Metric              | All Vars | DASH Excluded | DASH risk  |
|---------------------|----------|---------------|------------|
|                     | IPV      |               |            |
| TPR                 | 0.191 (0.009) | 0.190 (0.009) | 0.037 (0.039) |
| FPR                 | 0.036 (0.001) | 0.036 (0.001) | 0.019 (0.020) |
| PPV                 | 0.165 (0.008) | 0.165 (0.008) | 0.033 (0.034) |
| Accuracy            | 0.936 (0.001) | 0.936 (0.001) | 0.947 (0.018) |

Non-IPV
Table 2: Comparing predictive capability of i) the full variable set, ii) all variables except DASH, and iii) officer's DASH risk grading in terms of true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and overall accuracy. Mean and standard deviation based on 500-times 2-fold cross-validation.

| IPV                                           | Coef (std) | Odds-ratio |
|-----------------------------------------------|------------|------------|
| **Abuser's time since last incid**            | -0.500     | 0.978      |
| Abuser gender: male                           | 0.397      | 1.614      |
| Relationship type: ex                         | -0.305     | 0.737      |
| Victim gender: male                           | -0.263     | 0.729      |
| Abuser's count of charges in last 10 yrs*     | 0.232      | 1.122      |
| Question 9 (recent pregnancy): yes            | 0.187      | 1.319      |
| **Dyad's count of incids in last 2 yrs**      | 0.163      | 1.050      |
| Victim's count of incids as abuser in last 2 yrs* | 0.153      | 1.176      |
| Question 7 (conflict over child contact): yes | -0.148     | 0.801      |
| LSOA domestic abuse count in last 2 yrs*      | 0.146      | 1.114      |
| **Abuser's time since most recent crime**     | -0.134     | 0.999      |
| Variable Name                                      | Distribution | Revictimisation Prevalence | High Risk Prevalence |
|---------------------------------------------------|--------------|----------------------------|----------------------|
| IPV Officer Defined                               |              |                            |                      |
| Ethnicity of Victim                               |              |                            |                      |
| Ethnicity of Victim | 1 | 2-3 | 4-6 |
|---------------------|---|-----|-----|
| Asian               | 0.049 | 0.035 | 0.056 |
| Black               | 0.033 | 0.036 | 0.05  |
| European            | 0.574 | 0.044 | 0.043 |
| Unknown             | 0.344 | 0.022 | 0.036 |

| Non-IPV Officer Defined Ethnicity of Victim | 1 | 2-3 | 4-6 |
|---------------------------------------------|---|-----|-----|
| Asian                                       | 0.063 | 0.01  | 0.033 |
| Black                                       | 0.028 | 0.009 | 0.012 |
| European                                    | 0.574 | 0.012 | 0.013 |
| Unknown                                     | 0.335 | 0.009 | 0.015 |

| IPV Index of Multiple Deprivation | 1 | 2-3 | 4-6 |
|----------------------------------|---|-----|-----|
| 1                                | 0.371 | 0.043 | 0.052 |
| 2-3                              | 0.324 | 0.037 | 0.039 |
| 4-6                              | 0.196 | 0.027 | 0.033 |
| 7-10                             | 0.109 | 0.026 | 0.029 |

| Non-IPV Index of Multiple Deprivation | 1 | 2-3 | 4-6 |
|---------------------------------------|---|-----|-----|
| 1                                     | 0.389 | 0.010 | 0.016 |
| 2-3                                   | 0.318 | 0.011 | 0.017 |
| 4-6                                   | 0.186 | 0.010 | 0.010 |
| 7-10                                  | 0.107 | 0.008 | 0.011 |

Table 4: Summary statistics for each sensitive group. Distribution represents the relative size of each group so that, say, for the intimate-partner violence (IPV) group, the values for distribution over ethnic...
groups sums to 1. The second and third columns compare high risk prevalence and serious harm revictimisation prevalence for the subgroups.

| CURRENT SITUATION |  |
|-------------------|---|
| 1. Has the current incident resulted in injury? |  |
| 2. Are you very frightened? |  |
| 3. What are you afraid of? Is it further injury or violence? |  |
| 4. Do you feel isolated from family/friends, i.e. does (name of abuser(s)... ) try to stop you from seeing friends/family/Dr or others? |  |
| 5. Are you feeling depressed or having suicidal thoughts? |  |
| 6. Have you separated or tried to separate from (name of abuser(s)... ) within the past year? |  |
| 7. Is there conflict over child contact? |  |
| 8. Does (…) constantly text, call, contact, follow, stalk or harass you? |  |

| CHILDREN/DEPENDENTS |  |
|---------------------|---|
| 9. Are you currently pregnant or have you recently had a baby in the past 18 months? |  |
| 10. Are there any children, step-children that aren’t (…) in the household? Or are there other dependants in the household (i.e. older relatives)? |  |
| 11. Has (…) ever hurt the children/dependants? |  |
| 12. Has (…) ever threatened to hurt or kill the children/dependants? |  |

| DOMESTIC VIOLENCE HISTORY |  |
|----------------------------|---|
| 13. Is the abuse happening more often? |  |
| 14. Is the abuse getting worse? |  |
| 15. Does (…) try to control everything you do and/or are they excessively jealous? |  |
| 16. Has (…) ever used weapons or objects to hurt you? |  |
### Appendix: The DASH form.

| Question                                                                 |
|-------------------------------------------------------------------------|
| 17. Has (...) ever threatened to kill you or someone else and you believed them? |
| 18. Has (...) ever attempted to strangle/choke/suffocate/drown you?      |
| 19. Does (...) do or say things of a sexual nature that makes you feel bad or physically hurt you or someone else? |
| 20. Is there any other person that has threatened you or that you are afraid of? |
| 21. Do you know if (...) has hurt anyone else?                           |
| 22. Has (...) ever mistreated an animal or the family pet?               |
| **ABUSER(S)**                                                          |
| 23. Are there any financial issues? For example, are you dependent on (...) for money/have they recently lost their job/other financial issues? |
| 24. Has (...) had problems in the past year with drugs (prescription or other), alcohol or mental health leading to problems in leading a normal life? |
| 25. Has (...) ever threatened or attempted suicide?                     |
| 26. Has (...) ever breached bail/an injunction and/or any agreement for when they can see you and/or the children? |
| 27. Do you know if (...) has ever been in trouble with the police or has a criminal history? |
Figure 1 title: ROC Curves and AUC boxplots

Figure 1: ROC curves and AUC boxplots comparing the full variable set with variable subsets in terms of revictimisation prediction capability. The boxplots are grouped in pairs, with IPV on the left and non-IPV on the right. Note that in Figure 1B, descriptors vi) and ix), domestic abuse is abbreviated to DA. Based on 500-times 2-fold cross-validation.
Figure 2: Model calibration on full variable set. Grouping by 2 percentile increments, so that, e.g.: the left-most boxplot represents all those with predicted revictimisation probability in the range (0, 2%]. Due to tiny count in the higher percentiles, Based on 500-times 2-fold cross-validation.
A. Victim ODE

B. IMD decile

Figure 3 title: Model Fairness
Figure 3: Performance metrics for, A. ethnic subgroups as measured by officer-defined ethnicity (ODE) of the victim, and B. socio-demographics as measured by the Index of Multiple Deprivation (IMD) decile. Metrics are identified on the tabs to the right of the plots. They are true positive rate (TPR), positive predictive value (PPV), and false positive rate (FPR).

**Footnotes**

1. One criticism of the DASH form is that there is a lot of noise in the victim responses (Robinson, 2016; Medina Ariza et al., 2016). Question 27 on the DASH form was the only one for which we could compare victim responses to data resembling the truth about abuser criminal involvement. In this way we could gauge the level of agreement between police records and what the victim said. A contradiction between the two will have different explanations for different victims, but one possibility is that it is an indication of coercive control. —