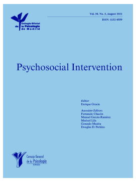
Colegio Oficial de la Psicología de Madrid

Psychosocial Intervention

https://journals.copmadrid.org/pi


Psychosocial Intervention

Predicting Domestic Abuse (Fairly) and Police Risk Assessment

Emily Turner[a], Gavin Brown[b], and Juanjo Medina-Ariza[b]

[a]*University of Manchester, UK;* [b]*University of Seville, Spain*

ARTICLE INFO

Received 28 December 2021
Accepted 7 May 2022
Available online 17 June 2022

Keywords:
Domestic abuse
Risk assessment
Machine learning
Algorithmic fairness
Police

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 Domestic Abuse, Stalking, and Honour Based Violence (DASH) risk assessment, is not identifying the most vulnerable victims. Instead, we tested several machine learning algorithms and propose a predictive model, using logistic regression with elastic net as the best performing, that incorporates information readily available in police databases, and census-area-level statistics. We used data from a large UK police force including 350,000 domestic abuse incidents. Our models made significant improvement upon the predictive capacity of DASH, both for intimate partner violence (IPV; AUC = .748) and other forms of domestic abuse (non-IPV; AUC = .763). The most influential variables in the model were of the categories criminal history and domestic abuse history, particularly time since the last incident. We show that the DASH questions contributed almost nothing to the predictive performance. We also provide an overview of model fairness performance for ethnic and socioeconomic subgroups of the data sample. Although there were disparities between ethnic and demographic subgroups, everyone benefited from the increased accuracy of model-based predictions when compared with officer risk predictions.

La predicción (equitativa) de la violencia doméstica y la evaluación policial de riesgo

Palabras clave:
Violencia doméstica
Evaluación de riesgos
Aprendizaje automático
Justicia algorítmica
Policía

RESUMEN

La evaluación de riesgo de las víctimas de abuso doméstico es crucial para poder ofrecerle a las mismas el nivel adecuado de asistencia. No obstante, se ha demostrado que el enfoque predominante en casi todas las fuerzas policiales británicas, que descansa en el uso de DASH (las iniciales en inglés del instrumento de evaluación de abuso doméstico, acoso y violencia por cuestión de honor), no sirve para identificar a las víctimas más vulnerables. En su lugar, este artículo evalúa varios algoritmos de aprendizaje automático y propone un modelo predictivo, usando como algoritmo con un mejor rendimiento una regresión logística con red elástica, que utiliza como fuente de información variables normalmente disponibles en los archivos policiales, así como en el censo de la población. Para desarrollar y evaluar este modelo usamos datos de un departamento policial responsable de un área metropolitana en el Reino Unido que incluía 350,000 incidentes de abuso doméstico. Nuestros modelos mejoran significativamente la capacidad predictiva de DASH, tanto para la violencia en la relación de pareja (AUC = .748) como para otras formas de abuso doméstico (AUC = .763). Las variables más influyentes en el modelo fueron medidas del historial delictivo y de violencia doméstica previa, en particular el tiempo transcurrido desde el último incidente. El artículo demuestra que el cuestionario DASH prácticamente no contribuye nada al rendimiento predictivo de nuestro modelo. El artículo también ofrece una evaluación del rendimiento en términos de equidad para distintos grupos étnicos y socioeconómicos en nuestra muestra. Aunque había disparidad entre estos subgrupos, todos ellos se beneficiaban de la mayor precisión predictiva resultante de usar nuestros modelos en lugar de las clasificaciones policiales basadas en DASH.

Psychosocial risk assessment in domestic abuse cases has become increasingly relevant in criminal justice. This is part of a broader trend towards risk assessment, no longer confined just within the penological context but also extending to sentencing and policing. Various domestic abuse risk assessment tools exist for use in different settings, by different professionals, and for different purposes (Messing & Thaller, 2013). Some take a structured professional judgement approach, whereas others (typically when deployed

Cite this article as: Turner, E., Brown, G., & Medina-Ariza, J. (2022). Predicting domestic abuse (fairly) and police risk assessment. *Psychosocial Intervention*, *31*(3), 145-157. https://doi.org/10.5093/pi2022a11

Funding: This paper was made possible thanks to a ESRC grant (ES/M01178X/1). Correspondence: jmedina11@us.es (J. Medina-Ariza).

by front line personnel with limited training) rely on actuarial methods. Given the repetitive nature of domestic abuse, these tools are oriented to identify people with a high risk of revictimization so that protective actions can be put in place to reduce this risk (Medina Ariza et al., 2016; Ybarra & Lohr, 2002).

Although research in domestic abuse risk assessment first developed in North America, it is in Europe where these practices have more quickly become national policy. 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 that "victims receive a timely and individual assessment, in accordance with national procedures, to identify specific protection needs" (article 22). Several European countries now require police officers to carry out risk assessment in domestic abuse cases. This has generated an increased research focus on issues around their implementation (Robinson et al., 2016) and their predictive quality (Lopez-Ossorio et al., 2016; Svalin, 2018; Turner et al., 2019).

Most police risk assessment tools rely on short questionnaires not requiring clinical training that are completed by officers after interviewing primarily the victim. Swedish police introduced B-Safer in the 1990s (Svalin, 2018); Spain developed a nationally centralized system (VIOGEN) for dynamic ongoing assessment in 2007 (Lopez-Ossorio et al., 2016); Portugal recently adapted the Spanish model (Machado et al., 2021); and police forces in the United Kingdom have been using a standardized tool called DASH since 2009 (Turner et al., 2019).

This paper used data from the UK to examine whether there is room to improve the existing systems to identify high risk victims. Specifically, can we improve the documented very low predictive performance of DASH using a machine learning method? Is it possible to do this relying on administrative data gathered by the police other than the information collected with the DASH questionnaires? What machine learning algorithm may be more appropriate for this application and data? Finally, and given the ongoing debate about the ethics of predictive policing applications, can we do this meeting basic requirements of fairness? We find that data outside these questionnaires significantly improve the predictions, with time elapsed from last incident being one of the more relevant predictors, and that basic requirements of fairness are met.

Literature Review

Domestic Abuse Risk Assessment in the UK

From 2009 onwards police in the UK has used DASH for assessing domestic abuse ("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", as defined by the Home Office). Most cases dealt with by the police under this definition are intimate partner violence incidents, but about a quarter of the cases relate to child to parent violence, sibling violence, and other forms of family violence. DASH aims to identify such defined domestic abuse victims at high risk of serious harm for referral to Multi Agency 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 used at the scene of a call. Where a victim answers 'yes' to 14 or more questions, police are advised to classify the case as high risk, but the final judgment calls upon the professional judgment of the officer.

Although the adoption of DASH was a pivotal moment for encouraging positive change towards the policing of domestic abuse in the UK, its implementation has been controversial. Some have been criticised the wording, phrasing, and design of the tool, and the quality of data gathering by officers – in what often are rushed circumstances and may be highly emotional contexts (Robinson et al., 2016). Police managers pressurized by reduced budgets have sometimes considered the process as time consuming and costly, whereas some officers see it as unnecessary red tape. Evidence-based policing scholars were unimpressed with the national adoption 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). It is also questionable whether a tool very much influenced by tools to predict intimate partner violence can be reliably used to predict other forms of domestic abuse.

Police auditing authorities in a landmark report published in 2014 recognised some of these problems (HMIC, 2014) and tasked the College of Policing, the professional police body in England and Wales, to review the system. The new tool aims to reduce the burden on officers and victims by reducing the number of items, rethinking what may be appropriate questions, and rephrasing them to enhance measurement quality. It is still in a pilot stage.

Using Data-driven Policing Ethically

The police already has a large volume of data 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. Aside from 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 risk. Inspired by the glitter of predictive policing practitioners we wonder whether we could develop better predictions applying machine learning methods to these administrative data sets (Grogger et al., 2020). Some, however, are calling for caution given the ethical implications of this (Kearns & Muir, 2019), which include the potential for bias and over-policing of certain individuals and communities (for a lengthy discussion see Ferguson, 2017).

The Machine Learning community as a whole has recently 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 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 public scrutiny. There are now countless reports outlining the current state of affairs and providing guidance on defining fairness, and the remaining considerable challenges (Berk et al., 2018; Chouldechova & Roth, 2018; Leslie, 2019; Partnership on Al, 2019; Veale et al. 2018). The dangers posed to human rights are various, particularly the amplification of institutional discrimination or the consolidation of lack of transparency on decision making (for details see Couchman, 2019; Leslie, 2019; Wachter et al., 2018), but 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 is that we are not yet anywhere near a satisfactory solution to ensuring fairness in an algorithmic society.

There are several types of fairness that can be protected in predictive modelling in order to deal with the potential of amplifying discrimination. '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., 2019), and the predictive accuracy of the model will likely also degrade. 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 '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 (Chouldechova & Roth, 2018; Friedler et al., 2016). Both group and individual fairness aim to ensure fairness 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 ('statistical fairness'). 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. However, a full exploration of fairness is beyond the scope of this paper.

Method

Data Pipeline

The data was provided by a large metropolitan police force in the UK and includes all people in the jurisdictional population for which there was an administrative (police) record of domestic abuse for a particular period of time. Our data sharing agreement and University Ethics Committee approval prevent us from identifying the police force providing the data. 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. This section describes how we constructed the analytical dataset (of around 84,000 index events) from the larger dataset of all incidents and the reasoning underpinning the construction of the analytical dataset. Of the original data, 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. Missing data in any other fields was handled, but the above mentioned fields were considered integral to our analysis.

In the police dataset, 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 (1%) of dyads (abuser-victim pairs) 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. As we created predictor variables out of two years of domestic abuse history, and also defined the outcome to capture subsequent incidents happening up to one year after the index incident (event we take as baseline for predicting revictimization), we only predict the outcome of incidents occurring between 2013 and 2015. Excluding incidents from 2011, 2012, and 2016 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 were 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 intimate partner violence (IPV) and non-IPV cases (most of which entail adolescent to parent violence). The UK protocols for risk assessment do not distinguish between various forms of abuse, yet it is reasonable to expect different risk factors will be relevant when predicting IPV and non-IPV events. 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.

Predictors

The predictor variables were mostly drawn from police and census data and were selected on the basis of what data was readily available 'and' has been previously identified by the literature as relevant risk factors. They are outlined below.

DASH (27 + 1 Variables)

The 27 DASH questions (see Appendix) 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, the officer assigns a risk grading. 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)

This information is gathered by officers when filling out an incident report independently of any risk assessment process, e.g., whether injuries were sustained, alcohol or drugs were involved. We also created variables to represent the event where the victim's answer to question 27 regarding the criminal history of the abuser contradicts police records on abuser criminal history.

Domestic Abuse History (20 Variables)

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 in 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, we could deduce that 92% of these were made in the context of a domestic abuse incident because there was a DASH form associated with the crime.

Personal Demographics (5 Variables)

These covered biographical details of victim and abuser (e.g., age of abuser, age difference between abuser and victim, gender, and victim relationship to abuser).

Geographical Demographics (4 Variables)

These variables are small area statistics at the level of LSOA (lower-layer super output area), a 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; employment; education, skills, and training; health and disability; crime; barriers to housing and services; and living environment deprivation. 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)

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 (ONS) 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. This is the score for 'assault with injury', but it would include other crimes of similar or greater severity. 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.

Revictimisation Outcome

We created several definitions of revictimisation, but for this paper report only on serious harm revictimisation (this is what officers are predicting when they grade a case as high risk with DASH) any time up to 365 days after the index event. Officers using DASH are guided by policy 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. This focus on harm has also been vindicated by recent work on policing domestic abuse (Sherman, 2018). DASH is victim-focused so that officers are encouraged to predict revictimization (rather than reoffending), thus, if the primary victim of the index incident was a primary or secondary victim at a new subsequent domestic abuse incident known to the police, we defined this as revictimization. Such defined (as any new incident within the year, regardless of whether there was a charge associated with it), the domestic abuse prevalence was 22.5% and 11.5% for IPV and non-IPV respectively. However, we are focusing our analysis here only on serious harm, by which we mean any violence against the person or sexual offences crime with an ONS severity score greater than or equal to 184. By this definition, the prevalence of 'serious harm revictimization' was 3.6% for IPV and 1.1% for non-IPV victims. This represented the 'ground truth', where ground truth is defined as that which we observe in the data, what we saw happened in terms of serious re-victimisation.

Analytical Procedures

There were small numbers of missing data. 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 (Van Buuren, 2018) to impute missing values for these fields.

We compared predictive algorithms based on six different machine learning models: logistic regression, naive Bayes, tree-augmented naive Bayes, random forests, gradient boosting, and weighted subspace random forests (for a gentle introduction on these see Kuhn & Johnson, 2013) as implemented in R, the programming language we used for all data cleaning and analysis. 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 (Yang & 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, 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.

Variable selection was required for logistic regression, naive Bayes, and tree-augmented naive Bayes. Logistic regression was paired with elastic net (Zou & 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 .50 indicates that the model is no better than random classification. 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 revictimisation cases above a 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.

Results

Can We Identify High-Risk Victims?

In short, yes we can. A classification model based on all the variables 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 .748 in the intimate partner violence (IPV) sample and .763 in 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 .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. Domestic abuse risk assessment instruments deemed 'good' in the literature achieve AUCs in the range of .67 and .73 (Jung et al., 2016; Messing & Thaller, 2013), indicating that our models surpassed expected performance.

Table 1. Comparison of Model Performance: Mean and Standard Deviation of 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

Model	IPV	Non-IPV
Logistic regression with elastic net	.748 (.004)	.763 (.017)
Naive Bayes (with proportional discretization)	.726 (.004)	.743 (.023)
Augmented Naive Bayes (with FUSINTER discretization)	.728 (.005)	.708 (.032)
Random forest	.722 (.002)	.712 (.020)
Gradient boosting	.743 (.004)	.760 (.019)
Weighted subspace random forest	.718 (.003)	.707 (.017)

The various algorithms are compared in Table 1. 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.

DASH vs. Our Model

The DASH form contributes almost nothing to the model. To establish this we rebuilt the model, excluding the DASH questions, DASH risk grading (based on the officer discretion), 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, Figure 1, and boxplots, Figure 2.

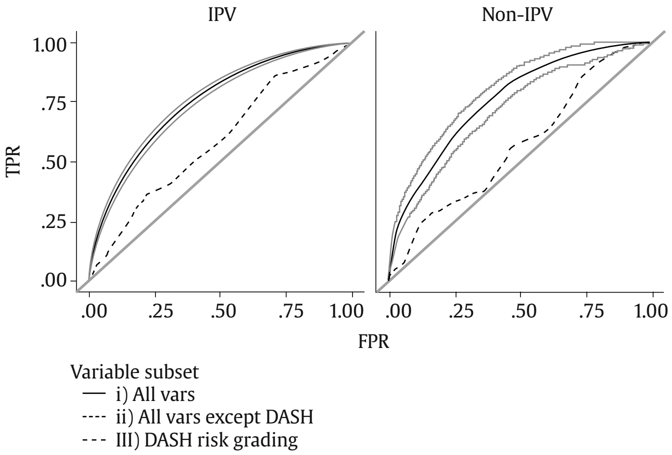


Figure 1. ROC Curves Comparing the Full Variable Set with Variable Subsets in terms of Revictimisation Prediction Capability.
Note. Based on 500-times 2-fold cross-validation. For the IPV group, the mean AUC on the ROC curve corresponding variable subsets i) and ii) is .748, with a standard deviation of .004; on variable subset iii) the AUC is .567 and standard deviation is .005. For the non-IPV group, the mean AUC on the ROC curve corresponding variable subsets i) and ii) is .763, with a standard deviation of .017; on variable subset iii) the AUC is .551 and standard deviation is .019.

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 1 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 in Figure 1. And thus the difference in mean AUC between these two boxplots is tiny (.0001 on IPV data and .0007 on non-IPV data) so that there is negligible difference between the sets of boxplots in Figure 2i and ii). We can deduce that DASH contributes almost nothing to the model.

In Section 3.2 we presented the predictor variables in sensible subgroupings. To understand the effectiveness of certain categories of predictors, we rebuilt the model on predictor subgroups, and the resultant model AUCs are set out in Figure 2iii) to ix). The boxplot pairs iii and iv of Figure 2 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 (Figure 2iii), 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, (Figure 2iv), which echoes previous findings (Robinson et al., 2016; Turner et al., 2019).

The remaining five pairs of boxplots, Figure 2v to ix, show the predictive capacity of the predictor variable subsets that were outlined in Section 3.2. 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 (Figure 2v), domestic abuse history (Figure 2vi), demographics (personal and geographic) (Figure 2vii), history of crimes and victimisation (Figure 2viii), and history of crimes, victimisations, and domestic abuse (Figure 2 ix). By far the most important variable subset is criminal and victimisation history combined with domestic abuse history (Figure 2ix). Note that all of these results are based on '500-times' 2-fold cross-validation.

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 variable set, but note that the results are near-identical for the variable 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.

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

Metric	All Vars	DASH Excluded	DASH risk
IPV			
TPR	.191 (.009)	.190 (.009)	.037 (.039)
FPR	.036 (.001)	.036 (.001)	.019 (.020)
PPV	.165 (.008)	.165 (.008)	.033 (.034)
Accuracy	.936 (.001)	.936 (.001)	.947 (.018)
Non-IPV			
TPR	.115 (.021)	.114 (.021)	.014 (.018)
FPR	.014 (.001)	.014 (.001)	.006 (.007)
PPV	.083 (.015)	.082 (.015)	.011 (.013)
Accuracy	.977 (.001)	.977 (.001)	.984 (.007)

Note. Mean and standard deviation based on 500-times 2-fold cross-validation.

Model Calibration

If a model is well-calibrated, it means that each level of predicted probability is representative of the underlying revictimisation rate. So for all the individuals for which we predict a revictimisation rate of 3%, the observed revicimisation rate is approximately 3% if the model is well calibrated. 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 3). 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 a revictimisation 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%.

What Predictors Were More Relevant?

Almost every one of the most influential variables were static, and variables concerning abuser criminal history are most dominant in count and influence. A total of 80 and 17 variables were selected in the IPV and non-IPV models respectively. 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. Furthermore, 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 the odds-ratios.

Table 3. Most Influential Variables in Intimate-Partner, and non-Intimate Partner Violence (IPV and non-IPV) Predictive Models

IPV	Coef (Std)	Odds-ratio
Abuser's time since last incident	-.500	0.978
Abuser gender: male	.397	1.614
Relationship type: ex	-.305	0.737
Victim gender: males	-.263	0.729
Abuser's count of charges in last 10 yrs[1]	.232	1.122
Question 9 (recent pregnancy): yes	.187	1.319
Dyad's count of incidents in last 2 yrs	.163	1.050
Victim's count of incidents as abuser in last 2 yrs[1]	.153	1.176
Question 7 (conflict over child contact): yes	-.148	0.801
LSOA domestic abuse count in last 2 yrs[1]	.146	1.114
Abuser's time since last crime	-.134	0.999
Dyad's count of serious charges in last yr[1]	.132	1.477
Victim's count of victimisations in last 10 yrs[1]	.132	1.111
Officer risk grading: Medium	.116	1.156
Abuser consumed alcohol: yes	.114	1.129
Non-IPV	Coef (Std)	Odds-ratio
Abuser's time since last incident	-.348	0.985
Dyad's count of incidents in last 2 yrs	.230	1.095
Abuser's time since 1st crime (last 10 yrs)	-.159	0.998
Dyad's count of charges in last yr[1]	.152	1.539
Dyad's count of charges in last 10 yrs[1]	.140	1.230
Victim's count of victimisations in last 10 yrs[1]	.138	1.112
Abuser's time since last crime	-.119	0.999

Note. Due to considerations of space, only variables with standardised coefficient magnitudes in excess of.100 are shown. Emboldened font indicates that the variables are common between the IPV and non-IPV models.

[1] Log-transform was applied prior to logistic regression modelling.

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 an abuser's last domestic abuse incident (where they were the abuser and regardless of how serious the incident was). The more recent the previous incident, the higher the risk. Similarly, in both IPV and non-IPV, the less time that has passed since the abuser was involved in crime (excluding crimes committed in the context of domestic abuse), the higher the risk of serious harm. The most influential geographical demographics variable for IPV is the LSOA-level count of domestic abuse incidents over the past 2 years. No geographic 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 perhaps be attributed to greater third-party intervention in cases of child conflict. We caution against over-interpretation of influential variables. Logistic regression modelling cannot establish causal relationships between predictors and the outcome. It merely identifies correlations.

Is the Model 'Fair'?

Although we could not answer this question in full, we could describe differences in how the model was working for protected groups, in terms of true positive rate (TPR), false positive rate (FPR), positive predictive value (PPV), and accuracy. These are metrics of 'statistical' fairness (Chouldechova & Roth, 2018) that provide the starting point for a conversation about fairness. 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). Instead, a positive prediction means that a victim will receive additional safeguarding support. 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.

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 (Jackson et al., 2012). 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, thus artificially inflating the revictimisation rate in an ODE category. 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 revictimizations rate by ethnic subgroup.

Table 4. Summary Statistics for Each Sensitive Group

Variable Name	Distribution	Revictimisation Prevalence	High Risk Prevalence
IPV Officer Defined Ethnicity of Victim			
Asian	.049	.035	.056
Black	.033	.036	.05
European	.574	.044	.043
Unknown	.344	.022	.036
Non-IPV Officer Defined Ethnicity of Victim			
Asian	.063	.01	.033
Black	.028	.009	.012
European	.574	.012	.013
Unknown	.335	.009	.015
IPV Index of Multiple Deprivation			
1	.371	.043	.052
2-3	.324	.037	.039
4-6	.196	.027	.033
7-10	.109	.026	.029
Non-IPV Index of Multiple Deprivation			
1	.389	.010	.016
2-3	.318	.011	.017
4-6	.186	.010	.010
7-10	.107	.008	.011

Note. Distribution represents the relative size of each group so that, say, for the intimate-partner violence (IPV) group, the sum of values for distribution over ethnic groups was 1. The second and third columns compare high risk prevalence and serious harm revictimisation prevalence for the subgroups.

The TPR, FPR, and 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 & 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 4). We present paired boxplots, with the predictive model on the left and the DASH form on the right. As we are primarily concerned with identifying true positives, TPR and PPV are arguably the most important metrics to compare by. 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 & Bleich, 2013; Kleinberg et al., 2019), it may be preferable to apply the model as opposed to remaining with the current procedure.

Discussion

Model Accuracy

We have shown that the predictive model provides a clear advantage over structured professional judgment and the DASH questionnaire. This is consistent with findings reported from Grogger

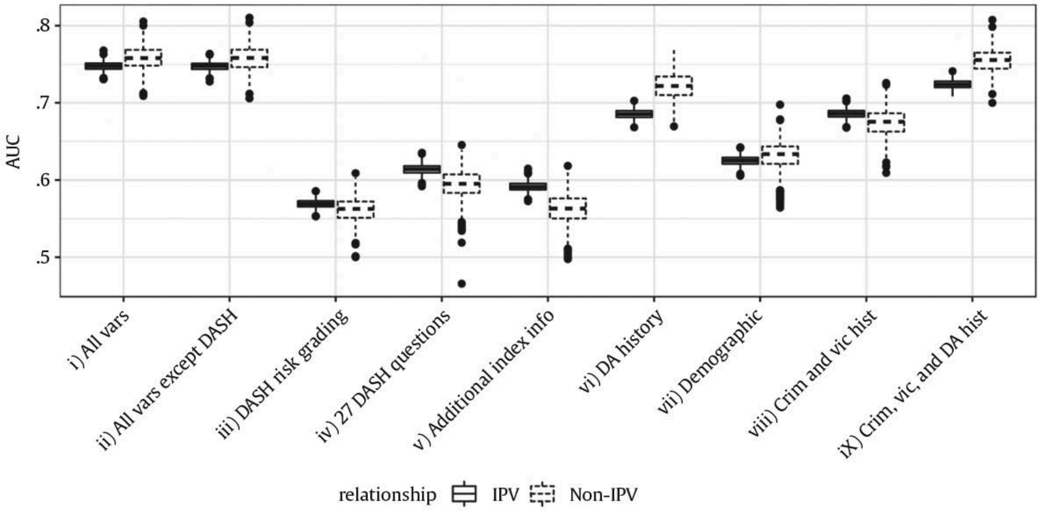


Figure 2. AUC Boxplots Comparing the Full Variable Set with Variable Subsets in Terms of Revictimisation Prediction Capability.
Boxplots are grouped in pairs, with IPV on the left and non-IPV on the right. Note that in descriptors vi) and ix), domestic abuse is abbreviated to DA. Based on 500-times 2-fold cross-validation.

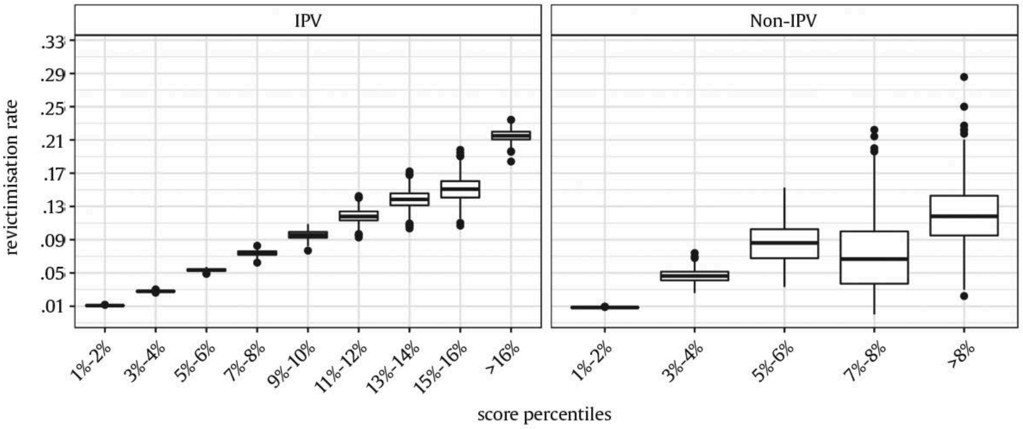


Figure 3. Model Calibration on Full Variable Set.
Note. Grouping by 2 percentile increments, so that, e.g., the left-most boxplot represents all those with predicted revictimisation probability in the range (0.2%). Based on 500-times 2-fold cross-validation.

et al. (2020) with data from Greater Manchester police. Our 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 of the categories criminal history and domestic abuse history, particularly time since the last incident. 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, there was negligible drop in model accuracy. When DASH variables were excluded, the AUC fell by only .0001 on IPV, and .0007 on non-IPV, data. The model is well-calibrated.

These results may look surprising, insofar as DASH includes risk factors considered as relevant in past literature (victim pregnancy, use of guns, strangulation, etc.). We suspect the poor performance of these risk factors is linked to problems with the way and the context in which DASH is administered. 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 perhaps 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 ensures appropriate investment in police training, and also the development of protocols for ensuring that the questioning is done in conditions that are conducive to establishing rapport with the victims and securing their trust.

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

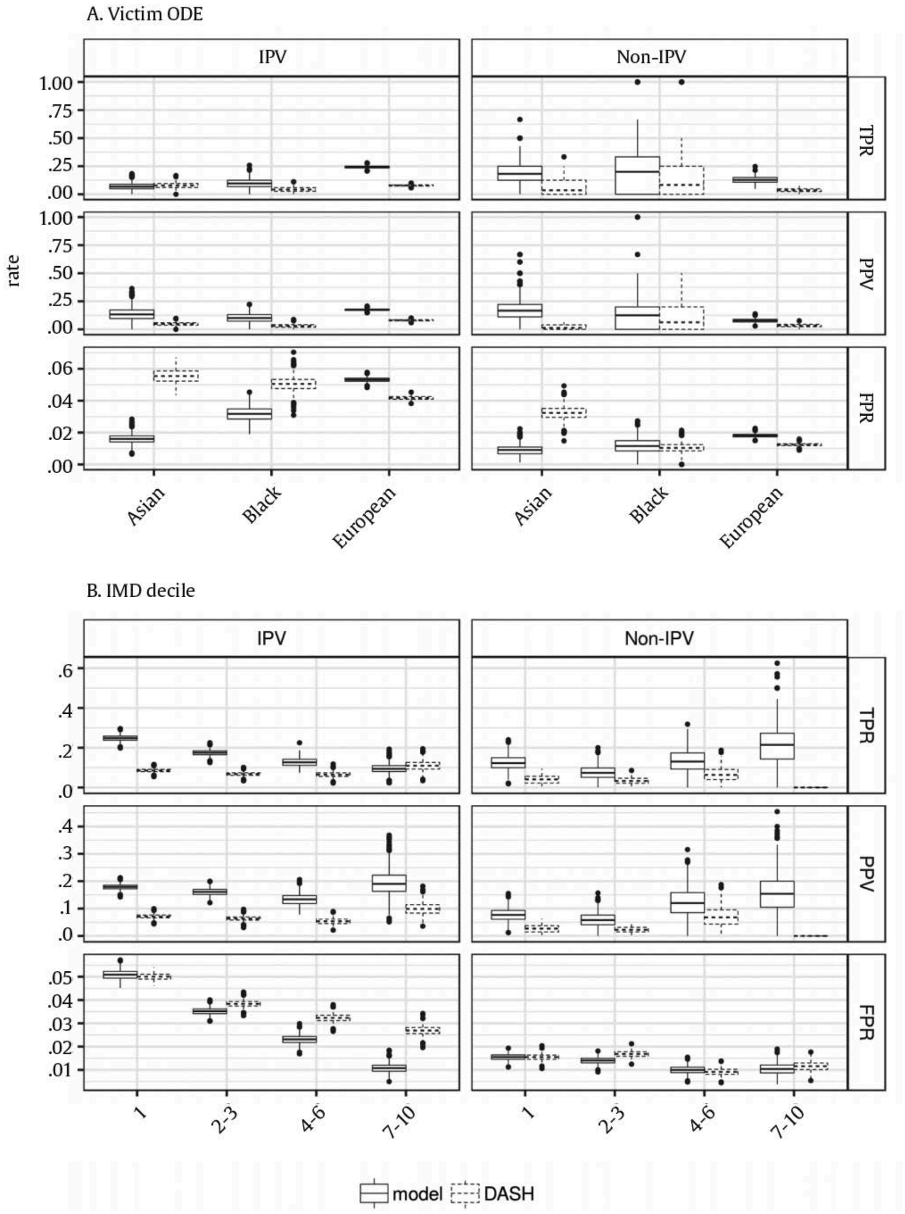


Figure 4. Model Fairness: Performance Metrics for, A. Ethnic Subgroups as Measured by Officer-defined Ethnicity (ODE) of the Victim, and B. Socio-demographics as Measured by the Index of Multiple Deprivation (IMD) Decile.
Note. Metrics are identified on the tabs to the right of the plots. They are true positive rate (TPR), positive predictive value (PPV), and false positive rate (FPR).

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 DASH. However, we suggest this occurred only to a minimal extent. Although information on the risk management actions per se was not available in this police force, 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. 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, indicates that the effect of intervention on the outcome is likely minimal.

Model Fairness

Machine learning models have a direct and serious impact on people's lives, and we are now reckoning with the consequences of this (Angwin et al., 2016; Couchman, 2019; Partnership on AI, 2019;

Commons Select Committee: Science and Technology Committee, 2018). While our work is situated within this debate, we could not propose to resolve these issues here. Instead, we provided an overview of several metrics of statistical fairness (Chouldechova & Roth, 2018), that compare differences between subgroups based on ethnicity and demographics. 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.

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 shortcoming in the analysis is that our data only captures incidents reported to the police. Survey indicates that 79% of victims do not report domestic abuse (ONS, 2018). It is also difficult to estimate the effect of coercive control (Stark, 2009). 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, varies within different wealth and ethnic demographics (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. 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).

Moreover our outcome only captures physical violence, whereas we know that coercive control is the most harmful form of abuse in terms of serious harm risk (Monckton-Smith et al., 2017; Sharp-Jeffs & Kelly, 2016). It is also the type of domestic abuse that mostly goes unrecognised by frontline officers (Robinson et al., 2016). Efforts are underway to equip officers with the training and tools required to better identify these dynamics (Wire & 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 (Myhill & Hohl, 2016; Stark, 2012).

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 the Black community were 9.5 times more likely than their White peers to be stopped and searched by police (ONS, 2019). As serious crimes are less subject to bias, it may be considered fairer to use charge data for these. However, 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.

So What? What's 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. At present, this information seems of better quality than that obtained through the use of police-administered questionnaires. Grogger et al (2020) using a similar approach to ours and finding similar results have suggested it may be advisable to just use criminal history to triage incoming incident calls and then use a more sensitive police questionnaire to tease out false positives and false negatives. Past work has also 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. It should also be clear from our analysis that different risk factors are relevant for IPV and non-IPV abuse. The continued reliance in one single risk assessment tool with common risk factors is far from optimal and we should move away from that approach.

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 only on administrative data. The Spanish experience with the VIOGEN system shows that it is technically possible to devise this sort of automatised systems, though this may be more challenging in decentralised police systems. 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 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 but also to use their discretion when needed to override the model recommendation. 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 can be integral for gathering information about key risk factors (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 in 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 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 gets us closer to achieving our societal goals (Stevenson & 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 suggests we need to be moving toward systems like those implemented by Spanish police that require recontacting 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).

Conflict of Interest

The authors of this article declare no conflict of interest.

Acknowledgement

We want to thank Prof. David Gadd for the comments he provided to earlier drafts of this paper.

References

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016, May 23). *Machine bias*. ProPublica.

Ashby, M. (2018). Comparing methods for measuring crime harm/severity. *Policing: A Journal of Policy and Practice, 12*(4), 439-454. https://doi.org/10.1093/police/pax049

Barrett, B. J., Peirone, A., Cheung, C. H., & Habibov, N. (2017). Pathways to police contact for spousal violence survivors. *Journal of Interpersonal Violence, 36*(1-2), 1-31. https://doi.org/10.1177/0886260517729400

Berk, R. A., & Bleich, J. (2013). Statistical procedures for forecasting criminal behaviour. *Criminology and Public Policy, 12*(3), 513-544. https://doi.org/10.1111/1745-9133.12047

Berk, R., Heidari, H., Jabbari, S., Kearns, M., & Roth, A. (2018). Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research, 50*(1), 3-44. https://doi.org/10.1177/0049124118782533

Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data, 5*(2), 153-163. http://doi.org/10.1089/big.2016.0047

Chouldechova, A., & Roth, A. (2018). *The frontiers of fairness in machine learning*. arXiv. https://arxiv.org/abs/1810.08810

Commons Select Committee: Science and Technology Committee. (2018, May 15). *Algorithms in decision-making*. HC 351 2017-2019.

Corbett-Davies, S., & Goel, S. (2018). *The measure and mismeasure of fairness: A critical review of fair machine learning*. arXiv. https://arxiv.org/abs/1808.00023

Couchman, H. (2019). *Policing by machine*. Liberty.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. *Proceedings of the 3rd Innovations in Theoretical Science Conference* (pp. 214-226). https://doi.org/10.1145/2090236.2090255

Ferguson, A. G. (2017). *The rise of big data policing*. NYU University Press.

Friedler, S. A., Scheidegger, C., & Venkatasubramanian, S. (2016). *On the (im)possibility of fairness*. arXiv. https://arxiv.org/abs/1609.07236

García, S., Luengo, J., Sáez, J. A., López, V., & 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. https://doi.org/10.1109/TKDE.2012.35

Gelman, A. (2008). Scaling regression inputs by dividing by two standard deviations. *Statistics in Medicine, 27*(15), 2865-2873. https://doi.org/10.1002/sim.3107

Grogger, J., Ivandic, R., & Kirchmaier, T. 2020. *Comparing conventional and machine learning approaches to risk assessment in domestic abuse cases* (Discussion Paper 1676). CEP.

HMIC. (2014). *Everyone's business: Improving the police response to domestic abuse*.

Home Affairs Committee. (2009). *The Macpherson Report - ten years on*. House of Commons.

Jackson, J., Bradford, B., Stanko, E. A., & Hohl, K. (2012). *Just authority? Trust in the police in England and Wales*. Routledge.

Jung, S., & Buro, K. (2016). Appraising risk for intimate partner violence in a police context. *Criminal Justice and Behavior, 44*(2), 240-260. https://doi.org/10.1177/0093854816667974

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

Kleinberg, J., Ludwig, J., Mullainathan, S., & Sunstein, C. R. (2019). Discrimination in age of algorithms. *Journal of Legal Analysis, 10*, 113-174. https://doi.org/10.1093/jla/laz001

Kuhn, M., & Johnson, K. (2013). *Applied predictive modelling*. Springer.

Kusner, M. J., Loftus, J. R., Russell, C., & Silva, R. (2017). *Counterfactual fairness*. arXiv. https://arxiv.org/abs/1703.06856

Kwak, H., Dierenfeldt, R., & 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. https://doi.org/10.1016/j.jcrimjus.2018.11.001

Leslie, D. (2019). *Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector*. The Alan Turing Institute.

López-Ossorio, J., González-Álvarez, J. L., & Andrés-Pueyo, A. (2016). Predictive effectiveness of the police risk assessment in intimate partner violence. *Psychosocial Intervention, 25*(1), 1-7. https://doi.org/10.1016/j.psi.2015.10.002

Machado, P., Pais, L., Felgueiras, S., & Quaresma, C. (2021). Frontline responses to high impact domestic violence in Portugal. In B. Lobnikar, C. Vogt, & J. Kersten (Eds.), *Improving frontline responses to domestic violence in Europe*. University of Maribor Press.

Medina Ariza, J. J., Robinson, A., & Myhill, A. (2016). Cheaper, faster, better: Expectations and achievements in police risk assessment of domestic abuse. *Policing, 10*(4), 341-350. https://doi.org/10.1093/police/paw023

Messing J., & Thaller, J. (2013) The average predictive validity of intimate partner violence risk assessment instruments. *Journal of Interpersonal Violence, 28*(7), 1537-1558. https://doi.org/10.1177/0886260512468250

Monckton-Smith, J., Szymanska, K., & Haile, S. (2017). *Exploring the relationship between stalking and homicide*. Suzy Lamplugh Trust.

Myhill, A. (2015). Measuring coercive control: What can we learn from national population surveys? *Violence Against Women, 21*(3), 355-375. https://doi.org/10.1177/1077801214568032

Myhill, A., & Hohl, K. (2016). The "Golden Thread": Coercive control and risk assessment for domestic violence, *Journal of Interpersonal Violence, 34*(21-22), 4477-4497. https://doi.org/10.1177/0886260516675464

Office for National Statistics. (2018). Domestic abuse: Findings from the Crime Survey for England and Wales: Year ending March 2017.

Office for National Statistics. (2019). Stop and search. Police powers and procedures England and Wales statistics.

Partnership on AI. (2019). *Report on algorithmic risk assessment tools in the U.S. criminal justice system*.

Peralta, D. (2015). *Data mining for the prediction of domestic violence* [Unpublished master's thesis]. University of Manchester.

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

Sharp-Jeffs, N., & Kelly, L. (2016). *Domestic homicide review (DHR) case analysis*. Standing Together/London Metropolitan University.

Sherman, L. (2018). Policing domestic violence, 1967-2017. *Criminology and Public Policy, 17*(2), 453-465. https://doi.org/10.1111/1745-9133.12365

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.

Stark, E. (2012). Looking beyond domestic violence: Policing coercive control. *Journal of Police Crisis Negotiations, 12*(2),199-217. https://doi.org/10.1080/15332586.2012.725016

Stevenson, M., & Doleac, J. (2019). *Algorithm risk assessment in the hands of humans*. SSRN. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3489440

Svalin, K. (2018). *Risk assessment of intimate partner violence in a police setting* (Unpublished doctoral dissertation). Malmo University.

Turner, E., Medina, J., & Brown, G. (2019). Dashing hopes? The predictive accuracy of domestic abuse risk assessment by police. *The British Journal of Criminology, 59*(5), 1013-1034. https://doi.org/10.1093/bjc/azy074

Van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

Veale, M., Van Kleek, M., & Binns, R. (2018). *Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making*. arXiv. https://arxiv.org/abs/1802.01029

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, 31*(2), 841-997. https://doi.org/10.2139/ssrn.3063289

Ward-Lasher, A., Messing, J., Cimino, A., & Campbell, J. (2018). The association between homicide risk and intimate partner violence arrest. *Policing, 14*(1), 228-242. https://doi.org/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. https://doi.org/10.1146/annurev-criminol-011518-024748

Xu, B., Huang, J.Z., Williams, G., Wang. Q., & Ye, Q. (2012). Classifying very high-dimensional data with random forests built from small subspaces. *IJDWM, 8*(2), 44-63. http://doi.org/10.4018/jdwm.2012040103

Yang, Y., & Webb, G. I. (2009). Discretization for naive-Bayes learning: Managing discretization bias and variance. *Machine Learning, 74*(1), 39-74. https://doi.org/10.1007/s10994-008-5083-5

Ybarra, L. M., & Lohr, S. L. (2002). Estimates of repeat victimization using the National Crime Victimization Survey. *Journal of Quantitative Criminology, 18*(1), 1-21. https://doi.org/10.1023/A:1013244611986

Zighed, D. A., Rabaséda, S., & Rakotomalala, R. (2003). FUSINTER: A method for discretization of continuous attributes. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems, 06*(03), 307-326. https://doi.org/10.1142/S0218488598000264

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, 67*(2), 301-320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

Appendix

The DASH Form for Risk Assessment

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/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?
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/dependents?
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?
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?
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?