doi:10.1093/bjc/azy074 BRIT. J. CRIMINOL. (2019) **59**, 1013–1034
Advance Access publication 13 June 2019

DASHING HOPES? THE PREDICTIVE ACCURACY OF DOMESTIC ABUSE RISK ASSESSMENT BY POLICE

Emily Turner*, Juanjo Medina and Gavin Brown

The Domestic Abuse, Stalking and Honour Based Violence (DASH) form is a standardized risk assessment implemented across most UK police forces. It is intended to facilitate an officer's structured professional judgment about the risk a victim faces of serious harm at the hand of their abuser. Until now, it has been an open question whether this tool works in practice. Here, we present the largest scale European study, making the case that the risk assessment tool is underperforming. Each element of the DASH questionnaire is, at best, weakly predictive of revictimization. Officer risk predictions based on DASH are little better than random and a logistic regression model that predicts the same outcome using DASH only provides modest improvement in performance.

Key Words: domestic abuse, risk assessment, police, predictive policing, machine learning

Introduction

This article evaluates the tool used by most British police forces to assess the risk for domestic abuse. One of the most notable reforms on policing domestic abuse internationally has been the introduction of standardized risk assessment. Repeat victimization is higher for crimes of domestic violence (Ybarra and Lohr 2002), so when cases come to the attention of the police, the opportunity for prevention of new incidents is salient. A number of studies also conclude that a 'power few' (Sherman 2007) perpetrators are responsible for most domestic abuse harm and that this harm is also concentrated in a minority of victims, therefore, valid prediction models to prospectively identify them may be a fruitful strategy (Bland and Ariel 2015; Barnham *et al.* 2017). Thus, although risk assessment can serve a variety of goals (Medina Ariza *et al.* 2016), trying to correctly identify the cases that are more likely to experience future incidents remains a central rationale. This is more so in a context of austerity policies in many countries, in which policing and social resources are under severe pressure, but also given national trends on domestic abuse. Since 2008, the decade-long downward trend has ceased, and for female victims, there has been an uptick, which is driven in part by a rise in the number of victimizations experienced by a minority of high-frequency victims (Walby *et al.* 2016). Risk assessment is intended to allow professionals to prioritize high-risk cases, allocating spare resources where they are most needed.

Yet, despite the introduction of domestic abuse risk assessment in policing, we still have a limited understanding of whether it works in practice. Here, we present the results from the largest scale European study. We open up with a review of the existing

*Emily Turner, Centre for Criminology and Criminal Justice (CCCJ), School of Law, University of Manchester, UK; emily.turner@manchester.ac.uk; Juanjo Medina, CCCJ, School of Law, University of Manchester, UK; Gavin Brown, Machine Learning and Optimisation, School of Computer Science, University of Manchester, UK.

literature on domestic abuse risk assessment and more recent studies about the main tool used by police forces in the United Kingdom. Using machine learning tools, we aim to establish a better understanding of police risk assessment and its effectiveness at identifying the 'right' victims and perpetrators by answering four questions:

(1) Are the police successfully distinguishing the high-risk cases?
(2) How does each question influence perception of risk?
(3) Can a statistical model that predicts new incidents of domestic abuse based on the Domestic Abuse, Stalking and Honour Based Violence (DASH) questions outperform officer risk assessment?
(4) Are all of the DASH questions required, or can DASH be shortened without reducing prediction accuracy?

We conclude that this tool offers limited value for correctly identifying high-risk victims. We reason that this is likely the result of the difficulties that arise in police–victim encounters. We also discuss the danger these tools present of continuing to emphasize a focus on incidents and measured harm.

Domestic abuse risk assessment

Early risk assessment tools were developed during the 1990s in North America with a focus on identifying female victims at a high risk of a new, particularly more serious, assault from their intimate partner. The best known include the Danger Assessment (Campbell *et al.* 2009), Spousal Assault Risk Assessment (SARA) (Kropp *et al.* 1995), the Domestic Violence Screening Inventory (Williams and Houghton 2004) and the Ontario Domestic Assault Risk Assessment (ODARA) (Hilton *et al.* 2010). These tools originated for use in clinical and probation settings and were adopted by victim advocates to guide safety planning. Subsequently, they were extended to policing practice.

Research on the validity of these tools has been carried out primarily in North America (Northcott 2012; Messing and Thaller 2014). Most validation studies were conducted by their developers with relatively small samples (around 500–2,000) of mostly white male adult perpetrators of intimate partner violence (IPV). This literature reports small effect sizes for their classification accuracy, with ODARA showing a moderate effect size (Hanson *et al.* 2007; Messing and Thaller 2013), and under certain circumstances, SARA (Helmus and Bourgon 2011). These effect sizes are similar to those of predictive tools for general violence (Coid *et al.* 2011).

These tools were developed when research on risk factors for IPV recidivism was in its infancy, which may explain its poor performance (Helmus and Bourgon 2011). For example, some include items that have subsequently been found not to be associated with re-assaults (Wong and Hisashima 2008; Hilton *et al.* 2010). The methods employed for the development of these tools have also been suboptimal and, as argued by Hilton *et al.* (2004), more oriented to psychological test construction than the development and testing of predictive models. ODARA is one of the few popular tools validated in a test sample (Hilton *et al.* 2004). More recently a new trend on risk assessment research has been to rely on machine learning tools to develop new more effective scoring rules for these tools, even in the domestic abuse setting (Berk *et al.* 2005; 2016), raising new

possibilities and new dilemmas, particularly regarding fairness and transparency (Berk *et al.* 2018; Ferguson 2017).

Research in the United States suggests that policing responses that are informed by the use of risk tools may reduce offending and increase victim satisfaction with the police (Messing *et al.* 2014), but other studies are more critical of both the effectiveness of policing based on predictions of individual recidivism (Saunders *et al.* 2016) and their ethical dimensions (Ferguson 2017). In any case, the development of these tools established a trend for policy reform that eventually reached Europe. By 2004, the European Parliament was already calling on Member States to take appropriate measures on gender violence including the development of 'adequate risk assessments'. Subsequently, the EU Victims' Right Directive (2012/29) urged Member States to ensure that 'victims receive a timely and individual assessment, in accordance with national procedures, to identify specific protection needs' (article 22).

Various European countries have now introduced risk assessment in the context of domestic abuse policing: Swedish police forces introduced B-Safer in the 1990s, an abbreviated version of SARA not requiring clinical training (Svalin 2018); Spain introduced a nationally centralized system for dynamic ongoing assessment in 2007 (Lopez-Ossorio *et al.* 2016); Portugal recently adapted the Spanish model; and police forces in the United Kingdom have been using a standardized tool called DASH since 2009. In this international context, the development of evidence about these tools seems critical, as policy bodies from the European Union have recognized.

Dash

Our article focuses on police use of DASH. DASH is also used by many other victim support organizations and was accredited by chief police officers in 2009. DASH aims to identify domestic abuse victims at high risk of serious harm (CAADA 2012) for referral to Multi-Agency Risk Assessment Conferences (MARAC)—multiagency panels producing coordinated action plans to increase victims' safety and to manage perpetrators behaviour. The extent to which DASH 'is used by responding officers and the way in which it is used, vary significantly from force to force' (HMIC 2014: 13). Research has identified three main models of implementation: (1) DASH is filled by frontline officers but the grading is later applied by a specialist; (2) frontline officers filled DASH and grade the incident with a specialist auditing a subset of cases; and (3) frontline officers complete DASH and grade the incident and a secondary risk assessor reviews grading for all cases (Robinson *et al.* 2016). Although grading procedures vary, typically attending officers will complete DASH when responding to a call or shortly afterwards.

DASH is a structured professional judgement scale. CAADA (2012) suggests a threshold of 14 as appropriate for classification as high risk, but emphasizes professional judgement and asks officers for their own assessment of risk. Being identified as high risk determines the level of intervention and support services provided to victims and, therefore, has potentially very real implications.

Although it is often referred to as an 'evidence-based' tool that 'saves lives',[1] there is not much published research estimating the classification error resulting from using DASH. We simply do not know if the classifications made with DASH are good

[1] See http://www.dashriskchecklist.co.uk.

enough (Pease *et al.* 2014), and we have no evidence either about its crime preventative impact. Some studies have suggested that only a small subset of factors measured by DASH are associated with 'recidivism' (Almond *et al.* 2017) and that DASH gradings are poor predictors of subsequent homicide (Chalkley and Strang 2017; Thornton 2017). A recent study concluded the tool is not applied consistently at the frontline and that often enough errors in recording contaminate the process (Robinson *et al.* 2016). Unsurprisingly, Her Majesty Inspectorate of Constabulary's (HMIC 2014) report on domestic violence *recommends that the effectiveness of this approach to risk management be reviewed.*

Data

Data pipeline

We used data provided by a large metropolitan police force in the United Kingdom. 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 its HMIC scores on the quality of services it provides (PEEL assessments).

The police force responded to ~350,000 domestic violence incidents between 2011 and 2016 inclusive. For each incident, we needed to be able to identify the perpetrator and primary victim involved, the type of relationship between perpetrator and victim (e.g. IPV or other), and whether there were any criminal charges associated with an incident. Additionally, we focussed on cases where officer risk grading (the officer prediction that the victim is facing either a high, medium or standard risk of serious harm) had been set and where neither the victim nor abuser had died or was too ill. There were 6 per cent of incidents missing perpetrator or victim identifiers and a further 8 per cent missing relationship type, 3 per cent missing a key field for merging on criminal charge data, 4 per cent where officer grading was not given, and in 0.01 per cent of the cases either the victim or abuser was dead or too ill. One or more of these fields were missing in 16 per cent of cases. These were excluded from the data.

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-focussed.

A small portion of perpetrator–victim pairs (~1 per cent) were recorded as being involved in more than one incident in a day. We did not know the time at which incidents occur, so we could not determine which incident occurred first. It is possible that some of these were duplicated records. Thus, where this occurred, only one incident was kept, and the rest were excluded. The incident with the most complete DASH questions was prioritized, and where completeness was the same for multiple incidents, the highest risk case was prioritized.

Of the remaining data, 30 per cent of the perpetrator-victim pairs were involved in more than one incident. We kept the first incident that a pair was involved in where we could see that they had not been involved in another incident together for at least two years prior to the incident in question. We excluded incidents from the years 2011 and 2012 because for these cases we could not confirm whether or not the pair was involved in an incident in the prior two years. By keeping only one incident per pair,

the assumption of independence of observations was preserved, which is required by the logistic regression algorithm. This also makes the analysis reproducible. Although it was not the focus of this article, we ran the same analysis for robustness purposes on a sample that was not exclusively made up of 'first' times. In this sample, if a perpetrator and victim were involved in multiple incidents, we randomly selected one of these incidents to be the index incident. So for these repeaters, the index incident could have been their first appearance on police records or it could have been the second, third etc.

As we are interested in serious harm that is caused up to 365 days after an index event, and as criminal charges data was only available up to the end of 2016, incidents occurring in 2016 were excluded because we could not see what the outcome was for these. Incidents from 2011, 2012 and 2016 make up 46 per cent of the data, and a further 26 per cent were excluded because they were a repeat appearance of a perpetrator–victim pair. After these final exclusions, we were left with ~86,000 incidents.

Of the remaining incidents, 19 per cent had the response 'not known' for *all 27 questions*. These were excluded from the model because we had no 'yes'/'no' answers to base our predictions on for these cases. An additional 11 per cent had between 1 and 26 'not known' answers. The missingness was missing not at random and hence was non-ignorable. These were retained in the model, and the 'not known' response treated as a third level to each predictor.

Finally, this data set 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 41,570 incidents. The non-IPV data set was made up of all other relationships and had 19,510 incidents.

Variables

The data consisted of 27 DASH questions, officer-defined risk and 24 recidivism/revictimization outcomes. The answers to the DASH questions were the predictor variables, each taking a value of 'yes', 'no' or 'not known'. The officer-defined risk assignment was either 'high', 'medium' or 'standard'. Official guidelines dictate that a grading of 'high' indicates belief that the victim is at risk of serious harm occurring at any time; 'medium' is a prediction that serious harm is unlikely unless circumstances change for the victim or perpetrator; and 'standard' is a prediction that there is no evidence indicating the likelihood of serious harm.

Our analysis aimed to evaluate how well the DASH assessments or alternative scoring of the DASH questionnaires can predict future harm. We used several definitions of future harm, but focus on serious harm revictimization for most of the discussion here because this is what officers are predicting when they grade a case as high risk. DASH is a victim-focussed tool so that officers are encouraged to predict revictimization 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 (regardless of whether the incident was flagged as domestic abuse in the police systems), we defined this as revictimization. We defined *serious harm* as any crime in the violence against the person or sexual offences category with a score greater than or equal to 184 on the Office

Table 1 *Revictimization prevalence and distribution of officer risk grading*

Description	IPV	Non-IPV
Outcome	Revictimization rate	
Serious harm revictimization within 365 days[a]	0.056	0.015
Officer risk grading	Risk distribution	
Standard	0.762	0.821
Medium	0.201	0.162
High	0.037	0.017

[a]Note the difference between the rate of revictimization and recidivism (causing serious harm, within 365 days). The recidivism rate is 0.071/0.057 (IPV/non-IPV), implying that 21%/68% of the recidivism cases were with a new victim. Particularly, for the non-IPV group, the difference between recidivism and revictimization rates seems very large. It may be due to deficiencies in the data. We suspect that the true rate of revictimization lies somewhere between the recidivism and revictimization rates and possibly is closer to the recidivism rate. When we used the recidivism outcome instead of the revictimisation outcome the overall conclusions of the article remained the same. See repository for further information.

for National Statistics Crime Severity Score (ONS score).[2] This is the 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 represents the 'ground truth', where ground truth is defined as that which we observe in the data rather than infer from a predictive model. We compared officer risk gradings to the ground truth to evaluate officer prediction accuracy. See Tables 1 and 2 for basic statistics on all variables.

Although this outcome provided the closest approximation of what the police are trying to predict, we could have defined future harm in other ways. Defining 'serious harm' as a crime with an ONS score of at least 184 may have set the threshold too high. The fact that the victim has gotten the police involved was already an indicator of serious distress. A second category of outcome lowered this threshold by defining revictimization as *any subsequent domestic violence incident*, regardless of whether or not there was a crime associated with it. The third, intermediary category defined *any* crime, rather than any incident, committed in the context of domestic violence as revictimization.

Apart from predicting whether a victim is at risk of further victimization, 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. It was worthwhile, thus, to evaluate whether DASH predicts perpetrator recidivism as well. This was called the recidivism outcome.

The two types of recurrence (recidivism and revictimization) and three harm definitions (serious crime, any crime and incident) were combined to make six categories of outcomes. For each of these outcomes, we examined four time limits: 30, 90, 180 and 365 days. This amounted to a total of 24 different outcomes. For reasons of available space, the figures in this essay focus on revictimization within 365 days causing serious harm, but we provide detailed results from evaluating the alternative outcomes in a public GitHub repository,[3] and our discussion later on will make cursory reference to those more detailed results.

[2] This is an experimental statistic that intends to capture the relative harm of an offence to society and that has been developed by the Office for National Statistics.

[3] https://gitlab.cs.man.ac.uk/emily-turner/dash_analysis.

Table 2 *Distribution of each of the 27 DASH questions*

Description	IPV		Non-IPV	
	Yes	Not known	Yes	Not known
Q.1 Has the current incident resulted in injury?	0.166	0.010	0.136	0.008
Q.2 Are you very frightened?	0.255	0.063	0.207	0.060
Q.3 What are you afraid of? Is it further injury/violence?	0.267	0.080	0.222	0.076
Q.4 Do you feel isolated from friends/family?	0.119	0.081	0.066	0.077
Q.5 Are you depressed/having suicidal thoughts?	0.147	0.085	0.108	0.083
Q.6 Have you separated/tried to separate in last year?	0.417	0.078	0.068	0.081
Q.7 Is there conflict over child contact?	0.129	0.071	0.048	0.074
Q.8 Do they constantly text/call/contact/ follow/stalk/harass you?	0.156	0.084	0.057	0.080
Q.9 Are you currently pregnant or have you had a baby in past 18 months?	0.163	0.068	0.054	0.073
Q.10 Are there any children/step-children not belonging to the abuser, or other dependents, in the household?	0.147	0.072	0.064	0.074
Q.11 Have they ever hurt the children/dependents?	0.024	0.082	0.032	0.082
Q.12 Have they ever threatened to hurt/kill children/dependents?	0.020	0.084	0.026	0.083
Q.13 Is abuse happening more often?	0.181	0.091	0.160	0.086
Q.14 Is abuse getting worse?	0.195	0.091	0.159	0.087
Q.15 Do they try to control everything you do or are they excessively jealous?	0.234	0.093	0.078	0.087
Q.16 Have they ever used weapons/objects to hurt you?	0.070	0.093	0.047	0.088
Q.17 Have they ever threatened to kill you/ someone else and you believed them?	0.074	0.094	0.052	0.089
Q.18 Have they ever attempted to strangle/ choke/suffocate/drown you?	0.104	0.094	0.035	0.089
Q.19 Do they do/say things of a sexual nature that make you feel bad or that physically hurt you or someone else?	0.059	0.098	0.018	0.091
Q.20 Is there another person who has threatened you or that you are afraid of?	0.031	0.094	0.029	0.089
Q.21 Do you know if they have hurt anyone else?	0.091	0.095	0.092	0.091
Q.22 Have they ever mistreated an animal or the family pet?	0.027	0.094	0.017	0.090
Q.23 Are there financial issues? For example, are you dependent on them for money/have they recently lost their job?	0.166	0.092	0.161	0.088
Q.24 Have there been problems in past year with drugs/alcohol/ mental health leading to problems in leading a normal life?	0.297	0.092	0.321	0.088
Q.25 Have they ever threatened/attempted suicide?	0.132	0.097	0.097	0.096
Q.26 Have they ever breached bail/injunction/other agreement for when they can see you/the children?	0.034	0.096	0.034	0.095
Q.27 Do you know if they have ever been in trouble with the police/has a criminal history?	0.349	0.089	0.343	0.089

Methods

Mutual information (MI) quantifies the bivariate relationships between each question, officer risk grading and each of the outcomes. It is a measure of mutual dependence between two variables (Cover and Thomas 2005). For example, I*(X; Y)* is the MI between *X* and *Y*. It tells us how much knowing about one variable, say, *X*, informs us about the values another variable, *Y*, may take. If *X* and *Y* are independent, knowing the value of one does not tell us anything about the other, so the MI will theoretically be equal to zero (in reality, it will be close to, but not exactly, zero).

MI is a component of the G-test statistic, which we use to determine whether a relationship between two variables is statistically significant. The G-test of independence is a generalized likelihood ratio test (Woolf 1957), calculated as *2*N**I*(X; Y)*, where *N* is

the number of observations in the data set. The G-statistic follows the chi-square distribution, so that it is asymptotically the same as the chi-square statistic.

The effect size, w, for the G-test is w = sqrt(2*N*I(X; Y)/N), which simplifies to w = sqrt(2*I(X; Y)). In this way, MI can be seen as the natural unit of effect size for the G-test (Rosenthal 1994). Cohen (1988) provides a guide on effect sizes for the chi-square statistic, describing a small effect size as $w = 0.1$, medium as $w = 0.3$ and large as $w = 0.5$. Since the G and chi-square statistics are asymptotically the same, we can use this guide on the G-test effect size too: $w = 0.1 \Leftrightarrow$ I(X; Y) = 0.005, $w = 0.3 \Leftrightarrow$ I(X; Y) = 0.045, $w = 0.5 \Leftrightarrow$ I(X; Y) = 0.125 (Sechidis *et al.* 2014).

The lower bound on MI is zero. The upper bound is less straightforward and depends on the marginal distribution of each variable. Since we want to compare degrees of relatedness between different combinations, we normalize the MI. By dividing the MI value by the minimum of the entropy values of X and Y (Kvalseth 1987), a normalized value that lies between 0 and 1 is achieved. A value of 1 signifies that two variables are completely correlated, whereas a value of 0 indicates independence. This value will only be large if the dependence between the variables is high, and is not also a function of the marginal distribution of each variable.

We apply eight machine learning models to predict future harm: logistic regression, naive Bayes, tree-augmented naive Bayes, decision tree, random forest, gradient boosting, k-nearest neighbour (k-NN) and support vector machine (SVM) with polynomial kernel. As we are primarily concerned with ranking incidents in terms of victim risk, we used receiver operating characteristic (ROC) curves and area under the ROC curve (AUC) to evaluate and compare predictive models.

The logistic regression model is built with forward stepwise Akaike Information Criteria variable selection (Akaike 1998) and, separately, also with elastic net regularization (Zou and Hastie 2005). At the model build stage, we observed that naive Bayes, tree-augmented naive Bayes, k-NN and SVM models also benefited from feature selection. For each variable, we calculated the AUC derived from a univariable logistic regression model that predicted the outcome with each variable individually and then ranked the variables by this value (Kuhn and Johnson 2013). Where feature selection was applied, we built the model on variable subsets and selected the subset that corresponded to the best model. The smallest subset had the best predictors (in terms of AUC) and the largest subset contained all of the features.

There is a class imbalance of 16:1/55:1 (IPV/non-IPV), which affected the performance of the decision tree algorithm. For this model, we applied class weights to balance the classes. Weights did not improve the performance of the two other tree-structured models, random forest and gradient boosting, so for these, we report the results for the unweighted outcome.

In each model build, a predicted probability of revictimization/recidivism is output for each observation. This is converted into a classification by choosing a threshold in the probability range and, for the revictimization case, predicting any observation with a probability above that threshold as revictimization and predicting the rest of the observations as non-revictimization. The true positive rate (TPR) and false positive rate (FPR) can then be calculated. TPR is the proportion of true revictimization cases that were correctly predicted as revictimization. The FPR is the proportion of true non-revictimization cases that were incorrectly predicted as revictimization. In other words, TPR represents the rate of revictimization detection, and the FPR represents the rate of

false alarms. The AUC represents the probability that a classifier will rank a randomly chosen revictimization cases above a randomly chosen non-revictimization case. We also refer to the positive predictive value (PPV), also known as precision, which is the proportion of revictimization predictions that were correct.

We applied five-time two-fold cross-validation (Dieterich 1998) to evaluate and compare each modelling approach. Thus, for each model, we display mean and standard deviation of the AUCs from 10 cross-validation builds. We provide the rate-wise consensus ROC curve and standard deviation on TPR and FPR across cross-validation runs. Logistic regression does not involve hyperparameter selection; however, for all other models, including logistic regression with elastic net, hyperparameters are selected using a further, nested cross-validation.

Findings

Are the police successfully identifying high-risk cases?

Performance of officer gradings was evaluated with reference to the 'ground truth' outcome that was described in the Data section. If an officer classified a case as high risk, they predicted that the perpetrator could cause 'serious harm' to the victim 'at any time'. Serious harm is defined as a criminal charge for a violent or sexual crime with an ONS harm score of 184 or more. It is critical to highlight that insofar as legislation, police and recording practices (Robinson *et al.* 2018a), and harm indexes underplay or inadequately capture the continued psychological and social harm that results from abuse, statistical models will only be able to adequately model incidents of physical harm. The harm must have occurred in the 365 days after the index event to come within our outcome definition. This outcome is different to the officer prediction insofar as the officer predicts that serious harm could befall the victim, whereas the outcome represents the event where serious harm occurred *and the police were involved*. This inherent limitation in the data highlights another challenge for predicting domestic abuse. As the most significant factor influencing a victim decision to report IPV to the police is the severity of the offence (Barrett *et al.* 2017; Smith 2017), this limitation should be more significant in our models focussing on less serious outcomes.

With this outcome, we evaluated officer performance. Table 3 displays the outcome rate for each grading. In the IPV data, the officers correctly ranked 5.7 per cent of the revictimization cases as high risk. This is the TPR of their predictions. The false negative rate (FNR) is at least 67.2 per cent, and given the significantly lesser amount of help received by the medium risk, as opposed to high risk, victim, we could consider

Table 3 *Displays, for a given revictimization outcome and data set, the distribution of officer grading*

Risk	IPV		Non-IPV	
	0	1	0	1
Standard	0.767	0.672	0.823	0.728
Medium	0.197	0.270	0.161	0.245
High	0.035	0.057	0.017	0.027

A 0 indicates no recidivism event and a 1 indicates that there was a recidivism event.

revictimization cases that were initially labelled as medium risk to be false negatives too, arriving at a total FNR of 94.3 per cent. Overall accuracy (if we consider a medium risk case where there was revictimization as a false negative) is 91.4 per cent, which is very poor considering the rate of non-revictimization is 94.4 per cent. These statistics mean that if officers were to always classify an incident as standard risk, they would be making less classification errors, but this is clearly not an acceptable solution when the costliness of failing to identify at-risk victims is taken into account. Similar observations were made on the non-IPV data.

We calculated the AUC for the univariable logistic regression model that used officer grading to predict revictimization for the IPV and non-IPV data. This permitted an easier comparison between officer performance and the performance of predictive models built on the DASH questions discussed later. The resultant two univariable models are called the IPV and non-IPV officer grading models. They yielded an AUC on their respective holdout data sets of 0.544/0.543 (IPV/non-IPV). An AUC of 0.5 indicates that a model will rank a randomly chosen positive case above a randomly chosen negative case with a probability of 0.5, making the model a random predictor. Thus, an AUC of 0.544 indicated that the officer grading model was performing little better than random.

Due to data limitations, this was probably an underestimate of the AUC, so that the police may be performing marginally better than it seems. It could be argued that the low predictive performance observed was *because* police interventions prevented further incidents. For example, if an incident is classified as high risk, it is referred to MARAC. Referrals to MARAC may reduce the risk of serious harm although the extent to which they do so is not well understood (Steel *et al.* 2011; McLaughlin *et al.* 2014). If we assume MARAC is effective, the revictimization rate for index incidents categorized as high risk would have been higher had there been no intervention. However, the extent to which this is so is obscured because we cannot know how many of the MARAC referrals would have been revictimization cases had the intervention not taken place. Essentially, while our interpretation of the true positives in the high-risk category was straightforward (these were correctly identified as high risk insofar as serious harm did subsequently occur), the interpretation of the false positives in the high-risk category was more challenging. Some false positives were cases misclassified as high risk, whereas others were genuine high-risk cases and the intervention perhaps meant that serious harm was prevented. Even if we could know what the revictimization rate in the high-risk index incidents would have been had the MARAC referral not been made, *it would not change our estimation of police performance dramatically*. High-risk cases only accounted for 3.7 per cent/1.7 per cent (IPV/non-IPV data) of the observations, and *94.3 per cent/97.5 per cent of the serious harm revictimization cases were wrongly classified as either standard or medium risk*. Furthermore, the prevalence of serious harm revictimization is 5.6 per cent/1.8 per cent, indicating that officer risk grading is poorly calibrated for IPV data in that it is *underpredicting revictimization by 34 per cent* (1–3.7/5.6).

Thus, it is important not to overestimate this limitation. We were also able to adjust for a wide variety of post-call risk management actions to these situations (including referrals to MARAC) in analysis of a much smaller but richer set of data from a different police force. The introduction of these risk management actions on the predictive models had no impact whatsoever in improving their performance (Peralta 2015). Other authors have reported similar results elsewhere (Svalin 2018; Ward-Lasher *et al.* 2018).

In the current analysis, we could only adjust for disposals and charges rather than more detailed information on risk management actions. We included a categorical variable describing disposals along with officer grading as predictors in a logistic regression model to predict the revictimization outcome. The AUCs for the univariable model and the two-variable model were 0.544 and 0.545 in the IPV data and 0.543 and 0.549 in the non-IPV data, respectively. Charges/disposals did not improve our ability to predict revictimization. An interaction term between officer risk grading and disposals did not improve the models either. This is consistent with the broader literature that suggests a non-substantial impact of police actions on future domestic abuse revictimization.

How does each question influence perception of risk?

A possible explanation for the poor predictiveness of DASH risk is that officers give greater weight to the less useful items when deciding the level of victim risk. Figure 1 provides the normalized mutual information between officer grading and each question. For comparison, the equivalent is provided for the revictimization outcome. This gives an indication of the types of questions that are most influential in shaping the officer's risk perception. Key influencers are victim-described fear, and whether or not the victim reported that the abuse is getting worse.

For the purpose of predicting revictimization, this information (as currently measured) is less important. Figure 1 also shows that, overall and in each data set, no single question is strongly predictive of the *outcome*. Although these values are very low, due to the large number of observations 25/1 (IPV/non-IPV data) questions are statistically significant at the 0.001 level and 27/3 are statistically significant at the 0.05 level (significance testing based on the chi-squared-distributed G-statistic). The questions concerning criminal, substance use, and mental health history are the 'best' predictors, although still very poor.

Can a machine learning model based on the DASH questions outperform officers?

Although, individually, none of the questions are that informative (see Figure 1), the question remains as to whether they are collectively informative. We undertook this analysis using a selection of machine learning models, listed in the Methods section. The officer grading is based on the answers to the DASH questions, so when we compare the officer grading to the question models, we are essentially comparing a professional judgement model and machine learning models that are all based on the same data: the DASH questions.

Each machine learning method was applied to each of the IPV and non-IPV data sets, to predict revictimization using the 27 DASH questions. All models, including the performance of officer risk grading, are compared in terms of mean AUC on holdout data sets that were not used in *any* stage of the model building process (Figure 2).

Logistic regression outperformed the more complex machine learning techniques with the exception, in the IPV case, of gradient boosting, where performance of gradient boosting was essentially equivalent to logistic regression with forward feature selection, with a 0.005 difference between respective AUCs. When model performance is similar, the simpler, more transparent model is favoured. A more parsimonious model

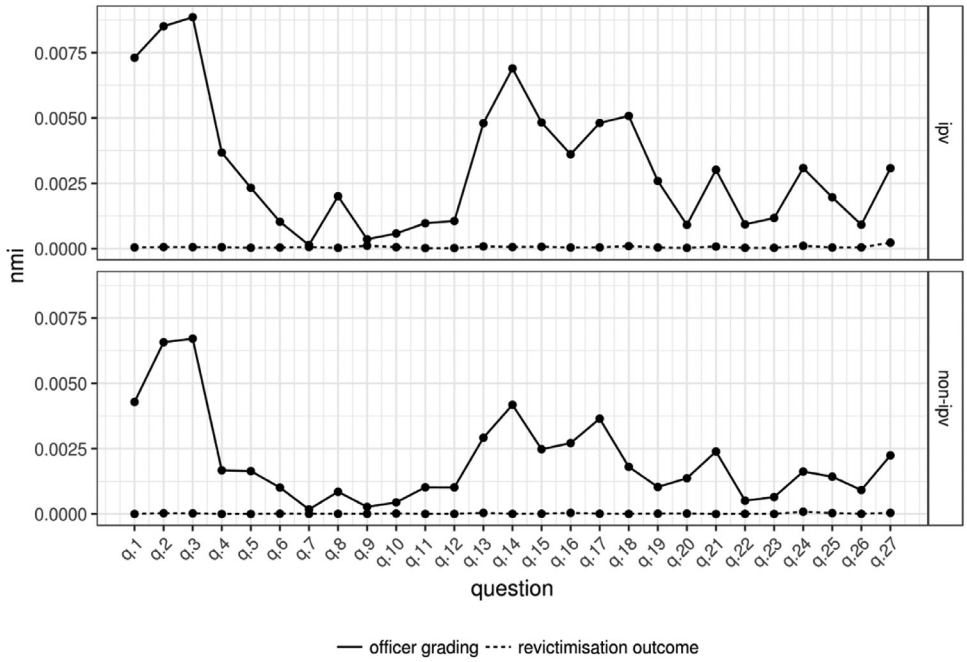

FIG. 1 For IPV and non-IPV data sets, the normalized mutual information between each question and officer grading (solid lines) and between each question and the serious harm revictimization outcome at 365 days (broken lines).

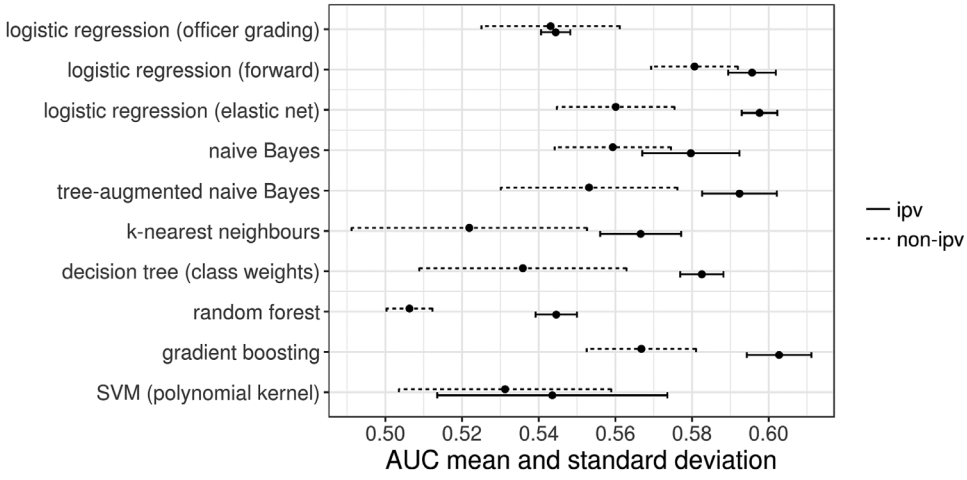

FIG. 2 Mean and standard deviation of holdout AUC for every model on IPV and non-IPV data.

is also favoured. Hence, we selected the forward feature selection method over elastic net for the logistic regression model for the IPV and non-IPV setting. We call these the IPV and non-IPV question models (as opposed to the *officer grading* models) from here on. The variable odds ratios for each question model are displayed in Table 4. The

Table 4 *IPV and non-IPV logistic regression odds ratios*

Description	Odds ratio	
	Yes	Not known
IPV data		
Q.1 Has the current incident resulted in injury?	1.081 (1.021–1.145)	0.875 (0.801–0.955)
Q.4 Do you feel isolated from friends/family?	1.156 (1.081–1.236)	0.809 (0.667–0.981)
Q.7 Is there conflict over child contact?	0.600 (0.556–0.648)	1.103 (0.904–1.345)
Q.8 Do they constantly text/call/ contact/follow/stalk/harass you?	0.861 (0.807–0.919)	1.333 (1.058–1.680)
Q.9 Are you currently pregnant or have you had a baby in past 18 months?	1.567 (1.485–1.652)	1.110 (0.932–1.323)
Q.10 Are there any children/step-children not belonging to the abuser, or other dependents, in the household?	1.207 (1.140–1.278)	0.873 (0.712–1.070)
Q.12 Have they ever threatened to hurt/ kill children/dependents?	0.598 (0.503–0.711)	0.683 (0.548–0.851)
Q.13 Is abuse happening more often?	1.179 (1.112–1.251)	1.204 (0.974–1.488)
Q.18 Have they ever attempted to strangle/ choke/suffocate/drown you?	1.278 (1.193–1.369)	0.888 (0.704–1.120)
Q.19 Do they do/say things of a sexual nature that make you feel bad or that physically hurt you or someone else?	0.912 (0.831–1.001)	1.677 (1.364–2.062)
Q.24 Have there been problems in past year with drugs/alcohol/mental health leading to problems in leading a normal life?	1.199 (1.139–1.261)	1.105 (0.931–1.312)
Q.27 Do you know if they have ever been in trouble with the police/has a criminal history?	1.629 (1.551–1.711)	1.209 (1.047–1.395)
Non-IPV data		
Q.1 Has the current incident resulted in injury?	0.662 (0.557–0.787)	0.760 (0.590–0.978)
Q.2 Are you very frightened?	1.307 (1.139–1.499)	0.828 (0.573–1.196)
Q.8 Do they constantly text/call/ contact/follow/stalk/harass you?	0.486 (0.366–0.647)	0.832 (0.480–1.442)
Q.10 Are there any children/step-children not belonging to the abuser, or other dependents, in the household?	1.348 (1.111–1.637)	2.453 (1.570–3.834)
Q.13 Is abuse happening more often?	1.363 (1.184–1.569)	0.986 (0.520–1.868)
Q.16 Have they ever used weapons/objects to hurt you?	1.990 (1.626–2.434)	1.256 (0.648–2.436)
Q.19 Do they do/say things of a sexual nature that make you feel bad or that physically hurt you or someone else?	1.733 (1.254–2.395)	2.308 (1.359–3.919)
Q.20 Is there another person who has threatened you or that you are afraid of?	0.290 (0.173–0.486)	2.520 (1.242–5.114)
Q.21 Do you know if they have hurt anyone else?	0.723 (0.596–0.878)	0.089 (0.045–0.176)
Q.24 Have there been problems in past year with drugs/alcohol/mental health leading to problems in leading a normal life?	1.581 (1.392–1.795)	3.509 (2.317–5.315)
Q.27 Do you know if they have ever been in trouble with the police/has a criminal history?	1.229 (1.084–1.394)	0.422 (0.283–0.629)

probability of revictimization for the reference level, where the victim answers 'no' to every question, was 0.036/0.013 (IPV/non-IPV). A total of 12 variables were retained in the IPV model build and 11 in the non-IPV build.

Random forests and other black box algorithms have been gaining popularity in recidivism prediction (Berk *et al.* 2016); however, our results show that they are not a silver bullet. This data set has a very high degree of noise relative to signal, and in this setting, more complex models are unable to extract sensible insights. A high-bias low-variance model may be more suited to a data set with very low predictive information in the variables.

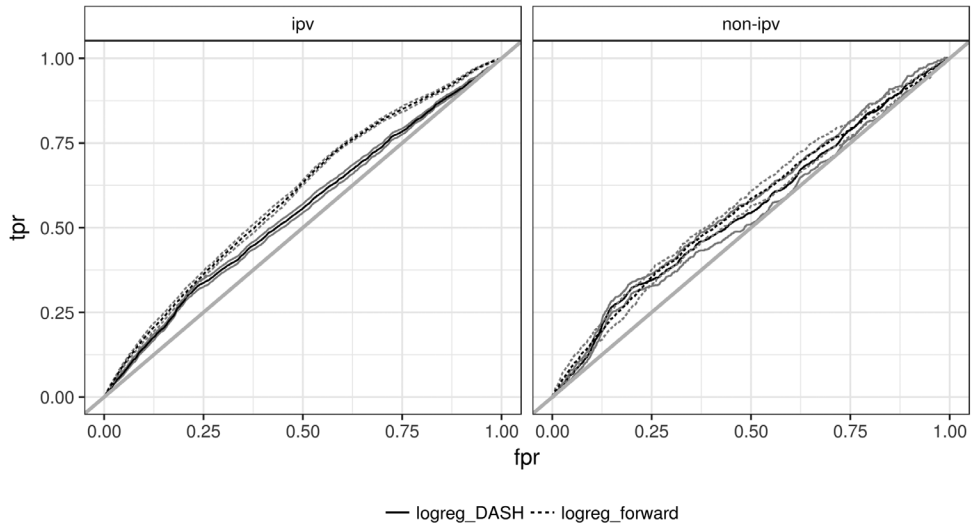

FIG. 3 ROC curves for the officer grading model (logreg_DASH) and best question model (logreg_forward). The grey lines represent the rate-wise standard deviation of TPR and FPR for every rate of event prediction.

The question model outperformed the officer grading model by small margin, with a difference of 0.051/0.038 (IPV/non-IPV) points in AUC. See Figure 3 for a comparison of the ROC curves of each model. In absolute terms, AUCs of 0.596/0.581 (IPV/non-IPV question models) indicate that we *did not achieve good discrimination between the revictimization and non-revictimization cases*.[4]

To gain a better understanding of what the difference in AUC between the grading and question models signify in the IPV case, we thresholded the predicted revictimization probabilities of the model to achieve standard- to medium- and high-risk classifications which could then be compared to officer performance. The standard and medium gradings were grouped together because our revictimization outcome represents the actualization of the high-risk forecast. The probabilities were thresholded such that the *distribution* of risk levels was the same as the distribution of officer grading in each of the data sets. In the IPV holdout data set, the risk grading distribution was 0.963/0.037 (standard to medium/high), so we classified the 3.7 per cent observations with the highest predicted probabilities of revictimization as high risk and the rest as standard to medium risk. Police predictions and model predictions were then compared in terms of TPR 0.086/0.108 (officer grading/question model), FPR 0.058/0.057, and precision 0.081/0.102. The question model made only modest improvements on the officer grading model. Consider that there were ~30,000 victims involved in identifiably IPV incidents in 2016. If the recidivism rate is 6 per cent, then there are ~1,800

[4] This analysis is based on 'first' encounters with the police (see Data section for definition). Similar analysis was conducted on a set of index events that is not limited to 'first' times. If a perpetrator and victim are involved in multiple incidents together, one incident is randomly selected to be the index event for that perpetrator–victim pair. On this second set of index events, the AUC on the officer grading is 0.569/0.563 (IPV/non-IPV), and for the question model, it is 0.613/0.580 (IPV/non-IPV). So similar conclusions are drawn with respect to this data sample: (1) the question models outperform officer grading; however, (2) the question models are still not discriminating well between revictimization and non-revictimization cases.

cases of serious harm recidivism within a year of the index event. The difference in TPR between officer grading and question model amounts to ~40 (30,000 * 0.06 * (0.108 – 0.086)) more cases accurately being identified as high risk if the questions model is used. The improvement in FPR means that there will be ~28 (30,000 * 0.94 * (0.058 – 0.057)) less non-recidivism cases misidentified as high risk.

Table 5 displays the ranges of AUCs for the question models on the IPV and non-IPV data for each of the 24 outcomes described in the Data section. There was no great improvement when the answers to the questions were used to predict recidivism (perpetrator involvement in a new domestic violence incident with same *or another* victim) instead of revictimization. Concerned that the predictiveness of the question answers decays the longer after the index incident the recidivism incident occurs, we also evaluated model performance on the revictimization and recidivism outcomes with time limits on the subsequent incident of 30, 90 and 180 days post-index and found no uplift in performance for the shorter time limits.

Can DASH be shortened?

Only a subset of DASH questions featured in each of the final model builds. In terms of ROC curves on the test data sets, there was negligible difference between the more parsimonious models and those which retain all 27 variables. It is ill-advised to use automatic variable selection as the only means of choosing variables for a model that is to be implemented in the field. However, it is adequate for our purpose, which is to demonstrate that not all of the questions are required to predict revictimization.

In the IPV setting, the strongest positive predictors of revictimization were 'yes' responses to questions 27, and 9, concerning abuser's criminal history, and victim's recent pregnancy; and 'not known' responses to question 19, on abuser's comments or behaviour of a sexual nature, and question 8, concerning controlling behaviour of the abuser. On the other hand, if a victim responds 'yes' to questions 7 or 12, regarding conflict over child contact, and whether the abuser has attempted to hurt the children or dependents, our model predicted that recidivism was *less likely* to occur.

In the non-IPV setting, both 'yes' and 'not known' responses to questions 19 and 24 are the stronger predictors of revictimization in the model, along with 'not known' responses to questions 10 and 20, concerning children or dependents, and threats made to the victim. Questions concerning victim fear (questions 2 and 3) were excluded from

Table 5 *AUC ranges across all time limit definitions (30, 90, 180, 365 days until subsequent incident), relating to all outcomes for the officer grading model, and logistic regression model with forward feature selection*

Revictimization or recidivism	Outcome harm level	IPV		Non-IPV	
		Officer grading	Question model	Officer grading	Question model
Revictimization	Incident	0.524–0.526	0.578–0.602	0.521–0.524	0.588–0.598
Revictimization	Any crime	0.537–0.547	0.608–0.635	0.547–0.551	0.597–0.616
Revictimization	Serious crime	0.537–0.546	0.575–0.596	0.543–0.586	0.517–0.581
Recidivism	Incident	0.525–0.528	0.585–0.600	0.518–0.524	0.603–0.614
Recidivism	Any crime	0.541–0.547	0.613–0.635	0.534–0.538	0.623–0.630
Recidivism	Serious crime	0.542–0.547	0.585–0.600	0.538–0.550	0.571–0.606

the IPV predictive model, and only weakly influence predictions in the non-IPV model. From Figure 1, we noted that these were the main influencers in officer risk perception.

Discussion and conclusion

There is no question that the introduction of risk assessment in the context of domestic abuse policing was a landmark moment in the development of responses to this phenomenon. There is widespread support for some form of risk assessment from both police and other stakeholders (Robinson *et al.* 2016). Yet, there is also an ongoing debate about the broader criminology of risk, the inherent limitations of any particular risk assessment regime, and whether we are misplacing our energies by emphasizing these practices (Walklate 2018). Our article recognizes the relevance of these debates but has a more specific but equally relevant focus. It simply aims to ask how effective are the existing tools to do what they were designed to do: identify high-risk victims.

Our results suggest that DASH is not enabling police officers to identify high-risk revictimization or recidivism cases. The risk assessment involves a balancing act where 27 pieces of information must be taken into consideration. We have shown that the answers to some questions have a large influence on this decision. Officers focus on the characteristics of the immediate domestic violence incident (see also Robinson *et al.* 2018b), yet these have been shown to be poorer predictors of recidivism. In our data, police performance in the identification of high-risk cases is little better than random and, for every recidivism and revictimization definition, is outperformed by the logistic regression model with forward feature selection. Police officers are underpredicting revictimization by a large margin. This is consistent with analysis that suggests that 'officers neither situated individual incidents in the context of coercive and controlling behaviour, nor recognized the tendency of some victims to minimize the abuse they were suffering' (Robinson *et al.* 2018a).

Some question answers that are less influential in officer risk perception are more predictive of recidivism; however, these are not, in absolute terms, good predictors (see Figure 1). Given the lack of signal in the data, the question models cannot identify new domestic abuse cases very well either. Correspondingly, the poor performance of officers is attributable, at least in part, to the fact that the tool they are working with is not performing as expected.

There are a few explanations for why DASH questions are not as predictive as expected. For one, it could be we are focussing on the wrong risk factors. On the other hand, the problem could lie at the point when the data are collected. As highlighted elsewhere 'responding to a call for service is … an often rushed and stressful endeavour, not always the best setting for establishing the rapport necessary for securing a full disclosure to sensitive questions' and 'the officers and citizens involved in these interactions are often encountering each other from very different gender, ethnic, and professional vantage points', which is likely to produce an 'endless combination of misunderstandings, judgement errors, and procedural mistakes' (Medina Ariza *et al.* 2016: 342). As HMIC reports, domestic homicide reviews and qualitative research on the topic have repeatedly documented that the quality and consistency of data gathering for risk assessments has been an ongoing problem. Research also continues to suggest that some officers display 'pejorative attitudes' and a lack of understanding of the dynamics of coercive control that may contribute to poor data gathering (Robinson

et al. 2018a). It is, therefore, probable that the reason for the poor predictive performance is linked to measurement error resulting from this. Indeed, measurement error in the features used in a predictive model is known to severely degrade our ability to accurately classify (Kuhn and Johnson 2013).

Although, for most of the questions, it is difficult to understand the extent to which the correct answers are being recorded, the answers to question 27, regarding perpetrator's criminal history, can be compared to something closer to the ground truth by using the criminal charges history available to police. In our analysis, we saw that when the form includes a 'yes' to question 27, this concurred with the police records 92.0 per cent of the time. On the other hand, if the answer is 'no', this contradicted police records 52.1 per cent of the time, and if the answer was 'not known', the perpetrator had a criminal history 75.8 per cent of the time. Note that the police records in our data are also an underestimation of the true proportion of perpetrators that have a criminal history because they only cover the metropolitan area. This highlights the problematic nature of seeking information from victims of domestic violence in what often are less than optimal circumstances, and the lack of cross-validation on the part of officers against existing records despite established policy requiring background and intelligence checks. Our preliminary analysis suggests that including a more accurate measurement of criminal history in a model significantly improves predictive performance. In the short term, all forces should put systems in place to ensure proper criminal background checks are done when DASH assessments are carried out.

Can this measurement error, more generally, be minimized? In theory, officers training could and should help. But it would need to be carefully crafted, thoroughly evaluated in the field in real-life conditions, and financially costly. The current suboptimal format, questionable content and often ambiguous wording of DASH, in addition to the situational attributes of police–victim encounters, significantly limits the effectiveness of training. Therefore, we believe our findings suggest the need for continued experimentation and evaluation with alternative tools such as those being tested by the College of Policing or, for other related purposes, the PPIT. Prediction quality might be improved by incorporating other information sources that do not rely on victim's statements. In a different setting, Dressel and Farid (2018) documented how a simple linear predictor with only two features (age and criminal history) was nearly equivalent to a well-known risk assessment tool (COMPAS) with its 136 features. We are working on using information other than DASH to explore this.

We are reporting here average effects. Preliminary analysis suggests there is significant variability across officers. Some officers seem to be better than others. It is not just that these 'good' officers were making better predictions. They were also working with better data insofar as the question answers that these officers received were more predictive of the outcome (based on bivariable mutual information analysis) than was the case for the officers performing less well. So the good officers might be better at speaking to victims and getting correct answers from them. More work is needed to understand the features and practices of these officers.

It is also very important to flag the many problems with our definition of harm. We believe it is fine, even important, for the police to use definitions of harm such as the one used in this study *as one* of many indicators for thinking about policing domestic abuse. At present, it is the only good indicator for which we have acceptable data to validate predictive models and preventing serious physical harm is indeed important.

But it would be highly questionable to think that these indicators of harm are complete, as many recent scholars seem to be doing. *They are not*, particularly in the context of domestic abuse. Measures such as the Cambridge Crime Harm Index or ONS harm index heavily weight physical harm that has been reported to the police. We know an important feature of domestic abuse and coercive control in particular is unmeasured harm. Placing such an unnuanced emphasis on recorded physical harm to evaluate *the policing* of domestic abuse (beyond the validation of predictive models) risks the continued neglect of unmeasured and less visible harm (Robinson *et al.* 2018a).

Algorithmic decision making has the potential to be a force for good. However, our findings suggest the necessity for a cautionary approach. If the data quality is poor, the prediction quality cannot be redeemed by the modelling approach, whether it be a basic logistic regression model or a state-of-the-art neural network. It is 'garbage in, garbage out', as we say in the field of machine learning. We not only encountered measurement error in recording the answers to DASH questions. We were also unable to measure true revictimization/recidivism rate. The revictimization and recidivism outcomes that were created from the data are more accurately described as revictimization/recidivism *where the police were involved*. Thus, our model is limited to predicting a type of recidivism that is serious enough to warrant police involvement and, indirectly, to predicting a type of victim that is willing to get the police involved—arguably, those experiencing coercive control are less likely to do this.

Another modelling challenge arises from the fact that there is heterogeneity in domestic violence incidents, abusers and victims. The problem of uniformity in risk assessment tools has been discussed by the more critical literature, and, clearly, a 'one size fits all' approach seems misguided (Walklate 2018). If there are subpopulations with contrasting characteristics in the data, then it may not make sense to apply the same model to them. Arguably, there are important subgroups within the IPV group and, similarly, within the non-IPV group. The incorporation of domestic violence and offender typologies into a predictive model could lead to better, subgroup-specific predictions.

It is also highly likely that model predictions vary based on protected characteristics, such as race and social demographics, leading to biased, unfair treatment of some subgroups (Berk *et al.* 2018). Were our models to be deployed by the police, then considerations from the burgeoning body of work on algorithmic fairness, and the unavoidable fairness-accuracy trade-off, would become crucial (Feldman *et al.* 2015; Chouldechova 2016; Hardt *et al.* 2016; Berk *et al.* 2018; Zafar *et al.* 2017). However, since in our study we only seek to establish the level of information contained in the DASH questions, we do not pursue fairness evaluations in this article. Further work will incorporate additional information available to the police into predictive models, and where a model appears viable, the various fairness considerations will become relevant.

Our final models are transparent, in that it is clear from the odds ratios how each variable is affecting model output. However, it is conceivable that when more variables are incorporated into the model, a more complex modelling approach will achieve greater accuracy. In this case, Bayesian networks could provide better that representation of the subgroup-dependent variable relationships, without compromising on transparency. This not only serves the strong jurisprudential argument for algorithmic transparency (Oswald *et al.* 2018; Wachter and Mittelstadt 2019), but also facilitates sense-checking and the incorporation of expert input, which is important when there

are suspected failures in the data, such as the aforementioned disparity between the answers provided to question 27 and police records of perpetrator criminal history. Where black box methods are applied, their performance should be compared with that of simpler models, such as a logistic regression model or decision tree.

Risk assessment cannot be reduced to prediction. Issues such as safeguarding, past harm identification and safety planning needs ought to be considered when designing the questions officers need to be posing to victims. Identifying victims at high risk of victimization is part of the picture. And to the extent that it is we need to continue developing better models. But we also need to learn from the way machine learning is applied in medical settings to support clinicians decision making around diagnosis and correct treatment referrals. Given police difficulties identifying patterns of abuse and coercive control with less visible harm, the machine learning potential for 'diagnostic' purposes in this context is clear. In medical settings targeted decision support tools are used to enforce physician behaviours that are known to improve clinical outcomes (Gage *et al.* 2001; Wang *et al.* 2018). Theoretically, that could also be possible in this context, but that, of course, would require addressing the elephant in the room: previously identifying successful responses to domestic violence victims and perpetrators.

Funding

This work was supported by the Economic and Social Research Council (grant number ES/M01178X/1).

Acknowledgement

We thank Dave Gadd, Amanda Robinson and Andy Myhill for the useful recommendations they provided on earlier drafts.

References

Akaike, H. (1998) 'Information Theory and an Extension of the Maximum Likelihood Principle'. In: Parzen, E., Tanabe, K., Kitagawa, G. (eds) *Selected Papers of Hirotugu Akaike. Springer Series in Statistics (Perspectives in Statistics).* Springer, New York, NY.

Almond, L., McManus, M., Brian, D. and Merrington, D. P. (2017), 'Exploration of the Risk Factors Contained Within the UK's Existing Domestic Abuse Risk Assessment Tool (DASH)', *Journal of Aggression, Conflict and Peace Research*, 9: 58–68.

Barnham, L., Barnes, G. and Sherman, L. (2017), 'Targeting Escalation of Intimate Partner Violence', *Cambridge Journal of Evidence-Based Policing*, 1: 116–42.

Barrett, B. J., Peirone, A., Cheung, C. H. and Habibov, N. (2017). Pathways to Police Contact for Spousal Violence Survivors: The Role of Individual and Neighborhood Factors in Survivors' Reporting Behaviors. *Journal of Interpersonal Violence.* doi:10.1177/0886260517729400

Berk, R., He, Y. and Sorenson, S. (2005), 'Developing a Practical Forecasting Screener for Domestic Violence Incidents', *Evaluation Review*, 29: 358–83.

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*, 1–42. doi 10.1177/0049124118782533

Berk, R., Sorenson, S. and Barnes, G. (2016), 'Forecasting Domestic Violence: A Machine Learning Approach to Help Inform Arraignment Decisions', *Journal of Empirical Legal Studies*, 13: 94–115.

Bland, M. and Ariel, B. (2015), 'Targeting Escalation in Reported Domestic Abuse: Evidence From 36,000 Callouts', *International Criminal Justice Review* 25: 30–53.

Campbell, J. C., Webster, D. W. and Glass, N. (2009), 'The Danger Assessment: Validation of a Lethality Risk Assessment Instrument for Intimate Partner Femicide', *Journal of Interpersonal Violence*, 24: 653–74.

CAADA (2012), *'CAADA-DASH Risk Identification Checklist (RIC) for MARAC Agencies'*, Coordinated Action Against Domestic Abuse.

Chalkley, R. and Strang, H. (2017), 'Predicting Domestic Homicide and Serious Violence in Dorset', *Cambridge Journal of Evidenced-Based Policing*, 1: 81–92.

Chouldechova, A. (2016), 'Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments', *Big Data*, 5: 1–17. doi 10.1089/big.2016.0047.

Cohen, J. (1988), *Statistical Power Analysis for the Behavioural Sciences*, 2nd edn. Routledge Academic.

Coid, J. W., Yang, M., Ullrich, S., Zhang, T., Sizmur, S., Farrington, D. and Rogers, R. (2011), 'Most Items in Structured Risk Assessment Instruments do not Predict Violence', *Journal of Forensic Psychiatry & Psychology*, 22: 3–21.

Cover, T. M. and Thomas, J. A. (2005), *Elements of Information Theory*. Wiley.

Dieterich, T. G. (1998), 'Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms', *Neural Computation*, 10: 1895–23.

Dressel, J. and Farid, H. (2018), 'The Accuracy, Fairness, and Limits of Predicting Recidivism', *Science Advances*, 4: 1–5.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C. and Venkatasubramanian, S. (2015), 'Certifying and Removing Disparate Impact', in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '15)*. ACM, New York, NY, USA, 259–268.

Ferguson, A. G. (2017), *The Rise of Big Data Policing*. NYU University Press.

Gage, B. F., Waterman, A. D., Shannon, W., Boechler, M., Rick, M. W. and Radford, M. J. (2001), 'Validation of Clinical Classification Schemes for Predicting Stroke', *Journal of the American Medical Association*, 285: 2864–70.

Hanson, R. K., Helmus, L. and Bourgon, G. (2007), *The Validity of Risk Assessments for Intimate Partner Violence: A Meta-analysis*, 29. Tech. rep., Public Safety Canada

Hardt, M., Price, E. and Srebro, N. (2016), 'Equality of Opportunity in Supervised Learning', in D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, eds, *Advances in Neural Information Processing Systems* 29, 3315–3323.

Helmus, L. and Bourgon, G. (2011), 'Taking Stock of 15 Years of Research on the Spousal Assault Risk Assessment Guide (SARA)', *International Journal of Forensic Mental Health*, 10: 64–75.

Hilton, N. Z., Harris, G. T. and Rice, M. E. (2010), *Risk Assessment for Domestically Violent Men*. American Psychological Association.

Hilton, N. Z., Harris, G. T., Rice, M. E., Lang, C., Cormier, A. C. and Lines, K. J. (2004), 'A Brief Actuarial Assessment for the Prediction of Wife Assault Recidivism: The Ontario Domestic Assault Risk Assessment', *Psychological Assessment*, 16: 267–75.

HMIC (2014), *Everyone's Business: Improving the Police Response to Domestic Abuse*. HMIC.

Kropp, P. R., Hart, S. D., Webster, C. D. and Eaves, D. (1994), *'Manual for the Spousal Assault Risk Assessment Guide'*, Vancouver: The British Columbia Institute Against Family Violence.

Kuhn, M. and Johnson, K. (2013), *Applied Predictive Modeling*. Springer.

Kvalseth, T. O. (1987), 'Entropy and Correlation: Some Comments', *IEEE Transactions on Systems, Man, and Cybernetics*, 17: 517–19.

Lopez-Ossorio, J. J., Gonzalez-Alvarez, J. L. and Andres-Pueyo, A. (2016), 'Eficacia Predictiva de la Valoracion Policial del Riesgo de la Violencia de Genero', *Psychosocial Intervention*, 25: 1–7.

McLaughlin, H., Robbins, R., Bellamy, C., Banks, C. and Thackray, D. (2014), *Domestic Violence, Adult Social Care, and MARACs: Implications for Practice*. Manchester Metropolitan University.

Medina Ariza, J. J., Robinson, A. and Myhill, A. (2016), 'Cheaper, Faster, Better: Expectations and Achievements in Police Risk Assessment of Domestic Abuse', *Policing*, 10: 341–50.

Messing, J. T. and Thaller, J. (2013), 'The Average Predictive Validity of Intimate Partner Violence Risk Assessment Instruments', *Journal of Interpersonal Violence*, 28: 1537–1558.

Messing, J. T., Campbell, J. C., Wilson, J. S., Brown, S., Patchell, B. and Shall, C. (2014), *Police Departments' Use of the Lethality Assessment Program: A Quasi-Experimental Evaluation*. Tech. rep., National Criminal Justice Reform Service, Washington DC

Messing, J. T. and Thaller, J. (2013), 'The Average Predictive Validity of Intimate Partner Violence Risk Assessment Instruments', *Journal of Interpersonal Violence*, 28: 1537–58.

Northcott, M. (2012), *'Intimate Partner Violence Risk Assessment Tools: A Review'*, Tech. rep., ON: Department of Justice Canada, Ottawa.

Oswald, M., Grace, J., Urwin, S. and Barnes, G. C. (2018), 'Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and "Experimental Proportionality"', *Information and Communications Technology Law*, 27: 223–50.

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

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

Robinson, A. and Clancy, A. (2015), *Development of the Priority Perpetrator Identification Tool for Domestic Abuse (Project Report)*. Cardiff University.

Robinson, A., Myhill, A., Roberts, J. and Tilley, N. (2016a), *Risk-Led Policing of Domestic Abuse and the DASH Risk Model*. College of Policing.

Robinson, A., Myhill, A. and Wire, J. (2018a), 'Practitioner (Mis)Understanding of Coercive Control in England and Wales', *Criminology and Criminal Justice*, 18: 29–49.

Robinson, A., Pinchevsky, G. and Guthrie, J. (2018b), 'A Small Constellation: Risk Factors Informing Police Perceptions of Domestic Abuse', *Policing and Society*, 28: 189–204.

Rosenthal, R. (1994), 'Parametric Measures of Effect Size', in H. M. Cooper and L. V. Hedges, eds., *The Handbook of Research Synthesis*, 231–44. Russell Sage Foundation.

Saunders, J., Hunt, P. and Hollywood, J. (2016), 'Predictions Put into Practice: A Quasi-experimental Evaluation of Chicago's Predictive Policing Pilot', *Journal of Experimental Criminology*, 12: 347–71.

Sechidis, K., Calvo, B. and Brown, G. (2014), 'Statistical Hypothesis Testing in Positive Unlabelled Data', *ECML PKDD*, 3: 66–81.

Sherman, L. (2007), 'The Power Few: Experimental Criminology and the Reduction of Harm', *Journal of Experimental Criminology*, 3: 299–321.

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.

Steel, N., Blakeborough, L. and Nicholas, S. (2011), *Supporting High-Risk Victims of Domestic Violence: A Review of Multi-Agency Risk Assessment Conferences (MARACs)*. Tech. rep., Home Office

Svalin, K. (2018), *'Risk Assessment of Intimate Violence in a Police Setting'*, doctoral dissertation, Malmo University.

Thornton, S. (2017), 'Police Attempts to Predict Domestic Murder and Serious Assaults: Is Early Warning Possible Yet?', *Cambridge Journal of Evidence-Based Policing*, 1: 64–80.

Wachter, S. and Mittelstadt, B. D. (2019), 'A Right to Reasonable Inferences: Re-thinking Data Protection Law in the Age of Big Data and AI', *Columbia Business Law Review*, forthcoming, available online at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3248829.

Walby, S., Towers, J. and Francis, B. (2016), 'Is Violent Crime Increasing or Decreasing? A New Methodology to Measure Repeat Attacks Making Visible the Significance of Gender and Domestic Relations', *British Journal of Criminology*, 56: 1203–34.

Walklate, S. (2018), 'Criminology, Gender and Risk: The Dilemmas of Northern Theorising for Southern Responses to Intimate Partner Violence', *International Journal for Crime, Justice and Social Democracy*, 7. doi: 10.5204/ijcjsd.v7i1.444. https://livrepository.liverpool.ac.uk/3013023/1/01_Walklate%20proof_27%20Nov%202017.pdf

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., Liu, S., Zeng, Y., Mehrabi, S., Sohn, S. and Liu, H. (2018), 'Clinical Information Extraction Applications: A Literature Review', *Journal of Biomedical Informatics*, 77: 34–49.

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.

Williams, K. R. and Houghton, A. B. (2004), 'Assessing the Risk of Domestic Violence Reoffending : A Validation Study', *Law and Human Behaviour* 28: 437–455.

Woolf, B. (1957), 'The Log Likelihood Ratio Test (the G-Test)', *Annals of Human Genetics*, 21: 397–409.

Wong, T. and Hisashima, J. (2008), 'Domestic Violence Exploratory Study on the DVSI and SARA'. Tech. Rep. October, Honolulu: Hawaii State Department of Health.

Ybarra, L. and Lohr, S. (2002), 'Estimates of Repeat Victimization Using the National Crime Victimization Survey', *Journal of Quantitative Criminology*, 18: 1–21.

Zafar, M. B., Valera, I., Rodriguez, M. G. and Gummadi, K. P. (2017), 'Fairness Constraints: Mechanisms for Fair Classification', in Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54.

Zou, H. and Hastie, T. (2005), 'Regularization and Variable Selection via the Elastic Net', *Journal of the Royal Statistical Society Series B Statistical Methodology*, 67: 301–20.