ESRC SDAI Application - Understanding domestic abuse using Big Data

1 Introduction

Domestic abuse is a complex phenomenon affecting people from all walks of life. It is increasingly recognised as a major public policy concern in many countries, including the UK (Prpic & Rosamund, 2018). While anyone can become a victim of domestic abuse, women are disproportionately affected, with more than 25% of women, and 15% of men in England and Wales reported to have experienced some form of domestic abuse since the age of 16 (Office for National Statistics, 2018). In the most extreme cases, domestic abuse can culminate in domestic homicide: in the period between April, 2017 and March, 2018, 70 people in England and Wales were killed by their current or former partner, and 90% of these victims were women (Office for National Statistics, 2019a). Domestic abuse has substantial mental health implications not only for the direct target of abuse (Ferrari et al., 2014), but also for children witnessing it (Callaghan, Alexander, Sixsmith, & Fellin, 2018). The economic cost of domestic abuse in England and Wales between April, 2016 and March, 2017 was estimated to be £66 billion, taking into account various components including missed workdays, health care costs, and reduced quality fo life for survivors (Oliver, Alexander, Roe, & Wlasny, 2019).

This research programme consists of four research projects that will deliver applicable, policy-relevant insights about domestic abuse. First, we will explore the demographic and socio-economic characteristics of victims of domestic abuse, and the predictors of suffering serious harm. Second, we will examine the reasons behind not just under-reporting from victims, but also police mis-recording of domestic abuse cases, and victim's perception of the police after reporting an incident. Third, we will examine the long-lasting effects of being a victim of domestic abuse or witnessing domestic abuse at home as a child. Finally, we explore the association between exogenous, time-varying factors, such as changes in alcohol consumption and financial resources, and the prevalence of domestic abuse.

2 Data sources

The research projects detailed below draw on various data sources. All of the projects rely on the CSEW, an annual, cross-sectional representative survey collecting information on victims of crime across England and Wales, including 30,000–40,000 households every year (Office for National Statistics, 2019b). The survey has a self-completion module containing questions about the respondent's experiences of domestic abuse throughout their life (since the age of 16), and in the past 12 months in particular. Given the serious problem of underreporting, one of the unique characteristic of domestic abuse, the CSEW is the most reliable source of information on the prevalence of domestic abuse in England and Wales. We have access to this dataset via the UK Data Service.

Some of the projects also benefit from an anonymised crime dataset provided by the West Midlands Police (WMP; the third largest police force in England, serving a population of 2.9 million; Home Office, 2018b), which includes all crimes and incidents recorded by the police force between 2010 and 2018, of which about 31% are domestic abuse-related. For each record in this dataset, we know the exact location, time, the gender, age, ethnicity and address of those involved, and the severity of the injury. Each person has a person identifier, allowing us to follow people over time. Access to this dataset is highly sensitive and strictly controlled. All researchers on this project have already been vetted by WMP and have been granted access to the required data fields. The PI, as data controller for the project is vetted to level 3, and the CoI's to level 2.

These two datasets have different merits. The CSEW provides us with rich demographic and socio-economic information on the respondent, and we can construct a control group from those individuals who have not experienced domestic abuse. However, it does not contain any information on the perpetrators, and has limited longitudinal information. In contrast, the WMP dataset has information on both the victim and the offender, and records all reported incidents for the same victim-offender pair, allowing us to analyse temporal patterns, and identify high-risk times of the year. We will not link the individuals in these datasets at any time.

One proposed project relies on a dataset containing credit card and current account spending of 20 million customers across the UK. We have access to this data as part of a data sharing agreement with one of the "Big 4" UK banks. As we know which neighbourhood (LSOA) the card owners live in, this data will allow us to construct a neighbourhood-level measure of spending on alcohol, gambling, and benefit payments throughout time. At no point will individual data be used, only geographic, LSOA level, aggregates. Finally, in analyses that rely on neighbourhood-level characteristics, we will use the Indices of Multiple Deprivation (IMD; Smith et al., 2015). The IMD ranks all 32,844 LSOAs in the UK based on their relative deprivation on various domains, including income, employment, education, health, crime, housing and living environment, providing us with a complex descriptor of the socio-economic characteristics of each neighbourhood in the country.

3 Research Programme

This research programme addresses four overarching topics, exploring different aspects of domestic abuse. Drawing on the extensive, rich respondent information in the CSEW, and the unique, temporal nature of the WMP data, this research complements, extends and deepens our understanding of domestic abuse. In outlining the analysis plan for each of these projects, we will specify a set of statistical analyses we plan to conduct. We will preregister all of our analysis plans and complement all analyses with robustness checks, and further exploratory regressions as the data allows (see, ***Trendl Football***, for an example of this approach in some of our prior work on domestic abuse).

1. Exploring the characteristics of survivors and the predictors of serious harm

Add note at the end of each section on publication plan, suggesting specific journal titles, (e.g., PNAS, Nature Human Behaviour, Psychological Science, some domain specific journals.

In this research project, we will first use the CSEW to investigate the demographic and socio-economic characteristics of survivors of domestic abuse. Exploring the risk factors in domestic abuse victimisation is key in designing effectively targeted policy measures. Previous research has identified a number of risk factors predicting domestic abuse victimisation (being female, young, unemployed, separated or divorced, living in single parent household and earning in the lowest income bracket), based on descriptive statistics derived from the CSEW data (Office for National Statistics, 2018). However, the CSEW provides a much broader range of information on the characteristics of the respondent, including age, sex, marital status, number of children, ethnicity, education, employment, income, benefit history, physical and mental health, frequency of going out, house and car ownership, self-reported well-being, and frequency of drug and alcohol use.

Investigating the explanatory power of this rich set of victim characteristics within one statistical model will provide us with a deeper understanding of the relative importance of these factors in predicting domestic abuse victimisation, compared to a purely descriptive statistics approach. We are also interested in how the predictive power of victim characteristics vary with the type of abuse suffered (physical abuse, threats, sexual abuse).

However, given the complex nature of domestic abuse, it is likely that there are important interactions between various victim characteristics in predicting the risk of domestic abuse victimisation (e.g., we would expect that not having stable employment is a more significant risk factor of domestic abuse victimisation for those who have young children). To uncover the structure of these interactions without having to a priori specify them – as is required in a logistic regression model – we will also use a random forest classification algorithm, a machine learning method that has the ability to detect non-linear relationships between variables, allowing us to identify particularly vulnerable subgroups of the population. Insights from both methods will provide us with the first extensive characterisation of the risk factors predicting domestic abuse victimisation in England, using a national-level, large representative sample like the CSEW.

Do we know enough about random forest classification to be confident about it's appropriateness? Could an expert reviewer find a reason to complain about it or criticize it's use? I ask because I don't know it well at all.

Table 1: Exploring the characteristics of survivors, and understanding the police response to reported cases, analysis plan

Research question	Dataset	Unit of analysis	Outcome variables	Explanatory variables	Model
What are the characteristics of domestic abuse victims?	CSEW	Individ-ual	Domestic abuse victimisation; Victimisation by type (physical, sexual abuse, threats)	demographic and socio-economic characteristics	Logistic regres-sion/random forest
What victim characteristics predict serious harm?	CSEW	Individ-ual (respondent-level)	Level of harm resulting from domestic abuse victimisation (no physical harm, minor, serious)	financial resources, mental health, dependence, social isolation, separation, feeling scared	Multinomial logistic regression
What perpetrator characteristics predict serious harm?	WMP	Perpetrator-victim pairs	Violent offences resulting in injury	previous self-harm and suicide attempts, violent and drug offences, breach of court orders, separation	Logistic regression with perpetrator-victim FE

We will also explore the victim and offender characteristics that predict serious harm. Insights from these analyses can inform decisions about the optimal timing and target audience for domestic abuse awareness campaigns to protect victims. Previous studies investigating risk factors for serious harm mostly relied on police data and focused on identifying high-risk offenders (Thornton, 2017).

Using a multinomial logistic approach, the CSEW data allows us to investigate the effects of financial independence and resilience (income bracket, employment stability), mental health (self-reported well-being), drug and alcohol dependence (frequency of usage), social isolation (self-reported frequency of going out, living alone or with the perpetrator, family disputes), recent separation, and feeling frightened (increased home security, reporting feeling scared). In the CSEW, the level of harm is identified by victims indicating that they required medical attention, or time off work.

Using the WMP data, we will also explore the risk factors associated with the perpetrator. The strength of the WMP data is that it has a temporal dimension, and we can follow the same perpetrator-victim pair over time. Using a logistic regression approach with perpetrator-victim fixed effects (FE), we will examine whether acts of serous harm can be predicted by previous self-harm and suicide attempts (which required police attendance), previous violent and drug offences, breaches of court orders, co-habitation with, and separation from the victim.

2. Understanding the decision to report and police mis-recording of domestic abuse

Short intro about aims, highlighting added value. In this project, we will first explore victims reporting behaviour. We will also use both the CSEW and the WMP dataset to explore the predictors of the decision to report and escape the abuse. We are not aware of previous research exploring this question in depth. Using the CSEW and logistic regression approach, we will investigate how the decision to report (or seek external help) depends on the individual-level, time-invariant characteristics of the survivor. Potential factors are trust in the justice system, social ties (living with family, time living in the area, member of Neighbourhood Watch), financial independence (own income, employment status, car ownership), number of children, mental health, and alcohol and drug dependency. The WMP data complements this with time-varying factors. We will use this to examine what are the potential characteristics and triggers that enable a victim to report ongoing, long-term abuse. We are interested in whether these reports are more likely to occur after certain days of the year (e.g., birthday of the victim or perpetrator, Christmas, New Years, Valentines etc.). This is particularly important knowledge for police, so that they can efficiently target awareness campaigns to encourage reporting.

We will also use the CSEW to explore the police's response to reported cases of domestic abuse. Increasing victim's trust in the Criminal Justice System (CJS) through improving law enforcement response to reported incidents is key in encouraging victims to report domestic abuse and prevent further harm. A recent inspection reported that WMP have failed to record 25% of reported crimes that were domestic abuse-related (Her Majesty's Inspectorate of Constabulary and Fires & Rescue Services, 2019). In this project, we will explore whether the extent of crime mis-recording depends on victim characteristics (e.g., a crime reported by a victim with a mental illness, substance dependency or language difficulties). Using the CSEW, we will be able to tell whether the domestic abuse reported by the respondent is likely to amount to a crime according to the list of notifiable offences (Home Office, 2018a). A multinomial regression approach with police force FE allows us to examine the extent to which police mis-reporting is different across broad crime types (sexual, physical violence, harassment and stalking), police forces, and depends upon the socio-economic characteristics of the victim. Complementing this analysis with a mixed effects (ME) random forest approach also reveals the roles of interactions between individual and police force characteristics. Lastly, these analyses can also be applied to examine the effect upon victim satisfaction with the police, and their perception of their personal security after reporting a crime.

Detailed plan

Table 2: Predicting serious harm and understanding the decision to report, analysis plan

Research question	**Dataset**	**Unit of analysis**	**Outcome variables**	**Explanatory variables**	**Model**
How does the police response depend on victim characteristics?	CSEW	Individ-ual	Police action (no action, warning, arrest, charge)	demographic and socio-economic characteristics of the victim	Multinomial logistic regression with FE/ ME random forest
What predicts victims' trust in the police?	CSEW	Individ-ual	Satisfaction with police action, perception of security after reporting	police action, demographic and socio-economic characteristics of the victim	Multinomial logistic regression with FE/ ME random forest
What time-invariant (victim) characteristics predict the decision to report?	CSEW	Individ-ual	Reported abuse to the police	severity/type of abuse, trust in CJS, social ties, financial status, children, mental health, dependencies	Logistic regression
What time-varying characteristics predict the decision to leave?	WMP	Perpetrator-victim pairs	Reported ongoing abuse to the police	time of year, specific holidays, length of abuse, severity of last occurrence	Logistic regression

Core outcomes

A deeper understanding of the victim and perpetrator characteristics that predict serious harm, and the time-invariant and time-varying factors affecting the decision to report the abuse to the police.

3. The long-lasting effects of domestic abuse

Short intro about aims, highlighting added value. Domestic abuse has long-lasting adverse effects on victims and those close to them. In this research project, our is aim to explore some of these consequences on the direct victims of domestic abuse and the children who live in the same household.

Detailed plan First, we will consolidate and extend research on the effect of witnessing domestic abuse during childhood. The 2016 CSEW included a module asking respondents about whether they witnessed domestic abuse at home as a child. Descriptive statistics of this dataset revealed that those who reported to have witnessed abuse as a child were significantly more likely to report long-term health problems, and live in a single parent household as an adult (Office for National Statistics, 2016). We will extend this using propensity score matching, a statistical approach that allows for causal inferences to be drawn about the effect of witnessing abuse whilst also controlling for a range of confounds, such as economic deprivation. Do we have any info on economic deprivation of where they grew up, or just where they live now?

Second, we will explore how children's behavioural problems are affected by witnessing domestic abuse at home. The CSEW questionnaire for 10-15 year olds records information about experiences with bullying, carrying knives, gang membership, school truancy, learning difficulties and health outcomes, including drug use and drinking behaviour. By applying propensity score matching, we will explore the causal effect of living in an abusive household.

We will complement this analysis using data on young offenders from the WMP. We will identify young offenders whose home addresses has been linked with a domestic abuse incident in the recent past. It is not possible to predict whether witnessing domestic abuse increases the chance of becoming a young offender, but we can examine whether the type of crime they engage in. For example, whether, after controlling for controlling for the socio-economic characteristics of the neighborhood (using the IMD), such individuals are more likely to commit similarly violent acts than other young offenders.

Core outcomes The first UK-based quantitative exploration of the causal effect of domestic abuse and childhood abuse on socio-economic outcomes.

Table 3: The long-lasting effects of domestic abuse, analysis plan

Research question	Dataset	Unit of analysis	Outcome variables	Explanatory variables	Model
How does witnessing domestic abuse as a child affect socio-economic outcomes in adulthood?	CSEW	Individual (respondent-level)	Educational attainment, employment status and health-related outcomes, including domestic abuse victimisation	Witnessing domestic abuse	Propensity score matching
How does experiences of past (but not recent) domestic abuse affect socio-economic outcomes in adulthood?	CSEW	Individual (respondent-level)	Current health and employment outcomes	Degree and type of domestic abuse suffered	Propensity score matching
How does living a household with domestic abuse affect behavioural outcomes in childhood?	CSEW	Individual (respondent-level)	Bullying, involvement with gangs, school truancy, carrying knives, drugs and alcohol usage, learning difficulties, health	Degree and type of domestic abuse in household	Propensity score matching
Are young offenders from abusive households more violent?	WMP, IMD	Young offenders (individual)	Type of offence (property-related, public order offence, violent)	Household LSOA socio-economic characteristics, abusive household or not	Multinomial logistic regression

4. Environmental factors influencing the prevalence of domestic abuse

Short intro about aims, highlighting added value. In this research project, we will investigate some of the environmental predictors of domestic abuse. While individual-level characteristics are likely to be the most important predictors of domestic abuse victimisation, the environment around the victim can play a key role in either exacerbating, or alleviating domestic abuse.

Detailed plan The CSEW includes rich information about the individual's environment, including the interviewer's perception of the street (signs of rubbish, vandalism, general condition of houses etc.), as well as the respondent's connection and attitude towards the neighbourhood (time living in the local area, noisy neighbours, teenagers hanging around, vandalism, drunken and anti-social behaviour, drug trafficking, abandoned cars, police presence in the area etc.). These properties of the environment can be used to predict domestic abuse victimisation, whilst controlling for the socio-economic characteristics of the respondent. Additionally, using a spatial regression framework, we will explore how deprivation in a neighbourhood (LSOA level, measured by the IMD) affects the prevalence of domestic abuse and the willingness to report. Using this statistical approach will alleviate the problem of spatial dependency in our dataset, and will allow us to derive a precise estimate of the effect. These analyses will extend our understanding of the neighbourhood-level predictors of domestic abuse victimisation, and complement our individual-level approach outlined in the first project to provide a comprehensive exploration of the factors affecting domestic abuse victimisation and willingness to report.

We are also interested in exploring the effect of factors including average alcohol consumption, and levels of financial stress at the community level, rather than the individual (measured by gambling expenditure and benefit dependency). In investigating this question, we will use the transaction dataset described earlier, which allows us to estimate a spending profile for each LSOA over time. This way we can identify time-specific changes in alcohol consumption, gambling expenses and benefit receipts in parts of the West Midlands, and investigate their effect on the number of reported domestic abuse cases in the area (from the WMP), using a spatial Poisson regression approach. Can we not do this for the whole country using the CSEW? I know the timecourse isn't great, but we have the data from individual years.

We shall complement this analysis by identifying external events that may affect alcohol consumption (e.g., festivals, weather, bank holidays, sport tournaments) measured by spending on alcohol in an LSOA in a given time period, and investigate whether it affects the number of reported domestic abuse incidents in that area. Previously, our research has identified a 60% increase in alcohol-related domestic abuse when the England national football team won, highlighting the profound effect exogenous events can have on the propensity for violence.

To explore the effect of financial shocks on domestic abuse, we will be using the same estimation approach, but we will focusing on gambling and benefit payments. Gambling is increasingly recognised as a serious health concern across the UK, and can have a profoundly adverse impact on family finances. We will use our transaction data and WMP dataset to assess the link between changes in spending on gambling in a given LSOA, and the prevalence of domestic abuse in that area. The recent reforms of the UK benefit system, in the form of the introduction of the Universal Credit (UC) had been criticised widely, due to the temporal financial strain it imposes on the least financially resilient people in society. Drawing on previous findings about the link between financial stress and domestic abuse, we will explore what effect the roll-out of the UC across the West Midlands had on the reported number of domestic abuse cases. The fact that the exact date of the roll-out varied across the seven metropolitan boroughs within the county allows for a more precise estimation of the effect of UC on the reported number of domestic abuse incidents in the West Midlands.

Core outcomes The first extensive analysis of the neighbourhood-level predictors of domestic abuse within the UK. An exploration of the exogenous, time-varying factors, including alcohol consumption and financial stress (gambling, universal credit) affecting the reported number of domestic abuse cases in the West Midlands.

Table 4: Environmental factors influencing the prevalence of domestic abuse, analysis plan

Research question	**Dataset**	**Unit of analysis**	**Outcome variables**	**Explanatory variables**	**Model**
How do environmental characteristics of the neighbourhood predict domestic abuse?	CSEW	Individual (respondent-level)	domestic abuse victimisation/willingness to report	Perception of neighbourhood, socio-economic characteristics	Logistic regression
How does neighbourhood deprivation predict domestic abuse?	IMD, CSEW	LSOA-level	domestic abuse victimisation/willingness to report (proportion of overall population)	Various measures of deprivation	Spatial Logistic regression
How do temporal changes in alcohol consumption and income affect domestic abuse?	WMP, Credit card data	LSOA, day/week-level	Number of reported domestic abuse cases	Spending on alcohol and gambling, changes in benefit receipts	Poisson/Negative binomial regression with spatial effects

References

Callaghan, J. E. M., Alexander, J. H., Sixsmith, J., & Fellin, L. C. (2018). Beyond "Witnessing": Children's Experiences of Coercive Control in Domestic Violence and Abuse. *Journal of Interpersonal Violence*, *33*(10), 1551–1581.

Crown Prosecution Service. (2017). *Domestic abuse.* Retrieved from https://www.cps.gov.uk/domestic-abuse

Ferrari, G., Agnew-Davies, R., Bailey, J., Howard, L., Howarth, E., Peters, T. J., ... Feder, G. S. (2014). Domestic violence and mental health: a cross-sectional survey of women seeking help from domestic violence support services. *Global Health Action*, *7*(1), 25519.

Her Majesty's Inspectorate of Constabulary and Fires & Rescue Services. (2019). *Reports - Rolling programme of crime data integrity inspections.* Retrieved from https://www.justiceinspectorates.gov.uk/hmicfrs/publications/west-midlands-crime-data-integrity-re-inspection-2018/

Home Office. (2018a). *Home Office counting rules for recorded crime.*

Home Office. (2018b). *Police funding for England and Wales 2015 to 2019* (Tech. Rep.). Home Office. Retrieved from https://www.gov.uk/government/statistics/police-funding-for-england-and-wales-2015-to-2019

Office for National Statistics. (2016). *Abuse during childhood - Office for National Statistics.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/abuseduringchildhood/findingsfromtheyearendingmarch2016crimesurveyforenglandandwales/pdf

Office for National Statistics. (2018). *Domestic abuse in England and Wales: year ending March 2018.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/bulletins/domesticabuseinenglandandwales/yearendingmarch2018/pdf

Office for National Statistics. (2019a). *Homicide in England and Wales: year ending March 2018.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/articles/homicideinenglandandwales/yearendingmarch2018/pdf

Office for National Statistics. (2019b). *User guide to crime statistics for England and Wales.* Retrieved from https://www.ons.gov.uk/peoplepopulationandcommunity/crimeandjustice/methodologies/userguidetocrimestatisticsforenglandandwales/pdf

Oliver, R., Alexander, B., Roe, S., & Wlasny, M. (2019). *The economic and social costs of domestic abuse.* Retrieved from https://www.gov.uk/government/publications/the-economic-and-social-costs-of-domestic-abuse

Prpic, M., & Rosamund, S. (2018). *Violence against women in the EU: State of play* (Tech. Rep.). Retrieved from http://www.europarl.europa.eu/thinktank/en/document.html?reference=EPRS{_}BRI(2018)630296

Safe Lives. (2015). *Ending Domestic Abuse - Getting it right first time.* Retrieved from http://www.safelives.org.uk/policy-evidence/getting-it-right-first-time

Smith, T., Noble, M., Noble, S., Wright, G., McLennan, D., & Plunkett, E. (2015). *The English Indices of Deprivation 2015.* Retrieved from https://www.gov.uk/government/statistics/english-indices-of-deprivation-2015

Thornton, S. (2017). Police Attempts to Predict Domestic Murder and Serious Assaults: Is Early Warning Possible Yet? *Cambridge Journal of Evidence-Based Policing*, *1*(2-3), 64–80.