Look-alike modelling in violence-related research: a missing data approach

Estela Capelas Barbosa[1* #a], Niels Blom[2], Annie Bunce[2]

[1] Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom.

[2] Violence and Society Centre, City, University of London, London, United Kingdom.

[#a] Current address: Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, United Kingdom.

Abstract

Violence as a phenomena has been analysed in silo due to difficulties in accessing data and concerns for the safety of those exposed. While there is some literature on violence and its associations using individual datasets, analyses using combined sources of data are very limited. Ideally data from the same individuals would enable linkage and a longitudinal understanding of experiences of violence and their (health) impacts and consequences. However, in the absence of directly linked data, look-alike modelling may provide an innovative and cost-effective approach to exploring patterns and associations in violence-related research in a multi-sectorial setting.

We approached the problem of data integration as a missing data problem to create a synthetic combined dataset. We combined data from the Crime Survey of England and Wales with administrative data from Rape Crisis, focussing on victim-survivors of sexual violence in adulthood. Multiple imputation with chained equations were employed to collate/impute data from different sources. To test whether this procedure was effective, we compared regressions analyses for the individual and combined synthetic datasets on a binary, continuous and categorical variables. Our results show that the effect sizes for the combined dataset reflect those from the dataset used for imputation. The variance is higher, resulting in fewer statistically significant estimates. We extended our testing to an outcome measures and finally applied the technique to a variable fully missing in one data source. Our approach reinforces the possibility to combine administrative with survey datasets using look-alike methods to overcome existing barriers to data linkage.

Introduction

It has been established for over 20 years that violence is a complex social problem and a public health issue [1-3], with implications for the health and social care systems, police and

justice systems [4], as well as significant productivity losses for those who experience it [5, 6]. Analysing data collected by these systems can aid understanding of the problem of violence and how to respond to it. In social research, analysing administrative records together with survey data has already enabled better measurements of violence experiences, capturing experiences of both victim-survivors and perpetrators across multiple points in time and social and economic domains [7].

Although some violence-related research has been carried out using matched or combined emergency departments and police data [8-11], most studies in violence-related research analyse data in silo due to difficulties in accessing data and concerns for the safety of those exposed [12, 13]. Particularly, data from third sector voluntary specialist support services for victims or perpetrators of violence has, to our knowledge, not been linked or combined with other datasets, as these services are keen to provide person-centred trauma-informed care and fear that information on their service users may be used against them in courts or by immigration authorities [14, 15].

From an analytical viewpoint, ideally, data from the same individuals would enable linkage and a longitudinal understanding of experiences of violence and their (health and inequalities) impacts. However, given safety concerns, data on people who have experienced violence is often pseudonymised before being made available for researchers, meaning records across sectors pertaining to the same individuals cannot be linked. Look-alike profiling may provide an innovative and cost-effective approach to exploring patterns and associations in violence-related research in a multi-sectorial setting.

Look-alike modelling has been extensively used to identify similar and new costumer and consumer target groups in marketing, e-commerce and advertising [16-18]. We apply costumer look-alike principles to violence-related research. Our goal is to propose an innovative method for data integration in this particularly sensitive research area, to move

beyond silo analyses, which could also be used in other research areas with similar issues. Effectively, this method allows for integrating additional information into one dataset based on its distribution and associations in another dataset, creating a new (synthetic) dataset. This methodology could also be used in other fields of social and economic research, where issues regarding pseudonymisation and missing information are also present.

In this paper, we approached the problem of data integration and look-alike profiling as a missing data problem, although we acknowledge that several other approaches are possible. We combined data from the Crime Survey for England and Wales (CSEW) with administrative data from three Rape Crisis Centres (RCC) in England, which are part of a Rape Crisis England and Wales (RCEW), focussing on victim-survivors of sexual violence in adulthood, in line with the understanding that a benefit of linking administrative and survey data is the improvement in imputation methods to fill in missing values in surveys [19]. Multiple imputation with chained equations were employed to collate and integrate data from these two different sources producing a synthetic dataset.

Theoretical framework

In theory, look-alike modelling is based on the principle that similar individuals have similar behaviours. While in economics this normally refers to consumption behaviour, for people experiencing violence it refers to their trajectories and help-seeking behaviours. Therefore, to explore similarities between individuals, one needs to look at socio-economic and demographic variables, as well as violence experience. Mathematically, in two different datasets A and B, there are a_{ij} and b_{ij} individual records. These records can be compared in multiple variables k to ascertain how similar their look-alike profiles are. Each component-wise or variable-wise comparison relies on a vector $C_{i,j} = [\, c_1^{i,j} , c_2^{i,j}, \ldots , c_k^{i,j}]$ that effectively produces a *comparison function* looking at the values of the record component or variable k in the two records a_{ij} and b_{ij}. In order to approach this data integration problem as a missing data

problem, one relies on a sequence of univariate imputation models, with fully conditional specifications of prediction equations. Formally, for imputation variables $X_1, X_2, \ldots, X_p$ and complete independent predictors C, so that:

$$X_1^{(t+1)} \sim g_1(X_1|X_2^{(t)}, \ldots, X_p^{(t)}, C, \varphi_1)$$
$$X_2^{(t+1)} \sim g_2(X_2|X_1^{(t+1)}, X_3^{(t)}, \ldots, X_p^{(t)}, C, \varphi_2) \quad (1)$$
$$\ldots$$
$$X_p^{(t+1)} \sim g_p(X_p|X_1^{(t+1)}, X_2^{(t+1)}, \ldots, X_{p-1}^{(t+1)}, C, \varphi_p)$$

Where t are iterations that converge at t = T and φ_j are the corresponding model parameters prior [20]. In our study, we created the vector $C_{i,j}$ based on the following variables: type of sexual violence experienced (type of SV), relationship to the perpetrator, health impact, employment status, housing tenure, number of dependants, relationship status (usually referred to as marital status in social research), ethnicity, age and gender. These variables were selected as they are considered to influence the journey of victim-survivors of sexual violence and their help-seeking behaviour.

Traditionally, multiple imputation (MI) is used to address missingness of data by generating plausible values derived from distributions and relationships among observed variables [21]. While MI has been widely used in statistical and economic analysis of clinical trials [22] and more recently social research [23], to our knowledge, it has not been used to produce a synthetic dataset. Our multiple imputation approach to data integration recognises that the reason for missing data may be different for each dataset A and B. This is particularly true in our empirical application, since we are using a population-level survey (CSEW) and administrative records from a victim support service (RCEW). Furthermore, while datasets A and B are completely independent in our case, the reasons for missingness may be correlated, as disclosing sexual abuse is still stigmatised in society [24-26]. Finally, our approach

recognises that the variables or covariates used for imputation may have non-normal distributions [27, 28].

Procedurally, multiple imputation replaces each missing value with a set of plausible values. Following Bayesian rules, the imputed values are drawn based on the conditional distribution of the missing observations given the observed data, reflecting the uncertainty associated with the missing data itself and parameters estimated in the imputation model [29]. Mathematically, let f_{ij} represent the variable you are interested in imputing for the ith individual within the jth cluster. In this case, $C_{i,j} = [c_1^{i,j}, c_2^{i,j}, \ldots, c_k^{i,j}]$, the *comparison function* and D_j, the cluster-level vector of covariates, are the predictors of missingness in variable f at individual and cluster-levels respectively. Then, a MI model can be specified as:

$$f_{ij} = \beta^f C_{ij} + \gamma^f D_j + \varepsilon_{ij}^f \quad (2)$$

Where β and γ are the vectors of the regression coefficients corresponding to individual and cluster-level covariates. The model assumes that the error term (ε) is normally distributed with variance σ^2. The imputation procedure generates multiple values for each missing observation based on the distributions for β, γ and σ conditioned on observed data. By combining two datasets A and B, based on the vector C and using multiple imputation, we are applying a look-alike modelling approach that may enable imputation of partially and completely missing data into a complete combined synthetic dataset.

Methods

We aimed to test our proposed approach to data integration by combining survey data from the Crime Survey for England and Wales (CSEW) with administrative data from Rape Crisis England & Wales (RCEW), focussing on victim-survivors of sexual violence in adulthood. This research was reviewed and approved by the IMJEE (International Politics, Music, Journalism, Economics, and English) research ethics committee from City, University

of London (ETH2122-2023 and ETH2122-0299). Informed verbal consent regarding future use of their data for research was obtained by case workers from Rape Crisis centres while working with service users and recorded in their case management system, in line with their a non-intrusive approach to data collection whereby only what is appropriate is asked and/or what survivors choose to disclose is recorded [30].

Datasets

The CSEW, previously known as the British Crime Survey, is a nationally representative face-to-face victimisation survey of about 35 thousand to 46 thousand respondents per survey wave, which started biannually from 1982 before becoming an annual survey from 2001[31]. The CSEW asks people aged 16 and over about their experience with household and personal crimes in the twelve months prior to the interview. Considering our focus on sexual violence, we only included individual level data from respondents who had reported being a victim-survivor of rape, attempted rape, wounding with sexual motive, and indecent assault. In order to include a sufficient number of incidents of sexual violence to do the data integration, we used CSEW data from 2001 to 2020.

The RCEW data comes from three RCCs in a region in eastern England and is based on routinely collected administrative data recorded in a centralised case management data system between April 2016 and March 2020. Information is self-reported by victim-survivors upon initial contact with RCEW, most commonly over the phone but sometimes online or face-to-face, and data are inputted to the RCEW database by frontline support workers. Rape Crisis centres collect individual level data for their service users in pre-determined coding categories based on a person-centred non-intrusive principle, which means frontline workers only ask questions that are appropriate, or rely on information victim-survivors choose to disclose [30]. Information collected typically includes socio-demographic and protected characteristics (gender, age, disability, ethnicity, nationality, sexuality, religion, marital status,

accommodation, employment, language, immigration status, socioeconomic status), experiences of sexual violence and abuse (SVA), victim-perpetrator relationship, impacts from experience of SVA, risk level, referral routes, engagement with different (statutory and non-statutory) services and contact with the criminal justice system. Data on experiences of SVA are collected in two main ways; information is gathered on the 'presenting incident' (the main experience of violence the victim-survivor is seeking support for at the time of initial contact with RCEW), and elsewhere in the database further details can be entered under 'incident summary' about separate 'incidents' or experiences of violence, if disclosed [30]. Most information is inputted into their case management system at the point of intake based on the victim-survivor's report and, where necessary, the assessment of the support worker. However, further information on the abuse can be collected and recorded at any point during the support journey, as appropriate. Case management and criminal justice data are collected in separate parts of the system, however, data are recorded under a client identification number, making it possible to merge case management and criminal justice data.

Considering our focus on sexual violence, we selected respondents (CSEW) or service users (RCEW) who have reported being a victim-survivor of rape (including attempted) or another form sexual violence and abuse (which included indecent assault and wounding with sexual motive). We selected respondents/service users with no missing values on vector *C* variables, which led to a sample of 1,232 incidents from 1,111 individuals in the CSEW, and 6,102 referral cases from 5,333 individuals in RCEW. In RCEW data, it included data for individuals who accessed the service more than once.

The comparison vector (C) variables

As previously mentioned, we created the vector $C_{i,j}$ based on the variables that were considered to influence victim-survivors' journeys and help-seeking behaviour the most. Thus, we needed to harmonise the following variables across CSEW and RCEW data: type of sexual

violence experienced, relationship to the perpetrator, health impact, employment status, housing tenure, number of dependants, relationship status, ethnicity, age and gender.

The type of sexual violence experienced in the CSEW was categorised into crime codes by professional coders based on respondents' responses to survey questions and narrative description of the incident. The categories are aimed to align with Home Office categorisation. We selected the following reported offences: rape, serious wounding with sexual motive, other wounding with sexual motive, attempted rape, and indecent assault. We categorised these into rape (including attempted) or some other form of sexual violence. In the RCEW data, sexual violence was categorised based on the information recorded at intake under 'presenting incident' and 'incident summary'. Once again, we categorised these into broader categories: rape (as an adult, including attempted rape); and some other form of sexual violence (including sexual assault, assault by penetration, voyeurism, sexual bullying, penetration by object, gang related sexual violence, forced sexual activity in public, exposed to sexual images, sexual harassment and sexual exploitation). Victim-survivors accessing RCEW services for other types of violence or abuse were excluded, including rape or sexual abuse during childhood.

For the variable victim-perpetrator relationship, respondents to CSEW were first asked whether they knew the perpetrator, and if so, what their relationship was at the time of the incident. The RCEW data recorded who the primary perpetrator was. This was categorised into domestic (such as [former] intimate partner or family member), acquaintances (including friends, colleagues), and strangers. If multiple perpetrators were mentioned, it was coded as the closer relationship (e.g. prioritising domestic over acquaintances).

The health impact of the incident was assessed in the CSEW by whether they were bruised, scratched, cut or injured in any way as a result of the incident. The health impact was measured in the RCEW data using information recorded under 'incident impact' and 'impact summary', for which we included physical health impacts of memory loss, physical injuries

and body problems, gynaecological disorder, and sexually transmitted infection. While these do not match directly between the two datasets, we only included a binary in our empirical application for whether there was (yes/no) a health impact on the victim-survivor.

Relationship status was categorised into whether respondents were in a co-residential relationship (either married or cohabiting), single/non-resident partner/widowed, or separated or divorced in the CSEW and RCEW. Ethnicity was coded as White and non-White, as further differentiation led to too small numbers in some categories. However, we acknowledge that the ethnicity categorisation of White/non-White may be problematic and any conclusion in this respect, limited [32]. Employment status in both datasets was assessed by whether people were employed, unemployed, students, or outside the labour force (e.g. a homemaker, retired, or unable to work due to illness). Gender was asked as whether the respondent was male or female[1] in the CSEW. In RCEW data, more detailed responses are given, including transgender female and transgender male, which were recoded into men and women. Finally, age was measured numerically in both datasets and we included in our analyses people over the age of 16. Table S1 in the supporting information summarises how variables were harmonised.

Analytical Strategy

To test whether approaching look-alike modelling as a missing data problem was effective, we compared regression analyses for the two datasets (CSEW and RCEW) and the combined synthetic dataset, which imputed data based on the comparison vector. As a proof of concept, we tested the approach using variables of different types (binary and continuous) that are observed in both datasets. Formally, our approach had three steps. First, we specified the same linear (OLS) or logistic regression (as appropriate) for dataset A (RCEW) and dataset B (CSEW). We then assumed one variable was missing from the combined integrated synthetic

[1] We acknowledge that female / male are correct categories for sex not necessarily gender. But we used the categories as asked by the CSEW as proxies for women / men.

dataset by generating a completely missing variable for dataset A, which we imputed, using multiple imputation with chained equations, based on the observed values for the combination vector in both datasets. This effectively imputed the (assumed) missing variable in dataset A based on the distribution and associations with other variables of the combination vector in dataset B.

We carried out this exercise for four variables, two that are very similarly measured – age (continuous) and gender (binary), one that is differently measured across datasets – health impact (binary), and lastly, we illustrated the potential of this method of combining data in a real-life application to a variable that only appears in one dataset (CSEW) – frequency of abuse (count). We acknowledge that the first two tests, using age and gender, are not particularly interesting from an analytical standpoint. Nonetheless, we wanted to start off with variables that were objectively measured as much as possible.

Results

Profiles comparison of sexual violence victim-survivors in CSEW and RC

Before conducting our look-alike exercises, we compared the profiles of sexual violence victim-survivors in CSEW and RCEW datasets (Table 1). The table shows some meaningful differences between the individuals pertaining to each dataset. Particularly, only 32% of sexual violence victim-survivors in the CSEW had been victims of rape, compared to 71% in the RCEW data. Relationship to the perpetrator was more likely to be domestic in the RCEW data compared to the CSEW (48% *vs* 25%, respectively) and perpetrators were far more likely to be strangers or to be unknown in the CSEW (42%), compared to only 12% of records in the RCEW dataset.

Furthermore, the CSEW recorded a physical injury in 39% of incidents, while this appeared in only 6% of cases in the RCEW dataset, which might reflect the different

263 measurements of physical health impact between these two data sources. Finally, there are
264 some differences in relationship status and employment status, with more single or widowed
265 people in RCEW data and more separated or divorced people in CSEW, and more unemployed
266 people and students in RCEW when compared to CSEW.
267

Table 1. Descriptive statistics of sexual violence victim-survivors in the Crime Survey for England and Wales (CSEW) and Rape Crisis England & Wales (RCEW) datasets

	CSEW		RCEW	
	%	Mean (SD)	%	Mean (SD)
Type of sexual violence				
Rape	32.4		70.7	
Other sexual violence and abuse	67.6		29.3	
Victim-perpetrator relationship				
Domestic	24.6		48.2	
Acquaintance	33.2		40.2	
Stranger or unknown	42.2		11.6	
Physical health impact				
No injury	60.9		94.0	
Injury	39.1		6.0	
Gender				
Male	9.6		7.9	
Female	90.4		92.1	
Relationship status				
Married/Cohabiting	21.7		16.8	
Single/non-resident relationship/Widowed	57.0		71.7	
Separated/Divorced	21.4		11.6	
Ethnicity				
White	91.6		91.0	
Non-White	8.4		9.0	
Employment status				
Employed	56.4		35.2	
Unemployed	7.6		38.2	
Outside labour force	30.4		11.7	
Student	5.6		14.9	
Housing tenure				
Homeowner/lives in own home	34.6		41.8	
Renter	59.8		18.4	
Other	5.6		39.7	
Age		32.9 (12.4)		34.1 (12.9)
Nr of dependents		0.6 (1.0)		0.8 (1.2)
N	1,232		6,102	

Source: Crime Survey for England and Wales (2001-2020) and Rape Crisis England & Wales (2016-2020)

Look-alike empirical application

Our first empirical application exercise pretended the variable *age* was missing from the combined dataset. Thus, we stipulated our comparison vector (C) as:

$$C_{i,j}^{1} = f[\text{type of SV, perpetrator relationship, health impact, employment status, housing tenure, number of dependants, relationship status, ethnicity and gender}] \quad (3)$$

Table 2 presents the results of a linear regression (OLS), looking at the associations between age as the dependent variable, and the independent variables for dataset A (RCEW), dataset B (CSEW) and the complete combined dataset inputting age based on our proposed approach. When comparing the associations with age between the original datasets, and the imputed synthetic dataset based on the variation observed in B, the results show that the effect sizes and direction for the imputed data reflects the results from the dataset used as the basis for imputation. For example, the type of SV was not associated with age in the original RCEW, but was in the CSEW. The imputed synthetic dataset reflects the CSEW dataset in that those who were victim-survivors of rape were younger on average. Reversely, while the perpetrator being an acquaintance compared to domestic was associated with younger people in RCEW, this was not the case for the CSEW, where no significant association was found, which was also the case in the imputed synthetic dataset. One coefficient was significantly related to age in both datasets, but not in the imputed version (stranger/unknown perpetrator). For all independent variables / controls, the standard errors were similar between the CSEW and the imputed synthetic dataset, which additional testing indicates is due to two opposing mechanisms which (partially) cancel each other out. That is, on the one hand, imputation may result in larger standard errors due to the uncertainty around the imputation; on the other hand, the bigger sample size of the imputation sample leads to smaller standard errors.

Table 2: Associations between age and other variables in RCEW data, CSEW data, and the imputed synthetic dataset. OLS models.

	Dataset A: RCEW original B(SE)	Dataset B: CSEW B(SE)	Synthetic: Dataset A imputed based on Dataset B B(SE)
Sexual violence (Ref: Other)			
Rape	-0.405	-1.949**	-1.761**
	(0.298)	(0.708)	(0.683)
Victim-perpetrator relationship (Ref: domestic)			
Acquaintance	-1.639***	-0.127	0.351
	(0.289)	(0.820)	(0.656)
Stranger or unknown	-1.674***	-1.658*	-0.747
	(0.439)	(0.833)	(0.899)
Gender (Ref: Female)			
Male	3.161***	2.831**	2.241**
	(0.501)	(1.009)	(0.686)
Health impact (Ref: No injury)			
Injury	0.634	0.531	0.918
	(0.559)	(0.676)	(1.262)
Relationship status (Ref: Married/cohabiting			
Single/widowed	-8.157***	-5.234***	-5.410***
	(0.376)	(0.765)	(0.452)
Separated/divorced	2.289***	7.999***	7.382***
	(0.512)	(0.916)	(1.002)
Ethnicity (Ref: White)			
Not White	-0.697	0.382	0.758
	(0.463)	(1.054)	(1.361)
Employment status (Ref: Employed)			
Unemployed	0.605	-2.299*	-2.120**
	(0.330)	(1.154)	(0.799)
Outside labour force	9.383***	4.033***	4.118***
	(0.459)	(0.690)	(0.822)
Student	-10.746***	-6.385***	-6.500*
	(0.427)	(1.335)	(2.811)
Housing tenure (Ref: Homeowner)			
Renter	1.380***	-4.351***	-4.768***
	(0.382)	(0.659)	(0.709)
Other	2.651***	-9.545***	-9.866***
	(0.324)	(1.363)	(1.582)
Nr of dependent	-0.625***	-2.486***	-2.274***
	(0.124)	(0.318)	(0.681)
Constant	40.074***	39.097***	38.591***
	(0.456)	(1.059)	(1.368)
Observations	6,102	1,232	6,102

Source: based on CSEW and RC datasets. *** p<0.001, ** p<0.01, * p<0.05

294

We then tested the approach on a binary variable, *gender*. For this, we stipulated that the comparison vector was specified as:

$$C_{i,j}^{2} = f[\text{type of SV, perpetrator relationship, health impact, employment status, housing tenure, number of dependants relationship status, ethnicity and age}] \quad (4)$$

Table 3 shows the results of logistic regressions looking at the associations between gender as a dependent variable and the independent variables for dataset A (RCEW), dataset B (CSEW) and the complete combined synthetic dataset. Similarly to what we saw in our analyses of age, the imputed dataset mimics the associations from the CSEW dataset. For example, men were less likely to experience rape than women, while stranger perpetrators where more strongly associated with male than female victim-survivors. Two important things stand out: Acquaintance (compared to domestic relationship) perpetrator was not associated with gender, nor was 'other' housing tenure (compared to homeowners) in the CSEW, but these do become significant in the imputed dataset. The latter is most likely due to the far higher prevalence of 'other' housing tenure in the RCEW dataset, making it more likely to reach statistical significance, while the former is likely due to the larger sample size of the imputed dataset compared to the original CSEW dataset.

Table 3: Associations between gender and other variables in RCEW data, CSEW data, and the imputed synthetic dataset. Logistic regression models.

	Dataset A: RCEW original B(SE)	Dataset B: CSEW B(SE)	Synthetic: Dataset A imputed dependent based on dataset B B(SE)
Sexual violence (Ref: Other)			
Rape	-1.113***	-0.840**	-0.747*
	(0.100)	(0.302)	(0.327)
Victim-perpetrator relationship (Ref: domestic)			
Acquaintance	0.646***	0.700	0.921***
	(0.106)	(0.404)	(0.197)
Stranger or unknown	0.189	1.130**	1.352***
	(0.181)	(0.394)	(0.303)
Health impact (Ref: No injury)			
Injury	-0.606*	0.345	0.152
	(0.258)	(0.248)	(0.271)
Relationship status (Ref: Married/cohabiting			
Single/widowed	-0.327*	-0.047	-0.048
	(0.128)	(0.246)	(0.252)
Separated/divorced	-0.620**	-1.494***	-1.661*
	(0.196)	(0.450)	(0.696)
Ethnicity (Ref: White)			
Not White	-0.331	0.247	-0.029
	(0.199)	(0.346)	(0.426)
Employment status (Ref: Employed)			
Unemployed	0.117	0.540	0.526
	(0.120)	(0.345)	(0.345)
Outside labour force	-0.037	-0.214	-0.253
	(0.167)	(0.269)	(0.288)
Student	-0.410*	-0.098	-0.108
	(0.191)	(0.452)	(0.374)
Housing tenure (Ref: Homeowner)			
Renter	0.144	0.424	0.511
	(0.137)	(0.238)	(0.320)
Other	0.005	0.786	1.123*
	(0.120)	(0.418)	(0.523)
Nr of dependent	-0.323***	-0.786***	-0.796***
	(0.056)	(0.197)	(0.218)
Age	0.023***	0.024**	0.029***
	(0.004)	(0.009)	(0.007)
Constant	-2.424***	-3.600***	-3.917***
	(0.232)	(0.589)	(0.430)
Observations	6,102	1,232	6,102

Source: based on CSEW and RCEW datasets. *** p<0.001, ** p<0.01, * p<0.05

312

We considered the consistencies across tables 2 and 3 as an indication that our proposed approach works for variables that are recorded similarly in the two datasets. We then extended our testing to an outcome measure which is *not* similarly recorded in CSEW and RCEW; health impact. For this, we specified the comparison vector as:

$$C_{i,j}^{3} = f[\text{type of SV, perpetrator relationship, employment status, housing tenure, number of dependants, relationship status, ethnicity, age, gender}] \quad (5)$$

Table 4 shows the results of logistic regressions looking at the associations between health impact for dataset A (RCEW), dataset B (CSEW) and the complete combined synthetic dataset. We acknowledge that this is a more meaningful regression specification than the previous two specified in the paper. However, since our approach is novel, we wanted to ensure that the approach worked for similarly recorded variables before testing for differently recorded ones. The results show the same as for the previous models, namely that the imputed dataset reflects the associations from the CSEW dataset in both magnitude, direction, and significance. This includes an association that was positive in the original RCEW dataset (dataset A), but negative in CSEW (dataset B), and thus are also negative in the imputed dataset, namely, whether the perpetrator was a stranger or unknown. In these models, the coefficients for single/widowed stand out, given the synthetic dataset presents a very similar association to that found in the RCEW dataset. This is again likely a result of a much higher prevalence of this group in the RCEW (and therefore in the synthetic dataset).

Table 4: Associations between health impact and other variables in RCEW data, CSEW data, and the imputed synthetic dataset. Logistic regression models.

	Dataset A: RCEW original B(SE)	Dataset B: CSEW B(SE)	Synthetic: Dataset A imputed dependent based on dataset B B(SE)
Sexual violence (Ref: Other)			
Rape	0.187	1.507***	1.541***
	(0.127)	(0.148)	(0.138)
Victim-perpetrator relationship (Ref: domestic)			
Acquaintance	0.224	-0.890***	-0.720***
	(0.121)	(0.178)	(0.122)
Stranger or unknown	0.401*	-1.137***	-0.992***
	(0.168)	(0.181)	(0.163)
Gender (Ref: Female)			
Male	-0.611*	0.232	-0.004
	(0.258)	(0.234)	(0.322)
Relationship status (Ref: Married/cohabiting			
Single/widowed	-0.305*	0.325	0.324*
	(0.151)	(0.185)	(0.149)
Separated/divorced	-0.280	0.183	0.125
	(0.206)	(0.221)	(0.241)
Ethnicity (Ref: White)			
Not White	0.043	-0.241	-0.351
	(0.187)	(0.254)	(0.309)
Employment status (Ref: Employed)			
Unemployed	0.342*	0.246	0.288
	(0.138)	(0.260)	(0.265)
Outside labour force	0.386*	0.300	0.241
	(0.189)	(0.157)	(0.212)
Student	0.306	-1.009**	-0.954**
	(0.183)	(0.348)	(0.362)
Housing tenure (Ref: Homeowner)			
Renter	-0.068	0.503**	0.614*
	(0.149)	(0.157)	(0.260)
Other	-0.524***	0.346	0.625
	(0.137)	(0.329)	(0.421)
Nr of dependent	0.002	0.193**	0.178**
	(0.051)	(0.075)	(0.067)
Age	0.006	0.005	0.010*
	(0.005)	(0.007)	(0.005)
Constant	-2.984***	-1.133**	-1.415***
	(0.281)	(0.353)	(0.219)
Observations	6,102	1,232	6,102

Source: based on CSEW and RCEW datasets. *** p<0.001, ** p<0.01, * p<0.05

333

Finally, in order to achieve our goal of combining data in a real-life application and producing a complete integrated dataset, we inputted a variable that only appears in CSEW and is, therefore, completely missing in RCEW; frequency of abuse. In this case, the comparison vector is:

$$C_{i,j}^{4} = f[\text{type SV, perpetrator relationship, health impact, employment status, housing tenure, number of dependants, relationship status, ethnicity, age, gender}] \quad (6)$$

The analyses estimating the number of sexual violence incidents or repetitions based on CSEW data reveals that rape (compared to other sexual violence) and incidents by acquaintances or strangers (compared to domestic perpetrators) are less likely to be repeated, and if repeated they are repeated fewer times. The imputed synthetic dataset reflects these associations. On the other hand, whilst in the CSEW, significant negative associations between sexual violence incidents and singles/widowed (versus married or cohabitors), non-White (compared to White) victim-survivors exist, these associations did not reach statistical significance in the imputed dataset. Lastly, while students did not have a higher number of sexual violence incidents compared to employed people in the CSEW, in the imputed synthetic dataset this was the case. This change in significance is likely due to the larger proportion of students in the RCEW (and therefore in the synthetic dataset).

Table 5: Associations between number of incidents or repetitions and other variables in CSEW, and the imputed synthetic dataset. Negative binomial models.

	Dataset B: CSEW B(SE)	Synthetic: Dataset A imputed dependent based on dataset B B(SE)
Sexual violence (Ref: Other)		
Rape	-0.585*	-0.706*
	(0.262)	(0.329)
Victim-perpetrator relationship (Ref: domestic)		
Acquaintance	-1.560***	-1.567***
	(0.275)	(0.324)
Stranger or unknown	-2.764***	-2.742***
	(0.299)	(0.398)
Gender (Ref: Female)		
Male	-0.539	-0.372
	(0.400)	(0.558)
Health impact (Ref: No injury)		
Injury	0.313	0.402
	(0.249)	(0.474)
Relationship status (Ref: Married/cohabiting		
Single/widowed	-0.531*	-0.649
	(0.267)	(0.377)
Separated/divorced	0.046	0.086
	(0.316)	(0.269)
Ethnicity (Ref: White)		
Not White	-0.869*	-0.977
	(0.389)	(0.543)
Employment status (Ref: Employed)		
Unemployed	0.129	-0.030
	(0.405)	(0.354)
Outside labour force	0.148	0.166
	(0.255)	(0.516)
Student	0.593	0.796*
	(0.510)	(0.363)
Housing tenure (Ref: Homeowner)		
Renter	0.086	0.081
	(0.233)	(0.484)
Other	0.083	0.097
	(0.566)	(0.624)
Nr of dependent	-0.041	-0.006
	(0.120)	(0.112)
Age	0.017	0.010
	(0.010)	(0.011)
lnalpha	2.166***	2.168***
	(0.084)	(0.138)
Constant	1.135*	1.339
	(0.525)	(0.684)
Observations	1,217	6,102

Source: based on CSEW and RCEW datasets. *** p<0.001, ** p<0.01, * p<0.05

Discussion

There are several implications from our proposed approach to combining data, based on look-alike principles, using multiple imputation methods. First, the initial distribution may be different between datasets, as was the case in the RCEW and CSEW, and while this does not appear to prevent meaningful analyses in the synthetic dataset, the sample sizes are important both in defining what dataset ultimately provides the basis for the synthetic dataset and also in interpreting some of the meaningful associations found. In general, in our proposed exercise, the associations mimic those of the CSEW (smaller sample size), which was used as the basis for imputation. However, where the prevalence of a certain group was much larger in the RCEW (larger sample size), this group was also larger in the synthetic version, meaning that there was an increased chance of significance. In order to test the robustness of our approach, we swapped datasets A and B, that is, we tested imputing data from the RCEW into the CSEW. This led to a synthetic dataset that was the size of the CSEW (1,232). While we found the same general findings, i.e. that magnitude and direction of effect sizes in the synthetic dataset mimicked those of the RCEW (used for imputation instead), standard errors were in general larger, meaning results were less likely to reach significance. This reinforces the importance of sample sizes (both in the imputing and in the imputed datasets).

A strength of the proposed method is that it enables the combining of data on different individuals based on similar characteristics, meaning that working with pseudonymised data is possible. This is relevant to any area of research where there are concerns around data-sharing, not only violence. Furthermore, our analyses have shown that results are fairly consistent regardless of the type of modelling used (OLS, logistic or negative binomials). Integrated survey and administrative data can strengthen study designs by providing more complete information on similar profiles, lessening response burden on participants, or by serving as a source of triangulated data [33].

The approach outlined involved a trade-off between the standardisation of variables required for imputation and the detail about individuals and experiences that is valued in research on violence. The need to standardise variables used for imputation meant that more nuanced understanding of experiences was lost. In our analyses, this was particularly relevant in terms of health impact. While our final coding only allowed for the inclusion of a binary, there is a wide literature on the impacts of sexual violence on physical health[34-36], and some of the final categories in the variables we used were much more aggregate than we would have liked. This was also the case for ethnicity, precluding analyses using an intersectional approach. Furthermore, we did not consider *time* (i.e. time of experience of sexual violence) as a variable in the comparison vector due to limited sample sizes, but we acknowledge that the understanding of experiences of violence varies over time, so ideally *time* should be a comparison-vector variable.

Our proposed data integration approach should be particularly useful for costing or burden of disease type of analyses, including calculating the societal burden of violence, given it enables taking a micro-costing approach, which produces more precise estimates [37]. Nonetheless, further applications, in particular to evaluate interventions, need further testing. Analyses using a longitudinal design are certainly not feasible if *time* is not used as a comparison-vector variable.

Similarly to all applications of multiple imputation, there are assumptions around the patterns of data missingness. While MI assumes data missing at random (MAR) or missing completely at random (MCAR), when using our approach to impute a variable that only appears in one dataset, there is a normative assumption that the synthetic dataset follows the same distribution (and the same pattern of missingness) as the dataset used for imputation.

Finally, there are numerous practice and policy implications for researchers, voluntary sector partner organisations, and the general population. Compared to traditional research, our proposed approach to data integration offers a cost-effective solution to breaking (data-related) silos in research. Further research should not only test different approaches to data integration, but also applications to evaluations by mutually engaging practitioners, policymakers, and researchers to foster a culture of research [33, 38] facilitating the refinement of techniques as well as producing real-world evidence based on integrated synthetic data.

Conclusion

This study has demonstrated that data integration between a survey (CSEW) and administrative records (RCEW) is possible using look-alike modelling principles and using multiple imputation by chained equations. Our results serve as a proof of concept, and the associations in the resulting synthetic dataset tend to mimic the dataset used for imputation in magnitude and direction. The regression results in the synthetic dataset also tend to yield larger standard errors, resulting in larger confidence intervals. This approach should be applicable for costing exercises as it permits micro-costing. Further applications of the approach should be the focus of future research.

References

1. Concha-Eastman A. Violence: a challenge for public health and for all. Journal of Epidemiology & Community Health. 2001;55(8):597-9.

2. Rosenberg ML, O'Carroll PW, Powell KE. Let's be clear: Violence is a public health problem. Jama. 1992;267(22):3071-2.

3. Assembly WH. Prevention of violence: Public health priority. 1996. p. 20-5 May 1996.

4. Blom N, Fadeeva A, Barbosa EC. The Concept and Measurement of Violence and Abuse in Health and Justice Fields: Toward a Framework Aligned with the UN Sustainable Development Goals. Social Sciences. 2023;12(6):316.

5. Waters HR, Hyder AA, Rajkotia Y, Basu S, Butchart A. The costs of interpersonal violence—an international review. Health policy. 2005;73(3):303-15.

6. Oliver R, Alexander B, Roe S, Wlasny M. The economic and social costs of domestic abuse. Home Office (UK). 2019.

7. O'Hara A, Shattuck RM, Goerge RM. Linking federal surveys with administrative data to improve research on families. The ANNALS of the American Academy of Political and Social Science. 2017;669(1):63-74.

8. Florence C, Shepherd J, Brennan I, Simon T. Effectiveness of anonymised information sharing and use in health service, police, and local government partnership for preventing violence related injury: experimental study and time series analysis. Bmj. 2011;342.

9. Shepherd JP, Ali M, Hughes A, Levers B. Trends in urban violence: a comparison of accident department and police records. Journal of the Royal Society of Medicine. 1993;86(2):87.

10. Sutherland I, Sivarajasingam V, Shepherd JP. Recording of community violence by medical and police services. Injury Prevention. 2002;8(3):246-7.

11. Faergemann C, Lauritsen JM, Brink O, Skov O. Trends in deliberate interpersonal violence in the Odense Municipality, Denmark 1991–2002.: The Odense study on deliberate interpersonal violence. Journal of forensic and legal medicine. 2007;14(1):20-6.

12. Mason R, Wolf M, O'Rinn S, Ene G. Making connections across silos: intimate partner violence, mental health, and substance use. BMC women's health. 2017;17(1):1-7.

13. Bunce A, Carlisle S, Capelas Barbosa E. The Concept and Measurement of Interpersonal Violence in Specialist Services Data: Inconsistencies, Outcomes and the Challenges of Synthesising Evidence. Social Sciences. 2023;12(7):366.

14. DAC. Safety before status: the solutions. The Domestic Abuse Commissioner's second report on supporting migrant survivors of domestic abuse. . https://www.gov.uk/government/publications/safety-before-status-the-solutions; 2022.

15. Imkaan, RCEW, Respect, SafeLives, Women's_Aid. Sector Sustainability Shared Standards: Shared Values That Apply across the VAWG Sector. . Bristol; 2016.

16. Chacko AM, Pranav BA, Madhvesh BV, Poornima A, editors. Customer Lookalike Modeling: A Study of Machine Learning Techniques for Customer Lookalike Modeling. Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020; 2021: Springer.

17. Rahman MM, Kikuta D, Abrol S, Hirate Y, Suzumura T, Loyola P, et al. Exploring 360-Degree View of Customers for Lookalike Modeling. arXiv preprint arXiv:230409105. 2023.

18. Peng Y, Liu C, Shen W. Finding Lookalike Customers for E-Commerce Marketing. arXiv preprint arXiv:230103147. 2023.

19. Medalia C, Meyer BD, O'Hara AB, Wu D. Linking survey and administrative data to measure income, inequality, and mobility. International journal of population data science. 2019;4(1).

20. StataCorp. Impute missing values using chained equations (manual). College Station, TX: Stata Press; 2023.

21. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing data. Jama. 2015;314(18):1966-7.

22. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. BMC medical research methodology. 2017;17(1):1-10.

23. Lall R. How multiple imputation makes a difference. Political Analysis. 2016;24(4):414-33.

24. Kennedy AC, Prock KA. "I still feel like I am not normal": A review of the role of stigma and stigmatization among female survivors of child sexual abuse, sexual assault, and intimate partner violence. Trauma, Violence, & Abuse. 2018;19(5):512-27.

25. Delker BC, Salton R, McLean KC, Syed M. Who has to tell their trauma story and how hard will it be? Influence of cultural stigma and narrative redemption on the storying of sexual violence. PloS one. 2020;15(6):e0234201.

26. Chakraborty T, Mukherjee A, Rachapalli SR, Saha S. Stigma of sexual violence and women's decision to work. World Development. 2018;103:226-38.

27. Yu L-M, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of semi-continuous data. Statistical methods in medical research. 2007;16(3):243-58.

28. Little R. Statistical analysis with missing data. statistical analysis with missing data, by RJA little and DB Rubin Wiley series in probability and stistics. New York, NY: Wiley. 2002;2002:1.

29. Gomes M, Díaz-Ordaz K, Grieve R, Kenward MG. Multiple imputation methods for handling missing data in cost-effectiveness analyses that use data from hierarchical studies: an application to cluster randomized trials. Medical decision making. 2013;33(8):1051-63.

30. Lovett J, Kelly L. Hidden Depths: a detailed study of Rape Crisis data. 2016.

31. ONS. Crime Survey for England and Wales. 1982 - 2022.

32. Innes A, Blom N, Bunce A, Fadeeva A, Manzur H, Thiara R, et al. Assessment of data and Risk of Bias when using data Ethnicity and Migration. In: Consortium UV, editor. 2023.

33. DeHart D, Shapiro C. Integrated administrative data & criminal justice research. American Journal of Criminal Justice. 2017;42:255-74.

34. Martin SL, Macy RJ, Young SK. Health and economic consequences of sexual violence. 2011.

35. Crofford LJ. Violence, stress, and somatic syndromes. Trauma, Violence, & Abuse. 2007;8(3):299-313.

36. Jina R, Thomas LS. Health consequences of sexual violence against women. Best practice & research Clinical obstetrics & gynaecology. 2013;27(1):15-26.

37. Gold MR. Cost-effectiveness in health and medicine: Oxford university press; 1996.

38. Duran F, Wilson S, Carroll D. Putting administrative data to work: A toolkit for state agencies on advancing data integration and data sharing efforts to support sound policy and program development. Farmington, CT: Child Health and Development Institute of Connecticut. 2005.