

1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17

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

Estela Capelas Barbosa^{1* #a}, Niels Blom², Annie Bunce²

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.

18 **Abstract**

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

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

39 **Introduction**

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

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

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

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

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

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

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

81 **Theoretical framework**

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

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

$$\begin{aligned} 95 \quad X_1^{(t+1)} &\sim g_1(X_1|X_2^{(t)}, \dots, X_p^{(t)}, C, \varphi_1) \\ 96 \quad X_2^{(t+1)} &\sim g_2(X_2|X_1^{(t+1)}, X_3^{(t)}, \dots, X_p^{(t)}, C, \varphi_2) \\ 97 \quad &\dots \\ 98 \quad X_p^{(t+1)} &\sim g_p(X_p|X_1^{(t+1)}, X_2^{(t+1)}, \dots, X_{p-1}^{(t+1)}, C, \varphi_p) \end{aligned} \quad (1)$$

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

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

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

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

$$126 \quad f_{ij} = \beta^f C_{ij} + \gamma^f D_j + \varepsilon_{ij}^f \quad (2)$$

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

134 **Methods**

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

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

145 **Datasets**

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

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

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

179 Considering our focus on sexual violence, we selected respondents (CSEW) or service
180 users (RCEW) who have reported being a victim-survivor of rape (including attempted) or
181 another form sexual violence and abuse (which included indecent assault and wounding with
182 sexual motive). We selected respondents/service users with no missing values on vector C
183 variables, which led to a sample of 1,232 incidents from 1,111 individuals in the CSEW, and
184 6,102 referral cases from 5,333 individuals in RCEW. In RCEW data, it included data for
185 individuals who accessed the service more than once.

186 **The comparison vector (C) variables**

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

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

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

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

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

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

218 Relationship status was categorised into whether respondents were in a co-residential
219 relationship (either married or cohabiting), single/non-resident partner/widowed, or separated
220 or divorced in the CSEW and RCEW. Ethnicity was coded as White and non-White, as further
221 differentiation led to too small numbers in some categories. However, we acknowledge that
222 the ethnicity categorisation of White/non-White may be problematic and any conclusion in this
223 respect, limited [32]. Employment status in both datasets was assessed by whether people were
224 employed, unemployed, students, or outside the labour force (e.g. a homemaker, retired, or
225 unable to work due to illness). Gender was asked as whether the respondent was male or
226 female¹ in the CSEW. In RCEW data, more detailed responses are given, including transgender
227 female and transgender male, which were recoded into men and women. Finally, age was
228 measured numerically in both datasets and we included in our analyses people over the age of
229 16. Table S1 in the supporting information summarises how variables were harmonised.

230 **Analytical Strategy**

231 To test whether approaching look-alike modelling as a missing data problem was
232 effective, we compared regression analyses for the two datasets (CSEW and RCEW) and the
233 combined synthetic dataset, which imputed data based on the comparison vector. As a proof of
234 concept, we tested the approach using variables of different types (binary and continuous) that
235 are observed in both datasets. Formally, our approach had three steps. First, we specified the
236 same linear (OLS) or logistic regression (as appropriate) for dataset A (RCEW) and dataset B
237 (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.

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

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

250 **Results**

251 **Profiles comparison of sexual violence victim-survivors in CSEW** 252 **and RC**

253 Before conducting our look-alike exercises, we compared the profiles of sexual
254 violence victim-survivors in CSEW and RCEW datasets (Table 1). The table shows some
255 meaningful differences between the individuals pertaining to each dataset. Particularly, only
256 32% of sexual violence victim-survivors in the CSEW had been victims of rape, compared to
257 71% in the RCEW data. Relationship to the perpetrator was more likely to be domestic in the
258 RCEW data compared to the CSEW (48% vs 25%, respectively) and perpetrators were far more
259 likely to be strangers or to be unknown in the CSEW (42%), compared to only 12% of records
260 in the RCEW dataset.

261 Furthermore, the CSEW recorded a physical injury in 39% of incidents, while this
262 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)
<i>Type of sexual violence</i>				
Rape	32.4		70.7	
Other sexual violence and abuse	67.6		29.3	
<i>Victim-perpetrator relationship</i>				
Domestic	24.6		48.2	
Acquaintance	33.2		40.2	
Stranger or unknown	42.2		11.6	
<i>Physical health impact</i>				
No injury	60.9		94.0	
Injury	39.1		6.0	
<i>Gender</i>				
Male	9.6		7.9	
Female	90.4		92.1	
<i>Relationship status</i>				
Married/Cohabiting	21.7		16.8	
Single/non-resident relationship/Widowed	57.0		71.7	
Separated/Divorced	21.4		11.6	
<i>Ethnicity</i>				
White	91.6		91.0	
Non-White	8.4		9.0	
<i>Employment status</i>				
Employed	56.4		35.2	
Unemployed	7.6		38.2	
Outside labour force	30.4		11.7	
Student	5.6		14.9	
<i>Housing tenure</i>				
Homeowner/lives in own home	34.6		41.8	
Renter	59.8		18.4	
Other	5.6		39.7	
<i>Age</i>		32.9 (12.4)		34.1 (12.9)
<i>Nr of dependents</i>		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)

268

269

270 **Look-alike empirical application**

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

$$273 \quad C_{ij}^1 = f[\text{type of SV, perpetrator relationship, health impact, employment status, housing} \\ 274 \quad \text{tenure, number of dependants, relationship status, ethnicity and gender}] \quad (3)$$

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

293

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

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

$$297 \quad C_{i,j}^2 = f[\text{type of SV, perpetrator relationship, health impact, employment status, housing} \\ 298 \quad \text{tenure, number of dependants relationship status, ethnicity and age}] \quad (4)$$

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

311

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

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

$$317 \quad C_{ij}^3 = f[\text{type of SV, perpetrator relationship, employment status, housing tenure, number of} \\ 318 \quad \text{dependants, relationship status, ethnicity, age, gender}] \quad (5)$$

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

332

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

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

$$338 \quad C_{i,j}^4 = f[\text{type SV, perpetrator relationship, health impact, employment status, housing tenure,} \\ 339 \quad \text{number of dependants, relationship status, ethnicity, age, gender}] \quad (6)$$

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

351

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

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

352 **Discussion**

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

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

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

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

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

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

407 **Conclusion**

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

416

417

418 **References**

- 419 1. Concha-Eastman A. Violence: a challenge for public health and for all. *Journal of*
420 *Epidemiology & Community Health*. 2001;55(8):597-9.
- 421 2. Rosenberg ML, O'Carroll PW, Powell KE. Let's be clear: Violence is a public health
422 problem. *Jama*. 1992;267(22):3071-2.
- 423 3. Assembly WH. Prevention of violence: Public health priority. 1996. p. 20-5 May
424 1996.
- 425 4. Blom N, Fadeeva A, Barbosa EC. The Concept and Measurement of Violence and
426 Abuse in Health and Justice Fields: Toward a Framework Aligned with the UN Sustainable
427 Development Goals. *Social Sciences*. 2023;12(6):316.
- 428 5. Waters HR, Hyder AA, Rajkotia Y, Basu S, Butchart A. The costs of interpersonal
429 violence—an international review. *Health policy*. 2005;73(3):303-15.
- 430 6. Oliver R, Alexander B, Roe S, Wlasny M. The economic and social costs of domestic
431 abuse. Home Office (UK). 2019.
- 432 7. O'Hara A, Shattuck RM, Goerge RM. Linking federal surveys with administrative
433 data to improve research on families. *The ANNALS of the American Academy of Political*
434 *and Social Science*. 2017;669(1):63-74.
- 435 8. Florence C, Shepherd J, Brennan I, Simon T. Effectiveness of anonymised
436 information sharing and use in health service, police, and local government partnership for
437 preventing violence related injury: experimental study and time series analysis. *Bmj*.
438 2011;342.
- 439 9. Shepherd JP, Ali M, Hughes A, Levers B. Trends in urban violence: a comparison of
440 accident department and police records. *Journal of the Royal Society of Medicine*.
441 1993;86(2):87.

- 442 10. Sutherland I, Sivarajasingam V, Shepherd JP. Recording of community violence by
443 medical and police services. *Injury Prevention*. 2002;8(3):246-7.
- 444 11. Faergemann C, Lauritsen JM, Brink O, Skov O. Trends in deliberate interpersonal
445 violence in the Odense Municipality, Denmark 1991–2002.: The Odense study on deliberate
446 interpersonal violence. *Journal of forensic and legal medicine*. 2007;14(1):20-6.
- 447 12. Mason R, Wolf M, O’Rinn S, Ene G. Making connections across silos: intimate
448 partner violence, mental health, and substance use. *BMC women's health*. 2017;17(1):1-7.
- 449 13. Bunce A, Carlisle S, Capelas Barbosa E. The Concept and Measurement of
450 Interpersonal Violence in Specialist Services Data: Inconsistencies, Outcomes and the
451 Challenges of Synthesising Evidence. *Social Sciences*. 2023;12(7):366.
- 452 14. DAC. Safety before status: the solutions. The Domestic Abuse Commissioner's
453 second report on supporting migrant survivors of domestic abuse. .
454 <https://www.gov.uk/government/publications/safety-before-status-the-solutions>; 2022.
- 455 15. Imkaan, RCEW, Respect, SafeLives, Women’s_Aid. Sector Sustainability Shared
456 Standards: Shared Values That Apply across the VAWG Sector. . Bristol; 2016.
- 457 16. Chacko AM, Pranav BA, Madhvesh BV, Poornima A, editors. Customer Lookalike
458 Modeling: A Study of Machine Learning Techniques for Customer Lookalike Modeling.
459 Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI
460 2020; 2021: Springer.
- 461 17. Rahman MM, Kikuta D, Abrol S, Hirate Y, Suzumura T, Loyola P, et al. Exploring
462 360-Degree View of Customers for Lookalike Modeling. *arXiv preprint arXiv:230409105*.
463 2023.
- 464 18. Peng Y, Liu C, Shen W. Finding Lookalike Customers for E-Commerce Marketing.
465 *arXiv preprint arXiv:230103147*. 2023.

- 466 19. Medalia C, Meyer BD, O'Hara AB, Wu D. Linking survey and administrative data to
467 measure income, inequality, and mobility. *International journal of population data science*.
468 2019;4(1).
- 469 20. StataCorp. *Impute missing values using chained equations (manual)*. College Station,
470 TX: Stata Press; 2023.
- 471 21. Li P, Stuart EA, Allison DB. Multiple imputation: a flexible tool for handling missing
472 data. *Jama*. 2015;314(18):1966-7.
- 473 22. Jakobsen JC, Gluud C, Wetterslev J, Winkel P. When and how should multiple
474 imputation be used for handling missing data in randomised clinical trials—a practical guide
475 with flowcharts. *BMC medical research methodology*. 2017;17(1):1-10.
- 476 23. Lall R. How multiple imputation makes a difference. *Political Analysis*.
477 2016;24(4):414-33.
- 478 24. Kennedy AC, Prock KA. “I still feel like I am not normal”: A review of the role of
479 stigma and stigmatization among female survivors of child sexual abuse, sexual assault, and
480 intimate partner violence. *Trauma, Violence, & Abuse*. 2018;19(5):512-27.
- 481 25. Delker BC, Salton R, McLean KC, Syed M. Who has to tell their trauma story and
482 how hard will it be? Influence of cultural stigma and narrative redemption on the storying of
483 sexual violence. *PloS one*. 2020;15(6):e0234201.
- 484 26. Chakraborty T, Mukherjee A, Rachapalli SR, Saha S. Stigma of sexual violence and
485 women's decision to work. *World Development*. 2018;103:226-38.
- 486 27. Yu L-M, Burton A, Rivero-Arias O. Evaluation of software for multiple imputation of
487 semi-continuous data. *Statistical methods in medical research*. 2007;16(3):243-58.
- 488 28. Little R. *Statistical analysis with missing data*. statistical analysis with missing data,
489 by RJA little and DB Rubin Wiley series in probability and stistics. New York, NY: Wiley.
490 2002;2002:1.

- 491 29. Gomes M, Díaz-Ordaz K, Grieve R, Kenward MG. Multiple imputation methods for
492 handling missing data in cost-effectiveness analyses that use data from hierarchical studies:
493 an application to cluster randomized trials. *Medical decision making*. 2013;33(8):1051-63.
- 494 30. Lovett J, Kelly L. *Hidden Depths: a detailed study of Rape Crisis data*. 2016.
- 495 31. ONS. *Crime Survey for England and Wales. 1982 - 2022*.
- 496 32. Innes A, Blom N, Bunce A, Fadeeva A, Manzur H, Thiara R, et al. Assessment of
497 data and Risk of Bias when using data Ethnicity and Migration. In: Consortium UV, editor.
498 2023.
- 499 33. DeHart D, Shapiro C. Integrated administrative data & criminal justice research.
500 *American Journal of Criminal Justice*. 2017;42:255-74.
- 501 34. Martin SL, Macy RJ, Young SK. Health and economic consequences of sexual
502 violence. 2011.
- 503 35. Crofford LJ. Violence, stress, and somatic syndromes. *Trauma, Violence, & Abuse*.
504 2007;8(3):299-313.
- 505 36. Jina R, Thomas LS. Health consequences of sexual violence against women. *Best
506 practice & research Clinical obstetrics & gynaecology*. 2013;27(1):15-26.
- 507 37. Gold MR. *Cost-effectiveness in health and medicine*: Oxford university press; 1996.
- 508 38. Duran F, Wilson S, Carroll D. *Putting administrative data to work: A toolkit for state
509 agencies on advancing data integration and data sharing efforts to support sound policy and
510 program development*. Farmington, CT: Child Health and Development Institute of
511 Connecticut. 2005.

512