Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Statistical Methods in Medical Research
2020, Vol. 29(4) 1212–1226
© The Author(s) 2019

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0962280219854482
journals.sagepub.com/home/smm
SAGE

Correcting for non-participation bias in health surveys using record-linkage, synthetic observations and pattern mixture modelling

Linsay Gray,[1] Emma Gorman,[1,2] Ian R White,[3] S Vittal Katikireddi,[1,4] Gerry McCartney,[5] Lisa Rutherford[6] and Alastair H Leyland[1]

Abstract

Surveys are key means of obtaining policy-relevant information not available from routine sources. Bias arising from non-participation is typically handled by applying weights derived from limited socio-demographic characteristics. This approach neither captures nor adjusts for differences in health and related behaviours between participants and non-participants within categories. We addressed non-participation bias in alcohol consumption estimates using novel methodology applied to 2003 Scottish Health Survey responses record-linked to prospective administrative data. Differences were identified in socio-demographic characteristics, alcohol-related harm (hospitalisation or mortality) and all-cause mortality between survey participants and, from unlinked administrative sources, the contemporaneous general population of Scotland. These were used to infer the number of non-participants within each subgroup defined by socio-demographics and health outcomes. Synthetic observations for non-participants were then generated, missing only alcohol consumption. Weekly alcohol consumption values among synthetic non-participants were multiply imputed under missing at random and missing not at random assumptions. Relative to estimates adjusted using previously derived weights, the obtained mean weekly alcohol intake estimates were up to 59% higher among men and 16% higher among women, depending on the assumptions imposed. This work demonstrates the universal value of multiple imputation-based methodological advancement incorporating administrative health data over routine weighting procedures.

Keywords

Missing not at random, multiple imputation, non-participation, pattern-mixture modelling, record-linkage, survey data

1 Introduction

Population health and health behaviour estimates are commonly derived from survey data to monitor trends and formulate and evaluate policies. However, bias may arise if the survey samples are not representative of the target population. Non-representativeness is of some concern when measures of association such as relative risk are being estimated[1] but of greater concern for population prevalence and quantity estimates,[2–4] such as for alcohol consumption.[5] A key aspect influencing the extent to which surveys are representative is the level of non-participation (unit non-response) among individuals included in the sampling frame. For instance, there is likely to be a group of harmful and dependent drinkers who may be disinclined to participate.

[1]MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, Glasgow, UK
[2]Department of Economics, Lancaster University, Lancaster, UK
[3]MRC Clinical Trials Unit at UCL, London, UK
[4]Directorate of Public Health and Health Policy, NHS Lothian, Edinburgh, UK
[5]NHS Health Scotland, Glasgow, UK
[6]ScotCen Social Research, Edinburgh, UK

Corresponding author:
Linsay Gray, MRC/CSO Social and Public Health Sciences Unit, University of Glasgow, 200 Renfield Street, Glasgow, G2 3AX, UK.
Email: linsay.gray@glasgow.ac.uk

Survey weights derived from inverse probability weighting[6] are usually applied in an attempt to correct for such unit non-response (as well as accounting for aspects of sampling design such as the oversampling of certain household types or geographical areas). However, these weights typically rely on a limited range of socio-demographic variables[7] and are based on the assumption that non-participants have equivalent behaviours to participants in the same socio-demographic category which is unlikely to be the case.

An alternative to the application of survey weights is multiple imputation (MI),[8] which is viable if the assumption that data are missing at random (MAR: the probability of missingness is unrelated to the unobserved data conditional on the observed data) holds. Alanya et al. applied MI to make unit non-response adjustments and compared it to weighting.[9] They found MI to compare favourably, though not consistently so. In another comparison with weighting, MI showed comparable performance in terms of bias but also yielded substantially lower variance estimates.[10] However, these papers made no allowance for the data being missing not at random (MNAR: the probability of missingness is related to the unobserved data). If the data are thought to be MNAR then an alternative approach is required, typically involving sensitivity analyses, and using methodology such as pattern mixture modelling[11] among others.[12,13]

Application of MI is strengthened if we can infer information on the absent non-participants. In the absence of whole population registers, as existing in Nordic countries,[14] nations typically lack individual-level data amenable to forming the bases of sampling frames. Thus, in countries such as the UK, individual non-participants cannot readily be identified and their routine health data extracted.

We propose a novel methodology that aims to improve addressing non-participation bias in national health survey data in order to obtain less biased estimates of alcohol consumption.[15,16] We consider both MAR and MNAR within a missing data framework, motivated by the possibility of non-participants differing in their alcohol consumption[17] from survey participants with the same socio-demographic variables and health outcome statuses. Our approach involves: (1) exploitation of record-linkage to hospital discharges and mortality; (2) survey–population comparisons which inform the creation of synthetic partial observations for non-participants; and (3) MI to generate refined estimates of weekly consumption of alcohol under assumptions of MAR (weaker than when based on survey data alone) and explorations of MNAR.[18] We illustrate the application using data from the 2003 Scottish Health Survey (SHeS) individually record-linked to administrative health information from the Scottish Morbidity Records (SMR), mortality data from the National Records of Scotland (NRS) and unlinked contemporaneous data for the entire population.

In the next section, we provide the context and motivation for the methodological approach described in section 3. In section 4, we report on the application before discussing the implications in section 5 and concluding in section 6.

2 Motivating example and data

2.1 Aim

We aim to devise and apply methodology to estimate sex-specific adult population mean alcohol consumption from national health survey data accounting for bias induced by non-participation.

2.2 SHeS

SHeS are a series of cross-sectional surveys designed to represent the Scottish population living in private households.[19] Socio-demographic data available in the surveys include sex, age group and Scottish Index of Multiple Deprivation (an area-based measure of deprivation collapsed into five equal population-weighted groups), collectively referred to here as 'socio-demographic characteristics'. Alcohol consumption is calculated in units (equivalent to 10 ml or 8 g of pure ethanol) per week. Pre-derived survey sampling weights which sum to the achieved sample total have been created to account for the stratified, multi-stage random sample survey design and departures from population estimates by sex and age.[19] We use the 2003 survey which had an adult response level of 60%.

2.3 Linked health outcomes

Baseline data on consenting SHeS participants (91%) have been confidentially linked to routinely-collected nationwide administrative health records available until the end of 2011 providing prospective follow up of around eight years. These include prospective SMR which record hospital discharges (~90% accurate diagnosis, 99% complete[20]) and mortality data using a probabilistic matching algorithm[21–24] (Figure 1).

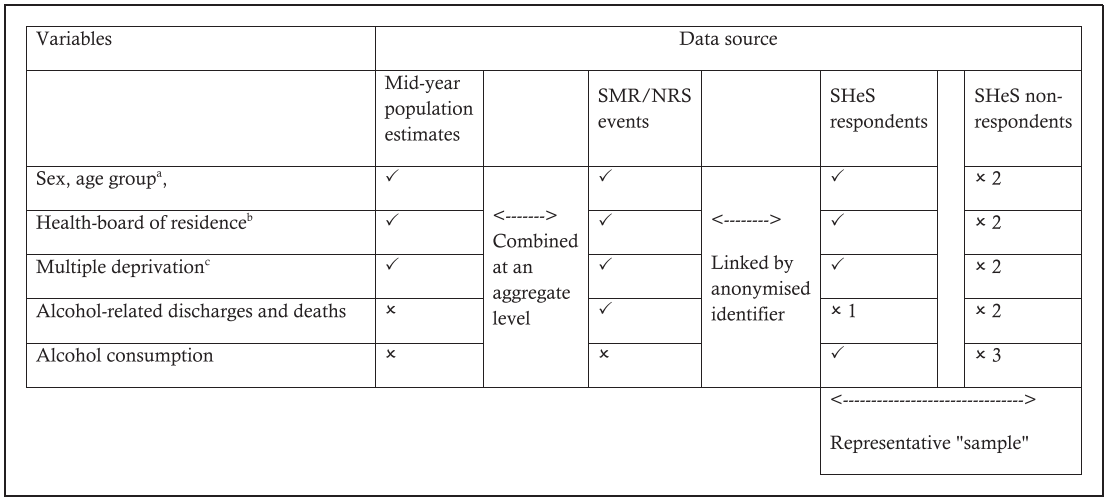

Figure 1. Available data from mid-year population estimates, Scottish Morbidity Records/National Records of Scotland, Scottish Health Survey data sources and desired data on SHeS non-respondents.

Table 1. Sex- and area deprivation group-specific breakdowns (%) for the general population of Scotland and participants[a] in the Scottish Health Survey 2003 aged 20 to 64 years consenting to linkage with inferred estimates for non-participants.

	Population		Scottish Health Survey					
			Respondents		Synthetic non-respondents		Inferred total	
Area deprivation group	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)
Least deprived	10.4	10.5	10.6	11.4	10.2	9.5	10.4	10.5
2	10.0	10.2	10.2	10.9	9.8	9.6	10.0	10.3
3	9.7	10.0	9.1	9.7	10.6	10.4	9.8	10.0
4	9.6	10.2	9.7	10.3	9.5	10.1	9.6	10.2
Most deprived	9.1	10.1	8.4	9.9	10.0	10.4	9.1	10.1
All groups	48.9	51.1	47.9	52.1	50.1	49.9	48.9	51.1

[a]Those participants consenting to record-linkage of their data.

2.4 Population data

For the general population, mid-year population estimates – available by sex, age group and area deprivation – for 2003 were used as denominators.[25] Numerator counts of morbidity and mortality events in the population during the eight years of follow-up were combined with mid-year population estimates – also by socio-demographic characteristics – to create an unlinked aggregate-level data set for the population for comparison with the record-linked survey data.

Two pertinent binary 'health outcome' variables were created from the morbidity and mortality data, the first indicating hospitalisation or death from an alcohol-related cause during the follow-up period (taken together as comprising alcohol-related harm, Supplemental Table 1) and the second indicating all-cause mortality during follow-up. The analyses were restricted to individuals aged 20 to 64 years in the survey year in an attempt to reduce the distortion of institution-dwelling communities (e.g. older people living in care homes) – which are not in the sampling frame – on the survey-population comparisons.

3 Methodology

Our approach to addressing non-participation bias in alcohol consumption estimates involved filling in the missing data in the survey in three stages marked as 1, 2 and 3 in Figure 1. The three stages are depicted in Figure 2 and described in detail in sections 3.2 to 3.4, with a worked example given in section 4.1. We compare the results of our approach with those obtained from the traditional survey-weighted results.

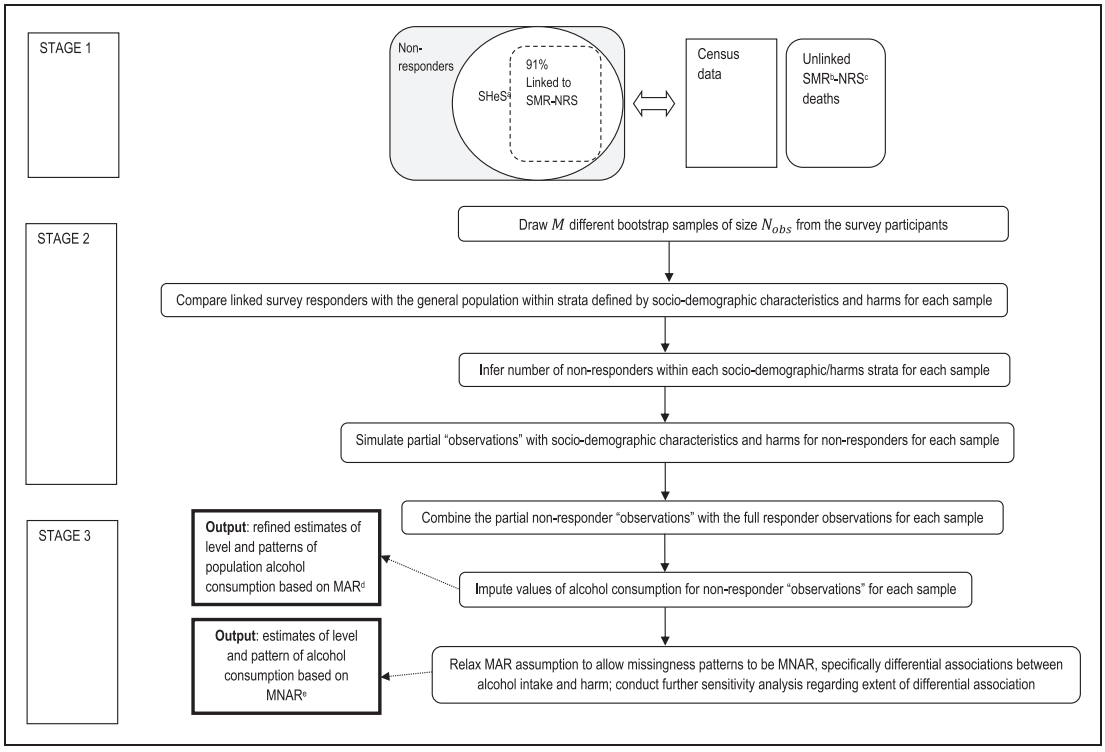

Figure 2. Summary of methodological strategy for addressing survey non-representativeness and refining alcohol consumption estimates. [a]SHeS: Scottish Health Survey; [b]SMR: Scottish Morbidity Record; [c]NRS: National Records of Scotland; [d]MAR: missing at random; [e]MNAR: missing not at random.

3.1 Notation

We use the following notation. Let N_{obs} be the number of linkage-consenting participants (henceforth referred to as 'participants' for simplicity of terminology) in the survey and let N_{mis} be the converse 'non-participants' (comprising those who did not participate and those who did participate but did not consent to record linkage). Let L denote the effective response level – the percentage of the sample who both responded to the survey and consented to linkage (determined by the product of the survey response level and the consent proportion and henceforth referred to as 'response level'). We infer the total survey size to be $N = N_{obs}/L$. Let $\boldsymbol{X}$ be the set of categorical random variables for the socio-demographic covariates; here X_1 is sex (1 for men and 2 for women), X_2 is age group and X_3 is area deprivation quintile. Let $\boldsymbol{H}$ be the set of random variables for the health outcomes; here H_1 corresponds to alcohol-related harm, and H_2 corresponds to all-cause mortality. Let Y denote the random variable for usual weekly alcohol consumption as a semi-continuous measure (detailed in section 3.4.1).

Let S denote the binary random variable for the source of the data, with $S = SHeS$ for the target survey sample (i.e. participants and non-participants combined) and $S = Pop$ for the administrative data. For the survey data, let R be the binary random variable for response, such that $R = 1$ for survey participants and $R = 0$ for survey non-participants. Pre-derived sampling weights for the survey participants are denoted w and have a mean of one. The observed linked survey data are therefore $(R\boldsymbol{X}, RY, R\boldsymbol{H})$ while the unobserved data are $((1-R)\boldsymbol{X}, (1-R)Y, (1-R)\boldsymbol{H})$. Let [.] denote a distribution. Finally, let M be the total number of repeated independent data sets arising from MI.

3.2 Stage 1: Using record linked data

In stage 1, record linkage of survey data to SMR and NRS data was used to determine the values of $\boldsymbol{H}$ for consenting survey participants. Surveyed individuals who did not give consent to record-linkage were treated as non-participants i.e. their survey data observations were excluded, as this was deemed a pragmatic approach.

3.3 Stage 2: Creation of synthetic observations for non-participants

In stage 2, we made inference on the non-participants by comparing the national health survey data with corresponding population data to identify deviations from representativeness in terms of $\boldsymbol{H}$ in addition to $\boldsymbol{X}$. We generated $N_{mis} = N - N_{obs}$ synthetic observations for the non-participants to the SHeS and filled in their values of $\boldsymbol{X}$ and $\boldsymbol{H}$, as follows.

We first assumed that $[\boldsymbol{X}, \boldsymbol{H}]$ – which we know from the population data – is the same in the target survey sample (irrespective of R) as in the population

$$\mathbf{A1}: \ [\boldsymbol{X}, \boldsymbol{H}|S = SHeS] \ = [\boldsymbol{X}, \boldsymbol{H}|S = Pop]$$

This assumption is valid if the sampling frame for the survey is representative of the general population.

It follows that the $\boldsymbol{X}$ and $\boldsymbol{H}$ characteristics of non-participants can be inferred by comparison of survey and general population data, exploiting the categorical nature of $\boldsymbol{X}$ and $\boldsymbol{H}$. Using the equality

$$\begin{aligned} P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS) = {} & P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 1)P(R = 1|S = SHeS) \\ & + P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0)P(R = 0|S = SHeS) \end{aligned}$$

we can write, using A1

$$P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0) = \frac{P(\boldsymbol{X}, \boldsymbol{H}|S = Pop) - \ P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 1)P(R = 1|S = SHeS)}{P(R = 0|S = SHeS)} \quad (1)$$

where each term on the right hand side can be estimated from the data: for estimation of $P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 1)$, simple weighted prevalences for each combination of $\boldsymbol{X}$ and $\boldsymbol{H}$ were used; $P(R = 1|S = SHeS)$ is L (for which weighting may be ignored). We hence identified the number of missing participants within each socio-demographic group in the survey as $N_{mis} \times P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0)$. The corresponding number of non-participant synthetic observations with assigned characteristics were generated for each $(\boldsymbol{X}, \boldsymbol{H})$ category, with usual weekly alcohol consumption, Y, set to missing, and w set to 1. Combining the participants with the synthetic observations for the non-participants provides a data set for imputation.

Three modifications were needed to this general method. First, the method as proposed does not allow for uncertainty in the survey-based estimates of $P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 1)$ arising from sampling variation. To accommodate such uncertainty, we drew M different bootstrap samples of size N_{obs} from the survey participants, yielding M different imputed datasets. Theoretically, the bootstrap method could have been extended to allow for uncertainty in $P(\boldsymbol{X}, \boldsymbol{H}|S = Pop)$ and $P(R = 1|S = SHeS)$, but we did not do this since these quantities were estimated with little imprecision.

Second, the calculated numbers of missing participants in each category were generally not integers. To avoid possible bias due to rounding, we applied random rounding which preserves the mean count. For example, if 2.6 missing participants were required in a particular category, then we took 3 missing participants with probability 0.6, and 2 missing participants with probability 0.4. This was performed separately in each imputed data set.

Third, estimates of $P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0)$ are in some instances negative due to sampling variation in cells with small numbers. This was handled by removing synthetic observations in the nearest neighbouring category or categories: for example, if a particular category $(\boldsymbol{X}, \boldsymbol{H}) = (x_1, x_2, x_3, h_1, h_2)$ required -3 synthetic observations, we identified the nearest category of synthetic individuals and randomly deleted 3 of them. The metric defining distance was the sum of the squared differences between the values of categories $(\boldsymbol{X}, \boldsymbol{H}) = (x_1, x_2, x_3, h_1, h_2)$ scaled by the squared standard deviation.

3.4 Stage 3: Imputing alcohol consumption for non-participants

Once the synthetic observations for the non-participants were created at Stage 2, the unit (person) non-response problem had been converted into an item (variable) non-response problem with the synthetic non-participant observations having data on socio-demographic characteristics and health outcomes but missing data on alcohol consumption. Imputation models for alcohol consumption could then be specified conditional on socio-demographic characteristics and health outcomes. Missing alcohol consumption observations among a small minority of participants ($n = 16$) were imputed in the same way as for non-participants.

The imputation approach we used begins by assuming that, given the fully observed data on health outcomes as well as socio-demographic characteristics, non-participation in the SHeS-SMR data set is MAR (note that this is already an improvement on standard methods based on unlinked data, for which MAR would not condition on health outcomes). We then accommodated the possibility of the data being MNAR by allowing the distribution of alcohol consumption to differ in a pre-specified manner between the non-participants and participants (given the fully observed characteristics including health outcomes). Sections 3.4.1 and 3.4.2 outline in turn the MI procedures based on MAR and MNAR.

For both MAR- and MNAR-based approaches, one stochastic imputation was performed for each of the M data sets of synthetic non-participant observations produced in stage 2, ultimately yielding M multiply imputed data sets. The imputed data sets were appended and weighted substantive analyses were performed using -mim-[26] in Stata 13.1 (StataCorp, Texas).

3.4.1 MI assuming MAR

Under MAR, conditional on socio-demographic characteristics and health outcomes, the distribution of alcohol consumption is independent of participation status. This is assumption **A2**:

$$\mathbf{A2}: \ [Y|\boldsymbol{X}, \boldsymbol{H}, S = SHeS, R = 0] \ = [Y|\boldsymbol{X}, \boldsymbol{H}, S = SHeS, R = 1]$$

Y is a semi-continuous variable characterised by a combination of zeros representing those who do not report drinking and a right skewed continuous distribution of positive consumption. We followed a previously adopted approach[27] to handle the nontrivial proportion of zero values of alcohol consumption by using a two-part model for Y, splitting it into the dichotomous drinking status variable D and continuous consumption variable Y_i^*, where

$$D = \begin{cases} 1 & \text{if } Y > 0 \\ 0 & \text{if } Y = 0 \end{cases} \tag{2}$$

$$Y^* = \begin{cases} g(Y) & \text{if } D = 1 \\ \textit{undefined} & \text{if } D = 0 \end{cases} \tag{3}$$

Testing of the log transformation $g(y) = \log(y)$ led to a left-skewed distribution of Y^* (a nontrivial proportion of values for alcohol consumption between 0 and 1 unit per week) and hence some extreme imputed values. Instead we used $g(y) = \log(y - k)$, the shifted log transformation[28] with a shift parameter, k, selected to eliminate the skew of Y^* ($\hat{k} = -1.9$, 95% CI: -2.3 to -1.6). Predictive mean matching with a potential match pool of size 10 was used to further improve the imputations.

Our model for $[Y|\boldsymbol{X}, \boldsymbol{H}, S = SHeS, R = 1]$ is therefore

$$\left.\begin{aligned} \textit{logit } p(D = 1) &= \alpha_D + \boldsymbol{\beta_D X} + \boldsymbol{\gamma_D H} + \zeta_D w \\ Y^*|D = 1, \boldsymbol{X}, \boldsymbol{H} &\sim N(\alpha_Y + \boldsymbol{\beta_Y X} + \boldsymbol{\gamma_Y H} + \zeta_Y w, \sigma^2) \end{aligned}\right\} \tag{4}$$

where $\alpha_D, \boldsymbol{\beta_D}, \boldsymbol{\gamma_D}, \zeta_D, \alpha_Y, \boldsymbol{\beta_Y}, \boldsymbol{\gamma_Y}, \zeta_Y, \sigma$ are regression parameters estimated from the complete data. These models were specified jointly in Stata using the conditional imputation option of -ice-,[29] within strata defined by sex and deprivation group to allow the association between harms and alcohol consumption to vary flexibly by sex and deprivation level. The w were entered as a continuous predictor in both equations in equation (4) as well as being included as sampling weights to account for survey design.[30] This procedure produced imputed data sets which allowed correctly for uncertainty in the parameters $(\alpha_D, \boldsymbol{\beta_D}, \boldsymbol{\gamma_D}, \zeta_D, \alpha_Y, \boldsymbol{\beta_Y}, \boldsymbol{\gamma_Y}, \zeta_Y, \sigma)$. Imputed values were back-transformed for use in the substantive model for Y.

3.4.2 MI assuming MNAR

We sought to change imputations of Y to reflect plausible differences between participants and non-participants in Y given a range of theoretically informed plausible departures from the MAR assumption. We did so by specifying how the conditional distribution of Y differs between participants and non-participants and hence altering the imputation model[31] in sensitivity analyses using a pattern mixture approach.[11]

We embed the MAR model in a wider class of models containing sensitivity parameters.[32,33] The sensitivity parameters describe the difference in the joint distribution of fully observed data on participants and partially

observed data on non-participants. Under MAR, the joint distribution $[Y|\boldsymbol{X},\boldsymbol{H}]$ is assumed equivalent for participants and non-participants alike according to assumption **A2**, whereas under MNAR, **A2** is relaxed.

Pattern mixture modelling offers a means to model the joint probability distribution of Y and R, allowing the distribution of Y to differ depending on whether or not Y is observed (equation (5)). Differences between the conditional distributions $[Y|\boldsymbol{X},\boldsymbol{H},S=SHeS,R=0]$ and $[Y|\boldsymbol{X},\boldsymbol{H},S=SHeS,R=1]$ are specified by a set of sensitivity parameters δ. These differences cannot be identified from our data themselves without making untestable distributional assumptions or parameter restrictions.[34] One way of tackling this is to specify the values based on plausible hypotheses about differences between participants and non-participants,[31] with reference to external data or expert opinion.[35] Examples of reference sources are given in section 3.4.3.

Our principal rationale for exploring MNAR concerns differential overall drinking levels, but the possibility remains that D may also deviate from MAR (for instance, for reasons such as lack of social cohesion, non-participants may have a different chance of drinking given their socio-demographics and health outcomes statuses). To assess the sensitivity of results to deviations from the MAR assumption for D, two categories of scenarios were considered for the imputation of D: first, imputing under a MAR assumption, and second, under an 'upper bound' scenario in which all non-participants are set as drinkers for comparison. The rationale for this being of more substantive interest is that those who do not respond in any given sociodemographic and harm group are more rather than less likely to drink, and that it gives an interval which we can be certain about even if we lack a specific plausible deviation from MAR. We then modified the imputation procedure for Y^* as detailed in section 3.4.3.

We consider the general specification which accommodates differential modification of the imputation model by H_1 and by a modifying variable, in our case X_1. Our choice of model is governed by the trade-off between increasing model accuracy against increased difficulty in eliciting plausible δ values

$$Y^*|D=1,\boldsymbol{X},\boldsymbol{H},R\sim N\big(\alpha_Y+\beta_Y\boldsymbol{X}+\gamma_Y\boldsymbol{H}+\zeta_Y w+\ (1-R)\big(\delta_0^{X_1}(1-H_1)+\delta_1^{X_1}H_1\big),\sigma^2\big) \tag{5}$$

Relative to participants with fully observed alcohol consumption, mean alcohol consumption is modified by $\delta_0^{X_1}$ among non-participants who do not experience alcohol-related harms, depending on X_1; and similarly by $\delta_1^{X_1}$ among the non-participants who experienced alcohol-related harms. Clearly, MAR is the case for which $\delta_0^{X_1}=\delta_1^{X_1}=0$ for all X_1.Various scenarios for the magnitudes and signs of the δ parameters are considered in section 3.4.3.

3.4.3 Specifying the parameters governing deviations from MAR

We considered two general approaches to specifying possible values for parameters $\delta_0^{X_1}$ and $\delta_1^{X_1}$. The first uses specific SHeS 'paradata' on fieldwork effort to secure participation (MNAR1 and MNAR2) and the second draws on existing literature-based subject-matter knowledge (MNAR3, MNAR4 and MNAR5). Within each of these, we explored both the MAR (e.g. MNAR1M) and 'upper bound' (e.g. MNAR1UB) approaches for imputing the drinking status (see footnote to Table 4 for full notation).

(a) ***Survey paradata-based approach***

We drew on continuum-of-resistance theory which is predicated upon the idea of a latent propensity to not participate.[2,3,36,37] Here, invited households who do not initially respond are re-approached one or more times, and the number of interviewer calls is recorded. Later responding participants can be theorised to be increasingly more like non-participants with the greater effort required to recruit them into the survey. We used the number of interviewer calls to a household as our proxy for non-participation propensity, where an individual who responded in three (the median number) or fewer attempts is considered an early-participant, and those that took four or more attempts are considered late-participants. Estimates of $\delta_0^{X_1}$ and $\delta_1^{X_1}$ are derived by estimating the mean difference in consumption between early- and late-participants among those who experience alcohol-related harms and those who do not, separately by sex, adjusting for age group and deprivation group. Taking the differences in consumption between early- and late-participants to inform us on the differences between participants and non-participants in this way is speculative in the absence of a more direct proxy but can be thought to represent a conservative MNAR estimate (MNAR1M). Setting the deviation from MAR to be equal to the adjusted difference between early- and late-participants resulted in values for $\delta_0^{X_1=1}=1.0$ and $\delta_1^{X_1=1}=23.1$ among men, and $\delta_0^{X_1=2}=0.75$ and $\delta_1^{X_1=2}=1.17$ among women, respectively.

We also considered the scenario in which the consumption deviation from MAR is twice the adjusted difference between early- and late-participants (MNAR2M and MNAR2UB).

(b) ***Literature-based approach***

A second form of sensitivity analysis considered a range of deviations from the MAR specification based on subject-matter knowledge. A survey in Scotland specifically sampled harmful and dependent drinking in-patients and out-patients attending alcohol addiction services in two Edinburgh hospitals, finding an estimated mean weekly consumption of 198 (95% CI: 185–211) units.[38] For our purposes we posit this to be a generalisable estimate of consumption among drinkers who have been hospitalised. We therefore considered the MNAR-based sensitivity analysis where the imputation model involves specifying $\delta_1^{X_1}$ such that that the resulting overall mean weekly consumption, among those experiencing alcohol-related harm, would equal approximately 198 units. This corresponds to a scenario where the deviation from MAR is five-times the observed sex-specific mean among those whose experienced harm ($\delta_1^{X_1=1} = 309.7$, $\delta_1^{X_1=2} = 91.8$; denoted MNAR5M and MNAR5UB). We also considered more moderate scenarios, where the deviation from MAR consumption was three-times the observed sex-specific mean among those whose experienced harm ($\delta_1^{X_1=1}$185.8, $\delta_1^{X_1=2} = 55.1$; MNAR4M and MNAR4UB;) and finally where the deviation from MAR was equal to the observed sex-specific mean among those whose experience harm ($\delta_1^{X_1=1} = 62.0$, $\delta_1^{X_1=2} = 18.3$; MNAR3M and MNAR3UB). $\delta_0^{X_1=1} = \delta_0^{X_1=2} = 0$ in all these scenarios.

4 Application

4.1 Non-participant synthetic observations (Stage 2)

The SHeS had an overall survey response level of 60% and a proportion of consent to record linkage in Stage 1 of 0.91 with N_{obs} =5381 participants aged 20 to 64 years consenting to linkage. This yielded an effective response level, $L = 54.6\%$. We therefore estimated the total number of participants which would have been observed under full response as $N = 9855$ and the number of non-participant synthetic observations to be generated in Stage 2 as N_{mis} =4474. We chose to draw $M = 70$ bootstrap samples to be imputed, based on the fraction of missing information of 70%.[39]

As a numerical example, consider the category of $(\boldsymbol{X}, \boldsymbol{H})$ defined by men, aged between 40 and 44, residing in the most deprived area quintile, who in 2003–2011 were admitted to hospital with an alcohol-related diagnosis but did not die (i.e., $H_1 = 1$, and $H_2 = 0$). In this category, $P(\boldsymbol{X}, \boldsymbol{H}|S = Pop) = 0.001135$, and using the first bootstrap sample, $P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 1) = 0.001021$. Since also $P(R = 1|S = SHeS) = 0.546$, $P(R = 0|S = SHeS) = 0.454$, equation (1) gives $P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0) = 0.001272$, and $N_{mis} \times P(\boldsymbol{X}, \boldsymbol{H}|S = SHeS, R = 0) = 5.691381$. This figure was randomly rounded up to 6.

After the creation of the synthetic observations, the combined samples were largely successful in reflecting the desired (population representative) socio-demographic composition and health outcome probabilities (Tables 1, 2 and 3).

4.2 MAR-based MI results (Stage 3)

A total of 4903 participants (91.1%) were classed as current drinkers with the remaining 478 participants (8.9%) considered non-drinkers (ex-drinkers or lifetime abstainers). Mean weekly consumption from the survey-weighted estimates was 21.8 units for men and 10.8 units for women. Imputing usual weekly alcohol consumption in Stage 3 using each of the created bootstrap sample data sets under a MAR assumption, resulted in an estimate of 22.4 units (3% increase) among men and 10.8 units (0% change) for women (MAR results in Table 4).

4.3 MNAR-based MI results

(a) ***Survey paradata-based approach***

The first scenario, in which the deviation from MAR is equal to this adjusted difference between early- and late-participants, yielded mean weekly consumption of 23.7 units among men and 11.1 units among women (Table 4, MNAR1M). For the second scenario, in which the deviation from MAR is twice the adjusted difference between

Table 2. Eight-year probabilities of alcohol-related harm in the population, in the Scottish Health Survey participants[a] and the synthetic non-participants in 2003 by sex and area deprivation group.

	Population		Scottish Health Survey					
			Respondents		Synthetic non-respondents		Inferred total	
Area deprivation group	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)
Least deprived	1.9	0.9	1.2	0.7	2.7	1.3	1.9	0.9
2	3.0	1.4	1.3	1.6	4.9	1.2	2.9	1.5
3	4.3	2.0	3.6	2.1	5.0	1.8	4.2	1.9
4	6.6	2.8	4.2	2.2	9.3	3.4	6.5	2.8
Most deprived	11.4	4.2	5.8	2.6	16.4	5.9	11.1	4.2
All groups	5.4	2.3	3.1	1.8	7.6	2.8	5.2	2.2

[a]Those participants consenting to record-linkage of their data.

Table 3. Eight-year probabilities of all-cause mortality in the general population, in the Scottish Health Survey participants[a] and the synthetic non-participants in 2003 by sex and area deprivation group.

	Population		Scottish Health Survey					
			Respondents		Synthetic non-respondents		Inferred total	
Area deprivation group	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)	Men (%)	Women (%)
Least deprived	2.5	1.7	1.1	1.4	3.7	2.1	2.2	1.7
2	3.3	2.2	2.4	1.8	4.2	2.5	3.2	2.1
3	4.4	2.8	2.6	1.7	5.7	3.6	4.1	2.6
4	5.7	3.5	3.8	2.3	6.8	4.2	5.1	3.1
Most deprived	8.3	4.5	7.6	4.6	7.4	3.4	7.5	4.1
All groups	4.8	2.9	3.3	2.3	5.6	3.2	4.4	2.7

[a]Those participants consenting to record-linkage of their data.

early- and late-participants, the figures were 24.9 units (14% increase) and 11.5 units (6% increase), respectively. Table 4, MNAR2M). The corresponding results under the assumption that all non-participants were drinkers gives figures of 25.0 units (15% increase) for men and 11.7 units (8% increase) for women in the first scenario (Table 4, MNAR1UB) and 26.2 units (20% increase) for men and 12.0 units (11% increase) for women in the second scenario (Table 4, MNAR2UB).

(b) ***Literature-based approach***

Among men, adjusted mean consumption under the literature-based scenarios ranged from 24.6 units (13% increase) in the most conservative sensitivity analyses (MNAR3M) to 33.3 units (53% increase) in the most extreme (MNAR5M). Among women, this range was smaller with corresponding figures of between 11.0 units (2% increase) and 11.9 units (10% increase), respectively (Table 3, MNAR3M and MNAR5M). The corresponding results under the assumption that all non-participants were drinkers gave figures ranging from 25.9 units for men and 11.6 units for women (Table 4, MNAR3UB) to 34.7 units for men and 12.5 units for women in the second scenario (Table 4, MNAR5UB).

5 Discussion

Our approach forms an important additional analytic strategy for addressing non-participation in population-sampled studies. The key innovations of our approach are the incorporation of auxiliary topic-relevant data into

Table 4. Weekly alcohol consumption estimates in the Scottish Health Survey 2003 participants[a] and the 'full sample' by sex under various assumptions about the missing data.

	Men					Women				
	δ_0	δ_1	Mean	(95% CI)	% change	δ_0	δ_1	Mean	(95% CI)	% change
Survey-weighted	–	–	21.8	(20.5–23.1)	–	–	–	10.8	(10.1 –11.6)	–
MAR	–	–	22.4	(20.3–24.4)	+3	–	–	10.8	(9.8–11.7)	0
Survey paradata-based approach										
MNAR1M	1.0	23.1	23.7	(21.8–25.6)	+9	0.75	1.17	11.1	(10.1–12.0)	+3
MNAR2M	2.0	26.2	24.9	(22.8–27.0)	+15	1.50	2.34	11.5	(10.5–12.4)	+7
MNAR1UB	1.0	23.1	25.0	(22.4–27.5)	+15	0.75	1.17	11.7	(10.6–12.7)	+9
MNAR2UB	2.0	26.2	26.2	(23.6–28.8)	+20	1.50	2.34	12.0	(11.0–13.1)	+11
Literature-based approach										
MNAR3M	0.0	62.0	24.6	(22.4–26.7)	+13	0.0	18.3	11.0	(10.0–12.0)	+2
MNAR4M	0.0	186	28.9	(26.4–31.5)	+33	0.0	55.1	11.5	(10.5–12.5)	+7
MNAR5M	0.0	310	33.3	(30.1–36.5)	+53	0.0	91.8	11.9	(10.8–13.0)	+10
MNAR3UB	0.0	62.0	25.9	(23.2–28.6)	+20	0.0	18.3	11.6	(23.2–28.6)	+7
MNAR4UB	0.0	186	30.3	(27.2–33.3)	+40	0.0	55.1	12.0	(10.9–13.1)	+11
MNAR5UB	0.0	310	34.7	(31.1–38.3)	+60	0.0	91.8	12.5	(11.4–13.6)	+16

[a]Those participants consenting to record-linkage of their data. δ_0: sex-specific missing not at random based addition to mean alcohol consumption among non-participants who do not experience alcohol-related harms; δ_1: sex-specific missing not at random based addition to mean alcohol consumption among non-participants who experience alcohol-related harms (δ multiples appear non-exact due to rounding); CI: confidence interval; MAR: missing at random; MNAR1M: missing not at random based on survey paradata approach using the adjusted difference between early- and late-participants assuming MAR drinking status; MNAR2M: missing not at random based on survey paradata using twice the adjusted difference between early- and late-participants assuming MAR drinking status; MNAR3M: conservative literature-guided missing not at random approach to deriving delta in which deviation from MAR is equal to the observed sex-specific mean among those whose experienced harm; MNAR4M: intermediate literature-based missing not at random approach to deriving delta in which the deviation from MAR is three-times the observed sex-specific mean among those whose experienced harm; MNAR5M: literature-based missing not at random approach to deriving delta in which the deviation from MAR is five-times the observed sex-specific mean among those whose experienced harm; M: assuming MAR drinking status; UB: the upper bound in which all non-participants are classed as drinkers; %-change: percentage change from survey-weighted estimate.

unit non-response correction in addition to the conventional socio-demographic data, combined with the creation of synthetic observations for non-participants and the application of pattern-mixture modelling to explore sensitivity to plausible departures from the MAR assumption. Resultant alcohol consumption estimates were sensitive to assumptions regarding both drinking status of non-participants and consumption level differences between participants and non-participants. The refined estimates were between 3% and 59% higher among men and up to 16% higher among women relative to the regular survey-weighted estimates. Given that survey-based alcohol consumption estimates scale up to approximately half those indicated by sales data,[40] our higher estimates appear to be most appropriate.

5.1 Strengths and limitations of this study

The first strength of this work is the utilisation of linked survey records enabling the extension of comparisons of participants and the general population from basic socio-demographic variables to health outcomes.[41] We circumvented the challenges associated with gaining rich data characterising the population, and non-participants in particular, by generating synthetic observations for non-participants. The second strength is the application of the much discussed but little implemented 'principled sensitivity analysis'[33] pattern mixture modelling to optimally[42] and transparently specify MNAR models.[43] In cases where no delta values are obviously more realistic than others, Rubin has emphasized the need for easily communicated models[18,43] which are particularly valued by policymakers;[44,45] we found it useful to impose assumptions in order to fix upon a plausible mechanism, considering specific conceivable scenarios in the context of the *a priori* information available.

Limitations include the possibility of distortion arising from survey participants not consenting to record linkage which could explain some of the disparities between health outcomes in the survey samples relative to the general population; however, this only affects 9% of participants and preliminary analyses suggest minimal differences between these groups (data available on request) indicating that this is unlikely to greatly distort

findings. The available alcohol-related harm outcome measures were restricted to the relatively extreme occurrences, hospitalisation and death, with no data on the more frequent occurrences of commonplace harms related to alcohol abuse such as nausea, cognitive impairment and missed working days. This may explain the relatively small changes seen in the MAR estimates despite large survey-population differences in alcohol-related harm.[46] Previous work on refinement of alcohol consumption data in the presence of non-participation has been based on Swedish data[41] which has also considered the implications for impact of the use of augmented data on estimates of consumption prevalence but was based on retrospective alcohol-related hospitalisation data. This offers an alternative approach which does not rely on the attendant passage of time required for follow-up data. This methodology alone is unable to address bias arising from participants mis-reporting their alcohol consumption. It is possible to account for such self-reporting bias by way of incorporation of sales data, for instance.[16]

5.2 Methodological strategy considerations

The following considers possible alternatives approaches in specific steps of the analyses:

(a) As an alternative to our procedure of generating synthetic observations and implementing MI, we could have applied weights or taken a Bayesian-based approach[47] based on health outcome statuses as well as socio-demographic characteristics. It is not clear how to implement MNAR methods with easily communicated models in these approaches.
(b) A possible alternative to creating multiple data sets of synthetic non-participants followed by single stochastic imputation on each is a nested MI procedure where more than one final imputed data set is generated for each first-stage imputed data set. This could be computationally efficient if stage 2 was very slow and stage 3 was relatively fast, which was not the case here, and could help to partition the fraction of missing information between stages 2 and 3, but would require alternative combining rules to Rubin's.[48]
(c) The assumptions and relative merits of our approaches to determining delta values for the pattern-mixture approach are inherently untestable, and there is an array of alternatives to the propensity- and literature-based scenarios, including: *(1) Other within-survey proxy non-participants:* e.g. those with other risky health behaviours such as heavy smoking: this may not form a useful reference point when considering plausibility of delta values; *(2) Expert opinion:*[35,49] we canvassed the broader international alcohol research community through a mailing list (administered by the Kettil Bruun Society) to informally elicit expert opinion; no useable information was gained from this channel; *(3) Record-linked cohort data*: the use of baseline alcohol data on cohort study subjects including those who drop out during follow-up (taken as proxies for non-participants on the basis that they may be somehow similarly 'disengaged') and those remaining in the study follow-up (acting as the corresponding counterparts for survey participants);[50] no such suitable data could be identified; *(4) Retail data*: such external sources of information could be used as the basis a plausible upper bound for population mean consumption; *(5) Worst-case bounds making no assumptions about the missing data*: completely assumption-free approaches, which consider all feasible values of the missing data, generate exceedingly wide bounds for continuous measurements like alcohol and are thus often not directly useful for policy purposes; approaches to narrow such bounds often rely on instrumental variables or longitudinal survey data which were not available in this case.[51]

5.3 Implications

National survey data are crucial resources for quantifying and monitoring trends in health related behaviours with information used for the development, implementation and evaluation of social and public health policy. As such, methodological improvements are of interest to a wide international audience of policy makers and researchers. The development of an effective post-hoc correction procedure for ever-worsening non-response in resource-intensive population-sampled studies offers an enhancement at no additional cost to data collection. This advanced methodology will potentially be applicable to existing and future surveys wherever there is the capacity to record-link surveys with administrative data. Presently, linkage of survey data to routine health records represents a cost-effective means of generating valuable longitudinal data but is performed in very few countries. In exploiting such linkage, our work demonstrates the extended utility of record linkage, providing further impetus for its wider uptake internationally.

Synthetic generation of survey non-participants is not necessary in countries with unique population identifiers and comprehensive linkage (such as the Nordic countries) with the ability to follow-up all individuals regardless of participation status.[14,52] However, possible ethical issues related to accessing outcome data of individuals who have chosen not to participate in a survey may mean that even in such countries stage 1 of our approach might be applicable. Regardless, stages 2 and 3 of our proposed methodology would be applicable in these settings. Our approach to the sensitivity analyses was specific to the context in terms of the estimate of interest and level of participation. Different applications will require distinct approaches to be formulated. In considering the most suitable derived estimates (the higher ones, in our case), we were guided by overall estimates obtainable from external alcohol retail data; dependent on the specific context of the wider applications, reference should be made, where possible, to such relevant sources.

The presented application suggests that non-response may contribute to the general under-estimation of alcohol consumption in survey estimates. There is scope for application to other survey-derived information, which can be discrete – cigarette smoking and obesity, for instance – for which only stage 3 of our procedure would need amending. The outcomes of choice in our application were alcohol-related harms and all-cause mortality on the basis of their strong association with alcohol consumption; single or multiple outcomes can be selected and good candidate outcomes for specific applications are those which have the strongest associations with survey items of interest. Further, non-health external data sources such a taxation or education records could be used to provide auxiliary information to correct for non-participation bias in other research areas. Moreover, this paper describes tackling a single incomplete variable; however, the method can be extended to multiple incomplete variables.

5.4 Further work

The current method requires that the informative data for creating the synthetic non-participants are categorical, since we are determining the missing numbers within discrete cells. It may be possible to incorporate continuous data – such as the number of health outcomes experienced – in a further stage by inferring the distribution among the non-participants such that a value could be assigned to each synthetic non-participant as a draw from that distribution and repeated across multiple replications to allow properly for uncertainty. This would be most appropriately performed by way of MNAR imputation to incorporate information about number of alcohol-related harms from the population comparison data.

Sensitivity analyses could potentially be used to address any differential consumption-outcome associations among area deprivation categories, i.e. allowing for the possibility of interaction effects suggested by the greater levels of alcohol-related harm among the more deprived for equivalent levels of consumption[53] or differential consumption-harm relationships by alcohol product type.[54] The application we describe focussed on a quantity estimate but there is growing recognition that non-representativeness can also lead to bias in estimates of associations.[55] We plan to develop, apply and test our methodology for association estimates.

A major alternative to the pattern-mixture approach to MNAR sensitivity analysis is the selection model approach.[56,57] Selection modelling expresses departures from MAR as coefficients in a logistic regression model for non-participation on alcohol consumption and other covariates: the sensitivity parameter may therefore be less intuitive than in the pattern-mixture framework,[32,35] and hence less easy to relate to subject-matter knowledge.[49] Shared parameter models,[58] in which the measurement of interest and missingness processes are joint modelled, offer yet another option which can be explored.

6 Conclusions

We offer a means to extend the addressing of non-representativeness in survey data beyond the use of conventional inverse probability weights by developing a methodology which harnesses administrative and record-linked data. The key advantage of our approach is the relaxing of the assumption that socio-demographically equivalent participants and non-participants are alike in other ways: the application of the MAR method to administrative health record-linked data is an improvement on the conventional application of survey weights, and the MNAR methods utilise the best available data to make plausible assumptions about how they might differ.

Acknowledgements

We thank Clare Beeston and Mark Robinson from NHS Health Scotland, Julie Landsberg from Scottish Government Health Analytical Services Division, Julie Ramsay from National Records of Scotland, Jim Sherval, from NHS Lothian and Michaela

Benzeval from University of Essex and Lesley Graham from Information Services Division (ISD) Scotland of the NHS National Services Scotland who were advisers on the project. We also thank ISD and the Scottish Government Health Analytical Services Division for performing the record linkage.

Authors' contribution

LG led in the conception of the study design, literature search, and prepared the first draft of the manuscript; EG undertook the analysis and contributed to the elements of the study design and all sections of the paper and the literature search; IRW, SVK, GM and LR contributed to all sections and the literature search; AHL was involved in the conception of the study design, literature search and contributed to all sections.

Ethics and dissemination

Ethics approval of the SHeS was given by the NHS Multi-Centre Research Ethics Committee (MREC03/0/19) and the supply and use of linked data was approved by the Privacy Advisory Committee to the Board of NHS National Services Scotland and Registrar General (PAC 47/12; IR2012-01837). A results-oriented account of the application features in specialist alcohol journal publication.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work is supported by the Medical Research Council Methodology Research Panel under the Population and Patient Data Sharing Initiative for Research into Mental Health grant number (MC_EX_MR/J013498/1). The Medical Research Council and the Chief Scientist Office of the Scottish Government Health Directorates (CSO) fund Linsay Gray, and Alastair H Leyland as part of the MRC/CSO Social and Public Health Sciences Unit's Measurement and Analysis of Socioeconomic Inequalities in Health programme (MC_UU_12017/13 and SPHSU13) and S Vittal Katikireddi as part of the Informing Healthy Public Policy programme (MC_UU_12017/15 and SPHSU15) at the MRC/CSO Social and Public Health Sciences Unit, University of Glasgow; S Vittal Katikireddi is also funded by a NHS Research Scotland Senior Clinical Fellowship (SCAF/15/02). Ian R White acknowledges support from the Medical Research Council (Unit Programme MC_UU_12023/21).

ORCID iD

Linsay Gray https://orcid.org/0000-0002-6918-5037

Supplemental material

Supplemental material for this article is available online.

References

1. Osler M, Kriegbaum M, Christensen U, et al. Rapid report on methodology: does loss to follow-up in a cohort study bias associations between early life factors and lifestyle-related health outcomes? *Ann Epidemiol* 2008; **18**: 422–424.
2. Lahaut VM, Jansen HA, van de Mheen D, et al. Estimating non-response bias in a survey on alcohol consumption: comparison of response waves. *Alcohol Alcohol* 2003; **38**: 128–134.
3. Meiklejohn J, Connor J and Kypri K. The effect of low survey response rates on estimates of alcohol consumption in a general population survey. *PLoS One* 2012; **7**: e35527.
4. Zhao J, Stockwell T and Macdonald S. Non-response bias in alcohol and drug population surveys. *Drug Alcohol Rev* 2009; **28**: 648–657.
5. Maclennan B, Kypri K, Langley J, et al. Non-response bias in a community survey of drinking, alcohol-related experiences and public opinion on alcohol policy. *Drug Alcohol Depend* 2012; **126**: 189–194.

6. Robins JM, Rotnitzky A and Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; **89**: 846–866.
7. Harkanen T, Kaikkonen R, Virtala E, et al. Inverse probability weighting and doubly robust methods in correcting the effects of non-response in the reimbursed medication and self-reported turnout estimates in the ATH survey. *BMC Public Health* 2014; **14**: 1150.
8. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *Brit Med J* 2009; **338**: b2393.
9. Alanya A, Wolf C and Sotto C. Comparing multiple imputation and propensity-score weighting in unit-nonresponse adjustments: a simulation study. *Public Opinion Quarterly* 2015; **79**: 635–661.
10. Peytchev A. Multiple imputation for unit nonresponse and measurement error. *Public Opin Quart* 2012; **76**: 214–237.
11. Little RJA. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993; **88**: 125–134.
12. Molenberghs G, Fitzmaurice G, Kenward M, et al. *Handbook of missing data methodology*. Boca Raton, FL: Chapman & Hall/CRC, 2014.
13. Tompsett DM and Leacy F. On the use of the not-at-random fully conditional specification (NARFCS) procedure in practice. 2018; **37**: 2338–2353.
14. Christensen AI, Ekholm O, Gray L, et al. What is wrong with non-respondents? Alcohol-, drug- and smoking-related mortality and morbidity in a 12-year follow-up study of respondents and non-respondents in the Danish Health and Morbidity Survey. *Addiction* 2015; **110**: 1505–1512.
15. Gray L, McCartney G, White IR, et al. Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol. *BMJ Open* 2013; **3**: e002647.
16. Gorman E, Leyland AH, McCartney G, et al. Adjustment for survey non-representativeness using record-linkage: refined estimates of alcohol consumption by deprivation in Scotland. *Addiction* 2017; **112**: 1270–1280.
17. Frankel MR, Battaglia MP, Balluz L, et al. When data are not missing at random: implications for measuring health conditions in the Behavioral Risk Factor Surveillance System. *BMJ Open* 2012; **2**: e000696.
18. Rubin DB. *Multiple imputation for nonresponse in surveys*. New York: J. Wiley & Sons, 1987.
19. Bromley C, Sproston K, Shelton N, et al. *The Scottish Health Survey 2003. Volume 4: technical report*. Edinburgh: The Stationery Office, 2005.
20. Harley K and Jones C. Quality of Scottish Morbidity Record (SMR) data. *Health Bull (Edinb)* 1996; **54**: 410–417.
21. Fleming M, Kirby B and Penny KI. Record linkage in Scotland and its applications to health research. *J Clin Nurs* 2012; **21**: 2711–2721.
22. Gray L, Batty GD, Craig P, et al. Cohort profile: the Scottish Health Surveys cohort: linkage of study participants to routinely collected records for mortality, hospital discharge, cancer and offspring birth characteristics in three nationwide studies. *Int J Epidemiol* 2010; **39**: 345–350.
23. Lawder R, Elders A and Clark D. *Using the linked Scottish Health Survey to predict hospitalisation & death. An analysis of the link between behavioural, biological and social risk factors and subsequent hospital admission and death in Scotland*. Technical Report, 2007. NHS Health Scotland & Information Services NHS NSS.
24. Hanlon P, Lawder R, Elders A, et al. An analysis of the link between behavioural, biological and social risk factors and subsequent hospital admission in Scotland. *J Public Health (Oxf)* 2007; **29**: 405–412.
25. National Records of Scotland, https://www.nrscotland.gov.uk/statistics-and-data/statistics/statistics-by-theme/population/population-estimates/2011-based-special-area-population-estimates/small-area-population-estimates/mid-2011-to-mid-2014/detailed-data-zone-tables-mid-2013 (2013, accessed 15 April 2019).
26. Royston P. Multiple imputation of missing values: update of ice. *Stat J* 2005; **5**: 527–536.
27. Schafer JL and Olsen MK. Modeling and imputation of semicontinuous survey variables. In: *Proceedings of the Federal Committee on statistical methodology research conference* 1999, pp.565–574, http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.159.7891
28. Box GE and Cox DR. An analysis of transformations. *J R Stat Soc Ser B (Meth)* 1964; 211–252.
29. Royston P and White IR. Multiple imputation by chained equations (MICE): implementation in Stata. *J Stat Softw* 2011; **45**: 1–20.
30. Seaman SR, White IR, Copas AJ, et al. Combining multiple imputation and inverse-probability weighting. *Biometrics* 2012; **68**: 129–137.
31. Carpenter J and Kenward M. *Multiple imputation and its application*. New York: John Wiley & Sons, 2012.
32. Daniels MJ and Hogan JW. *Missing data in longitudinal studies: strategies for Bayesian modeling and sensitivity analysis*. Boca Raton, FL: Chapman & Hall/CRC, 2008.
33. Kenward M, Goetghebeur E and Molenberghs G. Sensitivity analysis for incomplete categorical tables. *Stat Model* 2001; **1**: 31–48.
34. Little RJ and Rubin DB. *Statistical analysis with missing data*. New York: Wiley-Interscience, 2002.
35. White IR, Carpenter J, Evans S, et al. Eliciting and using expert opinions about dropout bias in randomized controlled trials. *Clin Trials* 2007; **4**: 125–139.
36. Lin -F I-F and Schaeffer NC. Using survey participants to estimate the impact of nonparticipation. *Public Opin Quart* 1995; **59**: 236–258.

37. Boniface S, Scholes S, Shelton N, et al. Assessment of non-response bias in estimates of alcohol consumption: applying the continuum of resistance model in a general population survey in England. *PLOS ONE* 2017; **12**: e0170892.
38. Black H, Gill J and Chick J. The price of a drink: levels of consumption and price paid per unit of alcohol by Edinburgh's ill drinkers with a comparison to wider alcohol sales in Scotland. *Addiction* 2011; **106**: 729–736.
39. White IR, Kalaitzaki E and Thompson SG. Allowing for missing outcome data and incomplete uptake of randomised interventions, with application to an Internet-based alcohol trial. *Stat Med* 2011; **30**: 3192–3207.
40. Catto S. *How much are people in Scotland really drinking? A review of data from Scotland's routine national surveys*. Glasgow: Public Health Observatory Division, NHS Health Scotland, 2008.
41. Ahacic K, Kareholt I, Helgason AR, et al. Non-response bias and hazardous alcohol use in relation to previous alcohol-related hospitalization: comparing survey responses with population data. *Subst Abuse Treat Prev Policy* 2013; **8**: 8–10.
42. Little Roderick J. Discussion. *J Off Stat* 2013; **29**: 363–366.
43. Van Buuren S. *Flexible imputation of missing data*. New York: CRC Press, 2012.
44. Katikireddi SV, Bond L and Hilton S. Perspectives on econometric modelling to inform policy: a UK qualitative case study of minimum unit pricing of alcohol. *Eur J Public Health* 2014; **24**: 490–495.
45. Katikireddi SV, Hilton S and Bond L. The role of the Sheffield Model on the minimum unit pricing of alcohol debate: the importance of a rhetorical perspective. *Evidence Policy* 2016; **12**: 521–539.
46. Gorman E, Leyland AH, McCartney G, et al. Assessing the representativeness of population-sampled health surveys through linkage to administrative data on alcohol-related outcomes. *Am J Epidemiol* 2014; **180**: 941–948.
47. Kopra J, Makela P, Tolonen H, et al. Follow-up data improve the estimation of the prevalence of heavy alcohol consumption. *Alcohol Alcohol* 2018; **53**: 586–596.
48. Reiter JP and Raghunathan TE. The multiple adaptations of multiple imputation. *J Am Stat Assoc* 2007; **102**: 1462–1471.
49. Carpenter JR, Kenward MG and White IR. Sensitivity analysis after multiple imputation under missing at random: a weighting approach. *Stat Methods Med Res* 2007; **16**: 259–275.
50. Dawson DA, Goldstein RB, Pickering RP, et al. Nonresponse bias in survey estimates of alcohol consumption and its association with harm. *J Stud Alcohol Drugs* 2014; **75**: 938–4114.
51. Manski CF. *Partial identification of probability distributions*. Berlin: Springer, 2003.
52. Christensen AI, Ekholm O, Gray L, et al. Response to Fergusson & Boden (2015): the importance of considering the impacts of survey non-participation. *Addiction* 2015; **110**: 1514–1515.
53. Katikireddi SV, Whitley E, Lewsey J, et al. Socioeconomic status as an effect modifier of alcohol consumption and harm: analysis of linked cohort data. *Lancet Publ Health* 2017; **2**: e267–e276.
54. Beeston C, Robinson M, Craig N, et al. *Monitoring and evaluating Scotland's alcohol strategy*. *Setting the scene: theory of change and baseline picture*. Edinburgh: NHS Health Scotland, 2011.
55. Munafo MR, Tilling K, Taylor AE, et al. Collider scope: when selection bias can substantially influence observed associations. *Int J Epidemiol* 2018; **47**: 226–235.
56. Kenward MG. Selection models for repeated measurements with non-random dropout: an illustration of sensitivity. *Stat Med* 1998; **17**: 2723–2732.
57. Little RJA. Selection and pattern-mixture models. In: Fitzmaurice G, Davidian M, Verbeke G, et al. (eds) *Longitudinal data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press, 2009, pp.409–431.
58. Follmann D and Wu M. An approximate generalized linear model with random effects for informative missing data. *Biometrics* 1995; **51**: 151–168.