BMJ open
accessible medical research

Use of record-linkage to handle non-response and improve alcohol consumption estimates in health survey data: a study protocol

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

To cite: 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. doi:10.1136/bmjopen-2013-002647

▸ Prepublication history for this paper are available online. To view these files please visit the journal online (http://dx.doi.org/10.1136/bmjopen-2013-002647).

Received 28 January 2013
Accepted 30 January 2013

[1]Social and Public Health Sciences Unit, MRC/CSO Social and Public Health Sciences Unit, Glasgow, UK
[2]NHS Health Scotland, Glasgow, UK
[3]MRC Biostatistics Unit, Cambridge, UK
[4]ScotCen Social Research, Edinburgh, UK

Correspondence to
Dr Linsay Gray;
l.gray@sphsu.mrc.ac.uk

ABSTRACT

Introduction: Reliable estimates of health-related behaviours, such as levels of alcohol consumption in the population, are required to formulate and evaluate policies. National surveys provide such data; validity depends on generalisability, but this is threatened by declining response levels. Attempts to address bias arising from non-response are typically limited to survey weights based on sociodemographic characteristics, which do not capture differential health and related behaviours within categories. This project aims to explore and address non-response bias in health surveys with a focus on alcohol consumption.

Methods and analysis: The Scottish Health Surveys (SHeS) aim to provide estimates representative of the Scottish population living in private households. Survey data of consenting participants (92% of the achieved sample) have been record-linked to routine hospital admission (Scottish Morbidity Records (SMR)) and mortality (from National Records of Scotland (NRS)) data for surveys conducted in 1995, 1998, 2003, 2008, 2009 and 2010 (total adult sample size around 40 000), with maximum follow-up of 16 years. Also available are census information and SMR/NRS data for the general population. Comparisons of alcohol-related mortality and hospital admission rates in the linked SHeS-SMR/NRS with those in the general population will be made. Survey data will be augmented by quantification of differences to refine alcohol consumption estimates through the application of multiple imputation or inverse probability weighting. The resulting corrected estimates of population alcohol consumption will enable superior policy evaluation. An advanced weighting procedure will be developed for wider use.

Ethics and dissemination: Ethics approval for SHeS has been given by the National Health Service (NHS) Multi-Centre Research Ethics Committee and use of linked data has been approved by the Privacy Advisory Committee to the Board of NHS National Services Scotland and Registrar General. Funding has been granted by the MRC. The outputs will include four or five public health and statistical methodological international journal and conference papers.

ARTICLE SUMMARY

Article focus

- To explore and address non-response bias in the health surveys, with a specific focus on alcohol consumption.

Key messages

- National health surveys provide estimates of behaviours in the population—such as levels of alcohol consumption—which inform health policies, but validity depends on their representativeness of the general population. Declining response levels mean that surveys may be increasingly less representative.
- This project aims to compare data from Scottish Health Surveys record-linked to administrative health data sources with corresponding general population data to resolve non-representativeness by using differentials to derive probabilities of alcohol-related hospitalisations and deaths in non-responders; the numbers missing from surveys will be identified by demographic subgroup to simulate observations for non-responders with corresponding alcohol-related harm probabilities and then multiply impute alcohol consumption.
- More accurate alcohol consumption estimation will lead to improved evaluation of interventions and enhanced information for policy. We shall ultimately devise a general application correction factor which will offer a valuable boost to survey-based research.

Primary subject heading: Public health.
Secondary subject heading: Addiction: health policy; mental health.

INTRODUCTION

The large scale of social harms linked to alcohol is increasingly recognised, with

ARTICLE SUMMARY

Strengths and limitations of this study

- The strengths of this work are the reliable utilisation of existing linked survey records and the extension of comparisons of responders and non-responders from basic sociodemographic variables to health outcomes.
- The limitations include the possibility of distortion from non-consent to record linkage of survey responders which could explain some of the disparities between alcohol-related harm outcomes in the survey samples relative to the general population; however, this only affects 7–15% of respondents and is unlikely to greatly distort findings. With the incomplete (around 96%) enumeration level, there is also uncertainty about the representativeness of the Census; although there is a concern that resultant underestimation of the population denominator estimates (but not of the alcohol-related hospitalisation and mortality) may lead to artificially elevated alcohol-related harm estimates (particularly for the most disadvantaged groups), this will be minimised by the limited extent of the population non-enumeration (around 4%).
- The scale of mismatch between survey and population estimates may vary over time because of differences in self-reporting (eg, greater home drinking or more binge drinking), making it increasingly difficult for respondents to estimate their consumption as well as differential non-response levels. Thus, although we may derive a correction method for a particular year, it is potentially invalid to apply it in future years. However, the differential non-response factor is likely to be predominant. Socioeconomic characteristics may change between the time of the survey and the hospitalisation or death event according to the social selection thesis,[50] but this is likely to account for only a very small number of individuals.

alcohol abuse being the most widely perceived social issue in Scotland.[1] Alcohol-related hospital admissions have quadrupled and death rates nearly tripled since the beginning of the 1980s[1]—relative increases which are the steepest in western Europe,[2] with detrimental repercussions for the well-being of the wider population. In response to the escalating problem, the Scottish Government (SG) has launched a strategic approach aimed at reducing alcohol-related harm and helping to address associated health inequalities. The approach encompasses a comprehensive range of interventions, service development and regulatory change—including the possibility of minimum unit pricing of alcohol—aimed largely at the whole population, alongside targeted interventions.[3] Given that alcohol harm is clearly linked to alcohol consumption at the individual[4 5] and population[6] levels, the Strategy aims to reduce population mean consumption, proportions exceeding weekly and daily sensible drinking guidelines, and the prevalence of dependent drinkers and ultimately reduce alcohol-related harm. The SG has tasked National Health Service (NHS) Health Scotland to lead a portfolio of studies—'Monitoring and Evaluating Scotland's Alcohol Strategy' (MESAS).[1] As well as the ultimate reduction of alcohol-related harms, a key outcome for the MESAS evaluation is whether alcohol consumption is reduced.[3] However, reliable ascertainment of alcohol consumption—which is useful in intervention planning as well as in evaluation—is problematic.

Alcohol retail sales data provide the most valid and reliable means of estimating total population alcohol consumption,[7] but they are limited to overall per capita consumption and do not give any information on the amount consumed by individual subgroups (demographic, socioeconomic or geographic) or on the patterns of drinking (eg, binge drinking); they also exclude alcohol purchased abroad and home brewed, and cannot distinguish between transactions made by visitors and residents. In contrast to sales data, health surveys, such as the Scottish Health Survey (SHeS),[8–14] provide estimates of population mean alcohol intake, drinking patterns and differential intake across subgroups.

However, a degree of error is unavoidable with such survey-based measures for two main reasons[15]: distorted self-reporting of intake (which tends to be under-reported for a variety of reasons including systematic underestimation and social desirability bias) and under-representation of groups associated with heavy drinking—men, younger individuals and those from deprived backgrounds, who have higher alcohol consumption than average, tend to be under-represented in surveys.[15] The SHeS suggests no association of alcohol intake with area deprivation[16]—for example, in 2008, 27% of men living in the most deprived quintile according to the Scottish Index of Multiple Deprivation (SIMD) self-reported consumption which exceeded binge drinking thresholds compared with 25% of those in the least deprived quintile.[1] However, the rates of alcohol-related mortality[17] and hospital admissions[18] are much higher in those living in the most deprived areas than in the least deprived areas: in 2009, alcohol-related death rates in the most deprived SIMD quintile (48/100 000 population) were six times those in the least deprived quintile (7/100 000 population); hospital admissions in 2009/2010 were 7.5 times as high.[1] We would thus expect alcohol consumption estimates to be higher with greater deprivation and question the lack of such an association apparent from survey data.

The discrepancy may be explained by one or more of the following: genuinely greater levels of alcohol-related harm among the more deprived for equivalent levels of consumption[5]; differential underestimation of self-reported consumption; a greater spread of drinking patterns within the most deprived areas, that is, a greater proportion of heavy drinkers *and* non-drinkers[19]—as indicated by SHeS data[1]—which averages the higher and lower consumption out in those communities when considering per capita consumption, potentially masking variation within deprivation strata; or differential sampling bias (either due to lower unit response levels in the most deprived areas or a similar response level across quintiles missing more extreme drinkers in the

Table 1 Response levels and alcohol consumption estimates in men in the Scottish Health Surveys, retail-based consumption estimates and population male alcohol-related mortality in Scotland 1995–2011

Survey data						National retail data	National mortality data
Survey year	Household response level (%)	Adult response level (%)	Achieved adult sample	Consent to linkage (%)	Mean alcohol units per week in men	Total volume of pure alcohol sold (1000 l)*	Number of male alcohol-related deaths†
1995	81	84	7932	93	20.1‡	41712	531
1998	77	76	9047	92	19.8‡	43770§	755
2003	67	54	8148	91	19.8¶	47175	1056
2008	61	54	6465	86	18.0¶	50346	971
2009	64	56	7531	85	17.5¶	50842	837
2010	63	55	7245	86	16.0¶	50524	909
2011	66	56	7544	86	15.0¶	48746	815

*Nielsen/CGA Strategy sales in Scotland dataset (off-trade sales in 2011 adjusted to account for the loss of discount retailers).[21]
†General Register Office for Scotland figures for 2011.[51]
‡The 1995 and 1998 surveys were prior to the significant change in the way in which alcohol consumption estimates were derived and are for men aged 16–64 only; thus, they are not comparable with those for 2003 onwards.
§Data not available for 1998—estimate interpolated from available figures for 1995 and 2000;
¶The estimates for the surveys from 2003 onwards are for men aged 16 and over.

most deprived quintile relative to those in the least deprived quintile). It is also possible that the association between alcohol consumption and harm differs between survey responders and non-responders, reflecting, for instance, differential patterns of consumption such as greater concentration of harmful binge drinking among non-responders for equivalent levels of overall consumption, or adverse combinations of different risk factors.[20]

Comparison with the UK sales data previously suggested that survey underestimation of alcohol intake may be as great as 50%,[15] and elevated sales estimates in recent years do not support SHeS-based time trends of reductions in alcohol consumption[21] (table 1). The apparent discrepancy could be explained, at least in part, by progressively increasing survey underestimation of alcohol intake as response levels have fallen—as low as 61% in 2008 at the household level compared with 81% in 1995 (table 1)—if the surveys have become increasingly less representative, especially for those living in deprived areas.[22] The inconsistency of drinking estimates from Scotland's surveys is thus of increasing concern as apparent population trends in consumption are potentially misleading. Addressing the issue is of wider importance for policy design and evaluation which rely on accurate and consistent monitoring of trends in population health.

Correction for under-representation of specific population subgroups can be made by procedures such as inverse probability weighting (IPW),[23] assuming data are 'missing at random' (MAR—see Statistical methodology section). However, the increasingly low response levels remain problematic if respondents and non-respondents with the same sociodemographic characteristics behave differently, for example, in terms of health-related behaviours. The SHeS reports use IPW based on limited sociodemographic characteristics, but since non-participation is likely to be related to heavy drinking,[15] this invalidates IPW based solely on sociodemographic characteristics and the MAR assumption: simply increasing the weight given to the young, deprived male respondents does not address the problem since those sampled are unlikely to be representative of the population of this subgroup.

Previous work on impacts of unit non-response based on studies with varying response rates has generally found that those of lower socioeconomic status in terms of employment,[24] income,[25] education[26] and area deprivation[27] are under-represented. Younger age groups,[28] men, single individuals and those with poorer health status[29] also tend to be under-represented, though this can vary.[28 30] Although estimates of association such as those between socioeconomic position and health outcomes are not generally distorted[29 31] (there are exceptions[26]), prevalence of behaviours related to poor health tends to be underestimated. In an Australian study, participants experienced 10% greater survival relative to the general population,[32] and a Finnish record-linkage study found that the risk of death was underestimated.[31] While previous work on impacts of survey non-response has focused on alcohol, in both a Canadian survey (47% response)[25] and a New Zealand survey (50% response),[27] among others,[33 34] alcohol consumption was found to be underestimated. A Danish survey-based cohort study found that the relatively healthy and affluent participants tended to be have lower risks of alcohol overuse and tobacco-related disease outcomes relative to non-participants.[29] The 'triangulation' of survey and sales data on alcohol to harmonise the survey-based consumption distribution with sales-derived per capita consumption has been demonstrated.[35]

Pilot work conducted by the group based on the 1995 SHeS with follow-up[36] to 2001 aimed to investigate whether respondents were representative of the Scottish population in terms of all-cause mortality and coronary heart disease (CHD) incidence or mortality.[37] Standardised rates of incidence and mortality were calculated by sex for respondents aged 40–64 at the time of the survey and a comparison dataset was created based on population estimates and event registers for the entire Scottish population. Male participants in SHeS had lower than expected mortality from CHD, and women had higher incidence of CHD. Differences were seen for all levels of deprivation but were more pronounced in the most deprived areas and were geographically patterned. This work demonstrated that even with a relatively high response level, participants differ from the population they are intended to represent and reflect a potentially serious bias in health surveys.[37] Separately, in an attempt to resolve the effect of alcohol abstainers in deprived areas, some reanalysis of SHeS data involved removing those who had not drunk in the previous week.[1] While this yielded some of the expected deprivation gradient in alcohol consumption, it was not enough to explain the inequalities in alcohol-related harm, indicating the need for further exploration of the discrepancy between consumption estimates and harms among the most disadvantaged groups (a fuller investigation of this is currently being pursued in a related project and potentially can be considered in an extension of this project).

The aim of this project is to inform the monitoring and evaluation of the SG's Alcohol Strategy by exploiting existing record-linked and population data resources and using advanced statistical methodology to quantify and address unit non-response induced imprecision of national health survey-based estimates of alcohol consumption (weekly intake; binge drinking; problem drinking) in the population of Scotland by age, sex, area deprivation and geographical region. While some attempt shall be made to account for distortion of survey-based estimates due to self-report bias, the main focus of this project is on departure from representativeness, particularly that arising from unit non-response.

METHODS AND ANALYSIS

SHeS are cross-sectional cluster-sampled surveys designed to provide data at both the national and regional level about the health of the population living in private households in Scotland (table 1).[8–14] Scotland is one of the very few countries to have created longitudinal information by way of record linkage of survey data. Individual SHeS data are confidentially linked to prospective and retrospective routine hospital admission data (Scottish Morbidity Records (SMR)) and mortality (from the National Records of Scotland (NRS; formerly the General Register Office of Scotland)).[36 38 39] Despite declining overall survey response levels, the percentage consenting to linkage is high and has remained above 85%.[36] The database is maintained by Information Services Division (ISD) of NHS Scotland; audits have shown that SMR data are around 90% accurate in identifying the correct diagnosis,[40] and SMR completeness is around 99%.[41]

Also available are administrative mortality and hospital admission data for the general population, as well as population estimates derived from routine data.

Robust protocols for identifying individuals with medical conditions attributable to alcohol have been defined and published by NRS/Office of National Statistics and ISD and are used to publish official statistics on alcohol mortality[42 43] and morbidity,[18] respectively. Through partnership with the market research agencies Nielsen Company and CGA Strategy,[22] we have privileged access to alcohol sales data at the national level for Scotland.

Linked SHeS-SMR/deaths

We plan to use the 1995, 1998, 2003, 2008, 2009 and 2010 SHeS records linked to SMR and NRS records, providing a maximum follow-up of around 16 years with adult sample sizes consenting to linkage of 7363, 8305 and 7425 for 1995, 1998 and 2003, respectively, and around 5560, 6400 and 6230 for 2008, 2009 and 2010, respectively. From the SHeS-SMR/NRS records, we have age, sex, area deprivation, health board region and estimates of weekly intake (including an indicator of heavy drinking), binge drinking and problem drinking (all from the survey-component; the latter two measures are available from 1998 onwards) and individually linked alcohol-related hospitalisation and mortality. We are missing all information on the SHeS non-responders, but we can infer their characteristics in terms of age, sex and deprivation based on population estimates (see step 3 below).

General population data

From NRS records, we have mid-year population estimates based on the decennial census (96% enumeration level), mortality, birth, immigration and emigration data. We have population denominators in all survey years for the whole of Scotland and corresponding alcohol-related hospitalisations (SMR) and deaths (NRS) in the general population data—all by age, sex, area deprivation and region—from those years through 2010 as numerators for comparison with the survey data (see step 1 below).

Statistical methodology

We propose to compare survey data and population data to examine how representative the respondents to the SHeSs are in terms of alcohol-related hospitalisations and deaths to inform the improvement of survey-based estimates of alcohol consumption (figure 1). This involves comparing linked records for the survey samples with combined census records, mortality and hospital admission data for the entire population by sociodemographic

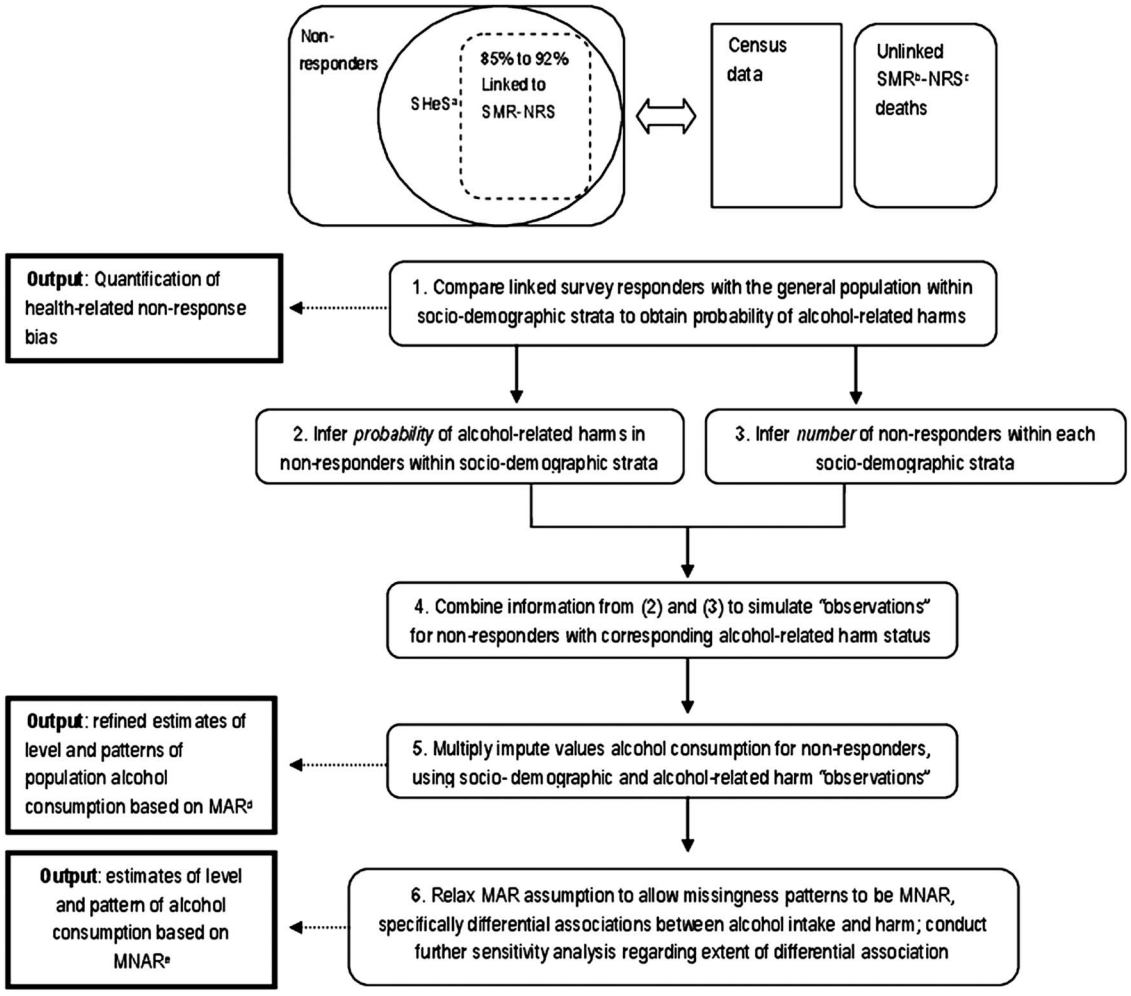

Figure 1 Summary of proposed methodological strategy for addressing survey non-representativeness and refining alcohol consumption estimates. [a]SHeS, Scottish Health Survey; [b]SMR, Scottish Morbidity Record; [c]NRS, National Records of Scotland; [d]MAR, missing at random; [e]MNAR, missing not a random.

subgroups. These comparisons inform on departures from representativeness mainly arising from bias induced by non-response. In the core set of analyses, we shall produce corrected alcohol consumption estimates, assessing the differential effects of varying response levels. We shall additionally develop an advanced correction procedure that can be tailored for different population subgroups and survey response levels for application to other surveys with record-linkage capacity. Finally, we shall inter-relate corrected survey-based consumption estimates and national alcohol sales data to ascertain self-report bias and obtain further refined estimates.

In missing data scenarios, there are a number of possible missingness mechanisms. Data can be missing completely at random (MCAR), MAR or missing not at random (MNAR). If missingness depends on the observed data but not on the unseen data, the missing observations are MAR. In this case, the individuals with complete data (the 'complete cases') are no longer representative and analysing complete cases gives biased estimates. However, under MAR, we can take the predictors of missingness into account in analyses using techniques such as multiple imputation (MI).[44] Imputation is the substitution of some value for a missing data item. Among the imputation techniques available, MI is considered to be superior as it makes reliable estimation of variances and CIs relatively easy. Once all missing values have been multiply-imputed, the datasets can then be analysed using standard techniques for complete data and combined using standard rules.

Alternatively, if the missingness depends on unobserved data (even after taking into account all the information in observed data), the observations are MNAR. In this case, we have to incorporate sensitivity analyses—such as a pattern mixture approach—into MI. A pattern mixture model allows different imputation models for each pattern of missing values under specified MNAR mechanisms with the potential for very general application.[45] We aim to achieve this in step 6 (below) by changing the imputations to allow them to represent likely differences in the

associations between alcohol consumption and alcohol-related hospitalisations and deaths in those observed compared with those with missing alcohol consumption data, by modifying the model intercept term before imputing.

The novel methodological approach which we will use is based on several assumptions: non-response in the (unlinked) SHeS dataset is MNAR; up to step 5 (below), we are assuming given alcohol-related hospitalisations and deaths for responders and non-responders, non-response in the SHeS-SMR dataset is MAR; step 6 goes one stage further, assuming that alcohol-related harm is greater for non-responders than responders for a given level of consumption and attempts to account for this differential relationship.

We propose to:

1. Compare rates of alcohol-related hospitalisations and deaths in the SHeS-SMR/NRS responders with corresponding rates in the general population for each sociodemographic category combination (age, sex, area deprivation and health board region).
2. From 1, estimate the probability of alcohol-related hospitalisations and deaths in the non-responders to the SHeS by each sociodemographic combination (figure 2).
3. From the denominator data of the general population, identify the number of missing respondents within each sociodemographic combination group in the survey.
4. From 2 and 3, simulate the observations for non-responders with the corresponding alcohol-related hospitalisation and death probabilities in each sociodemographic combination group. To our knowledge, this has not been performed previously.
5. Multiply impute unknown alcohol consumption in the simulated 'non-responders' based on sociodemographic characteristics and alcohol-related hospitalisations and deaths under the assumption that the consumption data are MAR.
6. Change the alcohol consumption imputations to reflect the likely difference between responders and non-responders in alcohol consumption for a given probability of alcohol-related hospitalisations and deaths using a pattern mixture model approach which assumes that the consumption data are 'MNAR', given the observed data. The effects of a range of differences will be explored, assuming, for instance, that the risk of mortality is 10% or 20% higher in the non-responders than the responders for equivalent levels of alcohol consumption.

We shall look separately at each round of the survey to see how the change in non-response affects the estimates of alcohol consumption. An advanced correction procedure, quite likely to involve weighting, will be developed that can act differently for different subgroups (especially by deprivation) and survey response levels for application to other surveys with record-linkage capacity.

Further work

Consideration will be given to alternative approaches. These will include use of the SHeS-SMR/NRS to build an imputation model for alcohol consumption, which would then be extrapolated to impute consumption for the entire population; this would be carried out with caution since we would be imputing a high fraction of the data. We shall perform validity checks on any systematic difference between survey participants as a whole and those not consenting to linkage. For instance, we can make overall comparisons of reported alcohol consumption and drinking patterns as well as sociodemographic factors. Additionally, we can potentially use sensitivity analyses to address any differential consumption-outcome associations among deprivation categories, that is, allowing for the possibility of genuinely greater levels of alcohol-related harm among the more deprived for equivalent levels of consumption.[5] Sensitivity parameters would be identified from literature reviews as well as detailed discussion with colleagues with experience in the alcohol field and other experts who could give critical feedback on proposed sensitivity parameters. Integration of corrected survey estimates of alcohol consumption with sales data will allow further refining of estimates.[35] There is a risk that the sociodemographic variables alone will not provide sufficient data for the response model for alcohol consumption. If modelling problems occurred indicating this as a limitation, we would seek the addition of marital status, which is associated with alcohol-related harm[20] and is available from hospital admissions, death certificates and population census records. Should our proposed approach of simulating age, sex and area data for non-responders fail, a method for IPW with MNAR would be considered.[46] Analyses may be complicated by the apparent dichotomy in the drinking behaviour of the most deprived groups who are the most likely not to drink at all, or to drink little within the moderate drinking category but also the most likely to drink at harmful levels.[19] We shall address this by considering a separate variable representing very heavy alcohol consumption, for which missing data would be directly imputed in addition to the other alcohol estimates. We shall also consider the incorporation of estimates of alcohol consumption among those admitted to hospital based on previously developed methodology.[47]

Implications

An optimal means of ensuring survey representativeness is attainment of high levels of response (based on an accurate and up-to-date sampling frame). While this has been achievable in the past, great efforts are required in survey conduct to maintain response levels of around two-thirds in the SHeS at the present time. Our proposed approach forms an important additional strategy to addressing non-response which is applied at the analysis stage.[48] The key innovations of this approach are the simulation of observations for non-responders, and the explicit incorporation of differential associations

$$P(\text{a-r h}^a \text{ in the non-respondents})^b = \frac{[P(\text{a-r h}^a \text{ in the general population})^b - \text{survey response proportion} * P(\text{a-r h}^a \text{ in the respondents})^b]}{(1 - \text{survey response proportion})}$$

Figure 2 Estimating the probability of alcohol-related hospitalisation/mortality in Scottish Health Survey non-respondents from alcohol-related hospitalisation/mortality data on respondents and on the general population of Scotland. [a]a-r h, alcohol-related harm—hospitalisation or mortality from alcohol-related causes; [b]P(x), probability of x.

for non-responders and responders for any given age/sex/deprivation/region combination by factoring in an alternative hospital admission/death rate for the non-responders by implementation of a pattern mixture-based approach. The latter attempts to find plausible sensitivity analyses of departures from data being MAR and fits with the paradigm of 'principled sensitivity analysis',[49] much discussed in the statistical literature but little implemented in practice.

Evaluation of public health policy such as strategies to tackle alcohol problems in Scotland (and beyond) will benefit from enhanced knowledge with the improved estimates of alcohol consumption and prevalence of harmful drinking and dependency which we aim to offer. The detection of changes in behaviour and harms in specific groups such as deprived groups and hazardous drinkers necessary to evaluate the effectiveness of, for instance, minimum unit pricing of alcohol relative to general duty rises will be supported. The accuracy of the assertion that there is a small proportion of the population who drink very heavily and who are responsible for the vast majority of harms may also be elucidated.

There is potential general application of this work beyond alcohol to other survey-derived information—tobacco, diet and physical activity, for instance. Data from population surveys are used extensively and methodological improvements are of interest to a wide international audience. The advanced correction procedure that we aim to create will potentially be applicable to existing and future surveys for improved addressing of non-response bias wherever there is the capacity to record-link surveys with administrative health data. Presently, the linkage of survey data to routine health records represents a cost-effective means of generating valuable longitudinal data, but it is performed in very few countries. In exploiting such linkage to improve conventional survey-based estimates, our work will demonstrate the extended utility of record linkage, providing further impetus for its wider uptake internationally. Simulation of demographic variables for survey non-responders 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 response status. The MI of survey data for non-responders and the pattern mixture aspects of our proposed methodology would nevertheless be applicable in these settings. The prospect of increasing the validity of survey data is increasingly valuable in the context of decreasing survey response, as well as increasing fiscal austerity.

Ethics and dissemination

Ethics approval of the SHeS has been given by the NHS Multi-Centre Research Ethics Committee (MREC03/0/19 for 2003; 07/MRE09/55 for 2008; 08/MRE09/62 for 2009–2011; reference numbers prior to 2003 are unavailable) and the supply and use of linked data have been approved by the Privacy Advisory Committee to the Board of NHS National Services Scotland and Registrar General (PAC 47/12; IR2012-01837). Funding for this work has been granted by the Medical Research Council Methodology Research Panel under the Population and Patient Data Sharing Initiative for Research into Mental Health (MR/J013498/1).

The outputs of the research will include a series of papers which are likely to include.

Public health papers:

1. A baseline assessment of the differential alcohol-related admissions/mortality in the survey samples relative to the general population.
2. Reporting of refined alcohol estimates.
3. Combination with sales data to ascertain self-report bias among responders, and further refine estimates.

Statistical methodological papers:

1. The novel application of pattern mixture modelling for refining survey estimates using record-linked data.
2. Establishing a correction methodology based on the non-response level which can be applied to future surveys.

Data sharing statement

The SHeS[8–14] and combined SHeS-SMR[36 38 39] have been created through substantial investment and are used extensively as the bases of secondary analysis by the research community; release of these anonymised resources is determined by ISD. The value added by this work is the corrective procedure methodology which will be published and hence available to researchers to replicate the enhanced data created by this project, as well as to produce similarly enhanced data from other record-linked surveys. Given this, neither is it possible for us to share, nor is there any benefit to the research community of having access to the specific file created.

Acknowledgements The authors would like to thank Lesley Graham from Information Services Division Scotland and Clare Beeston from NHS Health Scotland who are advisers on the project.

Contributors LG was involved in the conception of the study design, literature search and prepared the first draft of the manuscript; GM and IRW contributed to all sections of the paper; SVK contributed to all sections and the literature search; EG and LR contributed to the introduction and further work sections; AHL was involved in the conception of the study design, literature search and contributed to all sections. All authors read and approved the final manuscript.

Funding 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 (MR/J013498/1).

Competing interests GM is a member of the Scottish Government-funded MESAS evaluation. The remaining authors declare that they have no competing interests.

Ethics approval NHS Multi-Centre Research Ethics Committee and Privacy Advisory Committee to the Board of NHS National Services Scotland and Registrar General.

Provenance and peer review Not commissioned; internally peer reviewed.

REFERENCES

1. 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.
2. Leon DA, McCambridge J. Liver cirrhosis mortality rates in Britain from 1950 to 2002: an analysis of routine data. *Lancet* 2006;367:52–6.
3. Scottish Government. *Changing Scotland's relationship with alcohol. A framework for action*. Edinburgh, 2009.
4. Graham L. *Alcohol Statistics Scotland 2011*. Edinburgh: Information Services Division Scotland, 2011.
5. McDonald SA, Hutchinson SJ, Bird SM, *et al*. Association of self-reported alcohol use and hospitalization for an alcohol-related cause in Scotland: a record-linkage study of 23 183 individuals. *Addiction* 2009;104:593–602.
6. Norstrom T. Per capita alcohol consumption and all-cause mortality in 14 European countries. *Addiction* 2001;96(Suppl 1):S113–28.
7. Robinson M, Thorpe R, Beeston C, *et al*. A review of the validity and reliability of alcohol retail sales data for monitoring population levels of alcohol consumption: a Scottish perspective. *Alcohol Alcohol* 2013;48:231–40.
8. Dong W, Erens B. *Scotland's Health: Scottish Health Survey 1995 (2 Volumes)*. Edinburgh: The Stationery Office, 1997.
9. Shaw A, McMunn A, Field J. *The Scottish Health Survey 1998 (2 Volumes)*. Edinburgh: The Stationery Office, 2000.
10. Bromley C, Sprogston K, Shelton N. *The Scottish Health Survey 2003 (4 Volumes)*. Edinburgh: The Stationery Office, 2005.
11. Bromley C, Bradshaw P, Given L. *The Scottish Health Survey 2008 (2 Volumes)*. Edinburgh: The Scottish Government Health Directorate, 2009.
12. Bromley C, Given L, Ormston R. *The Scottish Health Survey 2009 (2 Volumes)*. Edinburgh: The Scottish Government Health Directorate, 2010.
13. Bromley C, Given L. *The Scottish Health Survey 2010 (2 Volumes)*. Edinburgh: The Scottish Government Health Directorate, 2011.
14. Bromley C, Sharp C, Given L. *The Scottish Health Survey 2011 (3 Volumes)*. Edinburgh: The Scottish Government Health Directorate, 2012.
15. 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.
16. Reid S. Volume 1: main report, chapter 3: alcohol consumption. In: Bromley C, Bradshaw P, Given L. *The Scottish Health Survey 2008*. Edinburgh: The Scottish Government Health Directorate, 2009;55–99.
17. Leyland A, Dundas R, McLoone P, *et al*. *Inequalities in mortality in Scotland 1981–2001. Occasional paper 16*. Glasgow: MRC Social and Public Health Sciences Unit, 2007.
18. Information Services Division. *Alcohol-related Hospital Statistics 2010*. Edinburgh: Information Services Division 2010.
19. Health Analytical Services Division Scottish Government. Alcohol consumption and harm across income groups, May 2010.
20. Lawder R, Grant I, Storey C, *et al*. Epidemiology of hospitalization due to alcohol-related harm: evidence from a Scottish cohort study. *Public Health* 2011;125:533–9.
21. Robinson M, Beeston C, Mackison D. *Monitoring and evaluating Scotland's alcohol strategy: an update of alcohol sales and price band analyses*. Edinburgh: NHS Health Scotland, 2012.
22. Robinson M, Craig N, Beeston C, *et al*. *Monitoring and evaluating Scotland's alcohol strategy: an update of alcohol sales and price band analyses*. Edinburgh: NHS Health Scotland, 2011.
23. Seaman SR, White IR. Review of inverse probability weighting for dealing with missing data. *Stat Methods Med Res* 2011 [Epub ahead of print].
24. Goldberg M, Chastang JF, Leclerc A, *et al*. Socioeconomic, demographic, occupational, and health factors associated with participation in a long-term epidemiologic survey: a prospective study of the French GAZEL cohort and its target population. *Am J Epidemiol* 2001;154:373–84.
25. Zhao J, Stockwell T, Macdonald S. Non-response bias in alcohol and drug population surveys. *Drug Alcohol Rev* 2009;28:648–57.
26. Lorant V, Demarest S, Miermans PJ, *et al*. Survey error in measuring socio-economic risk factors of health status: a comparison of a survey and a census. *Int J Epidemiol* 2007;36:1292–9.
27. Meiklejohn J, Connor J, Kypri K. The effect of low survey response rates on estimates of alcohol consumption in a general population survey. *PLoS ONE* 2012;7:e35527.
28. Schneider KL, Clark MA, Rakowski W, *et al*. Evaluating the impact of non-response bias in the Behavioral Risk Factor Surveillance System (BRFSS). *J Epidemiol Community Health* 2012;66:290–5.
29. 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;18:422–4.
30. Vinther-Larsen M, Riegels M, Rod MH, *et al*. The Danish Youth Cohort: characteristics of participants and non-participants and determinants of attrition. *Scand J Public Health* 2010; 38:648–56.
31. Harald K, Salomaa V, Jousilahti P, *et al*. Non-participation and mortality in different socioeconomic groups: the FINRISK population surveys in 1972–92. *J Epidemiol Community Health* 2007;61:449–54.
32. Hockey R, Tooth L, Dobson A. Relative survival: a useful tool to assess generalisability in longitudinal studies of health in older persons. *Emerg Themes Epidemiol* 2011;8:3.
33. 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;38:128–34.
34. Makela P, Paljarvi T. Do consequences of a given pattern of drinking vary by socioeconomic status? A mortality and hospitalisation follow-up for alcohol-related causes of the Finnish Drinking Habits Surveys. *J Epidemiol Community Health* 2008;62:728–33.
35. Rehm J, Kehoe T, Gmel G, *et al*. Statistical modeling of volume of alcohol exposure for epidemiological studies of population health: the US example. *Popul Health Metr* 2010;8:3.
36. 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;39:345–50.
37. Leyland AH, Finlayson A, Clark D, *et al*. Assessing the representativeness of health surveys. *Eur J Public Health* 2004; 14(4 Suppl):45.
38. 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;29:405–12.
39. Lawder R, Elders A, 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. Edinburgh: NHS Health Scotland & Information Services NHS NSS, 2007.
40. Harley K, Jones C. Quality of Scottish Morbidity Record (SMR) data. *Health Bull (Edinb)* 1996;54:410–17.
41. Managing Data Quality: SMR Completeness. http://www.isdscotland.org/Products-and-Services/Hospital-Records-Data-Monitoring/SMR-Completeness/ (accessed 10 Jan 2013).
42. Grant I, Springbett A, Graham L. *Alcohol attributable mortality and morbidity: alcohol population attributable fractions for Scotland*. Edinburgh: Information Services Division National Services Scotland, June 2009.
43. The Scottish Government Alcohol Information Scotland. ICD Codes.

44. Sterne JA, White IR, Carlin JB, *et al*. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ* 2009;338:b2393.
45. Little R. Pattern-mixture models for multivariate incomplete data. *J Am Stat Assoc* 1993;88:125–34.
46. Dufouil C, Brayne C, Clayton D. Analysis of longitudinal studies with death and drop-out: a case study. *Stat Med* 2004;23:2215–26.
47. NHS Quality Improvement Scotland. Understanding Alcohol Misuse in Scotland: HARMFUL DRINKING Final Report, 2008.
48. 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;126:189–94.
49. Kenward M, Goetghebeur E, Molenberghs G. Sensitivity analysis for incomplete categorical tables. *Stat Model* 2001;1: 31–48.
50. Hart CL, Davey Smith G, Upton MN, *et al*. Alcohol consumption behaviours and social mobility in men and women of the Midspan Family study. *Alcohol Alcohol* 2009;44:332–6.
51. General Register Office for Scotland. *Alcohol-related deaths, by sex and age-group, Scotland, 1979 to 2011*. Edinburgh: General Register Office for Scotland, 2012. http://www.gro-scotland.gov.uk/files2/stats/alcohol-related-deaths/ard-2011-table1.pdf (accessed 10 Jan 2013).