The impact of handling missing data on alcohol consumption estimates in the UK Women Cohort Study

U. Nur[1], N.T. Longford[2], J.E. Cade[3] and D.C. Greenwood[3]

[1]Cancer Research UK Cancer Survival Group, Non-Communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel St, London WC1E 7HT, UK.
[2] Departament d'Economia i Empresa, Universitat Pompeu Fabra (UPF) Ramon Trias Fargas 25-27, 08005 Barcelona, Spain
[3] Centre for Epidemiology and Biostatistics, University of Leeds, Leeds LS2 9LN, UK

Abbreviations: UKWCS, UK Women Cohort Study; FFQ, Food frequency questionnaire; WCRF, World Cancer Research Fund; MAR, Missing at random.

INTRODUCTION

Food supplies energy and provides essential nutrients needed for body functions. The three basic nutritional components of food are protein, carbohydrates and fats, known as the macronutrients. After being converted into simpler compounds, the body uses them as a source of energy.

Alcohol is also a source of energy but, unlike the macronutrients, it is not essential to the body. There has been a long debate about the effects of alcohol on the body. Moderate consumption of wine has been found to have a protective effect against coronary heart disease and cancer (1). The joint World Health Organization/Food and Agriculture Organization report (2) stated that low or moderate consumption of alcohol lowers the risk of coronary heart disease. However, other cardiovascular and health risks associated with alcohol do not favor a recommendation for its consumption. In developed countries, alcohol is considered as one of the main risk factors for cancers of the oral cavity, pharynx and esophagus, and 75 percent of such cancers are attributed to alcohol and tobacco (3) The same report states that excessive alcohol consumption is the main dietary risk factor related to cancer of the liver. Several studies have found that alcohol is the main dietary factor which increases the risk of breast cancer, with around 10 percent increase in the risk for an average one alcoholic drink per day(4). In some societies and communities, alcoholism is regarded as a stigma, especially among women, and inquiring about alcohol consumption in surveys presents considerable challenges to questionnaire design and interview protocol. Questions about alcohol consumption may be 'masked' by inserting them among questions about lifestyle, diet, smoking habits and exercise (5). Despite these arrangements, there is a lot of evidence that alcohol consumption is under-reported in surveys. Non-response, and the way it is handled in the analysis, is one reason for it.

Extensive and reliable datasets are essential to investigate the link between alcohol intake and disease. Missing values are a problem in most large-scale surveys that have

extensive questionnaires. They indicate that the data-collection protocol has not be adhered to, so less information was collected than planned. In a typical analysis, the incomplete records are either discarded or completed. Both these generic approaches, *data reduction* and *data completion*, are deficient; the former because some valuable information is not used, and the latter because by analyzing a single completion we pretend to have more information than was collected. The impact of missing data on estimates in epidemiological and biomedical studies is substantial (6-9).

The analysis of the complete records (by data reduction or listwise deletion) may yield inferences substantially different from those that would be obtained had no values been missing. In the study we analyze, the estimates of alcohol consumption based on complete records are biased. The practice established at present is to impute zero for each missing value for a subject's consumption. The rationale for this is that zero is the modal (most frequent) value. However, it is also the extreme (minimum) value that could be recorded, and so such imputation leads to under-representation of the alcohol consumption.

The mean alcohol nutrient intake estimated by the three methods we consider, data reduction, single imputation and multiple imputation, is 7.5, 8.6 and 11.3g/week (grams of net alcohol per week), respectively. Given the substantial sample size, in excess of 35,000, the differences among the estimates are mainly due to the bias of at least two of the estimates. The substantially greater estimate obtained by multiple imputation is due to exploiting the information in the incomplete records, in which the consumption declared tends to be greater than in the complete records.

Multiple imputation is based on a small number of alternative data completions that are generated by a process that entails some randomness. If this process faithfully reflects our uncertainty about the missing values, the multiple imputation estimator is nearly unbiased and nearly efficient. Important prerequisites for this are that the estimator used would have been

unbiased and efficient had the data been complete, and that its sampling variance would have been estimated without bias. In brief, multiple imputation limits the damage caused by the non-response (missing values), but cannot make up for the deficiencies in the complete-data estimator.

Analyzing the complete cases is the default approach for all those who do not appreciate the impact that missing values may have on the results. The approach forces the data into a rectangular form, which can be analyzed by the same method and software as was planned or contemplated prior to data collection. The price for this convenience is that a large fraction of the sample may be excluded. The retained (complete) cases may no longer be a representative sample, even if the original sample is, because the subjects with incomplete records may in some way be systematically different from those with complete records. This problem is addressed by (10) in conjunction with hot-deck imputation.

A missing response to a question about the quantity of alcohol consumed may be interpreted as zero – the respondent may have forgotten the instruction stated at the beginning of the questionnaire to draw a distinction between no consumption ('Enter zero as the response') and not responding for one reason or another. This motivates imputing zero for each missing value. This is clearly problematic for sequences of questionnaire items, when the respondent has given up completing the remainder of the questionnaire (dropped out), or was distracted and skipped a page or a section of the questionnaire. In this paper, we compare three methods for estimating the mean alcohol intake of the population represented by the UK Women Cohort Study (UKWCS) (11): data reduction, imputation of zero (the modal value) and multiple imputation, and discuss some extensions.

MATERIALS AND METHODS

The survey

The UK Women's Cohort Study (UKWCS) aims to make inferences about the relationship between diet and cancer incidence and mortality (from selected causes) in a group of UK women who were middle-aged in the mid-1990's. The original survey targeted women residing in England, Wales and Scotland. A 217-item food frequency questionnaire (FFQ) was sent to 65,000 women who earlier declared their support for the World Cancer Research Fund (WCRF). All women aged between 35-69 in 1995 and who described themselves as vegetarians in an earlier WCRF survey, were included in the cohort.

Each of these women was matched with a woman who declared that she eats meat and who was in the same 10-year age band; all fish eaters were also included (12, 13). Women were then contacted by post with a request to complete an extensive questionnaire about their diet and lifestyle. About 35,000 women responded, approximately a third of whom described themselves as being vegetarian, a third as red-meat eaters and a third as fish eaters.

The extent of missing data

Information on alcohol consumption was collected in two parts of the UKWCS questionnaire. The first part of the alcohol consumption questions consists of a block of five items in the form of FFQ. For each item there were ten response options ranging from "never" (coded as 0) to "six or more times per day" (coded 9), in response to the question:

"*How often have you eaten these foods in the last 12 months*?",

common to a long sequence of items. Information on alcohol consumption was also collected by asking to state the number of specified units (pints, glasses or measures) of each type of alcoholic beverage (beer or cider, wine, sherry or fortified wines, and spirits) per week. The question was then repeated, asking about the intake five years previously. For brevity, we

refer to this set of items as ‘recall’ (recent and five years ago). Note that beer and cider is treated as a single category in the recall items.

The rate of non-response to the recent recall items, a focus of this paper, ranged from 18 percent for wine to more than 52 percent for beer and cider. For the same question relating to five years ago, the response rate was just as low, from 18 percent for wine consumption to 52 percent for the consumption of beer and cider (Table 1).

A subject’s total alcohol nutrient intake

The overall nutrient intake of a subject is estimated by adding up the products of the reported frequency of each food by the amount of nutrient in a specified portion of that food. The total alcohol nutrient intake of a subject is then estimated by adding up the intake of the different types of alcohol consumed per week. For example, the total alcohol nutrient intake of a subject who reported consuming 2 pints of beer, 3 glasses of wine, 2 glasses of sherry and a glass of spirits per week is

$$\{(2 \times (2\times287) \times 4.53) + (3 \times 125 \times 9.25) + (2 \times 40 \times 16.65) + (23 \times 31.70)\} / 100$$

$$= 107.30 \text{ g/week.}$$

Here 287, 125, 40 and 23 are the quantities (masses), in grams, of a pint of beer or cider, a glass of wine, sherry and spirits respectively; 3.08, 5.98, 9.25, 16.65 and 31.70 are the quantities of the alcohol nutrient in 100 grams of beer, cider, wine, sherry and spirits, respectively. As beer and cider were combined in the same question, the nutrient intake in 100gms was calculated as the average nutrient of beer and cider (Table 2) (14).

Statistical analysis

When only a small fraction of the records are incomplete (say, five percent or less), data reduction may be a reasonable solution. The complexity of more involved methods for dealing with the missing values is hard to justify when they are unlikely to yield substantially different estimates

A blank response to an item about consumption of a food or beverage may in some circumstances be appropriately interpreted as 'no consumption'. The subject may have skipped the item believing that the blank response would be interpreted as such. Certainly, there is much less rationale for interpreting a blank response as any particular positive quantity, except perhaps as an excessive quantity, if the subject wishes not to disclose a habit she regards as undesirable or departing from some perceived norm. In this context, imputing zeros for missing responses to items about alcohol consumption is an easy target for criticism. For example, the survey would fail to capture information about the extent of binge drinking.

Multiple imputation

In multiple imputation, a model is posited for the association of the missing values with the recorded values. Replacements, called *plausible values,* are generated using this model for each missing value. We assume that a *complete-data analysis* (or method) is available – it is a method and software implementing it that would be appropriate (and efficient) if the collected data were complete; in most settings it is the analysis that would be applied if the data were complete. The analyst in charge may be familiar with this analysis and would like to apply it, ideally, without any alterations.

One set of plausible values completes a dataset, and it can be analyzed by the complete-data method. In multiple imputation, several (replicate) sets of plausible values are generated, with a completed dataset for each of them, yielding replicate complete-data

estimates and estimates of the associated sampling variances. The average of these completed-data estimates is the multiple-imputation (MI) estimate. Its sampling variance is estimated by the average of the completed-data sampling variances, inflated by the between-completion variance.

Let $\hat{\theta}_m$, $m = 1, \ldots, M$, be the set of completed-data estimates of a population quantity θ, and let $\hat{s}_m^2$ be the estimates of the completed-data sampling variances. Then the MI estimator of θ is defined as

$$\tilde{\theta} = \frac{1}{M}\sum_{m=1}^{M}\hat{\theta}_m$$

and its sampling variance is estimated by

$$\tilde{s}^2 = \hat{U} + (1+1/M)\hat{B},$$

where $\hat{U} = \frac{1}{M}\sum_{m=1}^{M} s_m^2$ and $\hat{B} = \frac{1}{M-1}\sum_{m=1}^{M}(\hat{\theta}_m - \tilde{\theta})^2$. The first term, $\hat{U}$, estimates the sampling variance of $\hat{\theta}$ that would be attained if the data were complete. The second term can be interpreted as the variance inflation due to non-response; $\hat{B}$ estimates this inflation if infinitely many completions were applied ($M \to \infty$), and the additional term, $\hat{B}/M$, is due to using only a finite number of completions (imputations).

The plausible values have to reflect the uncertainty about the missing values for which they are intended. In most applications this entails two sources of uncertainty: about the model parameters and about the missing values conditionally on the values of the model parameters. The former source is accounted for by using sets of plausible parameter values, drawn at random from the estimated sampling distribution. Each set of plausible values is based on a different set of plausible parameters. The latter source is due to the variation inherent in the posited model even when the model parameters are known. For example, in

the simple model of independent replicates of an event with binary outcomes (Yes/No), there is uncertainty about the probability of the positive outcome, but even if this probability were known there would still be uncertainty about the outcome because it is subject to chance. Similarly, in a linear regression model, there is uncertainty about the model parameters and uncertainty about the outcome due to the residual variation; the latter is present even when the model parameters are known.

Validity of the model for imputation is an important assumption of the MI method that cannot be ascertained. In many settings, we can merely define more general models, which improve the chances of attaining validity, or coming sufficiently close to it. For the theoretical background, see (15), and for applications and examples (16). Central to the applications is the assumption of the data *missing at random* (MAR), according to which, with a suitably specified conditioning, there are no systematic differences between the missing and the available data. Then the model that is applied for the available data, which can be fitted relatively easily, applies also to the missing data, providing us with a prescription for generating plausible values. If a model is valid and the condition of MAR is satisfied, then a more general model is also valid and MAR is also satisfied. This gives a rationale for using as complex models as is feasible for generating plausible values.

In a typical application of MI with an incompletely recorded continuous variable *y*, an ordinary regression model $y = \mathbf{X}\beta + \varepsilon$ is fitted to the complete records, assuming that with the conditioning in this regression the MAR condition is satisfied. A plausible residual variance $\tilde{\sigma}^2$ is then drawn from the scaled χ^2 distribution which estimates the sampling distribution of the residual variance estimate $\hat{\sigma}^2$. A plausible variance matrix of the regression parameter estimates is $\tilde{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$ and a plausible vector of regression parameters is generated as $\tilde{\beta} = \hat{\beta} + \gamma$, where γ is a vector drawn at random from the multivariate normal

distribution with zero mean and variance matrix $\tilde{\sigma}^2(\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}$. Finally, the plausible values are generated according to the 'plausible' model formula $\mathbf{y} = \mathbf{X}\tilde{\beta} + \varepsilon$, with a random sample ε from the normal distribution with zero mean and variance $\tilde{\sigma}^2$. Difficulties arise when there are missing values also in $\mathbf{X}$. We applied multiple imputation by chained equation (17) as data are incomplete on all alcohol variables. This method which is sometimes referred to as variable by variable multiple imputation, assumes that a multivariate distribution exists, without specifying a specific form for it, and that draws from it can be generated by Gibbs sampling the conditional distributions. i.e. the multivariate problem is split into a number of univariate problems. The procedure of generating plausible values proceeds as follows: -

- Fill in missing values for each incomplete variable by a starting value, in this application this is chosen the mean for the continuous variables.
- Discard the filled-in values from the first variable leaving the original missing values. The missing values are then imputed using linear regression, conditioning on the other four variables as described below.
- The filled-in values are discarded from the second variable. These missing values are then imputed using linear regression imputation.
- The procedure is repeated for each variable in turn. Once each variable has been imputed, we have then completed one iteration.
- The same procedure is repeated for several (in this case 10) iterations. This generates one complete dataset.
- For m completed datasets, repeat the procedure m times independently.

Application

Imputation models were specified to generate plausible values for each type in the recall section of the questionnaire by the simple linear regression of the current consumption on the consumption of the four alcohol types five years earlier, and the three remaining current alcohol consumption types. Note that the assumption of normality is particularly problematic because a large fraction of the outcomes are zeros. It is argued in the literature on MI (15, 18), that this assumption is unimportant. For an approach that addresses the problem of excess zeros among the outcomes, see (9).

The estimated correlations of the consumption of each alcohol type currently and five years earlier are 0.60, 0.81, 0.77 and 0.71 for beer, wine, sherry and spirits, respectively. They are based on all the available data. The strong association for each alcohol type suggests that the past consumption is useful in an imputation model for (missing) current consumption

Ten completed datasets were generated.

RESULTS

Complete-case analysis

The complete-case analysis of nutrient alcohol consumption is based on only 12,571 (36 percent) records that have complete data, see Table 3. Such a large reduction of the data raises two issues. First, having collected less data than anticipated, we have less information than planned. Second, the subjects who fail to respond (to an item or a block of items) may tend to differ (systematically) from subjects whose records are complete. A naïve analysis based on just one third of the data would very likely be biased, both for simple summaries and more involved inferences about associations of variables, e.g., those based regression models.

Imputing a default value

By imputing zeros for all missing responses in current consumption, all subjects could be included in the analysis. The estimate of the mean alcohol nutrient intake increased from

7.75g/day in the complete case analysis to 8.60g/day. This increase may at first appear counterintuitive; after all, we have imputed the smallest possible value for each missing value. Among the subjects included in this analysis, but not in the analysis of complete cases, there are disproportionately many high consumers; with their inclusion in the analysis, the estimate of the mean is greater. For example, if a subject reported that she currently consumes a lot of beer and wine, and did not respond to the questions about spirits and sherry, her record would not contribute to the analysis of complete cases. With zeros imputed for spirits and sherry consumption in her record, it now contributes to the single-imputation analysis by the lowest plausible amount.

Multiple imputation

The analysis by multiple imputation is based on information from 34,465 records. A small fraction of the subjects (902, 2.5 percent) were excluded from the analysis because the recall of the current alcohol consumption and that of five years earlier was missing for the four types of alcohol. We generated twenty sets of plausible values for the current consumption. Generating more sets does not present any problems as the (additional) data storage requirements are not excessive. We justify the choice of twenty sets post hoc, by comparing the estimated variance inflation $\hat{B}/M$ with the remainder of the sampling variance, $\hat{U}+\hat{B}$, which cannot be reduced.

The MI estimate of mean alcohol intake is much greater than with data reduction and single imputation, see Table 3. For example, the MI estimate of the alcohol intake is 13.84g/day and the zero-imputation estimate 8.60 g/day. The difference is due to imputing many large values (as opposed to zero as the default). Many plausible values are large because they are informed by the substantial consumption of the same type of alcoholic beverage in the past. The estimated standard errors obtained by the analysis of complete cases are greater than by the two other methods, because only a fraction of the records are

used. The estimated standard errors for the MI estimate are greater than for the zero-imputation estimate; however, the difference (0.056 vs. 0.088) is minute when compared to the likely bias. In any case the figure for zero-imputation underestimates the standard error because it is based on much more data than was collected. Apart from the rather complex theory in (15), we can argue that the MI estimator is more appropriate because some of the subjects who declared that they consumed a type of alcoholic beverage five years earlier are bound to have consumed some also recently. Evidence of this is borne out by the regression models used for imputation. Of course, single-imputation methods more complex than zero imputation can be devised. For example, the value from five years ago, when available, could be imputed for the current consumption. But every such method can be improved by its MI version in which the uncertainty about the fitted values is duly reflected.

DISCUSSION

We compared three methods of dealing with non-response in making inferences about the mean alcohol nutrient consumed and showed that ignoring non-response by reducing the data to complete records underestimates the mean. A lot of information contained in the non-empty records is discarded. Imputing a default value, in this case zero, also results in biased estimates, even though much more of the available information is brought to bear on the result. By pretending that we know the value of each missing item we underestimate the sampling variance; in our application this is of next to no importance when compared to the substantial bias we incur. The method of multiple imputation had two strengths: the information contained in most of the incomplete records is used and the estimates inherit the properties of the complete-data method – unbiasedness and unbiased estimation of the sampling variance. These properties are contingent on the appropriate model for imputation; however, the model we employed, is a substantial improvement on the model that can be associated with zero-imputation (‘missing’ mean zero), and the model associated with data

reduction (incomplete records are like complete records without any conditioning). A practical advantage of MI is that the software intended for the analysis when no non-response was anticipated can be used without any alterations, even though the application has to be repeated several times.

Even when applied with an imperfect model, MI is clearly superior to single imputation. The imperfection of the model we applied is due to its simplicity and obvious departure from the assumptions of normality. More complex models can ameliorate this problem.

The work associated with an application of MI can be split between an analyst with an expertise in MI who is acquainted with the data collection and non-response processes, who generates the sets of plausible values, and a (secondary) analyst whose expertise is only in the complete-data methods. The instructions that have to be given to this analyst (beyond those for analyzing a complete dataset) are simple and involve no complexity additional to an application of the complete-data method. The first analyst's product, sets of plausible values, can be used for several analyses.

TABLE 1. Response to the recall set of questions about alcohol consumption.

	Current intake			Intake five years before		
Alcohol	**Recorded**	**Missing**	**(%)**	**Recorded**	**Missing**	**(%)**
Beer or Cider	16,877	18,490	52.3	16,973	18,394	52.0
Wine	28,937	6,430	18.2	28,879	6,488	18.3
Sherry	17,122	18,245	51.6	17,459	17,908	50.6
Spirits	27,620	7,747	21.9	20,629	14,738	41.7

TABLE 2. Alcohol nutrient in a pint of beer and a glass of cider, spirits and sherry.

Alcohol	**Grams per pint/glass**	**Alcohol nutrient/100g**
Wine	125	9.25
Beer	287	3.08
Cider	287	5.98
Spirits	23	31.70
Sherry	40	16.65

TABLE 3. The impact of handling missing data by the complete case analysis, imputing zeros and multiple imputation, on alcohol nutrient intake.

Alcohol (g/day)	**Complete-Case Analysis**			**Imputing Zero**			**Multiple Imputation**		
	Obs.	**Est.**	**SE**	**Obs.**	**Est.**	**SE**	**Obs.**	**Est.**	**SE**
Wine	28,937	6.72	0.047	35,367	5.50	0.041	34,465	6.88	0.047
Beer/Cider	16,877	3.11	0.055	35,367	1.48	0.028	34,465	4.11	0.051
Spirits	27,620	1.45	0.019	35,367	1.11	0.015	34,465	1.54	0.019
Sherry	17,122	1.02	0.015	35,367	0.49	0.008	34,465	1.31	0.015
Total alcohol Intake g/week	12,571	7.75	0.098	35,367	8.60	0.056	34,465	13.84	0.088

References

1. Gronbaek M et al. Type of alcohol consumed and mortality from all causes, coronary heart disease, and cancer. Ann.Intern.Med. 2000;133:411-9.

2. World Health Organization, Food Agricultural Organization. Diet, Nutrition and the Prevention of Chronic Diseases. 2003. World Health Organization. WHO technical report series.
   Ref Type: Report

3. International Agency for Research on Cancer. Cancer: causes, occurrence and control. 1990. Lyon, IARC Scientific Publications.
   Ref Type: Report

4. Smith-Warner SA et al. Alcohol and breast cancer in women - A pooled analysis of cohort studies. Jama-Journal of the American Medical Association 1998;279:535-40.

5. Paton A. ABC of alcohol. BMJ Publishing Group, 1994.

6. Rubin DB, Schenker N. Multiple Imputation in Health-Care Databases - An Overview and Some Applications. Statistics in Medicine 1991;10:585-98.

7. Roth PL. Missing Data - A Conceptual Review for Applied Psychologists. Personnel Psychology 1994;47:537-60.

8. Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology 1995;142:1255-64.

9. Longford NT et al. Handling missing data in diaries of alcohol consumption. Journal of the Royal Statistical Society Series A-Statistics in Society 2000;163:381-402.

10. Nur U et al. Dealing with incomplete data in questionnaires of food and alcohol consumption. Statistics in Transition 2005;7:111-34.

11. Cade J et al. Costs of a healthy diet: analysis from the UK Women's Cohort Study. Public Health Nutrition 1999;2:505-12.

12. Pollard J et al. Lifestyle factors affecting fruit and vegetable consumption in the UK Women's Cohort Study. Appetite 2001;37:71-9.

13. Greenwood DC et al. Seven unique food consumption patterns identified among women in the UK Women's Cohort Study. European journal of clinical nutrition 2000;54:314-20.

14. Holland B et al. McCance and Widdowson's The Composition of Foods. London: The Royal Society of Chemistry and MAFF, 1991.

15. Rubin DB. Multiple Imputation for Nonresponse in Surveys. New York: John Wiley & Sons, 1987.

16. Clark TG, Altman DG. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. J.Clin.Epidemiol. 2003;56:28-37.

17. Van Buuren S, Boshuizen HC, Knook DL. Multiple imputation of missing blood pressure covariates in survival analysis. Statistics in Medicine 1999;18:681-94.

18. Rubin DB. Multiple imputation after 18+ years. Journal of the American Statistical Association 1996;91:473-89.