
Article:
Nur, UA, Longford, NT, Cade, JE orcid.org/0000-0003-3421-0121 et al. (1 more author) (2005) Dealing with incomplete data in questionnaires of food and alcohol consumption. Statistics in Transition, 7 (1). pp. 111-134. ISSN 1234-7655

DEALING WITH INCOMPLETE DATA IN QUESTIONNAIRES OF FOOD AND ALCOHOL CONSUMPTION

U.A.M. Nur[1], N.T. Longford[2], J.E. Cade[3] and D.C. Greenwood[4]

ABSTRACT

Missing data feature in most large-scale surveys of human populations, especially when extensive questionnaires are administered. We describe the established single-imputation procedures and explore multiple-imputation procedures for the missing values in the blocks of questions related to alcohol consumption in the UK Women's Cohort Study. We demonstrate how multiple imputation can be developed by adapting the established single-imputation procedures, ridding them of some of their weaknesses.

Key words: Complete-data analysis; food frequency questionnaire; incomplete data; missing information; multiple imputation.

1. Introduction

Alteration of the life style, and of the diet in particular, is generally perceived to have a great potential to prevent or postpone certain common conditions such as obesity, diabetes, as well as various forms of cancer. Appropriate diet may predispose the subject to greater capacity to resist the diseases and other conditions when they develop. These hypotheses can be evaluated by surveys in which the diet of subjects is recorded and related to diseases and conditions contracted in the future, to survival at some point in the future, or to death from one or a specified range of causes.

From a narrow statistical viewpoint, the ideal design for studying the association of diet and health would involve random allocation of subjects (or households) to treatment groups that would differ in how their diet is controlled.

[1] Department of Public Mental Health, Imperial College School of Medicine, London, UK.

[2] SNTL, Leicester, UK.

[3] Nuffield Institute for Health, University of Leeds, Leeds, UK.

[4] Biostatistics Unit, University of Leeds, Leeds, UK.

For instance, one group may be administered some intervention (encouragement), such as education about food.

Such designs, motivated by the principles for clinical trials (Pocock, 1982), are not realistic in the context of long-term large-scale dietary studies. The studies have to involve many subjects because of the variation in their diets, and they have to be conducted over a long period of time because the effect of the diet is realised only after years or decades. Also, of interest is the diet over a long period of time, sometimes even a substantial part of the lifetime. This introduces considerable difficulties in how to elicit information about the diet.

The weighed-intake diary, in which the weight of each food item about to be consumed is recorded, is generally regarded as the gold standard in food consumption surveys. If such a diary is completed diligently it provides very detailed information about the subject's diet in the designated period of time, usually one week. Diet diaries require at least a brief set of instructions, and are most effectively administered by completing at least the first day with the help of a health professional. Since completing the diary for a whole week requires considerable effort and diligence, many subjects are likely to drop out and not complete the diary for the entire period. The one-week snapshot may be sufficient if the subject has maintained a regular diet over a long period of time, but it will capture neither the long-term gradual changes in the diet nor the numerous departures from the subject's usual diet.

In large-scale surveys, low unit cost of data collection is a priority, and so modes of administration that require no face-to-face contact or detailed instructions are preferred. Food frequency questionnaires are suitable in such a setting because the questionnaire items have a simple format and the subjects can be presented a large number of them, inquiring about numerous food items and related details, pertaining to the last month, year or another period. Each question has the same preamble (lead-in passage), such as

In a typical week, how frequently do you eat ...

and a similar response format for each food item. For automated data entry, it is practical to have a set of response options, such as frequencies ranging from Never to Several times a day. On the one hand, the preamble of the questionnaire can emphasise that the items are about long-term diet; on the other hand, choosing the appropriate response option requires a judgement not coloured by the subject's desires or perceptions of an ideal diet or similar influences. Even if the judgement were perfect for all the items, each response category covers a wide range of quantities and patterns of consumption, and so the responses cannot be regarded as a precise summary of the subject's diet.

This paper describes an application of method of multiple imputation (MI, Rubin, 1987 and 1996; Longford, 2005) to the food frequency questionnaire (FFQ) and other questions about alcohol consumption in the UK Women's Cohort Study.

FFQ items involve rather vaguely defined categories (frequencies), but have relatively low rates of nonresponse. In contrast, questions about quantities (QA items) ask for more detail, but have much greater rates of nonresponse. We consider the setting in which a large number of analyses of varied complexity are planned, using the responses to the QA items in a variety of roles, but mainly as explanatory variables. Listwise deletion would discard too much information in the incomplete records and would fail to draw on the information in the FFQ responses about the missing values to the QA responses.

We regard MI as the only approach that is suitable for a large number of planned analyses, of any complexity, and that aspires to the standard of efficient estimation and honest assessment (unbiased estimation) of the precision. For an application in a context similar to this paper, see Longford et al. (2000). A comprehensive solution, imputation for all the missing values for the entire WCS database (over 600 variables), is beyond the scope of this paper. We discuss imputations for variables related to alcohol consumption. These are of particular importance both for statistical (data related) reasons and because of proven or conjectured association with cancer and other diseases. Throughout, we use the terminology of complete, incomplete, completed data (analysis), and missing at random (MAR) and missing not at random (MNAR), as defined in Little and Rubin (1987) and Rubin (1987).

The next section gives details of the Study and of the blocks of questionnaire items concerned with alcohol. The following section motivates the method by discussing the deficiencies of established single-imputation schemes and how they can be resolved by their MI adaptations. Section 4 gives details of the application, and the final section outlines the full potential of MI for dealing with FFQ data.

2. The UK Women's Cohort Study

The UK Women's Cohort Study (WCS) was designed to assess the relationship of diet on the occurrence of cancer and its mortality. The Study's subjects were recruited from the UK participants of a World Cancer Research Fund mail survey. The sample was drawn from among women aged 35–69 years in 1996, and was intended to over-represent women with vegetarian or vegan diet. The sample size of WCS is 35 374. The principal source of data in WCS are the responses to a questionnaire, administered in 1996, comprising a FFQ section with 211 questions grouped into 25 blocks of food categories. For instance, the first block, Bread/savoury biscuits, comprises nine questions about the frequency of eating

1. White bread and rolls
2. Brown bread and rolls
3. Wholemeal bread and rolls

4. Chapatis, nan, paratha
5. Papadums
6. Tortillas
7. Pitta bread
8. Crispbread (e.g., Ryvita)
9. Cream crackers, cheese biscuits.

The questions have the common preamble:

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

and have ten ordered response options, coded 0, ... , 9:
0. Never, 1. Less than once a month, 2. 1–3 times a month, 3. Once a week, 4. 2–4 times a week, 5. 5–6 times a week, 6. Once a day, 7. 2–3 times a day, 8. 4–5 times a day, 9. 6 or more times a day.

The FFQ section is followed by questions inquiring about how certain kinds of food are prepared (fruit, vegetables, meat), consumption of milk, alcohol, sugar, salt and other additives and supplements, dieting, smoking, physical activity, illnesses, education, employment, child-bearing history, and height, weight, and other physical size measurements.

Although most subjects are well motivated, having been recruited by unsolicited mail correspondence that they could have ignored, some respondents can be easily distracted while completing the extensive questionnaire. This is evident from the missing responses in the questionnaires. Although some instances of nonresponse can be attributed to intent, wishing not to disclose certain aspects of the diet, such as excessive alcohol consumption, many missing responses are most likely due to momentary distraction. The pattern of nonresponse is an evidence of this. There are some missing responses for every item, and when they are missing for one item (say, consumption of one type of alcoholic beverage), they are often not missing for a related item, such as another type of alcoholic beverage.

At some point in the future, when a proportion of the subjects have died, the survival will be related to the diet, with an adjustment for age and other relevant covariates. The proportion has to be large enough, so that the relevant hypotheses could be tested with sufficient power. A logistic regression analysis is planned, relating the survival (cancer) status to a summary of the subject's diet, with an appropriate adjustment for the confounding variables. A practical and well-motivated summary of the diet comprises the quantities of macro-nutrients, fat, protein, and carbohydrates, and micro-nutrients, such as vitamins and minerals, and quantities of other substances, such as fibre and net alcohol consumed. These quantities are calculated from tables that associate the frequency of consumption with (weekly) quantity, and the quantity of each nutrient per unit portion, piece or measure of the food item. This involves a lot of unavoidable approximation, as

food items belonging to the same category may contain different quantities of nutrients, subjects have uneven-sized portions, and so on.

Nonresponse is another problem. Routinely, nonresponse to an item is interpreted as no consumption. This is a questionable practice, as an extreme value (zero, the lowest possible) is imputed for each missing item. Single-imputation (SI) methods, in which each missing value is replaced by its estimate, the most likely value, or a guess (e.g., by an expert), are generally regarded as a practical solution. However, they are suitable with a limited class of analyses in which only linear transformations of the imputed values are used. This is not a problem when a minute fraction of the responses is missing. When the fraction is sizeable, such a practice results in a distortion of the inferences. Since the imputed values are not distinguished in the analysis from the genuine (recorded) values, we pretend to possess more information than was collected, and so the precision of the (completed-data) estimators is overstated. However, these estimators are also biased when they are non-linear functions of the imputed values, even when each of these values is an efficient and unbiased estimator of the respective missing value.

A rationale supporting the imputation of zero is that it is by far the most frequent category among the responses to most FFQ items; see, for instance, Hansson and Galanti (2000). The repertoire of food items available (and listed in the FFQ) is so wide that most subjects ever consume only a small fraction of them, and most food items are consumed by a small fraction of the subjects. Exceptions are the most common food items, such as white bread, milk, apples and beer.

Dealing with missing values by reducing the data to the complete records, or to records that are complete for each particular analysis or data summary, is not appropriate because too much information, contained in the non-empty incomplete records, is discarded. Although the sample size of WCS is quite large, it is not excessive for the detailed inferences sought about relatively rare events (cancer incidence and death from specific causes) in a population with varied patterns of diet and wide range of socio-economic circumstances and lifestyles. Thus, efficiency as a criterion for effective use of the data is of essence, and it is highly desirable to assess the precision of key estimators in the numerous planned analyses without any substantial distortion.

2.1. Missing data in alcohol items

Consumption of alcohol is recorded by FFQ, in a block of five questions inquiring about the frequency of consumption of

- wine
- beer, lager
- cider
- port, sherry, liqueurs
- spirits, e.g., whiskey, gin, vodka, brandy

and, in two blocks, about quantities of

- beer or cider (pints each week)
- wine (glasses each week)
- sherry/fortified wines (glasses each week)
- spirits (single measures each week)

consumed during a typical week (In a typical week, how much do you drink?), recently (in the last 12 months) and five years ago. (A pint is about 0.57 litre). We refer to them as quantity (QA) items. Note that the frequency and quantity questions are not in an exact correspondence — in FFQ, beer and cider are separate items, whereas in QA blocks they are included in a single question.

The two types of questions have complementary strengths; a cursory reflection is sufficient for responding to FFQ, while the QA questions are more probing. Thus, FFQ items might be expected to have a much higher response rate than QA items. This is confirmed in Table 1, where the response rates to the FFQ and QA items on alcohol are given.

The structure of the questions enables us to find evidence that imputing zero for missing values is not appropriate. For instance, if the response to a FFQ item indicates some (positive) consumption, it contradicts the imputed value of zero in the corresponding QA item. Also, the nonresponse to a FFQ item cannot plausibly be interpreted as Never when Never is entered as a response to another item of the same block. Table 1 gives univariate summaries of the response pattern. More detail is given in Table 2, where the frequencies of each pattern are listed. In general, a response pattern is defined as a sequence of indicators of response. Thus for FFQ, 11111 stands for response to all five items, and 00000 to no response to any of them.

Only 527 subjects (1.5%) have incomplete records on the five FFQ items. The most frequent incomplete pattern is that for the empty record (pattern 00000; 88 subjects). The patterns with one entry (00001, 00010, ..., 10000) occur in 76 cases, whereas the patterns with one item missing (01111, ..., 11110) in 207 cases. Of the latter 207 subjects with one missing item, 79 have not entered Never for any of the other four items. These subjects may have expected their nonresponse to be interpreted as zero. More than half of them (47 subjects) omitted the response to the question about cider, which is by far the least ‘popular’ of the alcoholic beverages in the Study (71% of the sample responded with Never). However, the presence of a Never response among the majority of the 207 subjects suggests a momentary lapse as a more likely reason for nonresponse. But a deliberate omission, not to disclose excessive consumption, cannot be ruled out either.

Table 1. Response rates for the FFQ and QA questions related to alcohol.

	Response rate (%)		
	FFQ	QA (now)	QA (5 years ago)
Wine	99.5	81.8	81.7
Beer	99.2	47.7	48.0
Cider	99.0		
Sherry	99.2	48.4	49.4
Spirits	99.4	78.1	58.3

The combined response and sign pattern for a set of four responses is defined as the sequence of four characters, each of them either M, 0 or P, standing for 'missing', zero or positive quantity, respectively. 25 444 subjects (72%) have the same pattern of responses for consumption at the two time periods, and 18 216 of them declared identical pairs of quantities (or omitted both responses to a pair of questions). 6199 subjects (17.5%) have only one discordant pair of responses each. Subjects tend to omit responses to the pairs of questions. This is confirmed by the four tables for the types of beverage. For instance, the relevant summary for beer and cider is given in Table 3.

This suggests that most failures to respond to QA items are intentional. We would expect such subjects not to respond to the corresponding FFQ item either, and this would be reflected in the table of the signs for the paired FFQ and QA questions. Since the FFQ questions about beer and cider correspond to a single QA question, their responses are combined. If only one response is available, it is adopted as the combined 'score'. If both responses are available, the higher one is adopted. Further, if both responses are positive (indicating some consumption) and equal to one another, the score is raised by one. The signs of this frequency score and quantity of current consumption of beer and cider are summarised in Table 4. The majority of those who responded Never to both FFQ items did not respond to the QA item, but almost half of those who indicated a positive frequency also failed to respond to the QA item. Thus, there is a considerable uncertainty about most missing values, and imputation according to a deterministic formula would inflate the information contained in the 'cleaned' data considerably.

Table 2. Response patterns for the FFQ and QA items related to alcohol. Patterns with over 2000 subjects are highlighted.

Numbers of subjects with response patterns								
FFQ								
Pattern	00000	00001	00010	00011	00100	00101	00110	00111
# subjects	88	22	7	6	2	0	1	1
	01000	01001	01010	01011	01100	01101	01110	01111
	2	6	1	1	2	1	2	29
	10000	10001	10010	10011	10100	10101	10110	10111
	43	37	24	23	0	5	1	29
	11000	11001	11010	11011	11100	11101	11110	**11111**
	4	26	4	76	11	45	28	**34 847**
QA — now								
	0000	0001	0010	0011	**0100**	**0101**	0110	0111
	3231	1152	674	217	**6746**	**3273**	1898	1302
	1000	1001	1010	1011	1100	1101	1110	**1111**
	697	278	97	85	1689	1181	409	**12 445**
5 years ago								
	0000	0001	0010	0011	**0100**	**0101**	0110	0111
	3129	1304	700	243	**6176**	**3462**	1917	1466
	1000	1001	1010	1011	1100	1101	1110	**1111**
	647	306	83	77	1677	1209	411	**12 567**

Table 3. Missing, zero and positive values of the responses to the QA item about beer and cider consumption at present and five years ago.

		QA 5 years ago		
		Missing	Zero	Positive
	Missing	17 315	234	944
QA now	Zero	329	9008	395
	Positive	753	494	5902

Table 4. Missing, zero and positive responses to the FFQ and QA questions about the consumption of beer and cider.

		QA now		
		Missing	Zero	Positive
	Missing	182	52	16
FFQ	Zero	9457	6610	49
	Positive	8854	3070	7084

3. Multiple imputation

Although the method of multiple imputation (MI) is well established in many agencies conducting and analysing large-scale surveys, especially in the USA (Rubin and Schenker, 1991), it has been applied much less in health-care and epidemiological surveys in the UK and Europe. The rationale for MI is applicable to all large-scale surveys with non-trivial levels of nonresponse (Rubin, 1996; Longford, 2001). The original motivation for MI is based on the premise that it is desirable to deal with the incompleteness of the data at the database construction stage, so that the secondary analysts, who often rely on standard statistical methods, can analyse the data without requiring any expertise for handling missing values or any specialised software. The method is not restricted by the type, complexity or the number of the planned (complete-data) analyses or by the software used for their application.

With MI, the planned analysis, designed for the data that has no missing values, is applied without any alterations to a small number of completed data sets. The completions (by sets of plausible values) are generated by the database constructor prior to the release of the database, and are used in every analysis. Although MI is not a recent invention (Rubin, 1987), the case for its application has only been strengthened with the advent of cheap and abundant computer processing speed and storage space, and as the economy of the programming effort (analyst's time) overtakes the amount of computing as the principal concern of statistical analysts.

Generating the sets of plausible values, used for completing the recorded data, carries a lot of responsibility because it has an impact on all the subsequent analyses. It is essential for the process applied to reflect the uncertainty about the missing values, so that the imputation would be proper (in the sense of Rubin, 1987), and at the same time to incorporate all the information about the processes that give rise to nonresponse. The process of nonresponse usually defies a simple description because the subjects fail to respond as a result of a wide range of intentions, motives, and momentary distractions or misinterpretation of the instructions. The questions they do not respond to are isolated items, sequences of items of varying length, whole blocks, all items from a certain point till the end of the questionnaire, or combinations of these patterns. The correct specification of the process of nonresponse is an important though usually unverifiable condition for the validity of the results associated with MI. This is often quoted as an objection to using MI. However, any SI procedure can be described by an underlying model; when SI and MI procedures are compared against the same standard, MI is superior because with it the completed-data analyses are valid for a

wider class of estimators (not only those linear in the data) and for a wider class of nonresponse mechanisms. Even if imperfectly, the uncertainty about the missing data is taken into account. Most SI procedures can be adapted by acknowledging that the imputed value is not determined with precision. For example, instead of imputing zero for every missing value, the imputed value is drawn from a distribution with a high probability of zero. We argue along these lines in the next section where we explore several procedures for dealing with missing values in the QA blocks related to alcohol consumption.

Frequency of alcohol consumption cannot be straightforwardly translated into the quantity consumed. On the one hand, occasional excessive consumption amounts to low frequency; on the other hand, some consumption in moderation is recorded as high frequency, even if much smaller quantity is consumed in total. With alcohol, the pattern of consumption may also be an important factor; regular moderate consumption and occasional excessive consumption may result in similar quantities consumed, yet they may have radically different long-term effects (Chan, Pristach and Welte, 1994).

The QA blocks have a potential to collect more detailed information for the assessment of the net alcohol consumed. Their drawback is the much lower response rate, as shown in Table 1. When QA responses are missing we may fall back on the corresponding FFQ response for a substitute. For instance, each of the ten response categories may be associated with a typical quantity of consumption, and this quantity used as the substitute. The MI version of this approach defines a 'substitute' distribution, derived from the respondents. The substitutes (imputed values) are drawn from this distribution. Further, the fact that this distribution is not known but is estimated, should also be reflected in the process of generating plausible (substitute) values, by drawing them from plausible distributions.

3.1. Associations among the responses

Substituting for the missing item a value recorded for a 'similar' variable is a straightforward SI procedure that exploits the similarity of the responses. An example of it is the method of 'bringing the last value forward' in longitudinal surveys. Its rationale is that the responses of most subjects change little (or not at all) from one time point to the next, so the 'previous' value is a good substitute for the current value. An application of this method to the QA items would impute for the missing values of the current consumption its counterpart from five years ago. The method can be applied also in reverse, imputing the current consumption for the consumption five years ago. Although only a small fraction of the values would be filled in in this way (see Table 3), it could still be used in conjunction with an imputation method for the pairs of items (say, wine consumption now and five years ago) when both are missing.

When a subject responds Never to a FFQ item, the logical value for the corresponding QA item is zero, so it is reasonable to impute it when the value is missing; see Table 4. In this way, we would deal with more than half of the missing values in the QA block of items. However, Never and positive consumption indicated by QA also occur, although rarely. Similarly, no current consumption in the QA block and a positive consumption in the FFQ block also occur. We prefer the information in the QA block, when available, and make no attempt to resolve these inconsistencies.

The responses to a FFQ item can be regarded as an effective stratification variable for the corresponding pair of QA items. This suggests imputing the mean QA response of the subjects who responded with the same category to the FFQ item. The subjects with missing responses to a QA item (recipients) belong to ten groups according to their response to the corresponding FFQ item. The subjects with recorded responses (donors) are classified similarly. For each recipient, the mean QA response of the corresponding group of donors is imputed. Such an imputation is appropriate only if the donors' within-group dispersions of the QA responses are small. In general, it is preferable to impute a random draw from the donors' responses — this is the hot deck method. Since the pairs of QA responses are correlated, hot deck should be applied to the two components simultaneously, by drawing a donor and imputing her pair of QA responses. When one QA response is missing and one is recorded, the draw should be made from the conditional distribution of the missing value given the recorded value. Conditioning is implemented by reducing the pool of donors further. Since the QA responses are on an ordinal scale, the conditioning can be implemented only by matching on a coarsened scale. The functional form of the conditional distribution is difficult to identify; it is supported on integers and some 'halves', such as 0.5, 1.5 and 2.5 (pints per week). Among the larger integers rounded numbers dominate.

This procedure can be improved by further conditioning, for instance, on the responses to the other QA items. This is limited only by the need to have a reasonably large pool of donors for every recipient. In the hot deck method, we draw, in effect, from the empirical (estimated) distribution of the QA responses in the relevant population. A more principled approach uses a plausible distribution instead of the estimated one. This corresponds to proper imputation, as defined by Rubin (1987).

The responses to the QA items are highly correlated within types of beverage, but much less so across types. The correlation matrix of the eight QA responses is estimated as

$$\frac{1}{1000}\begin{pmatrix} 1000 & 37 & 25 & 61 & \underline{721} & 37 & 30 & 54 \\ 37 & 1000 & 46 & 117 & 80 & \underline{770} & 78 & 133 \\ 25 & 46 & 1000 & 36 & 9 & 46 & \underline{680} & 38 \\ 61 & 117 & 36 & 1000 & 73 & 101 & 41 & \underline{690} \\ \underline{721} & 80 & 9 & 73 & 1000 & 71 & 23 & 99 \\ 37 & \underline{770} & 46 & 101 & 71 & 1000 & 101 & 178 \\ 30 & 78 & \underline{680} & 41 & 23 & 101 & 1000 & 76 \\ 54 & 133 & 38 & \underline{690} & 99 & 178 & 76 & 1000 \end{pmatrix} \quad \begin{bmatrix} \text{Now} & \text{BeerC} \\ & \text{Wine} \\ & \text{Sherry} \\ & \text{Spirits} \\ & \\ \text{-5Y} & \text{BeerC} \\ & \text{Wine} \\ & \text{Sherry} \\ & \text{Spirits} \end{bmatrix}$$

(the within-beverage correlations are underlined). The matrix was estimated by pairwise deletion; for instance, the estimate of its (1,6) element is based on the 15 612 subjects who responded to both questions BeerC Now and Wine –5Y. To appreciate the impact of missing values, we give the complete-case estimates of the correlation matrices for the two sets of four QA items. For the items about the current consumption (Now), the correlation matrix is estimated by

$$\frac{1}{1000}\begin{pmatrix} 1000 & 188 & 97 & 177 \\ 188 & 1000 & 213 & 255 \\ 97 & 213 & 1000 & 153 \\ 177 & 255 & 143 & 1000 \end{pmatrix}$$

and for the consumption five years ago (–5Y) by

$$\frac{1}{1000}\begin{pmatrix} 1000 & 262 & 164 & 239 \\ 266 & 1000 & 322 & 374 \\ 164 & 322 & 1000 & 267 \\ 239 & 374 & 267 & 1000 \end{pmatrix}.$$

These estimates, based on 12 445 and 12 567 subjects, respectively, differ from the correlations estimated by pairwise deletion by much more than their standard errors. This suggests that the bias in estimation due to incompleteness is a more serious concern than sampling variation.

We consider imputation for each type of beverage separately, and deal first with subjects who have no response for either of the corresponding QA items. For BeerC, we define a composite score for the FFQ items as follows. We take the higher of the responses to the FFQ items Beer and Cider, or the available response when the other one is missing. When the two responses are identical and positive (some consumption), we raise the score by one. The scores above 5 points (consumption at least once a day) are truncated. This yields the table of scores displayed in Table 5.

Recipients are the subjects with both QA responses missing, and donors are the subjects with both QA responses recorded. Some subjects are neither recipients nor donors, and so the numbers of recipients and donors do not add up to the numbers of all subjects within the columns. Hot deck draws for each recipient a donor from the pool of donors with the same FFQ score. As an alternative, a bivariate distribution could be fitted to the donors' values, and draws made from this distribution. However, a suitable (parametric) family of distributions is difficult to identify. The substantial mass at zero can be modelled by a mixture, but the remainder of the values are neither normally nor log-normally distributed. Also, all the values are either integers or halves. The conditional distributions of QA responses given FFQ scores have variances very close to their means.

Table 5. Tabulation of the recoded scores for recipients and donors used in the hot deck procedure to impute for missing consumption of beer and cider.

	FFQ score							
	Missing	0	1	2	3	4	5	**All**
Recipients	174	9152	3639	3239	819	229	63	17 315
Donors	65	6501	1698	2122	2081	2054	1278	15 799
All	250	16 116	5719	5964	3314	2540	1471	35 374

However, the Poisson distribution would not provide a good fit because the frequencies are distinctly not monotone; even and rounded numbers are more frequent. For instance, the distribution of the responses to the QA item about the current consumption of beer and cider for those with FFQ score 2 (1–3 times a month) is

Pints a week	0	$\frac{1}{2}$	1	$1\frac{1}{2}$	2	3	4+
Subjects	1214	560	489	4	53	4	14

Its mean is 0.41 and variance 0.39. For Beer and Cider, the bivariate hot deck deals with 17 141 subjects who have both QA responses missing but have a recorded response to FFQ item on beer or on cider. The usefulness of the stratification on the FFQ responses is illustrated on the summary of the means and percentages of zeros among the donated (imputed) values within the FFQ scores, displayed in Table 6. Higher score is associated with higher average consumption. Note that for each FFQ score the current consumption is lower than five years ago. (The FFQ score refers to the current consumption.) Also, the numbers of subjects whose imputed consumption is zero are greater for the current than for the past consumption.

With this procedure, we have failed to impute only for 174 subjects with no responses to both FFQ and both QA items about beer and cider and 1178 and 1082 subjects who have only one missing QA response, for the current and past

consumption, respectively. For the 174 subjects, prediction of their consumption is difficult because of the paucity of information about them. In any case, there are so few of them (0.5%) that imputation for them has a low priority.

Table 6. The means and percentages of zeros among the donated values of current and past consumption of beer and cider, by FFQ categories.

FFQ score							
	0	**1**	**2**	**3**	**4**	**5**	**All**
BeerC –Now							
Mean (pints a week)	0.01	0.09	0.39	0.94	1.69	5.10	0.18
% zeros	99.6	88.3	53.7	9.9	1.3	1.6	82.5
BeerC – 5 years ago							
Mean (pints a week)	0.04	0.26	0.88	1.47	2.41	6.13	0.37
% zeros	98.5	83.9	49.8	16.6	8.7	3.2	80.7

On the other hand, imputation for the subjects with one QA response is both easier and more important. The simplest procedure would impute the available QA response for the missing one. However, we have seen earlier that the consumption of beer has declined on average. This finding from Table 6 can be reinforced by comparing the pairs of QA responses when they are recorded. Among the 15 799 subjects with both QA responses recorded, the mean consumption has declined from 1.09 pints a week five years ago to 0.83 pints at present. 9008 subjects have declared zero for both items, 395 subjects have switched to consuming at present and 494 have stopped consuming beer since five years ago. Of the 6791 subjects who have declared a positive consumption on at least one occasion, 2232 (32.9%) have declared greater consumption and only 965 (14.2%) lower consumption in the past.

The imputation could be adjusted for the trend, by adjusting for the mean difference. As it is desirable to have rounded values, a random process could be employed to assign the direction of rounding to halves of pints. This would still not be sufficiently realistic, because halves of pints are reported only by the most moderate consumers of beer and cider. However, the trend in the consumption need not be uniform, as assumed by this procedure. The hot deck procedure can be adapted for these subjects by classifying the recipients' and donors' consumption into suitable categories (intervals). One such choice are the intervals (0, 1], [1.5,

2.5], [3, 5], [6, 10], 10+ (pints a week), and a category for ‘No consumption’. The two groups for the highest consumption contain the smallest numbers of subjects, but they are still sufficient for an effective application of the hot deck procedure. Table 7 gives the numbers of recipients and donors. The donors are common to both sets of recipients, although they donate values of different variables and form different pools.

Table 7. The numbers of recipients and donors in the hot deck procedure applied to the QA items about beer and cider.

	Category of consumption (pints a week)						
	0	$\frac{1}{2}-1$	$1\frac{1}{2}-2\frac{1}{2}$	3–5	6–10	10+	**All**
Imputing for the current consumption							
Recipients	234	368	264	210	85	17	1178
Donors	9502	3020	1290	1252	574	161	15 799
Imputing for the consumption 5 years ago							
Recipients	329	503	133	77	33	7	1082
Donors	9403	3857	1188	912	342	97	15 799

With this approach, the imputed values are drawn from existing values and they maintain the high correlation of the pairs of QA responses. For instance, most of the imputed values for the current consumption, given no consumption five years ago, are zeros (226 out of 234 in a particular replication), whereas the majority of the imputed values for subjects with high consumption five years ago are also high.

3.2. Why and how to impute multiply

The undoubted virtue of (nearly) completing the data by imputation is that any analysis is based on (nearly) all subjects and the method or algorithm planned for the analysis can be applied without any alterations. For some analyses, the sample size is, effectively, doubled. However, the reduced sampling variance is only one goal; the other one is a control over the bias, both in estimation and in assessment of the precision of the estimators.

If we treat the imputed values on par with the recorded values, we underestimate the sampling variance of the standard (complete-data) estimators

because we pretend that the imputed data has in fact been recorded. We assume that a replication of the response to a FFQ or QA item on the same subject would yield the same value, but we do not expect that a replication of the imputation procedure would duplicate the originally imputed values. Thus, the imputation process is associated with variation, and ignoring it causes overstating of the precision of the estimators. The variation can be assessed by replicating the process, followed by estimation. This naturally leads to MI. With M replicates of the imputation and estimation processes, we obtain estimates $\hat{\theta}_1,...,\hat{\theta}_M$ of the target θ, and estimated sampling variances $\hat{s}_1^2,..., \hat{s}_M^2$, evaluated by standard formulae that disregard the issue of missing values. These estimators are called completed-data because they would be appropriate if the data were complete and the plausible values were genuine observations. The MI estimator of θ is defined as the average of the completed-data estimates

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

and its sampling variance is estimated by

$$\tilde{s}^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{s}_m^2 + \frac{M+1}{M(M-1)}\sum_{m=1}^{M}(\hat{\theta}_m - \tilde{\theta})^2. \quad (1)$$

These estimators take account of the imputation process. The average of the estimated completed-data sampling variances, $\hat{W} = (\hat{s}_1^2 + ... + \hat{s}_M^2)/M$, is supplemented by the between-imputation sampling variance $\hat{B} = \sum_m(\hat{\theta}_m - \tilde{\theta})^2/(M-1)$, inflated (penalised) by the factor $1+1/M$ for using only M replicates.

These results apply under the following assumptions. The complete-data (original) method yields an efficient unbiased estimator of θ and its sampling variance s^2 is estimated without bias and with a sampling variance of smaller order of magnitude than s^4. Maximum likelihood estimators for moderate sample sizes usually satisfy these conditions. Next, the posited model for nonresponse is correct, and the imputation process based on it is proper; that is, it accurately reflects the uncertainty about the imputation model. This assumption of model correctness cannot be verified. We can merely claim that the model on which our imputation process is based is more realistic than the model implied by the established method (nonresponse ≡ zero). Hot deck is not a proper imputation method because not all the sources of uncertainty are reflected in the process. The method can be described as random draws (with replacement) from the empirical distribution defined by the

donor pool. In MI, imputations should be drawn from a plausible distribution of the donors' values. With the values of consumption, this is not straightforward to arrange because the (bivariate) distributions of present and past consumption in the various donor populations do not have a simple description. However, most of the sizes of the donor pools are very large, exceeding 1000, and so this contribution to the between-imputation variation is much smaller than the variation of the donors' values. For an example of a proper imputation in a similar context, see Longford et al. (2000).

Our concern is not in optimal estimation and full exploitation of the potential of MI, but a substantial improvement on how the WCS database would be analysed without introducing great demands on the complexity of the procedures. We regard it as essential that these improvements be well motivated, transparent, and acceptable to the wide community of secondary analysts. In Section 4.1, we apply the approximate Bayesian bootstrap, an (approximately) proper-MI adaptation of the hot deck procedure. It yields results very similar to the multiply applied hot deck.

4. Application

In this section, we compare the SI and MI estimates. For the current and past consumption of beer and cider, the estimates for various data reduction and imputation methods are listed in Table 8. Method 1 is based on the subjects who responded to both QA items about beer and cider, method 2 on those who responded to all eight QA items about alcohol consumption, and method 3 is based on case-wise deletion, making the maximum use of the responses. Although method 2 has the smallest sample size, its estimated standard error is smaller than for the other two data reduction methods. This is a consequence of the relative homogeneity among the diligent respondents — their mean is smaller and a higher proportion of them declared that they consume no beer or cider.

With imputation methods, the estimated standard errors are much smaller, but the estimates for method 4 (zero imputed for each missing value) and the hot deck method (5 or 6) differ substantially. In the hot deck method, many positive values for beer consumption are imputed. Methods 5 and 6 differ only in the 174 subjects who failed to respond to the FFQ and the two QA items about the consumption of beer and cider. The imputation for them has a negligible impact on the estimates.

The estimates with methods 2 and 4 are similar. With method 4 not only the imputed zeros but also numerous incomplete records are used. It appears that the consumption of beer and cider declared in the incomplete records tends to be higher than in the complete records. This can be confirmed by inspecting the data.

Method 7 entails repeated application of the hot deck procedure. Method 8 is discussed in Section 4.1. The standard errors are smaller for the current

consumption than for consumption in the past, reflecting the lower means. The standard errors estimated by methods 4–8 are very similar for the current consumption (0.0078–0.0084) and differ more for the past consumption.

The differences in the estimates are of a greater order of magnitude than the standard errors. The estimates could not possibly be all unbiased or apply to the same target (estimand). Indeed, the population represented by diligent responders may differ substantially from the population represented by all the subjects. Moreover, the latter population is not well defined; it is merely implied by the recruitment procedure. In all planned analyses, the bias is the principal concern; it is the dominant contributor to the mean squared error.

Table 8. Estimates of average beer consumption based on data reduction and imputation.

	Consumption (per week)					
	Current			**5 years ago**		
Method	Mean (pints)	Standard error	Sample size	Mean	Standard error	Sample size
Data reduction						
1. Both QA available	0.829	0.016	15 799	1.086	0.020	15799
2. All QA available	0.409	0.012	11860	0.540	0.015	11860
3. Each QA available	0.853	0.015	16881	1.162	0.020	16997
Single imputation						
4. Zero imputed	0.407	0.008	35374	0.558	0.010	35374
5. Hot deck	0.522	0.008	35200	0.785	0.012	35200
6. Hot deck and zeros	0.520	0.008	35374	0.780	0.011	35374
Multiple imputation						
7. Hot deck	0.525	0.008	35200	0.785	0.012	35200
8. Approximate Bayes bootstrap	0.524	0.008	35200	0.785	0.013	35200

The MI estimate is obtained by averaging the ten completed-data estimates; we obtained 0.525 and 0.785 pints for the current and past weekly consumption of beer, respectively, and 0.0083 and 0.0120 for the corresponding standard errors. Unlike the other estimates, these quantities are stochastic; apart from the data they depend on the random numbers used in the generation of the plausible values.

The contributions to the estimated sampling variance, $\hat{W}$ and $\hat{B}$, have relative sizes of about 10:1 for the current consumption and 7:1 for the past consumption, so the respective fractions of missing information, $\hat{B}(\hat{W} + \hat{B})$, are about 1/11 and 1/8. They are much smaller than the fractions of missing values (more than 50%) because there is little uncertainty about the missing values for the many subjects who most likely do not consume any beer or cider. This is partly due to the auxiliary information used, FFQ responses, although our assessment is

somewhat inflated because we applied an improper MI procedure, ignoring the between-imputation variation of the plausible donor distributions. However, with 10% of missing information, M=10 imputations are more than adequate; B/M contributes to the sampling variance $\tilde{s}^2$ by only about $1/111 \cong 1\%$. With M=5 the contribution would be 2%, still negligible. More information is missing about the past consumption than about current consumption because the auxiliary variable, FFQ score, refers to the latter and provides more information about it.

The original description of the MI procedure envisaged that the sets of plausible (donated) values would be generated once, prior to the public release of the database, and each analysis would then be based on them. An alternative afforded by fast computing is that each analysis is preceded by generating sets of plausible values and constructing the completed data sets. In this way, less storage space would be used. However, nowadays the storage space is readily available at a low cost.

4.1. Approximate Bayesian bootstrap

The distribution of the donor values is discrete, but with a different support, comprising a varying number of points, for each donor pool. Suitable underlying distributions are difficult to specify, and so it is more practical to employ a non-parametric approach with which to reflect the uncertainty about them. In the approximate Bayesian bootstrap (ABB, Heitjan and Little, 1991), a random sample is drawn from the donor pool with replacement, and recipients are assigned values by a random draw from this plausible donor pool. The plausible donor pool has the same size as the original donor pool. Several donor pools contain values that are either unique or occur only a few times. Values that might reasonably have occurred in another replication but are not present in the realised data set, will never be generated by ABB. This is an unsatisfactory feature of ABB, although it satisfies the standard of generating plausible values with the appropriate first two moments. See Rubin (1981) and Lipsitz, Zhao and Molenberghs (1998) for background. Rubin (1987) and Schafer (1997) propose several alternatives based on sampling from the donor's residuals in linear models.

For a large donor pool, it is not practical to implement the sampling with replacement literally, because it involves a large number of random draws. For values that occur frequently, their plausible probabilities can be generated from the normal approximation. In our application, most donor pools are large and the frequently occurring values account for a large fraction of the values. As a consequence, ABB yields results almost identical to the multiply applied hot deck. With M=5 replications, we obtained nearly the same estimates, 0.524 and 0.785 for the current and past consumption, respectively, with standard errors 0.0085 and 0.0124. In this instance, the aspect of proper-ness ensured by ABB is unimportant. It would become more prominent if the donor pools were defined by more detailed

conditioning because the probabilities implied by the compositions of the pools would be subject to more uncertainty.

Table 9 gives the MI estimates and their standard errors for the four types of alcoholic beverage. The results refer to the multiple application of the hot deck with M=10 replications. The fractions of missing information are lowest for wine (about 5%) and highest for sherry (about 1/3), reflecting the number of missing items. For all four types of beverage, the amount of missing information is higher for the past than for the present consumption, but the difference is substantial only for spirits; the response rate for them is much lower for the consumption in the past, see Table 1.

4.2. Sensitivity analysis

Since we are not privy to the processes that bring about nonresponse and our conjectures about the causes are not very well informed, the model for nonresponse is unlikely to be correctly specified. We have assumed that the values are missing at random (MAR), yet we have implied that this condition may not be satisfied. Sensitivity analysis explores the changes in the estimates (and standard errors) that result from altering the imputation process. Its purpose is to respond to concerns about the lack of validity of the posited model for nonresponse.

Suppose the principal concern is that the non-respondents tend to consume more than the amounts implied by the model used for imputation. This can be addressed by altering the imputations as follows. For the recipients in FFQ category k we use the donors from category k+1. As a more flexible alternative, we may assign the donor's category by a random process, such as deciding at random, with a set probability, whether to use the donors from the same category k, or from k+1. As a further generalisation, a binomial random variable B may be used to select the donor's category k+B, truncated at the highest category 5 (five times a week or more frequently). And finally, a conditional distribution of the donor pool category can be specified separately for each recipient category k. With it, we can address more complex concerns. For instance, the bias due to MNAR (missing not at random) for the higher categories k may act in one direction (upwards) and for the lower categories it the other (downwards).

Table 9. MI estimates of the mean consumption of the four types of alcoholic beverages in WCS.

		Consumption (per week)			
		Current		**5 years ago**	
	Units	Estimate	Standard error	Estimate	Standard Error
Beer and cider	pints	0.525	0.0084	0.785	0.0120
Wine	glasses	3.669	0.0267	3.629	0.0306

Sherry and fortified wines	glasses	0.793	0.0125	0.972	0.0148
Spirits and liqueurs	single measures	1.381	0.0168	1.746	0.0207

By way of an illustration, we explore the impact of the altered hot deck procedure in which a random half of the recipients in FFQ category k, k=0, 1, ..., 4, are assigned values from the donors in FFQ category k+1. The MI estimates of the mean consumption of beer and cider are 0.599 (present) and 0.889 (past), with respective standard errors 0.0085 and 0.0127. Of course, the means are higher than for the original MAR-based imputations. The increases, by 0.07 (present) and 0.10 (past) indicate how sensitive the estimates are to the specified departure of the posited nonresponse mechanism from MAR. The procedure can be repeated for other probabilities, but this is not necessary as a good guess can be made by extrapolation. However, the conclusion arrived at for one summary (mean) does not carry over to another summary, such as the corrected sum of squares, which is needed in a regression.

Another example of sensitivity analysis attends to the hypothesis that nonresponse should (in some cases) be interpreted as zero. A given proportion p of imputed values are replaced by zeros. With the proportion p in the range (0, 1), we obtain the range of estimates between methods 6 and 4 in Table 8. Whereas we can anticipate the result for the mean, the sensitivity with respect to other summaries is more difficult to predict. For instance, with a greater proportion of zeros, the regression on the net alcohol consumed, requiring the corrected sum of squares, is bound to be less stable because the relatively few high consumers will have a greater leverage.

5. Conclusion. The full potential of MI

We have applied multiple imputation to deal with the missing responses to the questions about alcohol consumption in a large-scale dietary survey. Other aspects of incompleteness of the recorded information can also be treated by MI. In the analyses planned in the future, the incidence or death from a particular disease will be related to a set of summaries of the diet by a logistic regression. A particular summary, such as the quantity of protein in the diet, is calculated from the FFQ items by converting the responses (ranges of frequencies) to numbers of portions, multiplying these by the size of a typical portion, and transforming the total consumed to its content of the nutrient concerned. This process relies on several simplifying assumptions: first, that the subjects make no misjudgements (deliberate or unintended) of the pattern of their consumption; second, that every subject's frequency is equal to the frequency in the declared FFQ response option; third, that all portions are of equal size and no food is discarded; fourth, that there is no variation in the composition of the foods covered by a single FFQ question; and that the nutrient is absorbed completely (or at a rate common to all subjects).

Further, of interest is long-term diet, and so information is required about long-term patterns of consumption. Insights can be gained by repeating the FFQ questionnaire after a few years, as was done in WCS in 2001, and by using alternative modes of data collection. The concerns raised apply to a varying extent to all modes of dietary data collection.

Dealing with all these concerns is a tall order, and would be difficult even with a much smaller survey and shorter questionnaire. Although the tables of portion sizes and nutrient content are standard, they are difficult to verify experimentally, and little information is available about the ranges (distributions) of the portion sizes in the population. See Hunter et al. (1988) and Conn, Rutishauser and Wheeler (1994) for studies exploring this issue. The tables are not complete, especially for some of the rarer micronutrients.

Insights into the extent of misjudgement can be gained by comparing the estimated quantity of nutrients with the subject's requirement, estimated from her height, weight and the information about lifestyle (occupation, amount and intensity of exercise, and the like). See Price et al. (1997) for a study exploring this issue. However, such a comparison compounds misjudgement of the frequency with the sizes of portions, content of nutrients in the food items, and related details. Also, for those who under-report, we have no methods for 'supplementing' FFQ with (randomly selected) positive responses.

An analysis of coarse data by MI was presented by Heitjan and Rubin (1990). In their application, children's ages were rounded to half or quarter of a year. For ages of small children (1–5 years), this makes a substantial impact on any regression analysis in which the age is a covariate. The setting with frequencies, or quantities of consumption reported on a coarse scale, is very similar, as are the distributions of the imputed quantities. Heitjan and Rubin (1990) demonstrated that the analysis is very sensitive to the assumptions about the underlying distribution of the ages, and it differs substantially from the trivial analysis in which the declared ages are taken at face value. The analysis of FFQ responses is exposed to the same threats to validity, compounded by the other sources of incomplete information. This cannot be remedied by more detailed questionnaires because they might result in even more extensive nonresponse.

We regard understanding of the processes associated with incomplete information as key — how variable are the portions within and between subjects, how consistent is the diet over a lifetime (or adulthood) and how strong and uniform are the secular trends in diet, how reproducible and accurate are the subjects' responses, and the like. The collection of such information and its integration in the analysis remain major challenges in the analysis of food frequency data.

Acknowledgement

Research for this article was supported by an ESRC Studentship (UAMN, at University of Leeds) and partly by the Campion Fellowship (NTL).

REFERENCES

CHAN, A. W., PRISTACH, E.A., and WELTE, J.W. (1994). Detection by the CAGE of alcoholism or heavy drinking in primary care outpatients and the general population. Journal of Substance Abuse 6, 123—135.

CONN, J.A., RUTISHAUSER, I.H.E., and WHEELER, C.E. (1994). Portion size data for foods consumed by a randomly selected sample of Geelong adults. Australian Journal of Nutrition and Diet 51, 58—65.

HANSSON, L.M., and GALANTI, M.R. (2000). Diet-associated risks of disease and self-reported food consumption: How shall we treat partial nonresponse in a food frequency questionnaire? Nutrition and Cancer 36, 1—6.

HEITJAN, D.F., and LITTLE, R.J.A. (1991). Multiple imputation for the Fatal Accident Reporting System. Applied Statistics 40, 13—29.

HEITJAN, D.F., and RUBIN, D.B. (1990). Inference from coarse data via multiple imputation with application to age heaping. Journal of the American Statistical Association 85, 304—314.

HUNTER, D.J., SAMPSON, L., STAMPFER, M.J., COLDITZ, G.A., ROSNER, B., and WILLETT, W.C. (1988). Variability in portion sizes of commonly consumed foods among a population of women in the United States. American Journal of Epidemiology 127, 1240—1249.

LIPSITZ, S.R., ZHAO, L.P., and MOLENBERGHS, G. (1998). A semiparametric method of multiple imputation. Journal of the Royal Statistical Society Series B 60, 127—144.

LITTLE, R.J.A., and RUBIN, D.B. (1987). Statistical Analysis with Missing Data. John Wiley and Sons, New York.

LONGFORD, N.T., ELY, M., HARDY, R., and WADSWORTH, M.E.J. (2000). Handling missing values in diaries of alcohol consumption. Journal of the Royal Statistical Society Series A 163, 381—402.

LONGFORD, N.T. (2001). Attitudes to immigration in an International Social Science Survey. Statistics in Transition 5, 267—280.

LONGFORD, N.T. (2005). Missing Data and Small-Area Estimation. Modern Analytical Equipment for the Survey Statistician. Springer-Verlag, New York; to appear.

POCOCK, S.J. (1982). Clinical Trials: A Practical Approach. John Wiley and Sons, Chichester, UK.

PRICE, G.M., PAUL, A.A., COLE, T.J., and WADSWORTH, M.E.J. (1997). Characteristics of the low-energy reporters in a longitudinal national dietary survey. British Journal of Nutrition 77, 833—851.

RUBIN, D.B. (1981). The Bayesian bootstrap. The Annals of Statistics 9, 130—134.

RUBIN, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. John Wiley and Sons, New York.

RUBIN, D.B. (1996). Multiple imputation aften 18+ years. Journal of the American Statistical Association 91, 473—489.

RUBIN, D.B., and Schenker, N. (1991). Multiple imputation in health-care databases: An overview and some applications. Statistics in Medicine 10, 585—598.

SCHAFER, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman and Hall, London.