Public Health Nutrition: 11(4), 361–370　　DOI: 10.1017/S1368980007000365

Comparing methods for handling missing values in food-frequency questionnaires and proposing *k* nearest neighbours imputation: effects on dietary intake in the Norwegian Women and Cancer study (NOWAC)

Christine L Parr[1,]*, Anette Hjartåker[2], Ida Scheel[3], Eiliv Lund[4], Petter Laake[1] and Marit B Veierød[1]

[1]Institute of Basic Medical Sciences, Department of Biostatistics, University of Oslo, PO Box 1122 Blindern, N-0317 Oslo, Norway: [2]Cancer Registry of Norway, Institute of Population-based Cancer Research, Norway: [3]Department of Mathematics, Statistics Division, University of Oslo, Norway: [4]Institute of Community Medicine, University of Tromsø, Norway

Submitted 1 July 2006: Accepted 17 April 2007: First published online 2 July 2007

Abstract

Objective: To investigate item non-response in a postal food-frequency questionnaire (FFQ), and to assess the effect of substituting/imputing missing values on dietary intake levels in the Norwegian Women and Cancer study (NOWAC). We have adapted and probably for the first time applied *k* nearest neighbours (KNN) imputation to FFQ data.

Design: Data from a recent reproducibility study were used. The FFQ was mailed twice (test–retest) about 3 months apart to the same subjects. Missing responses in the test FFQ were imputed using the null value (frequencies = null, amount = smallest), the sample mode, the sample median, KNN, and retest values.

Setting: A methodological substudy of NOWAC, a national population-based cohort.

Subjects: A random sample of 2000 women aged 46–75 years was drawn from the cohort in 2002 (response 75%). The imputation methods were compared for 1430 women who completed at least 50% of the test FFQ.

Results: We imputed 16% missing values in the overall test data matrix. Compared to null value imputation, the largest differences in estimated dietary intake were seen for KNN, and for food items with a high proportion of missing. Imputation with retest values increased total energy intake, indicating that not all missing values are caused by respondents failing to specify no consumption, and that null value imputation may lead to underestimation and misclassification.

Conclusion: Missing values in FFQs present a methodological challenge. We encourage the application and evaluation of newer imputation methods, including KNN, which may reduce imputation errors and give more accurate intake estimates.

Keywords
Food-frequency questionnaire
Missing values
Non-response
Imputation
Data quality
Bias

Postal food-frequency questionnaires (FFQs) have long been the standard tool for assessing diet in large-scale nutritional epidemiological studies. The method has many advantages but respondents may return incomplete FFQs, in particular when there is no in-person contact. Missing answers to individual food items, referred to as item non-response, present both computational and conceptual problems in the estimation of dietary intake. One approach to handling missing values is imputation, or the practice of 'filling in' plausible values for the skipped items. Imputation is practical because it creates a complete data set at the outset, which can be used as input for dietary intake calculation programs. It prevents loss of statistical power caused by subject exclusion and potential selection bias when the exclusion is related to characteristics of the subjects. On the other hand, it may seem conceptually problematic. Item non-response usually occurs for reasons unknown to the researcher, and imputation may distort estimates, standard errors and *P* values of tests[1]. However, statistical procedures for handling missing data are a developing field and the methods are improving[2], although there are few examples of more advanced methods being applied to FFQ data. Challenges may include a large number of variables,

*Corresponding author: Email c.l.parr@medisin.uio.no which must be aggregated to calculate the intake of food groups and nutrients, few respondents with complete data, dependent questions (e.g. frequency and portion size for the same food item) and an underlying missing-data mechanism that is not completely random. Most studies using FFQs do not describe how item non-response has been handled, as previously pointed out[3], but to assume no consumption[4–6], or to impute the median value[7], or a combination of the two[8] appear to be the most common practices after an initial exclusion of subjects. A few studies have evaluated different single imputation methods, including the null value compared to either the complete data[4,9] or the median value[3]. Multiple imputation (MI), a more modern procedure, was compared to single imputation and analysis of the complete data in the GISSI-Prevenzione study[10]. MI has also been applied in the Nurses' Health Study[11]. However, the published literature on the estimation of missing values in FFQs is still small, and there is currently no recommended practice.

The objective of our study was to investigate item non-response in a postal FFQ, and to assess how different methods for imputing missing values affect dietary intake levels using data from the Norwegian Women and Cancer study (NOWAC). We compare the common methods in nutritional epidemiology of imputing the null value, or the median value, to *k* nearest neighbours (KNN) imputation, a widely used procedure for missing entries in microarray data[12]. We have here adapted and applied KNN imputation to FFQ data, which is new to our knowledge. We also impute the mode value and use a repeated measurement of the FFQ on the same subjects to investigate the assumption that missing values imply no consumption.

The missing data problem in dietary intake calculations

The calculation of dietary intake from an FFQ is a chain of arithmetic operations and data aggregation. The first step is generally to estimate the food weights (grams per day) by converting the reported consumption frequencies to intakes per day and multiplying by the usual portion sizes (reported in the FFQ or determined by the investigator). The food weights are then added directly for each individual to create food group variables, e.g. 'dairy products'. The food weights are also multiplied by the nutrient and energy values per 100 g of food (adjusted for inedible waste) from a food composition table or database. The contribution from each food is subsequently added for each individual to create nutrient and energy intake variables. non-response to food frequencies or portion sizes will generate missing food weights and missing values in aggregated variables. An illustrative example is energy intake, a key variable in many analyses. Because most foods provide energy, total energy intake from the FFQ will usually be missing for respondents skipping one or more food item(s). If the majority has skipped at least one item, it may seem more reasonable to add the contributions from the available items rather than to report missing for energy intake. But to just add the non-missing items, or to impute missing with the null value, may lead to underestimation and biased results, unless it is certain that the skipped food items were not consumed. Some calculation programs add non-missing items automatically and make it easy to overlook the missing data problem, while others may require a complete data set. Since dietary intake is often calculated with programs that are questionnaire- or study-specific and not commercially available, it is impossible to know how non-response has been handled, unless explicitly stated. The EPIC-Norfolk study sets a good example by describing their computer program[13].

Subjects and methods

Study design

NOWAC is a national population-based cohort study primarily designed to study risk factors for cancer, with 102 443 women enrolled at age 30–70 years from 1991 to 1997. The cohort has been described in detail elsewhere[14]. A part of NOWAC composes the Norwegian sub-cohort in the European Prospective Investigation into Cancer and Nutrition (EPIC). Updated information about NOWAC can be found on the website (http://uit.no/kk/nowac/). Exposure information is collected by a self-instructive health and lifestyle questionnaire (eight pages) developed specifically for the cohort. The questionnaire is administered by post and optically read. The FFQ covers four consecutive pages within the larger questionnaire. The present study uses data from the reproducibility study of the FFQ[15]. The questionnaire was mailed twice (test and retest) to the same subjects, about 3 months apart in February/March and May/June 2002. In the present study, we have imputed missing values in the test FFQ. The retest was used to study how missing responses in the test were reported 3 months later.

Subjects

A random sample of 2000 women was drawn from the cohort for the reproducibility study. The procedure has been previously described[15]. Five women had not given informed consent to further contact and were therefore excluded. The retest questionnaire was returned by 1496 (75%) of the 1995 women, but three had left the entire FFQ section blank. The imputation methods were compared for 1430 (96%) of the 1496 women who had completed at least 50% of the test FFQ. The exclusion was done to study the effects of imputation in a sample likely to be included in a regular epidemiological analysis of

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

e.g. diet and cancer. All subject characteristics were based on self-reported information in the test and retest questionnaires, except for age, which was taken from the national population registry.

The FFQ

The FFQ structure and the dietary intake calculations have been described elsewhere[15], as well as the reproducibility and validity of the questionnaire[15,16]. In short, the FFQ is designed to assess habitual diet over the past year, with emphasis on fish and other traditional food items in the study population. The FFQ is mostly structured as smaller blocks or grids with two to nine similar items grouped together under a question heading, but with some single questions about only one food item. Food quantity is estimated by assigning standard portions, or by separate portion size questions. The response options for consumption frequencies and portion sizes are predefined and listed in increasing order with check boxes to facilitate completion and optical reading. For consumption frequencies, the first alternative is always 'never/rarely'.

The dietary intake was calculated from a total of 132 questions (consumption frequencies = 91, types of fat used on bread = 7, amounts = 28, time of year for the consumption of different species of fish = 6). The food groups were based on the classification system in the EPIC-SOFT program for conducting 24-hour dietary recalls in the EPIC study[17], but with some modifications[15]. The daily intake of food groups, energy, and nutrients was calculated using an analysis program developed at the Institute of Community Medicine, University of Tromsø, for SAS software.

Definition of missing values

All non-responses in an FFQ may not be considered missing values if e.g. respondents are directed to skip questions that are not relevant. In the present study, missing portion sizes were permitted if the consumption frequency was 'never/rarely'. Missing frequencies were permitted for fish if the preceding question about the time of year for consumption was 'never/rarely'. To identify users of cod liver oil supplements and alcohol, an introductory yes/no question was included, e.g. 'Do you take cod liver oil (liquid)?' If the answer was 'no', it was permitted to skip further questions about consumption.

The response option 'Do not use fat on bread' was listed before types (maximum 7) and the usual amount/layer (e.g. 'thin') on a slice of bread could be specified. 'Do not use fat on bread' and types of fat composed a group of eight separate 0/1 variables, each with one check box to confirm 'yes'. This layout presents a challenge because the answer 'no' cannot be distinguished from a missing value when the box is left open. However, if both the use of fat and all types of fat had open boxes, this was interpreted as missing information and defined as one missing value because either 'Do not use fat on bread' or at least one type should have been marked.

Item non-response in the FFQ was evaluated for 136 questions (132 about frequencies, amounts, types of fat and seasonality, and four yes/no questions about user status). But since the eight check boxes for fat on bread could give only one missing value, the maximum number of missing values was 129.

Methods used for imputing missing values

No consumption and the smallest portion size

The original NOWAC program for calculating dietary intake imputes missing consumption frequencies with the null value (no consumption) and missing portion sizes with the smallest portion, for a conservative intake estimate. Thus, the food quantity will be null for all missing frequencies. When all information about fat on bread is missing, the null value is imputed. If only type is missing, the most common type (soft margarine) is imputed. A factor of 0.5 corresponding to half the year is assigned for missing information about seasons for the consumption of different species of fish. When the initial yes/no questions about the use of cod liver oil supplements and alcohol are missing or create inconsistent responses (e.g. do not take supplements, but the frequency is every day), the frequency response has priority. The method described here was used as the reference and compared to the other methods described below. In all methods, permitted missing values were treated as null intake.

Mode and median

Substitution by the sample mode or median may be described as cross-sectional imputation techniques, since the values are taken from the available data in the same data set. Missing values for user status were imputed by the most frequent answer, i.e. to use fat on bread and drink alcohol, but not take cod liver oil supplements. Missing values in frequencies and amounts were then imputed based on reported or imputed user status (null value for non-users, and the mode/median for users). Most users specified one type of fat on bread. Therefore, the most common type was imputed.

Retest values

Missing values in the first FFQ measurement (test) were imputed by non-missing values in the second measurement (retest) of the same individual. This may be regarded as longitudinal imputation, although the time to the retest was relatively short. The consumption frequency and amount questions for a given food item were imputed as a pair, i.e. if one value was missing in the test then both values were taken from the retest. In the case where the pair of retest values (frequency and amount) was incomplete, the missing test value was imputed if the retest value was available. There were retest answers for

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

50% of the values missing in the test. Residual missing values in frequencies were treated as null intake and residual missing in amounts as the smallest portion size.

k nearest neighbours imputation

When applying KNN imputation to FFQ data, missing values for each respondent were imputed using values from the *k* most similar respondents. The idea behind KNN imputation[12] is to take advantage of positive correlations between rows. It is assumed that information about the missing values in row *p* is best provided by the *k* rows most similar to row *p* (the *k* nearest neighbours). A missing value in row *p* (here respondent *p*) in column *c* (here question *c*) is therefore imputed by averaging the values that the *k* nearest neighbours have in column *c*. The *k* nearest neighbours must have non-missing entries in column *c*. The similarity between row *p* and row *p*′ is measured by the Euclidean distance between the two rows (omitting the columns for which row *p* and/or row *p*′ have missing values), divided by the number of columns where both row *p* and row *p*′ have non-missing entries. KNN imputation performs best when there are strong positive correlations between rows.

To adapt KNN imputation to the FFQ setting, we modified the function *impute.knn* in the package *impute*[18] for the statistical software R. First, when computing the distance between two rows, we scaled all columns so that the columns contribute to the distance on the same scale. Second, because the FFQ data are categorical, the imputed values were rounded to the nearest category. Third, the set of seven questions regarding type of fat used on bread were highly dependent (most respondents reported only one type) so that we had to tailor the imputation to handle the dependency. Last, we introduced conditional imputation to prevent permitted missing responses, as defined earlier, from being imputed. A necessary condition for using KNN imputation on categorical variables is that the categories are either ordered or binary, which was the case in this study.

When applying KNN imputation, the number of nearest neighbours (*k*) must be specified. Troyanskaya *et al.*[12] have showed the optimal *k* to be between 10 and 20 for microarray data, and the default value in the software is 10. To determine a *k* value for the FFQ data, we calculated the root-mean-squared error (RMSE) for the KNN imputed data with *k* between 5 and 135, as well as for the data imputed with the null value and the mode and median (Fig. 1). The RMSE calculation was based on the subset of missing test data for which the retest data were non-missing. The non-missing retest data were used as the reference, since the true values for the responses missing in the test are unknown. Therefore, we also added the RMSE for test vs. retest for the subset of values which was non-missing in both data sets. Based on Fig. 1, three different values of the parameter *k* (10, 20 and 60) were selected to study the effects on the estimated energy and nutrient intake from the KNN imputed data (Appendix). The intake was stable, although a negligible decrease could be spotted for increasing values of *k*. In the main tables we used *k* = 20.

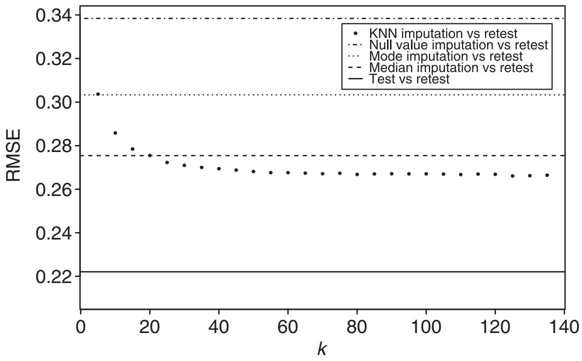

Fig. 1 RMSE for different values of the parameter *k* in KNN imputation of the test food-frequency questionnaire with non-missing retest data as the reference. The curve is compared to imputation with the null, mode and median values. The RMSE for test vs. retest for the subset of values which was non-missing in both data sets is also included ($n = 1430$) (KNN – *k* nearest neighbours; RMSE – root-mean-squared error)

Statistical analysis

The proportion of missing in the FFQ was calculated for each individual by dividing the number of missing values by the maximum number possible (129 minus the permitted missing values). The distribution was skewed, so the proportion of missing (%) is presented as the median value with lower and upper quartiles (Q1, Q3) by categories of selected background variables (Table 1). Estimated dietary intake is presented as the median value for the null imputed data, and as within-person differences for the other methods relative to the null imputation. The within-person differences did not fulfil the normality assumption and are therefore presented as both mean and median with quartiles (Q1, Q3). Distribution free confidence limits (95%) for the median differences were also calculated, but not included due to the large number of null values, for which the upper confidence limits were generally also null. We used SAS 9.1 for the data analysis and all imputations, except KNN, which was done with R software.

Results

Rate of missing in the FFQ

The overall data matrix for the 1496 respondents and 136 FFQ variables had 18% missing values for the test and 16% for the retest FFQ, after controlling for permitted missing values. The FFQ was fully completed by an equal proportion of respondents in the test and the retest (6%).

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

Table 1 Median (quartiles) for the proportion of missing values (%) in the food frequency questionnaire (test and retest) by selected background characteristics of the respondents, *n* = 1496*

	Test, % missing			Retest, % missing		
Characteristic	*n*	Median	(Q1, Q3)†	*n*	Median	(Q1, Q3)
Age (years)						
46–55	465	9	(3, 21)	465	8	(3, 17)
56–65	701	13	(5, 26)	701	11	(4, 23)
66–75	330	23	(12, 33)	330	18	(8, 32)
Household income (1000 NOK)‡						
<150	178	24	(11, 34)	181	20	(7, 33)
151–300	398	14	(6, 27)	422	12	(5, 25)
301–450	398	12	(5, 22)	397	9	(4, 20)
451–600	231	10	(3, 21)	222	9	(3, 18)
>600	179	9	(4, 20)	180	8	(3, 18)
Marital status§						
Married		–		1063	11	(5, 23)
Cohabiter		–		80	7	(3, 14)
Unmarried		–		42	8	(2, 26)
Divorced		–		143	9	(4, 24)
Widowed		–		157	16	(8, 30)
Health status						
Very good	378	14	(5, 25)	384	10	(4, 23)
Good	911	13	(6, 26)	927	12	(5, 24)
Poor/very poor	146	15	(7, 30)	139	10	(4, 25)
Try to lose weight						
Yes	540	12	(5, 25)	565	10	(4, 21)
No	956	15	(6, 28)	931	12	(5, 25)
Daily smoker						
Yes	336	12	(4, 26)	342	11	(4, 23)
No	1128	14	(6, 26)	1136	11	(5, 24)
Teetotaller						
Yes	177	20	(9, 34)	181	16	(6, 31)
No	1256	12	(5, 24)	1282	10	(4, 22)
Take cod liver oil supplements						
Yes	602	14	(6, 26)	613	11	(4, 23)
No	832	13	(5, 25)	844	10	(4, 23)
Days to return questionnaire¶						
6–9		–		390	11	(4, 24)
10–15		–		361	13	(5, 25)
16–37		–		369	12	(4, 25)
38–160		–		376	10	(4, 22)

* *n* may not total to 1496 for each characteristic due to missing values.
† Q1 = lower quartile (25th percentile), Q3 = upper quartile (75th percentile).
‡ Misprint in the questionnaire: the category <150 should have been ≤ 150. 1000 Norwegian kroner (NOK) ≈ 125 €.
§ Only presented for the retest, due to an optical reading error in the test.
¶ Could only be calculated for the retest.

After excluding individuals with >50% missing in the test, the test data matrix for the remaining 1430 (96%) respondents had 16% missing values, which were imputed. Respondents had a median value (Q1, Q3) of 13% (5, 25) missing values.

Table 1 shows how the proportion of missing values (%) in the test and retest FFQ varied by categories of selected background characteristics of the respondents. The median proportion increased with age and decreased with household income up to 450 000 NOK. The proportion was ≥16% in the oldest age group (66–75 years), in the lowest household income group (<150 000 NOK), and among widows and teetotallers. Health status, trying to lose weight, smoking status, supplement use and days used to return the FFQ appeared to have little or no effect.

The percentage of imputed values for consumption frequencies ranged from < 1% for potatoes to 50% for instant coffee (Table 2). The median value (Q1, Q3) was 12% (6, 25). All items that were not part of question blocks had ≤ 4% missing, e.g. potatoes, yoghurt, breakfast cereal, shellfish and eggs. Unspecific questions about 'other' items included at the end of some blocks, e.g. for fruits, vegetables, meat dishes and fish, had a relatively high percentage of missing with 27–38%. Most items with ≥30% missing were part of question blocks listing several types of the same item, e.g. types of milk, cheese, bread and coffee. Items with a high percentage of missing values in the test also tended to have a high percentage of missing in the retest.

Effects of imputation on dietary intake

Table 2 shows the daily intake of food groups after the missing values in the test FFQ were imputed. Imputation

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

Table 2 The intake of food groups (g day^{-1}) after imputing missing values in the test food-frequency questionnaire. Within-person differences (D) are presented as mean and median (quartiles) for imputation with retest values*, mode, median and KNN, relative to imputation with the null value, $n = 1430$

	Null value	Retest (D)			Mode (D)			Median (D)			KNN ($k = 20$) (D)		
Food group (range of missing for item)	Median	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)
Potatoes (<1%)	126	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)
Vegetables (2–27%)	102	7	0	(0, 4)	9	0	(0, 9)	10	0	(0, 12)	13	0	(0, 15)
Fruits (4–23%)	178	11	0	(0, 0)	24	0	(0, 42)	24	0	(0, 42)	24	0	(0, 42)
Dairy products	174	19	0	(0, 14)	2	0	(0, 6)	11	10	(0, 19)	77	54	(4, 132)
Milk (29–45%), yoghurt (2%), cheese (30–43%)	150	19	0	(0, 11)	2	0	(0, 6)	10	10	(0, 19)	77	54	(4, 132)
Cream desserts, milk-based puddings (1–7%)	20	1	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)
Cereal and cereal products	153	9	0	(0, 0)	10	0	(0, 4)	15	11	(0, 15)	20	11	(0, 26)
Bread, crisp bread (8–48%), breakfast cereal (3%)	122	8	0	(0, 0)	9	0	(0, 4)	14	11	(0, 11)	18	11	(0, 26)
Pasta and rice (3–6%)	22	1	0	(0, 0)	1	0	(0, 0)	1	0	(0, 0)	1	0	(0, 0)
Meat and meat products	91	5	0	(0, 2)	7	0	(0, 8)	7	0	(0, 8)	10	0	(0, 15)
Red meat and chicken (4–14%)	26	1	0	(0, 0)	2	0	(0, 0)	2	0	(0, 0)	2	0	(0, 0)
Processed meat (4–25%)	64	4	0	(0, 0)	6	0	(0, 8)	5	0	(0, 8)	8	0	(0, 10)
Fish and shellfish	112	9	0	(0, 9)	9	0	(0, 11)	11	5	(0, 16)	15	6	(0, 24)
Whole fish (filets, steaks) (4–38%) and shellfish (3%)	67	5	0	(0, 2)	7	0	(0, 0)	6	0	(0, 6)	8	1	(0, 11)
Fish products (3–25%)	41	4	0	(0, 1)	2	0	(0, 4)	5	1	(0, 8)	7	1	(0, 11)
Eggs (3%)	8	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)
Fat (margarine, butter) on bread	9	1	0	(0, 0)	1	0	(0, 0)	2	0	(0, 2)	2	0	(0, 3)
Orange juice, carbonated soft drinks and diluted syrups (12–33%)	31	2	0	(0, 0)	2	0	(0, 0)	2	0	(0, 0)	2	0	(0, 0)
Cakes (4–10%)	43	9	0	(0, 0)	0	0	(0, 0)	5	0	(0, 0)	23	0	(0, 43)
Coffee (boiled, filtered, instant) (16–50%)	300	17	0	(0, 0)	47	0	(0, 0)	47	0	(0, 0)	117	60	(0, 180)
Alcoholic beverages (wine, beer, spirits) (7–13%)	17	3	0	(0, 0)	0	0	(0, 0)	1	0	(0, 0)	3	0	(0, 0)
Condiments and sauces for fish (11–39%)	6	1	0	(0, 0)	0	0	(0, 0)	1	0	(0, 2)	2	1	(0, 4)
Sweets and salty snacks (2–13%)	22	1	0	(0, 0)	3	0	(0, 0)	2	0	(0, 0)	2	0	(0, 0)
Cod liver oil supplements (5–24%)	0	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)

KNN – *k* nearest neighbours imputation.

* Available for 50% of missing values. Residual missing treated as null intake.

† Missing in food frequency questions. Not specified for fat on bread since the intake is calculated from the frequency of bread consumption.

Table 3 The daily intake of energy and selected nutrients after imputing missing values in the test food-frequency questionnaire. Within-person differences (D) are presented as mean and median (quartiles) for imputation with retest values*, mode, median and KNN, relative to imputation with the null value, $n = 1430$

	Null value	Retest (D)			Mode (D)			Median (D)			KNN ($k = 20$) (D)		
Nutrient	Median	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)
Energy (kJ)	6407	415	220	(22, 526)	421	202	(27, 579)	638	455	(135, 943)	962	743	(230, 1486)
Protein (g)	71	5	2	(0, 7)	5	2	(0, 7)	7	5	(2, 11)	12	10	(3, 18)
Total fat (g)	57	4	2	(0, 6)	4	2	(0, 6)	7	5	(1, 10)	10	8	(2, 15)
Polyunsaturated fat (g)	10	1	0	(0, 1)	1	0	(0, 1)	1	1	(0, 2)	2	1	(0, 2)
Total carbohydrate (g)	176	10	4	(0, 12)	11	4	(0, 16)	15	9	(2, 22)	23	16	(4, 36)
Dietary fibre (g)	19	1	0	(0, 1)	1	1	(0, 2)	2	1	(0, 2)	2	1	(0, 3)
Sugar (g)	20	1	0	(0, 0)	2	0	(0, 0)	1	0	(0, 1)	2	0	(0, 4)
Alcohol (g)	1	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)
Retinol (RE, μg)	1102	60	21	(0, 69)	75	23	(2, 88)	103	61	(15, 134)	139	87	(23, 197)
Vitamin D (μg)	9	1	0	(0, 1)	1	0	(0, 1)	2	0	(0, 1)	2	1	(0, 2)
Vitamin E (mg)	8	1	0	(0, 1)	1	0	(0, 1)	1	0	(0, 1)	1	1	(0, 1)
Vitamin C (mg)	99	5	0	(0, 4)	9	1	(0, 13)	10	3	(0, 17)	12	3	(0, 17)
Calcium (mg)	584	60	16	(1, 71)	32	18	(2, 55)	90	84	(11, 145)	188	176	(50, 295)
% energy from protein	19	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 1)
% energy from fat	33	0	0	(0, 1)	0	0	(0, 0)	1	0	(0, 1)	1	0	(0, 1)
% energy from carbohydrate	47	0	0	(−1, 0)	0	0	(−1, 0)	0	0	(−1, 0)	−1	0	(−2, 0)
% energy from sugar	5	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(1, 0)
% energy from alcohol	0	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)	0	0	(0, 0)

KNN – k nearest neighbours imputation; RE – retinol equivalents.
* Available for 50% of missing values. Residual missing treated as null intake.

with retest values and the sample mode gave similar results compared to imputation with the null value. The median of the differences was null for all food groups, with some changes in the upper quartiles (Q3). The different imputation methods had little or no effect on the intake of food groups with a low percentage of missing, e.g. 'potatoes', 'cream desserts, milk-based puddings', 'pasta and rice', and 'red meat and chicken'. The median of the differences was >0 for some food groups when imputed with the sample median and KNN, most of which had >30% missing in one or more food items. KNN tended to give a higher intake than the other imputation methods, with some considerable differences. The median of the differences for the daily intake of 'milk, yoghurt and cheese' was 10 g (Q3 = 19) when imputed with the sample median, but 54 g (Q3 = 132) with KNN. For 'coffee' the value was 0 g (Q3 = 0) with the sample median, but 60 g (Q3 = 180) with KNN.

The differences in the estimated intake of 'milk, yoghurt and cheese' were reflected in the intake of calcium (Table 3). The median of the differences compared to the null value imputation (Q1, Q3) was 84 mg (11, 145) for the sample median, and 176 mg (50, 295) for KNN. For energy intake, the median of the differences was about 200 kJ for imputation with retest values and the sample mode, 455 kJ with the sample median and 743 kJ with KNN. The imputation methods did not change the % energy from protein, fat, carbohydrate, sugar and alcohol.

Discussion

The present study was undertaken to investigate item non-response in the FFQ developed for the NOWAC study and to assess effects of different imputation methods on dietary intake. Compared to null value imputation, the largest differences were seen for KNN, and for food items with a high proportion of missing values. Imputation with retest values increased total energy intake, indicating that not all missing values are caused by respondents failing to specify no consumption.

Item non-response

The proportion of missing within respondents was positively associated with age (data not shown). Likely explanations include impaired cognitive function or a diet with fewer food items. Other factors that seemed to increase non-response, including low income, being a widow or a teetotaller, were also associated with age. Somewhat surprisingly, health status did not seem to affect non-response or to be related to age.

In the present study, we observed a very low percentage of missing values in all items that were not part of a larger question block, or grid. FFQs with a non-grid format may be cognitively easier for respondents to complete, but it increases the page length and thus the costs for printing, scanning and mailing. One example is the 36-page dietary history questionnaire (DHQ) developed at the National Cancer Institute in USA. When compared to a shorter FFQ with a traditional format, the DHQ performed better for questions on portion sizes and dietary supplements, but not for consumption frequencies[19]. The proportion of missing/uninterpretable responses in the DHQ was low, but similar to the FFQ. In our study, several blocks of items also had relatively low proportions of missing, e.g. cakes (six items with 4–10% missing) and meat (nine items with 4–14% missing, except 'other meat dishes' with 27%). It could be that smaller blocks are cognitively not too demanding, or that the consumption awareness or desirability of some foods encourages responses. A complete non-grid format is not practical in all studies. One alternative is to mix single questions with smaller blocks as in the NOWAC questionnaire, and to put key foods as single questions or as the first item in a block.

Effects of imputation on dietary intake

To assume that missing values are due to respondents failing to specify the option for no consumption greatly simplifies dietary intake calculations and may be reasonable for certain food items, e.g. types of milk and fat on bread[6,20]. Some studies have validated this common assumption by follow-up telephone interview[3,20,21] or a resurvey[9]. In one study, the 'true' proportion of null consumption was found to vary greatly between food items from 0% (potatoes) to 96% (roe and fish pâté)[3]. Using our retest data to estimate the proportion of null consumption (i.e. the proportion of 'never/rarely' answers in the retest among the missing test values, for which there was a non-missing retest value), the range was 5% (carrots) to 86% (whole milk) (data not shown). Although the probability of null consumption may be high for some food items, it may be low for others. Food items could be consumed, but skipped due to lack of motivation, fatigue, oversight, difficult or sensitive questions, unclear instructions or other reasons. The increase in absolute intake when missing values are imputed with values from our retest or other resurveys[9,21] supports this.

Compared to imputation with the null value, the sample mode only increased the dietary intake slightly. In our study, 47% of the frequency variables had a mode value of null. The sample mode is the most frequent answer, but mode = 0 or 'never/rarely' may not reflect the probability of a food being consumed if the higher response categories add up to a larger proportion. Thus, the mode may be better for imputing variables on a nominal scale. For ordered categories, such as frequencies and portion sizes, the probability is better reflected in the median. Of the frequency variables, 29% had a median value of null. When using the sample median, we observed a higher intake, in particular for food groups with a high proportion of missing values. One study found marginal differences between imputation with the null and the sample median[3], but this study also had a low proportion of missing. KNN gave the highest intake. The largest differences were seen for the food groups 'milk, yoghurt, cheese' and 'coffee'. When inspecting the distribution of the consumption frequencies for each food item, the median of the KNN imputed values tended to be one category higher than the median values before imputation, but the highest category was never imputed with KNN.

It is interesting to note that the RMSE value (Fig. 1) was equal for imputation with KNN ($k = 20$) and the sample median, but with some clear differences in dietary intake between the methods. Oppositely, the energy and nutrient intake was stable for $k = 10$, 20 and 60, even though the RMSE values were different. A weakness of RMSE is that it does not necessarily measure the effect of imputation on the dietary intake or other outcomes. This has also been found for microarray data[22].

When comparing the effects of imputation on dietary intake (food groups and nutrients) we used the null imputed test data as the reference method for two main reasons: null imputation appears to be common practice and gives the most conservative intake estimate. Our reasons for not using e.g. the complete retest data as reference is that it would be difficult to interpret the effects of imputation separately from the underlying differences between the test and retest measurements[15]. Also, the reduction in sample size and statistical power would be too large for a meaningful analysis in this study since the retest data were complete for only 6% of the respondents. The percentage of missing test values in this small subsample ($n = 91$) was low and not representative of the study sample ($n = 1430$). Since our reference method can only be used for relative comparisons, we cannot conclude which imputation method is more accurate, only that the choice of method may affect dietary intake. We think the lack of an absolute reference method or a gold standard is a general problem in comparative studies of dietary intake. Therefore, our next step in the evaluation of the imputation methods would be to do a simulation study with a complete data set as the reference.

Imputation uncertainty

All imputation methods used in the present study fall into the category of single imputation methods, implying that each missing value is replaced by a single value. Single imputation methods are usually easy to implement, but ignore any uncertainty about the correct value to impute. This can be estimated by doing MI, or repeated simulations of the missing values[23]. MI is a model-based approach, relying on specific modelling assumptions, and the method may be difficult to use without proficiency in advanced statistics. To our knowledge, MI has been applied to FFQ data in two recent studies[10,11]. KNN is more sophisticated than the other single imputation

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

methods used here, since values are estimated for each individual, but without having to specify a strict model as with MI. In the present study we based the KNN imputation on FFQ variables only, but additional predictors could also have been used.

Both single and multiple imputation rely on a mechanism of missingness known as missing at random (MAR)[1], which requires that the probability that a value is missing is independent of the underlying value that is missing. However, if no consumption (or high consumption) is an important reason for missing values in FFQs, the predominant mechanism may not be MAR, but NMAR (not missing at random). Imputation methods using many predictors, such as MI and KNN, give better protection against departures from the MAR assumption than imputation with e.g. the sample median. The default limit in the function *impute.knn*[18] for doing imputation based on KNN is 50% missing in rows (here, respondents). For more than 50% missing a column mean is used. In the present study, we excluded individuals with >50% missing. However, the criteria used to exclude questionnaires vary between investigators[21]. If only a specific food group is of interest (e.g. alcohol), exclusion of subjects with missing values for all or most food group items (e.g. beer, wine and spirits) may be better than imputation. However, it is important to check that the exclusion does not significantly reduce the statistical power, or change the distribution of other exposure variables in the analysis.

Implications

Imputation may affect absolute intake levels and the ranking of subjects[9,21], which has implications for risk estimation in epidemiological studies, as well as for those who are defined as under- and overreporters and excluded from the analysis. When we excluded those in the lower (1%) and upper (99%) percentiles of energy intake after imputing the data (28 subjects with each method), 60% of the subjects were excluded by all methods (details not shown). Our data also show that by using different imputation methods, the median daily energy intake can be increased from 6.4 MJ for the null value to 7.3 MJ for KNN, with a median increase (Q1, Q3) of 11% (3, 24). This can affect the interpretation of FFQ data in general and in validation studies.

In conclusion, the calculation of dietary intake from FFQs is affected by the proportion of missing data and the imputation method used. As an overall imputation strategy, the null value is likely to lead to underestimation of dietary intake and misclassification. However, missing values cannot be estimated without error. We encourage the application and evaluation of more refined imputation methods, which are described in statistics literature, and which may reduce imputation errors and give more accurate intake estimates. To determine if KNN performs better than the other methods used, our next step would be to do a simulation study.

Acknowledgements

Sources of funding: The work of C.L.P. was supported by the Norwegian Foundation for Health and Rehabilitation.

Conflict of interest declaration: The authors declare no conflict of interest.

Authorship contributions: A.H., P.L., E.L. and M.B.V. were responsible for the initiation, design and funding of the present study. E.L. is the Principal Investigator of the Norwegian Women and Cancer study, and A.H. developed the food-frequency questionnaire. The imputation of the data was done by I.S. (KNN imputation) and C.L.P. (other methods). C.L.P. calculated the dietary intake, performed the statistical analysis, and drafted the manuscript, except for the section on KNN and Fig. 1, which were prepared by I.S. M.B.V. and P.L. contributed to the statistical analysis. All authors were involved in the interpretation of results and manuscript revision, and all authors approved the final version.

Acknowledgements: The authors are very grateful to Professor Arnoldo Frigessi at the University of Oslo for ideas, including the use of KNN imputation, and for fruitful discussions.

References

1 Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, 2nd edn. New Jersey: Wiley, 2002.

2 Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological Methods* 2002; **7**: 147–77.

3 Hansson LM, Galanti MR. Diet-associated risks of disease and self-reported food consumption: how shall we treat partial nonresponse in a food frequency questionnaire? *Nutrition and Cancer* 2000; **36**: 1–6.

4 Holmberg L, Ohlander EM, Byers T, Zack M, Wolk A, Bruce A, *et al.* A search for recall bias in a case–control study of diet and breast cancer. *International Journal of Epidemiology* 1996; **25**: 235–44.

5 Cade J, Thompson R, Burley V, Warm D. Development, validation and utilisation of food-frequency questionnaires – a review. *Public Health Nutrition* 2002; **5**: 567–87.

6 Johansson I, Hallmans G, Wikman A, Biessy C, Riboli E, Kaaks R. Validation and calibration of food-frequency questionnaire measurements in the Northern Sweden Health and Disease cohort. *Public Health Nutrition* 2002; **5**: 487–96.

7 Gaard M, Tretli S, Loken EB. Dietary fat and the risk of breast cancer: a prospective study of 25,892 Norwegian women. *International Journal of Cancer* 1995; **63**: 13–17.

8 Goldbohm RA, van den Brandt PA, Brants HA, van't Veer P, Al M, Sturmans F, *et al.* Validation of a dietary questionnaire used in a large-scale prospective cohort study on diet and cancer. *European Journal of Clinical Nutrition* 1994; **48**: 253–65.

9 Ahn Y, Paik HY, Ahn YO. Item nonresponses in mailed food frequency questionnaires in a Korean male cancer cohort study. *Asia Pacific Journal of Clinical Nutrition* 2006; **15**: 170–7.

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press

10 Barzi F, Woodward M, Marfisi RM, Tognoni G, Marchioli R. Analysis of the benefits of a mediterranean diet in the GISSI-prevenzione study: a case study in imputation of missing values from repeated measurements. *European Journal of Epidemiology* 2006; **21**: 15–24.

11 Michels KB, Rosner BA, Chumlea WC, Colditz GA, Willett WC. Preschool diet and adult risk of breast cancer. *International Journal of Cancer* 2006; **118**: 749–54.

12 Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, *et al.* Missing value estimation methods for DNA microarrays. *Bioinformatics* 2001; **17**: 520–5.

13 Welch AA, Luben R, Khaw KT, Bingham SA. The CAFE computer program for nutritional analysis of the EPIC-Norfolk food frequency questionnaire and identification of extreme nutrient values. *Journal of Human Nutrition and Dietetics* 2005; **18**: 99–16.

14 Lund E, Kumle M, Braaten T, Hjartaker A, Bakken K, Eggen E, *et al.* External validity in a population-based national prospective study – the Norwegian Women and Cancer Study (NOWAC). *Cancer Causes & Control* 2003; **14**: 1001–8.

15 Parr CL, Veierod MB, Laake P, Lund E, Hjartaker A. Test–retest reproducibility of a food frequency questionnaire (FFQ) and estimated effects on disease risk in the Norwegian Women and Cancer Study (NOWAC). *Nutrition Journal* 2006; **5**: 4.

16 Hjartaker A, Lund E, Bjerve KS. Serum phospholipid fatty acid composition and habitual intake of marine foods registered by a semi-quantitative food frequency questionnaire. *European Journal of Clinical Nutrition* 1997; **51**: 736–42.

17 Slimani N, Ferrari P, Ocke M, Welch A, Boeing H, Liere M, *et al.* Standardization of the 24-hour diet recall calibration method used in the European Prospective Investigation into Cancer and Nutrition (EPIC): general concepts and preliminary results. *European Journal of Clinical Nutrition* 2000; **54**: 900–17.

18 Hastie T, Tibshirani R, Narasimhan B, Chu G. *The Package Impute for Imputation of Microarray Data With R Software* [online], 2006. Available at http://cran.r-project.org/. Accessed 28 March 2007.

19 Subar AF, Ziegler RG, Thompson FE, Johnson CC, Weissfeld JL, Reding D, *et al.* Is shorter always better? Relative importance of questionnaire length and cognitive ease on response rates and data quality for two dietary questionnaires. *American Journal of Epidemiology* 2001; **153**: 404–9.

20 Kuskowska-Wolk A, Holte S, Ohlander EM, Bruce A, Holmberg L, Adami HO, *et al.* Effects of different designs and extension of a food frequency questionnaire on response rate, completeness of data and food frequency responses. *International Journal of Epidemiology* 1992; **21**: 1144–50.

21 Caan B, Hiatt RA, Owen AM. Mailed dietary surveys: response rates, error rates, and the effect of omitted food items on nutrient values. *Epidemiology* 1991; **2**: 430–436.

22 Scheel I, Aldrin M, Glad IK, Sorum R, Lyng H, Frigessi A. The influence of missing value imputation on detection of differentially expressed genes from microarray data. *Bioinformatics* 2005; **21**: 4272–9.

23 Schafer JL. Multiple imputation: a primer. *Statistical Methods in Medical Research* 1999; **8**: 3–15.

Appendix – The mean and median daily intake of energy and selected nutrients after imputing missing values in the test food-frequency questionnaire by *k* nearest neighbours imputation for different values for the parameter *k*. Within-person differences are presented as mean and median (quartiles) for *k* = 10 and *k* = 60 relative to *k* = 20, which is used in the main tables, *n* = 1430

	Mean			Median			Within-person differences (*k*20–*k*10)			Within-person differences (*k*20–*k*60)		
Variable	*k* = 10	*k* = 20	*k* = 60	*k* = 10	*k* = 20	*k* = 60	Mean	Median	(Q1, Q3)	Mean	Median	(Q1, Q3)
Energy (kJ)	7538	7501	7461	7370	7324	7297	−38	0	(−109, 73)	39	0	(−25, 116)
Protein (g)	85	85	85	83	82	82	0	0	(−2, 1)	0	0	(0, 1)
Total fat (g)	70	69	69	66	66	66	0	0	(−1, 1)	0	0	(0, 1)
Polyunsaturated fat (g)	12	12	12	12	12	12	0	0	(0, 0)	0	0	(0, 0)
Total carbohydrate (g)	203	202	201	198	198	197	−1	0	(−3, 2)	1	0	(−1, 3)
Dietary fibre (g)	22	22	21	21	21	21	0	0	(0, 0)	0	0	(0, 0)
Sugar (g)	24	24	24	22	22	22	0	0	(0, 0)	0	0	(0, 0)
Alcohol (g)	2	2	2	1	1	1	0	0	(0, 0)	0	0	(0, 0)
Retinol (RE, μg)	1327	1322	1317	1257	1246	1242	−5	0	(−14, 8)	4	0	(−4, 13)
Vitamin D (μg)	14	14	14	11	11	11	0	0	(0, 0)	0	0	(0, 0)
Vitamin E (mg)	12	12	12	9	9	9	0	0	(0, 0)	0	0	(0, 0)
Vitamin C (mg)	118	118	117	112	111	111	0	0	(0, 0)	0	0	(0, 0)
Calcium (mg)	814	809	803	783	777	770	−5	0	(−22, 13)	7	0	(−6, 16)
% energy from protein	19	19	19	19	19	19	0	0	(0, 0)	0	0	(0, 0)
% energy from fat	34	34	34	34	34	34	0	0	(0, 0)	0	0	(0, 0)
% energy from carbohydrate	46	46	46	46	46	46	0	0	(0, 0)	0	0	(0, 0)
% energy from sugar	5	5	5	5	5	5	0	0	(0, 0)	0	0	(0, 0)
% energy from alcohol	1	1	1	0	0	1	0	0	(0, 0)	0	0	(0, 0)

RE – retinol equivalents.

https://doi.org/10.1017/S1368980007000365 Published online by Cambridge University Press