

# Handling missing data in an FFQ: multiple imputation and nutrient intake estimates

Mari Ichikawa<sup>1</sup>, Akihiro Hosono<sup>1</sup>, Yuya Tamai<sup>1</sup>, Miki Watanabe<sup>1</sup>, Kiyoshi Shibata<sup>1,2</sup>, Shoko Tsujimura<sup>1</sup>, Kyoko Oka<sup>1</sup>, Hitomi Fujita<sup>1</sup>, Naoko Okamoto<sup>1</sup>, Mayumi Kamiya<sup>1</sup>, Fumi Kondo<sup>1</sup>, Ryoza Wakabayashi<sup>1</sup>, Taiji Noguchi<sup>1</sup>, Tatsuya Isomura<sup>1,3,4</sup>, Nahomi Imaeda<sup>1,5</sup>, Chiho Goto<sup>1,6</sup>, Tamaki Yamada<sup>7</sup> and Sadao Suzuki<sup>1,\*</sup>

<sup>1</sup>Department of Public Health, Nagoya City University Graduate School of Medical Sciences, 1 Kawasumi, Mizuho-cho, Mizuho-ku, Nagoya 467-8601, Japan: <sup>2</sup>Nagoya University of Economics, Inuyama, Japan: <sup>3</sup>Clinical Study Support Inc., Nagoya, Japan: <sup>4</sup>Institute of Medical Science, Tokyo Medical University, Tokyo, Japan: <sup>5</sup>Department of Nutrition, Faculty of Wellness, Shigakkan University, Obu, Japan: <sup>6</sup>Department of Health and Nutrition, School of Health and Human Life, Nagoya-bunri University, Inazawa, Japan: <sup>7</sup>Okazaki City Medical Association, Public Health Center, Okazaki, Japan

Submitted 22 May 2018: Final revision received 30 November 2018: Accepted 8 January 2019: First published online 26 February 2019

## Abstract

**Objective:** We aimed to examine missing data in FFQ and to assess the effects on estimating dietary intake by comparing between multiple imputation and zero imputation.

**Design:** We used data from the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study. A self-administered questionnaire including an FFQ was implemented at baseline (FFQ1) and 5-year follow-up (FFQ2). Missing values in FFQ2 were replaced by corresponding FFQ1 values, multiple imputation and zero imputation.

**Setting:** A methodological sub-study of the Okazaki J-MICC study.

**Participants:** Of a total of 7585 men and women aged 35–79 years at baseline, we analysed data for 5120 participants who answered all items in FFQ1 and at least 50% of items in FFQ2.

**Results:** Among 5120 participants, the proportion of missing data was 3.7%. The increasing number of missing food items in FFQ2 varied with personal characteristics. Missing food items not eaten often in FFQ2 were likely to represent zero intake in FFQ1. Most food items showed that the observed proportion of zero intake was likely to be similar to the probability that the missing value is zero intake. Compared with FFQ1 values, multiple imputation had smaller differences of total energy and nutrient estimates, except for alcohol, than zero imputation.

**Conclusions:** Our results indicate that missing values due to zero intake, namely missing not at random, in FFQ can be predicted reasonably well from observed data. Multiple imputation performed better than zero imputation for most nutrients and may be applied to FFQ data when missing is low.

## Keywords

FFQ  
Multiple imputation  
Missing data  
Item non-response

The FFQ is a dietary assessment tool commonly used in epidemiological studies. However, data collection via FFQ is associated with a serious issue. A long, self-administered, mailed questionnaire provides the opportunity of non-response. Missing values directly influence the estimation of dietary intake from the questionnaire. Several approaches have been used to deal with this problem, including deletion of cases with one or more missing values from the analysis (complete-case analysis) or filling

in the missing data with single values, such as the mean, median or mode (single imputation method). Although zero imputation is a common strategy for handling missing data in FFQ<sup>(1)</sup>, it is doubtful whether the use of this method is acceptable in obtaining unbiased estimates<sup>(2–5)</sup>.

Two studies reported that missing food items did not represent zero intake when the foods were frequently consumed<sup>(6,7)</sup>. In another study, some food items were left blank because of inattention or carelessness<sup>(4)</sup>. Additionally,

\*Corresponding author: Email ssuzuki@med.nagoya-cu.ac.jp

because of single imputation methods, such as zero imputation, where the subject's response and imputed value for non-response are treated in the same way, an underestimation of variance occurs<sup>(8)</sup>. Therefore, we believe that although the use of zero imputation is reasonable in some situations, it may give misleading results.

In the last decade, statistical methods for the analysis of missing data and the importance of not ignoring missing values in the analysis have gained attention among researchers. Rubin classified missing data into three categories: (i) missing at random (MAR); (ii) missing completely at random (MCAR); and (iii) missing not at random (MNAR)<sup>(8–10)</sup>. Data are considered MAR when the probability that the data are missing depends on the observed data but not on the missing data. MCAR is an important, special case of MAR when the probability that the data are missing does not depend on any data. In contrast, data are considered MNAR when the probability that the data are missing depends on the missing values themselves as well as on the observed data. For instance, MAR is when older subjects tend to skip more questions than younger subjects (age is known) and MNAR is when subjects with a higher or lower salary than the average leave questions pertaining to their salary unanswered because of their reluctance to answer. Some researchers have noted that missing data in FFQ are likely to be MNAR because some missing values might represent zero intake<sup>(1,5)</sup>. When considerable missing values due to zero intake are removed from missing data in FFQ, the remnant data might approach MAR. We considered how to predict missing values due to zero intake in the imputation process, because we cannot remove them from data in practice.

In the literature, missing data are indirect MNAR when the observed data can predict the reason for missing data<sup>(8–10)</sup>. For example, when subjects with a higher or lower salary than the average are likely to skip a question because of their salary, the missing data are MNAR. When we obtain correlates of their salary, like job and career, subsequently, indirect MNAR data are treated as MAR because the missing salary becomes predictable. In FFQ, missing values for food items that are infrequently consumed tend to represent zero intake; however, this phenomenon may vary with each food item<sup>(3,5–7)</sup>. Additionally, more information can help to predict missing values in multiple imputation<sup>(11)</sup>. The observed data explain how frequently each food item is consumed or is not. When the missing level is relatively low, the observed data can be used to predict the probability of zero intake to some degree, or can provide support to predict missing values by the distribution and pattern of intake. For these reasons, data may be indirect MNAR.

In the case of MAR or MCAR, multiple imputation is the preferred method to account for missing data and is recommended by several journals and medical research guidelines<sup>(12,13)</sup>. The use of this method provides several

advantages over conventional methods, such as avoiding loss of sample size and considering the uncertainty of imputed values, but only few nutritional epidemiological studies have applied this method for the missing data in FFQ<sup>(14,15)</sup>.

We aimed to examine missing data in FFQ, especially missing values due to zero intake, and to assess the effects on total energy and nutrient estimates. We used data from the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study to verify the missing data mechanism of FFQ and to compare nutrient estimates obtained with the two imputation methods. We did not conduct a resurvey and treated the baseline values as reference.

## Methods

### *Okazaki J-MICC study*

The present study was conducted as a part of the J-MICC study<sup>(16)</sup>. We recruited 7580 men and women (aged 35–79 years) between 2007 and 2011 from the Okazaki Public Health Center, Okazaki, Aichi, Japan. The self-administered health and lifestyle questionnaire, which contained an FFQ, was mailed prior to the health check-up at the centre. At baseline, all participants received the health check-up and submitted the questionnaire. The submitted questionnaires were checked three times by the investigators. When participants did not answer completely (e.g. leaving blanks or inconsistent answers), we queried them as possible. Between 2013 and 2017, we conducted the second survey for the 5-year follow-up. Mean period from baseline to the second survey was 5.3 years. At the second survey, some participants received the health check-up and submitted the questionnaire in the same way, whereas others provided the questionnaire only by mail. In this case, the questionnaires were neither checked nor resurveyed. Of the total participants, 5321 completed both the baseline (FFQ1) and second surveys (FFQ2). Data pertaining to 5151 participants (96.8%) who fully completed FFQ1 were included in the present analysis.

The study protocol was approved by the institutional review board of the Nagoya University Graduate School of Medical Science and Medical School. Written informed consent was obtained from each participant.

### *FFQ*

The Department of Public Health, Nagoya City University Graduate School of Medical Science and Medical School, developed and validated the FFQ<sup>(17–20)</sup>. Participants were asked how often on average over the previous year they consumed forty-seven foods and beverages. The alcohol item was divided into ten subtypes and had six possible responses ranging from 'never/rarely' to 'every day'. Because energy intake is derived mainly from staple foods

in the Japanese dietary style, each staple food item (rice, bread and noodles) was divided into each meal (breakfast, lunch and dinner), and six possible responses similar to those included for alcohol subtypes were also included. For other food and beverage items, there were eight possible responses ranging from 'never/rarely' to '3 or more times per day'; a portion size was specified for each staple food by meal and for the alcohol subtype (e.g. 1 cup or 1 slice of bread). The nutrient intake was calculated from a total of eighty-one questions (62 frequencies = 10 alcohols + 9 staple foods + 43 other foods; 19 portion sizes = 10 alcohols + 9 staple foods) by determining the daily intake frequency ('never/rarely' = 0; '1–3 times per month' = 0.1; '1–2 times per week' = 0.2; '3–4 times per week' = 0.5; '5–6 times per week' = 0.8; 'once per day' = 1.0; 'twice per day' = 2.0; '3 or more times per day' = 3.0).

In the present analysis, we calculated energy intake derived from alcohol only for current alcohol drinkers. For nutrient calculations, participants were excluded if answers to more than forty of the eighty-one questions in FFQ2 were missing or if their total energy intake in each FFQ was below 2510 kJ/d (600 kcal/d) or above 14 644 kJ/d (3500 kcal/d), according to previous studies<sup>(5,7)</sup>. The purpose of these criteria was to exclude, for example, those who are unwilling to answer the FFQ or who over-/under-report their dietary intake.

### Missing values and imputation

Missing alcohol items were permitted if the participants were not current alcohol drinkers; missing portion size was permitted if the participants answered with the frequency 'never/rarely' for the item. These permitted missing values were replaced with zero before imputations.

We filled the missing values in FFQ2 using the following imputation methods.

1. FFQ1 imputation (baseline value): missing values were imputed with the baseline value of the same individual in FFQ1.
2. Zero imputation: missing values were imputed with the frequency 'never/rarely' and portion size '1 unit' (standard size).
3. Multiple imputation: we conducted multiple imputation by chained equations<sup>(21,22)</sup>. Missing values for frequency and portion size in the FFQ were imputed as continuous variables (daily intake frequency and portion size per unit, respectively). Sixty data sets were created with a relative efficiency of >0.99 for each variable. Relative efficiency is an indicator of whether a sufficient number of data sets is created<sup>(23)</sup>. Although the usual practice or recommended method is to create ten to twenty data sets, a large number of data sets is expected to achieve stable estimates and to improve the validity of the significance tests. The imputation model included the following known or

suspected variables associated with missing values or outcomes<sup>(6,7)</sup>: age at the second survey (year, continuous); sex (men, women); alcohol drinking (current, former, never); smoking (current, former, never); questionnaire submission (at the centre, by mail); work (full-time workers, other workers, non-workers); education (elementary school/junior high school, high-school graduate, college or more); BMI (kg/m<sup>2</sup>, continuous) and physical activity at leisure time (MET-h/week, continuous; where MET is metabolic equivalent of task). Missing values for these covariates were also imputed. After nutrient calculations, we combined the results into a single estimate with an SE.

### Statistical analysis

We calculated the mean number of missing values in FFQ2 by category of background variables and evaluated the difference between categories, excluding the missing, by testing with one-way ANOVA. For each food item, the number and proportion of participants who responded with the frequency 'never/rarely' and portion size '1 unit' in FFQ2 are presented. For sensitivity analysis, we calculated the mean total energy and nutrient intakes for complete cases and all of FFQ2. We also calculated mean total energy and nutrient intakes by stratifying according to how the questionnaire was submitted (at the centre or by mail). Differences in total energy and nutrient intakes from zero imputation and multiple imputation compared with FFQ1 imputation were tested by paired *t* tests. We performed all data analyses and imputations using the statistical software package SAS version 9.4. PROC MI and PROC MIANALYZE procedures were used for multiple imputation and combining the results.

### Results

Of the 5151 participants who completely answered FFQ1, 3673 participants (71.3%) responded to all the questions in FFQ2 (Table 1). In FFQ2, the mean number of missing values among the 5151 participants was 3.3 per participant (4.1% of data matrix), and 99.4% of them had forty or fewer missing values in the eighty-one questions after allowing for the permitted missing values. The proportion of participants with one missing value was the largest in that of participants with at least one missing value. The number of participants decreased with increasing number of missing values, but a slight increment of participants with nine to twelve missing values was observed. Thirty-one participants were excluded because their FFQ2 had more than forty missing values.

Table 2 shows the personal characteristics that might be associated with the number of missing values in FFQ2. The mean number of missing values in FFQ2 increased with age. Similarly, a lower education level and a higher

**Table 1** Distribution of missing values in FFQ2 of the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study

No. of missing values	No. of participants	Proportion of participants (%)	Cumulative proportion of participants (%)
0	3673	71.31	71.31
1	196	3.81	75.11
2	178	3.46	78.57
3, 4	132	2.56	81.13
5, 6	85	1.65	82.78
7, 8	95	1.84	84.62
9, 10	140	2.72	87.34
11, 12	176	3.42	90.76
13, 14	75	1.46	92.22
15–17	72	1.40	93.61
18–20	74	1.44	95.05
21–23	48	0.93	95.98
24–26	49	0.95	96.93
27–29	51	0.99	97.92
30–34	66	1.28	99.20
35–39	10	0.19	99.40
40–59	25	0.49	99.88
60–81	6	0.12	100.00
Total	5151	100.00	100.00

physical activity level indicated more questions being missed in FFQ2 than those indicated by higher education level and lower physical activity. In addition, men, current alcohol drinkers, former smokers and non-workers were more likely to have a larger number of missing values in FFQ2. Obviously, participants who submitted the questionnaires by mail left many more blanks in FFQ2 than those who submitted the questionnaires at the centre.

We investigated the proportion of frequency 'never/rarely' in: (i) FFQ2; and (ii) FFQ1 among those who had a missing value in FFQ2 (Table 3). When proportion (i) was over 70%, it was observed that proportion (ii) was more than 80%. Most of the food items had almost the same degree of proportion (i) and (ii), but some food items (e.g. *chu-hi*, bread at lunch, broccoli) had a higher proportion (ii) than proportion (i).

We also investigated the proportion of portion size '1 unit' in: (i) FFQ2; and (ii) FFQ1 among those who had a missing value in FFQ2 (Table 4). We excluded the participants who answered 'never/rarely' for each item from the analysis of Table 4 because the distribution of portion size was divided into zero intake and others. Most of the participants responded with '1 unit' in FFQ2. When proportion (i) was over 70%, proportion (ii) was observed to be more than 60%. Proportions (ii) of food items excluding rice (bowl at dinner) and whiskey (double) were a little lower than proportions (i).

Table 5 shows the estimated dietary intakes of total energy and selected nutrients from FFQ1 and FFQ2. In complete-case analysis, the total energy decreased from 7082 kJ/d (1693 kcal/d) in FFQ1 to 6834 kJ/d (1633 kcal/d) in FFQ2. There were slight changes in the intakes of protein, fat, carbohydrate and dietary fibre. In all participants, zero-imputed FFQ2 presented the smallest total energy, whereas FFQ1 presented the highest total energy. The results of multiple imputation were almost identical to

estimates from FFQ1 imputation. Particularly, there were the smaller differences of total energy and carbohydrate (g) from multiple imputation to FFQ1 imputation than those from zero imputation. No significant differences between multiple imputation and FFQ1 imputation were observed for total energy, protein (% energy), fat (% energy), carbohydrate (g and % energy) and dietary fibre (g/1000 kcal). Less than forty participants were excluded from the total of 5120 participants because their total energy intake was <2510 kJ/d (<600 kcal/d) or >14 644 kJ/d (>3500 kcal/d).

As a sensitivity analysis, participants were stratified by questionnaire submission (see online supplementary material). Obviously, the results of participants who submitted a questionnaire at the centre were exactly similar between the imputation methods, because the proportion of missing values was too small (0.3% of data matrix; Supplemental Table 1). In participants who submitted by mail, where the proportion of missing values was 11.5% of the data matrix, multiple imputation presented smaller differences of total energy and nutrient intake except for alcohol to FFQ1 imputation than zero imputation, although these differences were larger than the results of all participants (Supplemental Table 2).

## Discussion

From the population of the second survey of the Okazaki J-MICC study, we examined the missing data in FFQ and then compared the results of some imputation methods for total energy and nutrient intakes. Missing food items that are not eaten often were likely to represent zero intake, and the number of missing values was associated with personal characteristics. The estimated total energy and nutrient intakes of the multiple imputation method were

**Table 2** Characteristics related to missing values in FFQ2 of the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study

Variable	No. of participants (n 5120)		No. of missing values		P†
	n	%	Mean	SD	
Age (years)					
< 55	1173	22.9	1.3	4.1	< 0.0001
55–64	1179	23.0	1.7	4.9	
65–74	2114	41.3	3.3	7.0	
≥ 75	654	12.8	7.5	9.8	
Sex					
Men	2865	56.0	3.3	7.3	< 0.0001
Women	2255	44.0	2.6	5.9	
Alcohol drinking					
Never	2177	42.5	1.8	4.0	< 0.0001
Former	196	3.8	3.8	6.5	
Current	2702	52.8	3.8	8.1	
Missing	45	0.9	12.4	6.8	
Smoking					
Never	2864	55.9	2.6	6.2	0.002
Former	1614	31.5	3.3	7.3	
Current	574	11.2	2.6	6.2	
Missing	68	1.3	13.1	9.4	
BMI (kg/m <sup>2</sup> )					
< 18.5	291	5.7	2.1	5.3	0.1
18.5–21.9	1733	33.8	2.9	6.6	
22.0–24.9	1948	38.0	3.0	6.8	
≥ 25.0	1110	21.7	3.0	6.8	
Missing	38	0.7	13.2	10.1	
Physical activity at leisure time					
Quartile 1	590	11.5	2.3	5.9	0.0004
Quartile 2	1259	24.6	2.7	6.2	
Quartile 3	1267	24.7	3.0	6.8	
Quartile 4	1264	24.7	3.6	7.2	
Missing	740	14.5	3.1	7.1	
Education					
Elementary school/junior high school	664	13.0	5.0	8.9	< 0.0001
High-school graduate	2170	42.4	3.2	6.8	
Some college or more	2253	44.0	2.1	5.5	
Missing	33	0.6	12.3	13.3	
Work					
Full-time workers	1144	22.3	1.0	3.6	< 0.0001
Other workers	1453	28.4	2.7	6.4	
Non-workers	2223	43.4	3.7	7.3	
Other/missing	300	5.9	6.4	9.7	
Questionnaire submission					
At the centre	3580	69.9	0.3	1.6	< 0.0001
By mail	1540	30.1	9.3	9.3	

†P value when the missing category is excluded.

found to be slightly different from those of the single imputation method (zero or baseline value). In particular, multiple imputation showed almost the same estimates as FFQ1 imputations (imputing baseline values). Our findings suggest that the application of multiple imputation to missing data in FFQ is a reasonable choice when the proportion of missing data is relatively low.

To the best of our knowledge, only two studies have investigated the potential effects of multiple imputation on missing data in FFQ. Barzi *et al.* applied repeated measures of the consumption of five food items in the GISSI-Prevenzione study and suggested that the use of multiple imputation is likely to provide valid estimates<sup>(14)</sup>. Fraser and Yan introduced guided multiple imputation to fill in a

random sub-sample of initially missing data and to adjust the imputation model using this extra information as a guide<sup>(11)</sup>. Compared with other methods (complete-case analysis, zero imputation and multiple imputation), the food intake frequencies showed a moderate difference. Although ascertaining whether the missing data in FFQ were actually MAR was a common problem in these studies, a lower proportion of missing values is likely to have minimized bias.

We investigated the mechanism of missing data in FFQ by comparing with baseline values. Most food items showed that the proportion of zero intake in the observed data was likely to be similar to the probability of zero intake in missing data; whereas for some food items this



**Table 3** Frequency of intake: response of 'never/rarely' in FFQ1 and FFQ2 of the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study

Item	No. of participants who responded with 'never/rarely'							
	No. of missing values in FFQ2 ( <i>n</i> 5120)		(i) In FFQ2†				(ii) In FFQ1 among those who had a missing value in FFQ2‡	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%		
Whiskey (double shot)§	303	5.9	2252	93.9	296	97.7		
Noodles (breakfast)	751	14.7	4065	93.0	710	94.5		
Whiskey (single shot)§	269	5.3	2238	92.0	260	96.7		
Bread (dinner)	782	15.3	3919	90.3	728	93.1		
Beer (250 ml can)§	362	7.1	2073	88.6	339	93.6		
Beer (633 ml bottle)§	363	7.1	2064	88.2	325	89.5		
Beer (500 ml bottle)§	350	6.8	1894	80.5	304	86.9		
Chu-hi§	242	4.7	1934	78.6	222	91.7		
Butter	140	2.7	3722	74.7	118	84.3		
Wine§	231	4.5	1756	71.1	203	87.9		
Liver	53	1.0	3274	64.6	31	58.5		
Sake§	163	3.2	1626	64.0	129	79.1		
Bread (lunch)	601	11.7	2503	55.4	432	71.9		
Margarine	78	1.5	2574	51.1	46	59.0		
Distilled spirit§	144	2.8	1204	47.1	111	77.1		
Noodles (dinner)	550	10.7	2146	47.0	350	63.6		
Fish eggs	30	0.6	2273	44.7	17	56.7		
Beer (350 ml can)§	236	4.6	972	39.4	162	68.6		
Rice (breakfast)	335	6.5	1701	35.5	241	71.9		
Dry radish	36	0.7	1711	33.7	11	30.6		
Tuna	33	0.6	1557	30.6	13	39.4		
Peanut, almond	42	0.8	1418	27.9	15	35.7		
Bread (breakfast)	325	6.3	1272	26.5	223	68.6		
Milk	53	1.0	1304	25.7	24	45.3		
Western-style confectioneries	38	0.7	1177	23.2	15	39.5		
Noodles (lunch)	311	6.1	1072	22.3	144	46.3		
Yogurt	45	0.9	1029	20.3	21	46.7		
Shellfish	27	0.5	923	18.1	3	11.1		
Japanese-style confectioneries	29	0.6	767	15.1	6	20.7		
Natto, soyabean	39	0.8	593	11.7	10	25.6		
Bone-edible small fish	12	0.2	587	11.5	2	16.7		
Rice (lunch)	238	4.6	542	11.1	93	39.1		
Ganmodoki, nama-age	19	0.4	559	11.0	3	15.8		
Ham, sausage, bacon	33	0.6	553	10.9	8	24.2		
Fish paste products	19	0.4	552	10.8	1	5.3		
Burdock, bamboo shoot	12	0.2	540	10.6	1	8.3		
Pumpkin	13	0.3	532	10.4	2	15.4		
Tofu	69	1.3	518	10.3	8	11.6		
Mayonnaise	6	0.1	506	9.9	1	16.7		
Cuttlefish, squid, octopus, shrimp, crab	18	0.4	464	9.1	2	11.1		
Citrus fruits	35	0.7	454	8.9	4	11.4		
Broccoli	22	0.4	406	8.0	4	18.2		
Rice (dinner)	65	1.3	390	7.7	15	23.1		
Coffee	20	0.4	355	7.0	4	20.0		
Green tea	25	0.5	345	6.8	2	8.0		
Chicken	39	0.8	329	6.5	6	15.4		
Other fruits	22	0.4	289	5.7	0	0.0		
Miso soup	24	0.5	201	3.9	0	0.0		
Seaweeds	18	0.4	170	3.3	1	5.6		
Beef, pork	10	0.2	145	2.8	2	20.0		
Carrot	10	0.2	130	2.5	0	0.0		
Deep-fried food	29	0.6	119	2.3	3	10.3		
Mushrooms	14	0.3	115	2.3	0	0.0		
Other green-yellow vegetables	17	0.3	109	2.1	3	17.6		
Egg	15	0.3	101	2.0	1	6.7		
Potatoes	6	0.1	90	1.8	1	16.7		
Japanese radish	24	0.5	87	1.7	0	0.0		
Green leafy vegetables	11	0.2	80	1.6	1	9.1		
Light fried food	48	0.9	44	0.9	1	2.1		
Fish	19	0.4	44	0.9	0	0.0		
Cabbage	7	0.1	38	0.7	1	14.3		
Other vegetables	4	0.1	16	0.3	0	0.0		

†Proportions when missing values in FFQ2 are excluded for each item.

‡Proportions to the number of missing values in FFQ2.

§Alcoholic drinks are only for current alcohol drinkers at the second survey (*n* 2702).

**Table 4** Portion size of intake: response of '1 unit' in FFQ1 and FFQ2 of the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study

Item (unit)	No. of missing values in FFQ2 ( <i>n</i> 5120)		No. of participants who responded with '1 unit'			
			(i) In FFQ2†		(ii) In FFQ1 among those who had a missing value in FFQ2‡	
	<i>n</i>	%	<i>n</i>	%	<i>n</i>	%
Noodles (bowl at lunch)	313	6.1	3532	98.4	270	86.3
Noodles (bowl at dinner)	310	6.1	2262	97.8	260	83.9
Noodles (bowl at breakfast)	70	1.4	262	95.3	42	60.0
Rice (bowl at breakfast)	138	2.7	2847	93.7	121	87.7
Rice (bowl at lunch)	233	4.6	3952	92.9	212	91.0
Rice (bowl at dinner)	155	3.0	4166	91.4	142	91.6
Beer (250 ml can)§	40	0.8	201	80.4	25	62.5
Beer (350 ml can)§	115	2.2	1150	79.1	80	69.6
Bread (slice or roll at breakfast)	134	2.6	2671	76.5	83	61.9
Beer (500 ml bottle)§	65	1.3	324	73.8	40	61.5
Beer (633 ml bottle)§	53	1.0	188	72.3	33	62.3
Bread (slice or roll at dinner)	80	1.6	273	69.5	37	46.3
Chu-hi (cup)§	40	0.8	335	66.2	20	50.0
Bread (slice or roll at lunch)	242	4.7	1221	62.8	139	57.4
Whiskey (single shot)§	25	0.5	112	62.6	7	28.0
Wine (glass)§	72	1.4	355	52.9	32	44.4
Sake (cup)§	53	1.0	468	52.3	13	24.5
Distilled spirit (cup)§	69	1.3	612	46.4	22	31.9
Whiskey (double shot)§	9	0.2	58	40.0	6	66.7

†Excluded when the frequency of each item is 'never/rarely'. Proportions when missing values in FFQ2 are excluded for each item.

‡Proportions to the number of missing values in FFQ2.

§Alcoholic drinks are only for current alcohol drinkers at the second survey (*n* 2702).

did not hold. The former indicates that missing values due to zero intake can be predicted from the observed distribution of how rarely a particular food is consumed in the population, and the data can be treated as MAR and not MNAR. The latter seems to indicate that there were other factors that influenced the missingness than just a zero intake; that is, missing data in FFQ may be MNAR. These findings are consistent with the results of previous studies that showed that each food item that is not eaten often is likely to represent zero<sup>(5-7)</sup>. For this reason, it would be reasonable to impute zero at least for rarely eaten foods. However, zero imputation can introduce an element of bias because it is impossible that all missing values are zero.

Additionally, the estimates of total energy and nutrient intakes determined with multiple imputation were equivalent to those determined with FFQ1 imputation, which suggests that multiple imputation can predict the missing values due to zero intake. Even if we determine the 'true' value for the missing food items, it might lie between the values determined with zero imputation and FFQ1 imputation. In the present study, the proportion of missing values in our data was relatively low, and a comparison of imputation methods might be much more difficult because of little freedom for any moderate difference to be displayed. To deal with this problem, we carried out stratified analysis and presented that multiple imputation had the smaller differences to FFQ1 values in total energy and nutrient intakes, except for alcohol, than zero imputation. The application of multiple imputation

might provide a more accurate estimate when the probability of missing data is less than 11.5%. For alcohol, several possible reasons exist. First, it seemed to be more difficult to predict and impute missing values than for the other items because the alcohol item has detailed questions. Additional covariates for more accurate imputation may be needed. Second, missing alcohol values were more likely to represent zero intake than other food items.

Furthermore, imputing a single value ignores the uncertainty of imputed values and can lead to biased results because of underestimating the variance<sup>(8)</sup>. Regarding the current trend in medical research, it is important to appropriately handle missing data and to avoid misleading results. Our findings suggest that multiple imputation may be employed for handling missing data in future studies involving FFQ.

The present study has several limitations. First, we treated baseline values in FFQ1 as a reference and did not perform a resurvey. Some studies have shown that total energy and food intakes decrease with age<sup>(24,25)</sup>; further, our results of complete-case analysis showed a decrease in total energy and nutrient intakes over a mean follow-up period of 5.3 years. To improve the accuracy of multiple imputation, the investigation of additional variables in the imputation model or of item-level and score-level multiple imputation<sup>(26,27)</sup> may be needed. However, because of the slight difference in nutrient estimates among imputation methods, this was not a problem in the present study.

Second, we used the complete-case data in FFQ1. Complete-case analysis may lead to biased results unless

**Table 5** Estimated dietary intakes of total energy and selected nutrients from FFQ1 and FFQ2 of the Okazaki Japan Multi-Institutional Collaborative Cohort (J-MICC) study

Nutrient	Complete-case analysis of FFQ2 ( <i>n</i> 3673)						All of FFQ2 ( <i>n</i> 5120)						Difference (B) – (A) ( <i>n</i> 5081, <i>n</i> <sub>alc</sub> 2547)	Difference (C) – (A) ( <i>n</i> 5085, <i>n</i> <sub>alc</sub> 2552)
	FFQ1 ( <i>n</i> 3656, <i>n</i> <sub>alc</sub> 1880)		FFQ2 ( <i>n</i> 3658, <i>n</i> <sub>alc</sub> 1823)		FFQ1 ( <i>n</i> 5098, <i>n</i> <sub>alc</sub> 2664)		(A) FFQ2 of imputing baseline values ( <i>n</i> 5087, <i>n</i> <sub>alc</sub> 2564)		(B) FFQ2 of zero imputation ( <i>n</i> 5084, <i>n</i> <sub>alc</sub> 2572)		(C) FFQ2 of multiple imputation ( <i>n</i> 5085, <i>n</i> <sub>alc</sub> 2576)			
	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE	Mean	SE		
Total energy														
kJ	7082	23.5	6834	22.6	7092	19.9	6890	19.6	6801	19.3	6891	19.8	-91.29***	2.61
kcal	1693	5.6	1633	5.4	1695	4.7	1647	4.7	1626	4.6	1647	4.7	-21.82***	0.62
Protein														
g	53.3	0.17	52.7	0.18	53.4	0.15	53.2	0.16	52.7	0.15	53.3	0.16	-0.54***	0.09*
% energy	12.7	0.03	13.0	0.03	12.7	0.03	13.1	0.03	13.1	0.03	13.1	0.03	-0.05***	0.01
Fat														
g	43.6	0.18	45.3	0.19	43.5	0.15	45.4	0.16	45.2	0.16	45.5	0.16	-0.22***	0.09**
% energy	23.8	0.10	25.6	0.11	23.6	0.08	25.5	0.10	25.7	0.10	25.5	0.10	0.24***	0.02
Carbohydrate														
g	251	1.1	237	1.0	251	0.9	240	0.9	236	0.9	240	0.9	-4.12***	-0.02
% energy	58.7	0.09	57.4	0.09	58.8	0.07	57.6	0.08	57.4	0.08	57.6	0.08	-0.25***	-0.01
Dietary fibre														
g	10.8	0.05	11.0	0.05	10.9	0.04	11.2	0.05	11.1	0.05	11.2	0.05	-0.07***	0.03***
g/4184 kJ (1000 kcal)	6.5	0.03	6.9	0.04	6.6	0.03	7.0	0.03	7.0	0.03	7.0	0.03	0.07***	0.01
Alcohol														
g	15	0.4	15	0.4	15	0.3	16	0.3	15	0.3	16	0.4	-0.44***	0.46**
g/4184 kJ (1000 kcal)	9	0.3	10	0.3	9	0.2	10	0.2	10	0.2	10	0.3	-0.15***	0.27**

*n* indicates the number of participants who had forty or fewer missing values and total energy intake between 2510 kJ/d (600 kcal/d) and 14 644 kJ/d (3500 kcal/d). *n*<sub>alc</sub> indicates the number of participants who consumed >0 g alcohol/d in *n* participants. Energy from alcohol was not included in the total energy.

\**P* < 0.05, \*\**P* < 0.01, \*\*\**P* < 0.001 for paired *t* test between imputation methods.



the mechanism of missing data is MCAR<sup>(8)</sup>. However, all questionnaires of the baseline survey were reviewed and the percentage of our participants who completely answered FFQ1 was 96.8%. Thus, the missing data of FFQ1 were considered to be MCAR and had no impact on our results.

Third, the questionnaires used in the second survey were submitted by two different approaches. Approximately 70% of the questionnaires were reviewed at the time of health check-up of the participants, whereas approximately 30% were submitted by mail and did not receive a review. Consequently, a higher proportion of our participants completely answered FFQ2 compared with that in some previous studies (34–41%)<sup>(7,28)</sup>, which would contribute to a more plausible imputation rather than any undesirable results from multiple imputation<sup>(29)</sup>.

## Conclusion

Zero imputation was a reasonable choice for a limited number of food items only when the missing values were due to 'zero intake'. To appropriately handle missing data in FFQ, we recommended multiple imputation, taking account of the uncertainty of imputed values. The present study confirmed that multiple imputation would be applicable to missing data in FFQ when missing level is relatively low. Particularly, if missing level exceeds about 10% of data, one should carefully consider the effect of the imputation method on epidemiological risk analysis with a disease outcome, at least for dietary intake.

## Supplementary material

To view supplementary material for this article, please visit <https://doi.org/10.1017/S1368980019000168>

## Acknowledgements

*Acknowledgements:* The authors would like to thank all the staff and participants for supporting this study. *Financial support:* This study was supported in part by the Grants-in-Aid for Scientific Research on Priority Areas of Cancer (No. 17015018), on Innovative Areas (No. 221S0001), and JSPS KAKENHI Grant Number JP (No. 16H06277) from the Japanese Ministry of Education, Culture, Sports, Science and Technology. The Japanese Ministry of Education, Culture, Sports, Science and Technology had no role in the design, analysis or writing of this article. *Conflict of interest:* None declared. *Authorship:* S.S. was responsible for the design and funding of the present study. T.Y. is the director of the Public Health Center of the Okazaki City Medical Association. N.I. and C.G. developed and provided the

calculation program for the FFQ. A.H. took the lead role in collecting data. K.S., S.T., M.I., Y.T., M.W., K.O., H.F., N.O., M.K., F.K., R.W. and T.N. also contributed to data collection. M.I. performed statistical analysis. S.S. and T.I. advised on drafting the manuscript. All authors have checked and approved the final version of the manuscript submitted and any revised version. *Ethics of human subject participation:* The study protocol was approved by the institutional review board of the Nagoya University Graduate School of Medical Science and Medical School. Written informed consent was obtained from each participant.

## References

1. Lamb KE, Olstad DL, Nguyen C *et al.* (2017) Missing data in FFQs: making assumptions about item non-response. *Public Health Nutr* **20**, 965–970.
2. Ahn Y, Paik HY & Ahn YO (2006) Item non-responses in mailed food frequency questionnaires in a Korean male cancer cohort study. *Asia Pac J Clin Nutr* **15**, 170–177.
3. Hansson LM & Galanti MR (2000) Diet-associated risks of disease and self-reported food consumption: how shall we treat partial nonresponse in a food frequency questionnaire? *Nutr Cancer* **36**, 1–6.
4. Kuskowska-Wolk A, Holte S, Ohlander EM *et al.* (1992) Effects of different designs and extension of a food frequency questionnaire on response rate, completeness of data and food frequency responses. *Int J Epidemiol* **21**, 1144–1150.
5. Parr CL, Hjartaker A, Scheel I *et al.* (2008) 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). *Public Health Nutr* **11**, 361–370.
6. Fraser GE, Yan R, Butler TL *et al.* (2009) Missing data in a long food frequency questionnaire: are imputed zeroes correct? *Epidemiology* **20**, 289–294.
7. Michels KB & Willett WC (2009) Self-administered semi-quantitative food frequency questionnaires: patterns, predictors, and interpretation of omitted items. *Epidemiology* **20**, 295–301.
8. Schafer JL & Graham JW (2002) Missing data: our view of the state of the art. *Psychol Methods* **7**, 147–177.
9. Enders CK (2010) *Applied Missing Data Analysis*. New York: Guilford Press.
10. Rubin DB (1996) Multiple imputation after 18+ years. *J Am Stat Assoc* **91**, 473–489.
11. Fraser G & Yan R (2007) Guided multiple imputation of missing data: using a subsample to strengthen the missing-at-random assumption. *Epidemiology* **18**, 246–252.
12. Klebanoff MA & Cole SR (2008) Use of multiple imputation in the epidemiologic literature. *Am J Epidemiol* **168**, 355–357.
13. Ware JH, Harrington D, Hunter DJ *et al.* (2012) Missing data. *N Engl J Med* **367**, 1353–1354.
14. Barzi F, Woodward M, Marfisi RM *et al.* (2006) Analysis of the benefits of a Mediterranean diet in the GISSI-Prevenzione study: a case study in imputation of missing values from repeated measurements. *Eur J Epidemiol* **21**, 15–24.
15. Rizzo NS, Sabate J, Jaceldo-Siegl K *et al.* (2011) Vegetarian dietary patterns are associated with a lower risk of

- metabolic syndrome: the Adventist Health Study 2. *Diabetes Care* **34**, 1225–1227.
16. Hamajima N & J-MICC Study Group (2007) The Japan Multi-Institutional Collaborative Cohort Study (J-MICC Study) to detect gene–environment interactions for cancer. *Asian Pac J Cancer Prev* **8**, 317–323.
  17. Tokudome S, Goto C, Imaeda N *et al.* (2004) Development of a data-based short food frequency questionnaire for assessing nutrient intake by middle-aged Japanese. *Asian Pac J Cancer Prev* **5**, 40–43.
  18. Tokudome Y, Goto C, Imaeda N *et al.* (2005) Relative validity of a short food frequency questionnaire for assessing nutrient intake versus three-day weighed diet records in middle-aged Japanese. *J Epidemiol* **15**, 135–145.
  19. Goto C, Tokudome Y, Imaeda N *et al.* (2006) Validation study of fatty acid consumption assessed with a short food frequency questionnaire against plasma concentration in middle-aged Japanese people. *Scand J Food Nutr* **50**, 77–82.
  20. Imaeda N, Goto C, Tokudome Y *et al.* (2007) Reproducibility of a short food frequency questionnaire for Japanese general population. *J Epidemiol* **17**, 100–107.
  21. van Buuren S (2007) Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res* **16**, 219–242.
  22. Van Buuren S & Oudshoorn C (2000) *Multivariate Imputation by Chained Equations: MICE V1.0 User's Manual*. TNO Report no. PG/VGZ/00.038. Leiden: TNO Prevention and Health.
  23. Enders CK (2010) *Applied Missing Data Analysis*, pp. 212–213. New York: Guilford Press.
  24. Otsuka R, Kato Y, Nishita Y *et al.* (2016) Age-related changes in energy intake and weight in community-dwelling middle-aged and elderly Japanese. *J Nutr Health Aging* **20**, 383–390.
  25. Wakimoto P & Block G (2001) Dietary intake, dietary patterns, and changes with age: an epidemiological perspective. *J Gerontol A Biol Sci Med Sci* **56**, Spec. No. 2, 65–80.
  26. Gottschall AC, West SG & Enders CK (2012) A Comparison of item-level and scale-level multiple imputation for questionnaire batteries. *Multivariate Behav Res* **47**, 1–25.
  27. Plumpton CO, Morris T, Hughes DA *et al.* (2016) Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Res Notes* **9**, 45.
  28. Caan B, Hiatt RA & Owen AM (1991) Mailed dietary surveys: response rates, error rates, and the effect of omitted food items on nutrient values. *Epidemiology* **2**, 430–436.
  29. White IR, Royston P & Wood AM (2011) Multiple imputation using chained equations: issues and guidance for practice. *Stat Med* **30**, 377–399.