

Statistical approaches for assessing the relative validity of a food-frequency questionnaire: use of correlation coefficients and the kappa statistic

LF Masson¹, G McNeill^{2,*}, JO Tomany², JA Simpson³, HS Peace², L Wei², DA Grubb⁴ and C Bolton-Smith⁵

¹Department of Public Health, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK:

²Department of Medicine & Therapeutics, University of Aberdeen, Polwarth Building, Foresterhill, Aberdeen AB25 2ZD, UK: ³Department of General Practice & Primary Care, University of Aberdeen, Foresterhill Health Centre, Westburn Road, Aberdeen AB25 2AY, UK: ⁴Rowett Research Institute, Bucksburn, Aberdeen AB21 9SB, UK: ⁵Medical Research Council Human Nutrition Research, Elsie Widdowson Laboratory, Fulbourn Road, Cambridge CB1 9NL, UK

Submitted 4 December 2001: Accepted 10 October 2002

Abstract

Objective: To compare different statistical methods for assessing the relative validity of a self-administered, 150-item, semi-quantitative food-frequency questionnaire (FFQ) with 4-day weighed diet records (WR).

Design: Subjects completed the Scottish Collaborative Group FFQ and carried out a 4-day WR. Relative agreement between the FFQ and WR for energy-adjusted nutrient intakes was assessed by Pearson and Spearman rank correlation coefficients, the percentages of subjects classified into the same and opposite thirds of intake, and Cohen's weighted kappa.

Subjects: Forty-one men, mean age 36 (range 21–56) years, and 40 women, mean age 33 (range 19–58) years, recruited from different locations in Aberdeen, Scotland.

Results: Spearman correlation coefficients tended to be lower than Pearson correlation coefficients, and were above 0.5 for 10 of the 27 nutrients in men and 17 of the 27 nutrients in women. For nutrients with Spearman correlation coefficients above 0.5, the percentage of subjects correctly classified into thirds ranged from 39 to 78%, and weighted kappa values ranged from 0.23 to 0.66.

Conclusions: Both Spearman correlation coefficients and weighted kappa values are useful in assessing the relative validity of estimates of nutrient intake by FFQs. Spearman correlation coefficients above 0.5, more than 50% of subjects correctly classified and less than 10% of subjects grossly misclassified into thirds, and weighted kappa values above 0.4 are recommended for nutrients of interest in epidemiological studies.

Keywords

Correlation coefficients
Kappa statistic
Food-frequency questionnaire
Relative validity
Weighed diet records
Epidemiological methods
Dietary assessment

The semi-quantitative food-frequency questionnaire (FFQ) is the primary dietary assessment method used in epidemiological studies¹. Results from such studies can be interpreted with greater confidence if the questionnaire has a quantified validity; i.e. the ability of the questionnaire to measure what it is intended to measure. To assess the true validity of an FFQ would require measuring with high accuracy the usual self-selected diet of free-living individuals over several months, which is not feasible. Therefore, researchers assess *relative* validity by comparing the FFQ with an alternative dietary assessment method with its own limitations.

In epidemiological studies, the odds ratio or relative risk of disease in relation to nutrient intake is the most common measure of association presented. Consequently,

FFQs must be able to rank individuals along the distribution of intake, so that individuals with low intakes can be separated from those with high intakes. Therefore, obtaining absolute nutrient intakes is not necessary. As long as FFQs can rank individuals, relative risk estimates will be accurate².

The usual method of assessing agreement in ranking between an FFQ and a reference method is to calculate Pearson correlation coefficients for log_e-transformed data or Spearman rank correlation coefficients for data that are not normally distributed. It is also common to classify subjects into categories (usually thirds, fourths or fifths) of intake by the FFQ and the reference method, and to calculate the percentage of subjects correctly classified into the same category and grossly misclassified into the

*Corresponding author: Email g.mcneill@abdn.ac.uk

opposite category. However, there is no consensus in the literature on the best statistical method for assessing the validity of dietary assessment tools. The correlation coefficient as a measure of absolute agreement was criticised by Bland and Altman³, but as FFQs are designed to rank individuals rather than to assess their absolute level of intake this objection does not apply. Herbert and Miller argued that one should assume that intakes obtained by different methods are related, and that the test of significance of the correlation coefficient should be the departure from perfect agreement rather than from no agreement⁴. Burema *et al.*⁵ argued that the correlation coefficient is preferable to cross-classification since the latter requires more than one value and can be inferred from the correlation coefficient if a bivariate normal distribution of intake is assumed.

With cross-classification, the percentages misclassified clearly illustrate the likely impact of measurement error; however, the percentage of agreement will include agreement that can be accounted for by chance. Cohen's weighted kappa statistic is a summary measure of cross-classification that allows for the agreement expected by chance and has the added advantage over the kappa statistic in that it allows for the degree of misclassification⁶. However, weighted kappa is still dependent on the number of categories used. Due to the lack of agreement on the best way of presenting results from validation studies^{5,7}, it is necessary to use more than one statistical method in order to give credence to the results⁸.

This study aimed to assess the relative validity of the Scottish Collaborative Group (SCG) FFQ, which was developed to provide a general tool for dietary assessment in the UK that would be applicable across a wide age range and be able to estimate total energy, macro- and micronutrient intakes. The SCG FFQ was compared with a 4-day weighed diet record (WR) and four statistical methods were used to assess the agreement in ranking: (1) the Pearson correlation coefficient, (2) the Spearman rank correlation coefficient, (3) the percentage of subjects classified into same and opposite thirds of intake and (4) Cohen's weighted kappa. The results from these four methods are presented simultaneously in order to illustrate the merits and limitations of each method, and the usefulness of applying a combination of statistical methods when assessing relative validity.

Subjects and methods

Subjects

Subjects were recruited from different locations in Aberdeen, and included men who had recently participated in a population-based study of coronary heart disease risk factors, women attending a Family Planning Clinic, and men and women recruited from a variety of public sector workplaces (police, fire and ambulance stations, regional council, hospital and university).

Forty-three men and 52 women (aged 19–58 years) agreed to take part, and 41 men, mean age 36 (standard deviation (SD) 9.8) years and 40 women, mean age 33 (SD 9.6) years completed the study. The mean body mass index was 26.1 kg m⁻² (range 18.9–34.2 kg m⁻²) in men and 25.1 kg m⁻² (range 17.7–40.3 kg m⁻²) in women. The study was approved by the Research Ethics Committee of Grampian Health Board and the University of Aberdeen, and all subjects gave written informed consent before taking part.

Food-frequency questionnaire

The SCG FFQ was developed from the FFQ used in the Scottish Heart Health Study and Monitoring Trends and Determinants in Cardiovascular Disease (MONICA) study⁹, which in turn was loosely based on the Medical Research Council Caerphilly questionnaire designed for use in studies of ischaemic heart disease¹⁰. The SCG FFQ has since been developed and employed in a range of epidemiological studies.

All subjects completed version 6.31 of the 150-item, semi-quantitative, SCG FFQ. Subjects were asked to specify the frequency of consumption of each food in the last 2–3 months as 'rarely or never', 'once or twice a month', or 1, 2, 3, 4, 5, 6 or 7 days per week, and to give information on the amount consumed by specifying how many measures per day they consumed of each food. A 'measure' was designed to be a small portion so that a single standard portion of a food would often be 2 measures. A colour photograph depicting examples of food measures accompanied the FFQ. The subjects were given the FFQ to complete at home, and all entries were checked with the subject around a week later, after the WR had been completed.

The FFQ data were scanned and verified using Teleform version 7 (Cardiff Software, Vista, CA, USA) and were processed using the Oracle Relational Database Management System (version 7). Average daily nutrient intakes were calculated using the UK National Nutrient Databank, which is based on *McCance & Widdowson's The Composition of Foods* (fifth edition)¹¹ and related supplements^{12–20}. The Royal Society of Chemistry/Crown copyright material from the Nutrient Databank and *McCance & Widdowson's The Composition of Foods* is used with the permission of The Royal Society of Chemistry and the Controller of Her Majesty's Stationery Office and with the agreement of the Minister of Agriculture, Fisheries and Food.

Weighed diet record

Subjects were asked to record the description and weight of all the food and drink they consumed over a period of 4 days, including three weekdays and one weekend day, starting several days (range –9 to +9 days) before or after completing the FFQ. Subjects were provided with a set of scales accurate to ±2 g and a food diary with which to

record their food descriptions and weights. Half way through the weighed record, the subjects were contacted by telephone to ensure that they were having no problems recording their food intake. Each entry in the food diary was checked with the subject to ensure that all foods were sufficiently described. Each food was given a code from the food tables^{11–20} and these codes and the weights consumed were entered into the Diet5 program (Univation Ltd, Aberdeen, UK), which calculated the average daily nutrient intakes.

Analysis

Nutrient intakes from dietary supplements were not included in any of the analyses. The Wilcoxon signed rank test was performed to test if the daily nutrient intakes were statistically significantly different between the methods.

Relative agreement between methods was assessed using correlation coefficients and cross-classification. Before calculating the Pearson correlation coefficient, the distribution of each nutrient intake was checked for normality. The nutrient intakes with a positively skewed distribution were \log_e -transformed. Alcohol intakes of 0 g were replaced with values of 0.1 g before \log_e transformation. Energy-adjusted nutrient intakes were calculated as the residuals from the regression of nutrient intake as the dependent variable on energy as the independent variable²¹.

\log_e transformation did not normalise all nutrient intake distributions. Skewness values greater than 1, indicating distributions that differ significantly from normal distributions, were found for intakes of alcohol, retinol, thiamin and iron in both sexes, for fat, saturated fatty acids (SFA), monounsaturated fatty acids (MUFA), folate, vitamin E, vitamin B₆ and niacin equivalents in women, and for energy and vitamin B₁₂ in men. Since the assumption of a normal distribution does not hold true for these nutrients, the Pearson correlation may give misleading results. Therefore, the non-parametric Spearman rank correlation coefficient was also calculated for energy-adjusted nutrient intakes. Ninety-five per cent confidence intervals (95% CI) for the correlation coefficients were calculated.

The ability of the FFQ to categorise subjects into equal thirds of energy-adjusted nutrient intake was assessed by calculating the percentages of subjects categorised into the same third and the extreme opposite third of intake by the FFQ and WR. The weighted kappa statistic (K_w)²² was calculated for each nutrient from the observed and expected proportions on the 3 × 3 table of frequencies using the formula:

$$K_w = \frac{Po_{(w)} - Pe_{(w)}}{1 - Pe_{(w)}}$$

where Po is the observed proportion of agreement and Pe is the expected proportion of agreement by chance. The weightings used were 1 for complete agreement – i.e. subjects classified into the same third

by the FFQ and WR; 0.5 for partial disagreement – i.e. subjects differing by one category; and 0 for complete disagreement – i.e. subjects differing by two categories. Values of kappa over 0.80 indicate very good agreement, between 0.61 and 0.80 good agreement, 0.41–0.60 moderate agreement, 0.21–0.40 fair agreement and <0.20 poor agreement²².

Ratios of energy intake to estimated basal metabolic rate (EI/BMR)²³ were calculated and compared with the cut-offs of 1.14 for the FFQ and 1.06 for the WR, based on 95% CI²⁴, to identify low energy reporting and therefore WRs that could not be representative of usual intake.

All analyses were carried out with the Statistical Package for the Social Sciences (SPSS) version 9.0 (SPSS, Inc., Chicago, IL, USA).

Results

For the 41 men and 40 women who completed the study, the mean ratios of EI/BMR by the FFQ were 1.56 (range 0.54–2.52) and 1.70 (range 0.98–2.84), respectively. Mean ratios by the WR were 1.32 (range 0.66–2.26) in men and 1.32 (range 0.73–2.09) in women. Twelve men and 10 women had EI/BMR values below the cut-offs, indicating low energy recording: of these, three men and three women reported being on a weight-reducing diet.

Absolute agreement

Table 1 shows the daily nutrient intakes from the FFQ and WR, and the relative differences in intakes between the two methods for all subjects. Intakes were significantly higher ($P < 0.05$) with the FFQ than with the WR for most nutrients, although the two methods gave similar intakes for fat (% of energy) and polyunsaturated fatty acids (PUFA) (% of energy) in both sexes, and for PUFA (g) in men. Alcohol had a median relative difference between the FFQ and WR of zero in men, and a negative difference in women. However, alcohol had wide interquartile ranges of differences, with large differences in median intakes between the methods. The largest median relative differences between the methods were for vitamin B₁₂, retinol and sugars in men, and for calcium, zinc, non-starch polysaccharides (NSP) and retinol in women.

Relative agreement

Tables 2 and 3 show, for men and women respectively, the Pearson and Spearman correlation coefficients calculated for each energy-adjusted nutrient intake. Pearson correlation coefficients tended to be higher than Spearman correlation coefficients, with Pearson correlation coefficients above 0.5 for 18 nutrients in men and 23 nutrients in women, whereas Spearman correlation coefficients were above 0.5 for 10 nutrients in men and 17 nutrients in women. Both Pearson and Spearman correlation coefficients were above 0.5 for SFA, NSP, alcohol, riboflavin, folate, iron, magnesium, potassium

Table 1 Median (interquartile range) crude daily nutrient intakes from the FFQ and WR, and relative differences between the FFQ and WR

Nutrient	Men (n = 41)				Women (n = 40)					
	FFQ		WR		FFQ		WR		Relative difference (%)	
	Median	(P ₂₅ , P ₇₅)	Median	(P ₂₅ , P ₇₅)	Median	(P ₂₅ , P ₇₅)	Median	(P ₂₅ , P ₇₅)	Median	(P ₂₅ , P ₇₅)
Energy (kJ)	12056	(10019, 13455)	9919**	(8664, 11708)	9465	(7480, 12670)	8167***	(6537, 8896)	25	(0, 51)
Fat (g)	104.9	(84.4, 126.5)	85.5**	(68.1, 104.8)	78.0	(57.6, 113.6)	66.2**	(53.5, 78.8)	25	(-3, 68)
Fat (% of energy)	33.3	(28.6, 36.2)	31.7	(27.6, 35.4)	32.4	(26.4, 35.9)	31.6	(26.9, 35.1)	0	(-11, 10)
SFA (g)	42.0	(31.7, 50.8)	30.1***	(23.0, 36.8)	31.7	(22.0, 46.0)	23.6***	(16.5, 29.7)	32	(-1, 75)
SFA (% of energy)	12.5	(10.8, 14.5)	10.6***	(9.8, 12.4)	12.3	(10.1, 14.2)	10.9*	(9.4, 13.4)	6	(-7, 22)
PUFA (g)	17.0	(13.7, 21.7)	17.0	(11.9, 20.8)	14.2	(9.5, 20.2)	11.8**	(7.8, 14.6)	22	(-5, 83)
PUFA (% of energy)	5.5	(4.6, 6.4)	5.7	(4.8, 6.8)	5.4	(4.4, 6.6)	5.3	(4.2, 6.7)	3	(-15, 29)
MUFA (g)	37.8	(30.2, 44.2)	28.6***	(23.9, 34.6)	28.1	(20.9, 39.4)	20.9***	(18.4, 26.2)	27	(4, 71)
MUFA (% of energy)	11.9	(10.3, 13.0)	10.5**	(9.6, 11.5)	11.7	(9.2, 12.9)	10.5**	(8.5, 12.2)	5	(0, 23)
Cholesterol (mg)	362	(288, 449)	234***	(207-324)	266	(171, 379)	176***	(143, 240)	27	(-2, 121)
Protein (g)	115.2	(92.4, 138.9)	94.1***	(76.4, 111.0)	94.8	(68.0, 125.5)	69.0***	(56.7, 78.6)	37	(3, 75)
Sugars (g)	169.7	(144.7, 211.7)	116.7***	(89.4, 149.7)	135.6	(92.6, 206.9)	92.0***	(66.1, 117.9)	42	(13, 101)
Starch (g)	177.3	(156.9, 212.2)	168.3*	(119.3, 197.2)	163.1	(123.1, 193.1)	125.7***	(106.1, 143.0)	31	(5, 58)
NSP (g)	21.0	(16.4, 24.4)	15.1***	(12.3, 18.6)	20.1	(15.3, 28.4)	13.2***	(10.0, 21.3)	53	(16, 109)
Alcohol (g)	15.3	(3.9, 25.2)	10.3	(2.9, 30.8)	6.2	(3.9, 19.5)	11.6**	(0.0, 28.8)	-21	(-49, 112)
Retinol (µg)	539	(346, 753)	302***	(180, 398)	391	(200, 562)	268**	(181, 378)	53	(-3, 131)
β-Carotene equivalents (µg)	1913	(1438, 3094)	1770	(782, 2818)	2798	(1509, 3752)	1525**	(1019, 2510)	32	(-22, 122)
Vitamin D (µg)	4.08	(2.67, 5.45)	3.13*	(1.58, 4.41)	2.63	(2.02, 4.29)	1.83*	(1.21, 2.60)	39	(-5, 86)
Vitamin E (mg)	9.55	(6.34, 13.76)	7.59*	(5.44, 10.38)	7.57	(5.50, 11.85)	5.83***	(5.07, 7.45)	46	(3, 88)
Thiamin (mg)	2.25	(1.83, 2.74)	1.67***	(1.40, 2.21)	2.08	(1.50, 2.70)	1.38***	(1.06, 1.91)	39	(6, 78)
Riboflavin (mg)	2.74	(2.21, 3.25)	1.95***	(1.45, 2.54)	1.97	(1.38, 2.91)	1.36***	(1.06, 2.00)	36	(1, 70)
Vitamin B ₆ (mg)	3.01	(2.55, 3.50)	2.42**	(1.83, 3.25)	2.57	(1.88, 3.34)	1.99***	(1.57, 2.34)	30	(1, 70)
Vitamin B ₁₂ (µg)	7.42	(4.87, 9.43)	4.54***	(3.28, 5.85)	4.53	(3.11, 7.22)	3.38*	(2.37, 5.30)	34	(-22, 125)
Vitamin C (mg)	112.0	(82.4, 129.9)	65.4**	(37.7, 107.8)	112.4	(85.8, 187.9)	90.3***	(63.1, 145.2)	35	(-11, 129)
Folate (µg)	353	(305, 431)	271**	(223, 390)	296	(215, 429)	222***	(155, 282)	32	(9, 91)
Niacin equivalents (mg)	56.36	(43.59, 63.74)	45.10**	(38.35, 54.00)	44.54	(34.55, 53.45)	32.55***	(27.93, 41.28)	26	(4, 69)
Iron (mg)	15.7	(13.3, 19.5)	13.6*	(11.8, 15.5)	13.7	(10.4, 18.2)	11.5*	(9.3, 14.6)	24	(-15, 52)
Calcium (mg)	1364	(1055, 1649)	980***	(773, 1245)	1023	(759, 1498)	756***	(528, 991)	54	(11, 82)
Magnesium (mg)	450	(391, 527)	335***	(260, 392)	388	(302, 512)	270***	(226, 337)	40	(13, 79)
Potassium (mg)	4610	(3963, 5455)	3336***	(2642, 4345)	4117	(3076, 5119)	3135***	(2368, 3539)	39	(11, 74)
Zinc (mg)	14.69	(10.42, 16.82)	9.40***	(7.95, 11.85)	11.33	(8.31, 14.94)	7.45***	(5.93, 9.13)	54	(10, 78)

FFQ – food-frequency questionnaire; WR – weighed diet record; P₂₅ – 25th percentile; P₇₅ – 75th percentile; Relative difference = [(FFQ – WR)/WR] × 100. SFA – saturated fatty acids; PUFA – polyunsaturated fatty acids; MUFA – monounsaturated fatty acids; NSP – non-starch polysaccharides. Significantly different from FFQ, Wilcoxon signed rank test: *, P < 0.05; **, P < 0.01; ***, P < 0.001.

Table 2 Pearson r and Spearman r_s correlation coefficients, percentages of subjects classified into the same and opposite thirds of intake, and weighted kappa (K_w) in 41 men

Nutrient	Pearson†		Spearman‡		Percentage classified in		K_w
	r	(95% CI)	r_s	(95% CI)	Same third	Opposite third	
Energy	0.35*	(0.04, 0.59)	0.24	(-0.01, 0.51)	34	12	0.12
Fat	0.54***	(0.27, 0.72)	0.42**	(0.13, 0.65)	44	15	0.21
SFA	0.55***	(0.29, 0.73)	0.59***	(0.34, 0.76)	51	5	0.38
PUFA	0.08	(-0.23, 0.38)	-0.07	(-0.24, 0.37)	22	22	-0.14
MUFA	0.33*	(0.03, 0.58)	0.36*	(0.01, 0.60)	37	10	0.16
Cholesterol	0.53***	(0.26, 0.72)	0.55***	(0.29, 0.73)	51	10	0.34
Protein	0.53***	(0.26, 0.72)	0.25	(-0.01, 0.52)	34	15	0.08
Sugars	0.41**	(0.11, 0.64)	0.17	(-0.15, 0.45)	37	20	0.08
Starch	0.57***	(0.32, 0.75)	0.45**	(0.16, 0.66)	46	10	0.29
NSP	0.64***	(0.41, 0.79)	0.65***	(0.43, 0.80)	54	5	0.43
Alcohol	0.83***	(0.70, 0.91)	0.72***	(0.53, 0.84)	61	2	0.53
Retinol	-0.49***	(-0.70, 0.22)	-0.13	(-0.36, 0.25)	29	22	-0.03
β -Carotene equivalents	0.16	(-0.16, 0.44)	0.11	(-0.21, 0.40)	44	22	0.12
Vitamin D	0.51***	(0.24, 0.71)	0.38*	(0.07, 0.61)	42	10	0.23
Vitamin E	0.21	(-0.10, 0.49)	0.23	(-0.08, 0.50)	39	15	0.15
Thiamin	0.83***	(0.70, 0.91)	0.37*	(0.07, 0.61)	46	17	0.19
Riboflavin	0.72***	(0.53, 0.84)	0.69***	(0.49, 0.83)	51	2	0.41
Vitamin B ₆	0.40**	(0.11, 0.63)	0.33*	(0.03, 0.58)	51	12	0.32
Vitamin B ₁₂	0.24	(-0.08, 0.51)	0.25	(-0.07, 0.52)	51	20	0.23
Vitamin C	0.64***	(0.41, 0.79)	0.48**	(0.20, 0.68)	37	12	0.15
Folate	0.56***	(0.31, 0.74)	0.55***	(0.30, 0.74)	46	10	0.29
Niacin equivalents	0.51***	(0.23, 0.70)	0.37*	(0.06, 0.60)	54	15	0.30
Iron	0.63***	(0.40, 0.79)	0.60***	(0.36, 0.77)	56	7	0.43
Calcium	0.52***	(0.25, 0.71)	0.49***	(0.21, 0.69)	54	7	0.39
Magnesium	0.61***	(0.37, 0.77)	0.66***	(0.45, 0.81)	54	2	0.45
Potassium	0.54***	(0.27, 0.72)	0.52***	(0.26, 0.71)	54	5	0.43
Zinc	0.54***	(0.28, 0.73)	0.54***	(0.27, 0.72)	39	7	0.23

SFA – saturated fatty acids; PUFA – polyunsaturated fatty acids; MUFA – monounsaturated fatty acids; NSP – non-starch polysaccharides.

† For \log_e -transformed, energy-adjusted nutrient intakes.

‡ For energy-adjusted nutrient intakes.

* Correlation is significant at the 0.05 level (two-tailed).

** Correlation is significant at the 0.01 level (two-tailed).

*** Correlation is significant at the 0.001 level (two-tailed).

and zinc in both sexes, and also for total fat, PUFA, MUFA, sugars, starch, vitamins E and C and calcium in women, and for cholesterol in men.

The Pearson correlation coefficient for thiamin was very high, but after excluding individuals who consumed Quorn (36.6 mg thiamin per 100 g¹¹), the correlation coefficient decreased from 0.83 to 0.35 in men ($n = 40$) and from 0.84 to 0.12 in women ($n = 37$), and was more similar to the Spearman correlation coefficient. The wide range of fat intakes in women also increased the Pearson correlation coefficients for total fat, SFA and MUFA. One woman consumed very little fat (7.4% of energy by WR or 11.6% of energy by FFQ), and excluding this individual from the analysis reduced the Pearson coefficients for total fat, SFA and MUFA from 0.83, 0.81 and 0.86 to 0.78, 0.64 and 0.78, respectively. Both thiamin and fat had skewness values above 1.0; however, relatively large differences between the Pearson and Spearman correlation coefficients were also found for some nutrients that had skewness values less than 1.0, e.g. fat, protein, sugars, starch, vitamin D, vitamin C and niacin in men, and riboflavin in women.

Spearman correlation coefficients were also calculated for the 29 men and 30 women who had EI/BMR values

above the cut-offs, and were found to be similar to those in all men and women (data not shown). After excluding the low energy reporters, the largest decrease in the Spearman correlation coefficient was for energy in both sexes (0.24 to 0.04 in men, and 0.39 to 0.22 in women). The largest improvements in the Spearman correlation coefficient were for vitamin D (0.38 to 0.52) in men and SFA (0.71 to 0.84) in women.

In men (Table 2), at least 50% were correctly classified into the same third of intake for 12 nutrients and 10% or less were grossly misclassified into the opposite third of intake for 14 nutrients. For all nutrients, a median of 46% of men were correctly classified (range 22% for PUFA to 61% for alcohol), and between 2 and 22% of men (median 10%) were grossly misclassified. In women (Table 3), at least 50% were correctly classified for 18 nutrients, and not more than 10% were grossly misclassified for 21 nutrients. For all nutrients, a median of 55% of women were correctly classified (range 35% for thiamin to 78% for magnesium), and 0–30% of women (median 5%) were grossly misclassified. In general, when more than 10% of individuals were grossly misclassified, the Spearman correlation coefficient was less than 0.5, and when more

Table 3 Pearson r and Spearman r_s correlation coefficients, percentages of subjects classified into the same and opposite thirds of intake, and weighted kappa (K_w) in 40 women

Nutrient	Pearson†		Spearman‡		Percentage classified in		K_w
	r	(95% CI)	r_s	(95% CI)	Same third	Opposite third	
Energy	0.40*	(0.11, 0.63)	0.39*	(0.09, 0.62)	58	13	0.37
Fat	0.83***	(0.70, 0.91)	0.64***	(0.41, 0.79)	53	8	0.37
SFA	0.81***	(0.67, 0.90)	0.71***	(0.51, 0.84)	65	5	0.54
PUFA	0.68***	(0.47, 0.82)	0.58***	(0.32, 0.75)	55	5	0.43
MUFA	0.86***	(0.75, 0.92)	0.66***	(0.44, 0.81)	43	8	0.26
Cholesterol	0.51***	(0.23, 0.71)	0.39*	(0.09, 0.62)	35	15	0.09
Protein	0.51***	(0.24, 0.71)	0.43**	(0.14, 0.65)	50	5	0.37
Sugars	0.72***	(0.53, 0.84)	0.72***	(0.52, 0.84)	63	3	0.54
Starch	0.55***	(0.29, 0.74)	0.52***	(0.24, 0.71)	45	5	0.32
NSP	0.73***	(0.53, 0.85)	0.76***	(0.59, 0.87)	60	0	0.54
Alcohol	0.70***	(0.50, 0.83)	0.79***	(0.63, 0.88)	60	0	0.54
Retinol	0.37*	(0.07, 0.61)	0.34*	(0.03, 0.59)	45	15	0.20
β -Carotene equivalents	0.51***	(0.24, 0.71)	0.44**	(0.14, 0.66)	40	10	0.20
Vitamin D	0.39*	(0.09, 0.62)	0.37*	(0.06, 0.61)	48	13	0.26
Vitamin E	0.52***	(0.25, 0.72)	0.52***	(0.24, 0.71)	60	5	0.49
Thiamin	0.84***	(0.72, 0.91)	-0.04	(-0.28, 0.35)	35	30	-0.08
Riboflavin	0.82***	(0.68, 0.90)	0.69***	(0.48, 0.82)	65	5	0.54
Vitamin B ₆	0.55***	(0.29, 0.74)	0.48**	(0.20, 0.69)	43	8	0.26
Vitamin B ₁₂	0.39*	(0.09, 0.62)	0.31	(-0.01, 0.56)	43	13	0.20
Vitamin C	0.68***	(0.47, 0.82)	0.59***	(0.34, 0.76)	60	10	0.43
Folate	0.78***	(0.62, 0.88)	0.73***	(0.53, 0.85)	68	3	0.60
Niacin equivalents	0.57***	(0.31, 0.75)	0.43**	(0.13, 0.65)	50	5	0.37
Iron	0.64***	(0.41, 0.79)	0.54***	(0.27, 0.73)	58	8	0.43
Calcium	0.78***	(0.63, 0.88)	0.75***	(0.57, 0.86)	70	5	0.60
Magnesium	0.73***	(0.54, 0.85)	0.71***	(0.52, 0.84)	78	8	0.66
Potassium	0.70***	(0.50, 0.83)	0.67***	(0.46, 0.81)	53	3	0.49
Zinc	0.61***	(0.37, 0.77)	0.57***	(0.31, 0.75)	55	5	0.43

SFA – saturated fatty acids; PUFA – polyunsaturated fatty acids; MUFA – monounsaturated fatty acids; NSP – non-starch polysaccharides.

† For log_e-transformed, energy-adjusted nutrient intakes.

‡ For energy-adjusted nutrient intakes.

* Correlation is significant at the 0.05 level (two-tailed).

** Correlation is significant at the 0.01 level (two-tailed).

*** Correlation is significant at the 0.001 level (two-tailed).

than 20% of individuals were grossly misclassified, the Spearman correlation coefficient was less than 0.3 and not statistically significant.

Weighted kappa values are shown in Tables 2 and 3. The weighted kappa tended to be higher in women than in men, ranging from -0.14 (PUFA) to 0.53 (alcohol) in men, and from -0.08 (thiamin) to 0.66 (magnesium) in women. The weighted kappa value was above 0.4 for NSP, alcohol, riboflavin, iron, magnesium and potassium in both sexes, and also for SFA, PUFA, sugars, vitamins E and C, folate, calcium and zinc in women.

Discussion

Studies of the validity of FFQs are often difficult to carry out due to the difficulties of obtaining a sufficiently large and representative sample of the population(s) to which the FFQ may be applied, and the lack of a 'gold standard' reference method. The subjects for this study were recruited from a wide range of settings in an attempt to obtain a broad cross-section of the local population, but we cannot rule out the possibility that those who

participated had a greater interest in and awareness of diet than those who would be included in an epidemiological study of diet and disease using the FFQ alone. Because of the relatively small number of participants, the confidence intervals on the measures of association between the FFQ and WR are wide and should be seen as only broad indicators of the performance of the FFQ for individual nutrients. The WR was carried out for four rather than seven days in an effort to reduce selection bias in recruitment, although this increases the possibility that the WR was not representative of longer-term dietary patterns reported on the FFQ, and therefore complete agreement between the two methods should not be expected.

Agreement between FFQ and WR

Brunner *et al.*²⁵ have suggested that correlations between FFQs and WRs of 'about 0.5 for most nutrients and 0.8 for alcohol' are 'good evidence that the FFQ has the ability to rank individuals', while Willett suggests that when FFQs are compared with multiple records of diet, correlation coefficients may reach 0.6–0.7 but that values of 0.8–0.9

are 'unlikely'¹. In this study we found Spearman correlation coefficients in the range 0.5–0.8 for SFA, NSP, alcohol, riboflavin, folate, iron, magnesium and potassium in both men and women. Spearman correlation coefficients were also above 0.5 for cholesterol in men and for PUFA, sugars, vitamin E, vitamin C, calcium and zinc in women. Correlation coefficients for retinol and β -carotene equivalents were lower than for other nutrients, particularly in men, which may reflect high within-person variation in intake of these nutrients²⁶. For these nutrients comparison of the FFQ with the 4-day WR may underestimate the validity of the FFQ-derived intakes.

For the majority of nutrients in this study the measures of agreement were higher for women than men, with major differences seen for PUFA and sugars. This finding contrasts with results of a recent study in British civil servants, which found little difference between men and women in the agreement between an FFQ and 7-day diet diaries for 19 nutrients²⁵. One possible reason for the difference is that the men in this study were less aware of their habitual diet than women. Another possibility is that the women in this study were better able to estimate their portion size than the men. This is supported by the fact that the results for women in this study are generally better than those found in two other recent UK studies of FFQs which do not include a measure of portion size^{25,27}, while those for men are similar to the results for men participating in the Whitehall II study²⁵ (Table 4). Although Willett²⁸ concluded that 'available data suggest that such questions [on portion size] do not add substantially to the assessment of dietary intake', a recent systematic review on portion size estimation found that measures of agreement between FFQs and the reference method were highest when subjects were able to estimate their own portion size⁸.

Table 4 Spearman correlation coefficients for selected nutrients from validation studies of other FFQs

Nutrient	Men		Women		
	Brunner <i>et al.</i> ²⁵ (n = 453)	Present study (n = 41)	Brunner <i>et al.</i> ²⁵ (n = 400)	Bingham <i>et al.</i> ²⁷ (n = 127)	Present study (n = 40)
Total fat	0.42	0.42	0.43	0.55	0.64
SFA	0.52	0.59	0.58	na	0.71
Starch	0.43	0.45	0.35	0.53	0.52
NSP*	0.62	0.65	0.60	0.57	0.76
β -Carotene equivalents	0.35	0.11	0.37	0.45	0.44
Vitamin E	0.41	0.23	0.33	na	0.52
Vitamin C	0.46	0.48	0.45	0.54	0.59
Folate	0.45	0.55	0.51	na	0.73
Iron	0.58	0.60	0.53	0.43	0.54
Calcium	0.48	0.49	0.44	0.50	0.75

FFQ – food-frequency questionnaire; SFA – saturated fatty acids; NSP – non-starch polysaccharides; na – not available.

*Englyst method except for Brunner *et al.*²⁵, where fibre was calculated using the Southgate method.

Comparison of measures of agreement

The data from this study illustrate how different measures of agreement do not necessarily give the same result. Pearson correlation coefficients were generally higher than the corresponding Spearman correlation coefficients, although the magnitude of the differences between these correlation coefficients varied between nutrients. The results suggest that log transformation of the data did not remove the influence of outlying data points on the Pearson coefficients, and that the Spearman coefficients may be more reliable since they use rank order and are therefore not as sensitive to extreme values as the Pearson coefficient.

The results also show that the correlation coefficients and percentages classified into thirds did not always correspond closely. For example, 54% and 5% of men were classified into the same and opposite third for both NSP and potassium while Spearman correlation coefficients were 0.65 for NSP compared with 0.52 for potassium (Table 2). However, as the confidence intervals on the correlation coefficients are relatively wide, some differences may be due to chance.

The data also show how the percentages correctly and incorrectly classified do not always correspond closely: for example, 8% of women were classified in the opposite third for both vitamin B₆ and magnesium, but 43% were correctly classified for vitamin B₆ compared with 78% for magnesium (Table 3). As cross-classification can group subjects with widely differing intakes into one category and subjects with very similar intakes into different categories if they are close to the cut-off point, agreement between the two approaches should not be expected, particularly in studies with small numbers of subjects in which misclassification of a few subjects can make a large difference to the percentages.

Weighted kappa values in this study were generally lower than both Pearson and Spearman correlation coefficients, suggesting that the levels of acceptability proposed for correlation coefficients should not be used for weighted kappa values. The weighted kappa values did not correspond directly with the correlation coefficients: for example, in women, the Spearman correlation coefficients for MUFA and potassium were very similar at 0.66 and 0.67 while the corresponding weighted kappa values were 0.26 and 0.49. Weighted kappa and cross-classification values agreed well, e.g. for all nutrients with weighted kappa values above 0.4, over 50% of subjects were correctly classified into thirds and less than 10% of subjects were classified into the opposite third. However, the same weighted kappa can be obtained for nutrients with different proportions of subjects in the same/opposite thirds, e.g. in men, $K_w = 0.12$ for energy and β -carotene equivalents, but the proportions in the same/opposite third were 34/12 and 44/22, respectively. Therefore, although weighted kappa is valuable in that it gives a single value to represent agreement, and adjusts for

chance agreement and the degree of disagreement, it is useful to present the weighted kappa value in association with the percentages, which are intuitively more meaningful.

Impact of measurement error on studies of diet–disease relationships

Most FFQs are designed for large-scale epidemiological studies of the association between diet and disease in which the expected effects of diet on the disease are often modest. Several authors have explored the impact of different values for the correlation coefficient between the surrogate and true dietary exposure in these studies. Willett²⁸ estimated that if a true relative risk is 2.0, the observed relative risk will be attenuated to 1.62 if the correlation between the estimated and true dietary exposure is 0.7, or to 1.32 if the correlation is 0.4, while Burley *et al.* argued that for correlations of below 0.3 or 0.4, attenuation will be so severe that it will be difficult to detect associations⁸. McKeown-Eyssen and Tibshirani²⁹ pointed out that measurement error also reduces the power of the studies and increases the required sample size. They suggest that the sample size of a case–control study should be increased by a factor of around 2 if the correlation between the estimate of nutrient intake and true intake is 0.7 or a factor of around 6 if the correlation is 0.4, compared with the sample size needed for a study in which habitual intake is measured without error. However, the correlations between FFQ-derived intakes and a reference method such as WR or single estimates of biomarkers are likely to be underestimates of the correlation between the FFQ and true intake due to the fact that the reference methods are not perfect measures of long-term intake. As a result, for nutrients with correlation coefficients between FFQ and WR of between 0.4 and 0.7, these calculations represent ‘worst case’ estimates of the impact of measurement error.

Recommendations for dietary validation studies

The fact that the measures of agreement differed between nutrients and between men and women indicates that there can be no single measure of the validity of a given FFQ for all subjects and all nutrients. In studies of diet–disease relationships using FFQs, a measure of validity for the nutrients of interest in subjects similar to the study population should always be reported. If the dietary data are used as a continuous variable the Spearman correlation coefficient should be given, whereas if the data are used in categories, the percentages classified into the same/opposite category and the weighted kappa may give a more appropriate indicator of performance.

For studies designed to establish the validity of a dietary assessment tool for a range of future epidemiological studies, we suggest that both the Spearman rank correlation coefficients and the percentages classified into the same/opposite third or fourth of the nutrient

measures be reported, with the weighted kappa values as a summary measure of cross-classification. Spearman correlation coefficients above 0.5, more than 50% of subjects correctly classified and less than 10% of subjects grossly misclassified into thirds, and weighted kappa values above 0.4 are desirable for nutrients of interest if the possibility of false-negative associations between diet and disease in epidemiological studies is to be minimised.

References

- 1 Willett WC. Future directions in the development of food frequency questionnaires. *Am. J. Clin. Nutr.* 1994; **59**(Suppl.): 171S–4S.
- 2 Sempos CT. Some limitations of semiquantitative food frequency questionnaires [invited commentary]. *Am. J. Epidemiol.* 1992; **135**: 1127–32.
- 3 Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of measurement. *Lancet* 1986; **1**(8476): 307–10.
- 4 Herbert JR, Miller DR. The inappropriateness of conventional use of the correlation coefficient in assessing validity and reliability of dietary assessment methods. *Eur. J. Epidemiol.* 1991; **7**(4): 339–43.
- 5 Burema J, van Staveren WA, Feunekes GIJ. Guidelines for reports on validation studies [letter to the editor]. *Eur. J. Clin. Nutr.* 1995; **49**: 932–3.
- 6 Cohen J. Weighted kappa: nominal scale agreement with provision for scaled disagreement or partial credit. *Psychol. Bull.* 1968; **70**: 213–20.
- 7 Garrow JS. Validation of methods for estimating habitual diet: proposed guidelines [editorial]. *Eur. J. Clin. Nutr.* 1995; **49**: 231–2.
- 8 Burley V, Cade J, Margetts B, Thomson R, Warm D. *Consensus Document on the Development, Validation and Utilisation of Food Frequency Questionnaires* [online]. Available at <http://www.leeds.ac.uk/nuffield/pubs/ffq.pdf>. Accessed September 2001.
- 9 Smith WCS, Crombie IC, Tavendale R, Irving JM, Kenicer MB, Tunstall-Pedoe H. The Scottish Heart Health Study – objectives and development of methods. *Health Bull.* 1987; **45**(4): 211–7.
- 10 Yarnell JWG, Fehily AM, Milbank JE, Sweetnam PM, Walker CL. A short dietary questionnaire for use in an epidemiological survey: comparison with weighed dietary records. *Hum. Nutr. Appl. Nutr.* 1983; **37A**: 103–12.
- 11 Holland B, Welch AA, Unwin ID, Buss DH, Paul AA, Southgate DAT. *McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1991.
- 12 Holland B, Welch AA, Buss DH. *Cereals and Cereal Products. Third Supplement to McCance & Widdowson's The Composition of Foods*, 4th ed. Cambridge: Royal Society of Chemistry, 1988.
- 13 Holland B, Welch AA, Buss DH. *Milk Products and Eggs. Fourth Supplement to McCance & Widdowson's The Composition of Foods*, 4th ed. Cambridge: Royal Society of Chemistry, 1989.
- 14 Holland B, Welch AA, Buss DH. *Vegetables, Herbs and Spices. Fifth Supplement to McCance & Widdowson's The Composition of Foods*, 4th ed. Cambridge: Royal Society of Chemistry, 1991.
- 15 Holland B, Welch AA, Buss DH. *Fruit and Nuts. First Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1992.

- 16 Holland B, Welch AA, Buss DH. *Vegetable Dishes. Second Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1992.
- 17 Holland B, Brown J, Buss DH. *Fish and Fish Products. Third Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1993.
- 18 Chan W, Brown J, Buss DH. *Miscellaneous Foods. Fourth Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1994.
- 19 Chan W, Brown J, Lee SM. *Meat, Poultry and Game. Fifth Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1995.
- 20 Chan W, Brown J, Church SM, Buss DH. *Meat Products and Dishes. Sixth Supplement to McCance & Widdowson's The Composition of Foods*, 5th ed. Cambridge: Royal Society of Chemistry, 1996.
- 21 Willett WC, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. *Am. J. Epidemiol.* 1986; **124**: 17–27.
- 22 Altman DG. *Practical Statistics for Medical Research*. London: Chapman and Hall, 1991; 404–9.
- 23 Department of Health. *Dietary Reference Values for Food Energy and Nutrients for the United Kingdom*. Report of the Panel on Dietary Reference Values of the Committee on Medical Aspects of Food Policy. London: The Stationery Office, 1991.
- 24 Goldberg GR, Black AE, Jebb SA, Cole TJ, Murgatroyd WA, Prentice AM. Critical evaluation of energy intake data using fundamental principles of energy physiology: 1. Derivation of cut-off limits to identify under-recording. *Eur. J. Clin. Nutr.* 1991; **45**: 569–81.
- 25 Brunner E, Stallone D, Juneja M, Bingham S, Marmot M. Dietary assessment in Whitehall II: comparison of 7d diet diary and food-frequency questionnaire and validity against biomarkers. *Br. J. Nutr.* 2001; **86**: 405–14.
- 26 Nelson M, Black AE, Morris JA, Cole TJ. Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. *Am. J. Clin. Nutr.* 1989; **50**: 155–67.
- 27 Bingham SA, Gill C, Welch A, Cassidy A, Runswick SA, Oakes S, *et al.* Validation of dietary assessment methods in the UK arm of EPIC using weighed records, and 24-hour urinary nitrogen and potassium and serum vitamin C and carotenoids as biomarkers. *Int. J. Epidemiol.* 1997; **26**(Suppl. 1): S137–51.
- 28 Willett W. *Nutritional Epidemiology*, 2nd ed. New York: Oxford University Press, 1998.
- 29 McKeown-Eyssen GC, Tibshirani R. Implications of measurement error in exposure for the sample sizes of case-control studies. *Am. J. Epidemiol.* 1994; **139**: 415–21.