A comparison of a food frequency questionnaire with a 24-hour recall for use in an epidemiological cohort study: results from the biomarker-based Observing Protein and Energy Nutrition (OPEN) study

Arthur Schatzkin,[1] Victor Kipnis,[2] Raymond J Carroll,[3] Douglas Midthune,[2] Amy F Subar,[4] Sheila Bingham,[5] Dale A Schoeller,[6] Richard P Troiano[4] and Laurence S Freedman[7]

Background Most large cohort studies have used a food frequency questionnaire (FFQ) for assessing dietary intake. Several biomarker studies, however, have cast doubt on whether the FFQ has sufficient precision to allow detection of moderate but important diet–disease associations. We use data from the Observing Protein and Energy Nutrition (OPEN) study to compare the performance of a FFQ with that of a 24-hour recall (24HR).

Methods The OPEN study included 484 healthy volunteer participants (261 men, 223 women) from Montgomery County, Maryland, aged 40–69. Each participant was asked to complete a FFQ and 24HR on two occasions 3 months apart, and a doubly labelled water (DLW) assessment and two 24-hour urine collections during the 2 weeks after the first FFQ and 24HR assessment. For both the FFQ and 24HR and for both men and women, we calculated attenuation factors for absolute energy, absolute protein, and protein density.

Results For absolute energy and protein, a single FFQ's attenuation factor is 0.04–0.16. Repeat administrations lead to little improvement (0.08–0.19). Attenuation factors for a single 24HR are 0.10–0.20, but four repeats would yield attenuations of 0.20–0.37. For protein density a single FFQ has an attenuation of 0.3–0.4; for a single 24HR the attenuation factor is 0.15–0.25 but would increase to 0.35–0.50 with four repeats.

Conclusions Because of severe attenuation, the FFQ cannot be recommended as an instrument for evaluating relations between absolute intake of energy or protein and disease. Although this attenuation is lessened in analyses of energy-adjusted protein, it remains substantial for both FFQ and multiple 24HR. The utility of either of these instruments for detecting important but moderate relative risks (between 1.5 and 2.0), even for energy-adjusted dietary factors, is questionable.

Keywords Attenuation factor, cohort study, dietary measurement error, doubly labelled water, energy intake, food frequency questionnaire, nutritional epidemiology, protein intake, 24-hour recall, urinary nitrogen

[1] Nutritional Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, Bethesda, MD, USA.

[2] Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD, USA.

[3] Department of Statistics, Texas A&M University, College Station, TX, USA.

[4] Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Bethesda, MD, USA.

[5] Medical Research Council, Dunn Human Nutrition Unit, Cambridge, UK.

[6] University of Wisconsin, Madison, WI, USA.

[7] Department of Mathematics, Statistics and Computer Science, Bar Ilan University, Ramat Gan, Israel, and Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel.

Correspondence: Dr Arthur Schatzkin, Nutritional Epidemiology Branch, National Cancer Institute, 6120 Executive Blvd–EPS 3040, Bethesda, MD 20892–7232, USA. E-mail: schatzka@mail.nih.gov

Much of the current evidence on diet and disease has been gathered from prospective cohort studies in which large numbers of individuals report their dietary habits and are monitored for subsequent development of specific diseases. A consensus is emerging that such prospective studies give more reliable results than the retrospective case-control approach.[1] Questions persist, however, regarding the most appropriate dietary report instrument to use in large cohort studies.[2–4]

Most large cohort studies have used a version of the food frequency questionnaire (FFQ), which has been shown to be sufficiently convenient and inexpensive to allow its use in tens or even hundreds of thousands of individuals. Day et al.[2] and Bingham et al.[4] have suggested use of a 7-day diet diary instead. Day et al.'s argument rests on data from a study of 179 individuals who completed two FFQ and two 7-day diaries and also provided six 24-hour urines for analysis of nitrogen, potassium, and sodium. Assuming that these urinary biomarkers give 'unbiased' measurements of the unobservable true intake, they showed that the diary was more closely correlated with the biomarker measurements for all three nutrients than was the FFQ. However, Day et al. could not study energy-adjusted nutrient intakes, because their study did not include a biomarker for energy intake.

In this paper, we describe the results of a study similar to that of Day et al., the Observing Protein and Energy Nutrition (OPEN) study.[5] Two essential differences between the OPEN study and that of Day et al. were (1) the addition of doubly labelled water (DLW) measurements to estimate energy expenditure, a surrogate for energy intake,[6] and (2) the use of two 24-hour recalls (24HR) instead of 7-day diaries. The design, therefore, allows us to investigate both absolute and energy-adjusted intakes, although unlike Day et al.,[2] our comparison is between 24HR and FFQ.

Methods

Study design

The OPEN study was conducted by the National Cancer Institute (NCI) from September 1999 to March 2000. All 484 participants (261 men, 223 women) were healthy volunteer residents of Montgomery County, Maryland (suburban Washington DC), aged 40–69.

A complete description of the study can be found elsewhere.[5] Briefly, each participant was asked to complete a FFQ and 24HR on two occasions. The FFQ was completed within 2 weeks of Visit 1 and approximately 3 months later, within a few weeks of Visit 3. The 24HR was completed at Visit 1 and approximately 3 months later at Visit 3. Participants received their dose of DLW at Visit 1 and returned 2 weeks later (Visit 2) to complete the DLW assessment. Participants provided two 24-hour urine collections, at least 9 days apart, during the 2-week period between Visit 1 and Visit 2, verified for completeness by the para amino benzoic acid (PABA) check method.[7] Since approximately 81% of nitrogen intake is excreted through the urine,[8] and nitrogen constitutes 16% of protein, the urinary nitrate (UN) values were adjusted, dividing by 0.81 and multiplying by 6.25, to estimate the individual protein intake.

In addition to the protocol for all study participants described above, we repeated the DLW procedure in 25 volunteers (14 men, 11 women). These participants received their second DLW dose at the end of Visit 2 and returned approximately 2 weeks later to complete the DLW assessment.

Dietary assessment methods

The food frequency questionnaire

In this study, we used the Diet History Questionnaire, an FFQ, developed and evaluated at NCI.[9–13] This FFQ is a 36-page booklet which queries frequency of intake over the previous year for 124 individual food items and asks portion size for most of these by providing a choice of three ranges. For 44 of the 124 foods, the FFQ asks from one to seven additional embedded questions about related factors such as seasonal intake, food type, (e.g. low-fat, lean, diet, caffeine-free), and/or fat uses or additions. The FFQ also includes six additional questions about use of low-fat foods, four summary questions, and ten dietary supplement questions.

The 24-hour recall

The employed 24HR was a highly standardized version utilizing the five-pass method, developed by the US Department of Agriculture for use in national dietary surveillance.[14] The recall data were collected in-person using a paper-and-pencil approach with standardized probes, food models, and coding. These data were linked to a nutrient database, the Food Intake Analysis System version 3.99, which obtains its database from updates to the 1994–1996 Continuing Survey of Food Intakes by Individuals.[15]

Biomarker measurements

Doubly labelled water

DLW, given orally at a dose of approximately 2 g 10 atom per cent $H_2^{18}O$ and 0.12 g 99.9 atom per cent ^2H_2O per kg of estimated total body water along with a subsequent 50 ml water rinse of the dose bottle, was used to assess total energy expenditure. Participants provided four spot urine samples, two shortly before and two shortly after the administration of the DLW dose. Participants $\geqslant 60$ years of age also provided a blood specimen due to the possibility of delayed bladder emptying. At the follow-up visit, approximately 2 weeks later, participants provided two more spot urine samples. Investigators at the University of Wisconsin Stable Isotope Laboratory determined energy expenditure via mass spectroscopic analysis of urine and blood specimens for deuterium and oxygen-18.[16–18]

Urine collections

In the 2-week period after Visit 1, participants collected their 24-hour urine on two separate occasions. To determine the completeness of urine collections, we asked study participants to take PABA tablets on each day they collected a 24-hour urine specimen. Investigators at the Dunn Nutrition Unit of the Medical Research Council in Cambridge, UK analysed UN and PABA. They analysed nitrogen by the Kjeldahl method and PABA by the colorimetric method. Collections with less than 70% PABA recovery were considered incomplete and removed from further analyses. Samples containing 70–85% PABA were also considered incomplete, but the content of analytes were proportionally adjusted to 93% PABA recovery.[5] To distinguish PABA from acetaminophen, taken by many participants, they used high protein liquid chromatography[19,20] to re-analyse PABA values deemed high (>110% recovery) by the colorimetric method.

Statistical methods

Attenuation resulting from measurement error

The effects of dietary measurement error on the estimation of disease risks are well known.[8] The most important concept is that of attenuation. Consider the disease model

$$R(D|T) = \alpha_0 + \alpha_1 T, \tag{1}$$

where R(D|T) denotes the risk of disease D on an appropriate scale (e.g. logistic) and T is the unobservable true long-term habitual intake of a given nutrient, also measured on an appropriate scale. The slope α_1 represents an association between the nutrient intake and disease. In logistic regression, for example, α_1 is the log relative risk (RR). Let λ be the slope in the linear regression of habitual intake, T, on reported intake, Q, based on the dietary instrument. If the instrument-based values Q are used in place of habitual intake, then instead of estimating the risk parameter α_1, one really estimates $\tilde{\alpha}_1 = \lambda\alpha_1$, the product of the slope λ and the true risk parameter α_1. Usually, in dietary studies, the value λ of is between zero and one, and so the effect of error in the instrument is to cause an underestimate of the risk parameter. This underestimation is called attenuation, and typically λ is called the attenuation factor. Values of λ closer to zero lead to more serious underestimation of risk. For the logistic regression disease model (1), a true RR of 2 for a given change in dietary exposure would be observed as $2^{0.4} = 1.27$ if the attenuation factor were 0.4, and as $2^{0.2} = 1.15$ if the attenuation factor were 0.2.

Sometimes, the RR is expressed for the standardized change of a certain amount of standard deviations of the distribution of dietary exposure, which is often interpreted as a comparison of quantiles.[21] In this case, the observed RR between quantiles will be attenuated by the Pearson correlation coefficient, $\rho(Q,T)$, between the reported and true intakes.

Measurement error also leads to loss of statistical power for testing the significance of the disease–exposure association. Approximately, the sample size required to reach the desired statistical power to detect a given risk is proportional to: $1/\{\rho^2(Q,T)\sigma_T^2\}$, or equivalently, $1/\{\lambda^2\sigma_Q^2\}$, where σ_Q^2 is the variance of the instrument-based reported intake and σ_T^2 is the variance of the true intake.[22] In particular, for a given instrument, the required sample size is inversely proportional to the squared attenuation factor, λ^2. For example, if the true attenuation factor were 0.2, the sample size, calculated to achieve the nominal power under the assumption that $\lambda = 0.4$, would be smaller by a factor of $0.4^2/0.2^2 = 4$. On the other hand, the comparison of the necessary sample sizes for different dietary assessment instruments should be based on the squared correlation coefficients between the corresponding instruments and truth.

Note that discrepancies between the reported and the true group mean intake do not in themselves affect the performance of an instrument in a cohort study. For example, an instrument that leads to <u>all</u> individuals under-reporting intake by exactly 25% would be no less useful than an instrument that gives the true intake for each individual, mainly because the ranking of the individuals would be unchanged.

Statistical analysis

Estimation of the attenuation factor λ and correlation coefficient $\rho(Q,T)$ requires collecting measurements on a second instrument, called the reference instrument, to compare with the main dietary instrument, in the same subset of individuals. Estimation of the attenuation factor requires that the adopted reference instrument have errors that are independent of both the true intake and errors in the instrument whose attenuation is being evaluated. Estimation of the correlation with true intake requires a more complex study design.[21,23] The conventional design requires that the reference instrument be unbiased, and that at least two independent repeat reference measurements be collected. Commonly in nutritional epidemiology, investigators have used multiple day food diaries or 24HR as reference measurements to evaluate FFQ, assuming that these dietary-report instruments satisfy all the above conditions and produce unbiased estimates of both the attenuation factor and correlation with true intake. There is now increasing evidence of jointly correlated biases in all dietary-report instruments, suggesting that none of them satisfies the requirements for a valid reference measure.[2,8,21,22,24–26]

In this paper we use a biomarker (M), either DLW, UN, or a combination of both, as the reference measurement. The evidence for both adjusted UN[8] and DLW[6] suggests that these are both valid, essentially unbiased reference instruments; that is, their errors have mean zero, and are unrelated to true intakes and errors in dietary-report instruments. We regard 24HR (F) as a second dietary instrument, on an equal footing with the FFQ (Q). Throughout, we applied the logarithmic transformation to energy and protein to make measurement error in the DLW and UN biomarkers additive and homo-scedastic and to better approximate normality.

We use the same statistical model as in our previous work.[8,25] Briefly, for individual i, let T_i denote usual nutrient intake, let Q_{ij} denote log nutrient intake as estimated from the jth repeat of the FFQ, j = 1, 2, let F_{ij} denote log nutrient intake as estimated from the jth repeat of the 24-hour recall, j = 1, 2, and let M_{ij} denote log nutrient intake as measured by the jth repeat of the biomarker, j = 1, 2. The statistical model specifies an error structure of the FFQ, 24HR, and biomarker, and is given by

$$\begin{aligned} Q_{ij} &= \mu_{Qj} + \beta_{Q0} + \beta_{Q1}T_i + r_i + \varepsilon_{ij} \\ F_{ij} &= \mu_{Fj} + \beta_{F0} + \beta_{F1}T_i + s_i + u_{ij} \\ M_{ij} &= \mu_{Mj} + T_i + v_{ij.} \end{aligned} \tag{2}$$

The model specifies that both the FFQ and 24HR values comprise (a) overall constant biases at the group level β_{Q0} and β_{F0}, respectively; (b) intake-related biases (i.e. those correlated with an individual's true intake), reflected by the slopes β_{Q1} and β_{F1} of the regressions of FFQ and 24HR, respectively, on true intake; (c) person-specific biases (the difference between total within-person bias and its intake-related component), r_i and s_i, that are independent of true intake T_i, have means zero, variances σ_r^2 and σ_s^2, respectively, and are correlated with the correlation coefficient ρ_{rs}, and (d) within-person random errors ε_{ij}, u_{ij} (reflecting variation between repeat measurements due to a variety of physiological and behavioral factors) with means zero and variances σ_ε^2, σ_u^2, respectively. The biomarker contains only within-person random error, u_{ij}, with mean zero and variance, σ_u^2. (Note that, for purposes of statistical modelling, any instrument measuring short-term intake—whether 24HR or a biomarker—includes deviations of short-term from longer-term intake as part of the error term.) Within-person random errors in all three instruments are assumed independent of each other

and of other terms in the model, except that 'within-pair' errors, $(\varepsilon_{ij}, u_{ij})$, $(\varepsilon_{ij}, v_{ij})$, and (u_{ij}, v_{ij}) are allowed to be correlated, if the corresponding measurements are taken contemporaneously. The model also includes time-specific group intercepts μ_{Qj}, μ_{Fj}, and μ_{Mj} for the FFQ, 24HR, and biomarker, respectively, which reflect possible differences among mean reported intakes over time and which sum to zero over j.

The model allows all its parameters to be estimated and tested. For example, the absence of overall group-level bias (type (a)) would be indicated by $\beta_{Q0} = \beta_{F0} = 0$. The absence of intake-related biases (type (b)) would be indicated by $\beta_{Q1} = \beta_{F1} = 1$. The absences of person-specific biases (type (c)) would be indicated by $\sigma_r^2 = \sigma_s^2 = 0$. The absence of a relationship between the person-specific biases on the two instruments would be indicated by $\rho_{rs} = 0$.

Model (2) specifies a much more parsimonious parameterization of dietary measurement error structure than the model considered by Plummer and Clayton,[21] but fits the data equally well.[8] The model used by Day *et al.*[2] is mathematically equivalent to model (2), but instead of introducing correlated person-specific biases in dietary-report instruments it allows within-person errors to be correlated both between instruments and between repeats within the same instrument.

In addition to absolute intakes, the OPEN study also allows us to investigate energy-adjusted intakes. We used two energy adjustment methods: nutrient density and nutrient residual.[1] Protein density was calculated as the percentage of energy coming from protein sources and then log transformed. The protein residual was calculated from the linear regression of protein on energy intake on the log scale. Both protein density and residual were calculated for each instrument using the protein and energy intakes as measured by this instrument. The convention used for dealing with biomarker-based derived measures is explained in the Appendix.

For all dietary variables, we excluded extreme outlying values that fell outside the interval given by 25th percentile minus twice the inter-quartile range to 75th percentile plus twice the inter-quartile range. For each variable and each instrument, no more than six outlying values for men and four for women were excluded from the analyses.

The estimates of the model parameters and their standard errors were obtained using the method of maximum likelihood under the assumption of normality of the random terms in the model. Standard errors were checked for accuracy by the bootstrap method. This method is similar to the method of moments used by Day *et al.*,[2] but is more efficient when the numbers of measurements per individual are not equal due to missing data and the normality assumptions are approximately correct.

Using the model parameters, the attenuation factor for the FFQ is expressed as

$$\lambda_Q = \frac{\text{cov}(T,Q)}{\text{var}(Q)} = \frac{\beta_{Q1}}{\beta_{Q1}^2 + \sigma_r^2 / \sigma_T^2 + \sigma_\varepsilon^2 / \sigma_T^2},$$

and the correlation of the FFQ and true intake is given by

$$\rho_{Q,\mathrm{T}} = \frac{\text{cov}(T,Q)}{\sqrt{\text{var}(T)\text{var}(Q)}} = \frac{\beta_{Q1}}{\sqrt{\beta_{Q1}^2 + \sigma_r^2 / \sigma_T^2 + \sigma_\varepsilon^2 / \sigma_T^2}}.$$

Both are estimated by replacing the parameters by their estimates based upon model (2). This is essentially equivalent to

adjusting for random within-person measurement error in the reference biomarker. Similarly, the attenuation factor for the 24HR and the correlation coefficient between the 24HR and true intake are estimated by plugging in the estimated parameters in the expressions

$$\lambda_F = \frac{\text{cov}(T,F)}{\text{var}(F)} = \frac{\beta_{F1}}{\beta_{F1}^2 + \sigma_s^2 / \sigma_T^2 + \sigma_u^2 / \sigma_T^2}$$

and

$$\rho_{F,\mathrm{T}} = \frac{\text{cov}(T,F)}{\sqrt{\text{var}(T)\text{var}(F)}} = \frac{\beta_{F1}}{\sqrt{\beta_{F1}^2 + \sigma_s^2 / \sigma_T^2 + \sigma_u^2 / \sigma_T^2}}.$$

When the main dietary-assessment instrument in the study is based on the average of a series of k repeat measurements, then, for the FFQ, σ_ε^2 is replaced by σ_ε^2/k and, for the 24HR, σ_u^2 is replaced by σ_u^2/k.

Results

Compliance with the study protocol was generally high, with nearly all patients completing the dietary report instruments and providing the necessary urine samples. DLW was successfully measured in 450 of the 484 participants (93%). In a substudy, the DLW procedure was repeated successfully in 24 out of 25 participants (96%). UN measurements were deemed complete by PABA analysis in 366 (76%) and 352 (73%) participants for the two collections.

In Table 1, we present sample sizes, medians, and quartiles for absolute energy, protein, and protein density, respectively. We note the 30–40% underreporting of median energy intake by the FFQ, as compared with the 10–20% underreporting for the 24HR. Median absolute protein intakes are underestimated by approximately 30% when using the FFQ, and by approximately 10% using the 24HR. In contrast, there is a slight overestimation of protein density by the FFQ and the 24HR, especially among women. Note, however that differences in group-mean reported nutrient intake do not necessarily invalidate an instrument for use in a cohort study. As explained in Methods, attenuation factor and the correlation with true intake are more important. We therefore examine the nature of the individual biases, and the attenuation and correlation with truth of the two instruments.

In Table 2 we present the between-person variance of true intake, the slopes β_{Q1} and β_{F1} associated with intake-related biases, the variances of person-specific biases (r and s), and the variances of within-person random error for the two instruments. Both instruments display across-the-board bias associated with individual intake, with the slopes consistently well below 1, leading to what is usually called a flattened slope phenomenon. If anything, energy adjustment appears to make this phenomenon even more pronounced. The bias appears somewhat more severe with the FFQ than with the 24HR for several gender–nutrient combinations, but in fact none of the differences are statistically significant at the 5% level.

Table 2 demonstrates substantial person-specific biases in both instruments. For measurements of energy or protein, variances of the person-specific bias for the FFQ are between 3–5 times higher than between-person variations in true intake. For the 24HR, variances of person-specific biases are considerably

Table 1 Medians and quartiles for energy, protein, and protein density assessed by biomarker,[a] 24-hour recall (24HR), and food frequency questionnaire (FFQ)

		Men			Women		
Nutrient	Instrument	N	Median	1st and 3rd quartiles	N	Median	1st and 3rd quartiles
Energy	Biomarker 1	245	2826	(2554, 3147)	206	2290	(2031, 2525)
(kcal/day)	Biomarker 2	13	2715	(2522, 3068)	11	2234	(1867, 2524)
	24HR 1	261	2577	(2085, 3108)	223	1937	(1565, 2438)
	24HR 2	260	2466	(1989, 3032)	222	1808	(1497, 2275)
	FFQ 1	260	1955	(1537, 2550)	222	1516	(1173, 1991)
	FFQ 2	259	1870	(1409, 2347)	221	1384	(1088, 1838)
Protein	Biomarker 1	192	411	(355,497)	174	308	(255, 374)
(kcal/day)	Biomarker 2	202	424	(352, 503)	150	299	(252, 367)
	24HR 1	261	376	(288, 476)	223	289	(217, 361)
	24HR 2	260	380	(286, 499)	222	271	(201, 358)
	FFQ 1	260	296	(226, 392)	222	226	(176, 306)
	FFQ 2	259	299	(205, 373)	221	207	(159, 281)
Protein	Biomarker 1	180	14.9	(12.7, 17.1)	160	13.9	(11.4, 16.3)
density (%)	Biomarker 2	189	14.8	(12.8, 17.1)	140	13.8	(11.2, 16.1)
	24HR 1	261	14.5	(11.9, 17.8)	223	14.9	(12.4, 17.4)
	24HR 2	260	15.5	(12.6, 18.3)	222	14.3	(12.0, 17.4)
	FFQ 1	260	15.4	(13.4, 17.0)	222	15.1	(13.1, 17.2)
	FFQ 2	259	15.5	(13.6, 17.1)	221	15.0	(13.1, 17.3)

[a] For Energy, Biomarker 1 refers to the doubly labelled water (DLW) measurement carried out on all participants, whereas Biomarker 2 refers to the DLW replication substudy completed on 24 participants. For Protein, Biomarker 1, and Biomarker 2, respectively, refer to urinary nitrogen determined in the first and second 24-hour urine collections. For Protein Density, Biomarker 1 was calculated from urinary nitrogen (1st collection) and doubly labelled water (on all participants); Biomarker 2 was calculated from urinary nitrogen (2nd collection) and DLW on all participants.

Table 2 Estimated between-person variation of true intake (σ_T^2), slopes in the regressions of reported on true intake (β_{Q1} or β_{F1}), variances of person-specific bias (σ_r^2 or σ_s^2), and within-person variation (σ_ε^2 or σ_u^2) for energy, protein, and protein density assessed by the food frequency questionnaire (FFQ) or 24-hour recall (24HR). Standard errors are given in parentheses. All variables are on the log scale

Nutrient	Gender	Between-person variance of true intake σ_T^2	Instrument	Slope in regression of reported true intake (β_{Q1} or β_{F1})	Variance of person-specific bias (σ_r^2 or σ_s^2)	Within-person variance (σ_ε^2 or σ_u^2)
Energy	M	0.026 (0.003)	FFQ	0.43 (0.16)	0.123 (0.012)	0.032 (0.003)
			24HR	0.63 (0.12)	0.034 (0.006)	0.053 (0.005)
	F	0.024 (0.003)	FFQ	0.22 (0.16)	0.112 (0.013)	0.039 (0.004)
			24HR	0.42 (0.12)	0.032 (0.008)	0.079 (0.007)
Protein	M	0.044 (0.006)	FFQ	0.67 (0.15)	0.133 (0.014)	0.037 (0.003)
			24HR	0.70 (0.11)	0.039 (0.009)	0.093 (0.008)
	F	0.037 (0.007)	FFQ	0.65 (0.21)	0.110 (0.015)	0.048 (0.005)
			24HR	0.60 (0.16)	0.026 (0.011)	0.120 (0.012)
Protein density	M	0.031 (0.005)	FFQ	0.46 (0.08)	0.016 (0.002)	0.012 (0.001)
			24HR	0.61 (0.11)	0.012 (0.005)	0.058 (0.005)
	F	0.035 (0.007)	FFQ	0.37 (0.11)	0.024 (0.004)	0.014 (0.001)
			24HR	0.39 (0.13)	0.012 (0.006)	0.068 (0.006)

smaller and comparable to between-person variations in true intake. Energy adjustment reduces person-specific biases, especially for the FFQ, where the variances of the person-specific bias are still higher compared with the 24HR, but only by 1.5- to 2-fold. However, even for protein density, person-specific biases in both instruments are still substantial and highly statistically significantly different from zero.

For absolute intakes, within-person random variation σ_ε^2 in the FFQ is of the same magnitude as between-person variation σ_T^2 of true intake. Similar to person-specific bias, it is considerably reduced by energy adjustment. As could be expected due to day-to-day variation in intake, within-person random variation σ_u^2 in the 24HR is substantially greater. Interestingly, relative to variation of true intake, it is only moderately reduced by energy adjustment. In contrast to person-specific bias,

within-person random error variance is higher for the 24HR than for the FFQ by 1.5- to 3-fold when measuring energy or protein and by 5-fold for protein density. Comparing variances of the different sources of error, it appears that for the FFQ the person-specific bias dominates, whereas for the 24HR the within-person variation dominates.

Table 3 shows the estimated attenuation factors and correlations with truth for the two instruments, for a single administration, averages of 2, 4, or 14 repeats, and the estimated theoretical maximum that can be attained as the number of repeats becomes very large (∞). When considering energy or protein, the FFQ's attenuation factors are very low (below 0.2), and repeated administrations of the instrument do not lead to much improvement. Attenuation factors for 24HR are somewhat better, and are improved by repeat administrations. With four repeats

Table 3 Estimated attenuation factors and correlations for energy, protein, and protein density assessed by the food frequency questionnaire (FFQ) and 24-hour recall (24HR) for different numbers of repeats of the instrument. Standard errors are given in parentheses. All variables are on the log scale

Nutrient	Gender	No. repeats	Attenuation factor		Correlation with true intake	
			FFQ	24HR	FFQ	24HR
Energy	M	1	0.080 (0.025)	0.176 (0.029)	0.199 (0.061)	0.342 (0.051)
		2	0.089 (0.028)	0.243 (0.039)	0.210 (0.064)	0.402 (0.059)
		4	0.094 (0.029)	0.300 (0.049)	0.216 (0.066)	0.446 (0.064)
		14	0.098 (0.031)	0.360 (0.062)	0.220 (0.067)	0.489 (0.071)
		∞	0.100 (0.031)	0.391 (0.070)	0.222 (0.068)	0.510 (0.074)
	F	1	0.039 (0.028)	0.096 (0.027)	0.098 (0.069)	0.210 (0.057)
		2	0.045 (0.032)	0.146 (0.041)	0.105 (0.074)	0.259 (0.070)
		4	0.048 (0.035)	0.197 (0.056)	0.109 (0.077)	0.300 (0.080)
		14	0.051 (0.037)	0.263 (0.078)	0.112 (0.079)	0.347 (0.093)
		∞	0.053 (0.038)	0.304 (0.095)	0.113 (0.080)	0.372 (0.101)
Protein	M	1	0.156 (0.034)	0.202 (0.032)	0.323 (0.067)	0.375 (0.054)
		2	0.173 (0.038)	0.289 (0.046)	0.340 (0.070)	0.449 (0.063)
		4	0.183 (0.040)	0.369 (0.059)	0.349 (0.072)	0.508 (0.070)
		14	0.190 (0.042)	0.459 (0.079)	0.357 (0.073)	0.566 (0.078)
		∞	0.194 (0.042)	0.509 (0.082)	0.360 (0.074)	0.597 (0.083)
	F	1	0.137 (0.041)	0.139 (0.035)	0.298 (0.088)	0.290 (0.070)
		2	0.159 (0.048)	0.223 (0.055)	0.321 (0.094)	0.367 (0.087)
		4	0.173 (0.052)	0.320 (0.080)	0.335 (0.098)	0.440 (0.103)
		14	0.184 (0.056)	0.463 (0.133)	0.346 (0.101)	0.529 (0.127)
		∞	0.189 (0.057)	0.563 (0.142)	0.351 (0.102)	0.584 (0.144)
Protein density	M	1	0.404 (0.066)	0.233 (0.040)	0.431 (0.063)	0.379 (0.057)
		2	0.489 (0.079)	0.361 (0.059)	0.474 (0.068)	0.472 (0.069)
		4	0.545 (0.089)	0.498 (0.085)	0.500 (0.071)	0.554 (0.080)
		14	0.594 (0.098)	0.683 (0.135)	0.522 (0.074)	0.649 (0.096)
		∞	0.617 (0.102)	0.802 (0.179)	0.532 (0.075)	0.703 (0.108)
	F	1	0.316 (0.084)	0.160 (0.051)	0.346 (0.087)	0.250 (0.076)
		2	0.377 (0.100)	0.266 (0.084)	0.378 (0.095)	0.322 (0.097)
		4	0.417 (0.111)	0.396 (0.127)	0.398 (0.099)	0.393 (0.118)
		14	0.452 (0.120)	0.610 (0.220)	0.414 (0.103)	0.487 (0.149)
		∞	0.467 (0.125)	0.777 (0.318)	0.421 (0.105)	0.550 (0.174)

the attenuations approach or exceed 0.3 except for energy intake among women.

Attenuation factors for protein density are considerably better. With a single administration of the FFQ, the attenuation factor is 0.32–0.40, greater than that for a single 24HR, which has attenuations of 0.16–0.23. However, averaging over repeated 24HR performs as well or better than a single FFQ. Four repeats of a 24HR have an attenuation of 0.40–0.50. Two repeats of the FFQ achieve attenuations of about the same level as four repeats of a 24HR. The theoretical maximum values indicate that substantial further improvements might be made by increasing the number of repeats of a 24HR, but only small gains are achieved with more repeats of the FFQ.

The overall pattern of results for estimated correlations of reported intakes with truth generally follows that of the attenuation factors. For the FFQ, correlations are very low (below 0.2) for energy, and are only slightly higher (around 0.3) for protein with not much improvement with repeat administration. Correlations improve with energy adjustment, although not as substantially as attenuation factors. For the 24HR, correlations are considerably better for energy, even without repeats, and become substantially better for protein and protein density with averaging over increasing number of repeats. Interestingly, although energy adjustment improves correlations for men, for women it leads to slightly lower correlations.

Discussion

Because different dietary-assessment instruments administered in diverse populations could produce different results, it is interesting to compare our results for protein intake with previous estimates of attenuation factors obtained from studies of dietary report instruments and urinary nitrogen measurements. Day et al.[2] (in their Table 9), report attenuation factors of 0.11 and 0.15 for a single and two administrations of a modified Willett FFQ, respectively, for men and women combined. These agree well with the estimates of 0.14–0.17 shown in our Table 3. They also report attenuation factors of 0.51 and 0.59 for one and two administrations, respectively, of a 7-day diet diary, which are considerably higher than our estimates for one or two administrations of the 24HR instrument (0.14–0.29, Table 3). However, one could argue that a 7-day diary should perform somewhat similarly to 7 repeats of a 24HR, and that two such diaries should perform roughly like 14 repeated 24HR. Our estimated attenuations for 7 (not shown in Table 3) and 14 repeated 24HR are 0.42 and 0.46 for men and 0.39 and 0.46 for women. The difference between these estimates and those of Day et al. are within the sampling errors of the two studies.

In a previous analysis of a study of 160 women conducted by the Medical Research Council Dunn Nutrition Unit in Cambridge, UK, we reported[8] an estimated attenuation factor of

0.19 for protein intake assessed by a FFQ similar to the one used in the Day et al. study. Again, this is a little higher than the 0.14 found in our OPEN study, but well within the margin of sampling error.

Willett[3] criticized Day et al.[2] for not adjusting for heterogeneity in their population. We have addressed this issue in our study by analysing males and females separately. We also performed the analyses not reported earlier that included as covariates age in 5-year groups and the logarithm of body mass index. These analyses did not change materially the results reported in this paper.

In his commentary on Kipnis et al.,[25] Willett[27] criticizes the OPEN study for underestimating within-person variation in our biomarkers by not including repeat measurements of DLW and UN at the informative interval of 6 or 12 months. Underestimated within-person variation in the reference biomarker would not affect the estimated attenuation factor, which is essentially equal to the regression slope of biomarker measurement on reported intake. But it would lead to overestimated between-person variation of true intake and therefore to underestimated correlation with true intake. It is important to note that, with a valid reference instrument that is unbiased and has errors independent of those in dietary-report measurements, such as DLW or UN, within-person variation will be correctly estimated with two independent repeat administrations. We have performed an analysis of studies with repeat UN and DLW measurements separated by different time intervals.[28] The results demonstrate that the OPEN design with two consecutive averages of DLW over 2-week periods each, as well as two UN repeats separated by at least 9 days, indeed produces independent biomarker replicates and unbiased estimates of within-person variation. The criticism may be justified, though, for estimating within-person variation in protein density. As explained in the Appendix, due to the convention used to derive biomarker-based reference measure for protein density, the correlation between reported and true intake may have been underestimated by at most approximately 4%.

While our results basically confirm the observations of Day et al., they also provide a partial answer to a challenge laid down by Willett,[3] who contended that the results of Day et al. were not convincing because they did not examine the energy-adjusted nutrient intakes that are used by many epidemiologists. When investigating the measurement of protein density, we have found that the performance of both the FFQ and the 24HR are considerably improved. A single administration of the FFQ has an estimated attenuation factor of 0.30–0.40, which may be improved slightly to 0.40–0.50 by administering the instrument twice. A single 24HR has an attenuation factor of 0.15–0.25, but this can be substantially improved by repeat administrations. Four repeats lead to attenuations of 0.40–0.50, and further improvements can be achieved by extra administrations. (The models that are used for these predictions do account for the tendency to report lower amounts on repeated administrations of a 24HR. See description of model (2) in Statistical Methods.)

Our results indicate that the FFQ cannot be recommended as an instrument for evaluating the absolute intakes of energy and protein in relation to disease. Even with two administrations of an FFQ, the attenuation factors for energy and protein in men are 0.089 and 0.173 (Table 3). A true RR of 2.0 for would be observed as $2.0^{0.077} = 1.05$ and $2.0^{0.173} = 1.13$ for absolute intakes of energy and protein in men. The attenuation would be even a little greater for women (Table 3). It seems unlikely that the exact form of the FFQ would change this conclusion. The FFQ used in this and the Day et al. study carried substantial differences, yet yielded similarly poor results. The attenuation factors are somewhat greater if multiple 24HR, rather than FFQ, are used for assessing absolute intakes of energy and protein (Table 3), but the attenuation (and consequent RR dilution) remains considerable.

If, however, our objective is to evaluate relations between protein density and disease, what is the optimum dietary assessment strategy? The OPEN data indicate that, for a single administration of an FFQ, a true RR of 2.0 would be observed as 1.33 in men and 1.24 in women (derived from Table 3). These attenuated RR certainly approach the limits of detection for observational epidemiological research. Studies would have to be very large to have adequate power to detect these associations. Moreover, uncontrolled confounding could easily account for a substantial portion of such modest excess risks. If we were interested in detecting a smaller, but entirely plausible and potentially important, RR of 1.6 for a nutritional factor and disease, that RR would reduce to 1.21 in men, 1.16 in women.

What about using two FFQ in a cohort study? This would increase costs substantially but might still be a feasible approach in future nutritional epidemiological research. Even with two FFQ, though, a true RR of 2.0 for protein density would be reduced to 1.40 in men and 1.29 in women; a RR of 1.6 would become 1.26 in men, 1.19 in women (derived from Table 3). Thus, for moderate RR, the administration of two FFQ would still leave us with substantial attenuation that challenges the capabilities of even our best epidemiological studies.

Even the use of multiple 24HR—a very expensive undertaking in a large cohort study due to the 'up front' administrative costs—would not provide a solution to this attenuation dilemma. For protein density, the attenuation factors associated with the administration of four 24HR recalls are comparable to those for administration of two FFQ (Table 3).

The OPEN data speak only to the 24HR in comparison to the FFQ. It is plausible that for protein density the multiple-day diary technique—which is considerably less expensive than the multiple 24HR approach because the diary data can be entered and analysed on a nested case-control basis at the end of a study—yields qualitatively less RR attenuation than that produced by either multiply repeated 24HR or the FFQ. Data in Table 3 show that use of 14 24HR would yield attenuation factors qualitatively greater than those for 2 FFQ or 4 24HR. Table 3 data also show that the correlation coefficients for 14 24HR are substantially greater than those for 2 FFQ or 4 24HR; this translates into greater power for detecting true RR (and reduced sample size requirements). To the extent that these results for 14 24HR can be extrapolated to two 7-day diaries, then it follows that use of the multiple-day diary might allow us to detect the modest RR that neither multiple 24HR nor FFQ could detect. We need new biomarker-based dietary assessment data, obtained from diverse populations, to evaluate this possibility.

Finally, we make two cautionary comments: (1) These results are based on a univariate analysis that relates disease risk to a single dietary variable. Whether deviations from these results are substantial in a multivariate analysis is yet to be determined. (2) Our results are based on only two nutritional factors, energy

and protein (and the ratio, protein density). Although we have no direct evidence, it is reasonable that our findings and conclusions with respect to both absolute and energy-adjusted intake can be extended to other dietary factors, especially other energy-containing macronutrients. Nevertheless, we need to develop new unbiased biomarkers of dietary factors other than energy and protein if we are to evaluate more comprehensively the strengths and limitations of our dietary assessment instruments.

Acknowledgements

References

[1] Willett WC. *Nutritional Epidemiology. 2nd Edn*. New York: Oxford University Press, 1998.

[2] Day NE, McKeown N, Wong MY, Welch A, Bingham S. Epidemiological assessment of diet: a comparison of a 7-day diary with a food frequency questionnaire using urinary markers of nitrogen, potassium and sodium. *Int J Epidemiol* 2001;**30**:309–17.

[3] Willett WC. Commentary: Dietary diaries versus food frequency questionnaires—a case of undigestible data. *Int J Epidemiol* 2001; **30**:317–19.

[4] Bingham S, Gill C, Welch A *et al.* Validation of dietary assessment methods in the UK arm of EPIC. *Int J Epidemiol* 1997;**26**:S137–51.

[5] Subar AF, Kipnis V, Troiano RP *et al.* Using intake biomarkers to evaluate the extent of dietary misreporting in a large sample of adults: the Observing Protein and Energy Nutrition (OPEN) study. *Am J Epidemiol* 2003;**158**:1–13.

[6] Schoeller DA. Measurement error of energy expenditure in free-living humans by using doubly labeled water. *J Nutr* 1988;**118**:1278–89.

[7] Bingham SA, Cummings JH. Urine nitrogen as an independent validatory measure of dietary intake: a study of nitrogen balance in individuals consuming their normal diet. *Am J Clin Nutr* 1985; **42**:1276–89.

[8] Kipnis V, Midthune D, Freedman LS, Bingham S, Schatzkin A, Carroll RJ. Empirical evidence of correlated biases in dietary assessment instruments and its implications. *Am J Epidemiol* 2001;**153**:394–403.

[9] Subar AF, Thompson FE, Smith AF *et al.* Improving food frequency questionnaires: a qualitative approach using cognitive interviewing. *J Am Dietet Assoc* 1995;**95**:781–88.

[10] Subar AF, Midthune D, Kulldorff M *et al.* An evaluation of alternative approaches to assigning nutrient values to food groups in food frequency questionnaires. *Am J Epidemiol* 2000;**152**:279–86.

[11] Subar AF, Ziegler RG, Thompson FE *et al.* Is shorter always better? Relative importance of dietary questionnaire length and cognitive ease on response rates and data quality. *Am J Epidemiol* 2001; **153**:404–09.

[12] Subar AF, Thompson FE, Kipnis V *et al.* Comparative validation of the Block, Willett, and National Cancer Institute food frequency questionnaires: The Eating at America's Table Study (EATS). *Am J Epidemiol* 2001;**154**:1089–99.

[13] Thompson FE, Subar AF, Brown CC *et al.* Cognitive research enhances accuracy of food frequency questionnaire reports: results of an experimental validation study. *J Am Dietet Assoc* 2002;**102**:212–25.

[14] Moshfegh AJ, Goldman J, LaComb R, Perloff P, Cleveland L. Research results using the new USDA automated multiple-pass method. *FASEB Journal* 2001;**15(4):**Part I.

[15] Tippett KS, Cypel YS (eds). *Design and Operation: The Continuing Survey of Food Intakes by Individuals and the Diet and Health Knowledge Survey, 1994–96. Continuing Survey of Food Intakes by Individuals 1994–96.* Nationwide Food Surveys Rep. No. 96–1. US Department of Agriculture, Agricultural Research Service, 1997.

[16] Racette SB, Schoeller DA, Luke AH, Shay K, Hnilicka JH, Kushner RF. Relative dilution spaces of ^{2}H- and ^{18}O-labeled water in humans. *Am J Physiol* 1994;**267**:E585–E590.

[17] Schoeller DA, Colligan AS, Shriver T, Avak H, Bartok-Olson C. Use of an automated chromium reduction system for hydrogen isotope ratio analysis of physiological fluids applied to doubly labeled water analysis. *J Mass Spectrom* 2000;**35**:1128–32.

[18] DA Schoeller, AH Luke. Rapid ^{18}O analysis of CO_2 samples by continuous flow isotope ratio mass spectrometry. *J Mass Spectrom* 1997;**32**:1332–36.

[19] Berg JD, Chesner I, Lawson N. Practical Assessment of the NBT-PABA pancreatic function test using HPLC determination of p-aminobenzoic acid in urine. *Ann Clin Biochem* 1985;**22**:586–90.

[20] Jakobsen J, Ovesen L, Fagt S, Pederson AN. Para-aminobenzoic acid used as a marker for completeness of 24 hour urine: assessment of control limits for a specific HPLC method. *Eur J Clin Nutr* 1997; **51**:514–19.

[21] Plummer M, Clayton D. Measurement error in dietary assessment an investigation using covariance structure models, Parts I and II. *Stat Med* 1993;**12**:925–48.

[22] Kipnis V, Carroll RJ, Freedman LS, Li L. Implications of a new dietary measurement error model for estimation of relative risk: application to four calibration studies. *Am J Epidemiol* 1999;**150**:642–51.

[23] Kaaks R, Riboli E, van Staveren W. Calibration of dietary intake measurements in prospective cohort studies. *Am J Epidemiol* 1995; **142**:548–56.

[24] Kroke A, Klipstein-Grobusch K, Voss S *et al.* Validation of a self-administered food frequency questionnaire administered in the European Prospective Investigation into Cancer and Nutrition (EPIC) Study: comparison of energy, protein and macronutrient intakes estimated with doubly labeled water, urinary nitrogen, and repeated 24-h dietary recall methods. *Am J Clin Nutr* 1999;**70**:439–47.

[25] Kipnis V, Subar AF, Midthiune D *et al.* The structure of dietary measurement error: results of the OPEN biomarker study. *Am J Epidemiol* 2003;**158**:14–21.

[26] Kipnis V, Midthune D, Freedman L *et al.* New statistical approaches to dealing with bias associated with dietary data. *Public Health Nutr* 2002;**5**:915–23.

[27] Willett W. Invited Commentary: OPEN questions. *Am J Epidemiol* 2003;**158**:22–24.

[28] Kipnis V, Subar AF, Schatzkin A *et al.* Reply to 'OPEN questions'. *Am J Epidemiol* 2003;**158**:25–26.

Appendix

Derived reference measures based on the observed biomarkers

In the OPEN study, the replications of the DLW measurement were available in only a small sample of 25 individuals (14 men and 11 women). This fact did not affect the results for total energy intake since the DLW measurements were remarkably consistent across replications. The coefficient of variation in the DLW measurements was only 5.1%, in effect indicating that energy expenditure was measured with very little error.

However, a technical difficulty arose in the analysis of energy-adjusted protein. The error in the biomarker-based derived reference measure was almost entirely influenced by the error in the UN measurements where the coefficient of variation was 17.6%. As a result, attempting to estimate the within-person variance of the derived reference measure as a parameter in the model led to relatively large standard errors in the main analysis and to instability in the procedure for bootstrap calculations.

Based on these facts, in dealing with the derived reference measurements for energy-adjusted protein, we used the following convention. When defining biomarker-based reference measures for nutrient density and nutrient residual, we used the first DLW observation with both the first and second repeat UN observations. In theory, this induced some correlation between repeat biomarker-based reference observations and therefore lead to somewhat underestimated within-person variation, but the measurement error in DLW was so small that this underestimation could be ignored in practice. For example, because the logarithm of the ratio of UN to DLW is equal to the difference between their corresponding logarithms, using the first DLW measurement underestimates within-person variation in the derived reference measure by within-person variance of DLW. From the components-of-variance analysis of the 25 subjects with two DLW observations, this variation was estimated as 0.0025. Subtracting this value from the estimates of between-person variation of true protein density intake (Table 2) would decrease them by approximately 8%. This would have very little effect on estimated model parameters that depend on between-person variation of true intake. For example, it would increase the estimated correlation coefficient between reported and true protein density by 4%.

Commentary: An OPEN assessment of dietary measurement errors

Martyn Plummer and Rudolf Kaaks

In epidemiological studies of chronic disease risk in relation to diet, a crucial question is whether assessments of dietary intake can accurately characterize an individual's habitual intake of foods and nutrients. Over the last two decades, this question has been addressed in numerous validation studies. The OPEN study[1] provides the best answer yet. It is a true landmark study both because of its size and the thoroughness of its design.

The main innovation of the OPEN study is the use of doubly labelled water (i.e. water made from stable isotopes of both hydrogen and oxygen) to measure energy expenditure. Doubly labelled water is extremely expensive to produce. It is therefore unlikely that the OPEN study will be replicated in the foreseeable future. With this lack of reproducibility in mind, we would like to consider the extent to which the conclusions of the OPEN study can be generalized.

Before considering the conclusions of OPEN, it is worth reviewing the evolution of dietary validation studies. Initially, dietary measurement validation studies were based on a comparison of two assessment methods, one of which (often based on a series of weighed food consumption records) was assumed to provide a perfectly valid intake measurement. This strong validity assumption was later relaxed by the use of statistical models for measurement error. These models impose certain constraints on the design of validation studies, where the nature of the constraint depends on the purpose of the study. If the aim is to correct for the 'attenuation' effect of measurement error on estimates of disease risk then two independent measurements are required, one of which must be unbiased. If the aim is to completely characterize the error properties of the dietary assessment methods then three independent measurements are required.[2,3] The difficulty is in finding three independent estimates of dietary intake.

In practice, three main categories of dietary assessment can be distinguished: questionnaires for assessment of habitual, long-term intake; methods based on recording of actual food intake on one or more days (e.g. weighed food records, 24-hour diet recall interviews), and biomarkers of diet. It is clear that the measurement errors of instruments in the first two categories are correlated, not least because the same food tables must be used when converting foods to nutrients. This leaves only certain biomarkers as a possible alternative. One set of biomarkers of particular interest consists of those based on the urinary recovery of chemical substances from diet. Such recovery-based markers allow the computation of absolute daily intakes of nutrients in