Public Health Nutrition: 5(6A), 915–923 DOI: 10.1079/PHN2002383

Part E. New statistical approaches to dealing with bias associated with dietary data

Bias in dietary-report instruments and its implications for nutritional epidemiology

Victor Kipnis[1,*], Douglas Midthune[1], Laurence Freedman[2], Sheila Bingham[3], Nicholas E Day[4], Elio Riboli[5], Pietro Ferrari[5] and Raymond J Carroll[6]

[1]Biometry Research Group, Division of Cancer Prevention, National Cancer Institute, Bethesda, MD 20892-7354, USA: [2]Bar Ilan University, Ramat Gan, Israel and Gertner Institute for Epidemiology and Health Policy Research, Tel Hashomer, Israel: [3]Medical Research Council, Dunn Human Nutrition Unit, Cambridge, UK: [4]Strangeways Research Laboratory, University of Cambridge, Cambridge, UK: [5]Unit of Nutrition and Cancer, International Agency for Research on Cancer, Lyon, France: [6]Department of Statistics, Texas A&M University, College Station, TX, USA

Abstract

Objective: To evaluate measurement error structure in dietary assessment instruments and to investigate its implications for nutritional studies, using urinary nitrogen excretion as a reference biomarker for protein intake.

Design: The dietary assessment methods included different food-frequency questionnaires (FFQs) and such conventional dietary-report reference instruments as a series of 24-hour recalls, 4-day weighed food records or 7-day diaries.

Setting: Six original pilot validation studies within the European Prospective Investigation of Cancer (EPIC), and two validation studies conducted by the British Medical Research Council (MRC) within the Norfolk cohort that later joined as a collaborative component cohort of EPIC.

Subjects: A sample of approximately 100 to 200 women and men, aged 35–74 years, from each of eight validation studies.

Results: In assessing protein intake, all conventional dietary-report reference methods violated the critical requirements for a valid reference instrument for evaluating, and adjusting for, dietary measurement error in an FFQ. They displayed systematic bias that depended partly on true intake and partly was person-specific, correlated with person-specific bias in the FFQ. Using the dietary-report methods as reference instruments produced substantial overestimation (up to 230%) of the FFQ correlation with true usual intake and serious underestimation (up to 240%) of the degree of attenuation of FFQ-based log relative risks.

Conclusion: The impact of measurement error in dietary assessment instruments on the design, analysis and interpretation of nutritional studies may be much greater than has been previously estimated, at least regarding protein intake.

Keywords
Dietary assessment methods
Measurement error
Biological markers

Researchers have long sought an association between diet and cancer. Animal experiments, international correlation studies and pooled results from some case–control studies suggest an important, albeit moderate, association between diet and major cancers. A number of highly visible large prospective studies have recently challenged this conventional wisdom in reporting no association between dietary fat and breast cancer[1], dietary fibre and colorectal cancer[2] and, most recently, fruit and vegetable intakes and colorectal cancer[3]. This controversy may be explained by a true lack of association between diet and cancer, the results of the international correlation and case–control studies being attributable to unmeasured confounding and recall bias, respectively. Alternatively, the prospective studies themselves may have serious methodological limitations, especially in the assessment of dietary measurement error and appropriate adjustment for it.

The food-frequency questionnaire (FFQ), which evaluates a person's usual intake over a defined period, is relatively inexpensive and easy to administer (mass mailings are feasible), and has become the dietary assessment instrument of choice for large-scale nutritional epidemiological studies. Over the years, investigators have

*Corresponding author: Email vk3b@nih.gov

recognised that the reported values from FFQs are subject to substantial error that can profoundly affect the design, analysis and interpretation of studies in nutritional epidemiology[4–6].

At the individual level, error in the FFQ may include both systematic and random components. In general, systematic error comprises both intake-related and person-specific biases[7]. Intake-related bias can be thought of as arising from correlation between error and true intake that usually manifests itself in a 'flattened' slope of the regression of reported on true intake. For example, given the socio-cultural pressure to follow a 'correct' dietary pattern, persons with a low intake of supposedly healthy food may be tempted to overreport their intake and those with a high intake of supposedly unhealthy food to underreport. Person-specific bias is the difference between an individual's reported intake averaged over many repeated measures and true average intake, after taking intake-related bias into account. It varies from person to person and may be caused by personality characteristics such as susceptibility to socio-cultural influences. Examples of random within-person errors are day-to-day fluctuations in dietary exposure, processing variation, recording mistakes, and so on.

Usually, the presence of dietary measurement error attenuates (biases towards unity) the estimates of disease relative risk when analysing single exposure variables, and reduces the statistical power of the corresponding significance test. An important relation between diet and disease, therefore, may be obscured.

Realisation of this problem has prompted the integration into large epidemiological investigations of validation/calibration sub-studies that involve a more intensive, but presumably more accurate, dietary-report method, called the 'reference' instrument. Typically, multiple-day food records, sometimes with weighed quantities instead of estimated portion sizes, or multiple 24-hour recalls have been chosen as reference measurements. FFQs have been validated against such instruments, and correlations between the FFQ and reference instrument, sometimes adjusted for within-person random error in the reference instrument, have been quoted as evidence of FFQ validity[8,9]. These reference instruments have also been used for adjusting FFQ-based relative risks for measurement error, using the 'linear approximation' version of the regression calibration approach that was introduced and made popular in epidemiology by Rosner *et al.*[10]. Because, in principle, this approach allows for all three types of FFQ error (intake-related bias, person-specific bias and within-person variation), it gained recognition as the best currently available method for correcting risk estimates for dietary measurement error.

The correct application of the regression calibration approach, however, requires that the adopted reference instrument satisfy some critical conditions. Although the reference instrument may be imperfect and contain errors of its own, these errors should be independent of (i) true intake and (ii) any error component in the FFQ[7]. Throughout this paper we take these two conditions as requirements for a valid reference instrument.

A great deal of accumulated evidence suggests that dietary-report reference instruments are unlikely to meet these requirements. Studies involving biomarkers that represent valid reference measurements ('reference' biomarkers), such as doubly labelled water (DLW) for measuring energy expenditure and urinary nitrogen (UN) for measuring protein intake, suggest biased reporting on food records or recalls (on average towards under-reporting)[11–17]. More importantly, individuals are shown to differ systematically in their reporting accuracy. This could mean that all dietary-report instruments involve bias at the individual level. Part of the bias may depend on true intake (intake-related bias), therefore violating the first requirement for a valid reference instrument. Part of the bias may also be person-specific and correlate with its counterpart in the FFQ, thereby violating the second requirement.

These findings have led to proposals for new dietary measurement error models that might explain the failure of large prospective studies to find an association between diet and cancer, even were an important association to exist[7,18,19]. For example, Kipnis *et al.*[7] proposed a model that includes person-specific bias in the dietary-report reference instrument and allows this bias to be correlated with person-specific bias in the FFQ. Using a sensitivity analysis, they showed that this correlation is critical for valid regression calibration and, if ignored, could lead to substantial residual attenuation of FFQ-based relative risks. In a subsequent paper, Kipnis *et al.*[19] generalised their model to also include intake-related bias in all dietary-report instruments and provided empirical evidence supporting their assumptions, based on the results from a calibration study that included UN biomarker for protein intake.

In this paper, we take this further and explore the structure of measurement error in different dietary assessment instruments and its implications for nutritional epidemiology, by analysing data from eight validation studies. In addition to repeat UN measurements, these studies included study-specific FFQs and the dietary-report reference instruments were 24-hour recalls, 4-day weighed food records or 7-day diaries.

Models and methods

Effect of measurement error

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 true long-term usual intake of a given food or nutrient, also measured on

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

an appropriate scale (e.g. logarithmic). The slope α_1 represents an association between the nutrient intake and disease. Let $Q = T + e_Q$ denote the intake obtained from an FFQ, where the difference between the reported and true intakes, e_Q, defines measurement error. Note that short-term variation in diet is included in e_Q, as well as systematic and random error components resulting from the instrument itself. We assume throughout that error e_Q is non-differential with respect to disease D, i.e. reported intake contributes no additional information about disease risk beyond that provided by true intake.

Fitting model (1) to reported intake Q, instead of true intake T, yields a biased estimate $\tilde{\alpha}_1$ of the exposure effect. To an excellent approximation[20], the expected observed effect is expressed as

$$E(\tilde{\alpha}_1) = \lambda_1 \alpha_1, \tag{2}$$

where the bias factor λ_1 is the slope in the linear regression of the true on observed intake,

$$T = \lambda_0 + \lambda_1 Q + \xi, \tag{3}$$

where ξ denotes random error. From the regression analysis,

$$\lambda_1 \equiv \frac{\mathrm{cov}(T, Q)}{\sigma_Q^2} = \rho_{Q,T} \frac{\sigma_T}{\sigma_Q},$$

where σ_T^2 and σ_Q^2 denote the variances of the true and reported intake, respectively, and $\rho_{Q,T}$ is the correlation coefficient between the true and reported intakes. Although in general λ_1 could be negative or greater than one in magnitude, in nutritional studies usually $\rho_{Q,T} \geqslant 0$ and $\sigma_T^2 \leqslant \sigma_Q^2$ so λ_1 lies between 0 and 1 and can be thought of as an attenuation of the true effect α_1 towards the null. Therefore, we call λ_1 the 'attenuation factor'.

Measurement error also leads to loss of statistical power for testing the significance of the disease–exposure association. Assuming that the exposure is approximately normally distributed on an appropriate scale, the sample size required to reach the desired statistical power for a given exposure effect is proportional to[21]

$$N \propto 1/\{\rho^2(Q, T)\sigma_T^2\} = 1/\{\lambda_1^2 \sigma_Q^2\}. \tag{4}$$

Thus, the attenuation factor also has a strong influence on the statistical power and is extremely important in the design of a study.

Adjustment for measurement error

Following equation (2), the unbiased (adjusted) effect can be calculated as $\hat{\lambda}_1^{-1}\tilde{\alpha}_1$, where $\hat{\lambda}_1$ is an unbiased estimate of the attenuation factor. Estimation of λ_1 usually requires simultaneous evaluation of additional dietary intake measurements made by the reference instrument in a calibration sub-study. Ideally, the reference measurement would be the 'gold standard' representing true intake (T), and the attenuation factor would be estimated as the slope of regression (3). Unfortunately, such a standard does not exist in dietary studies. Instead, consider the reference measurement

$$R = T + e_R,$$

where error e_R satisfies the two requirements for a valid reference instrument:

$$\mathrm{cov}(e_R, T) = 0 \tag{5}$$

and

$$\mathrm{cov}(e_R, e_Q) = 0. \tag{6}$$

Then

$$\lambda_1 \equiv \frac{\mathrm{cov}(T, Q)}{\sigma_Q^2} = \frac{\mathrm{cov}(R, Q)}{\sigma_Q^2},$$

and the attenuation factor can be estimated as the slope of the regression of the valid reference measurement on the FFQ.

Conventional approach

The above version of the regression calibration approach was first introduced in nutritional epidemiology by Rosner *et al.*[10], who suggested using multiple-day food records as a reference instrument. Since then this method has become the state-of-the-art approach to adjustment for dietary measurement error. Other dietary-report instruments, such as 24-hour dietary recalls, have also been used as reference instruments. In applications of this approach, it is usually assumed that the dietary-report reference instrument contains only within-person random error, and that this error is uncorrelated with error in the FFQ. More precisely, for person i and repeat measurement j, the conventional model for the dietary-report reference instrument F can be expressed as

$$F_{ij} = T_i + e_{Fij}, \tag{7}$$

where

$$E(e_{Fij}|T_i) = 0 \tag{8}$$

and

$$\mathrm{cov}(e_{Fij}, e_{Qij}) = 0. \tag{9}$$

Note that assumption (8) ensures that $\mathrm{cov}(e_{Fij}, T_i) = 0$ and thus, together with assumption (9), guarantees that requirements (5) and (6) for a valid reference instrument are satisfied. Assumption (8) is especially convenient because it leads to uncorrelated within-person repeat measurements if they are taken reasonably far apart. As a result, in addition to estimating the attenuation factor, the reference instrument can also be used to estimate the correlation between FFQ and true usual intake known as the FFQ validity[22].

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

A new dietary measurement error model

Based on recent evidence, as argued previously, it seems reasonable to question model (7)–(9) and to assume instead that all dietary-report instruments have systematic intake-related and person-specific biases. Moreover, because the same personality traits can influence person-specific biases in both the FFQ and the dietary-report reference instrument, one may anticipate that these two biases are positively correlated. A basic model that allows for this more general measurement error structure in all dietary-report instruments was introduced by Kipnis *et al.*[19] and can be specified as follows. For person $i = 1, 2, \ldots, n$ and repeat measurement j, consider

$$Q_{ij} = \beta_{Q0} + \beta_{Q1} T_i + r_i + \varepsilon_{ij}, \quad j = 1, \ldots, m_Q \qquad (10)$$

and

$$F_{ij} = \beta_{F0} + \beta_{F1} T_i + s_i + u_{ij}, \quad j = 1, \ldots, m_F \qquad (11)$$

where $\beta_{Q0} + (\beta_{Q1} - 1)T_i$ and $\beta_{F0} + (\beta_{F1} - 1)T_i$ represent intake-related biases, r_i and s_i denote person-specific biases, and ε_{ij} and u_{ij} denote within-person random errors in the FFQ and dietary-report reference instrument, respectively. True intake T has mean μ_T and variance σ_T^2, and error components r, s, ε and u have mean zero and variances σ_r^2, σ_s^2, σ_ε^2 and σ_u^2, respectively. All random variables on the right-hand side of equations (10) and (11) are assumed to be mutually independent with two important exceptions. First, as was explained before, person-specific biases r and s are allowed to be correlated. Second, due to short-term fluctuation of diet over time, within-person random errors ε and u may be correlated if different measurements are taken close in time.

The conventional model (7)–(9) is a special case of this more general model that assumes that both intake-related and person-specific biases in F are equal to zero, i.e. $\beta_{F0} \equiv 0$, $\beta_{F1} \equiv 1$ and $\sigma_s^2 \equiv 0$. It is interesting to note that the presence of population-level bias in F, such as average underreporting, does not violate requirements (5) and (6) for a valid reference instrument, only if this bias is independent of true intake, i.e. if $\beta_{F0} \neq 0$ but $\beta_{F1} = 1$. Under this condition, the additional presence of person-specific bias s also does not violate requirements (5) and (6), provided s is independent of person-specific bias r in the FFQ. Even correlation of random errors ε and u due to short-term fluctuation in diet does not make the reference instrument invalid if at least one of its repeat administrations is made not too close to the administration of the FFQ[22]. It is the presence of bias related to true usual intake ($\beta_{F1} \neq 1$) and/or the presence of person-specific bias correlated with its counterpart in the FFQ that makes the dietary-report instrument invalid as a calibration reference instrument. Note, though, that bias related to either true usual intake or person-specific bias s, even if it is not correlated with r, violates assumption (8) and therefore would lead to a biased estimate of the FFQ validity.

Unfortunately, the more general model (10) and (11) is not identifiable[7] without an alternative reference measurement that (i) is essentially unbiased and (ii) has errors that are unrelated to the errors in dietary-report instruments. In practice, these requirements may be satisfied only by certain 'reference' biomarkers that have a known functional relation to dietary intake which is independent of the amount of intake and other personal characteristics such as age, gender, body mass index, physical activity, etc.[23]. At the moment, there are only a few known biomarkers that satisfy this condition. For example, for *any* given person, the mean 24-hour UN excretion provides a practically constant proportion of 81% of dietary nitrogen intake[19,24]. After adjustment for this proportion, UN measurements become essentially unbiased at the individual level. Another known example is the measurement of total energy intake using DLW, provided that subjects in the study do not, on average, gain or lose weight[15].

Consider now the second requirement for a valid reference instrument. Aside from day-to-day variation in diet, potential sources of remaining within-person random

Table 1 Design characteristics of the studies

Study	Number of men/women	Dietary-report instrument	Number of repeats of FFQ measurements	Number of repeats of dietary-report measurements for men/women	Number of repeats of 24-hour urinary nitrogen samples for men/women
EPIC pilot studies:		24-hour recall			
France	0/119		2	0/12	0/4
Germany	49/55		2	12/12	4/4
Greece	43/38		2	12/12	3/3
Italy	59/158		2	10/12	3/6
Netherlands	68/68		2	12/12	4/4
Spain	46/47		2	12/12	4/4
MRC pilot study – Cambridge (UK)	0/160	4-day weighed food record	1	0/4	0/8
EPIC–Norfolk (UK)	64/93	7-day diary	2	2/2	6/6

FFQ – food-frequency questionnaire.

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

Table 2 Means and variances of log-transformed nitrogen intake estimates

Country	Gender	First FFQ mean	First FFQ variance	Second FFQ mean	Second FFQ variance	Average dietary-report* mean	Average dietary-report* variance	Average urinary nitrogen† mean	Average urinary nitrogen† variance
EPIC pilot – France	F	2.784	0.100	2.681	0.080	2.444	0.036	2.680	0.049
EPIC pilot – Germany	M	2.733	0.126	2.535	0.091	2.512	0.054	2.765	0.080
	F	2.485	0.181	2.282	0.072	2.241	0.059	2.554	0.047
EPIC pilot – Greece	M	2.424	0.264	2.388	0.209	2.508	0.095	2.655	0.065
	F	2.293	0.146	2.180	0.177	2.258	0.079	2.497	0.061
EPIC pilot – Italy	M	2.474	0.079	2.471	0.082	2.625	0.069	2.668	0.059
	F	2.436	0.082	2.392	0.094	2.459	0.059	2.525	0.047
EPIC pilot – Netherlands	M	2.717	0.064	2.643	0.051	2.681	0.032	2.793	0.050
	F	2.409	0.053	2.381	0.051	2.355	0.036	2.528	0.041
EPIC pilot – Spain	M	2.732	0.072	2.685	0.058	2.619	0.057	2.818	0.059
	F	2.436	0.070	2.439	0.082	2.359	0.048	2.631	0.043
MRC pilot – Cambridge (UK)	F	2.544	0.071	–	–	2.357	0.036	2.492	0.036
EPIC–Norfolk (UK)	M	2.654	0.056	2.608	0.074	2.586	0.035	2.745	0.026
	F	2.514	0.081	2.504	0.088	2.348	0.040	2.471	0.039

* Averaged over all replications of dietary-report reference instrument.
† Averaged over all replications of urinary nitrogen.

errors in reference biomarkers are mostly physiological and should be different from errors in dietary-report instruments. It is therefore reasonable to assume that errors in reference biomarker measurements are independent from dietary-report errors, provided that the two types of measurement are not taken contemporaneously.

Formally, for person i and repeat measurement j, the model for a reference biomarker (M) can be expressed as

$$M_{ij} = T_i + v_{ij}, \quad j = 1, \ldots, m_M \tag{12}$$

where v_{ij} denotes within-person random error with variance σ_v^2 and is assumed to be independent of true intake T_i and of error components in dietary-report instruments, with one exception. As before, if the biomarker measurement is taken contemporaneously with dietary-report measurements, short-term fluctuation in diet may induce non-zero correlation among within-person random errors ε_{ij}, u_{ij} and v_{ij}.

Validation data

The data come from six original pilot validation studies within the European Prospective Investigation of Cancer (EPIC)[25], the pilot validation study conducted by the British Medical Research Council (MRC)[26] and the validation study within the Norfolk cohort that joined later as a collaborative component cohort of EPIC[27]. The basic design of the eight validation studies is summarised in Table 1. Briefly, the six original EPIC validation studies in France, Germany, Greece, Italy, The Netherlands and Spain included country-specific FFQs, 24-hour recalls as the dietary-report reference instruments, and 24-hour urine collections. In most countries, FFQs were administered at the beginning and end of a one-year period. During this year, monthly 24-hour recalls were conducted, as far as possible covering equally all days of the week. In addition, at least three 24-hour urine collections were obtained at intervals of several months.

In the MRC pilot validation study, the principal measurements were a 4-day weighed food record (WFR) and two 24-hour urine collections obtained on each of four occasions (seasons) over the course of one year. Subjects were asked to make the first 24-hour urine collection on the third and fourth day of their WFR, and the second collection 3–4 days later. The study also included the Oxford FFQ that is based on the widely used Willett FFQ, modified to accommodate the characteristics of a British diet. This FFQ was administered once in season 3, one day before the start of the WFR.

The EPIC–Norfolk validation study covered a 9-month period. At baseline the participants filled out the first Oxford FFQ and completed the first 7-day diary (7DD) that was the dietary-report reference instrument. At the same time, the first series of two 24-hour urine collections (several days apart) were obtained. The other two series

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

Table 3 Parameter estimates (standard errors) for new and standard models

Country	Model	Attenuation factor	Correlation of Q and T	Variance of T	β_{Q1}	β_{F1}	Variance of r	Variance of s	Correlation of r and s
EPIC pilot – France	New	0.073 (0.081)	0.124 (0.136)	0.032 (0.007)	0.211 (0.234)	0.674 (0.127)	0.061 (0.011)	0.013 (0.005)	0.590 (0.152)
	Standard	0.233 (0.050)	0.422 (0.084)	0.029 (0.005)	0.764 (0.173)	1	0.046 (0.009)	0	0
EPIC pilot – Germany	New	0.103 (0.043)*	0.189 (0.075)*	0.046 (0.008)	0.349 (0.139)	0.375 (0.120)	0.063 (0.012)	0.039 (0.007)	0.505 (0.098)
		0.205 (0.081)	0.267 (0.102)						
	Standard	0.192 (0.044)	0.360 (0.067)	0.045 (0.008)	0.676 (0.132)	1	0.048 (0.011)	0	0
		0.386 (0.074)	0.511 (0.086)						
EPIC pilot – Greece	New	0.098 (0.056)	0.208 (0.115)	0.044 (0.010)	0.440 (0.251)	0.646 (0.177)	0.118 (0.025)	0.052 (0.012)	0.683 (0.099)
	Standard	0.336 (0.074)	0.557 (0.076)	0.071 (0.013)	0.925 (0.156)	1	0.066 (0.019)	0	0
EPIC pilot – Italy	New	0.158 (0.044)	0.259 (0.069)	0.032 (0.004)	0.424 (0.118)	0.586 (0.108)	0.039 (0.006)	0.031 (0.006)	0.375 (0.106)
	Standard	0.247 (0.049)	0.353 (0.065)	0.042 (0.006)	0.504 (0.103)	1	0.034 (0.006)	0	0
EPIC pilot – Netherlands	New	0.334 (0.071)	0.406 (0.079)	0.037 (0.006)	0.495 (0.015)	0.596 (0.080)	0.034 (0.005)	0.014 (0.003)	0.597 (0.096)
	Standard	0.443 (0.056)	0.625 (0.061)	0.027 (0.004)	0.883 (0.115)	1	0.022 (0.004)	0	0
EPIC pilot – Spain	New	0.166 (0.081)	0.221 (0.105)	0.040 (0.009)	0.295 (0.146)	0.342 (0.137)	0.031 (0.009)	0.036 (0.007)	0.951 (0.091)
	Standard	0.511 (0.064)	0.673 (0.061)	0.041 (0.008)	0.885 (0.115)	1	0.003 (0.009)	0	0
MRC pilot – Cambridge (UK)	New	0.187 (0.056)	0.284 (0.082)	0.031 (0.004)	0.430 (0.129)	0.766 (0.066)	≤ 0.065 (0.008)	0.012 (0.002)	≥ 0.349† (0.010)
	Standard	0.282 (0.054)	0.432 (0.076)	0.030 (0.004)	0.661 (0.131)	1	≤ 0.057 (0.007)	0	0
EPIC–Norfolk (UK)	New	0.068 (0.047)	0.120 (0.082)	0.024 (0.004)	0.212 (0.146)	0.614 (0.107)	0.042 (0.007)	0.019 (0.004)	0.545 (0.103)
	Standard	0.245 (0.047)	0.401 (0.070)	0.028 (0.004)	0.655 (0.129)	1	0.031 (0.006)	0	0

* In Germany the first and second replications of the food-frequency questionnaire (FFQ) have different variances, leading to different attenuation factors and correlations of Q and T.
† In MRC pilot there was only one administration of the FFQ, and therefore only the low limit for correlation of r and s could be estimated.

(two 24-hour urinary samples several days apart in each) were collected three and six months from the start. The second administrations of the FFQ and 7DD took place at random during the study period, so as not to coincide with each other and with urine collections.

In this paper, we study dietary nitrogen intake (g day^{-1}) that is calculated from protein intake by multiplying by 0.16. The biomarker was based on UN measurements that were adjusted by dividing by 81% to estimate the total nitrogen intake of every person.

In all of our analyses, we applied logarithmic transformation to the data to better approximate normality. For each study, Table 2 lists the means and variances of the log-transformed data for each administration of the corresponding FFQ and averaged over all replications of the dietary-report reference instrument and biomarker.

Model fitting

To be able to estimate all parameters in model (10)–(12), at least two repeat measurements from the FFQ, dietary-report reference instrument and biomarker are needed. Under this condition, the model can be generalised somewhat by allowing the repeat measurements for each instrument to have different group means and, if necessary, group variances, to adjust for possible time trends (e.g. seasonal effects) at the population level[19]. All validation studies except one had at least two repeated measurements for each administered instrument and therefore all of their model parameters could be estimated. Because the MRC pilot study contained only one FFQ administration, we could not estimate σ_ε^2 and σ_r^2 separately, but only their sum. Thus, for this study, we could estimate the covariance between r and s, and the correlation between $r+\varepsilon$ and s, but not the correlation $\rho(r,s)$. Note that the correlation between $r+\varepsilon$ and s provides the lower limit for $\rho(r,s)$, because ε is assumed to be independent of s.

To compare the results with the usual approach, we also fitted the conventional model, equations (7)–(9) and (10). We used the method of maximum likelihood assuming multivariate normal distribution for the data after logarithmic transformation. We checked the validity of the maximum-likelihood estimates against the bootstrap method with 500 replicates, with essentially the same results.

For those studies that included both males and females (all except EPIC–France and MRC pilot), the models were first fitted separately for each gender. Following this, overall study parameters were estimated for each study using gender as a covariate in the model. In such a model, the variance–covariance structure is assumed to be common to both genders, while the mean level of intake differs according to gender. Because both approaches produced essentially the same results, we present the overall study parameter estimates in this paper.

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

Results

Table 3 displays the estimates of the most important parameters for the new and conventional models. They include the attenuation factor λ_1, the correlation $\rho(Q, T)$ between the FFQ and true usual intake, the variance of true intake σ_T^2, the slopes β_{Q1} and β_{F1} that represent intake-related biases, the variances σ_r^2 and σ_s^2 of the person-specific biases in the FFQ and dietary-report reference instruments, respectively, and the correlation $\rho(r, s)$ between person-specific biases. For the reasons explained above, for MRC pilot, Table 3 provides the low limit for $\rho(r, s)$. The estimates were obtained for the version of both models that allowed for different mean levels in repeat measurements of all of the instruments, but assumed the same variances in all studies but one. In EPIC–Germany, the two administrations of the FFQ had highly statistically significantly different variances, and they were allowed to be different. As a result, Table 3 displays two estimates of λ_1 and $\rho(Q, T)$ for this study, one for each FFQ administration.

For all parameters, except σ_T^2, there were major differences between the new and the conventional models. First, all parameters of the more general error structure in the new model were statistically significantly different from their assumed values in the conventional model. The slopes of the regression of the dietary-report reference instruments on true intake, β_{F1}, assumed to be 1 in the conventional model, were statistically significantly smaller than 1, demonstrating the flattened slope phenomenon in the dietary-report reference instruments. In all eight studies, the variance σ_s^2 was statistically significantly different from zero, demonstrating the presence of person-specific bias in the dietary-report reference instruments. Although for most studies person-specific bias in the dietary-report instrument was somewhat smaller than person-specific bias in the FFQ, it was relatively large, with the variance from about 40% (EPIC–France, EPIC–Netherlands and MRC pilot study) to 80–100% of the variance of true intake. Most importantly, person-specific biases in the FFQ and dietary-report reference instruments were statistically significantly correlated, and their estimated correlation exceeded 0.5 in six out of seven studies where it could be estimated, with the highest value of 0.95 for EPIC–Spain.

Second, for all studies, the conventional approach suggested a slope, β_{Q1}, much closer to 1 in the regression of the FFQ on true intake, and substantially greater correlation $\rho(Q, T)$ between the FFQ and true usual intake compared with the new model, thereby overestimating the FFQ accuracy and validity.

The most important difference between the two models relates to the estimated attenuation factor λ_1, the parameter controlling the ability to detect diet–disease relationships using an FFQ. For all studies, the conventional approach yielded an attenuation factor substantially closer to 1 (i.e. much less attenuation) compared with the new model. In spite of relatively small sample sizes, the difference between the two attenuation factors was statistically significant in five out of eight studies. For each study, Table 4 displays the ratio of attenuation factors estimated by the conventional and new models. In half of the studies this ratio exceeded 3 and was below 1.5 for only one study (EPIC–Netherlands).

Table 4 Ratios of attenuation factors

Country	Ratio of attenuation factors (standard/new model)
EPIC pilot – France	3.192*
EPIC pilot – Germany	1.864
	1.883
EPIC pilot – Greece	3.429*
EPIC pilot – Italy	1.563
EPIC pilot – Netherlands	1.326
EPIC pilot – Spain	3.078*
MRC pilot – Cambridge (UK)	1.508*
EPIC–Norfolk (UK)	3.603*

* $P < 0.05$ in testing for a difference in attenuation factors yielded by the two models.

Discussion

We have applied a new, more general model for evaluating the structure of measurement error in dietary-report instruments to eight validation studies that included the UN biomarker for measuring nitrogen intake. The new model allows evaluation of whether conventional dietary-report reference methods meet two critical requirements for a valid reference instrument: (i) no correlation between its measurement error and true intake, and (ii) no correlation between its measurement error and that of the FFQ. We have documented consistent evidence that both requirements are violated due to the presence of both intake-related and person-specific biases in the dietary-report reference instruments and the correlation of the person-specific bias with that in the FFQ.

The statistical model that we have used relies on two assumptions about the UN biomarker for nitrogen intake: (a) that it is essentially unbiased and (b) that it does not contain errors related to errors in dietary-report instruments. Assumption (a) is supported by extensive literature on UN under various controlled-feeding situations[19,22,28–41]. Assumption (b) is based on the strong intuition that random discrepancies between this biomarker measurement and true intake are caused by physiological factors and therefore will be unrelated to errors in dietary-report instruments.

Using a flawed dietary-report reference instrument has very important implications for evaluating, and adjusting for, measurement error in the FFQ. From Table 3, the common approach yielded an estimated correlation between the FFQ-based nitrogen intake and true intake

https://doi.org/10.1079/PHN2002383 Published online by Cambridge University Press

substantially greater (by 36–240%) than the one estimated by the new model. This correlation is used as a measure of the FFQ validity and its squared value represents the loss in statistical power to test the significance of disease–exposure association. Thus the dietary-report reference instruments lead to serious overestimates of the statistical power of FFQ-based studies of diet and disease. Also, the flattened slope phenomenon estimated by the conventional model was less pronounced than when estimated by the new model.

Of most practical importance is the size of the differences in attenuation factors estimated by the new and conventional models. Following equation (2), the true effect of an exposure is calculated as the observed effect divided by the attenuation factor. From Table 4, the new model suggests that, for all studies but one, the true effect would be more than 50% greater, and for half the studies more than 200% greater, than the one estimated by the conventional approach.

There is also a much greater impact on the design of epidemiological studies. As follows from equation (4), for any two models, the ratio of the sample sizes for the same desired statistical power is the same as the squared ratio of their attenuation factors. Thus, the new model suggests that, based on half of the validation studies, the study size based on the conventional model should be increased by the factor of 9 or more. In other words, studies would have to be more than nine times as large as suggested by the conventional calculations in order to maintain nominal power.

Our results are limited in that they deal only with nitrogen intake, which is essentially the same as protein intake[42]. Whether the same results apply to intakes of other nutrients or foods is not clear. Direct examination of the question will not be possible until valid reference biomarkers for other dietary variables are found. A partial answer might be obtained by analysing measurement error structure for other nutrients, such as potassium, where there seem to be reliable reference biomarkers, or for total energy intake using DLW. If there are some common elements in the structure of measurement error for different dietary variables, it may be possible to adjust the new model so that the remaining parameters are estimable without additional biomarker measurements.

Another limitation of our study is that it does not deal with energy-adjusted protein intake (protein density or residual), which is often used in preference to absolute protein intake in nutrition analyses[8]. An answer to the question about the structure of measurement error and its effects on the analysis of energy-adjusted protein intake can only be obtained from biomarker studies that include both measurements of UN and DLW in the same participants.

In summary, our results suggest that the impact of measurement error in dietary-report instruments on the design, analysis and interpretation of nutritional studies may be much greater than has previously been suspected, at least regarding protein intake. Both the attenuation of relative risk estimates and the loss of statistical power in FFQ-based epidemiological studies may be substantially greater than previously estimated, due to the use of dietary-report methods as reference instruments. This means that current and past studies may be underpowered, and this may explain some of the null results that have been found in nutritional epidemiology. Until further research, including studies with simultaneous measurements of UN and DLW, gives a better understanding of the structure of dietary measurement error, the results of nutritional epidemiological studies should be interpreted with immense caution.

Acknowledgements

R.J.C.'s research was supported by a grant from the National Cancer Institute (CA-57030) and by the Texas A&M Center for Environmental and Rural Health via a grant from the National Institute of Environmental Health Sciences (P30-E509106).

References

1 Hunter DJ, Spiegelman D, Adami H-O, Beeson L, Van den Brandt PA, Folsom AR, *et al.* Cohort studies of fat intake and the risk of breast cancer – a pooled analysis. *N. Engl. J. Med.* 1996; **334**: 356–61.
2 Fuchs CS, Giovannucci EL, Colditz GA, Hunter DJ, Stampfer MJ, Rosner B, *et al.* Dietary fiber and the risk of colorectal cancer and adenoma in women. *N. Engl. J. Med.* 1999; **340**: 169–76.
3 Michels KB, Giovannucci E, Joshipura KJ, Rosner BA, Stampfer MJ, Fuchs CS, *et al.* Prospective study of fruit and vegetable consumption and incidence of colon and rectal cancers. *J. Natl. Cancer Inst.* 2000; **92**: 1740–52.
4 Beaton GH, Milner J, Corey P, McGuire V, Cousins M, Stewart E, *et al.* Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation. *Am. J. Clin. Nutr.* 1979; **32**: 2546–9.
5 Freudenheim JL, Marshall JR. The problem of profound mismeasurement and the power of epidemiological studies of diet and cancer. *Nutr. Cancer* 1988; **11**: 243–50.
6 Freedman LS, Schatzkin A, Wax J. The impact of dietary measurement error on planning sample size required in a cohort study. *Am. J. Epidemiol.* 1990; **132**: 1185–95.
7 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.
8 Willett W. *Nutritional Epidemiology*. New York: Oxford University Press, 1990.
9 Rosner B, Willett WC. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *Am. J. Epidemiol.* 1988; **127**: 377–86.
10 Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat. Med.* 1989; **8**: 1051–69.
11 Bandini LG, Schoeller DA, Cyr HN, Dietz WH. Validity of reported energy intake in obese and nonobese adolescents. *Am. J. Clin. Nutr.* 1990; **52**: 421–5.

12 Livingstone MBE, Prentice AM, Strain JJ, Coward WA, Black AE, Barker ME, *et al*. Accuracy of weighed dietary records in studies of diet and health. *Br. Med. J.* 1990; **300**: 708–12.
13 Heitmann BL. The influence of fatness, weight change, slimming history and other lifestyle variables on diet reporting in Danish men and women aged 35–65 years. *Int. J. Obes.* 1993; **17**: 329–36.
14 Heitmann BL, Lissner L. Dietary underreporting by obese individuals – is it specific or non-specific? *Br. Med. J.* 1995; **311**: 986–9.
15 Martin LJ, Su W, Jones PJ, Lockwood GA, Tritchler DL, Boyd NF. Comparison of energy intakes determined by food records and doubly labeled water in women participating in a dietary-intervention trial. *Am. J. Clin. Nutr.* 1996; **63**: 483–90.
16 Sawaya AL, Tucker K, Tsay R, Willett W, Saltzman E, Dallal GE, *et al*. Evaluation of four methods for determining energy intake in young and older women: comparison with doubly labeled water measurements of total energy expenditure. *Am. J. Clin. Nutr.* 1996; **63**: 491–9.
17 Black AE, Bingham SA, Johansson G, Coward WA. Validation of dietary intakes of protein and energy against 24 hour urinary N and DLW energy expenditure in middle-aged women, retired men and post-obese subjects: comparisons with validation against presumed energy requirements. *Eur. J. Clin. Nutr.* 1997; **51**: 405–13.
18 Prentice RL. Measurement error and results from analytic epidemiology: dietary fat and breast cancer. *J. Natl. Cancer Inst.* 1996; **88**: 1738–47.
19 Kipnis V, Midthune D, Freedman LS, Bingham S, Schatzkin A, Subar A, *et al*. Empirical evidence of correlated biases in dietary assessment instruments and its implications. *Am. J. Epidmiol.* 2001; **153**: 394–403.
20 Carroll RJ, Ruppert D, Stefanski LA. *Measurement Error in Nonlinear Models*. London: Chapman & Hall, 1995.
21 Kaaks R, Riboli E, van Staveren W. Calibration of dietary intake measurements in prospective cohort studies. *Am. J. Epidemiol.* 1995; **142**: 548–56.
22 Freedman LS, Carroll RJ, Wax Y. Estimating the relation between dietary intake obtained from a food frequency questionnaire and true average intake. *Am. J. Epidemiol.* 1991; **134**: 310–20.
23 Kaaks RJ. Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: conceptional issues. *Am. J. Clin. Nutr.* 1997; **65**(Suppl.): 1232S–9S.
24 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.
25 Kaaks R, Slimani N, Riboli E. Pilot phase studies on the accuracy of dietary intake measurements in the EPIC project: overall evaluation of results. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* 1997; **26**(Suppl. 1): S26–36.
26 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.
27 Day N, Oakes S, Luben R, Khaw KT, Bingham S, Welch A, *et al*. EPIC–Norfolk: study design and characteristics of the cohort. *Br. J. Cancer* 1999; **80**(Suppl. 1): 95–103.
28 Campbell WW, Crim MC, Dallal GE, Young VR, Evans WJ. Increased protein requirements in elderly people: new data and retrospective reassessments. *Am. J. Clin. Nutr.* 1994; **60**: 501–9.
29 Zanni E, Calloway DH, Zezulka AY. Protein requirements of elderly men. *J. Nutr.* 1979; **109**: 513–24.
30 Oddoye EA, Margen S. Nitrogen balance studies in humans: long-term effect of high nitrogen intake on nitrogen accretion. *J. Nutr.* 1979; **109**: 363–77.
31 Weller LA, Calloway DH, Margen S. Nitrogen balance of men fed amino acid mixtures based on Rose's requirements, egg white protein, and serum free amino acid patterns. *J. Nutr.* 1971; **101**: 1499–508.
32 Bunker VW, Lawson MS, Stansfield MF, Clayton BE. Nitrogen balance studies in apparently healthy elderly people and those who are housebound. *Br. J. Nutr.* 1987; **57**: 211–21.
33 Uauy R, Scrimshaw NS, Young VR. Human protein requirements: nitrogen balance response to graded levels of egg protein in elderly men and women. *Am. J. Clin. Nutr.* 1978; **31**: 779–85.
34 Castaneda C, Charnley JM, Evans WJ, Crim MC. Elderly women accommodate to a low-protein diet with losses of body cell mass, muscle function, and immune response. *Am. J. Clin. Nutr.* 1995; **62**: 30–9.
35 Cheng AHR, Gomez A, Bergan JG, Lee TC, Monckeberg F, Chichester CO. Comparative nitrogen balance study between young and aged adults using three levels of protein intake from a combination wheat–soy–milk mixture. *Am. J. Clin. Nutr.* 1978; **31**: 12–22.
36 Atinmo T, Mbofung CMF, Egun G, Osotimehin B. Nitrogen balance study in young Nigerian adult males using four levels of protein intake. *Br. J. Nutr.* 1988; **60**: 451–8.
37 Rand WM, Scrimshaw NS, Young VR. Retrospective analysis of data from five long-term, metabolic balance studies: implications for understanding dietary nitrogen and energy utilization. *Am. J. Clin. Nutr.* 1985; **42**: 1339–50.
38 Tarnopolsky MA, Atkinson SA, MacDougall JD, Chesley A, Phillips S, Schwarcz HP. Evaluation of protein requirements for trained strength athletes. *J. Appl. Physiol.* 1992; **73**: 1986–95.
39 Pannemans DLE, Wagenmakers AJM, Westerterp KR, Schaafsma G, Halliday D. Effect of protein source and quantity on protein metabolism in elderly women. *Am. J. Clin. Nutr.* 1998; **68**: 1228–35.
40 Wayler A, Queiroz E, Scrimshaw NS, Steinke FH, Rand WM, Young VR. Nitrogen balance studies in young men to assess the protein quality of an isolated soy protein in relation to meat proteins. *J. Nutr.* 1983; **113**: 2485–91.
41 Young VR, Wayler A, Garza C, Steinke FH, Murray E, Rand WM, *et al*. A long-term metabolic balance study in young men to assess the nutritional quality of an isolated soy protein and beef proteins. *Am. J. Clin. Nutr.* 1984; **39**: 8–15.
42 Matthews DE. Proteins and amino acids. In: Shils ME, Olson JA, Shike M, Ross AC, eds. *Modern Nutrition in Health and Disease*, 9th ed. Baltimore, MD: Williams & Wilkins, 1999; 11–48.