Public Health Nutrition: 5(6A), 969–976 DOI: 10.1079/PHN2002380

Part H. Advances in the statistical evaluations and interpretation of dietary data

Uses and limitations of statistical accounting for random error correlations, in the validation of dietary questionnaire assessments

Rudolf Kaaks[1,*], Pietro Ferrari[1], Antonio Ciampi[2], Martyn Plummer[1] and Elio Riboli[1]
[1]International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France:
[2]Department of Epidemiology and Biostatistics, McGill University, 1020 Pines Avenue West, Montreal, Quebec, Canada

Abstract

Objective: To examine statistical models that account for correlation between random errors of different dietary assessment methods, in dietary validation studies.
Setting: In nutritional epidemiology, sub-studies on the accuracy of the dietary questionnaire measurements are used to correct for biases in relative risk estimates induced by dietary assessment errors. Generally, such validation studies are based on the comparison of questionnaire measurements (Q) with food consumption records or 24-hour diet recalls (R). In recent years, the statistical analysis of such studies has been formalised more in terms of statistical models. This made the need of crucial model assumptions more explicit. One key assumption is that random errors must be uncorrelated between measurements Q and R, as well as between replicate measurements R_1 and R_2 within the same individual. These assumptions may not hold in practice, however. Therefore, more complex statistical models have been proposed to validate measurements Q by simultaneous comparisons with measurements R plus a biomarker M, accounting for correlations between the random errors of Q and R.
Conclusions: The more complex models accounting for random error correlations may work only for validation studies that include markers of diet based on physiological knowledge about the quantitative recovery, e.g. in urine, of specific elements such as nitrogen or potassium, or stable isotopes administered to the study subjects (e.g. the doubly labelled water method for assessment of energy expenditure). This type of marker, however, eliminates the problem of correlation of random errors between Q and R by simply taking the place of R, thus rendering complex statistical models unnecessary.

Keywords
Food-frequency questionnaires
Validation studies
Structural equation models

A central obstacle in nutritional epidemiology is the relative inaccuracy of subjects' habitual dietary intake estimates for specific food groups and nutrients. Random and uncorrelated measurement errors cause attenuation of relative risk estimates, and decrease the statistical power of studies. In addition, systematic (scaling) errors may cause problems of comparability across studies, or across sub-populations in multi-ethnic or multi-centre studies. To adjust for biases in relative risk estimates caused by measurement errors, it is increasingly being proposed that epidemiological studies of diet and disease risk should incorporate sub-studies for the validation and calibration of food-frequency questionnaire assessments of subjects' habitual dietary intake. In this paper, we review recent developments in statistical methods used for such studies, with special emphasis on structural equation modelling. Particular attention is given to problems that arise because random errors in dietary questionnaire measurements cannot be assumed to be uncorrelated with those of measurements obtained by weighed food consumption records or 24-hour diet recalls.

The concepts of 'validation' and 'calibration'

'Validation' is usually referred to as the evaluation of whether a measuring instrument really measures what is intended. When a *gold standard* measurement is available, validation consists of a comparison between the gold standard and a test measurement. If the errors in the test measurement are sufficiently small, it is considered valid, and may be substituted for the gold standard in future studies. This simple definition of validation cannot be applied in nutritional epidemiology, since a gold standard measurement does not exist. We therefore expand the

**Corresponding author:* Email kaaks@iarc.fr

definition of validation to mean evaluation of the measurement error properties of a test measurement. Validation thus takes place within the context of a given measurement error model. Increasingly complex measurement error models have been proposed to account for the various sources of error in dietary assessment. The evolution of these models is reviewed in this paper.

'Calibration' is, in general, the determination of the relationship between two measurement scales. In our case, these are the scales of measured diet and true habitual diet. In nutritional epidemiology, the term calibration is used for adjustments to the scale of questionnaire measurements of diet, such that relative risk estimates calculated for a quantitative difference in dietary intake level become unbiased. With some simplifying assumptions, calibration can be reduced to the problem of estimating a 'correction factor' that can be applied to naïve estimates of relative risk using the dietary assessment method in question[1].

Validation and calibration studies have somewhat different goals and imply different study designs[1,2]. Nevertheless, calibration studies require certain assumptions about the independence of measurement errors, and these assumptions may need to be verified in a validation study. It may be useful to think of a validation study as validating a particular measurement error model rather than validating a particular measurement method.

Dietary assessment errors and their effects on relative risk estimates

The dietary assessment instrument used most often in large-scale epidemiological studies is the food-frequency questionnaire. Initially, the measurement obtained from the questionnaire, Q, was assumed to be linked to the true habitual food intake value, T, and the measurement error, e_Q, by a simple measurement model which expresses Q as the sum of two *independent* random variables: $Q = T + e_Q$. This is known as the 'classical' measurement error model. Here the *latent variable* T, the true intake, is assumed to have finite variance σ_T^2 and errors are assumed to have mean zero, constant variance and to be uncorrelated when measurements are repeated on the same subjects. More specifically, for a particular realisation of the measurement Q on the ith individual and on occasion j, the classical model can be written as sum of two *independent* terms:

$$Q_{ij} = T_i + e_{Q,ij}, \tag{1}$$

with

$E[e_{Q,ij}] = 0$

(assumption of global unbiasedness on the group level);

$\text{Cov}[e_{Q,ij}, T] = 0$

(assumption of no correlation of errors with true intake levels); and

$\text{Cov}(e_{Q,ij}, e_{Q,i'j'}) = 0$ whenever $i \neq i'$ or $j \neq j'$

(assumption of uncorrelated errors).

This model contains some strong assumptions that are unlikely to hold in practice. To improve the model, three major modifications were eventually proposed[3–6].

First, the assumption of global unbiasedness ($E[e_{Q,ij}] = 0$) was dropped, as it was recognised that intake could be either systematically overestimated or underestimated on a group level. As an example, consider a questionnaire in which questions concerning a non-negligible source of alcohol (e.g. strong spirits) is absent; then clearly there is a bias in the measuring instrument leading to a systematic underestimation of alcohol consumption.

Second, it was recognised that also the assumption of no correlation of errors with true intake levels ($\text{Cov}[e_{Q,ij}, T_i] = 0$) may not hold in practice. It is possible, for a variety of reasons, that individuals consuming great quantities of alcohol tend to underestimate their alcohol intake more than individuals with more moderate alcohol intakes. A more accurate model should therefore include a term to represent a covariance between the error $e_{Q,ij}$ and T_i.

Third, it was recognised that even after accounting for a possible covariance between $e_{Q,ij}$ and T_i, individuals who respond to the questionnaire on multiple occasions may vary in their tendency to under- or overestimate true intake. This means that even those deviations from the true value that are uncorrelated with true intake level T can be decomposed into an individual specific bias $\delta_{Q,i}$, which is random only between individuals but not within, and a portion $\gamma_{Q,ij}$, which varies randomly within individuals from one occasion to another.

To accommodate the above departures from the original assumptions, a decomposition of the error term $e_{Q,ij}$ was postulated, including a linear relationship between individuals' questionnaire measurements and true intake levels:

$$e_{Q,ij} = \alpha_Q + \beta T_i + \varepsilon_{Q,ij},$$

where the total random error $\varepsilon_{Q,ij}$ is decomposed into $\delta_{Q,i}$ plus $\gamma_{Q,ij}$, and where $E[\delta_{Q,i}] = 0$, $E[\varepsilon_{ij}] = 0$, $\text{Var}(\delta_{Q,i}) = \sigma^2_{\delta Q}$, $Var(\gamma_{Q,i}) = \sigma^2_{\gamma Q}$, $\text{Var}(\varepsilon_{Q,ij}) = \sigma^2_{\varepsilon Q} = \sigma^2_{\delta Q} + \sigma^2_{\gamma Q}$, and all terms may be assumed to be mutually uncorrelated. This led to the following measurement model[4]:

$$Q_{ij} = \alpha_Q + \beta_Q T_i + \varepsilon_{Q,ij}, \tag{2}$$

with constant variance and uncorrelated error assumptions now holding for ε. This modified model implies a

https://doi.org/10.1079/PHN2002380 Published online by Cambridge University Press

more complex covariance structure for Q_{ij}:

$$\mathrm{Cov}(Q_{ij}, Q_{ij}) = \beta_Q^2\sigma_T^2 + \sigma_{\varepsilon Q}^2 = \beta_Q^2\sigma_T^2 + \sigma_{\delta Q}^2 + \sigma_{\gamma Q}^2$$

and

$$\mathrm{Cov}(Q_{ij}, Q_{ij'}) = \beta_Q^2\sigma_T^2 + \sigma_{\delta Q}^2 \quad \text{for } j \neq j'.$$

The parameters α_Q and β_Q express constant and proportional *scaling biases*, respectively. The term $\delta_{Q,i}$, which is systematic within subjects but varies randomly between them, has been termed *subject-specific* bias[7,8]. It is the presence of subject-specific biases that implies a non-zero correlation for replicate questionnaire measurements Q_{i1} and Q_{i2} taken from the same individual at different time points.

Within epidemiological studies, the effects of errors in dietary questionnaire assessment errors are several. First, and most importantly, random errors ε_Q cause attenuation of relative risk estimates. Under the simplifying model assumptions of approximate normality and a linear measurement error model as under equation (2) above, the magnitude of such bias depends on how the relative risk (RR) is expressed. If relative risks are expressed between quantile categories of dietary exposure, they will be biased by ρ_{QT}:

$$\log(\mathrm{RR}_{\mathrm{obs}}) = \rho_{QT}\log(\mathrm{RR}_{\mathrm{true}}), \qquad (3)$$

where ρ_{QT} is the correlation coefficient between individuals' questionnaire assessments and true intake levels. On the other hand, for absolute differences in dietary exposure on a continuous scale, log relative risk estimates will be biased by a factor

$$\lambda_Q = (1/\beta_Q)\rho_{QT}^2, \qquad (4)$$

where ρ_{QT}^2 represents the attenuation effect due to random errors, while $1/\beta_Q$ is the inverse of the proportional scaling bias in the dietary questionnaire assessments.

The design and analysis of validation and calibration studies

Classical approaches

In practice, ρ_{QT} or λ_Q can be estimated only by comparison of dietary questionnaire measurements with measurements obtained by an alternative technique. The major problem, however, is to find good reference measurements for such a comparison.

For a long time, it was assumed that accurate reference measurements could be obtained by asking subjects to record their current intake on a number of specific days, using weighed food consumption records, food consumption diaries or 24-hour diet recalls. It was thought that as long as food portions were assessed accurately by weighing or using a scale of pictures, such records would lead to highly accurate measurements of intake on each given day. With respect to the subjects' habitual dietary intakes in the longer term, the only major source of error left would then be due to within-subject, day-to-day variations in the actual intake of foods and nutrients. It was thus assumed that by increasing the number of recording days the mean of multiple food records would gradually converge to the individuals' true habitual intake values[9,10]. These considerations led to the formulation of the model:

$$R_{ij} = T_i + \varepsilon_{R,ij}, \quad \text{with } E[\varepsilon_R] = 0, \quad \mathrm{Cov}(T, \varepsilon_R) = 0, \quad \mathrm{Cov}(\varepsilon_{R,ij}, \varepsilon_{R,ik}) = 0 \quad (j \neq k). \qquad (5)$$

An additional assumption was statistical independence between the random errors of food consumption records and those of questionnaire assessments, hence: $\mathrm{Cov}(\varepsilon_R, \varepsilon_Q) = 0$. Based on the assumptions of the models of equations (2) and (5), a simple procedure to estimate the correlation ρ_{QT} was to:

1. calculate the coefficient of correlation $\rho_{Q\bar{R}}$ between measurements Q and the average of multiple days of food records ($\bar{R}$);
2. estimate the correlation $\rho_{T\bar{R}}$ between $\bar{R}$ and true intake levels, by means of an analysis of variance (to estimate within- and between-subject variances in the food records); and
3. to correct the crude estimate, $\rho_{Q\bar{R}}$, for the correlation between $\bar{R}$ and true intake levels (i.e. correcting for attenuating effects due to residual within-subject (day-to-day) variations in the average food record measurements); that is[11], $\rho_{QT} = \rho_{Q\bar{R}}/\rho_{T\bar{R}}$.

Initially, the practical implementation of calibration studies was described under the same assumptions as those of validation studies, namely that mean true intake levels of a population could be estimated correctly by average weighed food records or 24-hour diet recalls, and that random errors of such records or recalls would be independent of random errors in the dietary questionnaire (used for classification). In theory, the design of a calibration study can be lighter than that of a validation study, in that a single (non-replicated) reference measurement $R_i = T_i + \varepsilon_{R,i}$ per person (e.g. a single day's food consumption record or a single 24-hour diet recall) would be sufficient to estimate λ_Q, whereas validation studies would need at least one further measurement (e.g. a replicate measurement R_{i2} or a biochemical marker of diet)[1].

Validation in terms of structural equation models

In the mid-1990s, the statistical analysis of dietary validation studies received renewed attention. It was shown that statistical evaluation of dietary validation studies, based on the comparison of questionnaire measurements Q with replicate daily food consumption records R, could be analysed conveniently by simultaneously taking into account the measurement errors in

Table 1 Validation of questionnaire measurements by structural equation models: comparison of questionnaire assessments (Q) with replicate food consumption records (R_1 and R_2)

$$\begin{array}{ccc}
 & \begin{array}{ccc} Q & \qquad R_1 & \qquad R_2 \end{array} & \text{means} \\
\begin{array}{c} Q \\ R_1 \\ R_2 \end{array} & \begin{bmatrix} \beta_Q^2\sigma_T^2+\sigma_{\varepsilon Q}^2 & & \\ \beta_Q\sigma_T^2 & \sigma_T^2+\sigma_{\varepsilon R1}^2 & \\ \beta_Q\sigma_T^2 & \sigma_T^2 & \sigma_T^2+\sigma_{\varepsilon R2}^2 \end{bmatrix} & \begin{array}{c} \alpha_Q+\beta_Q\mu_T \\ \mu_T \\ \mu_T \end{array}
\end{array}$$

$$\begin{array}{ccc}
 & \begin{array}{ccc} Q & R_1 & R_2 \end{array} & \text{means} \\
\begin{array}{c} Q \\ R_1 \\ R_2 \end{array} & \begin{bmatrix} 3.32 & & \\ 1.50 & 2.55 & \\ 1.36 & 1.47 & 2.53 \end{bmatrix} & \begin{array}{c} 9.68 \\ 6.25 \\ 6.15 \end{array}
\end{array}$$

Parameter estimates

$\mu_T = 6.20$	$\alpha_Q = 3.64$	$\beta_Q = 0.97$
$\sigma_T^2 = 1.47$	$\sigma_{\varepsilon R}^2 = 1.07$	$\sigma_{\varepsilon Q}^2 = 1.93$
$\rho_{QT} = \sqrt{\dfrac{1}{1+(\sigma_{\varepsilon Q}^2/\beta_Q^2\sigma_T^2)}} = 0.69$	$\lambda = \dfrac{\beta_Q\sigma_T^2}{\beta_Q^2\sigma_T^2+\sigma_{\varepsilon Q}^2} = 0.43$	

both Q and R expressed by equations (2) and (5): this constitutes a simple structural equation model[4–6]. Structural equation models have been developed extensively in the field of psychometrics, which also suffers from the lack of gold standard measurements. In many situations, the number of unknown model parameters to be estimated equals the number of independent means, variances and covariances. Equating the theoretical and sample moments then results in a set of equations that can be solved directly to obtain the parameter estimates. This is illustrated in Table 1. In other situations, where the number of model parameters is smaller than the number of independent elements in the covariance matrix, a fitting algorithm can be used that is based on maximum likelihood or on minimum squared differences between the observed and theoretical moments.

The formulation of the measurement error problem within the framework of structural equation modelling has the considerable advantage of allowing generalisation to more complex study designs, such as those incorporating biomarkers of diet as a third type of measurement[4–6]. Furthermore, the introduction of structural equation modelling for dietary validation studies led to a greater awareness of essential model assumptions needed to allow estimation of the parameters of the error model. The most important of these assumptions is that *random* errors (ε) must be statistically independent between at least three measurements in the validation study, including the questionnaire measurement. This assumption means that correlations between any pair of these measurements should be entirely due to the fact that these all relate to the same latent variable – i.e. true intake – but not because the same subjects tended to make similar errors, in amount and direction, with each type of measurement. For example, if the validation study is based on the comparison of questionnaire measurements with multiple food consumption records, random errors must be assumed to be independent between measurements Q, R_i and R_j. Alternatively, a validation study may be based on the comparison between three different types of measurement – e.g. from a questionnaire (Q), food consumption records (R) and a biomarker (M) – assuming independence of random errors between the three different types of measurement[4,6,12]. Besides independence of random errors, the statistical models show that at least one measurement (R) must provide a reference scale – i.e. $R_{ij} = T_i + \varepsilon_{R,ij}$. Without these assumptions, statistical models cannot provide unique estimates for each of the unknown parameters in our measurement models, and the statistical model is called *unidentifiable*.

The problem of correlated random measurement errors for different instruments

As indicated in the definition of the model represented by equation (5), random errors for different instruments such as questionnaires and food records were at first assumed to be uncorrelated. However, since publication of the first papers on the application of structural equation models to the analysis of dietary validation studies, the assumption of uncorrelated random errors between questionnaire measurements Q_i and replicate weighed food consumption records R_{ij} and R_{ik} has been increasingly called into question. Doubts about the general applicability of this assumption were created especially by the comparison of total energy intake estimates with measurements obtained by the doubly labelled water technique[13,14] or of protein intake estimates with measurements based on 24-hour urinary nitrogen excretion[15]. These comparisons showed that, irrespective of the dietary assessment technique used, obese individuals tend to underreport their total food consumption more than lean subjects. Such

https://doi.org/10.1079/PHN2002380 Published online by Cambridge University Press

Table 2 Bias in estimates of the calibration factor λ_Q, when random errors of questionnaire assessment and reference measurements are correlated ($\rho_{\varepsilon R,\varepsilon Q} \neq 0$)

	ρ_{QT}					
	0.3		0.5		0.7	
	ρ_{RT}		ρ_{RT}		ρ_{RT}	
	0.3	0.5	0.3	0.5	0.3	0.5
$\rho_{\varepsilon R,\varepsilon Q} = 0.2$	3.02	2.10	2.10	1.60	1.65	1.35
$\rho_{\varepsilon R,\varepsilon Q} = 0.5$	6.06	3.75	3.75	2.50	2.62	1.99

(Multiplicative) bias $= 1 + \rho_{\varepsilon Q,\varepsilon R}\sqrt{[(1/\rho^2_{QT}) - 1][(1/\rho^2_{RT}) - 1]}$.

systematic differences in the subjects' tendencies towards underreporting will generally cause a positive correlation between the random errors of foods and nutrient intakes measured by questionnaires, food consumption records and 24-hour diet recalls.

It can easily be shown that a correlation between random errors ε_Q and ε_R will lead to an overestimation of the calibration factor λ_Q, if the statistical models used to estimate λ_Q do not simultaneously estimate (and hence adjust for) this error correlation. A general consequence of such bias in estimates of λ_Q is that calibration adjustments to dietary questionnaire measurements will provide only a partial correction for bias in relative risk estimates. The magnitude of bias in the estimated calibration factor λ_Q depends not only on the strength of the correlation $\rho_{\varepsilon Q,\varepsilon R}$ between random errors of Q and R, but also on the random error variances $\sigma^2_{\varepsilon Q}$ and $\sigma^2_{\varepsilon R}$ relative to the variance σ^2_T of true intake levels, and hence on the magnitude of the correlations ρ_{QT} and ρ_{RT} (Table 2).

The coefficient of correlation ρ_{QT} between questionnaire measurements and true intake levels will also tend to be overestimated if a positive correlation between random errors ε_Q and ε_R is ignored by the statistical model used for analysis. On the other hand, a positive correlation between the random errors $\varepsilon_{R,i1}$ and $\varepsilon_{R,i2}$ of replicate measurements R may cause an underestimation of the correlation ρ_{QT}. In most practical situations it will be unclear which of these two biases, upwards or downwards, predominates. More extensive theoretical simulations and sensitivity analyses, showing the magnitude of bias in estimates of λ_Q and ρ_{QT} when a correlation between the random errors ε_Q and ε_R is ignored (i.e. incorrectly assuming this correlation to be zero), have been presented by Spiegelman *et al.*[16], Wong *et al.*[17] and Kipnis *et al.*[8].

Biochemical markers of diet: a solution for the error correlation problem?

In order to avoid biases in the estimation of σ^2_T, $\sigma^2_{\varepsilon Q}$ and β_Q (and hence of ρ_{QT} and λ_Q), statistical models should take account of the correlation between the random errors in Q and R, as well as between the random errors in replicate measurements R_{i1} and R_{i2}. A fundamental problem, however, is that the number of error parameters to be estimated – which include the error covariances $\mathrm{Cov}(\varepsilon_Q, \varepsilon_R)$ and $\mathrm{Cov}(\varepsilon_{R,i1}, \varepsilon_{R,i2})$ – will be larger than the number of variances and covariances observed, as long as the validation/calibration study is based on comparisons between measurements of Q and R only. These statistical models thus remain unidentifiable. To solve this identifiability problem, a third type of measurement must be found, for which random errors can be assumed to be uncorrelated with those of Q and R, and for which random errors are also uncorrelated if the measurements are replicated over time in the same individuals. In practice, the only category of measurements that may fulfil these criteria is biochemical markers of dietary intake.

Much of the error occurring in the more traditional measurements of diet may be due to subjects' failure to recall or report accurately their intakes. Biochemical markers can be considered more 'objective' because they do not depend on a subject's report. It thus seems reasonable to assume that random errors in biomarker measurements will generally be independent of the random errors in questionnaire measurements of dietary intake or of food consumption records and 24-hour diet recalls. However, the assumption of independence of random errors in replicate marker measurements obtained from the same individuals is often more problematic.

Two categories of biomarker of diet can be distinguished: those based on a concentration and those based on recovery[12,18]. In the sections below, it is discussed whether and how these markers of diet can be of use in dietary validation/calibration studies, to overcome the problem of correlated random errors between questionnaire measurements and food consumption records.

Use of concentration-based markers

Concentration-based markers, as their name indicates, are based on the measurement of a concentration, at a given point of time, of a specific compound. Concentration measurements may be made in blood plasma (e.g. vitamin C, specific carotenoids), within a given blood lipid fraction (e.g. the relative fatty acid composition of circulating phospholipids), in an adipose tissue biopsy (e.g. the relative composition of fatty acids), in urine (e.g. sodium) or in other tissues (e.g. red blood cells) and body fluids (saliva). One key characteristic of this class of markers is that they do not have a time dimension; that is, their levels are measured and expressed without any time unit. A second characteristic is that these markers generally do not have the same quantitative relationship with dietary intake levels for *every* individual in a given study population. Concentration-based markers therefore cannot be translated into absolute intake levels per day, but at the very best can provide only a correlate of dietary intake levels. One consequence of this is that, if the objective is to estimate constant and/or proportional scaling factors α_Q and β_Q (for calibration), weighed food consumption

Table 3 Theoretical covariance matrix in validation studies with a concentration-based biomarker

$$
\begin{array}{c}
\\ Q \\ R \\ M_1 \\ M_2
\end{array}
\begin{array}{c}
\begin{array}{cccc} \quad Q \quad & \quad R \quad & \quad M_1 \quad & \quad M_2 \quad \end{array} \\
\begin{bmatrix}
\beta_Q^2\sigma_T^2+\sigma_{\varepsilon Q}^2 & & & \\
\sigma_T^2+\sigma_{\varepsilon Q,\varepsilon R}^2 & \sigma_T^2+\sigma_{\varepsilon R}^2 & & \\
\beta_Q\beta_M\sigma_T^2 & \beta_M\sigma_T^2 & \beta_M^2\sigma_T^2+\sigma_{\varepsilon M1}^2 & \\
\beta_Q\beta_M\sigma_T^2 & \beta_M\sigma_T^2 & \beta_M^2\sigma_T^2+\sigma_{\varepsilon M1,\varepsilon M2}^2 & \beta_M^2\sigma_T^2+\sigma_{\varepsilon M2}^2
\end{bmatrix}
\end{array}
$$

Parameters to be estimated

σ_T^2	β_Q	β_M	$\sigma_{\varepsilon Q}^2$
$\sigma_{\varepsilon R}^2$	$\sigma_{\varepsilon M}^2$	$\sigma_{\varepsilon R,\varepsilon Q}^2$	$\sigma_{\varepsilon M1,\varepsilon M2}^2$

records or 24-hour diet recalls must provide a reference scale (i.e. by assuming $\alpha_R = 0$ and $\beta_R = 1.0$).

Given the above, and dropping the subscripts for simplicity, we write the measurement error models as:

$$
\begin{aligned}
Q &= \alpha_Q + \beta_{QT} + \varepsilon_Q, \\
R &= T + \varepsilon_R, \\
M_1 &= \alpha_M + \beta_M T + \varepsilon_{M1}, \\
M_2 &= \alpha_M + \beta_M T + \varepsilon_{M2},
\end{aligned}
\tag{6}
$$

where $\mathrm{Cov}(\varepsilon_Q, \varepsilon_{M1}) = \mathrm{Cov}(\varepsilon_Q, \varepsilon_{M2}) = 0$ and $\sigma^2_{\varepsilon,M1} = \sigma^2_{\varepsilon,M2}$. It should be noted that here we are no longer assuming $\mathrm{Cov}(\varepsilon_Q, \varepsilon_R) = 0$ and $\mathrm{Cov}(\varepsilon_{M1}, \varepsilon_{M2}) = 0$. The population means and covariance matrix for the measurements Q, R, M_1 and M_2 are given in Table 3. Since the measurements M_1 and M_2 have the same variance, and because the covariances with measurements Q and R are also the same for M_1 and M_2, the covariance matrix in Table 3 contains only seven independent entries. However, there are eight unknown parameters in the model, so that the model is not identifiable.

The only way to make the model identifiable is to assume one or more parameter to be known. The most obvious additional assumption to be added would be $\mathrm{Cov}(\varepsilon_{M1}, \varepsilon_{M2}) = 0$, as other assumptions – e.g. $\mathrm{Cov}(\varepsilon_Q, \varepsilon_R) = 0$ or $\beta_M = 1.0$ – do not seem reasonable. Unfortunately, however, the assumption $\mathrm{Cov}(\varepsilon_{M1}, \varepsilon_{M2}) = 0$ is also unlikely to hold for concentration-based markers. The reason to reject this assumption is that between-subject variation in concentration-based markers is generally determined not only by dietary intake of a given compound, but also by variations in digestion and absorption, distribution over body compartments, endogenous synthesis and metabolism, and excretion. For example, plasma levels of β-carotene depend not only on intake levels, but also on factors affecting absorption (e.g. depending on cooking method and on amounts of co-ingested fats), internal metabolism (e.g. conversion into retinol, retinal or retinoic acids, by endogenous dioxygenases) and non-enzymatic internal breakdown of β-carotene because of smoking and other factors that may increase oxidative stress. Likewise, the fatty acid composition of plasma phospholipids or of adipose tissue depends not only on the intakes of specific fatty acids, but also on the internal synthesis of fatty acids from carbohydrates, and on the elongation and (de)saturation of polyunsaturated fatty acids. Generally, these non-dietary determinants are very likely to vary systematically between individuals, so that part of the random variations in the marker (i.e. variations not determined by diet) would tend to be correlated over time[12,18].

Assuming that there is positive correlation between the random errors ε_Q and ε_R, all that concentration-based biomarkers can add to validation studies is the estimation of an *upper limit* for ρ_{QT} and λ_Q[12], although such estimation may remain relatively imprecise if the marker does not correlate strongly with true intake levels[19].

Use of recovery-based markers

Recovery-based markers are based on precise and quantitative knowledge of the physiological balance between intake and output of a compound or chemical element. One example is the urinary excretion of nitrogen over a 24-hour period, which for *any* given individual in energy and protein balance is known to be approximately equal to 80% of nitrogen intake. Moreover, since nitrogen is present in the diet mostly in the form of protein, whereas the concentration of nitrogen in different types of protein is relatively constant, the 24-hour urinary nitrogen excretion can be translated into a valid estimate of an individual's daily protein intake. Another example is urinary excretion of potassium, which also represents a relatively constant proportion of potassium intake. Since a very large proportion of potassium intake comes from vegetables, the urinary potassium excretion can be used as an approximate marker for total vegetable consumption[20]. A third example is the doubly labelled water technique for the assessment of daily total energy expenditure, which for subjects in energy balance is very close to daily energy intake. This technique is based on the measured difference in recovery of [^{2}H] and of [^{18}O] in urine, after drinking a known amount of water that is doubly labelled with these two stable isotopes. From this difference in recovery, it can be computed how much CO_2 has been produced in the body by metabolism, and hence how much energy the body has produced[21].

Since the recovery (or difference in recovery of two different chemical elements, for the doubly labelled water method) is known to be a fixed proportion of intake for any given individual, the random variations in the marker that may occur over time can be assumed to be uncorrelated, provided that the time interval between successive biological samples is sufficiently large. In addition, since the quantitative relationship between recovery-based markers and dietary intake is known (especially for urinary nitrogen and the doubly labelled water technique), these markers can also provide a valid

https://doi.org/10.1079/PHN2002380 Published online by Cambridge University Press

Table 4 Theoretical covariance matrix in validation studies with a recovery-based biomarker, using structural equation models

$$\begin{array}{c c}
 & \begin{array}{cccc} Q & R & M_1 & M_2 \end{array} \\
\begin{array}{c} Q \\ R \\ M_1 \\ M_2 \end{array} &
\begin{bmatrix}
\beta_Q^2\sigma_T^2+\sigma_{\varepsilon Q}^2 & & & \\
\beta_Q\sigma_T^2+\sigma_{\varepsilon Q\varepsilon R}^2 & \sigma_T^2+\sigma_{\varepsilon R}^2 & & \\
\beta_Q\sigma_T^2 & \beta_R\sigma_T^2 & \sigma_T^2+\sigma_{\varepsilon M1}^2 & \\
\beta_Q\sigma_T^2 & \beta_R\sigma_T^2 & \sigma_T^2 & \sigma_T^2+\sigma_{\varepsilon M2}^2
\end{bmatrix}
\end{array}$$

Parameters to be estimated

σ_T^2	β_Q	$\sigma_{\varepsilon Q}^2$
$\sigma_{\varepsilon R}^2$	$\sigma_{\varepsilon M}^2$	$\sigma_{\varepsilon R,\varepsilon Q}^2$

reference scale. We can thus write the following measurement error models and model assumptions, for a validation study based on the comparison of questionnaire measurements (Q) with weighed food records (R) and two replicate measurements of a recovery-based marker:

$$\begin{gathered}
Q=\alpha_Q+\beta_Q T+\varepsilon_Q,\\
R=T+\varepsilon_R \text{ (alternatively, } R=\alpha_R+\beta_R T+\varepsilon_R),\\
M_1=T+\varepsilon_{M1},\\
M_2=T+\varepsilon_{M2},
\end{gathered} \tag{7}$$

where $\mathrm{Cov}(\varepsilon_Q,\varepsilon_{M1})=\mathrm{Cov}(\varepsilon_Q,\varepsilon_{M2})=0$, $\sigma^2_{\varepsilon,M1}=\sigma^2_{\varepsilon,M2}$ and $\mathrm{Cov}(\varepsilon_{M1},\varepsilon_{M2})=0$. The expected covariance matrix and parameters to be estimated in these models are shown in Table 4. Thus, for recovery-based markers, the only non-zero error correlation is that between ε_Q and ε_R, whereas for concentration-based markers non-zero correlations are allowed between ε_{M1} and ε_{M2} as well as between ε_Q and ε_R. It should also be noted that, using the above model assumptions, both the marker and the recording method are assumed to provide the same, valid reference scale for intake measurements. It would be possible to relax these assumptions, however, and to add for instance constant and proportional scaling biases to the measurement model for food consumption records (parameters α_R and β_R), to be estimated in the validation study.

For illustration, we applied this model to data from the pilot-phase validation studies of the European Prospective Investigation into Cancer and Nutrition (EPIC), for measurements of protein intake in men and women in Italy and in the Netherlands (Table 5). Measurements in this study were obtained by a food-frequency questionnaire (Q), by the average of 12 replicate 24-hour diet recalls (R) and by urinary nitrogen in two different urine samples. As shown in Table 5, the model of equation (7) gave mostly lower estimates for both ρ_{QT} and λ_Q compared with models without the biomarker, where the correlation between random errors ε_Q and ε_R was simply assumed to be zero. A similar analysis, using also urinary nitrogen excretion as a marker of protein intake, was performed by Plummer and Clayton[6].

Table 5 Estimates of the correlation coefficient ρ_{QT} and the calibration factor λ_Q, for questionnaire measurements of protein intake; EPIC pilot-phase data*

	Men			Women		
	n	ρ_{QT}	λ	n	ρ_{QT}	λ
Model with biomarkers						
Italy	59	0.35	0.25	158	0.25	0.15
The Netherlands	68	0.36	0.25	68	0.51	0.44
Model without biomarkers						
Italy	59	0.40	0.32	158	0.31	0.23
The Netherlands	68	0.64	0.43	68	0.50	0.39

*The data include two assessments by food-frequency questionnaire, at the beginning and at the end of a one-year period; twelve 24-hour diet recalls; and four urinary nitrogen measurements. For details of the study, see Kaaks *et al.*[23].

Discussion

We have reviewed developments over the last 10 years in the statistical analysis of dietary validation and calibration studies, with special emphasis on the use of structural equation models.

A major complication in validation/calibration substudies is that most probably random errors tend to be positively correlated between measurements obtained by food-frequency questionnaires and 'reference' measurements obtained by recording methods. Biochemical markers may solve this problem, but only for those nutrients for which markers are available that have uncorrelated random errors over time. Unfortunately, the latter assumption will generally hold only for recovery markers, and only very few such markers are available. This suggests a limited use of more complex structural equation models that take account of covariances (correlations) between random errors in measurements by food-frequency questionnaires and by recording methods.

An interesting observation is that markers based on recovery can also be translated into absolute daily intake levels, and thus can provide a valid reference scale. A paradoxical consequence of this is that such markers can simply replace reference measurements (R) based on subjects' reports: a structural equation model as in the example of Table 4 would remain perfectly identifiable after eliminating the measurements R from the covariance matrix. This means that, paradoxically, the problem of correlated measurement errors between measurements from questionnaires and from recording methods would be *de facto* eliminated, and the statistical analysis could also be based on much simpler structural equation models, as in the example of Table 1, comparing questionnaire measurements directly with replicate marker measurements. A recovery marker will simply

https://doi.org/10.1079/PHN2002380 Published online by Cambridge University Press

eliminate the problem of correlation between random errors ε_Q and ε_R of questionnaire assessments and consumption records by taking the place of *R*, plus the basic *assumption* that random errors in the marker are uncorrelated with those in measurements *Q*.

The question then is how, in the absence of further recovery-based markers for nutrients other than protein, energy or potassium, further progress may be made in the area of dietary validation and calibration studies. One possible way out of the problem of correlated random errors might be to assume that such correlation diminishes strongly when intake estimates are adjusted for total energy intake.

It is widely recognised that, before incriminating any specific nutrient or food in the aetiology of a disease, epidemiological studies should show that the nutrient or food remains associated with disease risk after the adjustment for total energy intake. The reason is that total energy intake is itself determined by factors such as body size, usual physical activity and metabolic efficiency, which may each also have effects on disease risk[22]. To account for the possible confounding effects of factors that lead subjects to eat more, or less, of all possible foods and nutrients, epidemiological analyses of diet–disease associations should be adjusted for total energy intake. Disease risk will thus no longer be related to absolute intake levels of nutrients or foods but rather to a measure of relative dietary composition, and methodological sub-studies on measurement error should also focus more on the validation or calibration of energy-adjusted intake levels of nutrients or foods. An interesting additional aspect of the total energy adjustment is that it might decrease substantially the correlation between random errors in different types of dietary intake assessment, if such correlation of errors were due mainly to variations in the degree of systematic underreporting by each assessment method. This will be true especially if one can assume that underreporting on average does not affect one type of nutrient or food more than another. More research should be done to verify if the latter assumption is generally reasonable, or to check whether at least it leads to a smaller degree of bias in estimates of ρ_{QT} and λ_Q than when random errors are assumed to be statistically independent for the non-adjusted intake variables.

References

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

2 Kaaks R, Riboli E. Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. European Prospective Investigation into Cancer and Nutrition. *Int. J. Epidemiol.* 1997; **26**(Suppl. 1): S15–25.

3 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.

4 Kaaks R, Riboli E, Esteve J, van Kappel AL, van Staveren WA. Estimating the accuracy of dietary questionnaire assessments: validation in terms of structural equation models. *Stat. Med.* 1994; **13**: 127–42.

5 Plummer M, Clayton D. Measurement error in dietary assessment: an investigation using covariance structure models. Part I. *Stat. Med.* 1993; **12**: 925–35.

6 Plummer M, Clayton D. Measurement error in dietary assessment: an investigation using covariance structure models. Part II. *Stat. Med.* 1993; **12**: 937–48.

7 van Staveren WA, Burema J. Dietary methodology: implications of errors in the measurement. *Proc. Nutr. Soc.* 1990; **49**: 281–7.

8 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.

9 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.

10 Nelson M, Black AE, Morris JA, Cole TJ. Between- and within-subject variation in nutrient intake from infancy to old age: estimating the number of days required to rank dietary intakes with desired precision. *Am. J. Clin. Nutr.* 1989; **50**: 155–67.

11 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.

12 Kaaks RJ. Biochemical markers as additional measurements in studies of the accuracy of dietary questionnaire measurements: conceptual issues. *Am. J. Clin. Nutr.* 1997; **65**: 1232S–9S.

13 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.

14 Schoeller DA. How accurate is self-reported dietary energy intake? *Nutr. Rev.* 1990; **48**: 373–9.

15 Bingham SA. Limitations of the various methods for collecting dietary intake data. *Ann. Nutr. Metab.* 1991; **35**: 117–27.

16 Spiegelman D, Schneeweiss S, McDermott A. Measurement error correction for logistic regression models with an 'alloyed gold standard'. *Am. J. Epidemiol.* 1997; **145**: 184–96.

17 Wong MY, Day NE, Bashir SA, Duffy SW. Measurement error in epidemiology: the design of validation studies I: univariate situation. *Stat. Med.* 1999; **18**: 2815–29.

18 Kaaks R, Riboli E, Sinha R. Biochemical markers of dietary intake. *IARC Sci. Publ.* 1997; 103–26.

19 Ferrari P, Kaaks R, Riboli E. Variance and confidence limits in validation studies based on comparison between three different types of measurements. *J. Epidemiol. Biostat.* 2000; **5**: 303–13.

20 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.

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

22 Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. *Am. J. Epidemiol.* 1986; **124**: 17–27.

23 Kaaks R, Riboli E, Slimani N. Pilot-phase studies on the accuracy of dietary intake measurements in the EPIC project: overall summary and evaluation of results. *Int. J. Epidemiol.* 1997; **26**: S26–36.

https://doi.org/10.1079/PHN2002380 Published online by Cambridge University Press