International Journal of Epidemiology
Vol. 26, No. 1(Suppl. 1)

Printed in Great Britain

Validation and Calibration of Dietary Intake Measurements in the EPIC Project: Methodological Considerations

RUDOLF KAAKS AND ELIO RIBOLI

Kaaks R (International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France). Validation and calibration of dietary intake measurements in the EPIC project: methodological considerations. *International Journal of Epidemiology* 1997; **26 (Suppl. 1):** S15–S25.

The statistical power of prospective studies on diet in relation to chronic disease risk can be improved by maximizing the variation in true intake levels actually distinguished—or 'predicted'—by dietary questionnaire assessments collected at baseline. This can be achieved by 1) developing a questionnaire method that provides measurements with the smallest possible random errors, thus maximizing the correlation of measured with true habitual intake levels; and 2) increasing the between-person variation in true dietary intake levels when combining multiple cohorts in populations with diverse consumption patterns. The first approach implies that, during the development or selection of the questionnaire method, correlations between measurements and true intake levels can be monitored; the second approach requires adjustment for between-centre differences in over- or underestimation of dietary questionnaire measurements. Besides optimizing the statistical power, it is important that the magnitude of the predicted variation in true intake level is estimated accurately, so as to allow unbiased estimations of relative risks. To meet these various objectives, substudies must be conducted for the 'validation' or 'calibration' of dietary questionnaire assessments, by comparison with additional measurements that have independent sources of error. This paper reviews the methodological considerations underlying the design and implementation of such substudies in the EPIC project, a collaborative multicentre study in nine Western European countries.

Keywords: prospective studies, diet, measurement error, validation, calibration, EPIC

Prospective cohort studies offer major methodological advantages for investigating the relation between diet, biochemical markers of nutritional or metabolic status, and cancer risk.[1] Since the incidence rates of individual forms of cancer are relatively low, however, this type of study must generally be very large—including several tens of thousands of individuals—for sufficient numbers of cases with specific forms of disease to accrue during 10 to 15 years' follow-up. As the cost of such large studies can be high, it is important to plan the study efficiently in order to obtain as much information as possible for a given investment of time and resources. In addition to the number of different exposures and disease outcomes measured, two major criteria for maximizing the informativeness of a study are the statistical power to test for associations between dietary exposures and disease risk, and the validity and precision with which the magnitude of such associations can be estimated.

A basic difficulty in epidemiological studies on diet is that measurements of individuals' habitual, long-term intake levels of foods and nutrients, usually obtained by means of a structured questionnaire, tend to have rather large errors. These errors entail substantial losses of statistical power, and often also cause bias in estimates of relative risk as a measure of diet-disease association. Substudies on the accuracy of the dietary exposure measurements can be used to improve the design of cohort studies (e.g. by selecting an optimal dietary questionnaire method), or to correct for the amount of bias in estimated measures of diet-disease association.[2] As the substudies may themselves require considerable investments of time and money, they should also be planned efficiently to obtain the best possible level of precision for a given investment.

In section 1 of this paper, we shall briefly present some basic model assumptions that are often made (at least implicitly) to estimate relative risk in epidemiological

International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France.

studies. In section 2, these basic assumptions are used to discuss the expected magnitudes of bias in relative risk estimates and of loss of statistical power, as a result of systematic or random errors in exposure measurements. In sections 3 and 4, we shall review key aspects of the use and design of substudies for the 'validation' or 'calibration' of dietary intake measurements in the context of EPIC, a multicentre cohort study on diet and cancer in nine Western European countries, and discuss the use of calibration studies in improving the between-centre comparability of dietary intake measurements.

MODEL ASSUMPTIONS

The estimation of relative risks in epidemiological studies is most often based on the assumption of a log-linear exposure-disease relation; that is, for a continuous exposure variable T,

$$\log(\text{disease rate at exposure level T}) = \gamma + \theta T \quad (1)$$

The slope of θ of this *'disease' model* represents the logarithm of the relative risk for a single unit increase in intake level. In a prospective cohort study, and given this disease model, the log-relative risk θ can be estimated by, for example, Poisson regression with disease status as a binary outcome variable. An unbiased estimation, however, requires that error-free measurements of exposure level T are available.

When the exposure level T (e.g. the individuals' habitual intake levels of a given nutrient) is measured with error (e.g. by means of a questionnaire), an unbiased estimation of the log-relative risk θ requires the simultaneous evaluation of the average magnitude of these errors, using additional dietary intake measurements obtained in a substudy. This simultaneous estimation process implies that additional model assumptions must be made about: i) types of measurement error that may occur (*'measurement' model*), and ii) the shape of the population distribution of true exposure values (*'exposure' model*).[3]

A very basic measurement model presents the individuals' dietary questionnaire measurements as the sum of true intake level (T) and an error (e_Q; we shall call this the 'total' measurement error):

$$Q = T + e_Q \quad (2a)$$

In many previous discussions of the effects of dietary assessment errors on estimation of relative risks, the errors e_Q were assumed simply to be random on a population level, with zero mean at any given value of T. It may be more realistic, however, to allow the mean of e_Q to be different from zero at the population level, so as to account for systematic over- or underestimations of intake. Furthermore, the average magnitude of the errors e_Q may depend on the individuals' true exposure levels T being measured; that is, there may be a covariance, γ, between the errors e_Q and true intake levels T of different individuals. A positive value for covariance γ then describes the subjects' tendency to overestimate more (or underestimate less; this will also depend on the value of the constant scaling parameter α_Q) if their true habitual intake level is higher than the population average, and to underestimate if their true intake level is lower; inverse tendencies are implied if the covariance γ is negative. Taking account of these additional assumptions, the measurement model can also be written as

$$Q = \alpha_Q + \beta_Q T + \varepsilon_Q \quad (2b)$$

Here, the coefficients α_Q and β_Q reflect constant and proportional '*scaling biases*', respectively, and subsume the parts of the individuals' total errors e_Q that can be taken as constant for all individuals, or that are linearly dependent (on a population level) of the individuals' true intakes level T. The term ε_Q—referred to as 'random' error—represents the remainder of the total error e_Q that is not constant for all individuals, and that is uncorrelated with T (i.e. $\text{Cov}(\varepsilon_Q, T) = 0$). The population (between-subject) variances of true intake levels and random errors will be denoted by σ_T^2 and $\sigma_{\varepsilon Q}^2$, respectively.

The population distributions of true, and measured intake are often assumed to be approximately normal. This joint normality, combined with a linear measurement model as in equation (2b), implies that the errors e_Q are also normally distributed, with constant variance regardless of the true intake level T. The assumptions of joint normality and of a linear measurement model, combined with the assumption that the overall disease rate is low, allow the expected bias in (log-) relative risk estimates and loss of statistical power due to measurement errors to be expressed by comparatively simple, closed-form equations. Moreover, these assumptions often also form the basis of statistical models to correct for bias in relative risks estimated from data actually observed, using information from substudies with additional dietary intake measurements.

In practice the population distributions of intake measurements often deviate from normality, which seems to be largely due to an increased error variance among subjects with higher true intake levels. Nevertheless, intake distributions of nutrients and major food groups can generally be normalized by some mathematical (e.g. logarithmic) transformation. Model assumptions (including that of linear relations between

measured and true intake levels) are then often assumed to apply to the transformed data.

Alternative statistical methods that do not require the assumption of joint normality of true and measured exposure distributions for the estimation of relative risks with correction for biases due to exposure measurement error have also been described. So far, however, these more complex methods have not found widespread use in nutritional epidemiology, and discussion of these methods is beyond the scope of this paper.

EFFECTS OF EXPOSURE MEASUREMENT ERRORS: BIAS AND LOSS OF POWER

Throughout this paper, exposure measurement errors are assumed to be of equal average magnitude for those who will eventually develop a given form of disease ('cases'), and those who do not ('controls'). In general, bias in regression estimates of log-relative risk θ will then arise only if, on average, over- or underestimations e_Q are different for individuals with high and low exposure measurements Q; that is, bias arises if the *total* measurement errors e_Q have either a negative or a positive covariance with the measurements Q.

Under the specific assumptions defined in the previous section (i.e. a linear measurement model, and joint normality of true and measured intake distributions), it can be derived that the regression of a binary disease outcome variable on measurements Q, rather than on true intake values T, will yield a log-relative risk estimate $\hat{\theta}^* = \lambda\theta$, biased by a factor λ with an expected value of, approximately

$$E(\lambda) = \frac{1}{\beta_Q}\rho^2_{QT} \qquad (3)$$

where

$$\rho_{QT} = \frac{1}{\sqrt{[1 + \sigma^2_{\varepsilon Q}/(\beta^2_Q\sigma^2_T)]}} \qquad (4)$$

is the correlation between questionnaire measurements and the individuals' true intake levels.[4]

The random part of the total measurement error, indicated by ε_Q in equation (2b) and uncorrelated with true intake level T, will always have a positive covariance with measurements Q; that is, positive random errors will be more frequent among individuals with intake measurements above the population mean, and negative errors will be more frequent among individuals with low intake measurements. This positive covariance between measurements Q and random measurement errors causes an underestimation of the slope θ, usually referred to as 'attenuation bias',[2] and indicated by the factor ρ^2_{QT} in equation (3).

The non-random part of the total measurement error can have either a positive or negative covariance with measurements Q, and may thus cause a bias in (log-) relative risk estimates that is either in the same direction, or opposite to the attenuation bias caused by the random part of the error. When the measurement errors have a positive covariance with true intake levels T, the proportional scaling factor β_Q will be greater than 1.0, and dietary questionnaire measurements above the population mean will be overestimated more than measurements below the mean. In this case, the bias in estimates of the log-relative risk θ caused by the proportional scaling error will go in the same direction as the attenuation bias caused by random errors (i.e. towards an underestimation of θ). When the measurement errors have a negative covariance with true intake levels T, the proportional scaling factor β_Q is smaller than 1.0 and its related bias in estimates of the log-relative risk θ will be opposite to that caused by the random errors ε_Q.

Wacholder[5] recently discussed the theoretical situation where the attenuation bias (by a factor ρ^2_{QT}) caused by the random parts of the measurement errors are cancelled out by an inverse, proportional scaling bias (i.e. $\beta_Q = \rho^2_{QT}$, so that the bias factor in equation (3) equals $\lambda = \rho^2_{QT}/\beta_Q = 1.0$). It can be shown (Appendix), that this is precisely the case where the covariance between the measurements Q and the total measurement errors e_Q equals zero. The total measurement errors e_Q are then also called '*berksonian*', after Berkson[6] who showed that in this particular situation the measurement errors cause no bias in estimates of the slope θ.

For dietary questionnaire measurements, the proportional scaling factor β_Q is generally expected to be rather close to 1.0, whereas the squared correlation ρ^2_{QT} is usually smaller than 0.50. Overall, therefore, the total errors e_Q in dietary questionnaire measurements are usually non-berksonian, and estimates of slope θ are biased predominantly towards zero by the attenuating effects of the random parts, ε_Q, of the measurement errors. A crucial observation, however, is that a multiplication of the scale of measurements Q by an arbitrary factor f changes the value of β_Q (to $f\beta_Q$), but leaves the value of the correlation ρ_{QT} unaffected. This implies that the measurement errors e_Q can always be *made* berksonian, by choosing the value of the rescaling factor f that will make β_Q equal to ρ^2_{QT}. In the Appendix it is shown that such a rescaling factor must be equal to the bias factor λ with the expected value indicated in equation (3). Thus, for questionnaire measurements rescaled to $Q' = \nu + \lambda Q$ (with an appropriately chosen value for ν), it can be shown that for any subgroup of

individuals the true intake level is, on average, equal to their rescaled questionnaire measurements (i.e. $E[T|Q'] = Q'$). Moreover, regressing outcome variable Y on rescaled measurements $Q' = \nu + \lambda Q$, rather than on original measurements Q, will yield (approximately) unbiased estimates of the slope θ. This method of re-scaling is equivalent to the method of 'linear approximation' described by Rosner *et al*.[7] for the correction of bias in logistic regression estimates of relative risk. In studies on diet, the method is increasingly being referred to as 'calibration', and the factor λ can thus also be called the 'calibration' factor.

Even in the situations where the bias factor λ equals 1.0—i.e. attenuation effects due to random errors ε_Q are balanced by a proportional scaling factor β_Q smaller than 1.0, so that the total measurement errors e_Q are berksonian and do not induce bias in regression estimates of log-relative risk θ—the presence of random errors ε_Q does affect the statistical power of a test for linear association between disease outcome Y and true intake level T (i.e. testing the null hypothesis that the log-relative risk θ equals zero). The power for this type of test can be shown to be a positive function of the variance in true dietary intake levels '*predicted*' by the questionnaire measurements,[8] where the predicted intake levels $X = E[T|Q]$ are defined as the mean true intake levels T (for groups of individuals) for given values of measurement Q. Under the specific model assumptions of section 1, the variance of the predicted intake levels can be shown to be equal to

$$\mathrm{Var}(E[T|Q]) = \rho_{QT}^2 \sigma_T^2 \qquad (5)$$

(Appendix). This equation confirms the common knowledge that, with decreasing values of the correlation ρ_{QT}, the variation in true intake levels distinguished correctly by the questionnaire measurements decreases, so that the statistical power progressively drops to zero. It must be noted that the variance of predicted intake levels is identical to the variance of perfectly calibrated questionnaire measurements $Q' = \nu + \lambda Q$ (since $E[T|Q] = E[T|Q'] = Q'$).[8]

VALIDATION AND CALIBRATION

As mentioned, the power of a statistical test for diet-disease association depends on the variation in true intake level actually distinguished by the dietary questionnaire measurements. Therefore, to optimize the power, the variation of predicted dietary intake levels should be made as large as possible. As indicated by equation (5), this can be achieved by maximizing the correlation ρ_{QT}, by maximizing the variance of true intake level σ_T^2, or by a combination of both approaches. In addition to optimizing the statistical power, it is desirable to obtain unbiased estimates of relative risk.

Developing a questionnaire that will yield dietary intake measurements with a high correlation ρ_{QT}, or selecting an optimal questionnaire amongst two or more candidate methods, implies that during this process the correlation ρ_{QT} can be estimated. Likewise, an unbiased estimation of the log-relative risk θ implies that the bias factor λ can be estimated. Both objectives—estimating the correlation coefficient ρ_{QT}, or estimating the bias factor λ—require the conduct of preliminary 'validity' substudies, or of 'calibration' substudies, in which questionnaire measurements are compared with additional dietary intake measurements obtained by independent methods. The practical design requirements for these two types of substudy differ somewhat, and are discussed in the next two subsections.

Selecting a Dietary Questionnaire: Preliminary Validity Studies

'Validation' is usually defined as the evaluation of whether a measuring instrument really measures what it is actually intended to. In studies on diet, however, a major practical complication is that no methods are available to obtain perfectly accurate measurements of the habitual intake levels of single individuals. Consequently, differences between measured and true intake levels cannot be evaluated for separate individuals, but one can estimate only to what extent *variation* in measurements reflects, on average, between-person variation in true, habitual intake levels. The proportion of the variance of questionnaire measurements Q that is associated with true intake level T is given by the square of the correlation coefficient ρ_{QT}, and the correlation coefficient ρ_{QT} is therefore also referred to as the 'validity coefficient' of questionnaire measurements.[2] The principal aim of preliminary validity studies in the EPIC project was to estimate this coefficient.

Especially because there are no methods for a fully accurate measurement of the true, habitual intake levels of foods or nutrients of free-living individuals, the analysis of dietary validity studies must be based entirely on model assumptions specifying the relation to true intake levels of the individuals' questionnaire measurements (as in equation (2b)), and of other types of measurement taken for the purpose of comparison. Plummer and Clayton[9,10] and Kaaks *et al*.[11] have discussed the design and analysis of dietary validity studies in general terms of latent variable models. It was concluded that estimation of the validity coefficient ρ_{QT} of dietary questionnaire measurements Q requires a comparison with at least two additional measurements (X_1, X_2) per person. A crucial assumption is that the

random errors in measurements Q,X_1,X_2 are independent, so that correlations between the measurements are due entirely to their association with the same, true intake variable.

In practice, the independence of random errors between the three measurements can only be assumed, but not proven. The likelihood that this assumption is valid can be increased, however, by taking measurements with different methods that have different (suspected) sources of error. The commonest approach to collecting the two (or more) additional measurements required for a dietary validity study is to obtain a number of replicate measurements of the actual daily food intakes at regular intervals during a one-year period, using weighed food records or 24-hour recalls. So far, validity studies have typically included about 100–200 subjects, using the average of 12–28 days of repeat daily food consumption records per person as reference measurements.[12,13] This was also the basic design of the validity studies conducted during the pilot phase of the EPIC project.[14] Following this study design, the validity coefficient ρ_{QT} is estimated from the correlation between questionnaire measurements and the individuals' averages of k daily intake records, R, with adjustment for the attenuating effects due to random, within-subject (day-to-day) variations in the latter.[15] It must be noted that in order to obtain valid results this approach requires random errors to be independent not only between questionnaire measurements and daily intake records, but also between the replicate daily intake measurements taken by the same method on the same individuals. Violation of the first assumption ($Cov(\varepsilon_Q,\varepsilon_R) \neq 0$) will lead to an overestimation of the validity coefficient ρ_{QT}; conversely, violation of the second assumption (i.e. $(Cov(\varepsilon_{Ri},\varepsilon_{Rj}) \neq 0$, for replicate measurements R_i, R_j taken on different occasions) will lead to an underestimation.

In the preliminary validity studies of the EPIC pilot phase, replicate samples of blood and (24-hour) urine were collected in addition to 24-hour recalls, for measurement of biochemical markers of dietary intake level.[16,17] Biochemical markers of diet have the appeal that variations uncorrelated with true intake level T (i.e. random 'errors' if we consider the markers as a measurement of intake) are likely to be truly independent of those of questionnaire measurements of habitual intake level. On the other hand, the correlation between marker and true intake level of a given food or nutrient is generally far from perfect, even when adjustments are made for attenuation due to within-person variations over time in the marker.[18] The observed sample correlation between marker and questionnaire measurements can thus generally be interpreted only as a *lower* limit for the correlation ρ_{QT}. Using structural equations models,[11] or an elementary factor analysis model,[18,19] another estimate of the validity coefficient ρ_{QT} can, however, be obtained from a triangular comparison between questionnaire measurements, records of daily intake and biochemical marker. The advantage of this triangular approach is that it avoids the assumption of independence between errors of replicate reference measurements taken by the same (recording) method on the same individuals. Nevertheless, the method still relies on the assumption of independence between the errors of different types of measurement. If one assumes that the only possible violation of model assumptions may be a certain level of correlation between random errors of questionnaire measurements and daily intake records, the estimated correlation ρ_{QT} can still be interpreted as an upper limit for the true validity coefficient.[18]

Estimating Predicted Intake Levels: Calibration

Besides selection of an optimal questionnaire method, using validity studies during a preliminary pilot phase, substudies with additional intake measurements are needed to estimate the magnitude of the calibration factor λ, to correct for bias in relative risk estimates. Substudies designed to estimate the factor λ or, which is equivalent, to estimate true intake levels predicted by the questionnaire measurements, are referred to as 'calibration' studies. The basic requirement for calibration studies is that for at least a subsample of the study population 'reference' measurements R must be available that are free of scaling bias (i.e. $R = T + \varepsilon_R$) with random errors ε_R that are independent of those of questionnaire measurements ($Cov(\varepsilon_R,\varepsilon_Q) = 0$). Under these assumptions, we can estimate the predicted intake level $E[T|Q]$ as the mean reference measurement conditional on the level of questionnaire measurement, $E[R|Q]$. Thus, questionnaire measurements are used to classify or rank individuals according to habitual dietary intake pattern, as characterized by the intake levels of specific food groups or nutrients, and additional reference measurements are used to estimate the mean true intake level of food or nutrients for subjects classified or ranked differently by the questionnaire. Assuming a continuous, linear relation between questionnaire measurements and true intake levels (as in equation (2b)), normal linear regression of R on Q can be used to estimate predicted intake levels as $X = E[R|Q] = \nu + \lambda Q = Q'$, which is equivalent to the rescaling (calibration) of questionnaire measurements mentioned in section 1 in this paper. Approximately unbiased estimates of the log-relative risk θ can be obtained by

regressing disease outcome Y on these estimates of predicted intake level.

More extensive discussions of the calibration method have been given elsewhere.[7,8,20] An additional note of interest may be that the estimates of predicted intake levels obtained by regression of reference measurements R on the questionnaire measurements can also be seen as 'empirical Bayes' ' estimates of exposure,[3] and that the calibration method has also been discussed in a more general context of Bayesian statistics, including the use of 'Gibbs sampling' methods.[21]

Weighed food records, or 24-hour recalls are generally thought to provide the best possible estimates of mean intakes of different food groups and nutrients at a population level, and thus seem to be optimal methods for taking reference measurements in a calibration study. It is important to realize that, with the assumed independence between the random errors of questionnaire and reference measurements (i.e. $\mathrm{Cov}(\varepsilon_R,\varepsilon_Q) = 0$), the random errors of reference measurements are not expected to cause bias in estimates of the calibration factor (attenuation bias is caused only by random errors in the predictor variable). The reference measurements therefore do not need to give precise evaluations of an individual's long-term, habitual intake level, and can be based even on a single day's food consumption record. Nevertheless, a sufficient number of records should be obtained, either by including a sufficiently large number of subjects in the calibration study, or by taking multiple records for each participant, for the calibration study to reach a minimum level of precision. It has been shown elsewhere that, for a given total number of daily food consumption records collected, the standard error of the estimated calibration factor λ will be smallest when the calibration study includes a maximum of individuals, with only a single record each.[22,23] In the EPIC study, where reference measurements are taken by 24-hour recall interview, this one-recall-per-person design is also optimal from a practical (logistic) and financial viewpoint: in many EPIC study centres 24-hour recalls can be obtained immediately when subjects come to a research centre to return a completed dietary questionnaire (mailed to them previously) and give a blood sample. Increasing the number of subjects in the calibration study when they make their first visit to a research centre is less expensive than re-inviting them for repeat interviews.

Calibration of dietary intake measurements will on average lead to approximately unbiased estimates of the log-relative risk θ, relating the probability of disease outcome Y to dietary exposure level. An unavoidable drawback of needing such bias correction procedure, however, is that the confidence intervals of adjusted θ-estimates obtained by regression of outcome Y on calibrated exposure measurements must be adjusted as well, to account for imprecision in the estimation of the calibration factor. Only in the theoretical case where the calibration factor λ is estimated with zero standard error is there no increase in the confidence interval of θ. In practice, however, the calibration factor λ is estimated with some degree of imprecision, and the confidence interval of a calibrated θ-estimate must be augmented to account for this imprecision (a closed-form equation for the adjusted variance of the calibrated θ-estimate has been derived by Rosner *et al*.).[7] A critical issue, therefore, is to decide how many subjects must be included in the calibration study for the factor λ to be estimated with a sufficiently small standard error. Ideally, calculations of sample size requirements would be based entirely on the criterion of a high relative efficiency, defined by the increase in width of the confidence interval for θ, compared to the confidence interval that would be obtained if the calibration was perfect (i.e. with absolute precision).[20] The use of this criterion alone may lead to excessive sample size estimations, however, if the confidence interval for the log-relative risk estimate θ before the calibration adjustment is expected to be very narrow (because there are strong relative risks for high *versus* low measured exposure levels, or because the expected numbers of cases are high).[23] An alternative is to use two complementary criteria of either a high relative efficiency, or a minimum value for a standardized test score computed as the calibrated θ-estimate divided by its (adjusted) standard error. Using these combined criteria, theoretical sample size requirements for calibration substudies depend only on the correlation between questionnaire and reference measurements.[23]

In the EPIC study, the minimum correlation between questionnaire measurements and a single 24-hour recall is expected to be around 0.20 for major nutrients (Table 1 gives results from the preliminary validity studies of the EPIC pilot phase). Given this minimum level of correlation, sample size requirements for calibration studies have been estimated to be around 4000 individuals per country, to obtain either a relative efficiency above 0.90, or a minimum value of 4.0 for a standardized score (and with roughly normal distribution) to test whether the calibrated log-relative risk estimate differs from zero. To optimize further the efficiency of the calibration studies, it was decided to recruit substudy participants by stratified sampling, weighting the numbers of participants proportionally to the expected cumulative total cancer incidence during a 10–15 years' follow-up in strata of age and sex.

TABLE 1 *Coefficients of correlation; between nutrient intake measurements obtained by dietary questionnaire and by a single 24-hour recall*

	Spain		France	Greece		Germany		Italy		Netherlands		UK
	Men	Women	Women	Men	Women	Men	Women	Men	Women	Men	Women	Women
Calories	0.53	0.46	0.24	0.45	0.27	0.34	0.20	0.15	0.15	0.42	0.37	0.30
% Energy from:												
Protein	0.33	0.41	0.31	0.28	0.20	0.27	0.28	0.27	0.23	0.43	0.32	0.38
Carbohydrates	0.46	0.46	0.37	0.23	0.10	0.38	0.34	0.25	0.34	0.51	0.47	0.53
Fat	0.26	0.36	0.24	0.12	0.07	0.30	0.18	0.14	0.17	0.41	0.31	0.35
Alcohol	0.73	0.49	0.38	0.34	0.34	0.50	0.55	0.56	0.55	0.57	0.61	0.75
Fibre	0.49	0.35	0.24	0.20	0.17	0.44	0.20	0.13	0.23	0.32	0.33	0.38
Vitamin C	0.40	0.40	0.24	0.16	0.11	0.08	0.14	0.07	0.23	0.23	0.25	0.35

CALIBRATION FOR BETWEEN-COHORT COMPARISONS

Besides developing a questionnaire method that yields dietary intake measurements with smallest possible random errors, a complementary way of increasing the variance of predicted dietary intake levels, thereby improving the power of statistical tests for association between dietary intake levels and disease risk, is to augment the between-person heterogeneity, σ_T^2, of true dietary exposure levels.

White *et al.*[24] recently discussed the fact that choosing a study population with a larger heterogeneity of dietary consumption patterns may actually have a double advantage. The first is that a wider range of true intake levels and an associated wider range of disease risks can be detected, which will increase the statistical power even when exposure levels are measured with absolute precision. The additional advantage is that the proportion of the variation in true intake level that is measured accurately (i.e. the correlation ρ_{QT} between measured and true intake levels) will also be increased. The latter can be seen directly from equation (4), which shows that the correlation ρ_{QT} depends on the ratio of the between-subject variances of random errors ($\sigma_{\varepsilon Q}^2$), and of true intake levels (σ_T^2). It must be noted, however, that White *et al.* assume that the variation in the magnitude of measurement errors does not also increase when a more heterogeneous study population is chosen. This assumption may not always be realistic. The heterogeneity of true dietary exposure levels may be increased, for example, by including individuals of different ethnic origin. A single type of questionnaire may not then provide equally accurate measurements of intake in different subgroups of individuals consuming very different types of food.

The very rationale of the EPIC project as a multicentre European study was to increase the total variation in true dietary exposures by including subjects from different regions in Europe, with diverse dietary intake patterns. Plummer *et al.*[20] and Kaaks *et al.*[22] have described how, under mild simplifying assumptions, the information obtained from such multicentre cohort studies can be decomposed into:

a. estimated relations between dietary exposures and disease risk at the individual level, within each of the study centres (cohorts) separately; and
b. an estimated 'ecological' relation between mean exposure measurements and average disease incidence at a cohort (group) level.

The evidence for diet-disease association in the form of intra-cohort relative risk estimates is strengthened by the overall increase in sample size, whereas the increased heterogeneity in exposure level by inclusion of diverse populations is captured by the between-cohort, ecological relation. Ideally, the within- and between-cohort estimates of relative risk corroborate one another, and may then be combined into a more powerful summary value. For optimal validity of this combined analysis it is obviously important to standardize measurements of dietary exposures and potential confounding factors carefully, and to take proper account of potential confounding or statistical interaction effects in the analysis. It must be noted that, in contrast to traditional ecological studies where estimates of exposure level, potential confounders and disease incidence rates are available only at an aggregate, population level, in a multicohort design these measurements can all be obtained at the level of individuals. Effects of

confounding of the ecological relations in a multicentre cohort study can therefore in principle be adjusted for properly in the analysis.

A possible complication in multicentre studies on diet is that dietary questionnaire measurements may not predict true intake differences with equal accuracy in all centres. Although a similar type of dietary questionnaire is used in the various EPIC study centres, the number and detail of questions about consumption of specific foods must be adapted to local habits, as dietary patterns and language vary substantially between countries. Moreover, the correlation ρ_{QT} between questionnaire assessments and true intake levels depends on the between-person variance of true intake level σ_T^2 (equation (4)), which may also differ between populations. There may thus be variation between study centres in the degree of bias in relative risk estimates induced by dietary assessment errors. Such distortions can, however, be corrected for by conducting dietary calibration studies, based on well-standardized reference measurements collected in a representative subsample of each main study population. Within individual cohorts, (log-)relative risk estimates can be adjusted for bias, after estimation of the appropriate calibration factors λ. If the calibration factors λ are estimated with sufficient accuracy, these adjustments may reduce between-cohort heterogeneity in relative risk estimates caused by dietary assessment errors.

The between-cohort, ecological relation may also be seriously distorted by measurement errors leading to differences in systematic over- or underestimation of mean dietary intake levels. Again, however, the ecological relation can be restored by substituting mean, standardized reference measurements for the mean dietary intake measurements obtained by the baseline questionnaires, thereby correcting for (or reducing) between-cohort differences in over- or underestimation of intake level.

DISCUSSION

This paper reviews methodological aspects of the use and design of substudies to evaluate the accuracy of dietary intake measurements in prospective cohort studies. The two central issues discussed are:

a. How to *maximize* the variation in the true intake levels of specific nutrients or food groups predicted by questionnaire measurements collected at baseline; this issue relates to obtaining an optimal power for statistical tests whether or not there are specific diet-disease associations.

b. How to estimate efficiently, and with sufficient accuracy, the *magnitude* of the variation in predicted intake level; this is equivalent to the question of how to estimate efficiently the calibration factor λ with a given level of precision, and relates to the correction for bias in relative risk estimates.

In the EPIC project, two types of substudy have been, or are being, conducted to address these issues in practice, namely 1) preliminary validity studies, conducted before the actual recruitment of the main study cohorts was started; and 2) calibration studies, conducted on a random subsample of cohort members, after their actual recruitment.

The principal objective of the preliminary validity studies conducted during the EPIC pilot phase (the results of which are presented in this Supplement) was to evaluate whether candidate questionnaire methods would measure a reasonable proportion of the between-person variation in true dietary intake level in a given study population. After this early stage evaluation, questionnaires could still be modified before actually being used in the main cohort studies. A secondary objective of the pilot-phase validity studies was to allow the various EPIC research centres to gain experience with the 24-hour recall, and to work on its standardization as a reference method for subsequent calibration studies. The data collected by the 24-hour recall method during the preliminary validity studies proved extremely useful for the development of a special computer software ('EPIC-SOFT'), designed to standardize the structure of the interview, and the number and detail of questions about foods consumed.

An essential difference in design requirements between validity studies (for estimation of the correlation ρ_{QT}) and calibration studies (for estimation of true intake levels predicted by questionnaire measurements), is that the former must be based on at least two additional intake measurements (e.g. two 24-hour recalls) per person, whereas only a single additional measurement (e.g. a 24-hour recall) is needed for the latter. In the validity studies of the EPIC pilot phase, the number of replicate 24-hour recalls per person was increased to 12. This was done to improve the precision of the validity studies, and because increasing the number of recalls to 12 per person was a logistically and financially more efficient way to improve the precision of estimated validity coefficients than increasing the numbers of study participants, with only two recalls each. By contrast, as the calibration studies are conducted on subgroups of cohort members with previously completed dietary questionnaires, the optimal (most cost-efficient) design of this type of substudy includes a larger number

of individuals with only a single reference measurement each.

An important advantage of calibration studies based on only a single 24-hour recall per person is that these may be conducted more easily on a truly representative subsample of cohort members. For calibration, such representativeness is indeed crucial, as the objective is to estimate in the substudy a mathematical function by which baseline questionnaire measurements can be translated (rescaled) into predicted true intake levels, and to use this function to correct for bias in crude relative risks estimated in the full cohort. In the preliminary validity studies of the EPIC pilot phase, where subjects were asked to comply with a much more intense schedule of dietary intake assessment, using 12 24-hour recalls and multiple samples of blood and urine, some self-selection of participants may have occurred. Thus, if those who volunteered to take part in the validity studies had a more than average motivation to respond accurately to dietary questionnaires, the accuracy of questionnaire measurements as estimated from the preliminary validity studies may have been overstated compared to the accuracy of questionnaire measurements in the main study cohort. Nevertheless, to the extent that the validity study is used only to develop an optimal questionnaire method, or to select an optimal version amongst several candidates, representativeness may be a less stringent requirement for preliminary studies, assuming that relative differences in the accuracy of methods are similar in the substudy and in the main study population.

The 24-hour recall method is considered ideal for intercultural comparisons of mean dietary intake levels, as it is an essentially open-ended method which allows a detailed reporting of amounts of very heterogeneous types of food or dishes.[25] Compared to weighed food consumption records, advantages of the 24-hour recall method are that participation rates are generally very high, and that the interviewer can monitor the completeness and quality of the subjects' responses, and elicit more detailed answers if needed. Nevertheless, it may be over-optimistic to assume, as was done in this paper, that 24-hour recalls provide truly unbiased measurements of mean dietary intakes at group level: underestimations may occur if subjects omit to report foods they have actually consumed. If this type of systematic over- or underestimation occurs, and particularly if these errors translate into a proportional scaling bias (i.e. $\beta_R \neq 1.0$), calibration will not transform errors in questionnaire measurements into truly berksonian error, and regression of disease outcome on the calibrated questionnaire measurements will not result in truly unbiased estimates of relative risk. In multicentre studies such as the EPIC project, however, the first objective of calibration is to estimate disease risk as a function of dietary intake differences expressed on a *similar* scale of measurement in all cohorts; that is, the aim is to improve the between-cohort *comparability* of relative risk estimates, and to improve the precision of and estimated ecological relation between mean intake levels and mean disease incidence rates. For this more limited objective, it is sufficient to assume that, to the extent that constant or proportional scaling biases occur in the 24-hour recalls, these biases will be of a relatively constant magnitude in all study cohorts.

Another possible complication, which can be considered only very briefly here, is that the assumption of independence of random errors of questionnaire and reference measurements (ε_Q and ε_R) may be violated. As mentioned, the independence of errors in practice can only be assumed, not proven. To increase the likelihood that the assumption is valid, a reference method should be chosen that has different suspected sources of error than the questionnaire used in the full cohort. Questionnaire measurements and 24-hour recalls both rely on an individual's capacity to remember and describe food consumption carefully. Nevertheless, the mental processes related to the long-term recall of average food consumption patterns, or to the very short-term recall of *actual* food consumption on the previous day are believed to be quite different.[26]

Ideally, reference measurements in calibration studies should be based on less subjective measures of dietary intake such as a biochemical marker. Although many biochemical markers measured in blood or other tissues are known to have some level of correlation with intake levels of specific nutrients or foods,[16,17] however, their quantitative relations with absolute daily intake levels are often unknown; that is, assuming that, for example, this relation is described well by the linear model

$$M = \alpha_M + \beta_M T + \varepsilon_M \qquad (7)$$

the values of the scaling factors α_M and β_M are unknown and may vary between populations. Most biochemical markers cannot therefore be used as reference measurements for the calibration of dietary intake measurements. An exception is the 24-hour urinary nitrogen excretion, which can be translated into an estimate of absolute level of protein intake.[27] In the EPIC study, 24-hour urines are being collected in subgroups of calibration study participants. Intake levels of protein estimated from the amounts of nitrogen excreted in these urine samples will be used to monitor the accuracy of between-country standardization of 24-hour recalls as a common measurement for calibration.

REFERENCES

[1] Riboli E. Nutrition and cancer: background and rationale of the European Prospective Investigation into Cancer and Nutrition (EPIC). *Ann Oncol* 1992; **3:** 783–91.

[2] Armstrong B, White E, Saracci R. *Principles of Exposure Measurement in Epidemiology*. Oxford: Oxford Medical Publications, 1992, pp. 63–64.

[3] Clayton D. Models for the analysis of cohort and case-control studies with inaccurately measured exposures. In: Dwyer J H (ed.). *Statistical Models for Longitudinal Studies on Health*. Oxford: Oxford University Press, 1988, pp. 1167–78.

[4] Cochran W G. Errors of measurement in statistics. *Technometrics* 1968; **10:** 637–66.

[5] Wacholder S. When measurement errors correlate with truth: surprising effects of non-differential misclassification. *Epidemiology* 1995; **6:** 157–61.

[6] Berkson J. Are there two regressions? *J Am Stat Assoc* 1950; **45:** 164–80.

[7] Rosner B, Willett W C, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989; **8:** 1051–69.

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

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

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

[11] Kaaks R, Riboli E, Estève J, van Kappel A L, van Staveren W A. Estimating the accuracy of dietary questionnaire assessments: Validation in terms of structural equation models. *Stat Med* 1994; **13:** 127–42.

[12] Riboli E, Elmstahl S, Saracci R, Gullberg B, Lindgärde F. The Malmö food study: Validity of two dietary assessment methods for measuring nutrient intake. *Int J Epidemiol* 1997; **26 (Suppl. 1):** S161–S173

[13] Willett W C, Sampson L, Stampfer M J *et al*. Reproducibility and validity of a semi-quantitative food frequency questionnaire. *Am J Epidemiol* 1985; **122:** 51–65.

[14] Riboli E, Kaaks R. The EPIC project: rationale and study design. *Int J Epidemiol* 1997; **26 (Suppl. 1):** S6–S14.

[15] Rosner B, Willett B. Interval estimates for correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing. *Am J Epidemiol* 1988; **127:** 377–86.

[16] Hunter D. Biochemical indicators of dietary intake. In: Willett W. *Nutritional Epidemiology*. New York: Oxford University Press, 1990, pp. 143–216.

[17] Riboli E, Rönnholm M, Saracci R. Biological markers of diet. *Cancer Surveys* 1987; **6:** 685–718.

[18] Kaaks R. Biochemical markers as an additional measurement in studies on the accuracy of dietary questionnaire measurements. *Am J Clin Nutr* (In Press), 1997.

[19] Ocké M, Kaaks R. Biomarkers as additional measurement in dietary validity studies: experiences with data from the EPIC study. *Am J Clin Nutr* (submitted).

[20] Plummer M, Clayton D, Kaaks R. Calibration in multicentre cohort studies. *Int J Epidemiol* 1994; **23:** 419–26.

[21] Richardson S, Gilks W R. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. *Am J Epidemiol* 1993; **138:** 430–42.

[22] Kaaks R, Plummer M, Riboli E, Estève J, van Staveren W A. Adjustment of bias due to errors in exposure assessments in multicenter cohort studies on diet and cancer: a calibration approach. *Am J Clin Nutr* 1994; **49:** 245S–50S.

[23] Kaaks R, Riboli E, van Staveren W A. Sample size requirements for calibration studies of dietary intake measurements in prospective cohort investigations. *Am J Epidemiol* 1995; **142:** 557–65.

[24] White E, Kushi L H, Pepe M S. The effect of exposure variance and exposure measurement error on study sample size: implications for the design of epidemiologic studies. *J Clin Epidemiol* 1994; **47:** 873–80.

[25] Witschi J C. Short-term recall and recording methods. In: Willett W C. *Nutritional Epidemiology*. New York: Oxford University Press, 1990, pp. 53–68.

[26] Cameron M E, van Staveren W A (eds). *Manual on Methodology for Food Consumption Studies*. Oxford: Oxford University Press, 1988.

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

APPENDIX

Assume questionnaire measurements are related to true intake level as

$$Q = T + e_Q$$

with $Var(T) = \sigma_T^2$, $Var(e_Q) = \sigma_{eQ}^2$, $Cov(T, e_Q) = \phi$. Following this model notation, the correlation between measured and true intake level can be written as

$$\rho_{QT} = Cov(Q,T)/\sqrt{[Var(Q)\ Var(T)]}$$

$$= (\sigma_T^2 + \phi) / \sqrt{[\sigma_T^2(\sigma_T^2 + 2\phi + \sigma_{eQ}^2)]}$$

As explained in the main text, the model can be rewritten as

$$Q = \alpha_Q + \beta_Q T + \varepsilon_Q$$

with

$$\beta_Q = 1 + Cov(e_Q, T)$$

$$= (\sigma_T^2 + \phi) / \sigma_T^2$$

Now suppose we transform the measurements Q into

$$Q' = v + \lambda Q$$

$$= v + \lambda(T + e_Q)$$

$$= v + T + (\lambda - 1)T + \lambda e_Q$$

$$= v + T + e_Q^*$$

(with $e_Q^* = (\lambda - 1)T + \lambda e_Q$)

We wish to determine λ so that $Cov(Q', e_Q^*) = 0$, that is,

$$Cov(Q',e_Q^*) = Cov(v + T + e_Q^*, e_Q^*)$$

$$= Cov(v + T, (\lambda - 1)\ T + \lambda e_Q) + Var((\lambda - 1)T + \lambda e_Q)$$

$$= (\lambda - 1)\sigma_T^2 + \lambda\phi + (\lambda - 1)^2\sigma_T^2 + \lambda^2\sigma_{eQ}^2 + 2(\lambda - 1)\lambda\phi$$

$$= \lambda\ [(\lambda - 1)(\sigma_T^2 + 2\phi) + \phi + \lambda\sigma_{eQ}^2]$$

$$= \lambda\ [\lambda(\sigma_T^2 + 2\phi + \sigma_{eQ}^2) - \sigma_T^2 - \phi]$$

$$= 0$$

Solutions to the last equation are $\lambda = 0$—a meaningless outcome if one wishes to correct for error by rescaling of measurements, as it implies a total absence of association between measured and true intake levels—or

$$\lambda = (\sigma_T^2 + \phi) / (\sigma_T^2 + 2\phi + \sigma_{eQ}^2)$$

$$= (\sigma_T^2 + \phi)^2 / [\sigma_T^2(\sigma_T^2 + 2\phi + \sigma_{eQ}^2)]) (\sigma_T^2 / (\sigma_T^2 + \phi))$$

$$= \rho_{QT}^2 / \beta_Q$$

The latter solution is the bias factor mentioned in equation (3) of the main text.

The variance of predicted intake levels is equal to the variance of perfectly calibrated questionnaire measurements; that is

$$Var(E[T|Q] = Var(E[T|Q']) = Var(Q')$$

$$= Var(v + \lambda Q)$$

$$= \lambda^2\ Var(Q)$$

$$= ((\sigma_T^2 + \phi) / (\sigma_T^2 + 2\phi + \sigma_{eQ}^2))^2\ (\sigma_T^2 + 2\phi + \sigma_{eQ}^2)$$

$$= (\sigma_T^2 + \phi)^2 / (\sigma_T^2 + 2\phi + \sigma_{eQ}^2)$$

$$= ((\sigma_T^2 + \phi)^2 / [\sigma_T^2(\sigma_T^2 + 2\phi + \sigma_{eQ}^2)]) \sigma_T^2$$

$$= \rho_{QT}^2 \sigma_T^2$$