STATISTICS IN MEDICINE, VOL. 8, 1051–1069 (1989)

CORRECTION OF LOGISTIC REGRESSION RELATIVE RISK ESTIMATES AND CONFIDENCE INTERVALS FOR SYSTEMATIC WITHIN-PERSON MEASUREMENT ERROR

B. ROSNER

Channing Laboratory, Department of Preventive Medicine and Clinical Epidemiology, Harvard Medical School, and Brigham and Women's Hospital, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

W. C. WILLETT

Department of Epidemiology, Harvard School of Public Health, Channing Laboratory, Harvard Medical School, and Department of Medicine, Brigham and Women's Hospital, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

AND

D. SPIEGELMAN

Channing Laboratory, Harvard Medical School, Departments of Biostatistics and Epidemiology, Harvard School of Public Health, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

SUMMARY

Errors in the measurement of exposure that are independent of disease status tend to bias relative risk estimates and other measures of effect in epidemiologic studies toward the null value. Two methods are provided to correct relative risk estimates obtained from logistic regression models for measurement errors in continuous exposures within cohort studies that may be due to either random (unbiased) within-person variation or to systematic errors for individual subjects. These methods require a separate validation study to estimate the regression coefficient λ relating the surrogate measure to true exposure. In the linear approximation method, the true logistic regression coefficient β^* is estimated by $\hat{\beta}/\hat{\lambda}$, where $\hat{\beta}$ is the observed logistic regression coefficient based on the surrogate measure. In the likelihood approximation method, a second-order Taylor series expansion is used to approximate the logistic function, enabling closed-form likelihood estimation of β^*. Confidence intervals for the corrected relative risks are provided that include a component representing error in the estimation of λ. Based on simulation studies, both methods perform well for true odds ratios up to 3·0; for higher odds ratios the likelihood approximation method was superior with respect to both bias and coverage probability. An example is provided based on data from a prospective study of dietary fat intake and risk of breast cancer and a validation study of the questionnaire used to asses dietary fat intake.

KEY WORDS Biometry Epidemiologic methods Nutrition

INTRODUCTION

It has long been appreciated that error in the measurement of individual exposure that is random with respect to disease status will tend to bias estimates of relative risk and other measures of association towards the null value in most commonly considered instances.[1] Errors in exposure measurement that are unrelated to disease status can be classified into two general types: (a) random (unbiased) within-person error and (b) systematic (biased) within-person error.

0277–6715/89/091051–19$09.50

Unbiased within-person error results from fluctuations in the level of exposure over time that may be due to biological and behavioural variation or to measurement error. The key feature is that the law of large numbers applies; if many replicate measurements n are obtained for an individual, then the mean of these replicates will provide an unbiased estimate of true exposure, and for large n will approach that subject's true long-term exposure. It has been assumed that errors in the measurement of blood pressure,[2] dietary intake[3,4] and urinary sodium excretion[5] are of this type.

If systematic within-person error is present, then unbiasedness does not hold and the mean of many repeated measures will not necessarily converge toward a person's true intake. For example, alcohol intake may be consistently under-reported by some, but not necessarily all, participants. If this error applies to all subjects equally (that is, all individuals over-report or under-report by a constant amount), it will not distort measures of association based on relative rankings, nor will it affect the statistical power of a study. However, systematic within-person error can vary among subjects, that is, repeated measures within a subject may be consistently too high or too low, but the degree and direction of this consistent error varies from person to person.

The use of dietary questionnaires is likely to result in systematic within-person error since repeated administrations may consistently under- (or over-) estimate intake due to conscious or unconscious under- (or over-) reporting. This may occur for many reasons, such as consistent misinterpretation of a question by a subject, or the omission of a food item from the questionnaire that is important for some individuals. Actually, most variables are probably subject to both random and systematic within-person error. For example, blood pressure measurements may have a random error component, but it is likely that certain individuals have systematically high (or low) measurements due to anxiety about the measurement or to anatomical features of their arm. Estimates of random within-person error can be obtained simply from one or more replicate measurements;[2-4] however, estimates of systematic within-person error must be obtained by comparison of the observed (surrogate) measurement with an independent measure of true exposure, that is, a measure of validity rather than reproducibility.

Methods to correct estimates of effect for measurement error have been described for simple and partial correlation coefficients,[6] logistic regression coefficients,[7-15] log-linear models,[16-19] proportional hazards models[20] and relative risk in the context of a 2×2 contingency table.[21,22] In many of these methods, it is assumed that error is purely due to random (unbiased) within-person variation; the possibility of systematic within-person error is generally ignored. In addition, derivations of standard errors or confidence intervals have generally not been provided, and when given, they have not considered error in the assessment of reproducibility (or validity). Confidence intervals are of particular interest when the point estimate is near the null value, since the width of the interval is then of primary interest.

In this paper, we provide two methods to correct logistic regression coefficients for error in the measurement of exposure, based on an independent assessment of validity. These methods incorporate correction for both random and systematic within-person error. In addition, confidence intervals for the corrected logistic regression coefficient are provided that include a component due to error in the measurement of validity. Examples are given using a prospective study of fat intake and risk of breast cancer.

METHODS

Suppose the model relating a single-dimensional true exposure X and the probability of disease D is of logistic form, whereby

$$\ln [Pr(D|X)/Pr(\bar{D}|X)] = \alpha^* + \beta^* X. \quad (1)$$

Furthermore, suppose that a linear relationship is assumed to exist between true exposure X and observed exposure Z that is of the form

$$X = \alpha' + \lambda Z + \varepsilon, \qquad \text{where } \varepsilon \sim N(0, \sigma^2). \tag{2}$$

Finally, we assume that the conditional distribution of X given Z and the marginal distribution of Z are the same for the main and validation study populations, and the conditional distribution of Z given X is the same for both diseased and non-diseased subjects (that is, $Pr(D|X, Z) = Pr(D|X)$). Our objective is to determine β^* using a data base consisting of (a) a main study population with only observed Z exposure levels, and (b) a validation study population, distinct from (a), where both true X and observed Z exposure levels are available. We will consider two approaches to achieve this objective.

Approach 1: linear approximation method

One intuitive and straightforward procedure is to use single imputation,[8,9,15] whereby we

(i) Estimate the parameters in equation (2) from the validation study population by ordinary least squares
(ii) Calculate an expected value of X ($E(X|Z) = \alpha' + \lambda Z$) for each person in the main study population based on the results in (i)
(iii) Substitute the expected value, $E(X|Z)$, for X for each person in the main study population and estimate β^* and associated confidence limits using equation (1) by ordinary logistic regression methods.

We denote this estimator by $\hat{\beta}_{imp}$. One problem with this method is that confidence intervals for β^* will be too narrow since the above method ignores errors in the estimation of X from Z.

Another approach which yields the same point estimate as the single imputation method, but where the imprecision of estimation of X given Z is reflected in the estimation of confidence intervals for β^*, is given as follows:

(i) Obtain an estimate of the logistic regression of disease on observed exposure (denoted by $\hat{\beta}$) from the main study
(ii) Obtain an estimate of the regression coefficient of X on Z from equation (2) using ordinary least squares based on the validation study data (denoted by $\hat{\lambda}$)
(iii) Estimate β^* by

$$\hat{\beta}_{lin} = \hat{\beta}_{imp} = \hat{\beta}/\hat{\lambda}. \tag{3}$$

We note that if the sample disease probability is small, then

$$Pr(D|X) \cong \exp(\alpha^* + \beta^* X)$$

$$Pr(D|Z) \cong \exp(\alpha^* + \beta^* \alpha' + \sigma^2 \beta^{*2}/2 + \beta^* \lambda Z). \tag{4}$$

Thus, $\hat{\beta}$ obtained from the 'naive' logistic regression of D on Z is an approximate maximum likelihood estimate of $\beta^*\lambda$, and division of $\hat{\beta}$ by the maximum likelihood estimate of λ can be expected to yield a consistent estimate of β^*.

Furthermore, if f is the conditional density of X given Z, then one can express the probability of disease conditional on Z in the form

$$\begin{aligned} Pr(D|Z) &= \int Pr(D|X)f(X|Z)\,dX \\ &= E\{\exp(\alpha^* + \beta^* X)/[1+\exp(\alpha^* + \beta^* X)]|Z\} \\ &= \int \{\exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z + \beta^* v)/[1+\exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z + \beta^* v)]\} \\ &\quad \times [\exp(-0{\cdot}5v^2/\sigma^2)/(\sqrt{(2\pi)}\sigma)]\,dv. \end{aligned} \tag{5}$$

Although it is impossible to express the integral on the right-hand side of equation (5) in closed form, one can expand $\ln[Pr(D|Z)/Pr(\bar{D}|Z)]$ as a first-order Taylor series about $\beta^* = 0$, whereby one obtains the approximation

$$Pr(D|Z) = \exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z)/[1+\exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z)]. \tag{6}$$

Finally, as pointed out by Armstrong,[9] the 0th-order Taylor series expansion of equation (5) around $\alpha^* + \beta^* E(X|Z)$ gives

$$Pr(D|Z) \cong \frac{\exp[\alpha^* + \beta^* E(X|Z)]}{1+\exp[\alpha^* + \beta^* E(X|Z)]} = \frac{\exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z)}{1+\exp(\alpha^* + \beta^*\alpha' + \beta^*\lambda Z)}. \tag{7}$$

Thus, based on the results in equations (4), (6) and (7), we would expect the linear approximation method to perform well when the disease is rare, when the absolute value of β^* is not large, or when measurement error is not too severe (that is, $\text{var}(X|Z)$ is small).

This solution provides the same estimate of β^* as in the imputation method described above; however, using this representation, confidence limits can be readily derived that take into account error in the estimation of λ (see below). One would expect that the linear approximation method should substantially reduce the bias obtained by using the uncorrected estimator $\hat{\beta}$, which is the observed logistic regression coefficient of disease on the surrogate measure Z, as an estimator of β^*.

Confidence interval and associated odds ratio for β^* based on the linear approximation estimator

Suppose that one obtains estimates of $\hat{\beta}$ for β from the main study and of $\hat{\lambda}$ for λ from the validation study. For the purpose of this development, we assume that the estimates of β and λ are independent random variables. To obtain the variance for the associated estimator $\hat{\beta}_{\text{lin}} = \hat{\beta}/\hat{\lambda}$, we use the delta method[23] and obtain

$$\text{var}(\hat{\beta}_{\text{lin}}) = (1/\hat{\lambda}^2)\text{var}(\hat{\beta}) + (\hat{\beta}^2/\hat{\lambda}^4)\text{var}(\hat{\lambda}), \tag{8}$$

where the estimates of $\text{var}(\hat{\beta})$ and $\text{var}(\hat{\lambda})$ are obtained using standard methods from the logistic regression model (1) and the linear regression model (2) respectively. It follows that 100% $X(1-\alpha)$ confidence limits for β^* and the associated odds ratio $\exp(\beta^*)$ are given respectively by

$$\hat{\beta}_{\text{lin}} \pm Z_{1-\alpha/2}\text{SE}(\hat{\beta}_{\text{lin}}) \tag{9}$$

and

$$\exp[\hat{\beta}_{\text{lin}} \pm Z_{1-\alpha/2}\text{SE}(\hat{\beta}_{\text{lin}})].$$

In some instances, the estimate X of truth μ_X may be subject to random (unbiased) within-person error, such as when the average of a small number of replicate measures is used to estimate true

exposure, whereby $X=\mu_X+e_X$, where $e_X \sim N(0, \sigma_X^2)$. Since the estimate of truth X is utilized in equation (2) as the dependent variable, $\hat{\lambda}$ still provides an unbiased estimate of the regression coefficient of μ_X on Z. However, random error inherent in X will result in an increased variance of $\hat{\lambda}$ and in a wider confidence interval for the corrected relative risk.

It should be noted that in the special circumstance when $\hat{\beta}=0$ (that is, the estimated odds ratio = 1·0), error in the estimation of λ will not be incorporated in the corrected confidence limits, since the second term of equation (8) will be zero.

Approach 2: likelihood approximation method

We have also considered a likelihood-based approach in an attempt to reduce the bias in the uncorrected estimator. We note from equation (5) that $Pr(D|Z)$ cannot be expressed in closed form if $Pr(D|X)$ is of logistic form and $f(X|Z)$ is normal. To address this problem, we write the logistic function in equation (1) in the form $\ln[Pr(D|X^*)/Pr(\bar{D}|X^*)]=(\alpha^*+\beta^*\bar{X})+\beta^*X^*$, where $X^*=X-\bar{X}$, and $\bar{X}$ is the sample mean of true exposure in the validation study. We then approximate $\ln[Pr(D|X^*)]$ by a second-order Taylor series expansion about $X^*=0$, whereby we have

$$Pr(D|X^*)\cong c_0\exp(c_1X^*-c_2X^{*2}), \tag{10}$$

where

$$c_0=\exp(\alpha^*+\beta^*\bar{X})/[1+\exp(\alpha^*+\beta^*\bar{X})]$$
$$c_1=\beta^*/[1+\exp(\alpha^*+\beta^*\bar{X})]$$
$$c_2=c_1^2c_0/[2(1-c_0)].$$

The advantage of this approach is that the approximate logistic function and the normal density are in the same form in equation (5). If we reparameterize equation (2) in terms of $X^*=X-\bar{X}$ and $Z^*=Z-\bar{Z}$, where $\bar{Z}$ is the sample mean of observed exposure in the validation study, then we have $X^*=\lambda Z^*+\varepsilon$, and upon integration in equation (5) one can obtain a closed-form approximation for $Pr(D|Z^*)$ as follows:

$$\hat{Pr}(D|Z^*)\cong c_0'\exp(c_1'z^*-c_2'z^{*2}), \tag{11}$$

where

$$c_0'=[c_0/(1+2\sigma^2c_2)^{1/2}]\exp[(\sigma^2c_1^2/2)/(1+2\sigma^2c_2)]$$
$$c_1'=c_1\lambda/(1+2\sigma^2c_2)$$
$$c_2'=\lambda^2c_2/(1+2\sigma^2c_2).$$

We wish to estimate c_0', c_1', c_2' and thereby c_0, c_1, c_2 from equation (11) and then β^* from equation (10). For this purpose, we note that if we fit a logistic regression of disease probability as a function of Z^* and use a second-order Taylor series expansion for $\ln[Pr(D|Z^*)]$, then it follows that

$$\hat{\hat{Pr}}(D|Z^*)=\exp(\alpha+\beta\bar{Z}+\beta Z^*)/[1+\exp(\alpha+\beta\bar{Z}+\beta Z^*)]$$
$$\cong c_0''\exp(c_1''Z^*-c_2''Z^{*2}), \tag{12}$$

where

$$c_0''=\exp(\alpha+\beta\bar{Z})/[1+\exp(\alpha+\beta\bar{Z})]$$
$$c_1''=\beta/[1+\exp(\alpha+\beta\bar{Z})]$$
$$c_2''=c_1''^2c_0''/[2(1-c_0'')].$$

By equating (c_0', c_0''), (c_1', c_1'') and (c_2', c_2'') in equations (11) and (12), and solving for c_0, c_1 and c_2, we obtain

$$\hat{c}_2 = \hat{\beta}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})/\{2\hat{\lambda}^2[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})]^2 - 2\hat{\beta}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})\hat{\sigma}^2\} \tag{13}$$

$$\hat{c}_1 = 2\hat{c}_2\{\hat{\lambda}[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})]/[\hat{\beta}\exp(\hat{\alpha} + \hat{\beta}\bar{Z})]\}$$

$$\hat{c}_0 = (1 + 2\hat{\sigma}^2\hat{c}_2)^{1/2}\{\exp(\hat{\alpha} + \hat{\beta}\bar{Z})/[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})]\} \exp\{[-\hat{\sigma}^2\hat{c}_1^2/2]/(1 + 2\hat{\sigma}^2\hat{c}_2)\}.$$

From equation (10), we estimate β^* as a function of c_0 and c_1 as follows:

$$\hat{\beta}_{\text{lik}} = \hat{c}_1/(1 - \hat{c}_0). \tag{14}$$

Confidence interval for β^* based on the likelihood approximation estimator

To obtain confidence limits for $\hat{\beta}_{\text{lik}}$, we can use the delta method, taking into account the variability of $\hat{\alpha}$, $\hat{\beta}$, $\hat{\lambda}$, $\bar{Z}$ and $\hat{\sigma}^2$. The variance formula for $\hat{\beta}_{\text{lik}}$ and the corresponding confidence interval is given in the Appendix.

We note that Schafer[15] also proposed an approximate likelihood-based estimator of β^* which is obtained using iterative methods based on the EM algorithm.[24] This estimator performed well in the simulation study presented; however, the complex computations required may be difficult to implement. In addition, although an estimate of λ can be used in this algorithm, the additional uncertainty introduced by such estimation is not considered.

SIMULATION STUDY

We conducted a simulation study to investigate the properties of the estimators proposed above. In particular, the bias, coverage probability and mean square error (MSE) were considered for (a) the 'uncorrected' estimator ($\hat{\beta}_u$), (b) the linear approximation estimator given in equation (3) ($\hat{\beta}/\hat{\lambda}$), and (c) the likelihood approximation estimator given in equation (14) ($\hat{\beta}_{\text{lik}}$). Specifically, for each parameter combination (α^*, β^*, λ), we generated 1000 main and validation study data bases. Each main study data base (D, Z) consisted of 1000 observations and was generated as follows:

(i) The true exposure X_i was randomly generated from a N(0, 1) distribution, that is $X_i \sim N(0, 1)$.

(ii) The observed exposure Z_i was generated conditional on true exposure X_i according to the model $Z_i = X_i + e_i$, where $e_i \sim \text{N}[0, (1-\lambda)/\lambda]$. This is a special case of equation (2), where $\alpha' = 0$ and $\sigma^2 = 1 - \lambda$.

(iii) The disease outcome D_i was generated from a Bernoulli random variable conditional on X_i as in equation (1), whereby $D_i = 1$ with probability $\exp(\alpha + \beta X_i)/[1 + \exp(\alpha + \beta X_i)]$, and $= 0$ otherwise.

A total of 1000 observations were created for each main study data base using (i)–(iii), that is $i = 1, \ldots, 1000$. Each validation study data base (X, Z) consisted of an additional 100 (X_i, Z_i) pairs generated as in (i) and (ii) above. The parameters $\hat{\alpha}$, $\hat{\beta}$ were then estimated from the logistic regression of D on Z using the main study data base, while the parameters $\hat{\lambda}$, $\bar{Z}$ and $\hat{\sigma}^2$ were estimated from the validation study data base using ordinary least squares. The uncorrected, linear (equation (3)) and likelihood (equation (14)) approximation estimators were then computed for each data base.

Table I. Percentage bias, mean square error and coverage probability of the uncorrected, imputation, linear approximation and likelihood approximation estimators: cohort study* simulation results (1000 iterations)

True odds ratio¶	λ	Percentage bias†			Mean square error‡			Coverage probability§			
		$\hat{\beta}_u$ #	$\hat{\beta}/\hat{\lambda}$ #	$\hat{\beta}_{lik}$ #	$\hat{\beta}_u$	$\hat{\beta}/\hat{\lambda}$	$\hat{\beta}_{lik}$	$\hat{\beta}_u$	$\hat{\beta}_{imp}$ #	$\hat{\beta}/\hat{\lambda}$	$\hat{\beta}_{lik}$
1·5	0·3	−24·7	0·8	1·2	0·087	0·079	0·082	5·8	94·2	95·7	96·0
	0·5	−18·2	0·7	0·9	0·050	0·043	0·044	47·6	94·6	95·9	96·1
	0·7	−12·1	−0·9	−0·9	0·030	0·029	0·029	80·4	96·2	96·4	96·4
2·0	0·3	−38·7	−0·6	0·5	0·245	0·073	0·080	0·0	94·0	96·6	96·8
	0·5	−29·1	1·4	2·1	0·128	0·048	0·050	7·2	92·7	94·3	94·7
	0·7	−19.2	−0·2	0·1	0·059	0·030	0·030	52·8	94·0	94·8	94·9
3·0	0·3	−54·9	−6·1	−2·2	0·638	0·090	0·112	0·0	85·1	91·4	92·5
	0·5	−43·7	−4·1	−1·9	0·339	0·050	0·055	0·0	88·4	92·9	93·5
	0·7	−29·6	−2·8	−1·7	0·136	0·033	0·034	13·8	90·1	93·0	93·3
4·0	0·3	−64·0	−13·4	−6·0	1·050	0·109	0·143	0·0	76·7	85·8	88·8
	0·5	−53·1	−11·5	−7·4	0·582	0·060	0·061	0·0	80·2	87·1	90·6
	0·7	−37·3	−6·9	−4·6	0·230	0·036	0·036	2·7	88·5	91·3	93·2
5·0	0·3	−70·0	−21·1	−10·0	1·451	0·147	0·194	0·0	64·1	78·0	87·7
	0·5	−59·9	−18·2	−11·9	0·842	0·094	0·091	0·0	66·7	79·8	86·8
	0·7	−43·8	−11·8	−8·3	0·343	0·050	0·046	0·2	79·8	86·1	90·8

* Disease probability for a person with average exposure level $= \exp(\alpha^*)/[1+\exp(\alpha^*)] = 0{\cdot}05$.

† Percentage bias for uncorrected estimator $= 100\% \times [\exp(\bar{\beta}_u)/\exp(\beta^*) - 1]$, where $\bar{\beta}_u = \Sigma\hat{\beta}_{u,i}/1000$. Percentage bias for the other estimators is defined similarly.

‡ Mean square error for uncorrected estimator $= \Sigma(\hat{\beta}_{u,i} - \beta^*)^2/1000$; mean square error for the other estimators is defined similarly.

§ Coverage probability = percentage of 1000 samples for which the 95 per cent confidence interval for β^* based on each respective estimator included the true parameter value (β^*).

¶ True odds ratio $= \exp(\beta^*)$.

\# $\hat{\beta}_u$ = uncorrected estimator; $\hat{\beta}/\hat{\lambda}$ = linear approximation estimator (equation (3)); $\hat{\beta}_{lik}$ = likelihood approximation estimator (equation (14)); $\hat{\beta}_{imp}$ = imputation estimator.

The percentage bias and mean square error (MSE) for the uncorrected estimator were computed using the formulae:

$$\text{percentage bias} = 100\% \times \left[\exp\left(\sum_{i=1}^{1000} \hat{\beta}_{u,i}/1000\right) \Big/ \exp(\beta^*) - 1\right] \tag{15}$$

$$\text{MSE} = \sum_{i=1}^{1000} (\hat{\beta}_{u,i} - \beta^*)^2/1000,$$

where $\hat{\beta}_{u,i}$ is the uncorrected estimator for the ith data base. Furthermore, 95 per cent confidence intervals for β^* were computed based on the uncorrected estimator, and the percentage of such intervals which included the true parameter value (β^*) was assessed and denoted by the 'coverage probability'. The percentage bias, MSE and coverage probability were computed in a similar manner for the linear and likelihood approximation estimators. In addition, the coverage probability was computed for the imputation estimator (bias and MSE are the same as for the linear approximation estimator). The following parameter combinations were studied: $\exp(\alpha^*)/[1+\exp(\alpha^*)] = 0{\cdot}05$ corresponding to a typical disease probability encountered in a cohort study, and 0·50 corresponding to an unmatched case-control study; $\exp(\beta^*) = 1{\cdot}5, 2{\cdot}0, 3{\cdot}0, 5{\cdot}0$; $\lambda = 0{\cdot}3, 0{\cdot}5, 0{\cdot}7$. The cohort study simulation results are presented in Table I, while the corresponding case-control results are presented in Table II. In addition, the percentage bias and coverage probability are displayed for the case of $\lambda = 0{\cdot}5$ as a function of the true odds ratio for both the cohort and case-control studies in Figures 1 and 2.

Table II. Percentage bias, mean square error and coverage probability of the uncorrected, imputation, linear approximation and likelihood approximation estimators; case-control study* simulation results (1000 iterations)

True odds ratio¶	λ	Percentage bias†			Mean square error‡			Coverage probability§			
		$\hat{\beta}_u$ #	$\hat{\beta}/\hat{\lambda}$ #	$\hat{\beta}_{lik}$ #	$\hat{\beta}_u$	$\hat{\beta}/\hat{\lambda}$	$\hat{\beta}_{lik}$	$\hat{\beta}_u$	$\hat{\beta}_{imp}$ #	$\hat{\beta}/\hat{\lambda}$	$\hat{\beta}_{lik}$
1·5	0·3	−24·8	1·1	3·1	0·082	0·021	0·029	0·0	91·0	95·8	96·4
	0·5	−18·8	−0·6	0·3	0·046	0·011	0·013	0·6	91·7	94·5	95·1
	0·7	−11·4	0·3	0·9	0·017	0·007	0·007	41·3	94·0	95·3	95·5
2·0	0·3	−39·2	−2·4	4·7	0·249	0·032	0·066	0·0	83·7	91·8	94·5
	0·5	−30·4	−2·5	1·8	0·134	0·014	0·020	0·0	88·6	92·9	95·9
	0·7	−20·0	−1·9	0·6	0·054	0·010	0·012	5·0	88·8	93·4	94·4
3·0	0·3	−55·8	−12·0	10·4	0·667	0·064	0·200	0·0	62·9	80·9	94·8
	0·5	−45·5	−9·9	4·5	0·370	0·034	0·055	0·0	67·4	83·4	95·7
	0·7	−31·5	−6·1	3·0	0·147	0·019	0·026	0·0	79·2	86·8	95·9
4·0	0·3	−65·2	−22·1	13·5	1·113	0·116	0·314	0·0	45·3	66·9	93·5
	0·5	−55·0	−17·9	6·9	0·642	0·066	0·095	0·0	50·8	70·7	95·9
	0·7	−40·1	−12·0	4·4	0·267	0·034	0·039	0·0	66·1	78·9	96·0
5·0	0·3‖	−71·2	−31·2	15·9	1·554	0·201	0·840	0·0	28·9	54·1	92·3
	0·5	−61·7	−25·4	7·9	0·926	0·118	0·140	0·0	32·5	55·9	94·9
	0·7	−46·5	−18·0	5·0	0·398	0·060	0·057	0·0	50·1	66·0	96·1

* Disease probability for a person with average exposure level $= \exp(\alpha^*)/[1+\exp(\alpha^*)] = 0{\cdot}50$.

† Percentage bias for uncorrected estimator $= 100\% \times [\exp(\bar{\beta}_u)/\exp(\beta^*) - 1]$, where $\bar{\beta}_u = \Sigma\hat{\beta}_{u,i}/1000$. Percentage bias for the other estimators is defined similarly.

‡ Mean square error for uncorrected estimator $= \Sigma(\hat{\beta}_{u,i} - \beta^*)^2/1000$; mean square error for the other estimators is defined similarly.

§ Coverage probability = percentage of 1000 samples for which the 95 per cent confidence interval for β^* based on each respective estimator included the true parameter value (β^*).

¶ True odds ratio $= \exp(\beta^*)$.

\# $\hat{\beta}_u$ = uncorrected estimator; $\hat{\beta}/\hat{\lambda}$ = linear approximation estimator (equation (3)); $\hat{\beta}_{lik}$ = likelihood approximation estimator (equation (14)); $\hat{\beta}_{imp}$ = imputation estimator.

‖ $\hat{\beta}_{lik}$ was not computable for two out of 15,000 simulated data sets, both in the case of a true odds ratio = 5·0 and $\lambda = 0{\cdot}3$; this was due to a negative estimate for c_3 (see Appendix).

As shown in Tables I and II and Figures 1 and 2, the uncorrected estimator of the odds ratio was grossly biased under all conditions. Similarly, the coverage probability was extremely low, being zero under many conditions. The mean square error was also consistently greatest for the uncorrected estimator compared with the corrected estimators.

The linear approximation estimator ($\hat{\beta}/\hat{\lambda}$) was essentially unbiased for true odds ratios up to 3·0. With a true odds ratio of 5·0, the method exhibited a moderate degree of negative bias, being up to −21 per cent for the cohort study and −31 per cent for the case-control study simulations with $\lambda = 0{\cdot}3$. The coverage probability was approximately 95 per cent for true odds ratios up to 2·0. At true odds ratios of 5·0, the coverage probability was somewhat reduced, being 78–86 per cent for the cohort study and 54–66 per cent for the case-control study simulations. Interestingly, the imputation method (which does not take error in the estimation of λ into account) resulted in noticeably smaller coverage probabilities than the linear approximation method. At an odds ratio of 5, coverage probabilities were 64–80 per cent for the cohort study and 29–50 per cent for the case-control study simulations.

The likelihood approximation estimator ($\hat{\beta}_{lik}$) yielded the least biased estimate of the odds ratio; even with a true odds ratio of 5·0, the bias was only at most −12 per cent for the cohort study and +16 per cent for the case-control simulations. The coverage probability was also closest to the correct value using the likelihood estimator; the probability was approximately 95 per cent for all odds ratios for the case-control study and for odds ratios up to 3·0 in the cohort study; at an odds

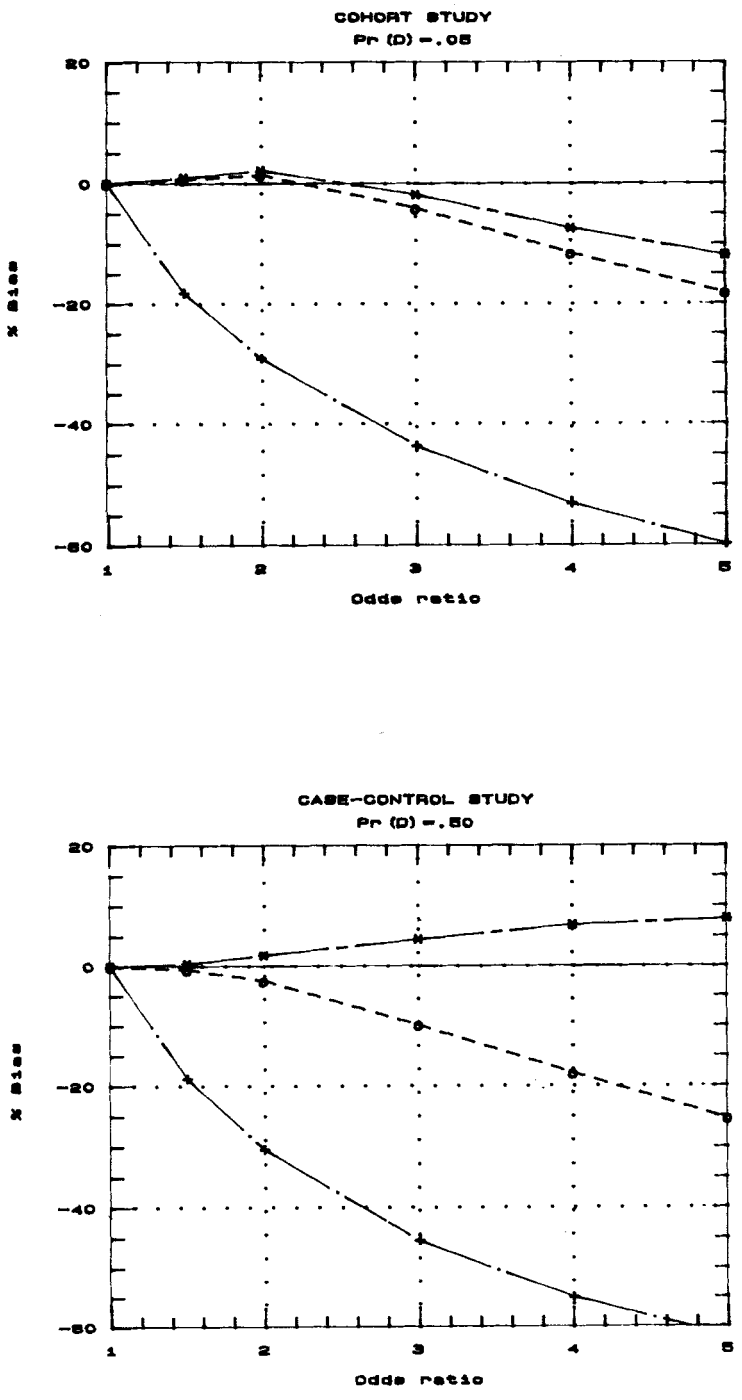

Figure 1. Percentage bias as a function of the true odds ratio for the uncorrected, linear and likelihood approximation estimators ($\lambda = 0{\cdot}5$)

—·— + Uncorrected estimator

– – – – □ Linear approximation estimator (imputation estimator)

———— * Likelihood approximation estimator

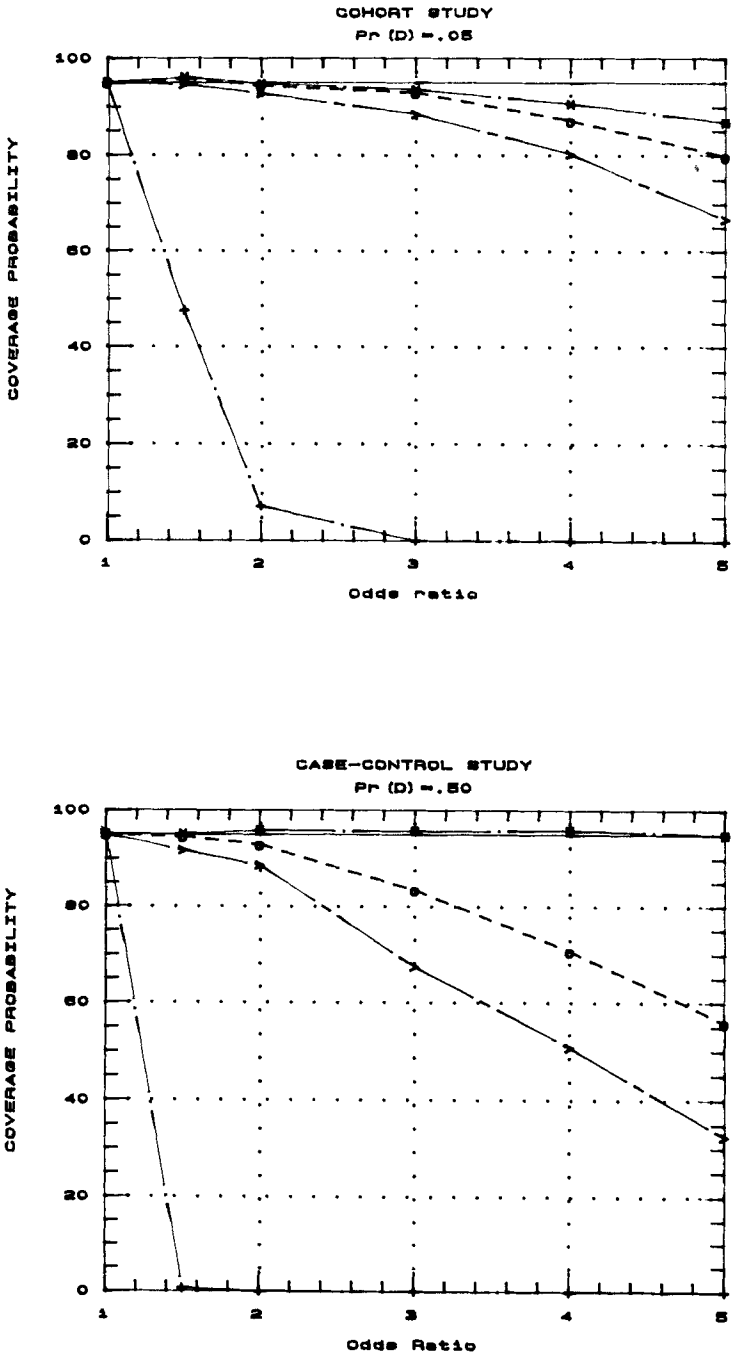

Figure 2. Coverage probability as a function of the true odds ratio for the uncorrected, linear approximation, likelihood approximation and imputation estimators ($\lambda = 0{\cdot}5$)

—·— + Uncorrected estimator

- - - - □ Linear approximation estimator

—·— * Likelihood approximation estimator

— - - — > Imputation estimator

ratio of 5·0, the coverage probability was 87–91 per cent for the cohort study simulations. For all estimators in both Tables I and II, as λ increases the bias decreases and the coverage probability generally increases because the 'measurement error' variability $= 1 - \lambda$ decreases. In addition, we also examined plots of the empirical density functions for each estimator based on the cohort study simulation results. The distribution of each of the estimators is approximately normal; thus, asymptotic 95 per cent coverage probability should be approximately valid for the sample sizes considered. Also, the variance of the corrected estimator is larger than for the uncorrected estimator.

EXAMPLE 1: RELATIVE RISK OF BREAST CANCER IN RELATION TO CALORIE-ADJUSTED SATURATED FAT INTAKE CORRECTED FOR ERRORS IN MEASUREMENT (CONTINUOUS VARIABLE)

In a previous study, we prospectively examined the relationship between dietary saturated fat intake as assessed by a food frequency questionnaire in 1980 and risk of breast cancer from 1980 to 1984, among a cohort of 89,538 women in the Nurses Health Study.[25] To measure dietary fat in this study, we utilized a self-administered semi-quantitative food frequency questionnaire that in 1981 had been subjected to a detailed validation study[26] among 173 cohort members. To represent true dietary intake in the validation study, we used an average of four one-week diet records based on weighed food intake collected by each subject at three-month intervals over a one-year period. We assumed that the large number of days per subject would be sufficient to dampen within-person variation and that the three-month intervals would account for any important seasonal variation. The main and validation study populations were comparable with respect to age (years: main study population mean $\pm$ SD $= 46{\cdot}7 \pm 7{\cdot}2$; validation study population mean $\pm$ SD $= 45{\cdot}9 \pm 7{\cdot}4$), and calorie-adjusted saturated fat intake as assessed in the 1980 questionnaire (grams/day: main study population mean $\pm$ SD $= 25{\cdot}8 \pm 5{\cdot}9$; validation study population mean $\pm$ SD $= 25{\cdot}0 \pm 5{\cdot}7$).

In the main study, we employed a multiple logistic model with breast cancer cumulative incidence as the dependent variable and calorie-adjusted saturated fat intake (a continuous variable)[27] as the primary predictor variable. Age and alcohol intake were also included as covariates. In this model, the coefficient for calorie-adjusted saturated fat intake represents the effect on breast cancer incidence of a daily increase of one gram of saturated fat with total calories held constant. Since the average saturated fat intake was approximately 25 grams daily, one gram is a very small increment. Thus, we computed the relative risk for a 10 gram increase in daily calorie-adjusted saturated fat intake, which approximates the difference between the means of the top and bottom quintiles as assessed by the diet record data.[26] To estimate λ, we utilized the data provided by the 173 women in the valdation study; calorie-adjusted saturated fat intake measured by the average of four one-week diet records (X) was regressed on the calorie-adjusted saturated fat intake measured by the questionnaire after completion of the diet records (Z). Two different correction procedures were used. For each procedure, we used the estimate of β from the main Nurses Health Study data and the estimate of λ from the validation substudy; for the likelihood approximation method, we also used estimates of α from the main study and of $\bar{Z}$ and σ^2 from the validation substudy. These parameters were used in equation (3) for the linear approximation estimator and equation (14) for the likelihood approximation estimator.

Based on the full 28 days of diet records, the estimated slope ($\hat{\lambda}$) representing the change in diet record calorie-adjusted saturated fat intake (grams/day) associated with a one gram change in daily saturated fat intake measured by the questionnaire was 0·468, with a standard error of 0·048. One assumption of the linear regression model in equation (2) is that there is homogeneity of

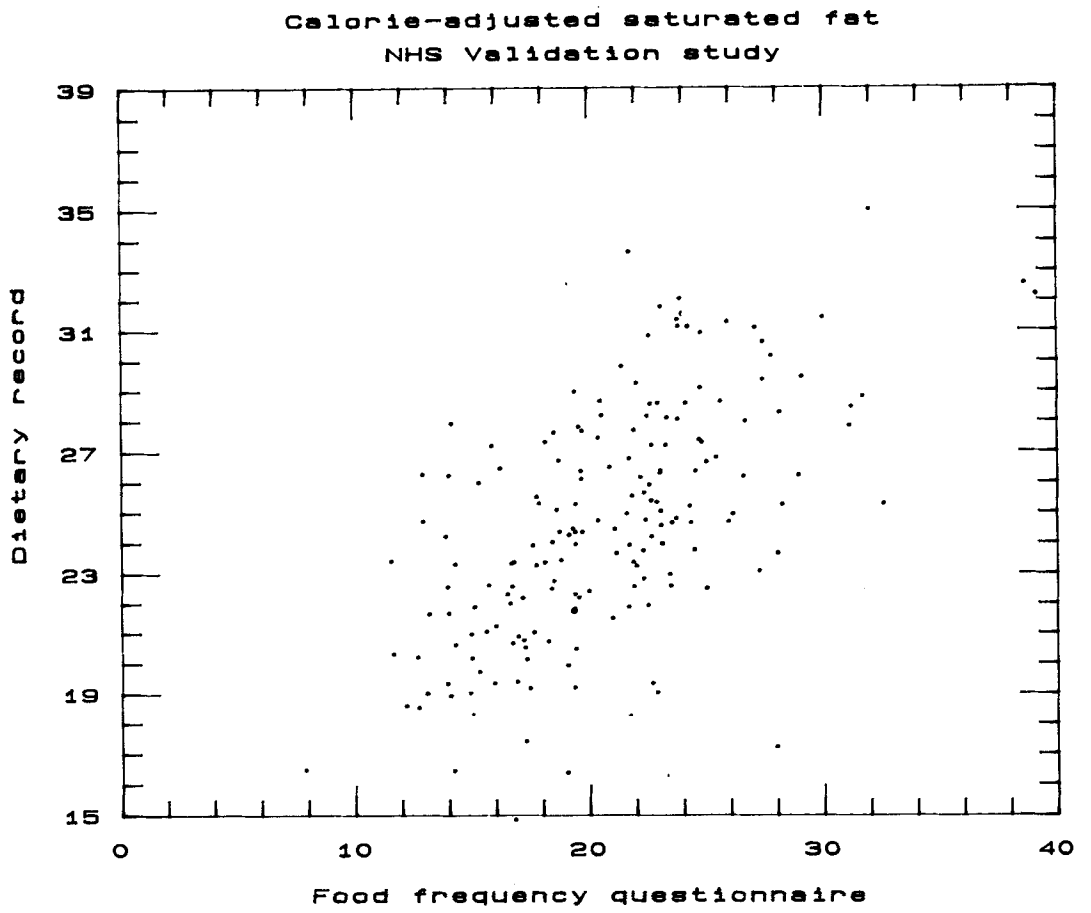

Figure 3. Plot of calorie-adjusted saturated fat intake for the diet record versus calorie-adjusted saturated fat intake for the food frequency questionnaire (validation study data from the Nurses Health Study)

variance for all levels of food frequency calorie-adjusted saturated fat (Z) from the validation study. To investigate this issue, we present in Figure 3 a plot of the linear regression of diet record on food frequency calorie-adjusted saturated fat intake based on the validation study data base. It appears from this plot that the residuals will be approximately homoscedastic across different levels of Z. Based on the actual data in the total cohort, the observed relative risk for a 10 gram increase in calorie-adjusted saturated fat intake was 0·92 (95 per cent confidence limits (CL) from 0·80 to 1·05). Using both the linear and likelihood approximation methods, the corrected relative risk was 0·83 with 95 per cent CL from 0·61 to 1·12 (see Table III), providing an estimate of the true effect due to a 10 gram increase in calorie-adjusted saturated fat intake given our main study data and allowing for measurement error. As expected, the point estimate of the corrected relative risk is further from unity than the observed relative risk, and the width of the 95 per cent confidence interval has increased.

In this example, we assumed that the 28 days of diet recording in the validation study were sufficient to dampen within-person variation, thus providing a close approximation to true intake for an individual. To provide an example of the situation where X (true exposure) is still subject to within-person variation, we sampled two days (one day from each of weeks 1 and 3) and four days

Table III. Relative risks of breast cancer for a 10 gram/day increase in calorie-adjusted saturated fat intake, adjusted for measurement error

	Validation study*			Main study†	
Example	Days of diet record	$\hat{\lambda} \pm$SE	Observed RR (95% CL)	Corrected‡ (linear) RR (95% CL)	Corrected§ (likelihood) RR (95% CL)
$\hat{\beta}_{10g} = -0{\cdot}0878$ SE $= 0{\cdot}0712$	28	$0{\cdot}468 \pm 0{\cdot}048$	0·92 (0·80–1·05)	0·83 (0·61–1·12)	0·83 (0·61–1·12)
	4	$0{\cdot}554 \pm 0{\cdot}073$	0·92 (0·80–1·05)	0·85 (0·66–1·10)	0·85 (0·66–1·10)
	2	$0{\cdot}540 \pm 0{\cdot}090$	0·92 (0·80–1·05)	0·85 (0·65–1·11)	0·85 (0·65–1·11)

* Validation study data are based on 173 women participating in the Nurses Health Study.[26] Each woman completed four one-week diet records over a one-year period. The four days are used in this analysis were sampled one from each week; the two days were sampled one from each of weeks 1 and 3.

† Main study data are based on 590 cases of breast cancer occurring among 89,538 women participating in the Nurses Health Study, aged 34–59 years and followed for four years.[25] Logistic model included calorie-adjusted saturated fat intake in grams/day as a continuous variable, age (34–39/40–44/45–49/50–54/55–59) and alcohol intake (0/0·1–1·4/1·5–4·9/5·0–14·9/15 + grams/day).

‡ Based on the linear approximation estimator given in equation (3).

§ Based on the likelihood approximation estimator given in equation (14).

(one from each week) from the total 28 days of diet recording. In Table III, we have provided corrected relative risk estimates using these two or four days of diet recording to estimate λ. As expected, the estimates of λ were similar, but their standard errors were larger; in this instance, however, the degree of variation in the standard error of λ had only a small influence on the confidence limits for the corrected relative risks. Since, by chance, the value for λ was somewhat lower using all 28 days, the confidence intervals were actually slightly narrower with fewer days.

EXAMPLE 2: RELATIVE RISK OF BREAST CANCER IN RELATION TO CALORIE-ADJUSTED SATURATED FAT INTAKE CORRECTED FOR ERRORS IN MEASUREMENT (QUINTILE ANALYSIS)

In epidemiologic analyses, continuous variables are frequently categorized into groups based on rank, such as quartiles or quintiles. To illustrate the correction of relative risks based on quintiles, we again utilize the relationship between calorie-adjusted saturated fat intake and breast cancer risk among Nurses Health Study participants.[25] We employed a multiple logistic model with a five-level continuous variable (coded as 1–5) to represent quintiles of calorie-adjusted saturated fat intake and included age and alcohol intake as covariates. In this model, the coefficient for saturated fat (β) indicates the effect of a one quintile increase in calorie-adjusted saturated fat intake, and $\exp(4\beta)$ is the relative risk for quintile 5 versus quintile 1. To estimate the validity of the dietary questionnaire in the same scale as was used in the cohort analysis, we divided the 173 subjects participating in this substudy into quintiles of calorie-adjusted saturated fat intake based on their diet record data and also into quintiles based on their 1981 questionnaire data. We then regressed diet record saturated fat intake in quintiles (X) on questionnaire saturated fat intake in quintiles (Z) to obtain the estimated regression coefficient ($\hat{\lambda}$) and its variance.

Table IV. Relative risks of breast cancer corrected for errors in measurement of dietary saturated fat. Relative risks are for highest quintile of calorie-adjusted intake compared with lowest quintile, based on a multiple logistic model

	Validation study*			Main Study†	
Example	$\hat{\lambda}$ (SE)	$\hat{\beta}$ (SE)	Observed RR (95% CL)	Corrected‡ (linear) RR (95% CL)	Corrected§ (likelihood) RR (95% CL)
Quintile¶	0·587 (0·062)	−0·0410 (0·030)	0·85 (0·67–1·07)	0·76 (0·50–1·13)	0·76 (0·50–1·14)

* Validation study data are based on 173 women participating in the Nurses Health Study.[26] Each woman completed four one-week diet records over a one-year period. The four days used in this analysis were sampled one from each week; the two days were sampled one from each of weeks 1 and 3.

† Main study data are based on 590 cases of breast cancer occurring among 89,538 women participating in the Nurses Health Study, aged 34–59 years and followed for four years.[25] Logistic model included calorie-adjusted saturated fat intake in grams/day as a continuous variable, age (34–39/40–44/45–49/50–54/55–59) and alcohol intake (0/0·1–1·4/1·5–4·9/5·0–14·9/15 + grams/day).

‡ Based on the linear approximation estimator given in equation (3).

§ Based on the likelihood approximation estimator given in equation (14).

¶ Scale of measurement is 1–5. For 'quintile' values 1–5, the categories were assigned consecutive integer values and then used as a continuous variable.

Based on the full 28 days of diet records, the estimated slope relating diet record to questionnaire saturated fat intake ($\hat{\lambda}$) in the validation study was 0·587, and the observed relative risk in the main study for quintile 5 versus quintile 1 was 0·85 (95 per cent CL from 0·67 to 1·07) (see Table IV). Using either the linear or likelihood approximation method, the corrected relative risk was 0·76 with an upper 95 per cent confidence limit of 1·13 and 1·14 respectively for the two methods.

Although groupings of continuous data such as quintiles are widely used in epidemiologic studies, this typically results in categories with uneven intervals between their midpoints. For example, if the underlying distribution is normal, then the intervals between the midpoints of the extreme categories and their adjacent categories will be greater than the intervals between the middle categories; this is due to the greater density of data near the mean. The use of quantiles (such as quintiles) for a dependent variable is also not strictly correct, since this violates the assumption of normality. To be consistent with this assumption, we retained the same five groups, but assigned values of $-1{\cdot}28$, $-0{\cdot}52$, 0·0, 0·52, 1·28 to them (corresponding to normal deviates of the quintile midpoints) before fitting the logistic model and the simple regression model (equation (2)). In this way, quintiles were treated as grouped normally distributed data. Results were similar with this weighting scheme to those for quintiles weighted in the customary manner.

We note that another possible method of assessing the relationship between breast cancer incidence and saturated fat consumption is to include both total saturated fat and total caloric intake as independent variables in the logistic regression based on the main study population, and to correspondingly perform a multivariate regression of (X_1, X_2) on (Z_1, Z_2) based on the validation study population, where X_1, X_2 are total saturated fat and total caloric intake from the diet record and Z_1, Z_2 are the corresponding variables from the food frequency questionnaire. This method would allow one to consider the effect of the probable correlation between measurement error associated with these two variables on the point estimate and standard error of

the regression coefficient for saturated fat from the logistic model (corrected for measurement error). However, the problem of performing error correction for logistic regression coefficients in a multivariate setting is beyond the scope of this paper.

DISCUSSION

We have provided methods to correct relative risks obtained from a logistic model for errors in measurement of exposure in cohort studies. These methods are applicable to errors due to simple random (unbiased) error in exposure as well as to systematic within-person error. Since both types of error may commonly occur simultaneously in epidemiologic studies, a method which does not require the assumption of strictly random within-person error has obvious advantages. This approach cannot, of course, correct for systematic bias in exposure measurement related to disease status, which is likely to affect case-control studies.

The situation in which measurement error is strictly due to random (unbiased) within-person variation can be considered as a special case of the model in equation (2). In this situation, the validation study can consist of two or more (R) repeated measures of observed exposure among a subsample of subjects and λ can be estimated by the intraclass correlation among replicates[28] for persons in the validation study (see equation (19)).

One way of considering both systematic (biased) and random (unbiased) within-person error is in the context of the classical errors-in-variables model. In particular, suppose we measure observed exposure on R occasions for each individual, and let Z_{ij} be the jth replicate for the ith person. We can represent Z_{ij} in the form

$$Z_{ij} = X_i + \alpha_i + \varepsilon_{ij}, \tag{16}$$

where $X_i \sim N(\mu_X, \sigma_X^2)$ represents true exposure, $\alpha_i \sim N(0, \sigma_\alpha^2)$ represents systematic within-person error, and $\varepsilon_{ij} \sim N(0, \sigma^2)$ represents random within-person error. We can depict the regression of X_i on Z_{ij} in the form

$$X_i = \alpha' + \lambda Z_{ij} + \varepsilon_i. \tag{17}$$

From equations (16) and (17), it follows that

$$\lambda = \sigma_X^2/(\sigma_X^2 + \sigma_\alpha^2 + \sigma^2). \tag{18}$$

Furthermore, the intraclass correlation coefficient among the R replicates is given by

$$\rho_{\text{intraclass}} = (\sigma_X^2 + \sigma_\alpha^2)/(\sigma_X^2 + \sigma_\alpha^2 + \sigma^2). \tag{19}$$

If error is strictly random (unbiased), then $\sigma_\alpha^2 = 0$, $\lambda = \rho_{\text{intraclass}}$, and one can estimate λ from an intraclass correlation coefficient based on replicate determinations $Z_{i1}, \ldots, Z_{iR}$ without the need for performing a validity study to determine X. Clearly, if both systematic and random error are present, then the parameters in equations (18) and (19) will not be the same and one needs to perform a validity study to determine X and appropriately estimate λ.

The derivation of the linear and likelihood estimators was based on the assumption of normality of exposure variables. However, it is often the case that epidemiologic analyses are performed based on standardized variables such as quintiles (for example see Table IV). To examine the robustness of the procedures in this setting, we performed simulation studies similar to those given in Tables I and II, but quantifying the true and observed exposures into quintiles. The results of these simulations were similar to those given in Tables I and II.

It will be apparent that the value of λ does not directly provide an interpretable measure of validity, since it reflects differences in the scale of the true and surrogate measurements as well as

the correlation between them. For example, in Table III, part of the correction is related to a change in scale, since the standard deviation of calorie-adjusted saturated fat intake measured by the questionnaire is larger than the standard deviation according to the diet record. Even if the two measures were perfectly correlated, the larger standard deviation of the questionnaire would result in a value of λ less than one. Although less perfect measures will usually result in larger variances, this is not necessarily true; the value of λ can thus exceed one (this can also result simply from a change in units). If both the 'true' (X) and surrogate (Z) measures have the same standard deviation, then the value for λ is equivalent to the correlation coefficient relating X and Z, and is comparable to the measure of misclassification discussed by Walker and Blettner.[29]

We have assumed that the validation study population is representative of the main study population and that the estimates of β (obtained from the main study) and λ (obtained from the validation study) are independent. To minimize concern regarding the generalizability of the validity estimate to the main study population, the validation study would ideally be conducted among a subsample of the main study participants. However, it may be possible, with caution, to use an estimate of validity from a completely external source in some instances, if the populations are generally similar. As long as the validation study includes only a small fraction of the total study population (for example, in Table III it represents approximately 0·2 per cent of the total study population), the assumption that the estimates of β and λ are independent will be approximately satisfied. In this setting, the determination of the appropriate number of subjects for the validation study merits further consideration.

The effect of including covariates in a multiple logistic model on the correction of relative risks deserves comment. In general, the covariates should also be considered in the design and analysis of the validation study; this would be particularly important for variables that are strongly associated with the primary exposure. One approach would be to adjust both X and Z for relevant covariates before regressing X on Z as in equation (2). (This adjustment can be accomplished by obtaining the residual from a regression model with these covariates as independent variables and X as the dependent variable; similar residuals can be computed using the same covariates where Z is the dependent variable.) In Example 1, calorie-adjusted saturated fat was minimally correlated with age and alcohol use, so we did not feel compelled to make our estimate of λ conditional on these variables, although this could have been done. Furthermore, we have assumed that these covariates are measured without error. Error in the measurement of covariates can have important effects on the relationship of primary interest;[6] this topic is beyond the scope of our paper.

In summary, we have provided two methods to correct relative risk estimates obtained from a logistic model for error in the measurement of exposure in cohort studies, based on an estimate of validity from a separate substudy. These approaches yield confidence intervals for the relative risk that incorporate a component due to uncertainty in the estimate of validity. While technical improvement in the measurement of exposure is the most fundamental way to enhance the accuracy of relative risks, these approaches may also improve our estimation of effects in epidemiologic studies when exposure measurement remains imperfect.

Note

The methods discussed in this paper can be implemented on either a programmable calculator or a micro or larger computer. Programs are available on request for either an HP-41C programmable calculator or in MS-DOS BASIC or FORTRAN 77 for this purpose. Please send either a blank magnetic card, a $5\frac{1}{4}$ in. floppy diskette, or mainframe magnetic tape specifications, if such programs are desired.

APPENDIX: CONFIDENCE LIMITS FOR THE LIKELIHOOD APPROXIMATION ESTIMATOR

We represent var $(\hat{\beta}^*)$ in the form $f(\hat{\alpha}, \hat{\beta}, \hat{\lambda}, \bar{Z}, \hat{\sigma}^2)$ and use the delta method to obtain approximate confidence limits for $\hat{\beta}^*$ as given in equation (14). Since

(a) $\hat{\beta}^*$ is a function of $\hat{\alpha}$, $\hat{\beta}$ from the main study and $\hat{\lambda}$, $\bar{Z}$, $\hat{\sigma}^2$ from the validity study,
(b) We assume independence between parameters estimated from the main and validity studies, and
(c) $\hat{\lambda}$, $\bar{Z}$ and $\hat{\sigma}^2$ are always statistically independent,

it follows that

$$\begin{aligned}\operatorname{var}(\hat{\beta}^*) \cong (\partial f/\partial\hat{\alpha})^2 \operatorname{var}(\hat{\alpha}) + (\partial f/\partial\hat{\beta})^2 \operatorname{var}(\hat{\beta}) + (\partial f/\partial\hat{\lambda})^2 \operatorname{var}(\hat{\lambda}) \\ + (\partial f/\partial\bar{Z})^2 \operatorname{var}(\bar{Z}) + (\partial f/\partial\hat{\sigma}^2)^2 \operatorname{var}(\hat{\sigma}^2) + 2(\partial f/\partial\hat{\alpha})(\partial f/\partial\hat{\beta}) \operatorname{cov}(\hat{\alpha}, \hat{\beta}). \qquad (20)\end{aligned}$$

Let

$$\begin{aligned}\hat{c}_3 &= (1 + 2\hat{\sigma}^2\hat{c}_2) \\ c_4 &= \exp(\hat{\alpha} + \hat{\beta}\bar{Z})/[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})] \\ \hat{c}_5 &= -\hat{\sigma}^2\hat{c}_1^2/(2\hat{c}_3) \\ \hat{c}_6 &= \hat{\lambda}^2[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})]^2 - \hat{\beta}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})\hat{\sigma}^2.\end{aligned}$$

Since $\hat{\beta}^* = \hat{c}_1/(1 - \hat{c}_0)$, we have that

$$\partial f/\partial\hat{\alpha} = (\partial\hat{c}_1/\partial\hat{\alpha})/(1 - \hat{c}_0) + (\partial\hat{c}_0/\partial\hat{\alpha})[\hat{c}_1/(1 - \hat{c}_0)^2]. \qquad (21)$$

It is straightforward but tedious to show that

$$\begin{aligned}\partial\hat{c}_2/\partial\hat{\alpha} &= [\hat{\beta}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})/(2\hat{c}_6)]\{1 - \{2\hat{\lambda}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})] - \hat{\beta}^2 \exp(\hat{\alpha} + \hat{\beta}\bar{Z})\hat{\sigma}^2\}/\hat{c}_6\} \\ \partial\hat{c}_1/\partial\hat{\alpha} &= 2\hat{\lambda}(\hat{\beta}\hat{c}_4)^{-1}[(\partial\hat{c}_2/\partial\hat{\alpha}) - \hat{c}_2(1 - \hat{c}_4)] \\ \partial\hat{c}_5/\partial\hat{\alpha} &= (-\hat{\sigma}^2\hat{c}_1/\hat{c}_3)[\partial\hat{c}_1/\partial\hat{\alpha} - (\hat{\sigma}^2\hat{c}_1/\hat{c}_3)\partial\hat{c}_2/\partial\hat{\alpha}] \\ \partial\hat{c}_0/\partial\hat{\alpha} &= \hat{c}_4 \exp(\hat{c}_5)[\hat{c}_3^{1/2}(1 - \hat{c}_4) + \hat{c}_3^{-1/2}\hat{\sigma}^2\partial\hat{c}_2/\partial\hat{\alpha} + \hat{c}_3^{1/2}\partial\hat{c}_5/\partial\hat{\alpha}]. \qquad (22)\end{aligned}$$

Thus, we can evaluate $\partial f/\partial\hat{\alpha}$ in equation (21) using the results in equation (22). In a similar manner, we have that

$$\begin{aligned}\partial f/\partial\hat{\beta} &= (\partial\hat{c}_1/\partial\hat{\beta})/(1 - \hat{c}_0) + (\partial\hat{c}_0/\partial\hat{\beta})[\hat{c}_1/(1 - \hat{c}_0)^2] \\ &\vdots \qquad\qquad \vdots \\ \partial f/\partial\hat{\sigma}^2 &= (\partial\hat{c}_1/\partial\hat{\sigma}^2)/(1 - \hat{c}_0) + (\partial\hat{c}_0/\partial\hat{\sigma}^2)[\hat{c}_1/(1 - \hat{c}_0)^2], \qquad (23)\end{aligned}$$

where

$$\begin{aligned}\partial\hat{c}_6/\partial\hat{\beta} &= \exp(\hat{\alpha} + \hat{\beta}\bar{Z})\{2\hat{\lambda}^2\bar{Z}[1 + \exp(\hat{\alpha} + \hat{\beta}\bar{Z})] - 2\hat{\beta}\hat{\sigma}^2 - \hat{\beta}^2\bar{Z}\hat{\sigma}^2\} \\ \partial\hat{c}_2/\partial\hat{\beta} &= [\exp(\hat{\alpha} + \hat{\beta}\bar{Z})\hat{\beta}/\hat{c}_6]\{1 + \hat{\beta}\bar{Z}/2 - [\hat{\beta}/(2\hat{c}_6)]\partial\hat{c}_6/\partial\hat{\beta}\} \\ \partial\hat{c}_1/\partial\hat{\beta} &= 2\hat{\lambda}(\hat{\beta}\hat{c}_4)^{-1}[\partial\hat{c}_2/\partial\hat{\beta} - \hat{c}_2/\hat{\beta} - \hat{c}_2\bar{Z}(1 - \hat{c}_4)] \\ \partial\hat{c}_5/\partial\hat{\beta} &= -(\sigma^2\hat{c}_1/\hat{c}_3)[\partial\hat{c}_1/\partial\hat{\beta} - (\hat{c}_1\hat{c}_3^{-1}\sigma^2)\partial\hat{c}_2/\partial\hat{\beta}] \\ \partial\hat{c}_0/\partial\hat{\beta} &= \hat{c}_4 \exp(\hat{c}_5)[\hat{c}_3^{-1/2}\hat{\sigma}^2\partial\hat{c}_2/\partial\hat{\beta} + \hat{c}_3^{1/2}\partial\hat{c}_5/\partial\hat{\beta} + \hat{c}_3^{1/2}\bar{Z}(1 - \hat{c}_4)]\end{aligned}$$

$$\partial\hat{c}_2/\partial\hat{\lambda} = -\hat{\beta}^2 \exp(\hat{\alpha}+\hat{\beta}\bar{Z})[1+\exp(\hat{\alpha}+\hat{\beta}\bar{Z})]^2\hat{\lambda}/\hat{c}_6^2$$

$$\partial\hat{c}_1/\partial\hat{\lambda} = 2(\hat{\beta}\hat{c}_4)^{-1}(\hat{c}_2+\hat{\lambda}\,\partial\hat{c}_2/\partial\hat{\lambda})$$

$$\partial\hat{c}_5/\partial\hat{\lambda} = -(\hat{\sigma}^2\hat{c}_1/\hat{c}_3)[\partial\hat{c}_1/\partial\hat{\lambda} - (\hat{\sigma}^2\hat{c}_1/\hat{c}_3)\partial\hat{c}_2/\partial\hat{\lambda}]$$

$$\partial\hat{c}_0/\partial\hat{\lambda} = \hat{c}_4 \exp(\hat{c}_5)[\hat{c}_3^{-1/2}\hat{\sigma}^2\,\partial\hat{c}_2/\partial\hat{\lambda} + \hat{c}_3^{1/2}\,\partial\hat{c}_5/\partial\hat{\lambda}]$$

$$\partial\hat{c}_2/\partial\bar{Z} = [\hat{\beta}^3 \exp(\hat{\alpha}+\hat{\beta}\bar{Z})/(2\hat{c}_6)]\{1-\exp(\hat{\alpha}+\hat{\beta}\bar{Z})\{2\hat{\lambda}^2[1+\exp(\hat{\alpha}+\hat{\beta}\bar{Z})] - \hat{\beta}^2\hat{\sigma}^2\}/\hat{c}_6\}$$

$$\partial\hat{c}_1/\partial\bar{Z} = 2\hat{\lambda}(\hat{\beta}\hat{c}_4)^{-1}[\partial\hat{c}_2/\partial\bar{Z} - \hat{c}_2\hat{\beta}(1-\hat{c}_4)]$$

$$\partial\hat{c}_5/\partial\bar{Z} = -(\hat{\sigma}^2\hat{c}_1/\hat{c}_3)(\partial\hat{c}_1/\partial\bar{Z} - \hat{\sigma}^2\hat{c}_1\hat{c}_3^{-1}\partial\hat{c}_2/\partial\bar{Z})$$

$$\partial\hat{c}_0/\partial\bar{Z} = \hat{c}_4 \exp(\hat{c}_5)[\hat{\sigma}^2\hat{c}_3^{-1/2}\,\partial\hat{c}_2/\partial\bar{Z} + \hat{c}_3^{1/2}\hat{\beta}(1-\hat{c}_4) + \hat{c}_3^{1/2}\,\partial\hat{c}_5/\partial\bar{Z}]$$

$$\partial\hat{c}_2/\partial\hat{\sigma}^2 = [\hat{\beta}^2 \exp(\hat{\alpha}+\hat{\beta}\bar{Z})/\hat{c}_6]^2/2$$

$$\partial\hat{c}_1/\partial\hat{\sigma}^2 = 2[\hat{\lambda}(\hat{\beta}\hat{c}_4)^{-1}]\partial\hat{c}_2/\partial\hat{\sigma}^2$$

$$\partial\hat{c}_5/\partial\hat{\sigma}^2 = -(\hat{c}_1/\hat{c}_3)[\hat{c}_1/2 + \hat{\sigma}^2\partial\hat{c}_1/\partial\hat{\sigma}^2 - (\hat{\sigma}^2\hat{c}_1/\hat{c}_3)(\hat{c}_2+\hat{\sigma}^2\partial\hat{c}_2/\partial\hat{\sigma}^2)]$$

$$\partial\hat{c}_0/\partial\hat{\sigma}^2 = \hat{c}_4 \exp(\hat{c}_5)[\hat{c}_3^{-1/2}(\hat{c}_2+\hat{\sigma}^2\,\partial\hat{c}_2/\partial\hat{\sigma}^2) + \hat{c}_3^{1/2}\,\partial\hat{c}_5/\partial\hat{\sigma}^2].$$

Finally, if there are n subjects in the validation study, then it follows that

$$\begin{aligned} \operatorname{var}(\hat{\lambda}) &= \hat{\sigma}^2 \Big/ \sum_{i=1}^{n}(Z_i-\bar{Z})^2 \\ \operatorname{var}(\bar{Z}) &= \sum_{i=1}^{n}(Z_i-\bar{Z})^2/[n(n-1)] \\ \operatorname{var}(\hat{\sigma}^2) &= 2\hat{\sigma}^4/(n-2). \end{aligned} \tag{24}$$

Thus, upon (a) evaluating equations (21) and (23), (b) obtaining $\hat{\alpha}$, $\hat{\beta}$, $\operatorname{var}(\hat{\alpha})$, $\operatorname{var}(\hat{\beta})$, $\operatorname{cov}(\hat{\alpha}, \hat{\beta})$ from the main study, and (c) obtaining $\hat{\lambda}$, $\bar{Z}$, $\hat{\sigma}^2$ (residual mean square), $\operatorname{var}(\hat{\lambda})$, $\operatorname{var}(\bar{Z})$, $\operatorname{var}(\hat{\sigma}^2)$ from the validation study, one can compute $\operatorname{var}(\hat{\beta}^*)$ using equation (20). This yields a 95 per cent confidence interval for the true relative risk given by $\exp[\hat{\beta}^* \pm 1{\cdot}96\,\mathrm{SE}(\hat{\beta}^*)]$.

ACKNOWLEDGEMENTS

This work was supported by research grants CA 40356, CA 40935 and CA 42059 and HL 07427 from the National Institutes of Health. We wish to acknowledge Mauricio Hernandez for programming assistance involved with the preparation of this manuscript. We also wish to thank the referees for their thoughtful comments concerning the manuscript.

REFERENCES

1. Kleinbaum, D. G., Kupper, L. L. and Morganstern, H. *Epidemiologic Research*, Lifetime Learning, Belmont, CA, 1982.
2. Rosner, B. and Polk, B. F. 'Predictive values for routine blood pressure measurements in screening for hypertension', *American Journal of Epidemiology*, **117**, 429–442 (1983).
3. Liu, K., Stamler, J., Dyer, A., McKeever, J. and McKeever, P. 'Statistical methods to assess and minimize the role of intra-individual variability in obscuring the relationship between dietary lipids and serum cholesterol', *Journal of Chronic Diseases*, **31**, 399–418 (1978).
4. Beaton, G. H., Milner, J., Corey, P., McGuire, V., Cousins, M., Stewart, E., De Ramos, M., Hewitt, D., Grambsch, P. V., Kassim, N. and Little, J. A. 'Sources of variance in 24-hour dietary recall data: implication for nutrition study design and interpretation', *American Journal of Clinical Nutrition*, **32**, 2456–2459 (1979).

5. Liu, K., Cooper, R., McKeever, J., McKeever, P., Byington, R. and Soltero, I. 'Assessment of the association between habitual salt intake and high blood pressure: methodological problems', *American Journal of Epidemiology*, **110**, 219–226 (1979).
6. Kupper, L. 'Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies', *American Journal of Epidemiology*, **120**, 643–648 (1984).
7. Carroll, R. J., Spiegelman, C. H., Gordon Lan, K. K., Bailey, K. T. and Abbott, R. D. 'On errors-in-variables for binary regression models', *Biometrika*, **71**, 19–25 (1984).
8. Stefanski, L. A. and Carroll, R. J. 'Covariate measurement error in logistic regression', *Annals of Statistics*, **13**, 1335–1351 (1985).
9. Armstrong, B. 'Measurement error in the generalized linear model', *Communications in Statistics – Simulation and Computation*, **14**, 529–544 (1985).
10. Armstrong, B. G. and Oakes, D. 'Effects of approximation in exposure assessments on estimates of exposure–response relationships', *Scandinavian Journal of Work Environment and Health*, **8** (supplement 1), 20 (1982).
11. Burr-Doss, D. 'Errors-in-variables in binary regression – Berkson case', Technical Report No. 104, Division of Biostatistics, Stanford University, 1985.
12. Clark, R. R. 'The errors-in-variables problem in the logistic regression model', unpublished PhD thesis, University of North Carolina, Chapel Hill, NC, 1982.
13. Prentice, R. L. 'Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors', *Journal of the American Statistical Association*, **81**, 321 (1986).
14. Whittemore, A. S. and Grosser, S. 'Regression methods for data with complete covariates', in Moolgavkar, S. H. and Prentice, R. L. (eds), *Modern Statistical Methods in Chronic Disease*, Wiley-Interscience, 1986.
15. Schafer, D. W. 'Covariate measurement error in generalized linear models', *Biometrika*, **79**, 385–391 (1987).
16. Chen, T. T. 'Log linear models of categorized data with misclassification and double sampling', *Journal of the American Statistical Association*, **72**, 481–488 (1979).
17. Espeland, M. A. and Odoroff, C. L. 'Log linear models for doubly sampled categorical data fitted by the EM algorithm', *Journal of the American Statistical Association*, **80**, 663–670 (1985).
18. Espeland, M. A. 'A general class of models for discrete multivariate data', *Communications in Statistics – Simulation and Computation*, **15**(2), 405–424 (1986).
19. Espeland, M. A. and Hui, S. L. 'A general approach to analyzing epidemiological data that contain misclassification errors', *Biometrics*, **43**, 1001–1012 (1987).
20. Prentice, R. L. 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika*, **69**, 331–342 (1982).
21. Barron, B. A. 'The effects of misclassification on the estimation of relative risk', *Biometrics*, **33**, 414–418 (1977).
22. Selen, J. 'Adjusting for errors in classification and measurement in the analysis of partly and purely categorical data', *Journal of the American Statistical Association*, **81**, 75–81 (1986).
23. Armitage, P. *Statistical Methods in Medical Research*, Blackwell Scientific, London, 1971.
24. Dempster, A. P., Laird, N. M. and Rubin, D. B. 'Maximum likelihood from incomplete data via the EM algorithm', *Journal of the Royal Statistical Society B*, **39**(1), 1–38 (1977).
25. Willett, W. C., Stampfer, M. J., Colditz, G. A., Rosner, B., Hennekens, C. H. and Speizer, F. E. 'Dietary fat and risk of breast cancer', *New England Journal of Medicine*, **316**, 22–28 (1987).
26. Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekens, C. H. and Speizer, F. E. 'Reproducibility and validity of a semi-quantitative food frequency questionnaire', *American Journal of Epidemiology*, **122**, 51–65 (1985).
27. Willett, W. C. and Stampfer, M. J. 'Total energy intake: implications for epidemiologic analyses', *American Journal of Epidemiology*, **124**(1), 17–27 (1986).
28. Donner, A. and Koval, J. J. 'The estimation of intraclass correlation in the analysis of familial data', *Biometrics*, **36**, 19–26 (1980).
29. Walker, A. M. and Blettner, M. 'Comparing imperfect measures of exposure', *American Journal of Epidemiology*, **121**, 783–790 (1985).