ELSEVIER

REVIEW

The Correction of Risk Estimates for Measurement Error

S. A. BASHIR, PhD, AND S. W. DUFFY, MSc

PURPOSE: The methods available for the correction of risk estimates for measurement errors are reviewed. The assumptions and design implications of each of the following six methods are noted: linear imputation, absolute limits, maximum likelihood, latent class, discriminant analysis and Gibbs sampling.
METHODS: All methods, with the exception of the absolute limits approach, require either repeated determinations on the same subjects with use of the methods that are prone to error or a validation study, in which the measurement is performed for a number of persons with use of both the error-prone method and a more accurate method regarded as a "gold standard."
RESULTS: The maximum likelihood, latent class and absolute limits methods are most suitable for purely discrete risk factors. The linear imputation methods and the closely related discrimination analysis method are suitable for continuous risk factors which, together with the errors of measurement, are usually assumed to be normally distributed.
CONCLUSIONS: The Gibbs sampling approach is, in principle, useful for both discrete and continuous risk factors and measurement errors, although its use does mandate that the user specify models and dependencies that may be very complex. Also, the Bayesian approach implicit in the use of Gibbs sampling is difficult to apply to the design of the case-control study. *Ann Epidemiol 1997;7:154–164.*
© 1997 by Elsevier Science Inc.

KEY WORDS: Risk estimates, measurement error, biostatistics, epidemiology review.

INTRODUCTION

Suppose an explanatory variable X needs to be measured. Often a true or exact measurement of X cannot be made, so a surrogate measure Z is used instead. This surrogate measurement (which can be made on discrete or continuous variables) is equal to X plus some error that will need to be taken into consideration in statistical analysis. Primarily there are two types of error: systematic errors and random errors. Systematic error can occur in many forms; an example would be a pair of scales that has not been set correctly and always overmeasures by a fixed amount. Random error can occur for various reasons; for example, one can cite temporal variation in subjects of some factor in the blood or the laboratory's measurement of this factor. (Note that control materials tend to be used in laboratories to calibrate measurement methods and prevent systematic error.) In some studies (e.g., case-control) there may be the additional problem of differential misclassification, when exposure measurement varies by outcome status (e.g., diseased or non-diseased).

Hereafter case-control studies will be discussed for the most part, although there are applications to other designs. Two possible designs can be used in the analysis of case-control studies: the "gold standard," where the imperfect measure is validated against a method assumed to be error free, and the "repeat measurement." Both designs can be internal or external. The external design is often used in a large study with precise estimates of error, but it entails more assumptions that introduce the constraint that prevalences in the external study and in the main study must be equal. The internal design may not have precise estimates of errors, but fewer assumptions are needed.

Measurement error has traditionally been modeled in two ways: the classical method and Berkson formulation. The classical measurement error formulation models the problem in terms of the conditional distribution of surrogates Z given the true measurement X. The Berkson (1) formulation models the problem in terms of the conditional distribution of X given Z.

From the International Agency for Research on Cancer, Lyon, France (S.A.B.); and the MRC Biostatistics Unit, Institute of Public Health, University Forvie Site, Cambridge, England.

Address reprint requests to: S. A. Bashir, International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon cedex 08, France.

Received May 24, 1995; accepted October 16, 1996.

655 Avenue of the Americas, New York, NY 10010

1047-2797/97/$17.00
PII S1047-2797(96)00149-4

TABLE 1. Observed frequencies by disease state, risk factor, and confounder status (measured situation—the table without the primes would represent the true situation)

	Confounder level 1		Confounder level 2	
	Exposed	Not exposed	Exposed	Not exposed
Cases	a_1'	b_1'	a_2'	b_2'
Controls	c_1'	d_1'	c_2'	d_2'

This review looks at primarily six methods for the correction of risk estimates for measurement errors: absolute limits, linear imputation, maximum likelihood, latent class analysis, discriminant analysis, and Gibbs sampling. Almost all analytic methods for dealing with this problem are variants of one of these six. In the description of individual methods in this review, the methods specifically designed for measurement error (such as linear imputation) are treated as fully as space permits. For more general methods, the description is limited to specific applications that commonly arise.

ABSOLUTE LIMITS

Walker and Lanes (2) developed the work of Cox and Elwood (3) and Blettner and Wahrendorf (4) to obtain the "Absolute limits" method. This method is best suited to categorical risk factors and is typically employed in the situation of a binary risk factor and a binary confounder (Table 1). The aim of this method is to find the limiting values (i.e., the maximum and minimum values that give physically possible expected counts) for the misclassification proportions given the observed data. This method is radically different from the other methods considered, in that only data from the study are needed. Validation or replication studies are not necessary. A description of the absolute limits method follows.

Let u be the proportion of persons properly members of stratum 2 who are correctly classified into stratum 2 by the imperfect measures and v be the corresponding proportion of stratum 1. From the definition of u and v, the misclassified counts can be represented in terms of the true (unknown) counts. For example, for exposed cases we have

$$a_1' = a_1 v + a_2(1 - u) \text{ and } a_2' = a_1(1 - v) + a_2 u. \quad (1)$$

Similar expressions hold for the other cells. These can be solved to give expressions for the true counts in terms of the misclassified counts u and v. For example,

$$\hat{a}_1 = \frac{a_1' u - a_2'(1 - u)}{u + v - 1}. \quad (2)$$

It is now possible to derive the true odds ratio (OR) for stratum 1 as

$$OR_1 = \frac{\hat{a}_1\hat{d}_1}{\hat{b}_1\hat{c}_1} = OR_1' \frac{\left(l - \frac{a_2'(1 - u)}{a_1' u}\right)\left(l - \frac{d_2'(1 - u)}{d_1' u}\right)}{\left(1 - \frac{b_2'(1 - u)}{b_1' u}\right)\left(l - \frac{c_2'(1 - u)}{c_1' u}\right)}, \quad (3)$$

where OR_1' is the observed OR for stratum 1.

Using equation 3 poses a problem because the true measures and the proportions u and v are unknown. However, the mismeasured values can be used to obtain the admissible values of u and v. The absolute limits are those that contain all values which give physically possible (i.e., nonnegative and no greater than the total number of cases or controls) quantities for the true counts.

By imposing the constraint that u and v are both nonnegative and using all equations of the type of equation 2 with the fact that each true measure has to be greater than or equal to zero, limits on u and v can be obtained, such as

$$v \geqslant \frac{a_1'}{a_1' + a_2'} \quad \text{if } u + v > 1 \quad (4)$$

and

$$v \leqslant \frac{a_1'}{a_1' + a_2'} \quad \text{if } u + v < 1 \quad (5)$$

The discussion above deals with nondifferential misclassification. This approach is expandable to differential misclassification. The needed formulae are given by Walker and Lanes (2), and further developments are given by Marinos and colleagues (5).

The absolute limits method appears at first sight to give maximum and minimum feasible values for the misclassification probabilities without making any assumptions. However, this is not strictly the case. The method assumes that misclassified cell counts cannot yield negative expected underlying true cell counts. That is, observed and expected misclassified cell counts are assumed to be equivalent. Although this assumption is necessary and justifiable, it should be noted that it can be broken. Suppose $u = v = 0.7$ and that the true cell counts are $a_1 = 50$ and $a_2 = 50$. It is possible in practice to observe $a_1' > 70$, although the probability thereof is fairly small. Such a situation would lead to the constraint that $v \geqslant \frac{a_1'}{a_1' + a_2'}$ (i.e., the true value is excluded). The likelihood that this method would lead to wildly inappropriate conclusions is low provided that the error rates are relatively low (of the order of $\leqslant 0.2$).

Example. Qizilbash and colleagues (6) performed a case-control study of stroke in Oxfordshire, U.K. The results with respect to risk factor lipoprotein (a) (Lp(a)) and confounding factor apolipoprotein B (apo-B) are shown in Ta-

TABLE 2. Case-control data classified by two binary factors: risk factor (RF) is lipoprotein (a) (LP(a)), and confounding factor (CF) is apolipoprotein B (apo-B)[a]

	CF+		CF−	
Group	RF+	RF−	RF+	RF−
Cases	24	14	25	21
Controls	47	38	27	51

[a] Plus sign indicates factor is present; minus sign indicates factor is absent.

ble 2. Note that Lp(a) and apo-B are continuous and have been dichotomized above and below the median to represent the factors being present and absent, respectively.

The feasible regions for u and v are as follows: for $u + v < 1$; $0 \leq u \leq 0.365$ and $0 \leq v \leq 0.4$ and for $u + v > 1$; $0.6 \leq u \leq 1$ and $0.635 \leq v \leq 1$.

LINEAR IMPUTATION

This method is most readily applicable to continuous explanatory variables. The linear imputation methods of correcting for measurement errors have been expounded by MacMahon and colleagues (7) and Rosner and colleagues (8). The two approaches are based on the same principle but use different strategies in tackling the underlying problem; therefore, they are reviewed separately.

MacMahon-Peto Method (Categorized Data)

For some continuous true risk factor X, two imprecise measurements Z_1 and Z_2 are made at two different points in time. Z_1 is measured before Z_2. Z_1 will be referred to as the baseline measure, and Z_2 as the repeat measure. Each member of the study population is categorized, often by quantiles of the distribution of the baseline measure. Relative risks are calculated from an overall logistic or log-linear regression analysis.

The slope of real association between the usual measure and disease is systematically underestimated due to the random error in the characterization of the risk factor. The top and bottom categories of the baseline measure include study subjects whose baseline measurements happen to be lower and higher respectively, than their usual levels. This bias causes a systematic dilution of the apparent importance of the risk factor. This bias is often referred to as regression dilution bias and can be corrected for in two ways.

The first method involves adjusting the range of the explanatory variable. For all individuals in a given category according to the first determination, we calculate the average value of X by the second determination. This average value at second determination, which is conditional on the category at the first determination, is our estimate of the true mean for those individuals. Suppose $Z_1 = X + \epsilon_1$ and $Z_2 = X + \epsilon_2$, where ϵ_1 and ϵ_2 are independent of each other and of X. Thus ϵ_1 and ϵ_2 are the measurement errors. Then, if ϵ_1 and ϵ_2 each have mean zero,

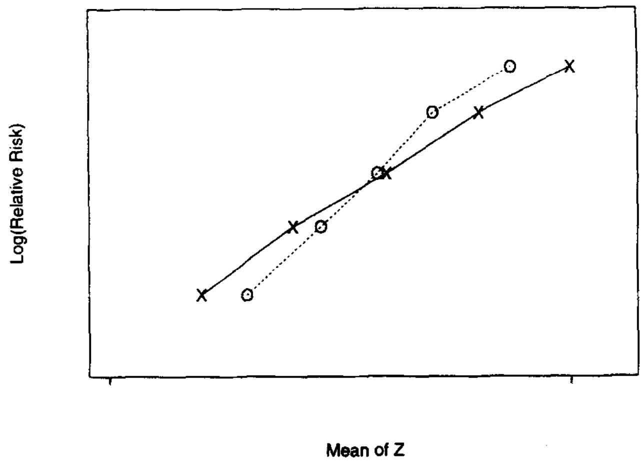

FIGURE 1. An example of the regression dilution bias (X = Z_1 and O = Z_2). O estimates the true level and so gives a corrected steeper slope to the risk gradient.

$$\begin{aligned} E(Z_2|Z_1\epsilon(a,b]) &= E(X + \epsilon_2 \,|X + \epsilon_1\epsilon(a,b]) \\ &= E(X|X + \epsilon_1\epsilon(a,b]) + E(\epsilon_2|X + \epsilon_1\epsilon(a,b]) \\ &= E(X|Z_1\epsilon(a,b]), \qquad (6) \end{aligned}$$

since ϵ_2 has mean zero and is independent of X and ϵ_1.

The mean of Z_2 conditional on category of Z_1 will not in general be the same as the mean of Z_1 but will be closer to the overall mean of all individuals. This phenomenon is known as regression to the mean. The method yields a steeper gradient to the risk function without adjusting the estimates of relative risk. This can be seen in Figure 1, where the values for Z_2 conditional on Z_1 estimate the true mean values in each category. Variance estimation has not been adequately addressed in the literature.

The second method involves explicit numerical adjustment of the slope of the risk function, and for this we require the relative risk to be estimated as a single linear trend. If, by the above method, we find that the range of the Z_2 means for the categories is a factor p times the range of the Z_1 means, we can multiply the trend parameter by $1/p$ to obtain a corrected trend. This can be intuitively seen by thinking of the uncorrected risk function as

$$\ln(RR) = \alpha + \beta_u X \qquad (7)$$

and

$$\beta_U \approx \frac{Max(\ln(RR)) - Min(\ln(RR))}{Max(\bar{Z}_1) - Min(\bar{Z}_1)}. \qquad (8)$$

But we want the true β estimated by β_T:

$$\beta_T = \frac{Max(\ln(RR)) - Min(\ln(RR))}{Max(\bar{X}) - Min(\bar{X})} . \qquad (9)$$

and as above $\bar{Z}_2$ conditional on $\bar{Z}_1$ is an unbiased estimator of X. Since $Max(Z_2) - Min(\bar{Z}_2) = p(Max(Z_1) - \mathrm{Min}(\bar{Z}_1)$, we have

$$\beta_T = \frac{\beta_U}{p} . \qquad (10)$$

Example. Lp(a) was used as a continuous risk factor from a case-control study of stroke in Oxfordshire (6). There were 97 cases and 165 controls with a single measurement and 66 controls with two measurements (validation group). Selection of three categories (tertiles with cut-off points at 68 and 249 international units) for the parametric method relative risks (relative to the baseline or lowest category) gives 1.603 (95% confidence interval (CI, 0.872–2.947) for category 2 and 2.025 (95% CI, 1.113–3.686) for category 3 (those with the highest level of Lp(a)). The means of the first measure with the three categories are 26, 153, and 710 international units. The means of the second measure conditional on the category of the first are 37, 183, and 587, respectively. The parametric method yields a corrected OR of 1.08 (95% CI, 1.00–1.17) compared with an uncorrected estimate of 1.07 (95% CI, 1.00–1.13).

ROSNER'S METHOD (CONTINUOUS RISK FACTOR)

Rosner and colleagues (8) described two methods, again applicable to continuous risk factors, for the correction of measurement error, namely the linear approximation method and the likelihood approximation method. These methods required a validation study, in which a sample of individuals are measured by the error-prone method and by a gold standard, assumed to be error-free.

Assume that the true exposure is related to probability of disease D in the form of the following logistic regression model.

$$\ln[Pr(D|X)/Pr(\bar{D}|X)] = \alpha^* + \beta^* X , \qquad (11)$$

and also assume there is a linear relationship between X and the surrogate measure (i.e., the determination that is subject to error) Z, so that

$$X = \alpha' + \lambda Z + \epsilon, \qquad \text{where } \epsilon \sim N(0,\sigma^2) . \quad (12)$$

It is assumed that $Pr(D|X,Z) = Pr(D|X)$, which in turn implies that the conditional distribution of Z given X is the same for subjects with and without the disease. We also 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. Using a main study population with known surrogate measures Z and an external validation study population with known values for the true and surrogate measure, the parameter β^* is to be determined.

We will describe the linear imputation method. Rosner and colleagues (8) also described the likelihood approximation method, which is based on the same principle but is more mathematically sophisticated.

There are two practical strategies within the linear approximation method that lead to an estimate of β^*, the imputation and linear procedures. It should be noted that both of these procedures lead to the same estimate of β^*.

The imputation procedure uses the validation study population to obtain ordinary least squares estimates for parameters in equation 12. These estimates are used to obtain expected values of X in the main study population (i.e., $E(X|Z) = \alpha' + \lambda Z$). These expected values are substituted into equation 11 to obtain β_{imp}, an estimate of β^* by ordinary logistic regression. The associated confidence interval for β^* will be too narrow because the errors in the estimation of X from Z are ignored.

The linear procedure obtains an estimate $\hat{\beta}$, the logistic regression of disease on observed exposure, from the main study. Then λ is estimated by ordinary least squares from equation 12 in the validation study. Now β^* is estimated as

$$\hat{\beta}_{lin} = \hat{\beta}_{imp} = \hat{\beta}/\hat{\lambda} . \qquad (13)$$

To obtain a confidence interval and associated OR for β^*, it is assumed that estimates of $\hat{\beta}$ and $\hat{\lambda}$ are independent random variables. The delta method (9) is used to obtain var $(\hat{\beta}_{lin})$,

$$var(\hat{\beta}_{lin}) = (1/\hat{\lambda}^2)var(\hat{\beta}_{imp}) + (\hat{\beta}^2_{imp}/\hat{\lambda}^4)var(\hat{\lambda}) , \qquad (14)$$

where standard methods from the logistic regression model (equation 11) and linear regression model (equation 12) are used to obtain estimates of *var* $(\hat{\beta})$ and *var* $(\hat{\lambda})$, respectively. This interval reflects the imprecision in the estimation of X given Z and hence is wider than that obtained by the imputation procedure.

Example. We again use the example of Lp(a) in the case-control study of stroke in Oxfordshire (6). The gold standard here is an average of two measurements on separate occasions. The linear approximation method yields a relative risk of 1.06 (95% CI, 1.00–1.12) compared with an uncorrected estimate of 1.07 (95% CI, 1.00–1.13).

MAXIMUM LIKELIHOOD

The maximum likelihood methods of correcting for measurement errors have been expounded by Duffy (10–12), Greenland (13–16) and Palmgren (17) and their colleagues. The first group of investigators aimed at analysis of unmatched studies, whereas the second also consider the problem in the context of matched case-control studies. These

TABLE 3. Case-control data classified by two binary factors: risk factor (RF) and confounding factor (CF)[a]

	CF+		CF−	
Group	RF+	RF−	RF+	RF−
Cases	*a*	*b*	*e*	*f*
Controls	*c*	*d*	*g*	*h*

[a] Plus sign indicates factor is present; minus sign indicates factor is absent.

methods are generally more applicable to discrete risk factors, and especially to binary factors.

Duffy's Method (Unmatched Studies)

The method proposed by Duffy, Rohan, and Day (10) is reviewed in this section. This method is typically employed in the situation of a binary risk factor and a binary confounder (Table 3) and repeat measurements (Table 4). It is assumed that the case-control status is measured without error but that misclassification can occur in the risk factor (RF) and confounding factor (CF). Table 4 includes a sample of subjects for whom a repeat measure of each factor is made.

Let p_a, p_b, . . ., p_h represent the true probabilities corresponding to risk factor (RF) and confounding factor (CF) status, i.e.,

$$p_a = Pr(\text{RF present, CF present} \mid \text{case})$$
$$p_b = Pr(\text{RF absent, CF present} \mid \text{case}),$$

and so on. Misclassification is assumed to be equally likely in either direction for both the risk factor and the confounding factor, i.e., the present and absent states are equally likely to misclassified. Let

$$\alpha = Pr(\text{RF classified correctly}), \text{ and}$$
$$\gamma = Pr(\text{CF classified correctly}).$$

Let p_a', p_b', . . . , p_h' be the probabilities of the observed states corresponding to the true states p_a, p_b, . . ., p_h. Conditional on α and γ, the probability of a case being observed with both risk and confounder present is

$$p_a' = \alpha\gamma p_a + (1 - \alpha)\gamma p_b + \alpha(1 - \gamma)p_e + (1 - \alpha)(1 - \gamma)p_f \qquad (15)$$

Similar equations can be derived for p_b', p_c', . . ., p_h'.

TABLE 4. Repeat data on risk and confounding factors

	First	Repeat determination	
Factor	determination	Present	Absent
Risk factor	Present	i_1	j_1
	Absent	k_1	l_1
Confounding factor	Present	i_2	j_2
	Absent	k_2	l_2

TABLE 5. Repeat data on risk and confounding factors—risk factor is lipoprotein (a) (Lp(a)), and counfounding factor is apolipoprotein B (apo-B)

	First	Repeat determination	
Factor	determination	Present	Absent
Risk factor	Present	30	8
	Absent	5	30
Confounding factor	Present	28	10
	Absent	9	23

Maximum likelihood estimates are

$$\hat{\alpha} = \frac{1}{2} \pm \frac{1}{2}\left(1 - \frac{2m_1}{M_1}\right)^{1/2} \text{ and}$$
$$\hat{\gamma} = \frac{1}{2} \pm \frac{1}{2}\left(1 - \frac{2m_2}{M_2}\right)^{1/2}, \qquad (16)$$

where $m_1 = j_1 + k_1$ and $M_1 = i_1 + j_1 + k_1 + l_1$. M_2 and m_2 are the corresponding quantities for the confounder. Estimates of p_a, . . . , p_h are calculated as the solution of two sets of four simultaneous linear equations.

Simple calculations lead to the point estimates with the constraints $\alpha > 0.5$ and $\gamma > 0.5$. The Mantel-Haenszel relative risk (18) estimate is

$$\hat{\psi} = \left[\frac{\hat{a}\hat{d}}{\hat{E}_1} + \frac{\hat{e}\hat{h}}{\hat{E}_2}\right] \Big/ \left[\frac{\hat{b}\hat{c}}{\hat{E}_1} + \frac{\hat{f}\hat{g}}{\hat{E}_2}\right], \qquad (17)$$

where $\hat{E}_1 = D_1(\hat{p}_a + \hat{p}_b) + D_2(\hat{p}_c + \hat{p}_d)$, $\hat{E}_2 = D_1(\hat{p}_e + p_f) + D_2(\hat{p}_g + \hat{p}_h)$, $D_1 = a + b + e + f$, and $D_2 = c + d + g + h$. $\hat{E}_1$ is the estimated true number of subjects with the confounder present and E_2 is the true number of subjects with the confounder absent.

Duffy and colleagues (10) used the profile likelihood, as described by Elton and Duffy (12), to obtain interval estimates.

Example. Using Tables 2 and 5 where the risk factor is Lp(a) and the confounding factor is apo-B, table 5 shows the risk and confounding factor status with respect to repeat data from an external validation study. The uncorrected OR estimate for Lp(a), adjusted for apo-B, is 1.79 (95% CI, 1.04–3.09) and the corrected odds ratio is 2.31 (95% CI, 1.15–4.64).

GREENLAND'S METHOD (MATCHED STUDIES)

In a matched pair case-control study, let a risk factor x have K levels, and let m_{ij} denote the number of pairs with $x = i$ for the case and $x = j$ for the control. Let π_{ij} be the probability that a case with true status $x = j$ is classified as $x = i$, and τ_{ij} the corresponding probability for a control.

Assume we have estimates p_{ij} and q_{kl} of π_{ij} and τ_{kl}; the study pairs form a separate random sample of all case-control

TABLE 6. Matched case-control study of breast cancer in Moscow

	Controls	
Cases	Family history	No family history
Family history	2	15
No family history	5	117

pairs, and measurements are independent within and across pairs. Let t_{ij} be the cell counts corresponding to m_{ij}.

The relation between **m** and **t** is $E(\mathbf{m}) = \Gamma\mathbf{t}$, that is,

$$E(m_{ij}) = \sum_{jl} \pi_{ij}\tau_{kl}t_{jl} \,. \tag{18}$$

Let **C** be the maximal likelihood estimate of Γ with corresponding elements $p_{ij}q_{kl}$, and $\mathbf{V}_m$ the estimated covariance matrix of **m**.

If C is invertible, equation 18 leads to an estimate of **t**: $\hat{\mathbf{t}} = \mathbf{C}^{-1}\mathbf{m}$. Greenland (13) has shown via Selen (19) that the asymptotic covariance matrix of $\hat{\mathbf{t}}$ is given to the first order by

$$\begin{aligned}\mathrm{Cov}^A(\hat{\mathbf{t}}) &= \mathrm{Cov}^A[\mathbf{C}^{-1}E(\mathbf{m})] \\ &\quad + E(\mathbf{C})^{-1}\mathrm{Cov}(\mathbf{m})E(\mathbf{C})^{-1}\,,\end{aligned} \tag{19}$$

where asymptotic covariances are denoted by Cov^A.

Formula 19 is discussed by Greenland (13), who goes on to say that if "variability in the estimated classification rates is negligible . . . the covariance-matrix estimate for **t** is simply $\mathbf{C}^{-1}\ \mathbf{V}_m\ \mathbf{C}^{-1T}$, ..." A form of $\mathbf{V}_m$ is also given by Greenland (13).

Example. Zaridze and colleagues (20) performed a case-control study of breast cancer in Moscow. Although not published, the results with respect to family history of breast cancer were as shown in Table 6. Validation results from another study suggest that the false-positive rate of reporting family history of breast cancer is zero, and the false-negative rate 0.15 (21). Assuming that these rates hold for both cases and controls, this gives $\pi_{11} = \tau_{11} = 0.85$; $\pi_{12} = \tau_{12} = 0$; $\pi_{21} = \tau_{21} = 0.15$; and $\pi_{22} = \tau_{22} = 1$.

The uncorrected OR estimate is 3.00 (95% CI, 1.09–8.25). Greenland's correction yields an OR of 3.15 (95% CI, 1.35–7.37).

LATENT CLASS ANALYSIS

The use of Latent Class Analysis for correcting measurement error was developed by Clayton (22), Kaldor and Clayton (23), and Liu and Liang (24). Although the latent class methods also yield maximal likelihood estimates, they use a very different algorithm from those in the foregoing section and entail a different practical approach. Hence we review them separately. Here we describe the simplest approach as given by Kaldor and Clayton (23).

The Latent Class/Logistic Model

Suppose there is one latent risk factor C measured by a surrogate *W*, and a disease state *D*. We describe the simplest case, where both C and *W* are dichotomous. This results in two 2 × 2 tables (one each for cases and controls), with cell counts represented by m_{dwc}, given the values of *d*, *w*, and *c*.

It is assumed that m_{dwc} is generated by a log-linear model of the form

$$\log\mu_{dwc} = \theta + \theta^D_d + \theta^C_c + \theta^{DC}_{dc} + \theta^W_w + \theta^{WC}_{wc}, \tag{20}$$

where $E[m_{dwc}] = \mu_{dwc}$, $\theta^D_0 = \theta^C_0 = \ldots = 0$ and $\theta^D_1 = \theta^D$, etc. The parameters in this model are unknown.

Defining π as the prevalence of the risk factor among controls, γ and δ as the sensitivity and specificity of W as a measurement of C, respectively, gives

$$\pi = \frac{(e^{\theta C} + e^{\theta C + \theta W + \theta CW})}{(1 + e^{\theta W} + e^{\theta C} + e^{\theta C + \theta W + \theta CW})} \tag{21}$$

$$\gamma = \frac{e^{\theta W + \theta CW}}{1 + e^{\theta W + \theta CW}} \tag{22}$$

$$\delta = \frac{1}{1 + e^{\theta W}} \,. \tag{23}$$

Equations 22 and 23 show that with this model, the sensitivity and specificity do not depend on disease status and hence correspond to nondifferential misclassification. Collapsing the table over *W* (i.e., combining *W* within the cases and controls) results in the OR

$$\psi = \frac{(\mu_{101} + \mu_{111})(\mu_{000} + \mu_{010})}{(\mu_{001} + \mu_{011})(\mu_{100} + \mu_{110})} = e^{\theta DC} \,. \tag{24}$$

Hence, the logarithm of the OR is θ^{DC} for the relationship between the latent class C and the disease status *D*. Estimation is via the EM algorithm (25).

Example. We shall use a fictitious example to illustrate this method. Assume that the true probabilities of a risk factor being present among cases and controls are 0.6 and 0.2, respectively. This gives us a true OR of 6. Further assume that there is a misclassification rate of 10%. From this information we can create a fictitious example where cases ($n = 100$) and controls ($n = 100$) are measured twice. Then we have 49 cases with risk factor present (RF+) twice and 33 with risk factor absent (RF−) twice, and nine cases who are RF+ at first measure and RF− at second measure (and vice versa). Similarly for controls, we have 17 controls who are RF + twice and 65 controls who are RF− twice and the nine of each with disagreements either way. This gives a corrected OR of 6.217 (95% CI, 3.311–11.677) compared with an uncorrected OR (corresponding to a single measure) of 3.93 (95% CI, 2.162–7.146).

TABLE 7. Variables used in the discriminant analysis approach

Variable	Description
D	Disease status (d = 0 or 1, controls and case respectively)
x_{ij}	True covariate value in stratum j for cases (i = 1) and controls (i = 0)
μ_j	Population mean among controls in stratum j of true covariate
Δ	Population within-stratum mean difference in true covariate values in cases and controls
Σ	Within-stratum covariance matrix of true covariate
z_{ijk}	k th measured value of x_{ij}
π_j	Disease probability given stratum j
Ω	Covariance matrix for error e_{ijk}

DISCRIMINANT ANALYSIS

The discriminant analysis method of correcting for misclassification errors was developed by Armstrong and colleagues (26). A discriminant analysis approach is used to relate normally distributed risk factors to the probability of being a disease case, with the risk factors subject to normally distributed measurement errors. The model allows for matching into strata.

Under the multivariate discriminant analysis model x_{ij} is MVN($\mu_j + i\Delta$, Σ) (i.e., has a multivariate normal distribution with mean $\mu + i\Delta$ and covariance matrix Σ). Measurements z_{ijk} repeated n_{ij} times, provide information on x_{ij}, where

$$z_{ijk} = x_{ij} + e_{ijk}, \qquad k = 1, \ldots, n_{ij}. \tag{25}$$

Variables used in the development of the method are defined in Table 7.

Here e_{ijk} are identically distributed independent of each other and of x_{ij} and $e_{ijk} \sim MVN\ (\gamma + i\delta, \Omega)$, γ is the systematic (mean) error, and δ is the systematic difference in error between cases and controls. For matching stratum j, reported intakes z_{0jk} and z_{1jk} from controls and cases respectively, are independent normal variates where $z_{0jk} \sim N(\mu_j + \gamma, \Sigma + \Omega)$ and $z_{1jk} \sim N(\mu_j + \gamma + \Delta + \delta, \Sigma + \Omega)$.

Suppose for simplicity that $n_{ij} = n$, constant across strata. It can be shown (26) that the true OR β is estimated as

$$\beta = \Lambda_n^{-1}\beta^* - \Sigma^{-1}\delta^T, \tag{26}$$

where

$$\Lambda_n = (I_p + n^{-1}\Sigma^{-1}\Omega)^{-1} \tag{27}$$

is the covariance matrix of the purely random (nonsystematic error).

Estimation

If in equations 26 and 27, Σ and Ω are known (and hence Λ_n), then $\beta^\ddagger$ can be estimated by the logistic regression of disease status on transformed covariates $\bar{z}_{.}\ \Lambda_n$. β can be estimated from equation (26) if δ is also known. The variance of $\beta^\ddagger$ can be estimated via the variance of $\hat{\beta}^*$ as follows:

$$v\hat{a}r(\hat{\beta}^\ddagger) = (\Lambda_n^{-1})v\hat{a}r(\hat{\beta}^*)(\Lambda_n^{-1})^T \tag{28}$$

This variance estimate leads to a confidence interval for $\beta^\ddagger$ and hence the OR in the usual manner (assuming $\beta^\ddagger$ is normally distributed).

In reality Ω and Σ are not known and need to be estimated by $\hat{\Omega}$ and $\hat{\Sigma}$, respectively. $\hat{\Omega}$ is estimated from the repeated measures either via subjects within the study or separately by an external validation group. $\hat{\Sigma}$ is estimated by subtracting $\hat{\Omega}$ from the estimated within-stratum variance $\hat{\Omega} + \hat{\Sigma}$ of the observed covariates, which can be obtained by MANOVA (multivariate analysis of variance). Estimators $\hat{\Omega}$ and $\hat{\Sigma}$ can be used to estimate Λ_n, which in turn lead to an estimate for $\hat{\beta}^\ddagger$ from $\hat{\beta}^*$. Armstrong and colleagues (26) suggested using resampling methods.

Example. Armstrong and colleagues (26) illustrated the discriminant analysis method by using a dietary case-control study of colorectal cancer conducted in Canada (27). These data included 171 cases and 171 controls, a repeat study on 52 subjects, and a validation study on 16 subjects. The investigators did a trivariate analysis using calories, protein, and fat. The uncorrected OR for calories was 1.07 (95% CI, 1.01–1.13), for protein 0.75 (95% CI, 0.56–1.00), and for fat 1.00 (95% CI, 0.90–1.11). The variance corrected estimate for calories was 1.21 (95% CI, 0.98–1.49), for protein 0.38 (95% CI, 0.06–2.35), and for fat 0.93 (95% CI, 0.78–1.11). The variance and recall bias-corrected odds for calories was 1.22 (95% CI, 0.99–1.51), for protein 0.26 (95% CI, 0.04–1.60), and for fat 0.95 (95% CI, 0.79–1.13).

GIBBS SAMPLING (BAYESIAN MODELS)

The Gibbs sampling methods for correction of measurement errors were developed by Thomas and colleagues (28) and Richardson and Gilks (29) and involve making inferences about the marginal distribution of the parameters (from their conditional distribution) of interest via the Gibbs sampler (30). It should be noted that in using the Gibbs sampling technique the model parameters are regarded as random variables with probability distributions. As a starting point, information about the prior distribution of the parameters is used in conjunction with Bayes' theorem to obtain the posterior distribution.

The material below summarizes a more detailed exegesis given by Richardson and Gilks (29). The relationship between some risk factors X and disease status Y is unknown. Exact measures of X are not available. However some surrogate measures Z and X are recorded. Estimation of the relative risk, which describes the relationship between the factors X and Y is required, via the surrogate measures of Z, taking into account all of the uncertainty in X.

Some information on the effects of the risk factors on the general population is assumed to be available at the start of the study to specify a prior distribution for X. Additional data are obtained throughout the study on X via the recording of the surrogate measures Z, which reduces the uncertainty in X. The disease status Y contains information on the relative risks, which is strengthened by the surrogates Z. It is assumed that the variables Y_i and Z_i are conditionally independent given X_i for any individual *i*.

The parameters that connect the risk factors, surrogates, and disease fall into three groups: (*i*) group β, the epidemiological parameters (e.g., relative risk, OR, or the logarithm of these), which model the link between risk factors and disease status; (*ii*) group γ, the measurement error parameters (e.g., measurement error variance), which model the link between the risk factors and their surrogates; and (*iii*) group π, the exposure parameters (e.g., population mean and variance), which models the population distribution of X.

The model components are therefore:

$$\text{disease model } [Y_i|X_i,C_i,\beta] \tag{29}$$

$$\text{measurement model } [Z_i|X_i,\lambda] \tag{30}$$

$$\text{exposure model } [X_i|C_i,\pi] \tag{31}$$

As a part of the Bayesian method, the parameters β, λ, and π require prior distributions (denoted by [β], [λ] and [π] respectively. To complete the structure, the joint distribution of all the variables is written as the product of the model conditional:

$$[\beta][\lambda][\pi]\prod_i[X_i|C_i,\pi]\prod_i[Z_i|X_i,\lambda]\prod_i[Y_i|X_i,C_i,\beta]\,. \tag{32}$$

Two further implications of the preceding equations are that Y_i is independent of all the X_i', $i' \neq i$, conditionally on X_i, C_i, and β and secondly that conditional on parameters λ and the true exposure X_i, the surrogate measures Z_i are independent among individuals.

INFERENCE

Model parameters, β, λ, π and unobserved data $\{X_i\}$ are referred to as parameters below. Primarily we are interested in the marginal posterior distribution of the epidemiological parameter β given the data (since the value of the other parameters is unknown). This results in a high-dimensional integral, which in many circumstances is not soluble. Now the Gibbs sampling method can be used to make inferences about the parameters of interest by generating samples from the joint posterior distribution of all the parameters (β, λ, π, and $\{X_i\}$) given the data.

The Gibbs sample method starts by choosing an arbitrary starting value for each parameter. Then, taking each parameter in turn, the starting value is updated by a new value from its conditional distribution (or density) given the data and the current value of all the other parameters in the model. A complete cycle of the Gibbs sampler consists of updating all the unknown variables in the model once. The updating cycle is repeated a large number of times, and the samples generated are eventually considered as being from the joint posterior distribution of all the parameters. To derive the conditional distribution for each parameter, one simply uses the fact that it is proportional to the product of the terms that contain that term in the joint distribution (31). Note that a variable iteratively sampled in this way tends to be a variable sampled from the joint unconditional distribution as the number of iterations become large. Either a validation group or a repeat measures strategy may be used. Each determined the exact form of the measurement model component above.

Example. Richardson and Gilks (29) simulated 1000 subjects with one measurement on instrument 1 and 200 subjects with two measurements on instrument 1 and one on instrument 2, where instrument 2 is more accurate than instruction 1. There are two risk factors, X (measured with error) and C (measured accurately). The true (simulated) relative risk for X (RR_x) is 2.45 and for C (RR_C) is 3.32, and the Gibbs sampling approach results in RR_X = 3.62 (2.27–6.42) and $\hat{RR}_C$ = 3.52 (2.68–4.47) (with the 95% credible interval in brackets, i.e., the 2.5% and the 97.5% percentiles). These can be compared to the relative risks from true covariates before error is incorporated, where RR_X^T = 3.17, and RR_C^T = 3.38. The estimates for X are somewhat higher than the true value. The corresponding risk estimates for C are similar to the true value.

DISCUSSION

Although the potential bias from mismeasurement is well known, analysis which takes account of mismeasurement is comparatively rare for two reasons. First, many researchers are unsure of the reliability of methods of analysis that take account of mismeasurement. Second, such nonstandard statistical analysis is difficult. The following are therefore crucial questions for work in this area. (*i*) How do we obtain the results of analyses correcting for mismeasurement credible? (*ii*) How do we make the analyses accessible and easy to perform?

The first question concerns the reliability of estimates. The answer must lie in giving estimates from different methods, relying on different assumptions, and perhaps more importantly on confidence intervals that reflect uncertainty in the estimation of mismeasurement parameters as well as in the relative risk parameters of the main study. The second question reflects the absence of user-friendly software, which is an important target for research in this area. Bashir and

TABLE 8. Summary table of correction methods

Method, reference	Design, limitations	Types of variables	Assumptions	Ease of use
Absolute limits, (2)	Case-control studies; no need for validation or replications.	Discrete; dichotomous variables preferred.	Observed misclassified cell counts giving negative expected actual cell counts are impossible; nondifferential misclassification.	Readily applicable with minimal computing for nondifferential error case; becomes very convoluted if data are complex or differential error is assumed.
Linear imputation (7)	Cohort studies with repeat determination preferred.	Continuous	Conditional independence; correction known absolutely.	Readily applicable with minimal computing.
Linear imputation (8)	External validation; prospective or retrospective studies.	Continuous; normally distributed risk factor and measurement error.	Relationship between true and measured value, same in main and validation study.	Point estimate, easy; interval estimation, more difficult.
Maximum likelihood (10, 11)	Case-control (unmatched), with external repeat determination.	Categorical (best for bindary risk and confounding factors).	Conditional independence of repeat classification; independence between misclassification of each factor; symmetric misclassification for external validation only.	Point estimate, easy; exact interval estimation is very computer intensive; approximations are available.
Maximum likelihood (13, 14)	Matched-pair case-control studies, expandable to cohort studies; external validation.	Categorical risk factor.	Independence of classification of matched pairs; can deal with differential and asymmetric misclassification.	Readily applicable; point and interval estimation is easy.
Discriminant analysis (26)	Case-control with repeat data; external validation needed for bias between cases and controls.	Multivariate normally distributed data (continuous).	Repeated classifications are conditionally independent; nondifferential random error; systematic bias possible.	Versatile but algebraically complex; reduces to Rosner's linear imputation method in special case.
Latent class analysis (23)	Case-control studies (in principle adaptable to cohort studies); internal repeat data.	Categorical.	Conditional independence of repeat determination.	Point estimate, easy; interval estimation is computer intensive; approximations are vailable.
Gibbs sampling (29)	Theoretically no limitations; in practice, case-control design raises difficulties.	Discrete and/or continuous.	Model parameters are random variables with probability distributions; conditional independence of repeat data; broad regularity assumptions.	Depends on whether the Gibbs sampler converges or not; computer intensive.

Duffy (32) have written software to implement the methods of absolute limits (2), linear imputation (7, 8), maximal likelihood (10, 13), and latent class analysis (a simple case) (23). Rosner and collaborators (31) have written some software for the method of linear imputation. Gilks and colleagues (33) described a Bayesian software package (BUGS, Bayesian inference using Gibbs sampling), which can be used to apply the Bayesian method of correcting for measurement error (34, 29). Use of this software will require some programming, however. There is no standard software to implement the discriminant analysis method in a single analysis.

Where possible, real data have been used to demonstrate the above methods. These demonstrations illustrate the range of possible measurement error problems and result in corrections ranging from 4% to 180%. Note also that what appear to be minor changes in slope per unit (for example Lp(a) in the discussion of the McMahon-Peto method) represent major changes in slope in the logarithmic scale. If the relationships with risk were described per 10 units instead of per single unit, the correction would appear to have considerably greater magnitude.

The methods reviewed above differ in terms of the appropriate field of application and of the assumptions involved. All of the methods assume that the disease status is error-free. All of the methods involving repeat determinations by the error-prone method involve the assumption of independence of repeat determinations conditional on the true value. Methods involving external validation against a "gold standard" (e.g., Rosner's methods of linear imputation) assume that the relationship between true and observed values is the same in the main study as in the validation study. As commonly used, most methods assume nondifferential mismeasurement. The absolute limits methods assume that the observed and the expected cell counts are equivalent. In theory, all methods could be adapted for differential mismeasurement (i.e., different error distributions for those with and without disease), but the methodology is most developed for this use in the absolute limits and the discriminant analysis approaches.

The maximum likelihood, latent class, and absolute limits methods are most suitable for discrete risk factors. The linear imputation methods and the closely related discriminant analysis method are suitable for continuous risk factors which, together with the errors of measurement, are usually assumed to be normally distributed (although this is not necessary for the McMahon-Peto method). The Gibbs sampling approach is in principle useful for both discrete and continuous risk factors and measurement errors, although this usage does entail that the user specify models and dependencies that may be very complex. Also, it is difficult to apply the Bayesian approach implicit in the use of Gibbs sampling to the design of the case-control study.

These attributes are summarized in Table 8. In principle, Gibbs sampling provides the most flexible approach in terms of the type of data used, but the other methods have advantages in terms of ease of computing and simplicity. In practice, the method chosen will depend partly on the problem at hand and partly on the inclination of the individual researcher, but mostly on the tools immediately available to perform the analysis.

For the duration of this work, S. A. Bashir was supported by a studentship from the Medical Research Council.

REFERENCES

1. Berkson J. Are there two regressions? J Am Stat Assoc. 1950;45:164–180.
2. Walker AM, Lanes SF. Misclassification of covariates. Stat Med. 1991;10:1181–1196.
3. Cox B, Elwood JM. The effect on stratum specific odds ratios of nondifferential misclassification of a dichotomous covariate. Am J Epidemiol. 1991;28:202–207.
4. Blettner M, Wahrendorf J. What does an observed relative risk convey about possible misclassification? Methods Inf. Med. 1984;23:378-40.
5. Marinos AT, Tzonou AJ, Karantzas ME. Experimental quantiles of epidemiological indices in case-control studies with non-differential misclassification. Stat Med. 1995;14:1291–1306.
6. Qizilbash N, Duffy SW, Rohan TE. Repeat measurement of case-control data: Correcting risk estimates for misclassification due to regression dilution of lipids in transient ischaemic attacks and minor ischaemic strokes Am J Epidemiol. 1991;133:832–838.
7. MacMahon S, Peto R, Cutler J, et al. Blood pressure, stroke, and coronary heart disease: Part 1. Lancet. 1990;335:765–774.
8. Rosner B, Willet WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med., 1989;8:1051–1069.
9. Lindley DV. Introduction to Probability and Statistics from a Bayesian viewpoint (Part 1). Cambridge: Cambridge University Press; 1969.
10. Duffy SW, Rohan TE, Day NE. Misclassification in more than one factor in a case-control study: A combination of Mantel-Haenszel and maximum likelihood approaches. Stat Med. 1989;8:1529–1536.
11. Duffy SW, Maximovitch DM, Day NE. External validation, repeat determination, and precision of risk estimation in misclassified exposure data in epidemiology. J Epidemiol Community Health. 1992;46:620–624.
12. Elton RA, Duffy SW. Correcting for the effect of misclassification bias in a case-control study using data from two different questionnaires. Biometrics. 1983;39:659–665.
13. Greenland S. On correcting for misclassification in twin studies and other matched-pair studies. Stat Med. 1989;8:825–829.
14. Greenland S. The effect of misclassification in matched-pair case-control studies. Am J Epidemiol. 1982;116:402–406.
15. Greenland S. The effect of misclassification in the presence of covariates. Am J Epidemiol. 1980;112:564–569.
16. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched studies. Int J Epidemiol. 1983;12:93–97.
17. Palmgren J, Ekholm A. Exponential family nonlinear models for categorical data with errors of observation. Appl Stochastic Models Data Analysis. 1987;3:111–124.
18. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. J Nat Cancer Inst. 1959;22:719–748.
19. Selen J. Adjusting for errors in classification and measurement in the

analysis of partly and purely categorical data. J Am Stat Assoc. 1986;81–75.

20. Zaridze D, Lifanova Y, Maximovitch D, Day NE, Duffy SW. Diet, alcohol consumption and reproductive factors in a case-control study of breast cancer in Moscow. Int J Epidemiol. 1991;48:493–501.

21. Duffy SW, Roberts MM, Elton RA. Risk factor for breast cancer: Relevance to screening. J Epidemiol Community Health. 1983;37: 127–131.

22. Clayton D. Using test-retest reliability data to improve estimates of relative risk; An application of latent class analysis. Stat Med. 1985;4:445–455.

23. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. Stat Med. 1985;4:327–335.

24. Liux, Liang K-Y. Adjustment for non-differential misclassification error in the generalized linear model. Stat Med. 1991;10:1197–1211.

25. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. J Roy Stat Soc [B]. 1977;39:1–38.

26. Armstrong BG, Whittemore AS, Howe GR. Analysis of case-control data with covariate measurement error: Application to diet and colon cancer. Stat Med. 1989;8:1151–1163.

27. Jain M, Davies GM, Grace FG, Howe GR, Miller AB. A case-control study of diet and colo-rectal cancer. Int J Cancer. 1980;26:757–768.

28. Thomas DC, Gauderman J, Kerber R. A non-parametric Monte Carlo approach to adjustment for covariate measurement errors in regression analysis. Biometrics (in press). 1997.

29. Richardson S, Gilks WR. A Bayesian approach to measurement error problems in epidemiology using conditional independence models. Am J Epidemiol. 1993;138:430–442.

30. Casella G, George EI. Explaining the Gibbs Sampler. The American Statistician. 1992;46:167–174.

31. Rosner B, Spiegelman D, Willet WC. Correction of logistic regression relative risk estimates and confidence interval for random within-person measurement error. Am J Epidemiol. 1992;136:1400–1413.

32. Bashir SA, Duffy SW. Correction of risk estimates for measurement error in epidemiology. Methods Inf Med. 1995;34:503–510.

33. Gilks WR, Thomas A, Spiegelhalter DJ. A language and program for complex Bayesian modelling. The Statistician, 1994;43:159–177.

34. Richardson S, Gilks WR. Conditional independence models for epidemiological studies with covariate measurement error. Stat Med. 1993;12:1703–1722.