J Clin Epidemiol Vol. 41, No. 9, pp. 923–937, 1988
Printed in Great Britain.

0895-4356/88 $3.00 + 0.00

ESTIMATION OF TEST ERROR RATES, DISEASE PREVALENCE AND RELATIVE RISK FROM MISCLASSIFIED DATA: A REVIEW

S. D. WALTER[1] and L. M. IRWIG[2]*

[1]Department of Clinical Epidemiology & Biostatistics, McMaster University Medical Centre, Hamilton, Ontario, Canada L8N 3Z5 and [2]Department of Public Health, University of Sydney, N.S.W. 2006, Australia

(*Received* 28 *October* 1985; *in revised form* 2 *November* 1987)

Abstract—We review methods for the analysis of categorical clinical and epidemiological data, in which the observations are subject to misclassification. Under certain conditions, it is possible to estimate error parameters such as sensitivity, specificity, relative risk, or predictive value, even though no definitive classification (gold standard) is available. The parameter estimates are obtained by modelling the data, using maximum likelihood, with or without some constraints. The models recognize that the true classification of an individual is unknown, and so are sometimes referred to as "latent class" models.

The latent class approach provides a unified framework for various methods found in a dispersed literature, characterising each by the number of populations or subgroups in the data, and the number of observations made on each individual; the statistical degrees of freedom are implied by the sampling design. Data sets with less than three replicate observations per individual necessarily require constraints for parameter estimation to be possible. Data sets with three or more replicates lead directly to estimates of the misclassification rates, subject to some simple assumptions.

Some more complex problems are also discussed, including data where the response variable has more than two levels, sequential and irregular designs and the effects of assumption violations.

Biostatistical methods Misclassification Sensitivity Specificity Relative risk
Prevalence Error rates Reliability Models Categorical data Contingency tables

1. INTRODUCTION

Clinicians and epidemiologists often measure or classify individuals according to the presence or absence of a disease, signs or symptoms, or exposure to risk factors. In practice, errors in measurement or classification can occur for many reasons, including the use of subjective clinical judgment, technical imperfections of a diagnostic test, memory loss, deliberate misstatement or interpretational errors by interviewers or patients, and clerical errors.

Errors of measurement or misclassification in exposure variables, outcomes or confounders lead to bias in estimated indices of association such as the relative risk, and distortion of the p-values of their statistical tests of significance. Even modest probabilities of misclassification can have a substantial impact [1–5].

The ideal way of assessing the probability of misclassification with a particular method of observation is to compare it to a "definitive" or "gold standard" method which is error free [6]. For diagnostic testing, this is often not possible because of cost or risk to the subjects or ethical considerations. It has been found that in about 1/3 of medical articles describing diagnostic test evaluation, no well-defined gold standard was used [7]. Similarly, for epidemiologic risk factors, there is often no definitive method of measurement available.

In situations where error-free measures are difficult to obtain, reliability (i.e. reproducibility or repeatability) is often assessed by comparing the classifications of a set of individuals by two

* Based in part on work done at the Institute for Biostatistics of the South African Medical Research Council.

or more observers or methods. It is sometimes tempting to assume that one of the methods is indeed error free, for instance that a senior clinician always makes an accurate diagnosis, at least relative to junior students! However such an assumption clearly biases the estimated error rates in the other observations.

Another strategy might be to analyse the level of inter-observer agreement, perhaps compared to the majority opinion [8, 9]. This method is useful if the number of observers is large, when the estimation of all the observer-specific error rates may be computationally difficult or unstable. It is also possible to identify those observers who agree least often with the majority. Such observers are not necessarily any less accurate, but may be using different criteria for their classification [9].

One difficulty in the interpretation of agreement analyses is that the commonly-used indices of agreement, such as the kappa statistic [10] depend on the true prevalence of the attribute in the data [11–13]. Kappa will tend to be lower in populations where the attribute prevalence is very high or very low, even if observer error rates remain constant. The same difficulty applies to other indices of agreement [14], making the analysis of agreement an inherently less attractive option than direct estimation of misclassification rates, if the latter is feasible [11].

In this paper we review various ways in which observer error rates can be estimated in some generic clinical and epidemiological settings. Unless otherwise stated, we assume that *all* of the measurements are subject to potential error. The misclassification probabilities of the observers, will be estimated using various models which assume the "true" classifications to be unknown values.

In many situations, there is also interest in estimating the prevalence of an attribute (e.g. disease, symptoms, risk factor exposure), or in functions of prevalence such as relative risk. The estimability of these parameters depends on the number of populations or distinguishable groups of individuals sampled, the pattern of observations (especially the number of observations made on each individual), and how many statistical constraints are imposed. It will be shown that most of the methods which have been suggested for data with one or two observers require constraints in order to render a subset of parameters estimable. In contrast, when there are three or more observers, all the relevant parameters can be estimated without constraints.

2. TERMINOLOGY AND NOTATION

Suppose that in a population, a proportion θ of persons are actually "positive" for an attribute of interest; θ is known as the prevalence. For a given observer, the probability that an individual who is actually negative will be classified positive will be denoted by α, this being the false positive rate for that observer; $1-\alpha$ is known as the specificity, the probability that a truly negative individual is correctly classified. The corresponding parameters for truly positive individuals are β (the false-negative rate), and the sensitivity $1-\beta$, (the probability of correct classification for truly positive individuals). This follows the notation of Hui and Walter [15].

We will usually refer to the classifications as "observations"; depending on the context, an observation might be a diagnostic test, one of several alternative data sources (e.g. patient interview or hospital records), or a repeat classification by the same method on a different occasion. For situations where there is more than one observer, parameters will be subscripted accordingly.

Also of interest are the predictive values. The positive predictive value PV+ is the probability that an individual observed as positive is actually a true positive, and PV− is the probability that an apparently negative individual is truly negative. PV+ and PV− depend on the prevalence θ, and so may vary between populations, even though the sensitivity and specificity remain constant [16, 17]. PV+ and PV− are useful when one is interested in the likelihood of a correct classification of individual subjects in a specified population. However, sensitivity and specificity are more useful as indicators of the reliability of the observations as a whole.

We denote by n_{ijk} the number of individuals in a particular combination of classifications by a set of observers, with $i, j, k \ldots = 1$ denoting the positive results, and 0 the negative results. The total number of individuals will be denoted by N.

As a simple example, Table 1 shows the cross-classification of N individuals, according to a fallible observer 1, and according to a definitive observation 2 made without error. The true prevalence of the attribute is $\theta = (n_{11} + n_{01})/N$, the proportion of individuals

Table 1. Comparison of a fallible observer (1) with an error-free classification method (2)

		Error Free Observation 2		
		+	−	Total
Observer 1	+	n_{11}	n_{10}	$n_{11}+n_{10}$
	−	n_{01}	n_{00}	$n_{01}+n_{00}$
		$n_{11}+n_{01}$	$n_{10}+n_{00}$	N

Prevalence: $\theta = (n_{11}+n_{01})/N$
False positive rate $= \hat{\alpha} = n_{10}/(n_{10}+n_{00})$
False negative rate $= \hat{\beta} = n_{01}/(n_{11}+n_{01})$
Positive predictive value: $\hat{PV+} = n_{11}/(n_{11}+n_{10})$
Negative predictive value: $\hat{PV-} = n_{00}/(n_{01}+n_{00})$

classified positive by the definitive method. Similarly the false positive and false negative rates can be estimated as $\hat{\alpha} = n_{10}/(n_{10}+n_{00})$ and $\hat{\beta} = n_{01}/(n_{01}+n_{11})$ respectively; their denominators are based on the error-free classification according to observation 2. Finally, PV+ and PV− may be estimated, using the fallible observations 1 to define appropriate denominators. The parameters may be estimated directly in this situation, because of the availability of the error-free method 2. However in the more typical case where all the observations are potentially misclassified, a more general approach is needed, as is described in the following section.

3. REVIEW OF PROBLEMS INVOLVING MISCLASSIFICATION

The number of observers (R) and the number of populations or sub-populations which are sampled (S) determine the number of cross-classifications into which the data are grouped, and hence the number of degrees of freedom (df) which are available for parameter estimation. We will review estimation methods for a variety of typical situations, each characterised primarily by their values of R and S; for each problem we will calculate the number of independent parameters involved, and the number of df implied by the sampling design. If there are too many parameters to be estimated from the available df, constraints must be applied, for instance by regarding some of the parameters as known constants.

3.1. General problem: R observations per individual; S populations

We first restrict attention to binary observations. (Extensions to multilevel responses are given in Section 4.2). In the most general case, the prevalence and the sensitivity and specificity of the observers may vary across the S populations. For example, suppose the (sub)-populations are defined as women in various age groups, and the observations are to detect breast cancer using a mammographic screening device. Because of changes in breast tissue mass and density with age, the error rates may depend on age.

For situations with a single binomial response and arbitrary numbers of observers and populations (R and S), there are R false positive rate parameters, R false negative rate parameters and a prevalence parameter associated with each population, or $S(2R+1)$ parameters in total over all populations. If the observations are independent between observers, there are 2^R possible combinations of results for each subject. Regarding the sample sizes as fixed gives $2^R - 1$ df from each population, or $S(2^R-1)$ df in total. Estimability of all the parameters therefore depends essentially on R, because S is a common factor. The number of parameters and df for the first few values of R in one population ($S = 1$) are:

Number of observers R :	1	2	3	4	5 ...
Number of parameters $2R+1$:	3	5	7	9	11 ...
Number of df 2^R-1:	1	3	7	15	31 ...

Thus $R = 3$ is the minimum number of observers for which all parameters may be estimated without further assumptions, for any number of populations.

When all of the observers are imperfect on sensitivity and/or specificity, the "true" state of each person remains unknown. However the log likelihood of the data can be expressed as

$$L = \sum_{s=1}^{S} \sum_{\mathbf{x}} n_s(\mathbf{x}) \ln\left[\theta_s \prod_{r=1}^{R} \beta_{rs}^{1-x(r)}(1-\beta_{rs})^{x(r)} + (1-\theta_s)\prod_{r=1}^{R} \alpha_{rs}^{x(r)}(1-\alpha_{rs})^{1-x(r)}\right]. \quad (1)$$

Here α_{rs} and β_{rs} denote the false-positive and false-negative rates for observer r in population s, $x(r)$ denotes the classification of an individual by observer r, and the second summation is over all combinations of observations by the set of observers; $n_s(\mathbf{x})$ is the number of individuals in population s who receive a given set of classifications $\mathbf{x}$. This likelihood supposes that

Table 2. Summary of methods for estimation of sensitivity, specificity, prevalence and related parameters (binary data)

Author(s) (Ref.)	Examples of sampling design or application	No. of observers (R)	No. of pop. (S)	df	Parameters	No. of parameters (P)	Constraints	No. of constraints (C)	$P-C$	Estimated parameters
Quade *et al.* [3] Landis and Koch [8, 9] Walter [12] Dawid and Skene [18] White and Landis [20]	Estimation of sensitivity, specificity and prevalences; observer agreement analyses	$\geqslant 3$	S	$S(2^R-1)$	$\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}$	$S(2R+1)$	None	0	$S(2R+1)$	All
Quade *et al.* [3] Rogan and Gladen [28]	Cross-sectional survey with a fallible measurement; prevalence estimate required	1	1	1	α, β, θ	3	α, β known	2	1	θ
Gart and Buck [32] Greenberg and Jekel [33] Staquet *et al.* [34]	Simultaneous use of two diagnostic tests; comparison of a new test with an established test	2	1	3	$\alpha_1, \alpha_2, \beta_1, \beta_1, \theta$	5	α_2, β_2 known	2	3	$\alpha_1, \beta_1, \theta$
Chinn and Burney [38]	Estimation of "average" probability of correct classification	2	1	3	$\alpha_1, \alpha_2, \beta_1, \beta_2, \theta$	5	$\alpha_1 = \alpha_2 = \beta_1 = \beta_2$	3	2	α, θ
Goldberg and Wittes [29]	Estimation of true number of disease cases from population screening data	2	1	3	$\alpha_1, \alpha_2, \beta_1, \beta_2, \theta$	5	$\alpha_1 = \alpha_2 = 0$	2	3	β_1, β_2, θ
Barron [39]	Estimation of odds ratio associating two clinical entities	2	1	3	$\alpha_1, \alpha_2, \beta_1, \beta_2$ and 3 cell probabilities	7	$\alpha_1, \alpha_2, \beta_1, \beta_2$ known	4	3	Odds ratio

Murphy [40]	Cross-classification of disease state by a risk attribute; estimation of predictive values and relative risk; misclassification of risk attribute only	2	1	3	α, β, θ	3	None	0	3	α, β and θ or PV+ PV− and relative risk
Yanagawa and Gladen [42]	Repeat classification of disease state at two different times; estimation of incidence and remission rates	2	1	3	α, β, baseline prevalence θ_1 incidence rate (I), remission rate (R)	5	(i) α, β known (ii) $R = 0$, $\alpha = 0$	2 2	3 3	θ_1, I, R β, θ_1, I
Yanagawa and Gladen [42]	Repeat classification of state at 3 time points	3	1	7	as above	5	None	0	5	All
Chinn and Burney [38]	Classification of disease state at two different times; estimation of "average" probability of correct classification	2	1	3	α_1, α_2, β_1, β_2, θ_1, θ_2	6	$\alpha_1 = \alpha_2 = \beta_1 = \beta_2$	3	3	α, θ_1, θ_2
Copeland *et al.* [47]	Case–control odds ratio estimation, corrected for misclassification of risk exposure	1	2	2	α, β, θ_1, θ_2	4	α, β known	2	2	θ_1, θ_2 and Odds ratio
Hui and Walter [15]	Comparison of two diagnostic tests; odds ratio and relative risk estimation	2	2	6	α_1, α_2, β_1, β_2, θ_1, θ_2	6	None	0	6	All
Green [57]	Prospective relative risk estimation when only a subset of individuals have definite disease classification	2	2	3	α, β; disease incidences I_1 and I_0	4	I_0 small	1	3	Relative risk and PV+

Table 3. Assessment of pleural thickening by three independent radiologists for 1692 males

Reader*			Number of men	Frequency notation
1	2	3		
−	−	−	1513	n_{000}
−	−	+	21	n_{001}
−	+	−	59	n_{010}
−	+	+	11	n_{011}
+	−	−	23	n_{100}
+	−	+	19	n_{101}
+	+	−	12	n_{110}
+	+	+	34	n_{111}

* + denotes "positive" (pleural thickening present); − denotes "negative" (pleural thickening absent).

(i) all observers observe all subjects in all populations; (ii) the errors of classification are independent between subjects; (iii) the errors of classification are independent within subjects and between observers, conditional on the true state; and (iv) the sample sizes in each population are regarded as fixed. These assumptions are required for all methods reviewed, unless otherwise indicated.

The EM algorithm [18, 19] has been suggested to estimate the parameters of model (1). By adopting initial probabilistic estimates of each subject's true (but unknown) state, provisional estimates of the misclassification probabilities and the prevalence may be obtained directly, as a simple generalisation of the situation in Table 1; these estimates are then used to calculate improved estimates of the true status, and the process is iterated until convergence occurs. An alternative approach is to use the GSK methodology [8, 9, 20]. Finally the same model has been used in a "latent class analysis" using logistic regression [21]. The term "latent class" refers to the fact that the true state variable is always hidden or unknown, even though probabilistic estimates can be made for it.

Increasing the number of observers above 3 will cause an excess of *df*. For instance, with four observers there is an excess of 6 *df* over the 9 required for the parameters. These additional *df* may be used for a goodness-of-fit χ^2 test of the model (cf. Section 5.1). The amount of computation for parameter estimation goes up rapidly with the number of observers. Using three to five observers seems a reasonable compromise, this being enough to allow complete parameter estimability, but not so large as to pose computational problems.

Table 2 gives a synposis of methods of estimation of misclassification probabilities, disease prevalence and related parameters. This is for the general case as discussed above and for which an example is given below, as well as for other situations where there are fewer than 3 observers, as discussed in Section 3.3.

3.2. Examples with three or more observations per individual ($R \geqslant 3$)

Data of this kind can occur in a number of ways. First, several diagnosticians may independently classify a set of patients; for example, if nurses, radiologists and other physicians carry out screening diagnoses for cancer [22]. Similarly, Dawid and Skene [18] studied five anaesthetists who rated the same patients on their fitness for surgery. A second possibility is where the same diagnostic test may be used several times, e.g. the recommended sequence of six stool guaiac tests for colon cancer [23]. Third, there may be several different diagnostic tests for the same disorder, which may be used simultaneously or in sequence, e.g. Mantoux, tine, imotest, and "monovacc" methods for tuberculin sensitivity [24].

Consider the numerical example of Table 3. Three experienced readers independently evaluated the chest X-rays of 1692 male employees in asbestos mines and mills, taken at each worker's annual examination. Using the ILO U/C International Classification of Radiographs of Pneumoconioses [25], the readers assessed the presence/absence of pleural thickening [26]. Hence there are $R = 3$ observers in $S = 1$ population, implying 7 parameters in total. Under the independence assumption, the probability that a true positive individual is classified positive by all three readers is $(1-\beta_1)(1-\beta_2)(1-\beta_3)$, where β_r is the false-negative probability for observer r ($r = 1, 2, 3$). Similarly, the probability that a true negative individual is classified in this way is $\alpha_1\alpha_2\alpha_3$. Using a similar argument for each combination of classifications, as in the general likelihood equations (1), allows the complete likelihood to be computed. After numerical maximisation, we obtain parameter estimates, with an approximate variance–covariance matrix derived by standard maximum likelihood (ML) methods.

For these data, the parameters and their approximate standard errors are given in the first two rows of Table 4. We may note that observers 1 and 3 have very similar error probabilities, and that observer 2 has both higher false positive and false negative rates. This is also

Table 4. Parameter estimates for the data of Table 3

	Parameter						
	$\hat{\alpha}_1$	$\hat{\alpha}_2$	$\hat{\alpha}_3$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\theta}$
ML estimate	0.011	0.035	0.010	0.235	0.356	0.251	0.054
Standard error	0.004	0.005	0.003	0.112	0.171	0.119	0.023
Majority agreement estimate	0.014	0.037	0.013	0.145	0.250	0.158	0.045

reflected in a lower level of chance-corrected agreement, as measured by κ [10], in pairs involving observer 2. These results show either that observer 2 is inherently less accurate, or that he is using different diagnostic strategies, despite the attempts to standardise observers in their operational definitions of "abnormality". The findings are also consistent with correlated errors for readers 1 and 3.

Numerical iteration is required in the ML estimation process, and so it is useful to have suitable starting values for each parameter. We used initial estimates based on the majority opinion among the observers. For instance, the proportion of subjects with at least two "positive" X-ray assessments was used as the initial estimate of prevalence. Similarly the initial false positive rate for each observer was taken as the proportion of times he rated positive among subjects where the other two assessments were negative. These estimates (also shown in Table 4) gave satisfactory ML convergence after three iterations, which is typical of our experience with other data sets.

In some small or ill-conditioned data sets, convergence may be faster and the solution more stable if it can be assumed that the observations all have the same sensitivity and specificity. An example with 3 similar measurements is given by Quade *et al.* [3], who also provide a simple iterative computing algorithm for this case.

3.3. Methods for data with less than three observations per individual ($R < 3$)

As indicated in Section 3.1, for $R < 3$ observers there are insufficient *df* to permit the simultaneous estimation of all the parameters. Nevertheless there are a number of common situations where estimates are required of sensitivity, specificity, or prevalence, but without the luxury of three separate observations. For instance, one may wish to assess the performance of only one or two diagnostic tests; and even in studies of agreement in subjective diagnosis, there may be occasions when only two observers are available. For such problems, several methods have been suggested to estimate subsets of parameters after the imposition of certain constraints. The form of these constraints varies, but a common option is to regard some parameters as known, and then to estimate the remainder. If the number of parameters is p, and the number of constraints is c, then in order to estimate those parameters whose values are not directly implied by the constraints, we must have that the number of *df* available is at least $p + c$. The methods reviewed below deal with 1 or 2 observations on individuals from 1 or 2 populations.

3.3.1. One observation per individual, one population ($R = 1$, $S = 1$). Here subjects from a single source are classified by a single observation as positive or negative for an attribute. Typical frequency data may be summarised simply by the total numbers of positives (n_1) and negatives (n_0). There are 3 parameters—the prevalence θ of the attribute, and the false-positive and false-negative rates (α and β). This type of data would arise if a diagnostic test is used to detect sub-clinical disease, for instance abnormally high intra-ocular pressure as an indication of glaucoma [27]. Regarding the total number of individuals observed (N) as fixed, the number of "positives" n_1 determines the number of "negatives" n_0, and vice versa, because $n_1 + n_0 = N$; thus only 1 *df* is available, and so two constraints must be applied if parameter estimation is intended.

A common option is to impose the two constraints by regarding the sensitivity $1 - \beta$ and specificity $1 - \alpha$ as known, and then to estimate the prevalence θ. This would be appropriate if one were evaluating the prevalence in various populations, using a well-established screening test with known error probabilities. Rogan and Gladen [28] give algebraic expressions for $\hat{\theta}$ when α and β are given: $\hat{\theta}$ can occasionally be negative. They also demonstrate that the alternative of simply using the propor-

tion of positive observations (n_1/N) yields a very biased estimate of θ; specificity errors are usually a more important source of bias in the prevalence estimate than are errors of sensitivity [3].

3.3.2. Two observations per individual, one population ($R = 2, S = 1$). This type of data arises frequently when two diagnostic tests are being compared, or the agreement of two subjective raters is being assessed. For example, mammography and physical examination may be used as screens for early breast cancer [29]; the stress ECG and arteriography are both tests for coronary artery disease; two psychiatrists may differ in their diagnostic categorisation of a series of patients [30]; and ultrasound and venography are both tests for venous occlusion [31].

Typical data may be displayed as in Table 1, except that now both observations 1 and 2 are regarded as subject to error. Regarding N as fixed, there are 3 *df* for parameter estimation. There are five parameters—the test error rates α_1, α_2, β_1 and β_2, and the prevalence θ. Therefore at least two constraints are required. A wide variety of alternative constraints have been suggested, as outlined below.

When the data arise from two different methods of observation (as in the comparison of diagnostic tests), one possibility is to regard the sensitivity and specificity of one (say method 2) as known; this is appropriate if a new method is to be validated against an established criterion with known measurement properties. Assuming α_2 and β_2 known imposes the necessary two constraints, and several authors have all given identical formulae for $\hat{\alpha}_1$, $\hat{\beta}_1$ and $\hat{\theta}$ in this situation [32–34]. Standard errors of these are also available [32, 33] and predictive values may be estimated [34].

A special case of this approach is when α_2 and β_2 are assumed to be zero, i.e. that observation 2 is error-free. This is an expedient assumption when a third assessment of status is not possible, and when observation 2 is felt to be "definitive". For instance, arteriography is regarded as a "gold standard" diagnosis for coronary artery disease, against which the less invasive stress ECG method may be compared [35]. Other examples of this kind include: stool guaiac tests for colon cancer vs the barium enema as a "gold standard" [23]; thermography results vs tissue biopsy in the detection of minimal breast cancer [36]; and examination by a school nurse vs a specialist for hearing loss [37]. One danger here is that if test 2 is assumed to be error free, but is actually subject to error, the error rates for test 1 will be overestimated.

Some other approaches have been proposed for this type of data. Firstly, both tests being compared may be pathognomonic, so that they are regarded as having perfect specificity ($\alpha = 0$). Staquet *et al.* [34] give formulae for $\hat{\beta}_1$, $\hat{\beta}_2$, $\hat{\theta}$, and predictive values for this case. Secondly, it is possible to obtain approximate estimates of α_1 and α_2, without assuming anything about β_1 or β_2, if θ is known to be low or assumed so; correspondingly, populations with high prevalence yield approximate estimates of β_1 and β_2, without knowledge of α_1 and α_2 [33]. Some parameters may be estimated in various two-stage sampling designs, where the results of one observation are known before the second observation is made, on a sample basis [34]; again, however, the error rates of one test must be assumed known in order to estimate the other parameters. Finally, Chinn and Burney [38] have proposed another alternative constraint structure for this problem; they assume that $\alpha_1 = \alpha_2 = \beta_1 = \beta_2$, i.e. the probability of correct classification is constant. The estimated parameters are the prevalence and common sensitivity (or specificity), based on the 2 *df* remaining after application of 3 parameter constraints and loss of 1 *df* because of symmetry in the table when the error rates are the same for both observers.

A related problem arises from disease screening data. The screening method may misclassify persons, either by falsely labelling normals as (false) positives, or by missing true cases of disease (false negatives). Goldberg and Wittes [29] have described the use of capture–recapture methodology to estimate the number of true positives (the preclinical cases of disease) and negatives which exist in the screened sample, when two alternative means of diagnosis are used simultaneously, for instance in the detection of preclinical breast cancer using mammography and physical examination. In general, there are four error parameters, and one prevalence, giving 5 parameters in total. Goldberg and Wittes assume the false positive rates to be zero for both screening modalities: this then renders the remaining parameters estimable from the 3 *df* in the 2×2 table of screening results.

Related methods have been suggested for situations where the two observations are on different characteristics of individuals in the

same population. Barron [30] investigates the effects of misclassification on the odds ratio relating two clinical attributes (e.g. two different diseases). Individuals are classified on each attribute into the usual fourfold table with 3 *df*. There are now 7 parameters: 3 independent cell probabilities in the table, and α_1, α_2, β_1 and β_2 as before. Barron shows that if the error probabilities are taken as independent and known (4 constraints), a corrected odds ratio estimate may be derived from the empirical data.

Finally Murphy [40] considers a fourfold table generated by classifying individuals cross-sectionally according to the presence of a disease and/or risk attribute. The available 3 *df* are sufficient to estimate the sensitivity/specificity of the attribute interpreted as a marker for the disease, and the disease prevalence. Murphy also derives a simple relationship between PV+, PV− and the relative risk of disease for individuals attributed positive compared to negative. Bennett [31, 41] gives a test of the hypothesis that two (or more) diagnostic tests have equal predictive values, but does not discuss estimation.

Yanagawa and Gladen [42] discuss estimation when a single diagnostic test is applied to the same individuals at two or more points in time. As before, the "test positive" rate is a very biased estimate of the prevalence at each time point, and one that is more affected by specificity errors than sensitivity. If there are two times, 5 parameters are involved (the prevalence at the first time point, the incidence and remission rates between the two times, and the test sensitivity and specificity), but only 3 *df* are available. Two alternative pairs of constraints are possible: (i) the sensitivity and specificity are known, or (ii) the remission rate is zero and the specificity is 1. An example of the latter is provided, using onchoceriasis data. A similar approach is given by Chinn and Burney [38]; after assuming that sensitivity and specificity are equal for both observations, (3 constraints), estimates can be obtained for the prevalence at each of the two time points, and the common probability of correct classification.

Yanagawa and Gladen add that if three time points are used, there are 7 *df* still with 5 parameters; complete estimation is then possible [42, 43]. More general models have been developed to estimate the error rates of screening tests administered on several occasions [44, 45]; these models also allow one to estimate the disease incidence rate and the duration of its preclinical interval.

3.3.3. One observation per individual, two populations ($R = 1$, $S = 2$). An example of this common design is the case–control study, where cases and controls are classified on exposure to an antecedent risk factor. Misclassification may distort the odds ratio relating exposure to disease status. If the misclassification rates are the same for cases and controls, the odds ratio will be biased towards the null value; with different misclassification rates for the cases and controls, bias in either direction can occur [46].

Assuming fixed total numbers of subjects in each of the two groups, a binary classification gives 2 *df* in total. When sensitivity and specificity are assumed constant across populations, the four parameters are the common false positive rate α, the common false negative rate β, and the population-specific prevalences θ_1, and θ_2. So again two constraints are required for estimation purposes.

As for the problem with $R = 2$, $S = 1$, a common solution is to assume α and β known: Copeland *et al.* [47] then describe how to estimate the "true" numbers of positives and negatives in each group, and hence how to recalculate the odds ratio or relative risk, "corrected" for the misclassification effect. Greenland and Kleinbaum [48] describe a similar approach; they point out that *a priori* estimates of the misclassification rates are also subject to error, so that the corrected or "error-free" tables may still indicate incorrect levels of association. Greenland [49] has proposed an equivalent method for matched pair data without replication; this requires prior estimates of the error rates, and the solution to 4 simultaneous equations. Finally, rather than assuming any particular values for the error rates, Blettner and Wahrendorf [50] consider the possible ranges for the probability of correct classification in case–control studies, given the empirical misclassified data. Equal reliability is assumed for the cases and controls, leading to a range of possible values for the relative risk.

3.3.4. Two observations per individual, two populations ($R = 2$, $S = 2$). In this problem, we have two 2×2 data tables cross-classifying the two observations, one from each of the two populations. The two groups might be from different geographic areas, or be sub-groups (e.g. sex, race) of the same population. As an example, the Mantoux and tine tests (the two

observations) for tuberculosis were both administered to individuals in two populations [15].

There are now four misclassification probabilities as before (α_1, α_2, β_1 and β_2), and a prevalence in each population (θ_1 and θ_2), making 6 parameters in total. There are also 6 *df* (3 *df* from each population), and so no constraints are required. As before, we have assumed the sensitivity and specificity for each observer to be constant across populations. Closed form ML estimates of the 6 parameters and their variance–covariance matrix may be obtained [15].

A further example of this type of data occurs in case–control studies, if there are two assessments of exposure in each group. Marshall and Graham [51] have proposed that only individuals with concordant assessments be used to estimate the exposure–disease odds ratio. However this method gives a biased estimate, in contrast to the ML method which is asymptotically unbiased and efficient [52]. In their discussion of latent class analysis, Kaldor and Clayton [21, 53] give an example of data where replicate measurements are available for some or all of the cases and/or controls. They demonstrate that obtaining replicate measurements on even a modest proportion of subjects leads to substantially improved estimation of case–control odds ratios.

4. OTHER DESIGNS

4.1. Irregular observational designs

All of the above methods have supposed that all of the observers categorise all study subjects, but there are several practical situations where each observer classifies only a subset of individuals, leading to a less regular design. The usual effect of departing from the regular layout is to reduce the available *df*, thereby imposing further limitations on the number of estimable parameters.

A common example of an irregular design is from sequential observations. Here the early observations determine whether or not an individual will go on to be observed later in the sequence. For instance, a sequence of diagnostic tests may be available, typically having increasing accuracy but at increasing cost or risk to the patient; only individuals with early abnormal results progress to the later stages in the sequence. An example is the sequence of tests recommended for spina bifida: two tests for elevated alpha-fetoprotein (AFP), ultrasound, amniocentesis, and amniography [54]. Another example of this kind is when children are given an initial multiple puncture test (e.g. tine) for TB which, only if negative, is followed by a Mantoux test [55].

Sequential designs which introduce a new method of observation at each step are generally overparameterised. At each stage, 2 new parameters (the sensitivity and specificity of the new observation) are introduced, but only 1 *df* is added. Only if a sufficient number of repeat uses are made of the *same* test can all the relevant parameters be estimated; under an assumption that the error probabilities remain constant on repeated uses, the accumulation of an extra single *df* at each stage will eventually produce a total *df* which exceeds the number of parameters. For example, suppose the same test is applied repeatedly only to those positive in the previous step. Three parameters are involved: α, β and the population prevalence θ. Each step provides 1 *df*, and so therefore 3 steps are required as a minimum for parameter estimation.

On the other hand, if the assumption of constant error probabilities is not valid, the situation is then similar to using a different test method at each stage, and again there is a deficit of *df*. Thus although sequential strategies may often be desirable for routine clinical practice, they are inadequate in general for the initial assessment of performance.

An irregular design was used by Rudd *et al.* [56] to compare several tests for TB sensitivity. All study subjects received the Mantoux test, but were randomised to receive either the tine test *or* the imotest in addition. The data thus consist of two 2×2 tables, each having 3 *df*, giving 6 *df* in total. Subjects whose initial Mantoux was negative, and whose other test was positive were retested by Mantoux; the retests provide an additional 2 *df*, one from each of the tine and imotest groups. The data in total then have enough *df* (8) to estimate all the parameters (3α's, 3β's and 2θ's).

A final example of an irregular design is that proposed by Green [57] to estimate the relative risk of disease in exposed versus unexposed individuals. Disease classification is made on the basis of a fallible test T, for a sample of individuals from each of the two exposure groups; this yields 2 *df*. In addition, it is supposed that a correct classification is made only

for individuals in the unexposed group who have a positive result on T; this yields a further 1 df (giving 3 df in total) and an estimate of PV+ for the unexposed. There are 4 parameters; the α and β for T, and the true disease incidence rates for the exposed and unexposed. Green imposes a constraint by supposing the disease incidence in unexposed individuals to be small, and then shows that a relative risk estimate, adjusted for misclassification, may be obtained as a function of the crude relative risk and PV+ for the unexposed. Green gives examples of this technique in associating coronary atherosclerosis with smoking and serum cholesterol.

Green's method was extended by Begg [58], who shows how an unbiased estimate of the odds ratio may be obtained by restricting the analysis to those individuals with positive results on T, and for whom an error-free classification is made; this is done without information on sensitivity, specificity, or the incidence among the unexposed. He also notes that the relative risk can be estimated by assuming the test specificity to be 1. Finally, if the sensitivity and specificity are assumed known, unbiased estimates of the disease incidence among exposed and unexposed persons may be derived. It has also been shown that an unbiased relative risk estimate can be obtained if PV− is 1 [59]. The method used by Green and Begg is similar to previous work [60, 61] using a two stage sampling procedure with an error free classification for a random subsample.

4.2. Response variable with more than two categories

Although it is always possible to reduce data into a binary form (e.g. normal/abnormal), there is often a more detailed classification available, usually as a multilevel discrete variable. A multilevel response may be more meaningful substantively, but it will necessarily involve additional statistical parameters, and hence a more elaborate design if they are all to be estimated. As before, an option is to assume some of the parameters to be known.

An example of a data set with four response categories is discussed by Spiegelhalter and Stovin [62] where up to three biopsies had been taken from a series of cardiac transplant patients. Each biopsy was categorised by a single pathologist on a four point scale indicating their assessment of the likelihood that organ rejection had taken place. It was required to estimate the probability that rejection had taken place for a patient with a given set of biopsy results.

Assuming the various biopsy classifications to have equal reliability, there are $4 \times 3 = 12$ independent misclassification probabilities, and 3 independent prevalence parameters. Ignoring the order of observations, there are 20 possible combinations of results from a set of 3 biopsies (4 possibilities where all three results are the same, 12 where exactly two different results occur among the three, and 4 where 3 different results are given), and 10 possible combinations in patients who had only 2 biopsies. In the data, 15 and 8 combinations respectively were actually observed. After grouping several sets of small frequency cells and allowing for 2 constraints implied by the sample sizes, 19 df were available for estimation. Arguing that it would be impossible to observe a biopsy state which is worse than the true rejection state of the patient, Speigelhalter and Stovin assumed that 6 of the false positive rates were zero. This left 6 false negative rates and 3 "prevalences" (describing the true distribution across the 4-point scale), giving 9 parameters to be estimated in total. A goodness of fit test was then possible on the remaining 10 df.

Another example is given by Dawid and Skene [18], where 5 anaesthetists rated patients' suitability for surgery on a 4 point scale; also, one of the 5 anaesthetists made 3 independent ratings of each patient. In this problem there are 3 "prevalence" parameters and 60 misclassification rates (12 for each rater). A very large number of df is available, actually 20×4^4 (ignoring the order of the 3 independent replicates by one rater), so that the data will be sparsely distributed across all the possible combinations of ratings. This implies that the estimated parameters will likely be very unstable, except in very large samples. Dawid and Skene examined stability by selective removal of observers and/or patients from the data. They also remark that the usual large sample properties of maximum likelihood estimates are unlikely to hold good.

In general, if R observers all rate the same individuals on a K point scale, there are $K^R - 1$ df available. There are $K - 1$ "prevalence" parameters, and $K(K-1)$ misclassification parameters for each observer, making $(K-1)(RK+1)$ in total. The following table shows the number of parameters and available df for low R and $K = 3$ and 4:

		Number of observers (R)		
		1	2	3
$K = 3$	Number of parameters:	8	14	20
	Number of *df*:	2	8	26
$K = 4$	Number of parameters:	15	27	39
	Number of *df*:	3	15	63

This shows that, as for binary data, 3 observers is the minimum number required for full parameter estimability. In fact this is true for any K. For large K, the number of parameters is approximately K^2R, and the *df* K^R, so R must exceed 2 for full estimability. Note that both the number of parameters and the *df* increase rapidly with K, implying that a larger sample is needed for stable estimation when a multi-point scale is used. Because of this, it may be preferable to adopt a continuous measurement approach, rather than increasing K unduly. One set of constraints which has been suggested [38] is that the probability of correct classification is constant for all persons, and that each of the $K - 1$ possible misclassifications is equally likely. Although these are strong assumptions to make, they do have the effect of eliminating all but two parameters, which have closed form estimators.

5. DEPARTURES FROM ASSUMPTIONS

Most of the statistical methods described above involve some simplifying assumptions. Simplification is, in some ways, a virtue, because a simple model may be understood more easily. More complex models may be a more accurate representation of the real world situation, but they are correspondingly more difficult to evaluate, often because there is insufficient data to test all the component parts of the model. Some of the literature dealing with the major assumptions is described below.

5.1. Correlation of errors

All of the methods described above assume implicitly or explicitly that the errors of classification are independent between observers, conditional on the true state of the individual. This is a convenient assumption statistically, but one which in practice may be dubious. For instance, there may be extreme subgroups of patients whose disease status is relatively easy to diagnose, and for whom misclassification is unlikely in comparison to other patients with "borderline" disease. This is especially true if the true underlying disease state is actually continuous rather than discrete.

The likelihood method might be generalised to incorporate correlated errors, but this will be at the expense of introducing more parameters. Error correlations, if present, are most likely to be positive; for example, clinicians with similar training are likely to misclassify the same patients in similar ways. The assumption of independent errors is anti-conservative if there are in fact positively correlated errors, because there will be an empirically higher level of agreement among observers than would be expected with independence; the misclassification probabilities will then be underestimated. An observer identified as having larger misclassification rates may erroneously be considered less skillful, when this result has actually arisen because of correlated errors between the other observers. An alternative method which uses the relative accuracy of observers has been suggested as a solution to this problem [63].

One can sometimes test the assumption of independent errors by using a goodness of fit test on the closeness of the observed data to their expected frequencies based on the independence assumption. This test is feasible if there are excess *df* remaining after the parameters of the independence model have been estimated. For instance, such a test may be carried out when there are 4 or more observers in a balanced design in one population [12]. As mentioned earlier, the test may have low power. Also a significant lack of fit need not be due to an inter-test dependence; other departures from the model (e.g. increasing sensitivity of the observers over time) might also lead to a lack of fit.

Rindskopf *et al.* [64] suggest the goodness of fit test as a way of validating the model. They give an example with 4 diagnostic tests for myocardial infarction, and argue that a satisfactory fit of the likelihood model to the data supports a binary representation of the disease. If a poor fit occurs, Rindskopf suggests dividing the data into homogeneous subgroups so that the within-group error correlation might be reduced. This approach is obviously limited by the reduced power of the subgroup tests of goodness of fit, because of smaller sample sizes in each.

Very little analytic work has been done on the effect of error correlations on parameter estimates. Thibodeau [65] has developed bounds for the sensitivity and specificity of a fallible diagnostic test in comparison to a reference test, when the errors are correlated. The bounds are

determined by the magnitude of the inter-test correlation, and on constraints on cells in the fourfold table of data.

Vacek [66] examined the effect of error correlations for data with $R = 2$, $S = 2$; specifically he examined the robustness of the Hui–Walter ML estimates (which assume independent errors) for this situation, under various assumptions about the true error structure. Error covariances were introduced into a modified likelihood. Positive error covariances generally lead to underestimation of the misclassification rates if the unmodified likelihood is used, but the bias in the prevalence estimate can be in either direction. Interestingly, the true prevalence has no effect on the biases in the misclassification estimates, and the prevalence in one population has no effect on the bias of the other prevalence estimate.

5.2. Assumption of constant sensitivity/specificity

A second assumption of many of the methods discussed here is that the sensitivity/specificity values for a given method of observation remain constant over various population subgroups, and in particular do not vary with changes in exposure or disease prevalence. In many situations this assumption is reasonable, but in others the assumption is probably made only for mathematical convenience.

In one of the few analyses to address this problem, Goddard [67] allowed the sensitivity of a test for schistosomiasis to depend on prevalence. He felt that higher disease prevalence would correspond to a higher intensity of infection, hence decreasing the chance of a false negative result. (This again represents an issue which arises when disease is measured as a dichotomized state, when the underlying disease is a continuous spectrum.) A negative exponential relationship was assumed between the false negative rate and the disease prevalence. As before, the assumption in question can be relaxed by suitable generalisation of the likelihood, incorporating additional parameters in the process. In this case, the number of new parameters was minimised by assuming the exponential sensitivity function to apply.

6. MEASUREMENT ERROR IN CONTINUOUS DATA

This paper has been restricted to a review of misclassification in discrete data, although analogous work has been done on random measurement error for continuous data. The emphasis on discrete aspects of misclassification was deliberate because, as Howe [68] notes, continuous risk variables are often discretised anyway in the analysis of medical data sets; this allows the calculation of indices such as relative risk for categories or risk, without the necessity of assumptions concerning the shape of dose response relationships.

The reader interested in the effects of errors of measurement in continuous variables in this context is directed to the work of Howe [68], Kupper [1], Walker [5], and others [69–72]. These authors conclude that, as for discrete data, random measurement error leads to attenuation of estimates of effect, and that a substantial increase in sample size may be needed to maintain power.

Random measurement error in a confounder may seriously bias measures of effect [1]. Methods exist for obtaining unbiased estimates of effect if data are available on the random measurement error variance [70, 71].

7. DISCUSSION

We have seen how the various designs for investigating the reliability of clinical data may be characterised by the number of misclassification and prevalence parameters, and by the number of statistical *df* available for their estimation. Subsets of parameters can be estimated in designs which have too many parameters to be estimated simultaneously but only by imposing constraints. If possible, it is desirable to use additional observers, or independent replicates of the same observers. Three observers is the minimum for which all parameters can be estimated; with more than 3 observers the number of *df* exceeds the number of parameters, allowing a goodness-of-fit test of errors. Using a response scale with more than 2 points increases the *df* faster than the number of parameters, but a larger sample will be needed to permit stable estimation of all the parameters. Irregular designs, where not all observers classify all sample subjects, may limit the number of parameters which are estimable without constraints.

Acknowledgements—This research was supported in part by a Canadian National Health Research and Development Program through a National Health Scientist Award. The authors would like to acknowledge the assistance of the X-ray readers mentioned in Section 3.2, namely Dr G. K. Sluis-Cremer, Dr R. Glyn Thomas, and Professor A. Solomon. Dr D. L. Sackett, Dr C. H. Goldsmith and Dr N. J. Birkett of McMaster University provided some helpful comments on an early draft of the paper.

REFERENCES

1. Kupper LL. Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies **Am J Epidemiol** 1984; 120: 643–648.
2. Fung KY, Howe GR. Methodological issues in case–control studies III; the effect of joint misclassification of risk factors and confounding factors upon estimation and power. **Int J Epidemiol** 1984; 13: 366–370.
3. Quade D, Lachenbruch PA, Whaley FS *et al.* Effects of misclassifications of statistical inferences in epidemiology. **Am J Epidemiol** 1980; 111(5): 503–515.
4. Gregorio DI, Marshall JR and Zielezny M. Fluctuations in odds ratios due to variance differences in case–control studies. **Am J Epidemiol** 1985; 121(5): 767–774.
5. Walker AM. Misclassified confounders. **Am J Epidemiol** 1985; 122: 921.
6. Sackett DL, Haynes B, Tugwell P. In: **Clinical Epidemiology.** Boston: Little, Brown; 1985.
7. Sheps SB, Schechter MT. The assessment of diagnostic tests. **JAMA** 1984; 252(17): 2418–2422.
8. Landis JR, Koch GG. The measure of agreement for categorical data. **Biometrics** 1977; 33: 159–174.
9. Landis JR, Koch GG. An application of hierarchial kappa-statistics in the assessment of majority agreement among multiple observers. **Biometrics** 1977; 33: 363–374.
10. Cohen J. A coefficient of agreement for nominal scales. **Educ Psychol Meas** 1960; 20: 37–46.
11. Kraemer HC. Ramification of a population model for kappa as a coefficient of reliability. **Psychometric** 1979; 44: 461–472.
12. Walter SD. Measuring the reliability of clinical data: the case for using three observers. **Rev Epidemiol sante publique** 1984; 32: 206–211.
13. Thompson WD. Design issues in the assessment and control of misclassification errors. Paper given at the **1982 Meeting of the Society Epidemiology Research** Cincinnati, Ohio.
14. Armitage P, Blendis LM, Smyllie HC. The measurement of observer disagreement in the recording of signs. **J R Stat Soc A** 1966; 129: 98–109.
15. Hui SL, Walter SD. Estimating the error rates of diagnostic tests. **Biometrics** 1980; 36: 167–171.
16. Feinstein AR. Clinical Biostatistics XXXI. On the specificity, sensitivity and discrimination of diagnostic tests. **Clin Pharmacol Ther** 1975; 17: 104–116.
17. Galen RS, Gambino SR. **Beyond Normality: the Predictive Value and Efficiency of Medical Diagnoses.** Wiley: New York; 1975.
18. Dawid AP, Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. **Appl Stat** 1979; 28: 20–28.
19. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. **JRSS B** 1977; 39: 1–38.
20. White AA, Landis JR. A general categorical data methodology for evaluating medical diagnostic tests. **Commun Stat** 1982; 11: 567–605.
21. Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. **Stat Med** 1985; 4: 327–335.
22. Simpson PR, Chamberlain J, Gravelle HSE. Choice of screening tests. **J Epidemiol Commun Hlth** 1978; 32: 166–170.
23. Neuhauser D, Lewicki AM. What do we gain from the sixth stool guaiac? **N Engl J Med** 1975; 293: 226–228.
24. Gutjahr P, Jung H. Detecting tuberculin sensitivity. **Lancet** 1982; 768.
25. International Labour Office. **ILO U/C International Classification of Radiographs of Pneumoconioses 1971.** Geneva, Switzerland; 1972.
26. Irwig LM, duToit RSJ, Sluis-Cremer GK *et al.* Risk of asbestosis in crocidolite and mosite mines in South Africa. **Annals NY Acad Sci** 1979; 330: 35–52.
27. Thorner RM, Remein QR. **Principles of Screening for Disease.** Washington, D.C.: Government Printing Office; 1961. P.H. Monogr. No. 67, p. 24.
28. Rogan WJ, Gladen B. Estimating prevalence from the results of a screening test. **Am J Epidemiol** 1978; 107: 71–76.
29. Goldberg JD, Wittes JT. The estimation of false negatives in medical screening. **Biometrics** 1978; 34: 77–86.
30. Fleiss JL. **Statistical Methods for Rates and Proportions.** New York: Wiley, 1981.
31. Bennett BM. On tests for equality of predictive values for *t* diagnostic procedures. **Stat Med** 1985; 4: 535–540.
32. Gart JJ, Buck AA. Comparison of a screening test and a reference test in epidemiologic studies. II: A probabilistic model for the comparison of diagnostic tests. **Am J Epidemiol** 1966; 83(3): 593–602.
33. Greenberg RA, Jekel JF. Some problems in the determination of false negative rates of tuberculin tests. **Am Rev Resp Dis** 1969; 100: 645.
34. Staquet M, Rozencweig M, Lee YJ *et al.* Methodology for the assessment of new dichotomous diagnostic tests. **J Chron Dis** 1981; 34: 599–610.
35. Weiner DA, Ryan TJ, McCabe CH *et al.* Exercise stress testing: correlations among history of angina, ST-segment response and prevalence of coronary artery disease in the coronary artery surgery study (CASS). **N Engl J Med** 1979; 301: 230–235.
36. Moskowitz M, Milbrath J, Gartside P *et al.* Lack of efficacy of thermography as a screening tool for minimal and stage I breast cancer. **N Engl J Med** 1974; 295: 249–252.
37. Bay KS, Flathman D, Nestman L. The worth of a screening program: an application of a statistical decision model for the benefit evaluation of screening projects. **Am J Public Hlth** 1976; 66: 145–150.
38. Chinn S, Burney PGJ. On measuring repeatability of data from self-administered questionnaires. **Int J Epidemiol** 1987; 16(1): 121–127.
39. Barron BA. The effects of misclassification on the estimation of relative risk. **Biometrics** 1977; 33: 414–418.
40. Murphy JR. The relationship of relative risk and positive predictive value in 2×2 tables. **Am J Epidemiol** 1983; 117(1): 86–89.
41. Bennett BM. On comparisons of sensitivity, specificity and predictive value of a number of diagnostic procedures. **Biometrics** 1972; 28: 703–800.
42. Yanagawa T, Gladen BC. Estimating disease rates from a diagnostic test. **Am J Epidemiol** 1984; 119: 1015–1023.
43. Yanagawa T, Kasagi F. Estimating prevalence and incidence of disease from a diagnostic test. In: Matusita, K. Ed. **Stat. Theory and Data Analysis** Amsterdam: Elsevier; 1985.
44. Day NE, Walter SD. Simplified models of screening for chronic disease: Estimation procedures for mass screening programs. **Biometrics** 1984; 40: 1–14.
45. Walter SD, Day NE. Estimation of the duration of a preclinical disease state using screening data. **Am J Epidemiol** 1983; 118(6): 865–886.
46. Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. **J Am Stat Assoc** 1975; 70: 561–567.
47. Copeland KT, Checkoway H, McMichael AJ *et al.*

Bias due to misclassification in the estimation of relative risk. **Am J Epidemiol** 1977; 105: 488–495.
48. Greenland S, Kleinbaum DG. Correcting for misclassification in two-way tables and matched-pair studies. **Int J Epidemiol** 1983 12(1): 93–97.
49. Greenland S. The effect of misclassification in matched-pair case–control studies. **Am J Epidemiol** 1982; 116: 402–406.
50. Blettner M, Wahrendorf J. What does an observed relative risk convey about possible misclassification? **Meth Inform Med** 1984; 23: 37–40.
51. Marshall JR, Graham S. Use of dual responses to increase validity of case–control studies. **J Chron Dis** 1984; 36: 125–136.
52. Walter SD. Use of dual responses to increase validity of case–control studies: A commentary. **J Chron Dis** 1984; 37(2): 137–139.
53. Clayton D. Using test–retest reliability data to improve estimates of relative risk; an application of latent class analysis. **Stat Med** 1985; 4: 445–446.
54. Chinchilli VM. Estimates of sensitivity and specificity in a multistage screen for medical diagnosis. **Biometrics** 1983; 39: 333–340.
55. Ackerman-Liebrich U. Tuberculin sensitivity testing (letter). **Lancet** 1982; Oct. 23: 934.
56. Rudd RM, Gellert AR, Venning M. Comparison of mantoux, tine, and 'imotest' tuberculin tests. **Lancet** 1982; Sept. 4: 515–518.
57. Green MS. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. **Am J Epidemiol** 1983; 117(1): 98–105.
58. Begg CB. Estimation of risks when verification of disease status is obtained in a selected group of subjects. **Am J Epidemiol** 1984; 120: 328–329.
59. Lawrence C, Greenwald P. Epidemiologic screening: a method to add efficiency to epidemiologic research. **Am J Epidemiol** 1977; 105: 575–581.
60. Tenenbein A. A double sampling scheme for estimating from misclassified binomial data. **J Am Stat Assoc** 1970; 65: 1350–1361.
61. Hochberg Y. On the use of double sampling schemes in analysing categorical data with misclassification errors. **J Am Stat Assoc** 1977; 72: 914–921.
62. Speigelhalter DJ, Stovin PGI. An analysis of repeated biopsies following cardiac transplantation. **Stat Med** 1983; 2: 33–40.
63. Irwig LM, Groeneveld HT, Pretorius JPG *et al.* Relative observer accuracy for dichotomized variables. **J Chron Dis** 1987; 38: 899–906.
64. Rindskopf D, Rindskopf W. The value of latent class analysis in medical diagnosis. **Stat Med** 1986; 5: 21–28.
65. Thibodeau LA. Evaluating diagnostic tests (Abstract). **Biometrics** 1981; 37(1): 192.
66. Vacek PM. The effect of conditional dependence on the evaluation of diagnostic tests. **Biometrics** 1985; 41: 959–968.
67. Goddard MJ. On allowing for diagnostic imperfections in assessing effectiveness of treatment for schistosomiasis. **Int J Epidemiol** 1977; 6(4): 381–389.
68. Howe HR. The use of polychotomous dual response data to increase power in case–control studies: an application to the association between dietary fat and breast cancer. **J Chron Dis** 1985; 38: 663–670.
69. Gardner MJ, Heady JA. Some effects of within-person variability in epidemiological studies. **J Chron Dis** 1973; 26: 781–795.
70. Shepard DS. Reliability of blood pressure measurements: implications for designing and evaluating programs to control hypertension. **J Chron Dis** 1981; 34: 191–209.
71. Richardson DH, Wu D. Least squares and grouping method estimators in the errors in variables model. **J Am Stat Assoc** 1970; 65: 724–748.
72. Fuller WA, Hidiroglou MA. Regression estimation after correcting for attenuation. **J Am Stat Assoc** 1978; 73: 99–104.