THEORY AND METHODS

A simple model for potential use with a misclassified binary outcome in epidemiology

S W Duffy, J Warwick, A R W Williams, H Keshavarz, F Kaffashian, T E Rohan, F Nili, A Sadeghi-Hassanabadi

J Epidemiol Community Health 2004;**58**:712–717. doi: 10.1136/jech.2003.010546

See end of article for authors' affiliations

Correspondence to: Professor S W Duffy, Cancer Research UK Department of Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, Queen Mary University of London, Charterhouse Square, London EC1M 6BQ, UK; stephen.duffy@cancer.org.uk

Accepted for publication 23 October 2003

Study objective: Error in determination of disease outcome occurs in epidemiology, but such error is not usually corrected for in statistical analysis. A method of correction of risk estimates for misclassification of a binary disease outcome is developed here.

Methods: The method is a simple, closed form correction to the logistic regression estimate. A closed form variance estimate is also developed.

Setting: The method is illustrated in two studies, a cross sectional survey of cervicitis in Iran in 1996–97, as determined by inflammation on cervical smear specimens, and a case-cohort study of benign proliferative epithelial disease of the breast, in Canada 1980–88.

Main results: The method provides corrected odds ratio estimates and corrects the spurious precision conferred by misclassification.

Conclusions: The method is easy to apply and potentially useful, although potential failures of the assumptions involved should be borne in mind. It is necessary to give careful consideration to the plausibility or otherwise of the assumptions in the context of the individual study. Correction for misclassification of disease outcome may become more common with the development of readily applicable methods.

Misclassification of risk factors in epidemiology has received considerable attention in the biostatistical literature over the past 30 years.^{1–4} Misclassification of clinical outcome, for example disease status, has received less attention, although examples of such misclassification do occur.^{5–7} Some work has been done on its effects in the case of the two by two contingency table, but these are not specifically aimed at correction of risk estimates for the biasing effects.^{8–10} A reasonable formal approach to such problems, particularly when repeated determinations of disease status are performed, is the use of over-dispersion or latent variable models.^{11–12} A powerful and practical tool in taking into account additional error such as in determination of disease status is Monte Carlo simulation.^{13–14}

It would be of some use in epidemiology to derive a method that directly relates to measures of relative rate of disease, such as odds ratios. In this paper, therefore, we develop a simple correction to the odds ratio to take account of misdetermination of disease status. The correction relates to a simple logistic regression model and has a closed form estimate. The correction is illustrated by two examples. The first considers risk factors for cervicitis, as determined by cytological observation of inflammation on cervical smears. The determination of inflammation is subject to misclassification. The second considers the association between dietary fat intake and risk of benign proliferative epithelial disorders of the breast. These disorders are identified histologically and are also subject to misclassification.

Our method is not a definitive solution to the problem of misclassification of disease state. It shows the effect of such misclassification on the odds ratio under certain assumptions, with an implied correction to the odds ratio, also dependent on those assumptions. Whether the assumptions, and therefore the correction, are reasonable will vary depending on the disease outcome and the method of its determination.

METHODS

We develop the algebra for both symmetric misclassification and the more realistic situation of asymmetric misclassification. For simplicity of layout, we take the case of symmetric misclassification first.

Suppose we have a disease state ($0 =$ no disease, $1 =$ disease) that is subject to random misclassification, with a probability of correct classification of α ($0 \leq \alpha \leq 1$) and a probability of misclassification of $1 - \alpha$.

We have data on disease state and on a binary risk factor in N individuals. We assume for simplicity that the risk factor is measured without error, although this is not absolutely necessary. Under certain assumptions, misclassification of the risk factor can be corrected for after correction for the misclassification of the disease outcome (see Discussion). Denote the risk factor by x ($x = 0$ for exposure negative, $x = 1$ for exposure positive). Let the overall true probability of disease be p . Under the logistic model

$$E(y_1) = \beta_0 + \beta_1$$

where y_1 is the log odds of case status with exposure positive and β is the log odds ratio. We also have

$$E(y_0) = \beta_0$$

where y_0 is the log odds of case status with exposure negative.

Now note that the odds ratio is invariant to the direction of dependence in the logistic regression model. Thus if z_1 is the log odds of exposure positive status given disease present and z_0 the log odds given disease absent, we have

Abbreviations: BV, bacterial vaginosis; BPBD, benign proliferative breast disease

$$E(z_1) = \beta'_0 + \beta_1$$

and

$$E(z_0) = \beta'_0$$

That is, the intercept in the logistic regression equation differs, but the regression coefficient (the log odds ratio) is the same for dependence of disease state on risk factor as for risk factor on disease state. Then

$$\beta_1 = E(z_1) - E(z_0)$$

However, if the observed probability of positive disease is p^* , then

$$p^* = p\alpha + (1-p)(1-\alpha) \quad (1)$$

where p is the true probability of positive disease status. Thus of those with observed positive disease status, the proportion who truly are disease cases is $p\alpha/p^*$. Similarly among those observed to be disease negative, the proportion truly negative is $(1-p)\alpha/(1-p^*)$. Therefore, among those ostensibly positive, instead of observing z_1 we observe w_1 that has expectation

$$\frac{p\alpha E(z_1)}{p^*} + \frac{(1-p)(1-\alpha)E(z_0)}{p^*}$$

Among those ostensibly negative, instead of observing z_0 , we observe w_0 which has expectation

$$\frac{(1-p)\alpha E(z_0)}{1-p^*} + \frac{p(1-\alpha)E(z_1)}{1-p^*}$$

Thus, instead of observing $\hat{\beta}_1$, which has expectation $E(z_1) - E(z_0)$ we observe $\hat{\beta}_1^*$ that has expectation

$$E(z_1) \left\{ \frac{p\alpha}{p\alpha + (1-p)(1-\alpha)} - \frac{p(1-\alpha)}{p(1-\alpha) + \alpha(1-p)} \right\} - E(z_0) \left\{ \frac{(1-p)\alpha}{(1-p)\alpha + p(1-\alpha)} - \frac{(1-p)(1-\alpha)}{p\alpha + (1-\alpha)(1-p)} \right\} \quad (2)$$

In the case of asymmetric misclassification, with α_1 representing the correct classification probability for a true positive and α_2 that for a true negative, we would observe

$$E(z_1) \left\{ \frac{p\alpha_1}{p\alpha_1 + (1-p)(1-\alpha_2)} - \frac{p(1-\alpha_1)}{p(1-\alpha_1) + \alpha_2(1-p)} \right\} - E(z_0) \left\{ \frac{(1-p)\alpha_2}{(1-p)\alpha_2 + p(1-\alpha_1)} - \frac{(1-p)(1-\alpha_2)}{p\alpha_1 + (1-\alpha_2)(1-p)} \right\}$$

If PPV denotes positive predictive value of the observed outcome and NPV its negative predictive value, both formulas can be re-expressed as

$$(E(z_1) - E(z_0))(PPV + NPV - 1)$$

So if we estimate the naive odds ratio $\hat{\beta}_1^*$ (even by logistic regression of disease state on risk factor), we can estimate the corrected odds ratio $\hat{\beta}_1$ by

$$\hat{\beta}_1 = \hat{\beta}_1^* / (PPV + NPV - 1)$$

The correction factor can be expressed in terms of p^* and α . In the case of symmetric misclassification, from equation (1) we have

$$p = \frac{p^* - 1 + \alpha}{2\alpha - 1}$$

And so, after some routine algebra

$$\beta_1^* = (z_1 - z_0) \frac{(\alpha - p^*)(\alpha - 1 + p^*)}{(2\alpha - 1)p^*(1 - p^*)} \quad (3)$$

$$\hat{\beta}_1 = \hat{\beta}_1^* / \left\{ \frac{(\alpha - p^*)(\alpha - 1 + p^*)}{(2\alpha - 1)p^*(1 - p^*)} \right\} \quad (4)$$

where p^* is the observed disease prevalence in the study. The corresponding formula for asymmetric misclassification is

$$\hat{\beta}_1 = \hat{\beta}_1^* / \left\{ \frac{(\alpha_1 - p^*)(\alpha_2 - 1 + p^*)}{(\alpha_1 + \alpha_2 - 1)p^*(1 - p^*)} \right\}$$

If we have repeated data a, b, c, d on disease status as in table 1, we can estimate α as

$$\hat{a} = 0.5 + 0.5 \sqrt{\frac{N - 2(b + c)}{N}} \quad (5)$$

This is shown in the appendix, which also shows that the variance of α can be estimated as

Table 1 Notation for repeated data on disease outcome

Second determination	First determination		Total
	Disease absent	Disease present	
Disease absent	a	b	a+b
Disease present	c	d	c+d
Total	a+c	b+d	N

$$\hat{V}(\hat{\alpha}) = - \left(\frac{8(a+d)\hat{\alpha}(1-\hat{\alpha})}{(2\hat{\alpha}^2 - 2\hat{\alpha} + 1)^2} - \frac{(b+c)(2\hat{\alpha}^2 - 2\hat{\alpha} + 1)}{\hat{\alpha}^2(1-\hat{\alpha})^2} \right)^{-1} \quad (6)$$

The variance of our corrected estimate in the case of symmetric misclassification is therefore

$$V(\hat{\beta}_1) = V(\hat{\beta}_1^*) \left\{ \frac{(2\hat{\alpha} - 1)p^*(1-p^*)}{(\hat{\alpha} - p^*)(\hat{\alpha} - 1 + p^*)} \right\}^2 + \hat{\beta}_1^{*2} V \left\{ \frac{(2\hat{\alpha} - 1)p^*(1-p^*)}{(\hat{\alpha} - p^*)(\hat{\alpha} - 1 + p^*)} \right\}$$

Assuming that p^* is fixed by design or is from a study sufficiently large that it can be assumed to have negligible variance, we have

$$V(\hat{\beta}_1) = V(\hat{\beta}_1^*) \left\{ \frac{(2\hat{\alpha} - 1)p^*(1-p^*)}{(\hat{\alpha} - p^*)(\hat{\alpha} - 1 + p^*)} \right\}^2 + \frac{(\hat{\beta}_1^* p^*(1-p^*))^2}{(\hat{\alpha} - p^*)^4 (\hat{\alpha} - 1 + p^*)^4} \{ 2(\hat{\alpha} - p^*)(\hat{\alpha} - 1 + p^*) - (2\hat{\alpha} - 1)^2 \}^2 V(\hat{\alpha}) \quad (7)$$

For asymmetric misclassification, assuming that α_1 and α_2 are estimated independently, the variance is estimated as

$$V(\hat{\beta}_1) = V(\hat{\beta}_1^*) \left\{ \frac{(\hat{\alpha}_1 + \hat{\alpha}_2 - 1)p^*(1-p^*)}{(\hat{\alpha}_1 - p^*)(\hat{\alpha}_2 - 1 + p^*)} \right\}^2 + (\hat{\beta}_1^* p^*(1-p^*))^2 \left\{ \frac{V(\hat{\alpha}_1)}{(\hat{\alpha}_1 - p^*)^4} + \frac{V(\hat{\alpha}_2)}{(\hat{\alpha}_2 - 1 + p^*)^4} \right\} \quad (8)$$

Thus we have a closed form estimate of the multiplicative correction for misclassification of the variance of the corrected estimator.

This method applies only to logistic regression as it depends on the directional invariance of the odds ratio. However, it can still be used in a variety of study designs, provided logistic regression gives a reasonable approximation to the method of choice, for example log-linear or Cox regression. This essentially depends on the rare disease assumption.

RESULTS

Example 1 Cross sectional study of cervicitis

Surveys of cervical cytology in women in Southern Iran found a large proportion (70%–90%) of women with cervicitis.^{15 16} It is of some interest to identify aetiological factors responsible for this. Table 2 shows cervicitis status by bacterial vaginosis (BV) in a cross sectional survey of women in Southern Iran.¹⁷ Overall the observed prevalence of cervicitis is 85%.

Logistic regression gives an unconditional prevalence log odds ratio of 0.48 for the association of BV with cervicitis. This implies an odds ratio of 1.62 (95% confidence intervals (CI): 1.15 to 2.26). Table 3 shows repeat determination of cervicitis by two independent cytologists, in a sample of 72 women from Southern Iran. Assuming symmetric misclassification, this yields an estimate of α of 0.91 with variance 6.90×10^{-4} . From equation (3) our corrected estimate of disease prevalence is 0.93. From equation (4), the corrected estimate of the log odds ratio is

$$\hat{\beta}_1 = 0.48 \times 2.29 = 1.10$$

with variance 1.16. Thus the corrected odds ratio estimate is 3.00 (95% CI: 0.35 to 25.27).

In view of the very high prevalence, one might wish to avoid the assumption of symmetric misclassification. In the absence of further data such as a gold standard validation study, the repeated data in table 3 would not be sufficient to estimate α_1 and α_2 . We can, however, do so if we treat the second reader as if he/she were an expert panel or a definitive test method, and assume that the second reader invariably gives the correct response. In this case, we would estimate α_1 as $50/56 = 0.89$, with binomial variance $(0.89 \times 0.11)/56 = 0.0017$ and α_2 as $10/16 = 0.63$ with variance 0.0146 . This gives a corrected OR of 5.55 (95% CI: 0.21 to 149.02). Note that it is not uncommon to encounter the situation in epidemiology where because of expense or difficulty, the gold standard is only available on a small minority of subjects. This is replicated in the current example, where the overall study size is 1121 subjects, but determination by the second cytologist is only available for 72 of them (6%).

Example 2 Diet and benign proliferative breast disease

Here we consider a case-cohort study of diet and benign proliferative breast disease (BPBD). The study population comprised 545 cases and 4921 non-cases (so $p^* = 0.10$), a

Table 3 Repeat determination of cervicitis in 72 women

Second cytologist	First cytologist		Total
	No	Yes	
No	10	6	16
Yes	6	50	56
	16	56	72

Table 2 Observed cervicitis and bacterial vaginosis

Cervicitis	With bacterial vaginosis (%)	Without bacterial vaginosis (%)	Total
No	66 (12)	105 (18)	171 (15)
Yes	479 (88)	471 (82)	950 (85)
	545 (100)	576 (100)	1121 (100)

sub-cohort of 56 537 women whose dietary intakes were assessed during the Canadian national breast screening study.¹⁸ Table 4 shows the results with respect to total fat intake. Approximating the poisson (log-linear) regression by logistic regression gives an odds ratio of 0.87 (95% CI: 0.69 to 1.10). That is

$$\hat{\beta}_1^* = -0.14$$

$$V(\hat{\beta}_1^*) = 0.0139$$

The poisson analysis gave the same estimate of β but with a slightly lower variance of 0.0126. Cases of BPBD were determined by local pathologists, although some of the biopsied cases and non-cases underwent review by a reference pathologist. Table 5 shows the cross tabulation of local and reference pathologists' findings. Assuming the reference pathologist to be correct, we would estimate the correct classification probability for those with true status positive to be $\alpha_1 = 267/280 = 0.95$, and the correct classification probability for those with true status negative to be $\alpha_2 = 101/328 = 0.31$, an apparent violation of the assumption of symmetric misclassification.

These figures, however, relate to true positives and negatives who actually underwent biopsy because of a radiological or palpable abnormality. Biopsied subjects contain a large proportion of positive cases and those referred to the reference pathologist may not be a representative sample even of the biopsied subjects. The overall true correct classification probabilities are related to α_1 and α_2 by

$$P(\text{Observe} + | +) = P(\text{Observe} + | +, \text{biopsy})$$

$$P(\text{biopsy} | +) + P(\text{Observe} + | +, \overline{\text{biopsy}})$$

$$(1 - P(\text{biopsy} | +)) = \alpha_1 P(\text{biopsy} | +)$$

and

$$P(\text{Observe} - | -) = P(\text{Observe} - | -, \text{biopsy})$$

$$P(\text{biopsy} | -) + P(\text{Observe} - | -, \overline{\text{biopsy}})$$

$$(1 - P(\text{biopsy} | -)) = 1 - P(\text{biopsy} | -)(1 - \alpha_2)$$

As the subjects received some form of breast screening or instruction in breast self examination, it is reasonable to approximate $P(\text{biopsy} | +)$ by 1.0. That is, we assume that those who have the disease outcome undergo biopsy. We therefore take α_1 as our estimate of α , with variance $0.95 \times 0.05 / 280 = 0.00017$. Estimation of the correct classification of truly negative subjects is rendered difficult by a lack of knowledge about $P(\text{biopsy} | -)$, which can only be estimated after considerably more algebra and further assumptions.

Table 4 Cases of benign proliferative breast disease and person years by total fat intake

Total fat intake (g/day)	Cases	Person years
≤102.4	449	160858
>102.4	96	39688

Table 5 Cross tabulation of findings of case status as determined by local pathologist and findings of reference pathologist

Reference pathologist's finding	Case status determined locally		
	Non-case	BPBD case	Total
Non-case	101	227	328
BPBD case	13	267	280

However, if it is approximately equal to the proportion in the sub-cohort who were biopsied with an observed negative result, it is approximately 0.1, which would yield a correct classification probability estimate of just above 0.93. The assumption is reasonable because with repeated breast screening around 10% of subjects might be expected to have at least one suspicious finding requiring further diagnostic examination. Thus the assumption of a common α equal to 0.95 seems reasonable. This gives a corrected log risk ratio of

$$\hat{\beta}_1 = -0.14 / \left\{ \frac{0.95 - 0.1}{(0.9)(0.1)(1 - 0.1)} \right\}$$

$$= \frac{-0.14}{0.52} = -0.27$$

with variance calculated from equation (7) as

$$V(\hat{\beta}_1) = 0.0548$$

Thus the corrected odds ratio estimate is 0.76 with 95% CI (0.48 to 1.21).

DISCUSSION

We have developed a method for adjusting risk estimates for misclassification of the disease end point. The method has the advantage of simplicity, with a closed form estimate. Although the method can adjust for both symmetric and asymmetric misclassification, sometimes the data required for the latter are not available. In our example of cervicitis, we obtained asymmetric misclassification probabilities by using the assumption that the second reader's observation was invariably correct. In practice, one might be as reluctant to make this assumption as to adopt the symmetric misclassification model. To obtain reliable asymmetric misclassification estimates, one would require either a validation study against a gold standard measure, a third determination in the repeated data, or the assumption that true prevalence in the repeated data study was identical to that in the main study. Otherwise, one has to fall back on the assumption of symmetric misclassification.

Table 6 shows the effect of violation of the equality assumption. When the assumption is satisfied, the corrected estimate is consistent. Where it is not, with an absolute difference of more than 0.1 the corrected estimate can be substantially in error. Thus the symmetry assumption is not advisable if it is strongly suspected that there is an absolute

Key points

This paper demonstrates a simple method for correction for bias attributable to misclassification of disease outcome in epidemiological studies.

Table 6 Effect of violation of equality assumption when $p=0.3$, $p(\text{exp}|\text{case})=0.6$, $p(\text{exp}|\text{control})=0.3$ and $\text{OR}=3.5$

α_1	α_2	Obs OR	Est α	Corrected OR	% error
0.95	0.95	2.88	0.95	3.50	0
0.90	0.95	2.59	0.94	3.13	-10
0.80	0.95	2.69	0.91	3.96	+13
0.70	0.95	2.40	0.89	4.27	+22
0.95	0.85	2.33	0.89	3.04	-13
0.90	0.85	2.33	0.87	3.37	-4
0.80	0.85	2.13	0.84	3.58	+2
0.70	0.85	1.95	0.81	4.26	+22
0.85	0.85	2.23	0.85	3.50	0

difference of more than 10% between the two misclassification probabilities.

The correction for misclassification in the cervicitis example yielded both substantial changes to the odds ratio and a considerable widening of the confidence intervals. This illustrates several points. Firstly, in the presence of misclassification, unadjusted for, there may be a significant element of spurious precision. The results from the adjusted analysis in the cervicitis study indicate that there is much more uncertainty about the relation between cervicitis and BV than an uncorrected analysis would suggest. Secondly, if the observed proportion of negative cases is close to the false negative rate, there is a considerable loss of information and consequently a wide range of uncertainty in estimation.

It is likely that, as in many epidemiological situations, most of the assumptions are incorrect. None of symmetric misclassification, equality of misclassification probability between repeat measures, or the existence of a perfect method of disease determination is likely to hold absolutely, although one or other of them may be a reasonable approximation. A good starting point is to consider the possibility of serious misdetermination of disease state. Where such misdetermination is likely on the basis of knowledge of the process of diagnosis and classification, it might then be wise to derive the estimates above, in the first instance to assess the likely magnitude of the effect of such errors on the odds ratio estimates and to gain a more realistic estimate of the uncertainty around these. The credibility of the corrected odds ratios as estimates of the true effects on risk will depend on careful consideration of the plausibility of the assumptions *as approximations* (because they will very rarely hold exactly true) in the context of the particular disease and method of determination. The corrected odds ratios should not be interpreted as reliable estimates of the true effects unless such consideration indicates, and all qualifications on the assumptions should be reported.

In the presence of misclassification of the risk factor in addition to that of the disease outcome, the same relation between the observed and true log odds ratios would apply, that is

$$\hat{\beta}_1 = \hat{\beta}_1^* / (PPV + NPV - 1)$$

In this case, however, the PPV and NPV refer to positive and negative predictive values of observed for true risk factor rather than disease state. If the two misclassification

processes can be assumed independent, the corrections could then be made serially. If however the probability of misclassification of disease status is related to whether or not the risk factor is misclassified, simple closed form estimates of the type proposed here are not applicable.

To date, misclassification of disease status has received little attention in epidemiology, compared with misclassification of risk factors. With the existence of simple methods for correction for the phenomenon, it may attract more attention in the future. The phenomenon of misclassification of disease status certainly exists—there are potential applications in psychiatric epidemiology, classification of dyskariosis on cervical smear cytology, stroke type, and numerous other areas.

Authors' affiliations

S W Duffy, J Warwick, Cancer Research UK Department of Epidemiology, Mathematics and Statistics, Wolfson Institute of Preventive Medicine, Queen Mary University of London, London, UK

A R W Williams, Department of Pathology, University of Edinburgh, UK

H Keshavarz, Centers for Disease Control, Atlanta, USA

F Kaffashian, Cancer Intelligence Unit, Strangeways Research Laboratory, Cambridge, UK

T E Rohan, Department of Epidemiology and Social Medicine, Albert Einstein College of Medicine, New York, USA

F Nili, A Sadeghi-Hassanabadi, Shiraz University of Medical Sciences, Shiraz, Iran

Funding: We thank the estate of the late Ali Reza Soudavar for financial support for Homa Keshavarz.

Conflicts of interest: none declared.

APPENDIX

ESTIMATION OF A SYMMETRIC MISCLASSIFICATION RATE AND THE STANDARD ERROR OF THE ESTIMATE FROM REPEATED MEASUREMENT

With true prevalence p and symmetric misclassification probability α , the probability of agreement when the end-point is measured twice is

$$P_1 = \alpha^2 p + (1 - \alpha)^2 (1 - p) + (1 - \alpha)^2 p + \alpha^2 (1 - p) = \alpha^2 + (1 - \alpha)^2$$

The probability of disagreement is

$$P_2 = \alpha(1 - \alpha) + (1 - \alpha)\alpha = 2\alpha(1 - \alpha)$$

Working in terms of the probability of agreement or disagreement enables us to dispense with the nuisance parameter for the prevalence. This is desirable as we may not wish to assume the prevalence in the validation sample to be equal to that in the main study. With repeated data as in

Policy implications

With the existence of simple methods for correction for the phenomenon, it may attract more attention in the future.

table 1, the log-likelihood is therefore

$$L = (a + d) \ln(\alpha^2 + (1 - \alpha)^2) + (b + c) \ln(\alpha(1 - \alpha))$$

Equating the first derivative to zero, the maximum likelihood estimate is a solution of

$$0 = \frac{(a + d)(4\alpha - 2)}{(\alpha^2 + (1 - \alpha)^2)} + (b + c) \left(\frac{1}{\alpha} - \frac{1}{1 - \alpha} \right)$$

So

$$0 = (a + d)(2\hat{\alpha}^2 - 2\alpha) + (b + c)(2\alpha^2 - 2\alpha + 1)$$

Assuming that correct classification is more likely than incorrect, this quadratic equation solves to give

$$\hat{\alpha} = 0.5 + 0.5 \sqrt{\frac{N - 2(b + c)}{N}}$$

The second derivative is

$$\begin{aligned} \frac{d^2 L}{d\alpha^2} &= \frac{(a + d)(4(\alpha^2 + (1 - \alpha)^2) - (4\alpha - 2)^2)}{(\alpha^2 + (1 - \alpha)^2)^2} \\ &\quad - \frac{(b + c)(\alpha^2 + (1 - \alpha)^2)}{\alpha^2(1 - \alpha)^2} \\ &= \frac{8(a + d)\alpha(1 - \alpha)}{(2\alpha^2 - 2\alpha + 1)^2} - \frac{(b + c)(2\alpha^2 - 2\alpha + 1)}{\alpha^2(1 - \alpha)^2} \end{aligned}$$

We estimate the variance of the maximum likelihood estimate of α as

$$V(\hat{\alpha}) = - \left(\frac{d^2 L}{d\alpha^2} \right)^{-1}$$

REFERENCES

- 1 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**:620-4.
- 2 Greenland S. The effect of misclassification in the presence of covariates. *Am J Epidemiol* 1980;**112**:564-9.
- 3 Rosner B, Spiegelman D, Willett WC. Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *Am J Epidemiol* 1980;**132**:734-45.
- 4 Richardson S, Gilks WR. Conditional-independence models for epidemiologic studies with covariate measurement error. *Stat Med* 1983;**12**:1703-22.
- 5 Howell S, Theodor M, Pacey NF, et al. Quality assurance in cytology. Rescreening of previously negative smears from high grade squamous intraepithelial lesions. *Acta Cytol* 1997;**41**:1085-90.
- 6 Glaser SL, Dorfman RF, Clarke CA. Expert review of the diagnosis of Hodgkin disease in a population based cancer registry—interobserver reliability and impact on incidence and survival rates. *Cancer* 2001;**92**:218-24.
- 7 Abramson JH, Abramson ZH. *Making sense of data*. New York: Oxford University Press, 2001.
- 8 Chavance M, Dellatolas G, Lellouch J. Correlated nondifferential misclassifications of disease and exposure: application to a cross-sectional study of the relation between handedness and immune disorders. *Int J Epidemiol* 1992;**21**:537-46.
- 9 Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 1992;**3**:210-15.
- 10 Brenner H, Savitz DA, Gefeller O. The effects of joint misclassification of exposure and disease on epidemiologic measures of association. *J Clin Epidemiol* 1993;**46**:1195-202.
- 11 Shih JH, Albert PS. Latent model for correlated binary data with diagnostic error. *Biometrics* 1999;**55**:1232-5.
- 12 Wang Y, Taylor JMG. Jointly modelling longitudinal and event time data with application to acquired immunodeficiency syndrome. *J Am Stat Assoc* 2001;**96**:895-905.
- 13 Phillips CV. Quantifying and reporting uncertainty from systematic errors. *Epidemiology* 2003;**14**:459-66.
- 14 Phillips CV, LaPole LM. Quantifying errors without random sampling. *BMC Medical Research Methodology* 2003;**3**:9.
- 15 Keshavarz H, Duffy SW, Sotodeh-Maram E, et al. Factors relating to cervicitis in Qashgha'i nomadic women of Southern Iran. *Rev Epidemiol Sante Publique* 1997;**45**:279-85.
- 16 Duffy SW, Keshavarz H, Sadeghi-Hassanabadi A, et al. Prevalence of cervicitis in women living in southern Iran. *Rev Epidemiol Sante Publique* 1999;**47**:86.
- 17 Keshavarz H, Duffy SW, Sadeghi-Hassanabadi A, et al. Risk factors for and relationship between bacterial vaginosis and cervicitis in a high risk population for cervicitis in Southern Iran. *Eur J Epidemiol* 2001;**17**:89-95.
- 18 Rohan TE, Jain M, Miller AB. A case-cohort study of diet and risk of benign proliferative epithelial disorders of the breast (Canada). *Cancer Causes Control* 1998;**9**:19-27.

APHORISM OF THE MONTH

“The following is my entire establishment... your servant, William Henry Duncan”

How frequently do we hear public health practitioners claiming that they cannot do what is needed without more dedicated resources, and how often are we told that public health systems can only cope with three or four priorities? To which my answer is, which of the Ten Commandments is optional? And when we refer back to William Henry Duncan's pioneering work in Liverpool in the 1840s and 50s, when it must have seemed that he would be overwhelmed by the cholera and the slum conditions, his response to Edwin Chadwick's inquiry about his resource base was as described above. Public health is about mobilising and working through others if it is about anything. Duncan worked with the Health of Towns Association, with its preponderance of participants from the business community, the churches, other civic leaders, and even the occasional other doctor to mobilise those resources needed to tackle the threats to health.

JRA