Journal of Epidemiology and Community Health 1992; **46**: 620–624

External validation, repeat determination, and precision of risk estimation in misclassified exposure data in epidemiology

Stephen W Duffy, Dmitrii M Maximovitch, Nicholas E Day

Abstract

***Study objective*—The aim was to quantify the difference in precision of risk estimates in epidemiology between the situations where misclassification of exposure is corrected for by external validation and where it is corrected for by internal repeat measurement. Precision was measured in terms of the expected width of the 95% confidence interval on the odds ratio.**

***Design*—In a hypothetical case-control study, first with 100 cases and 100 controls, then with 100 cases and 1000 controls (the latter to approximate the cohort study situation), expected estimated odds ratios and confidence intervals were calculated based on postulated underlying true odds ratios and misclassification error rates. The sizes of the confidence intervals using the two design strategies were compared, based on the same number of subjects receiving internal repeat measurements as were used in the external validation study.**

***Main results*—Confidence intervals obtained using internal repeat measurement were considerably narrower than those using external validation. Both methods yielded approximately correct point estimates.**

***Conclusions*—In terms of precision, it is preferable to correct for misclassification using internal repeat measurement rather than external validation.**

J Epidemiol Community Health 1992; **46**: 620–624

Mismeasurement of explanatory variables is a basic problem in epidemiology. It results in incorrect estimates of the effect on disease risk of exposure variables, and reduces the ability to control for confounding.[1] It affects both prospective cohort studies and retrospective case-control studies. In the latter, other sources of bias, notably recall and selection bias, may also distort the results, so that mismeasurement is only one of a number of sources of error to be taken into account. In the former, however, we would expect that careful study design and execution would eliminate other forms of bias, and that measurement error would be the major source of distortion in observing the correct "state of nature". In addition, in prospective studies, one is reasonably sure that equal mismeasurement will apply to disease affected and unaffected individuals. In retrospective studies the problem of differential mismeasurement may be present.

The question that arises, in both prospective and retrospective studies, is what observations on the degree of measurement error should be made for purposes of accounting for mismeasurement in the analysis. In this paper, we consider the effect of mismeasurement on estimates of risk in the absence of confounding. The problem of controlling for confounding in the face of measurement error we leave to a later paper.

In the absence of confounding, measurement error which applies equally to diseased and disease free individuals biases estimates of risk towards the null hypothesis. For reviews of the problem and approaches to its solution, see Espeland and Hui[2] and Chen.[3] While research should be planned to minimise the likelihood of such errors, in assessment of, say, smoking, diet, alcohol consumption, or sexual behaviour, errors of measurement can never be ruled out. The researcher must therefore try to correct for such errors at the stage of data analysis.

Two broad strategies are available to correct for mismeasurement. The first is to estimate the required correction to risk estimates from a validation study, independent of the epidemiological study under consideration, possibly using a "gold standard".[4] The second is to measure the quantities subject to error repeatedly within the epidemiological study.[5] Under certain assumptions, both methods will yield unbiased estimates of the odds ratio. In this paper we consider a binary risk factor, measured with error, and investigate the variances of the estimates derived from each method of correction and the corresponding confidence interval widths. Since many prospective studies generate data of the nested case-control type, we consider case-control data with a range of case to control ratios.

Methods

THE PROBLEM

Assume a case-control study with a binary risk factor subject to measurement error that is equally likely in either direction and is the same for cases as for controls. Let the probability of correct classification of the risk factor be α. Let $p_1 = P(\text{true RF}+ \mid \text{case})$ and $p_2 = P(\text{true RF}+ \mid \text{control})$. For estimation to be possible, α must exceed each of p_1, p_2, $1-p_1$ and $1-p_2$.

EXTERNAL VALIDATION STUDY

Suppose we have case-control data and validation against a "gold standard" of the form in table I. Ignoring measurement error, we obtain an odds ratio estimate $\hat{\varphi} = ad/bc$.

The logarithm of the odds ratio has variance asymptotically estimated as

$$\text{Var}\,(\ln(\hat{\varphi})) = 1/a + 1/b + 1/c + 1/d$$

Medical Research Council Biostatistics Unit, Institute of Public Health, University Forvie Site, Robinson Way, Cambridge CB2 2SR, United Kingdom
S W Duffy
N E Day
Department of Epidemiology, All-Union Cancer Research Centre, Academy of Medical Sciences, Moscow, Russia
D M Maximovitch

Correspondence to:
Dr Duffy

Accepted for publication
March 1992

Let $p_1' = P(\text{observed RF+} \mid \text{case})$ and $p_2' = P(\text{observed RF+} \mid \text{control})$; then

$$p_1' = p_1\alpha + (1-p_1)(1-\alpha),$$
$$\text{and } p_2' = p_2\alpha + (1-p_2)(1-\alpha)$$

and the likelihood function is

$$L = [(1-p_2)\alpha + p_2(1-\alpha)]^a [p_2\alpha + (1-p_2)(1-\alpha)]^c [(1-p_1)\alpha + p_1(1-\alpha)]^b [p_1\alpha + (1-p_1)(1-\alpha)]^d \alpha^{(e+h)}(1-\alpha)^{(f+g)} \quad (1)$$

Differentiating the log likelihood yields maximum likelihood estimates:

$$\hat{p}_1 = \frac{\hat{p}_1' - 1 + \hat{\alpha}}{(2\hat{\alpha}-1)} \; ; \; \hat{p}_2 = \frac{\hat{p}_2' - 1 + \hat{\alpha}}{(2\hat{\alpha}-1)} \quad (2)$$

and

$$\hat{\alpha} = \frac{e+h}{e+f+g+h} \quad (3)$$

We estimate p_1' as $d/(b+d)$ and p_2' as $c/(a+c)$ as usual. The odds ratio is estimated as

$$\hat{\varphi} = \frac{\hat{p}_1(1-\hat{p}_2)}{\hat{p}_2(1-\hat{p}_1)} \quad (4)$$

The covariance matrix of $\hat{p}_1$, $\hat{p}_2$, and $\hat{\alpha}$ can be estimated by inverting the matrix of second derivatives and, using the approximation $\text{Var}[\ln(x)] = \text{Var}(x)/x^2$, we can estimate the variance of the logarithm of the odds ratio. The partial derivatives involve only elementary calculus, but they are lengthy and are omitted here. They are available from the authors (SWD), and the second derivatives for the slightly more complicated situation of repeat determinations are given in the appendix.

REPEAT DETERMINATIONS WITHIN THE STUDY

Suppose now that we have data of the form given in table II. Assuming independence of repeat determinations conditional on true state, we have, for cases measured twice:

$$P(\text{RF+ twice} \mid \text{case}) = p_1\alpha^2 + (1-p_1)(1-\alpha)^2 \quad (5)$$

$$P(\text{RF- twice} \mid \text{case}) = (1-p_1)\alpha^2 + p_1(1-\alpha)^2 \quad (6)$$

$$P(\text{RF-, RF+} \mid \text{case}) = 2\alpha(1-\alpha) \quad (7)$$

Table I Notation for case-control data with a binary risk factor and external validation against a "gold standard" method of measurement assumed to be error free

	Case-control data	
Exposure status	*Controls*	*Cases*
RF −	a	b
RF +	c	d

External validation data. Exposure status measured by:

	Imperfect measure	
Gold standard	*RF+*	*RF−*
RF+	e	f
RF−	g	h

RF = risk factor

Table II Notation for a case-control study with repeat determinations of risk factor status on a subset of cases and controls

	Measured twice			*Measured once*	
Group	*RF+ twice*	*RF− twice*	*RF+, RF−*	*RF+*	*RF−*
Cases	a	e	c	x	y
Controls	b	f	d	z	w

RF = risk factor

For cases measured once, we have:

$$P(\text{RF+} \mid \text{case}) = p_1\alpha + (1-p_1)(1-\alpha) \quad (8)$$

$$P(\text{RF-} \mid \text{case}) = (1-p_1)\alpha + p_1(1-\alpha) \quad (9)$$

The same expressions with p_1 replaced by p_2 apply to controls. The likelihood is therefore:

$$L = [p_1\alpha^2 + (1-p_1)(1-\alpha)^2]^a [p_2\alpha^2 + (1-p_2)(1-\alpha)^2]^b [2\alpha(1-\alpha)]^{c+d} [p_1(1-\alpha)^2 + (1-p_1)\alpha^2]^e [p_2(1-\alpha)^2 + (1-p_2)\alpha^2]^f [p_1\alpha + (1-p_1)(1-\alpha)]^x [(1-p_1)\alpha + p_1(1-\alpha)]^y [p_2\alpha + (1-p_2)(1-\alpha)]^z [(1-p_2)\alpha + p_2(1-\alpha)]^w \quad (10)$$

Analytic estimates are not obtainable (unless all subjects have risk factor status measured twice), but the log likelihood can be numerically maximised to give estimates of p_1, p_2, and α. The odds ratio φ is estimated in turn as in equation (3). The variances of the estimates are estimated by inversion of the matrix of second derivatives of the log likelihood. The second derivatives of the likelihood function are given in the appendix, together with analytic solutions for $\hat{p}_1$, $\hat{p}_2$, and $\hat{\alpha}$ in the case where all cases and controls are measured twice. The variance of the log odds ratio is estimated as:

$$\text{Var}(\hat{\varphi}) = \text{Var}[\ln(\hat{p}_1) + \ln(1-\hat{p}_2) - \ln(\hat{p}_2) - \ln(1-\hat{p}_1)] = b_{11}/[\hat{p}_1(1-\hat{p}_1)]^2 + b_{22}/[\hat{p}_2(1-\hat{p}_2)]^2 - 2b_{12}/[\hat{p}_1(1-\hat{p}_1)\hat{p}_2(1-\hat{p}_2)] \quad (11)$$

where b_{ij} are the elements of minus the inverse of the matrix of second derivatives.

COMPARISON OF THE TWO METHODS

Given numbers of cases and controls, the true odds ratios, correct classification probability α and the control prevalence, expected numbers in categories were calculated using equations (1), (3), and (5) − (9). From these, we calculated the estimated odds ratios and 95% confidence intervals expected for each method and for a variety of true odds ratios, correct classification probabilities and control prevalences. Note that the correct classification probability must exceed both case and control probabilities of positive and negative risk factor status.

Results

For various odds ratios, control risk factor prevalences and misclassification rates, table III shows expected point and 95% confidence interval estimates for the odds ratio in a case-control study with 100 cases and 100 controls, for three strategies: ignoring misclassification, using an external validation study also comprised of 200 subjects, and using repeat determinations of risk factor status on all 200 subjects. The case where $\alpha = 1{\cdot}0$ represents no measurement error. It can be seen that both strategies yield approximately correct point estimates, but the repeat determination strategy yields the smaller confidence intervals, in the case of $\alpha = 0{\cdot}9$ approaching those where there is no mismeasurement.

Table III Expected estimated odds ratios and 95% confidence intervals in a case-control study with 100 cases and 100 controls, with risk factor subject to error which is (1) ignored, (2) estimated from an external validation study of size 200, and (3) estimated by repeat determination on all cases and controls

True OR[a]	α[b]	Control prevalence	(1) Ignoring error OR (95% CI[a])	(2) External validation OR (95% CI)	(3) Repeat determination OR (95% CI)
6·0	1·0	0·2	6·0 (3·1, 11·3)		
		0·3	6·0 (3·2, 11·1)		
		0·35	5·9 (3·1, 10·9)		
	0·9	0·2	3·9 (2·1, 7·2)	6·0 (2·4, 14·9)	6·0 (2·9, 12·4)
		0·3	4·1 (2·2, 7·5)	6·2 (2·6, 14·3)	6·2 (3·0, 12·4)
		0·35	4·0 (2·2, 7·3)	5·9 (2·5, 13·9)	6·0 (2·9, 12·1)
	0·8	0·3	2·8 (1·5, 5·0)	5·9 (1·7, 19·7)	5·9 (2·4, 14·2)
		0·35	2·8 (1·5, 5·0)	6·1 (1·7, 20·7)	6·1 (2·4, 15·0)
4·0	1·0	0·2	4·0 (2·1, 7·5)		
		0·3	4·0 (2·2, 7·2)		
		0·35	4·0 (2·2, 7·2)		
	0·9	0·2	2·8 (1·5, 5·2)	4·0 (1·6, 10·1)	4·0 (1·9, 8·2)
		0·3	2·9 (1·6, 5·2)	3·9 (1·7, 8·7)	3·9 (1·9, 7·6)
		0·35	2·9 (1·6, 5·2)	3·9 (1·7, 8·5)	3·9 (1·9, 7·6)
	0·8	0·3	2·3 (1·2, 4·0)	4·0 (1·2, 13·1)	4·0 (1·7, 9·3)
		0·35	2·3 (1·2, 4·0)	4·0 (1·2, 12·5)	4·0 (1·7, 9·2)
	0·7	0·35	1·7 (0·9, 3·0)	3·9 (0·5, 28·2)	3·9 (1·1, 13·3)
2·0	1·0	0·2	2·0 (1·0, 3·8)		
		0·3	2·0 (1·1, 3·6)		
		0·35	2·0 (1·1, 3·6)		
	0·9	0·2	1·6 (0·8, 3·0)	1·9 (0·6, 5·5)	1·9 (0·9, 4·1)
		0·3	1·7 (0·9, 3·1)	2·0 (0·8, 4·6)	2·0 (1·0, 3·9)
		0·35	1·8 (1·0, 3·2)	2·1 (0·9, 4·5)	2·1 (1·0, 4·0)
	0·8	0·3	1·5 (0·8, 2·7)	2·0 (0·5, 8·3)	2·0 (0·9, 4·6)
		0·35	1·5 (0·8, 2·7)	2·0 (0·5, 6·9)	2·0 (0·8, 4·4)
	0·7	0·35	1·3 (0·7, 2·4)	2·1 (0·03, 137·0)	2·1 (0·6, 6·4)

[a]OR = odds ratio; CI = confidence interval.
[b]α = probability of classifying risk factor correctly.

Table IV shows the expected point and interval estimates when the validation study is of size 40 and the repeat measures are performed on 20 cases and 20 controls. The latter strategy again yields the smaller confidence intervals. In the case of repeat measures the point estimates vary rather more, up to a factor of 10% in either direction, than in the case of external validation. This is because in application of equations (5)–(9), rounding error becomes substantial when calculating integer expected cell sizes for the small numbers with repeated measures (20 cases and 20 controls). In real life, neither approach would yield exact point estimates of α, and therefore of φ.

Tables V and VI give the corresponding results when there are 100 cases and 1000 controls, as an approximation to a cohort study, where the majority of the variance comes from the individuals with disease (the cases). The repeat determination strategy is better in terms of the size of the confidence interval on the odds ratio, although this advantage is considerably attenuated when only small numbers of cases and controls have repeat determinations (table VI). Note also that the rounding error is no longer a problem in the repeat determination case when there are large numbers of individuals with repeat determinations.

To check whether these results hold for more robust interval estimates, we recalculated the 95% confidence intervals using the profile likelihood[6] for external validation and repeated measurement in five situations. Results are shown in table VII. These results correspond to those of table III, and the qualitative conclusion that internal repeat measurement leads to greater precision remains the same. It should be noted, however, that for both strategies the upper point of the confidence interval can be very high when measurement is poor. This is because the estimated probability of correct classification is falling almost as low as one of the case or control positive or negative prevalences at the higher values in the interval.

Table IV Expected estimated odds ratios and 95% confidence intervals in a case-control study with 100 cases and 100 controls, with risk factor subject to error which is (1) estimated from an external validation study of size 40, and (2) estimated by repeat determination on 20 cases and 20 controls

True OR[a]	α[b]	Control prevalence	(1) Ignoring validation OR (95% CI[a])	(2) Repeat determination OR (95% CI)
6·0	0·9	0·2	6·0 (2·1, 17·2)	6·6 (2·5, 17·3)
		0·3	6·2 (2·3, 15·9)	5·8 (2·5, 13·2)
		0·35	6·0 (2·2, 15·5)	6·0 (2·6, 13·9)
	0·8	0·3	5·9 (1·3, 25·7)	6·0 (1·8, 19·7)
		0·35	6·1 (1·3, 27·0)	5·4 (1·7, 16·9)
4·0	0·9	0·2	4·0 (1·4, 11·3)	4·3 (1·7, 10·8)
		0·3	3·9 (1·6, 9·3)	4·1 (1·8, 9·1)
		0·35	3·9 (1·6, 9·0)	3·9 (1·8, 8·4)
	0·8	0·3	4·0 (1·0, 16·1)	3·7 (1·3, 10·2)
		0·35	4·0 (1·0, 14·8)	3·7 (1·3, 10·1)
	0·7	0·35	3·9 (0·3, 47·9)	4·2 (0·5, 33·2)
2·0	0·9	0·2	1·9 (0·6, 5·8)	2·0 (0·8, 5·0)
		0·3	2·0 (0·8, 4·7)	2·1 (0·9, 4·3)
		0·35	2·1 (0·9, 4·6)	2·1 (1·0, 4·2)
	0·8	0·3	2·0 (0·4, 9·8)	1·9 (0·7, 4·9)
		0·35	2·0 (0·5, 7·6)	1·9 (0·7, 4·7)
	0·7	0·35	2·1 (0, 3×10^{13})	1·9 (0·5, 6·6)

[a]OR = odds ratio; CI = confidence interval.
[b]α = probability of classifying risk factor correctly.

Discussion

The above results indicate that in the situation considered, in terms of sensitivity, as expressed by the distance of the lower confidence limit from unity, the repeat measurement strategy is superior to the use of an external validation study. Spiegelman and Gray[7] obtain an analogous result for continuous exposure variables, comparing internal with external validation against a better method of exposure measurement. It is perhaps not intuitively obvious why internal repeat measurement should yield better precision, but an analogy can be seen in the case of estimation of a mean from a sample of continuous data subject to measurement error. Suppose we observe data $z = x + \varepsilon$, where ε represents measurement error and is distributed as normal with mean zero and variance σ_ε^2, independent of x, where x has variance σ^2. To estimate the mean, if we take a single sample of size n, the variance of the sample mean will be $(\sigma^2 + \sigma_\varepsilon^2)/n$, even if σ_ε^2 is known exactly from an independent source. If, on the other hand we take repeated measures on the data z to estimate σ_ε^2, the variance of the sample mean will be $\sigma^2/n + \sigma_\varepsilon^2/2n$. This illustrates the principle that repeated measurement affords an opportunity to lower the contribution to the standard error of the component of variance within individuals.

The above principle appears to be general, although the specific results above apply only to the case of a binary risk factor, with non-differential symmetric misclassification and conditional independence of repeat determinations. The epidemiologist should attempt to design the study and questionnaire to minimise the risk of differential measurement error between cases and controls. It should be noted that the conditional independence assumption is very difficult to avoid if any meaningful inference is to take place. Further, work on misclassification in two con-

founded risk factors indicates that maximum likelihood estimates of odds ratios corrected for misclassification are fairly robust to departures from the symmetry assumption,[6] used here for ease of computation. It must be admitted, however, that in practice it may be necessary to estimate more than one parameter for misclassification, for rates which vary by exposure status or by levels of a concomitant factor.

Since the results we give refer to asymptotic approximation of the variances, we checked the results with 500 computer simulations for about a quarter of the situations considered, and for both the external validation and the repeat measures approaches. The results of the simulations were essentially the same as the analytic results.

The results of tables V and VI are of interest first because case-control studies are frequently designed with several controls per case, but perhaps more importantly because they approximate the situation of the cohort study where most of the variance results from the cases of the disease. These tables also show smaller confidence intervals to be associated with the repeat determination approach. Thus it is likely that the strategy of taking repeat determinations within the study is also preferable in terms of precision for cohort studies.

It is likely that the same result holds for matched studies, on the basis that it holds within each matched set, although a likelihood based analysis would be difficult. An alternative might be a Mantel-Haenszel approach.[8]

A further possibility in the repeat determinations strategy is to take a third determination. This may not always be feasible but if it is, it can facilitate either estimation from more complicated models of misclassification or a further increase in power when recruitment of more subjects is not feasible, or when measurement is feared to be poor. Suppose three determinations are made on each subject. Let a = number of cases positive three times; b = number of cases positive twice; c = number of cases positive once; and d = number of cases negative all three times. Let e, f, g and h be the corresponding numbers of controls. The likelihood is:

$$
\begin{aligned}
L = {} & [p_1\alpha^3 + (1-p_1)(1-\alpha)^3]^a \\
& [p_1\alpha^2(1-\alpha) + (1-p_1)\alpha(1-\alpha)^2]^b \\
& [p_1\alpha(1-\alpha)^2 + (1-p_1)\alpha^2(1-\alpha)]^c \\
& [p_1(1-\alpha)^3 + (1-p_1)\alpha^3]^d \\
& [p_2\alpha^3 + (1-p_2)(1-\alpha)^3]^e \\
& [p_2\alpha^2(1-\alpha) + (1-p_2)\alpha(1-\alpha)^2]^f \\
& [p_2\alpha(1-\alpha)^2 + (1-p_2)\alpha^2(1-\alpha)]^g \\
& (p_2(1-\alpha)^3 + (1-p_2)\alpha^3]^h
\end{aligned}
$$

For the five situations in table VII, the profile likelihood for three determinations yielded confidence intervals (3·1, 12·1), (2·0, 9·0), (1·4, 10·5), (0·9, 3·9), and (0·9, 4·6). Thus the third determination has given a further increase in precision, and in all but the third of the five situations has yielded confidence intervals close to those derived in the absence of measurement error (table III). The lower points of the confidence interval are generally moved further from the null, but the difference is considerably smaller than that between one and two measurements as shown in table III. It has, however, given the interval estimate more stability when measurement is poor.

Comparing tables V and VI, the confidence intervals resulting from a repeat sample of size 220 (20% of the study size) are similar to those

Table V Expected estimated odds ratios and 95% confidence intervals in a case-control study with 100 cases and 1000 controls, with risk factor subject to error which is (1) ignored, (2) estimated from an external validation study of size 1100, and (3) estimated by repeat determination on all cases and controls

True OR[a]	α[b]	Control prevalence	(1) Ignoring error OR (95% CI[a])	(2) External validation OR (95% CI)	(3) Repeat determination OR (95% CI)
6·0	1·0	0·2	6·0 (3·9, 9·3)		
		0·3	6·0 (3·7, 9·5)		
		0·35	5·9 (3·6, 9·5)		
	0·9	0·2	3·9 (2·5, 6·0)	6·0 (3·4, 10·5)	6·0 (3·6, 9·8)
		0·3	4·1 (2·6, 6·5)	6·2 (3·3, 11·3)	6·2 (3·6, 10·4)
		0·35	4·0 (2·5, 6·3)	5·9 (3·1, 11·4)	6·0 (3·4, 10·3)
	0·8	0·3	2·8 (1·8, 4·3)	5·9 (2·5, 13·6)	5·9 (3·0, 11·3)
		0·35	2·8 (1·8, 4·4)	6·1 (2·4, 15·3)	6·1 (3·0, 12·2)
4·0	1·0	0·2	4·0 (2·6, 6·1)		
		0·3	4·0 (2·5, 6·1)		
		0·35	4·0 (2·5, 6·2)		
	0·9	0·2	2·8 (1·8, 4·4)	4·0 (2·3, 7·0)	4·0 (2·4, 6·5)
		0·3	2·9 (1·9, 4·5)	3·9 (2·2, 6·8)	3·9 (2·3, 6·4)
		0·35	2·9 (1·8, 4·5)	3·9 (2·1, 6·8)	3·9 (2·3, 6·4)
	0·8	0·3	2·3 (1·4, 3·5)	4·0 (1·8, 8·6)	4·0 (2·2, 7·4)
		0·35	2·3 (1·4, 3·5)	4·0 (1·8, 8·8)	4·0 (2·1, 7·5)
	0·7	0·35	1·7 (1·1, 2·6)	3·9 (1·1, 12·9)	3·9 (1·5, 9·4)
2·0	1·0	0·2	2·0 (1·2, 3·1)		
		0·3	2·0 (1·3, 3·1)		
		0·35	2·0 (1·3, 3·1)		
	0·9	0·2	1·6 (1·0, 2·5)	1·9 (1·0, 3·5)	1·9 (1·1, 3·2)
		0·3	1·7 (1·1, 2·7)	2·0 (1·1, 3·5)	2·0 (1·2, 3·3)
		0·35	1·8 (1·1, 2·7)	2·1 (1·2, 3·5)	2·1 (1·2, 3·3)
	0·8	0·3	1·5 (0·9, 2·3)	2·0 (0·9, 4·3)	2·0 (1·1, 3·7)
		0·35	1·5 (0·9, 2·3)	2·0 (0·9, 4·1)	2·0 (1·1, 3·6)
	0·7	0·35	1·3 (0·8, 2·0)	2·1 (0·6, 6·3)	2·1 (0·9, 4·6)

[a]OR = odds ratio; CI = confidence interval.
[b]α = probability of classifying risk factor correctly.

Table VI Expected estimated odds ratios and 95% confidence intervals in a case-control study with 100 cases and 1000 controls, with risk factor subject to error which is (1) estimated from an external validation study of size 220, and (2) estimated by repeat determination on 20 cases and 200 controls

True OR[a]	α[b]	Control prevalence	(1) External validation OR (95% CI[a])	(3) Repeat determination OR (95% CI)
6·0	0·9	0·2	6·0 (3·3, 11·0)	6·1 (3·4, 10·6)
		0·3	6·2 (3·2, 11·7)	5·9 (3·2, 10·6)
		0·35	6·0 (3·0, 11·7)	6·0 (3·1, 11·4)
	0·8	0·3	5·9 (2·4, 14·5)	5·9 (2·6, 13·4)
		0·35	6·1 (2·2, 16·2)	5·9 (2·4, 14·1)
4·0	0·9	0·2	4·0 (2·2, 7·2)	4·0 (2·3, 6·9)
		0·3	3·9 (2·2, 6·9)	3·9 (2·2, 6·9)
		0·35	3·9 (2·1, 6·9)	3·9 (2·2, 6·9)
	0·8	0·3	4·0 (1·8, 9·0)	4·1 (1·9, 8·5)
		0·35	4·0 (1·7, 9·2)	3·9 (1·8, 8·3)
	0·7	0·35	3·9 (1·0, 14·0)	4·3 (1·1, 15·4)
2·0	0·9	0·2	1·9 (1·0, 3·6)	1·9 (1·0, 3·4)
		0·3	2·0 (1·1, 3·5)	2·1 (1·2, 3·5)
		0·35	2·1 (1·2, 3·5)	2·1 (1·2, 3·5)
	0·8	0·3	2·0 (0·9, 4·4)	2·1 (1·0, 4·1)
		0·35	2·0 (0·9, 4·2)	2·0 (1·0, 3·9)
	0·7	0·35	2·1 (0·6, 6·6)	2·0 (0·7, 5·5)

[a]OR = odds ratio; CI = confidence interval.
[b]α = probability of classifying risk factor correctly.

Table VII Expected estimated odds ratios and 95% confidence intervals calculated from the profile likelihood in a case-control study with 100 cases and 100 controls, with risk factor subject to error which is (1) estimated from an external validation study of size 200, and (2) estimated by repeat determination on all cases and controls

True OR[a]	α[b]	Control prevalence	(1) External validation OR (95% CI[a])	(3) Repeat determination OR (95% CI)
6·0	0·9	0·2	6·0 (2·6, 16·6)	6·0 (2·9, 12·8)
4·0	0·8	0·3	4·0 (1·5, 13·1)	4·0 (1·8, 9·9)
4·0	0·7	0·35	3·9 (0·9, 41·1)	3·9 (1·2, 12·1)
2·0	0·9	0·2	1·9 (0·8, 5·2)	1·9 (0·9, 4·2)
2·0	0·7	0·35	2·1 (0·4, 7·8)	2·1 (0·9, 18·3)

[a]OR = odds ratio; CI = confidence interval.
[b]α = probability of classifying risk factor correctly.

resulting from an external validation study of size 1100 (100% of the study size). In Table V, the variance of α estimated externally is very small indeed. Even if its variance were zero, that is if α were known absolutely, the repeat measurement strategy would yield a smaller confidence interval on the odds ratio. The same phenomenon is observed in a case-control study with 100 cases and 100 controls (tables III and IV). Thus it may be that if precision of estimation is crucial, the researcher may find it worthwhile to take a repeat sample even if external validation information is already available.

This collaboration was made possible by the Joint Agreement of the Governments of the former USSR and the UK on Cooperation in the Field of Medicine and Public Health. We thank Professor D G Zaridze of the All-Union Cancer Research Centre of the Academy of Medical Sciences, Moscow, for cooperation.

1 Kaldor J, Clayton D. Latent class analysis in chronic disease epidemiology. *Stat Med* 1985; **4**: 327–35.
2 Espeland MA, Hui SL. A general approach to analysing epidemiologic data that contain misclassification errors. *Biometrics* 1987; **43**: 1001–12.
3 Chen TT. A review of methods for misclassified categorical data in epidemiology. *Stat Med* 1989; **8**: 1095–106.
4 Rosner B, Willett W C, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Stat Med* 1989; **8**: 1051–69.
5 Qizilbash N, Duffy SW, Rohan TE. Repeat measurement of case-control data: correcting risk estimates for misclassification due to regression dilution of lipids in transient ischaemic attacks and minor ischaemic strokes. *Am J Epidemiol* 1991; **133**: 832–8.
6 Duffy SW, Rohan TE, Day NE. Misclassification in more than one factor in a case-control study: a combination of Mantel-Haenszel and maximum likelihood techniques. *Stat Med* 1989; **8**: 1529–36.
7 Spiegelman D, Gray R. Cost-efficient study designs for binary response data with Gaussian covariate measurement error. *Biometrics* 1991; **47**: 851–69.
8 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**: 366–70.

Appendix

When repeat measures are taken on a subset of cases and controls, the second derivatives with respect to α, p_1 and p_2 are:

$$\frac{\delta 1n(L)}{\delta\alpha^2} = \frac{2a((p_1\alpha^2+(1-p_1)(1-\alpha)^2)-2(p_1-1+\alpha)^2)}{(p_1\alpha^2+(1-p_1)(1-\alpha)^2)^2} + \frac{2b((p_2\alpha^2+(1-p_2)(1-\alpha)^2)-2(p_2-1+\alpha)^2)}{(p_2\alpha^2+(1-p_2)(1-\alpha)^2)^2} + \frac{(c+d)(-2\alpha^2+2\alpha-1)}{(\alpha(1-\alpha))^2} + \frac{2e((p_1(1-\alpha)^2+(1-p_1)\alpha^2)-2(\alpha-p_1)^2)}{(p_1(1-\alpha)^2+(1-p_1)\alpha^2)^2} + \frac{2f((p_2(1-\alpha)^2+(1-p_2)\alpha^2)-2(\alpha-p_2)^2)}{(p_2(1-\alpha)^2+(1-p_2)\alpha^2)^2} - \frac{x(2p_1-1)^2}{(p_1\alpha+(1-p_1)(1-\alpha))^2} - \frac{y(2p_1-1)^2}{((1-p_1)\alpha+p_1(1-\alpha))^2} - \frac{z(2p_2-1)^2}{(p_2\alpha+(1-p_2)(1-\alpha))^2} - \frac{w(2p_2-1)^2}{((1-p_2)\alpha+p_2(1-\alpha))^2}$$

$$\frac{\delta^2 1n(L)}{\delta p_1^2} = -(2\alpha-1)^2\left\{\frac{a}{(p_1\alpha^2+(1-p_1)(1-\alpha)^2)^2} + \frac{e}{(p_1(1-\alpha)^2+(1-p_1)\alpha^2)^2} + \frac{x}{(p_1\alpha+(1-p_1)(1-\alpha))^2} + \frac{y}{((1-p_1)\alpha+p_1(1-\alpha))^2}\right\}$$

$$\frac{\delta^2 1n(L)}{\delta p_2^2} = -(2\alpha-1)^2\left\{\frac{b}{(p_2\alpha^2+(1-p_2)(1-\alpha)^2)^2} + \frac{f}{(p_2(1-\alpha)^2+(1-p_2)\alpha^2)^2} + \frac{z}{(p_2\alpha+(1-p_2)(1-\alpha))^2} + \frac{w}{((1-p_2)\alpha+p_2(1-\alpha))^2}\right\}$$

$$\frac{\delta^2 1n(L)}{\delta\alpha\delta p_1} = 2\alpha(1-\alpha)\left\{\frac{a}{(p_1\alpha^2+(1-p_1)(1-\alpha)^2)^2} - \frac{e}{(p_1(1-\alpha)^2+(1-p_1)\alpha^2)^2}\right\} + \frac{x}{(p_1\alpha+(1-p_1)(1-\alpha))^2} - \frac{y}{((1-p_1)\alpha+p_1(1-\alpha))^2}$$

$$\frac{\delta^2 1n(L)}{\delta\alpha\delta p_2} = 2\alpha(1-\alpha)\left\{\frac{b}{(p_2\alpha^2+(1-p_2)(1-\alpha)^2)^2} - \frac{f}{(p_2(1-\alpha)^2+(1-p_2)\alpha^2)^2}\right\} + \frac{z}{(p_2\alpha+(1-p_2)(1-\alpha))^2} - \frac{w}{((1-p_2)\alpha+p_2(1-\alpha))^2}$$

$$\frac{\delta 1n(L)}{\delta p_1\delta p_2} = 0$$

When repeat determinations are performed on all cases and all controls ($x=y=z=w=0$), maximum likelihood estimates are analytically obtainable as follows:

$$\hat{\alpha} = 0{\cdot}5 + 0{\cdot}5\sqrt{\frac{N-2(c+d)}{N}}$$

$$\hat{p}_1 = \frac{a\hat{\alpha}^2 - e(1-\hat{\alpha})^2}{(2\hat{\alpha}-1)(a+e)}$$

$$\hat{p}_2 = \frac{b\hat{\alpha}^2 - f(1-\hat{\alpha})^2}{(2\hat{\alpha}-1)(b+f)}$$

Work is in hand to produce user friendly programs for the analyses used above. Copies of the present, "unfriendly" versions are available from the authors (SWD).