The effect of misclassification on the estimation of association: a review

MICHAEL HÖFLER¹

¹ Max Planck Institute of Psychiatry, Clinical Psychology and Epidemiology, Munich, Germany

Abstract

Misclassification, the erroneous measurement of one or several categorical variables, is a major concern in many scientific fields and particularly in psychiatric research. Even in rather simple scenarios, unless the misclassification probabilities are very small, a major bias can arise in estimating the degree of association assessed with common measures like the risk ratio and the odds ratio. Only in very special cases – for example, if misclassification takes place solely in one of two binary variables and is independent of the other variable ('non-differential misclassification') – is it guaranteed that the estimates are biased towards the null value (which is 1 for the risk ratio and the odds ratio). Furthermore, misclassification, if ignored, usually leads to confidence intervals that are too narrow. This paper reviews consequences of misclassification. A numerical example demonstrates the problem's magnitude for the estimation of the risk ratio in the easy case where misclassification takes place in the exposure variable, but not in the outcome. Moreover, uncertainty about misclassification can broaden the confidence intervals dramatically. The best way to overcome misclassification is to avoid it by design, but some statistical methods are useful for reducing bias if misclassification cannot be avoided.

Key words: misclassification, measurement error, mental disorders, diagnoses, relative risk

Introduction

What one measures in empirical sciences is usually not the 'reality' or the phenomenon of interest itself but a more-or-less biased 'projection' of it. In medical research, behavioural sciences and other fields the measurement of any characteristic is subject to potential error. Especially in health sciences, measurement error is common not only for psychological variables (Cohen, 1988) but also for physiological phenomena (Brenner, 1996). One reason for this is that error-free measurements or measurements considered as 'gold-standard', respectively, are often much more expensive than error-prone measurements. In the case of categorical variables such as binary diagnoses, measurement error is referred to as 'misclassification'.

In psychiatric research, misclassification is an important issue, for instance, when deriving diagnoses of mental disorders if the best diagnostic standards are

not available (Cohen, 1988; Tsuang and Tohen, 2002). For example, diagnoses derived from the full Composite Diagnostic Interview (CIDI; WHO, 1997) are often regarded as gold-standard diagnoses, and symptoms assessed, for example, with the Composite International Diagnostic Screener (CID-S; Wittchen et al., 1999) can be regarded as crude measurements of diagnoses derived from the full CIDI. In general, misclassification can arise from choosing instruments, informants or their perspective or an inadequate combination or mixture of these aspects when designing a study (Kraemer, Measelle, Ablow, Essex, Boyce and Kupfer, 2003).

In many medical applications concerning diagnosis, however, several measurements exist, none of which is accepted as gold standard. In such cases, the terms and methods to address misclassification described in this paper do not apply. Instead, by using

appropriate methods of factor analysis, several outcome measures can often be usefully combined into one or more dimensions that reflect essential common features of the different measures and reduce variance across different measurements (see, for example, Dunn, 2000; Kraemer et al., 2003).

Erroneous measurements usually yield biased results for both the estimation of marginal parameters such as prevalence or incidence rates and the estimation of associations. Errors in measurement are often classified into 'random' and 'systematic' errors. Random errors refer to the case where mismeasurement does not introduce a bias into the estimation of a rate – the error is zero in expectation indicating that, on average, there is no error. Nevertheless, random errors almost always introduce a bias in the estimation of the degree of association. Furthermore, despite bias in point estimates of the magnitude of association, misclassification usually leads to an exaggerated precision in estimates; for instance confidence intervals that ignore misclassification can be far too narrow. The reason is that classification errors add further noise to the data (Reade-Christopher and Kupper, 1991; Neuhaus, 1999).

In this paper I will primarily review the impact of misclassification on point estimates of association regarding common measures like the risk ratio or the odds ratio. In a numerical example, the severity of the problem is examined making different assumptions about the mechanism of misclassification. In the final section, some statistical methods suitable to reduce bias introduced by misclassification are discussed. As we shall see, however, the best method to deal with misclassification is to avoid it by design.

Statistical theory of misclassification

During the last three decades statistical theory has produced many findings about the impact of misclassification on the estimation of associations between categorical variables. Assume that one wants to estimate the degree of association between two binary variables, X and Y . Each of them is coded as 0 if the condition is absent and coded as 1 if the condition is present. X and Y denote the variables for the 'true value' and X^* and Y^* represent the corresponding variables for the observed values or *surrogates* of X and Y that are prone to errors.

X and Y may represent two variables considered at the same temporal footing. Imagine that the research aim is to analyse the degree of comorbidity between two diagnoses and that the data come from a cross-

sectional study. Alternatively, X may be an exposure variable or a potential risk factor that precedes Y . Here, the temporal ordering could be established, preferably, by using a prospective study design. X and Y may denote the presence of a particular mental disorder such as major depression (MDD) in parents and in offspring, respectively. The magnitude of association between X and Y here would represent the degree of familial aggregation of the disorder under consideration. More generally, the presence of the disorder in the relatives of affected probands could be analysed with regard to the presence of the disorder in relatives of unaffected probands. Depending, for instance, on the kind of sample used (clinical versus general population samples) and the source of information (direct versus indirect information) used for deriving diagnoses, misclassification was often found to be a serious concern here (for example, Cohen, 1988; Kendler et al., 1991; Chapman TF et al., 1994; Kendler, 1995, 1995; Sullivan et al., 1996).

The amount of misclassification, say in X , is expressed by the size of misclassification probabilities. In the example above, imagine that X^* denotes an error-prone method of deriving a diagnosis for MDD in the parents and that X represents an appropriate gold standard. Then there are two misclassification probabilities in each variable: $P(X^* = 1 | X = 0)$ denotes the conditional probability that MDD in the parents is observed although there is actually no MDD (false positive rate), and $P(X^* = 0 | X = 1)$ is the probability that there is apparently no MDD although the diagnostic criteria are actually met (false negative rate). Accordingly, the corresponding probabilities of measuring without error

$$\begin{aligned} P(X^* = 1 | X = 1) \\ \text{and} \\ P(X^* = 0 | X = 0) \end{aligned}$$

are the sensitivity and specificity, respectively, of the erroneous instrument as measures or surrogate of the 'true' value or gold standard.

Misclassification is broadly classified in three different ways. First, misclassification can take place in one variable only or among both variables concerned. In our example, the diagnosis of MDD in parents (X) could be subject to misclassification, but not the diagnosis of the offspring (Y), or vice versa, or both X and Y could be erroneous. Second, the misclassification (for example, in X) could vary across $Y = 0$ and $Y = 1$. For instance, the probability of a false positive diagnosis

among the parents might depend on the presence of a true diagnosis of MDD among the offspring, that is

$$P(X^* = 1 \mid X = 0, Y = 0)$$

might differ from

$$P(X^* = 1 \mid X = 0, Y = 1).$$

The latter case is called 'differential misclassification'; the first case is referred to as 'non-differential misclassification'. In our example, when using the so-called 'family history method' (informants are asked about disorders in relatives), it was indeed often observed that affected children were more sensitively reporting MDD in relatives (Orvaschel et al., 1982; Kendler et al., 1991; Chapman et al., 1994; Kendler and Roy, 1995) – but not always (Lieb et al., 2002). Third, if both variables are subject to misclassification, the misclassification probabilities in X and Y could be correlated or uncorrelated.

Theoretical results on misclassification

In the easiest case, only one of two binary variables (X or Y , respectively) is misclassified, and misclassification is non-differential. Here, the magnitude of association in terms of common measures like the risk ratio

$$RR = \frac{P(Y = 1 \mid X = 1)}{P(Y = 1 \mid X = 0)}$$

or the odds ratio is always biased toward the null value (which is 1 here). The degree of error in estimation is proportional to the magnitude of misclassification probabilities and inversely proportional to the relative frequency of both variables (Copeland, Checkoway, McMichael and Holbrook, 1977; Gladen and Rogan, 1979). Flegal, Browne and Haas (1986) have shown that if the sum of sensitivity (Se) and specificity (Sp) is less than 1 the direction of the risk ratio is reversed, thus, a risk ratio that is truly > 1 appears to be < 1 and vice versa. (In that odd scenario the instrument works worse than throwing a coin.) In the case where misclassification takes place in the exposure X they have also demonstrated that, if the prevalence of the exposure $p_x = P(X = 1)$ is high as compared to the risk ratio (precisely: if $p_x > 1/[1 + (\sqrt{RR})]$), then the sensitivity of X^* as a surrogate of the true exposure X affects the bias more than the specificity does; otherwise the specificity is more influential. The apparent risk ratio, the risk ratio between X^* and Y , here depends on the outcome rates via the true risk ratio only.

The above finding of bias towards the null value in marginal associations (only one factor is assessed irrespective of other factors) generalizes to the case where two covariates are jointly analysed, only one of which is misclassified (Greenland, 1980). It does not generalize, however, to the case where at least one of two variables takes on more than two categories. Counterexamples were found where the true values and the surrogate measurement were related in a non-monotonic way across ordered categories (for example, in Weinberg, Umbach and Greenland, 1994). If misclassification is differential, the bias often becomes stronger and considerable underestimation is possible as well (Goldberg, 1975; Barron, 1977; Copeland et al., 1977). Another scenario is the situation where a control variable is misclassified but X and Y are error free. In this situation, non-differential misclassification (with respect to exposure and outcome) can also yield exaggeration of an association (Fleiss, Levin and Cho Paik, 2003, Chapter 7).

If misclassification takes place in both binary variables, things become more complicated. Non-differential misclassification with uncorrelated misclassification probabilities yields bias towards the null value in a similar way as in the case above (Gladen and Rogan, 1979; Brenner, Savitz and Gefeller, 1993). Barron (1977) has provided formulas for the apparent risk ratio in this case. Brenner et al. (1993) have investigated several scenarios with non-differential misclassification but correlated misclassification probabilities. A major bias could occur in both directions. Over-estimation occurred together with a low exposure rate and a higher correlation between the error probabilities. Kristensen (1992) has found similar results for odds ratios.

Neuhaus (1999) has investigated the effect of misclassification on the outcome variable Y within logistic regression models, which correspond with odds ratios. He had assumed that all the covariates $X_1, \dots, X_k$ were measured without error. In this situation, non-differential misclassification means that the misclassification probabilities in Y are equal across the joint values of the covariates $X_1, \dots, X_k$. Under weak assumptions, again, non-differential misclassification yielded under-estimation. Differential misclassification in Y , on the other hand, could have introduced a strong bias in both directions. This was particularly the case if both the sensitivity ($P(Y^* = 1 \mid Y = 1)$) and the specificity ($P(Y^* = 0 \mid Y = 0)$) of the erroneous outcome were not high. He has, furthermore, shown that even for a sensitivity of

80% and a specificity of 90%, a double sample size may be required to achieve the same statistical precision in the result.

In a recent paper, Davidov, Faraggi and Reiser (2003) have investigated the effect of misclassification in a covariate, X_1 , that is differential with respect to the outcome Y within a logistic regression model. They have shown that such misclassification in X_1 may cause considerable bias in the estimation of association between Y and each of the other binary covariates $X_2, \dots, X_k$ in the model – provided that Y and X_1 are related. In general, risk ratios and odds ratios are known to be particularly sensitive to misclassification in the context of other measures of association in 2×2 tables (Kraemer, 1985).

Example

In the following numerical example, it is assumed for simplicity that misclassification takes place only in exposure X and that we want to assess the bias from observing the surrogate X^* instead of the true exposure X . Suppose that adolescents' and young adults' lifetime MDD status (Y) was investigated in a cohort study with the best available instrument and only direct information; for assessing the MDD status of their parents (X), however, indirect information from the offspring had also to be used due to the study design (Lieb et al., 2002). In this example, the true risk ratio – that between X and Y – is always assumed to equal 2, indicating a twofold risk for the outcome Y in the presence of the exposure X . Then the apparent risk ratio $P(Y = 1 | X^* = 1)/P(Y = 1 | X^* = 0)$ depends on the relative frequency of the exposure X , the probability of the outcome under $X = 1$ and $X = 0$, and the sensitivity and specificity of X^* as a surrogate of X . In the case of non-differential misclassification the apparent risk ratio depends on the probability of the outcome under $X = 0$ and $X = 1$ only via the ratio of these probabilities, that is, via the true risk ratio

$$\frac{P(Y = 1 | X = 1)}{P(Y = 1 | X = 0)}$$

In the appendix, formulas are provided for the scenarios covered in this example. In Table 1 misclassification is taken to be non-differential (sensitivity and specificity of X^* are equal for $Y = 1$ and $Y = 0$).

In most cases considered here, and particularly if the exposure is a rare condition (10%), the risk ratio

was considerably underestimated. In the worst case, an exposure frequency of 10% and sensitivity and specificity both of 80%, a risk ratio only of 1.27 was observed. For low-prevalent exposures a poor sensitivity (80%), however, was not found to be of major concern as long as specificity was high. For instance, an exposure rate of 10%, a specificity of 80% and a perfect sensitivity (100%) yielded an observed risk ratio of 1.96. With increasing frequency of exposure, the impact of poor sensitivity increased relative to the impact of poor specificity.

Under differential misclassification the results became much worse and more complicated: the apparent risk ratio also depends on the risk for Y under true exposure ($P(Y = 1 | X = 1)$) and under truly no exposure ($P(Y = 1 | X = 0)$), not only on their ratio. In Table 2a, misclassification is assumed to be higher if the outcome event Y is present, that is, if sensitivity and/or specificity are higher in cases of the outcome as compared to non-cases.

Here, differential sensitivity yielded an overestimation provided that specificity was perfect in both $Y = 1$ and $Y = 0$ and that the rates of exposure and outcome were low. The degree of estimation error, however, was found to be moderate with apparent risk ratios around 2.2, and largely independent of the exposure rate as well as of the outcome frequencies. Differential specificity alone, on the other hand, yielded dramatically underestimated risk ratios. Particularly for low rates in exposure (10%) – as well as the outcome (10% versus 5%) – this was even found to change the sign of association; risk ratios < 1 emerged. In the presence of both differential sensitivity and specificity, the effect of specificity on diminishing the apparent risk ratio dominated in the range of values covered here – especially if, again, the rates of exposure and outcome were very low.

In Table 2b the misclassification probabilities were assumed to be higher in those cases that are affected by the outcome.

Now the findings for differential sensitivity and specificity changed. Differential specificity alone yielded more severely overestimated risk ratios than those found in Table 1b: risk ratios up to 5.83 emerged and, even in the case with the highest rates for X and Y , the risk ratio was considerably elevated with 2.54. The effect of differential sensitivity alone was also more pronounced. The worst case with lowest rates for X and Y resulted in a risk ratio of 0.44.

Table 1. Effect of non-differential misclassification in the exposure (true risk ratio = 2)

Exposure rate $P(X = 1)$	Sensitivity $P(X^* = 1 X = 1)$	Specificity $P(X^* = 0 X = 0)$	Apparent risk ratio between Y and X^*
0.1	0.99	0.99	1.91
0.1	0.95	0.95	1.67
0.1	0.90	0.90	1.48
0.1	0.80	0.80	1.27
0.1	0.99	1.00	2.00
0.1	0.95	1.00	1.99
0.1	0.90	1.00	1.98
0.1	0.80	1.00	1.96
0.1	1.00	0.99	1.92
0.1	1.00	0.95	1.69
0.1	1.00	0.90	1.53
0.1	1.00	0.80	1.36
0.2	0.99	0.99	1.96
0.2	0.95	0.95	1.80
0.2	0.90	0.90	1.65
0.2	0.80	0.80	1.42
0.2	0.99	1.00	2.00
0.2	0.95	1.00	1.98
0.2	0.90	1.00	1.95
0.2	0.80	1.00	1.91
0.2	1.00	0.99	1.96
0.2	1.00	0.95	1.83
0.2	1.00	0.90	1.71
0.2	1.00	0.80	1.56
0.3	0.99	0.99	1.97
0.3	0.95	0.95	1.85
0.3	0.90	0.90	1.72
0.3	0.80	0.80	1.49
0.3	0.99	1.00	1.99
0.3	0.95	1.00	1.96
0.3	0.90	1.00	1.92
0.3	0.80	1.00	1.85
0.3	1.00	0.99	1.98
0.3	1.00	0.95	1.90
0.3	1.00	0.90	1.81
0.3	1.00	0.80	1.68

Discussion

In a numerical example, it was demonstrated that misclassification in the exposure variable alone can have serious consequences on estimating the risk ratio even if sensitivity or specificity rates were not lower than 80% – values that some researchers might consider as satisfactory. In the case of differential misclassification, sensitivity and specificity rates of 90% can already be

sufficiently low to produce high bias. Differential misclassification has been found to be a major problem in studies on the familial aggregation of disorders (see Szatmari and Jones, 1999, and the references therein). More complicated, and, thus often more realistic scenarios (misclassification in both variables and/or correlated misclassification probabilities) tend to magnify the problem. Using other measures for the magnitude of

Table 2a. Misclassification in exposure higher if the outcome event is not present

Exposure rate $P(X = 1)$	Risk if unexposed $P(Y = 1 X = 0)$	Risk if exposed $P(Y = 1 X = 1)$	Sensitivity under $Y = 1$		Specificity under $Y = 1$		Apparent risk ratio $P(Y = 1 X^* = 1) / P(Y = 1 X^* = 0)$	
0.1	0.05	0.10	1.00	0.90	1.00	0.90	1.04	
0.1	0.05	0.10	0.90	0.80	0.90	0.80	0.94	
0.1	0.05	0.10	0.80	0.70	0.80	0.70	0.88	
0.1	0.05	0.10	1.00	0.90	1.00	1.00	2.22	
0.1	0.05	0.10	0.90	0.80	1.00	1.00	2.22	
0.1	0.05	0.10	0.80	0.70	1.00	1.00	2.23	
0.1	0.05	0.10	1.00	1.00	1.00	0.90	0.98	
0.1	0.05	0.10	1.00	1.00	0.90	0.80	0.94	
0.1	0.05	0.10	1.00	1.00	0.80	0.70	0.92	
0.1	0.10	0.20	1.00	0.90	1.00	0.90	1.06	
0.1	0.10	0.20	0.90	0.80	0.90	0.80	0.96	
0.1	0.10	0.20	0.80	0.70	0.80	0.70	0.90	
0.1	0.10	0.20	1.00	0.90	1.00	1.00	2.19	
0.1	0.10	0.20	0.90	0.80	1.00	1.00	2.19	
0.1	0.10	0.20	0.80	0.70	1.00	1.00	2.19	
0.1	0.10	0.20	1.00	1.00	1.00	0.90	1.01	
0.1	0.10	0.20	1.00	1.00	0.90	0.80	0.96	
0.1	0.10	0.20	1.00	1.00	0.80	0.70	0.93	
0.2	0.05	0.10	1.00	0.90	1.00	0.90	1.44	
0.2	0.05	0.10	0.90	0.80	0.90	0.80	1.24	
0.2	0.05	0.10	0.80	0.70	0.80	0.70	1.10	
0.2	0.05	0.10	1.00	0.90	1.00	1.00	2.25	
0.2	0.05	0.10	0.90	0.80	1.00	1.00	2.22	
0.2	0.05	0.10	0.80	0.70	1.00	1.00	2.20	
0.2	0.05	0.10	1.00	1.00	1.00	0.90	1.31	
0.2	0.05	0.10	1.00	1.00	0.90	0.80	1.21	
0.2	0.05	0.10	1.00	1.00	0.80	0.70	1.13	
0.2	0.10	0.20	1.00	0.90	1.00	0.90	1.45	
0.2	0.10	0.20	0.90	0.80	0.90	0.80	1.25	
0.2	0.10	0.20	0.80	0.70	0.80	0.70	1.11	
0.2	0.10	0.20	1.00	0.90	1.00	1.00	2.22	
0.2	0.10	0.20	0.90	0.80	1.00	1.00	2.18	
0.2	0.10	0.20	0.80	0.70	1.00	1.00	2.16	
0.2	0.10	0.20	1.00	1.00	1.00	0.90	1.34	
0.2	0.10	0.20	1.00	1.00	0.90	0.80	1.23	
0.2	0.10	0.20	1.00	1.00	0.80	0.70	1.15	

Table 2b. Misclassification in exposure higher if the outcome event is present

Relative frequency of exposure X $P(X = 1)$	Risk if unexposed $P(Y = 1 X = 0)$	Risk if exposed $P(Y = 1 X = 1)$	Sensitivity		Specificity		Apparent risk ratio $P(Y = 1 X^* = 1) / (P(Y = 1 X^* = 0))$
			$P(X^* = 1 X = 1, Y = 1)$	$P(X^* = 1 X = 1, Y = 0)$	$P(X^* = 0 X = 0, Y = 1)$	$P(X^* = 0 X = 0, Y = 0)$	
0.1	0.05	0.10	0.90	1.00	0.90	1.00	2.82
0.1	0.05	0.10	0.80	0.90	0.80	0.90	1.99
0.1	0.05	0.10	0.70	0.80	0.70	0.80	1.66
0.1	0.05	0.10	0.60	1.00	1.00	1.00	1.78
0.1	0.05	0.10	0.80	0.90	1.00	0.90	0.81
0.1	0.05	0.10	0.70	0.80	1.00	0.80	0.44
0.1	0.05	0.10	1.00	1.00	0.90	1.00	3.07
0.1	0.05	0.10	1.00	1.00	0.80	1.00	4.31
0.1	0.05	0.10	1.00	1.00	0.70	1.00	5.83
0.1	0.10	0.20	0.90	1.00	0.90	1.00	2.71
0.1	0.10	0.20	0.80	0.90	0.80	0.90	1.94
0.1	0.10	0.20	0.70	0.80	0.70	0.80	1.63
0.1	0.10	0.20	0.90	1.00	1.00	1.00	1.80
0.1	0.10	0.20	0.80	0.85	1.00	0.90	0.86
0.1	0.10	0.20	0.70	0.80	1.00	0.80	0.46
0.1	0.10	0.20	1.00	1.00	0.90	1.00	2.93
0.1	0.10	0.20	1.00	1.00	0.80	1.00	3.94
0.1	0.10	0.20	1.00	1.00	0.70	1.00	5.13
0.2	0.05	0.10	0.90	1.00	0.90	1.00	2.29
0.2	0.05	0.10	0.80	0.90	0.80	0.90	1.88
0.2	0.05	0.10	0.70	0.80	0.70	0.80	1.61
0.2	0.05	0.10	0.90	1.00	1.00	1.00	1.74
0.2	0.05	0.10	0.80	0.90	1.00	0.90	1.07
0.2	0.05	0.10	0.70	0.80	1.00	0.80	0.68
0.2	0.05	0.10	1.00	1.00	0.90	1.00	2.60
0.2	0.05	0.10	1.00	1.00	0.80	1.00	3.33
0.2	0.05	0.10	1.00	1.00	0.70	1.00	4.25
0.2	0.10	0.20	0.90	1.00	0.90	1.00	2.26
0.2	0.10	0.20	0.80	0.90	0.80	0.90	1.86
0.2	0.10	0.20	0.70	0.80	0.70	0.80	1.60
0.2	0.10	0.20	0.90	1.00	1.00	1.00	1.76
0.2	0.10	0.20	0.80	0.85	1.00	0.90	1.15
0.2	0.10	0.20	0.70	0.80	1.00	0.80	0.71
0.2	0.10	0.20	1.00	1.00	0.90	1.00	2.54
0.2	0.10	0.20	1.00	1.00	0.80	1.00	3.18
0.2	0.10	0.20	1.00	1.00	0.70	1.00	3.96

association (odds ratio, risk difference, Cohen's Kappa and so forth) would, in general, not yield qualitatively different results (Greenland, 1980; Kraemer, 1985; White, 1986), but the relative impact of poor sensitivity and specificity would change because these measures weight sensitivity and specificity differently (Kraemer et al., 1999).

The best way to deal with misclassification is to avoid it by using a study design that is known to reduce the danger of misclassification: 'You can't fix by analysis what is screwed up in design' (Light, Singer and Willett, 1990). The danger of differential misclassification is reduced if researchers are blind, for instance, to the value of X when assessing Y (White, 1986). Moreover, using different sources of information for exposure and outcome strongly reduces the probability that misclassification errors are correlated (Kristensen, 1992).

If the data have already been collected and misclassification cannot be avoided there are two major approaches to address the problem. The first method builds on the availability of gold-standard measurements in a randomly selected subsample, a so-called 'validation sample' (also called 'double sampling'). Thereby, the relationship between the error-free – or gold-standard – variable and the respective surrogate variable can be extrapolated to those observations where the precise measurement is not available (Fleiss, Levin and Peik, 2003). Statistically, this is closely related to a missing data problem where, for instance, the EM algorithm is known to be a powerful tool (Schafer, 1997). Here, correct statistical inference (for example, valid confidence intervals) strongly depends on the study design and has to take into account the uncertainty in the predicted values (see, for example, Greenland, 1988; Lyles, 2002). Note that, for the design of validation substudies, reducing the sample size (N) and applying the gold standard measurement to all cases can be more cost efficient than measuring the error-prone variable among all N cases and the gold standard measurement among a subsample only. This may apply even if the gold standard measurement is much more costly than the surrogate (Greenland, 1988).

In the other approach, often referred to as 'sensitivity analysis', assumptions about the underlying mechanism of misclassification are incorporated into statistical models. This aims to calculate expected values of the magnitude of association in the present data set under the postulated scenario. Such methods are described by, among others, Copeland et al. (1977), Neuhaus (1999)

and Van den Hout and Van der Heijden (2002). For Bayesian approaches where prior distributions on misclassification probabilities are specified see Gustafson (2004) and Greenland (2005). In general, valid statistical inference is roughly equal between and cross-sectional and cohort studies but for case-control studies the situation is different (Brenner et al., 1993). Here, sampling is conditional on the outcome status Y (or its surrogate Y*; see, for example, Lyles, 2002). For instance, Neuhaus (1999) and Paulino et al. (2003) describe methods for the situation in which misclassification takes place only in the outcome Y. For matched case-control studies see Rice (2003) and Duffy et al. (2003).

In practice, when no validation subsample is available researchers should reflect on potential scenarios of misclassification that might underlie the data under investigation. Sometimes external information is available that might be valid for the present study as well. Scientists should generally assess how robust their results are against a variety of realistic error mechanisms. This would often yield more careful but also more robust conclusions. Because of the uncertainty inherent in such procedures the best way to overcome misclassification, however, remains to evade it by appropriate design.

Acknowledgement

I wish to thank Evelyn Alvarenga for language editing.

References

- Barron BA. The effects of misclassification on the estimation of relative risk. *Biometrics* 1977; 33: 414–18.
- Brenner H. How independent are multiple 'independent' diagnostic classifications? *Statistics in Medicine* 1996; 15: 1377–86.
- Brenner H, Savitz DA, Gefeller O. The effects of joint misclassification of exposure and disease on epidemiologic measures of association. *Journal of Clinical Epidemiology* 1993; 46: 1195–202.
- Chapman TE, Mannuzza S, Klein DF, Fyer AJ. Effects of informant mental disorder on psychiatric family history data. *American Journal of Psychiatry* 1994; 151: 574–9.
- Cohen PC. The effects of instruments and informants on ascertainment. In Dunner D, Gershon ES, Barrett JE (eds) *Relatives at Risk for Mental Disorder*. New York: Raven Press, 1998.
- Copeland KT, Checkoway H, McMichael AJ, Holbrook RH. Bias due to misclassification in the estimation of relative risk. *American Journal of Epidemiology* 1977; 105: 488–95.
- Davidov O, Faraggi D, Reiser B. Misclassification in logistic regression with discrete covariates. *Biometrical Journal* 2003; 5: 541–53.

- Duffy SW, Rohan TE, Kandel R, Prevost TC, Rice K, Myles JP. Misclassification in a matched case-control study of c-erbB-2 overexpression and breast cancer. *Statistics in Medicine* 2003; 22: 2459–68.
- Dunn D. *Statistics in Psychiatry*. London: Arnold, 2000.
- Flegal KM, Browne C, Haas JD. The effects of exposure misclassification on estimates of relative risk. *American Journal of Epidemiology* 1986; 123: 736–51.
- Fleiss JL, Levin B, Cho Paik M. *Statistical Methods for Rates and Proportions*. 3 edn. New York: Wiley.
- Gladden B, Rogan WJ. Misclassification and the design of environmental studies. *American Journal of Epidemiology* 1979; 109: 607–16.
- Goldberg JD. The effects of misclassification on the bias in the difference between two proportions and the relative odds in the fourfold table. *Journal of the American Statistical Association* 1975; 561–7.
- Greenland S. The effect of misclassification in the presence of covariates. *American Journal of Epidemiology* 1980; 112: 564–9.
- Greenland S. Statistical uncertainty due to misclassification: implications for validation substudies. *Journal of Clinical Epidemiology* 1988; 41: 1167–74.
- Greenland S. Multiple bias modelling with analysis from observational data. *Journal of the Royal Statistical Society A* 2005; 168: 267–306.
- Gustafson P. *Measurement Error and Misclassification in Statistics and Epidemiology*. London: Chapman & Hall CRC, 2004.
- Kendler KS. Is seeking treatment for depression predicted by a history of depression in relatives? Implications for family studies of affective disorder. *Psychological Medicine* 1995; 25: 807–14.
- Kendler KS, Roy MA. Validity of a diagnosis of lifetime major depression obtained by personal interview versus family history. *American Journal of Psychiatry* 1995; 152: 1608–13.
- Kendler KS, Silberg JL, Neale MC, Kessler RC, Heath AC, Eaves LJ. The family history method: Whose psychiatric history is measured? *American Journal of Psychiatry* 1991; 148: 1501–4.
- Kraemer HC. The robustness of common measures of 2 × 2 associations to bias due to misclassifications. *American Statistician* 1985; 39: 286–90.
- Kraemer HC, Kazdin AE, Offord DR, Kessler RC, Jensen PS, Kupfer DJ. Measuring the potency of risk factors for clinical or policy significance. *Psychological Methods* 1999; 4: 257–71.
- Kraemer HC, Measelle JR, Ablow JC, Essex MJ, Boyce WT, Kupfer DJ. A new approach to integrating data from multiple informants in psychiatric assessment and research: Mixing and matching contexts and perspectives. *American Journal of Psychiatry* 2003; 160: 1566–77.
- Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 1992; 3: 210–15.
- Lieb R, Isensee B, Höfler M, Pfister H, Wittchen HU. Parental major depression and the risk of depression and other mental disorders in offspring – a prospective-longitudinal community study. *Archives of General Psychiatry* 2002; 59: 365–74.
- Light RJ, Singer JD, Willett JB. *By Design: Planning Research on Higher Education*. Cambridge MA: Harvard University Press, 1990.
- Lyles RH. A note on estimating crude odds ratios in case-control studies with differentially misclassified exposure. *Biometrics* 2002; 58: 1034–7.
- Neuhaus JM. Bias and efficiency loss due to misclassified responses in binary regression. *Biometrika* 1999; 86: 843–55.
- Orvaschel H, Thompson WD, Belanger A, Prusoff BA, Kidd KK. Comparison of the family history method to direct interview. *Journal of Affective Disorders* 1982; 4: 49–59.
- Paulino CD, Soares P, Neuhaus J. Binomial regression with misclassification. *Biometrics* 1999; 55: 670–5.
- Reade-Christopher SJ, Kupper LL. Effects of exposure misclassification on regression analyses of epidemiologic follow-up study data. *Biometrics* 1991; 47: 535–48.
- Rice K. Full-likelihood approaches to misclassification of a binary exposure in matched case-control studies. *Statistics in Medicine* 2003; 22: 3177–94.
- Schafer JL. *Analysis of Incomplete Multivariate Data*. London: Chapman & Hall, 1997.
- Sullivan PF, Wells JE, Joyce PR, Bushnell JA, Mulder RT, Oakley-Browne MA. Family history of depression in clinic and community samples. *Journal of Affective Disorders* 1996; 40: 159–68.
- Szatmari P, Jones MB. Effects of misclassification on estimates of relative risk in family history studies. *Genetic Epidemiology* 1999; 16: 368–81.
- Tsuang MT, Cohen T (eds) *Textbook in Psychiatric Epidemiology*, 2 edn. New York: John Wiley & Sons, 2002.
- Van den Hout A, Van der Heijden PGM. Randomized response, statistical disclosure control and misclassification: a review. *International Statistical Review* 2002; 70: 269–88.
- Weinberg CR, Umbach DM, Greenland S. When will nondifferential misclassification of an exposure preserve the direction of a trend? *American Journal of Epidemiology* 1994; 140: 565–71.
- White E. The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria. *American Journal of Epidemiology* 1986; 124: 816–25.
- WHO. *Composite International Diagnostic Interview (CIDI, version 2.1)*. Geneva: World Health Organization, 1997.
- Wittchen HU, Höfler M, Gander F, Pfister H, Storz S, Üstün TB, Müller N, Kessler RC. Screening for mental

disorders: performance of the Composite International Diagnostic Interview (CID-S). *International Journal of Methods in Psychiatric Research* 1999; 8: 59–70.

Appendix: formula for the apparent risk ratio under misclassification of the exposure

As suggested by an anonymous reviewer, if the misclassification takes place in the exposure X only and is non-differential with regard to the outcome Y , then the apparent risk ratio can be expressed as:

$$RR^* = \frac{p_Y + \delta R/q}{p_Y - \delta R / (1 - q)},$$

where p_Y is the probability that $Y = 1$, with $p_X = P(X = 1)$, $q = p_X Se + (1 - p_X) (1 - Sp)$ is the probability that X^* equals 1 (observable),

$$\delta = P(X = 1, Y = 1) - P(X = 1) P(Y = 1)$$

is a measure of association between the true exposure X and Y , and $R = (Se + Sp - 1)$ is an indicator of the quality of X^* to measure X .

The benefit from this way to express the apparent risk ratio RR^* is that it shows that RR^* and RR are equal if and only if X and Y are independent ($\delta = 0$) or else if $R = 1$, that is, X^* perfectly measures X . If R approaches 0, RR^* approaches 1.

The case of non-differential missclassification is not shown here because of notational complexity but is available on request from the author.

Correspondence: Michael Höfler, Clinical Psychology and Epidemiology, Max Planck Institut of Psychiatry, Kraepelinstr. 2-10, 80804 München, Germany.

Telephone (+49) 89 30622-537.

Fax (+49) 89 30622-544.

Email hoefler@mpipsykl.mpg.de.