International Journal of Epidemiology
© International Epidemiological Association 1996

Vol. 25, No. 6
Printed in Great Britain

LEADING ARTICLE

Basic Methods for Sensitivity Analysis of Biases

SANDER GREENLAND

Greenland S (Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095-1772, USA). Basic methods for sensitivity analysis of biases. *International Journal of Epidemiology* 1996; **25:** 1107–1116.

Background. Most discussions of statistical methods focus on accounting for measured confounders and random errors in the data–generating process. In observational epidemiology, however, controllable confounding and random error are sometimes only a fraction of the total error, and are rarely if ever the only important source of uncertainty. Potential biases due to unmeasured confounders, classification errors, and selection bias need to be addressed in any thorough discussion of study results.

Methods. This paper reviews basic methods for examining the sensitivity of study results to biases, with a focus on methods that can be implemented without computer programming.

Conclusion. Sensitivity analysis is helpful in obtaining a realistic picture of the potential impact of biases.

Keywords: bias, case-control studies, misclassification, relative risk, odds ratio, validity

Despite the importance of systematic errors and biases, quantitative methods that take account of these biases have seen much less development than methods for addressing random error. There are at least two reasons for this. First, until recently, randomized experiments supplied most of the impetus for statistical developments. These experiments were concentrated in agriculture, manufacturing, and clinical medicine, and often could be designed in such a way that systematic errors played little role in the final results. The second reason for the limited development of methods for addressing bias is more fundamental: Most biases can be fully analysed only if certain additional 'validation' data are available, but such data are not usually collected and are often very limited when they are available. As a result, investigators must resort to less satisfactory partial analyses, or eschew any quantitative assessment of bias.

Quantitative assessments of bias can provide valuable insight into the importance of various sources of error, and can help one better assess the uncertainty of study results. Such assessments may demonstrate that certain sources of bias cannot plausibly explain a study result, or that a bias explanation cannot be ruled out. I here review some basic methods for such assessment. Because methods for bias assessment are not available in packaged programs, the review focuses on a few methods that can be carried out with a hand calculator. While more advanced methods are available, they have seen almost no published applications, and this will undoubtedly remain the case until major software packages incorporate them. Until that time, I believe that the methods given here will constitute the most practical approaches.

Department of Epidemiology, UCLA School of Public Health, Los Angeles, CA 90095–1772, USA.

Analyses of Confounding

Suppose one conducts an analysis of an exposure X and a disease D, adjusting for the recorded confounders, but one is aware of an unmeasured potential confounder. For example, in a case-control analysis of occupational exposure to resin systems (resins) and lung cancer mortality among male transformer assembly workers,[1] the authors could adjust for age and year of death but had no data on smoking. Upon control of age and year at death, a positive association was observed for resins exposure and lung cancer mortality (Odds Ratio [OR] = 1.77, 95% confidence interval [CI] = 1.18–2.64). What role could confounding by smoking have played in this observation?

For simplicity, suppose resin exposure and smoking are measured as simple dichotomies: X = 1 for resin exposed, 0 otherwise; Z = 1 for smoker, 0 otherwise. There is no reason to suspect that resins workers would have smoked more than other workers in the study. Nonetheless, we might wish to know how large the

TABLE 1 *Crude data for case-control study of occupational resins exposure (X) and lung cancer mortality*[a]

	X = 1	X = 0	Total
Cases	$A_{1+} = 45$	$A_{0+} = 94$	$M_{1+} = 139$
Controls	$B_{1+} = 257$	$B_{0+} = 945$	$M_{0+} = 1202$

[a] Ref. 1.

resins-smoking association would have to be so that control of smoking would have removed the resins-lung cancer association. The answer to this question depends on a number of things, among them: 1) The resins-specific associations of smoking with lung cancer risk (i.e. the associations within levels of resins exposure), 2) the resins-specific prevalences of smoking among the controls, and 3) the prevalence of resins exposure among the controls. The latter (resins) prevalence is observed, but we can only speculate about the first two quantities.

It is this speculation, or educated guessing, that forms the basis for sensitivity analysis. One may assume various plausible combinations of values for the smoking-lung cancer association and resins-specific smoking prevalences, and see what sorts of values result for the smoking-adjusted resins-lung cancer association. If all the latter values are substantial, one would have a basis for doubting that the unadjusted resins-lung cancer association is due to confounding by smoking. Otherwise, one has to admit confounding by smoking as a plausible explanation for the observed resins-lung cancer association.

To limit the calculations, I will use the crude data in Table 1 for illustration. Fortunately, there was no evidence of important confounding by age or year in these data, probably because the controls were selected from other chronic disease deaths. For example, the crude OR was 1.76 versus an adjusted OR of 1.77. If it were necessary to stratify on a confounder, however, one would have to repeat the computations given below for each stratum and then combine the results across strata.

Consider the general notation for the expected stratified data given in Table 2. One can use known values of the stratum-specific prevalences to fill in this table and solve for the assumed common OR relating exposure to disease,

$$OR_{DX} = A_{11}B_{01}/A_{01}B_{11} = (A_{1+} - A_{11})(B_{0+} - B_{01})/(A_{0+} - A_{01})(B_{1+} - B_{01}).$$

Suppose the smoking prevalences among the exposed and unexposed populations are estimated or assumed to be P_{Z1} and P_{Z0} and the odds ratio relating the confounder and disease within levels of exposure is OR_{DZ}. Assuming the control group is representative of the source population, one sets $B_{11} = P_{Z1}B_{1+}$ and $B_{01} = P_{Z0}B_{0+}$. Next, to find A_{11} and A_{01}, one solves the pair of equations.

$$OR_{DZ} = \frac{A_{11}(B_{1+} - B_{11})}{(A_{1+} - A_{11})B_{11}} \text{ and } OR_{DZ} = \frac{A_{01}(B_{0+} - B_{01})}{(A_{0+} - A_{01})B_{01}}.$$

These have solutions

$$A_{11} = OR_{DZ}A_{1+}B_{11}/(OR_{DZ}B_{11} + B_{1+} - B_{11}) \quad (1)$$

and

$$A_{01} = OR_{DZ}A_{0+}B_{01}/(OR_{DZ}B_{01} + B_{0+} - B_{01}). \quad (2)$$

Having obtained A_{11}, A_{01}, B_{11}, and B_{01}, one can simply plug these numbers into Table 2 and directly compute the exposure disease odds ratios OR_{DX}. The answers from each smoking stratum should agree, and both should be computed to check the computations.

The preceding estimate of OR_{DX} is sometimes said to be 'indirectly adjusted' for Z, because it is the estimate of OR_{DX} that one would obtain if one had data on the confounder Z and disease D that displayed the assumed prevalences and confounder odds ratio OR_{DZ}. A more precise term for the resulting estimate of OR_{DX} is 'externally adjusted', because it makes use of an estimate of OR_{DZ} obtained from sources external to the study data. The prevalences must also

TABLE 2 *General layout (expected data) for sensitivity analysis and external adjustment for a dichotomous confounder Z*

	Z = 1		Total	Z = 0		Total
	X = 1	X = 0		X = 1	X = 0	
Cases	A_{11}	A_{01}	M_{11}	$A_{1+} - A_{11}$	$A_{0+} - A_{01}$	$M_{1+} - M_{11}$
Controls	B_{11}	B_{01}	M_{01}	$B_{1+} - B_{11}$	$B_{0+} - B_{01}$	$M_{0+} - M_{01}$

be obtained externally; occasionally (and preferably) they may be obtained from a survey of the underlying source population from which the subjects were selected. (Because I assumed the OR are constant across strata, the results do not depend on the exposure prevalence.)

To illustrate external adjustment with the data in Table 1, suppose that the smoking prevalences among the resins exposed and unexposed were 70% and 50%. Then

$$B_{11} = P_{Z1}B_{1+} = 0.70(257) = 179.9$$

and

$$B_{01} = P_{Z0}B_{0+} = 0.50(945) = 472.5.$$

Taking $OR_{DZ} = 5$ for the resins-specific smoking-lung-cancer OR, equations 1 and 2 yield

$$A_{11} = 5(45)179.9/\,[5(179.9) + 257 - 179.9] = 41.45$$

and

$$A_{01} = 5(94)472.5/\,[5(472.5) + 945 - 472.5] = 78.33.$$

Plugging these results into Table 2 we get the stratum-specific resins-lung cancer OR

$$OR_{DX} = \frac{41.45(472.5)}{179.9\ (78.33)} = 1.39 \quad \text{and}$$

$$OR_{DX} = \frac{(45 - 41.45)(945 - 472.5)}{(257 - 179.9)(94 - 78.33)} = 1.39,$$

which agree (as they should). We see that smoking confounding could account for much but not most of the crude resins odds ratio under the assumed values of OR_{DZ}, P_{Z1}, and P_{Z0}.

For a sensitivity analysis, we repeat the above external adjustment process using other plausible values for the prevalences and the confounder effect. Table 3 presents a summary of results using other values for the resins-specific smoking prevalences and the smoking OR. The Table also gives the smoking-resins OR

$$OR_{XZ} = P_{Z1}(1 - P_{Z0})/(1 - P_{Z1})P_{Z0}.$$

It is apparent that there must be a substantial exposure-smoking association to remove most of the exposure-cancer association. Since there was no reason to expect an exposure-smoking association at all, Table 3 supports the notion that the observed resins-cancer association is probably not entirely due to confounding by the dichotomous smoking variable. (The above analysis does not fully address confounding by smoking, however; to do so would require consideration of a polytomous smoking variable.)

The above method readily extends to cohort studies. For data with person-time denominators T_{ji}, we use the T_{ji} in place of the control counts B_{ji} in the above formulas to obtain an externally adjusted rate ratio. For data with count denominators N_{ji}, we use the N_{ji} in place of the B_{ji} to obtain an externally adjusted risk ratio.

The basic idea of sensitivity analysis and external adjustment for confounding by dichotomous variables was initially developed by Cornfield *et al.*[2] and further elaborated by Bross,[3,4] Schlesselman,[5] Yanagawa,[6] Axelson and Steenland,[7] and Gail *et al.*[8] (see also the correction to Schlesselman by Simon[9]). Extensions of these approaches to multiple-level confounders have also been developed.[10–12] Although most of these methods assume the OR or risk ratios are constant across strata, it is possible to base external adjustment on other assumptions.[8]

TABLE 3 *Sensitivity of externally adjusted resins-cancer odds ratio OR_{DX} to choice of P_{Z1} and P_{Z0} (smoking prevalences among exposed and unexposed), and OR_{DZ} (resins-specific smoking-cancer odds ratio)*

P_{Z1}	P_{Z0}	OR_{XZ}	OR_{DZ}		
			5	10	15
0.40	0.30	1.56	OR_{DX} = 1.49	OR_{DX} = 1.42	OR_{DX} = 1.39
0.55	0.45	1.49	1.54	1.49	1.48
0.70	0.60	1.56	1.57	1.54	1.53
0.45	0.25	2.45	1.26	1.13	1.09
0.60	0.40	2.25	1.35	1.27	1.24
0.75	0.55	2.45	1.41	1.35	1.33

Analyses of Misclassification

Nearly all epidemiological studies suffer from some degree of measurement error, which is usually referred to as classification error or misclassification when the variables are discrete. The impact of even modest amounts of error can be profound, yet rarely is the error quantified. A formal analysis of elementary situations can be done, however, using basic algebra,[13–15] and more extensive analyses can be done using a computer package or language that performs matrix algebra routines (such as GAUSS, MATLAB, SAS, and S-Plus).[16,17] I will here focus on the elementary methods for dichotomous variables, which can reveal much about the bias due to misclassification. I will then briefly discuss methods that allow use of validation study data, in which classification rates are themselves estimated from a sample of study subjects.

Consider first the estimation of exposure prevalence from a single sample of subjects (such as a control series). Define

$X = 1$ if exposed, 0 if not

$X^* = 1$ if classified as exposed, 0 if not.

We then have the following four probabilities:

Se = probability someone exposed is classified as exposed
= sensitivity = $\Pr(X^* = 1 \mid X = 1)$

Fn = Probability someone exposed is classified as unexposed
= False-negative probability = $\Pr(X^* = 0 \mid X = 1) = 1 - Se$

Sp = Probability someone unexposed is classified as unexposed
= Specificity = $\Pr(X^* = 0 \mid X = 0)$

Fp = Probability someone unexposed is classified as exposed
= False-positive probability = $\Pr(X^* = 1 \mid X = 0) = 1 - Sp$.

Suppose B_1 subjects are truly exposed and B_0 subjects are truly unexposed. Then

B_1^* = Expected number of subjects classified as exposed
$= SeB_1 + FpB_0$ (3)

and B_0^* = Expected number of subjects classified as unexposed
$= FnB_1 + SpB_0$. (4)

Note that Se + Fn = Sp + Fp = 1, and so the total is unchanged by the misclassification:

$$M_0 = B_1 + B_0 = (Se + Fn)B_1 + (Sp + Fp)\,B_0$$
$$= SeB_1 + FpB_0 + FnB_1 + SpB_0 = B_1^* + B_0^*.$$

In most studies, one observes only the misclassified counts B_1^* and B_0^*. If we assume that the sensitivity and specificity are equal to Se and Sp (with Fn = 1 – Se and Fp = 1– Sp), we can estimate B_1 and B_0 by solving equations 3 and 4. From equation 4, we get

$$B_0 = (B_0^* - FnB_1)/Sp.$$

We can substitute the right side of this equation for B_0 in equation 4, which yields

$$B_1^* = SeB_1 + Fp(B_0^* - FnB_1)/Sp$$

which we then solve for B_1 to get

$$B_1 = (SpB_1^* - FpB_0^*)/(SeSp - FnFp). \quad (5)$$

Finally, we get $B_0 = M_0 - B_1$.

Three important points should be noted about these results:

1) The B_1 and B_0 are only estimates obtained under the assumption that the true sensitivity and specificity are Se and Sp. To make this clear, we should have denoted the solutions by $\hat{B}_1$ and $\hat{B}_0$; for notational simplicity, we have not done so.

2) These solutions will be undefined if SeSp = FnFp, and will be negative if SeSp < FnFp. The latter result means that the classification method (with sensitivity Se and specificity Sp) is worse than random, in the sense that simply tossing a coin (even an unfair one) to classify people as exposed or unexposed would do better. Tossing a coin to classify subjects would yield the same probability of being classified as exposed or unexposed regardless of true exposure, so that Se = Fp, Sp = Fn, and hence SeSp = FnFp.

3) If one has reasonable estimates of the predictive values for the exposure measurement, as when one has an internal validation study, it can be better to perform the analysis using those values. This point will be discussed below.

In most situations, one can assume that a classification (measurement) is better than random. Sensitivity analysis for exposure classification then proceeds by applying formula 5 for various pairs Se and Sp to the non-case data B_1^* and B_0^*, and by applying the analogous

TABLE 4 *Corrected resins-lung cancer mortality odds ratios (OR_{DX}) under various assumptions about the resins exposure sensitivity (Se) and specificity (Sp) among cases and controls*

Cases		Controls				
Se	Sp	Se:	0.90	0.80	0.90	0.80
		Sp:	0.90	0.90	0.80	0.80
0.90	0.90		2.34[a]	2.00	19.3	16.5
0.80	0.90		2.83	2.42[a]	23.3	19.9
0.90	0.80		1.29	1.11	10.7[a]	9.1
0.80	0.80		1.57	1.34	12.9	11.0[a]

[a] Non-differential misclassification.

formulas to estimate A_1 and A_0 from the observed (misclassified) case counts A_1^* and A_0^*:

$$A_1 = (SpA_1^* - FpA_0^*)/(SeSp - FnFp), \qquad (6)$$

from which it follows that $A_0 = M_1 - A_1$, where M_1 is the observed case total. These formulas may be applied to case-control, closed-cohort, or prevalence-survey data. For person-time follow-up data, formula 5 has to be modified by substituting T_1, T_0, T_1^* and T_0^* for B_1, B_0, B_1^* and B_0^*. The formulas may be applied within strata of confounders as well. After application of the formulas, one may compute 'corrected' stratum-specific and summary effect estimates from the estimated true counts. Finally, one tabulates the corrected estimates obtained by using different pairs (Se, Sp), and thus obtains a picture of how sensitive the results are to various degrees of misclassification.

In the above description, I assumed that the misclassification was non-differential, that is, the same values of Se and Sp applied to both the cases (equation 6) and the non-cases (equation 5). This may be a reasonable assumption in many cohort studies, although it is not guaranteed to hold. It is less often reasonable in case-control studies; for example, if cases are more likely to recall exposure (correctly or falsely) than controls, the sensitivity will be higher or the specificity lower for cases relative to controls. When differential misclassification is suspected, one can simply extend the sensitivity analysis by calculating results in which different values of Se and Sp are used for cases and non-cases.

As a numerical example, let us correct the resins-lung cancer data in Table 1 under the assumption that the case sensitivity and specificity are 0.9 and 0.8, and the control sensitivity and specificity are 0.8 and 0.8. This assumption means that exposure detection was somewhat better for cases. (Because this was a record-based study with other cases as controls, it seems unlikely that the actual study would have had such differential misclassification.) From equations 5 and 6 we get

$$B_1 = [0.8(257) - 0.2(945)]/[0.8(0.8) - 0.2(0.2)] = 27.67$$
$$B_0 = 1202 - 27.67 = 1174.33$$
$$A_1 = [0.8(45) - 0.2(94)]/[0.9(0.8) - 0.1(0.2)] = 24.57$$
$$A_0 = 139 - 24.57 = 114.43.$$

These yield a corrected OR of 24.57(1174.33)/114.43(27.67) = 9.1. This value is much higher than the uncorrected OR of 1.8, despite the fact that exposure detection was better for cases.

By repeating the preceding calculation, we obtain a resins-misclassification sensitivity analysis for the data in Table 1. Table 4 provides a summary of the results of this analysis.

As can be seen, under non-differential misclassification, the corrected OR estimates are always further from the null than the uncorrected estimate computed directly from the data (which corresponds to the corrected estimate assuming Se = Sp = 1, no misclassification). This result reflects the fact that, if the exposure is dichotomous, the misclassification is non-differential, and nothing else is misclassified, the bias produced by misclassification is always towards the null. We caution, however, that this rule does not automatically extend to situations in which other variables are misclassified or the exposure is polytomous.[18]

Table 4 also illustrates that, even if one assumes cases are always more likely to be classified as exposed than non-cases, the corrected estimates may be higher than the uncorrected estimate. This outcome reflects the fact that recall bias does not always result in an upwardly biased OR. It is also apparent that the specificity is a much more powerful determinant of the observed

OR than is the sensitivity in this example; this is because the exposure prevalence is low.[19] Finally, the example shows that the uncertainty in results due to the uncertainty about the classification probabilities can easily overwhelm statistical uncertainty: The uncorrected confidence interval in the example extends from 1.2 to 2.6, whereas the misclassification-corrected OR range above 10 if we allow specificities as low as 0.8, even if we assume the misclassification is non-differential.

Disease misclassification. Consider first the estimation of the incidence proportion from a closed cohort or prevalence from a cross-sectional sample. The above formulas can then be directly adapted by redefining Se, Fn, Sp, and Fp to refer to disease: Let

D = 1 if diseased, 0 if not

D* = 1 if classified as diseased, 0 if not

Se = Probability someone diseased is classified as diseased

= Disease sensitivity = $\Pr(D^* = 1 \mid D = 1)$

Fn = False-negative probability = 1 – Se

Sp = Probability someone non-diseased is classified as non-diseased

= Disease specificity = $\Pr(D^* = 0 \mid D = 0)$

Fp = False-positive probability = 1 – Sp.

Suppose A and B are the true number of diseased and nondiseased subjects, and A* and B* are the numbers classified as diseased and non-diseased. Then equations 3, 4 and 5 give the expected relations between A, B and A*, B* with A, B replacing B_1, B_0, A*, B* replacing B_1*, B_0*, and N = A + B = A* + B* replacing M_0. With these changes, equation 5 becomes

$$A = (SpA^* - FpB^*)/(SeSp - FnFp), \qquad (7)$$

and B = N – A. These equations can be applied separately to different exposure cohorts and within strata, and 'corrected' summary estimates can then be computed from the corrected counts. Results of repeated application of this process for different pairs of Se, Sp can be tabulated to provide a sensitivity analysis. Also, the pair Se, Sp can either be kept the same across exposure cohorts (non-differential disease misclassification) or allowed to vary across cohorts (differential misclassification).

The situation is not quite the same for person-time follow-up data. Here, one must replace the specificity Sp and false-positive probability Fp with a different concept, that of the *False-positive rate*, Fr:

Fr = No. of false-positive diagnoses (noncases diagnosed as cases) per unit person-time.

We then have

$$A^* = SeA + FrT \qquad (8)$$

where T is the true person-time at risk. Also, false-negatives (of which there are FnA) will inflate the observed person-time T*; how much depends on how long the false negatives are followed. Unless the disease is very common, however, the false negatives will add relatively little person-time and one can take T to be approximately T*. Upon doing so, one need only solve equation 8 for A:

$$A = (A^* - FrT^*)/Se, \qquad (9)$$

and then get a corrected rate A/T*. Sensitivity analysis proceeds (similarly to before) by applying equation 9 to the different exposure cohorts, computing corrected summary measures, and repeating this process for various combinations of Se and Fr (which may vary across subcohorts).

The preceding analysis of follow-up data is simplistic, in that it does not account for possible effects if exposure accelerates or decelerates the time from incidence to diagnosis. As discussed elsewhere,[20] these effects (which can be subtle) have generally not been correctly analysed in the medical literature.

Often studies make special efforts to verify case diagnoses, so that the number of false positives within the study will be negligible. If such verification is successful, one can assume Fp = 0, Sp = 1, and equations 7 and 9 will then simplify to A = A*/Se. If we examine a risk ratio RR under these conditions, then, assuming non-differential misclassification, the observed RR* will be

$$RR^* = \frac{A_1^*/N_1}{A_0^*/N_0} = \frac{SeA_1/N_1}{SeA_0/N_0} = \frac{A_1/N_1}{A_0/N_0} = RR.$$

In other words, with perfect specificity, non-differential disease misclassification will not bias the risk ratio. Assuming the misclassification negligibly alters person-time, the same will be true for the rate ratio.[21] The preceding facts have an important implication for case-control studies. Suppose cases are carefully screened to remove false positives, and controls are

selected to represent the people or person-time at risk of becoming cases. Then, assuming the disease is uncommon, non-differential disease misclassification will not bias the case-control OR as an estimate of the risk ratio or rate ratio.

Suppose now the cases cannot be screened, so that in a case-control study there may be many false cases (positives). It would be a severe mistake to apply the disease correction equation 7 to case-control data if (as is almost always true) Se and Sp were determined from other than the study data themselves,[17] because the use of different sampling probabilities for cases and controls alters the sensitivity and specificity within the study relative to the source population. To see this, suppose all apparent cases A_1^*, A_0^* but only a fraction f of apparent noncases B_1^*, B_0^* are randomly sampled from a closed cohort in which disease had been classified with sensitivity Se and specificity Sp. The expected numbers of apparent cases and controls selected at exposure level j would then be

$$A_j^* = SeA_j + FpB_j \text{ and } f \cdot B_j^* = f(FnA_j + SpB_j).$$

The numbers of true cases and non-cases at exposure level j in the case-control study are

$$SeA_j + >f \cdot FnA_j = (Se + f \cdot Fn)A_j \text{ and}$$
$$FpB_j + f \cdot SpB_j = (Fp + f \cdot Sp)B_j,$$

while the numbers of correctly classified cases and non-cases in the study are SeA_j and $f \cdot SpB_j$. The sensitivity and specificity in the study are thus

$$SeA_j/(Se + f \cdot Fn)A_j = Se/(Se + f \cdot Fn)$$
$$\text{and } f \cdot SpB_j/(Fp + f \cdot Sp)B_j = f \cdot Sp/(Fp + f \cdot Sp).$$

The study specificity can be extraordinarily far from the population specificity. For example, if Se = Sp = 0.90 and controls are 1% of the population at risk, the study specificity will be 0.01(0.90)/(0.1 + 0.01(0.90)) = 0.47. Use of the population specificity 0.90 instead of the study specificity 0.47 in a sensitivity analysis could produce extremely distorted results.

Confounder misclassification. The effects of dichotomous confounder misclassification can be explored using the methods discussed above for dichotomous exposure misclassification.[22] One may simply apply equations 5 and 6 to the confounder within strata of the exposure (rather than exposure within strata of the confounder) and then compute an exposure-effect summary from the corrected data. The utility of this approach is limited, however, because most confounder adjustments involve more than two strata. With more than two strata, matrix-correction formulas[17] can be used for sensitivity analysis.

Misclassification of multiple variables. So far, I have assumed that only one variable requires correction. In many situations, age and sex (which tend to have negligible error) are the only important confounders, the cases are carefully screened, and only exposure remains seriously misclassified. There are, however, many other situations in which not only exposure but also major confounders (such as smoking level) are misclassified. Disease misclassification may also coexist with these other problems, especially when studying disease subtypes.

In examining misclassification of multiple variables, it is commonly assumed that the classification errors for each variable are independent of errors in other variables.[17] This is a different assumption from that of non-differentiality, which asserts that errors for each variable are independent of the true values of the other variables. Neither, either one, or both assumptions may hold, and both have different implications for bias. The old generalization that 'non-differential misclassification of exposure always produces bias towards the null' is false if the errors are dependent,[23,24] or if exposure has multiple levels.[18]

If all the classification errors are independent across variables we can apply the correction equations in sequence for each misclassified variable, one at a time. For example, in a prevalence survey one may first obtain semi-corrected counts by correcting for exposure misclassification from equations 5 and 6, then further correct these counts for disease misclassification using equation 7. One could also apply the correction for disease first, and then correct exposure. If, however, the classification errors are dependent across variables, one must turn to more complex correction methods based on matrix algebra. The same methods are also needed for corrections of polytomous (multilevel) variables.

Use of validation substudy data. Up to this point I have assumed that Se and Sp are educated guesses, perhaps suggested by external literature. Suppose now that classification probabilities can be estimated directly from an internal validation subsample of the study subjects. A number of statistically efficient ways of using these data are available, including two-stage and missing-data analysis methods. From such methods, correctly classified counts may be estimated using maximum likelihood or related techniques, and full statistics

(including confidence intervals) can be obtained for the resulting effect estimates. Robins *et al.*[25] review and compare a number of such methods.

An important caution in interpreting the results from these and other formal correction methods is that the methods typically assume the validation standard is measured without error. If this is not true i.e. if the validation measurement taken as the truth is itself subject to error—then the corrected estimate will also be biased, possibly severely so.[26] Because most validation measurements are indeed subject to error, there is a role for sensitivity analysis as a supplement to more formal correction methods, even when internal validation data are available, because sensitivity analysis allows us to see the impact of deviations from the validation study results.

When internal validation data are available one may use simpler and more efficient correction formulas based on predictive values, rather than those based on sensitivity and specificity.[27–30] When such data are not available, however, there are reasons for preferring the earlier correction formulas. First, published reports of the performance of instruments usually provide only sensitivity and specificity. The second reason (which may explain the first) is that predictive values heavily depend on true prevalences, which are unknown and which can vary widely from study to study.[31] Therefore, there is rarely a sound basis for extrapolating predictive values from one study or clinical setting to the next. In contrast, arguments can often be made that the sensitivity and specificity of an instrument will be roughly stable across similar populations, at least within levels of disease and covariates such as age, sex, and socioeconomic status. One should not take such arguments for granted, as variations in sensitivity and specificity will occur under many conditions.[32] Nonetheless, variations in sensitivity and specificity will also produce variations in predictive values.[31]

The preceding reason for preferring sensitivity and specificity disappear when corrections are based on internal validation data, since there is then no issue of generalization across populations. In such situations a strong argument can be made for using the predictive-value approach.[29,30]

SELECTION BIAS

Selection bias (including response and follow-up bias) is mathematically perhaps the simplest to deal with, and yet is often the hardest to address convincingly in practical terms. Two extreme and opposite misconceptions should be dispelled immediately. Some early writings implied that selection bias, like confounding, could always be controlled if one obtained data from subjects on factors affecting selection; other writings implied that this was never possible. The truth is that some forms of selection bias ('selection confounding') can be controlled like confounding; other forms can be impossible to control without external information that is rarely (if ever) available.

An example of controllable selection bias is that induced by matching: one need only control the matching factors to remove the bias.[33] Other examples include two-stage studies, in which (like matching) intentionally biased selection is used and the bias is controlled in the analysis.[25,34,35] Examples of ordinarily uncontrollable bias occur when case-control matching is done on factors affected by exposure or disease, such as intermediate factors or disease symptoms or signs, whose population distribution is unknown.[36]

Selection bias is controllable when the factors affecting selection are measured on all study subjects, and either a) these factors are antecedents of both exposure and disease, and so can be controlled like confounders; or b) one knows the joint distribution of these factors (including exposure and disease, if they jointly affect selection) in the entire source population, and so can adjust for the bias using special techniques. A condition equivalent to (b) is that one knows the selection probabilities for each level of the factors affecting selection. Unfortunately, this situation is rare in practice. It usually occurs only when the study incorporates features of a population survey, as in two-stage designs[34,35] and randomized recruitment.[37] In conventional studies, one can usually only control as appropriate and hope no other factors (such as intermediates or disease symptoms) have influenced selection.

There is well known decomposition for the OR[15,33,35] that can be used for sensitivity analysis. Suppose S_{Aj} and S_{Bj} are the probabilities of case and non-case selection at exposure level j. (Alternatively, in a density sampled case-control study, let S_{Bj} be the person-time rate of control selection at exposure level j.) Then the population case counts can be estimated by A_j/S_{Aj} and the population non-case counts (or person–times) can be estimated by B_j/S_{Bj}. Therefore, the corrected OR or rate ratio estimate comparing exposure level j to level 0 is

$$\frac{(A_j/S_{Aj})(B_0/S_{B0})}{(A_0/S_{A0})(B_j/S_{Bj})} = \frac{A_jB_0}{A_0B_j}\left(\frac{S_{Aj}S_{B0}}{S_{A0}S_{Bj}}\right)^{-1}. \qquad (10)$$

In words, a corrected estimate can be obtained by dividing the sample OR by a selection-bias factor $S_{Aj}S_{B0}/S_{A0}S_{Bj}$. Equation 10 can be applied within strata

of confounders, and the selection-bias factor can vary across the strata.

The obstacle to any application is determining or even getting a vague idea of the selection probabilities S_{Aj} and S_{Bj}. Again, these usually can be pinned down only if the study in question incorporates survey elements to determine the true population frequencies. Otherwise, a sensitivity analysis based on equation 10 will have to encompass a broad range of possibilities. Equation 15 does provide one minor insight: No bias occurs if the selection-bias factor is 1. One way the latter will occur is if disease and exposure affect selection independently, in the sense that $S_{Aj} = t_A u_j$ and $S_{Bj} = t_B u_j$, where t_A and t_B are the marginal selection probabilities for cases and non-cases, and u_j is the marginal selection probability at exposure level j (in density case-control studies,[33] t_B would be the marginal rate of control selection). Occasionally one may reason that such independence will hold, or independence can be forced to hold through careful sampling. In other situations there may be good reasons to question the assumption.[38]

In summary, selection bias is mathematically of relatively simple form, but does not seem to lend itself readily to quantitative resolution because necessary external information is usually lacking. It may be no surprise then, that many subject matter controversies (such as those once surrounding exogenous oestrogens and endometrial cancer) have come down to disputes about selection bias in case-control studies.

COMBINED CORRECTIONS

Sensitivity analyses for different biases may be combined into joint analyses. One should, however, give thought to the proper ordering of the cell corrections, since order can make a difference. For example, suppose one wishes to make corrections for uncontrolled smoking confounding and exposure misclassification, and we have external data indicating a likely joint distribution for smoking and exposure. Suppose also that these external data were themselves based on exposure measurements misclassified in a manner similar to the study data (as would be the case if the data came from the same cohort as the study data). The external adjustment for smoking would then yield a hypothetical smoking-stratified table of misclassified exposure by disease, which then must be corrected for misclassification. In other words, the smoking stratification should precede the misclassification correction. On the other hand, if the joint distribution of smoking and exposure used for external adjustment was a purely hypothetical one referring to the true exposure, the misclassification correction should precede the construction of the hypothetical smoking-stratified table.

DISCUSSION

Sensitivity analysis is a quantitative extension of the qualitative speculations which characterize good discussions of study results. In this regard, it can be viewed as an attempt to bridge the gap between conventional statistics, which are based on implausible randomization and random error assumptions,[39] and the more informed but informal inferences that recognize the importance of biases but do not attempt to estimate their effects.

It is possible to apply sensitivity analyses to *P*-values[12] and confidence limits,[12,40] for example by repeatedly applying conventional formulas to the corrected data obtained from each scenario. A problem with such approaches, however, is that they may convey an unduly pessimistic or conservative picture of the uncertainty surrounding results. For example, the lowest lower 95% limit and highest upper 95% limit from a broad ranging analysis will contain a very wide interval that could have a coverage rate much greater than 95%. This problem occurs in part because sensitivity analyses treat all scenarios equally, regardless of plausibility. A more coherent approach would integrate the results using explicit prior distributions for the bias parameters (confounder prevalences, sensitivities, specificities, selection probabilities, etc.).[41] Unfortunately, such an approach would demand too much labour in prior specification and computation to be adopted anytime soon. In the meantime, the basic techniques reviewed here can help one critically evaluate the plausibility of claims that biases could or could not have been responsible for a study result.

ACKNOWLEDGEMENTS

The author is grateful to Corinne Aragaki, George Maldonado, and Wendy McKelvey for their helpful comments.

REFERENCES

[1] Greenland S, Salvan A, Wegman D H, Hallock M F, Smith T J. A case-control study of cancer mortality at a transformer-assembly facility. *Int Arch Occup Environ Health* 1994; **66:** 49–54.

[2] Cornfield J, Haenszel W, Hammond E C, Lillienfeld A M, Shimkin M B, Wynder E L. Smoking and lung cancer: Recent evidence and discussion of some questions. *J Natl Cancer Inst* 1959; **22:** 173–203.

[3] Bross I D J. Spurious effects from an extraneous variable. *J Chron Dis* 1966; **19:** 637–47.

[4] Bross I D J. Pertinency of an extraneous variable. *J Chron Dis* 1969; **20:** 487–95.

[5] Schlesselman J J. Assessing effects of confounding variables. *Am J Epidemiol* 1978; **108:** 3–8.

[6] Yanagawa T. Case-control studies: Assessing the effect of a confounding factor. *Biometrika* 1984; **71:** 191–94.

[7] Axelson O, Steenland K. Indirect methods of assessing the effect of tobacco use in occupational studies. *Am J Ind Med* 1988; **13:** 105–18.

[8] Gail M H, Wacholder S, Lubin J H. Indirect corrections for confounding under multiplicative and additive risk models. *Am J Ind Med* 1988; **13:** 119–30.

[9] Simon R. Re: 'Assessing effect of confounding variables'. *Am J Epidemiol* 1980; **111:** 127–28.

[10] Greenland S. Quantitative methods in the review of epidemiologic literature. *Epidemiol Rev* 1987; **9:** 1–30.

[11] Flanders W D, Khoury M J. Indirect assessment of confounding: Graphic description and limits on effect of adjusting for covariates. *Epidemiology* 1990; **1:** 239–46.

[12] Rosenbaum P R. *Observational Studies*. New York: Springer Verlag, 1995.

[13] Copeland K T, Checkoway H, Holbrook R H, McMichael A J. Bias due to misclassification in the estimate of relative risk. *Am J Epidemiol* 1977; **105:** 488–95.

[14] Greenland S. The effect of misclassification in matched-pair case-control studies. *Am J Epidemiol* 1982; **116:** 402–06.

[15] Kleinbaum D G, Kupper L L, Morgenstern H. *Epidemiologic Research: Principles and Quantitative Methods*. Van Nostrand, 1984.

[16] Barron B A. The effects of misclassification on the estimation of relative risk. *Biometrics* 1977; **33:** 414–18.

[17] Greenland S, Kleinbaum D G. Correcting for misclassification in two-way tables and matched-pair studies. *Int J Epidemiol* 1983; **12:** 93–97.

[18] Dosemeci M, Wacholder S, Lubin J. Does nondifferential misclassification of exposure always bias a true effect toward the null value? *Am J Epidemiol* 1990; **132:** 746–49.

[19] Drews C D, Greenland S. The impact of differential recall on the results of case-control studies. *Int J Epidemiol* 1990; **19:** 1107–12.

[20] Greenland S. A mathematical analysis of the 'epidemiologic necropsy.' *Ann Epidemiol* 1991; **1:** 551–58.

[21] Poole C. Exception to the rule about nondifferential misclassification (abstract). *Am J Epidemiol* 1985; **122:** 508.

[22] Savitz D A, Baron A E. Estimating and correcting for confounder misclassification. *Am J Epidemiol* 1989; **129:** 1062–71.

[23] Kristensen P. Bias from nondifferential but dependent misclassification of exposure and outcome. *Epidemiology* 1992; **3:** 210–15.

[24] Chavance M, Dellatolas G, Lellouch J. Correlated nondifferential misclassifications of disease and exposure. *Int J Epidemiol* 1992; **21:** 537–46.

[25] Robins J M, Rotnitzky A, Zhao L P. Estimation of regression coefficients when some regressors are not always observed. *J Am Stat Assoc* 1994; **89:** 846–66.

[26] Wacholder S, Armstrong B, Hartge P. Validation studies using an alloyed gold standard. *Am J Epidemiol* 1993; **137:** 1251–58.

[27] Tennenbein A. A double sampling scheme for estimating from binomial data with misclassification. *J Am Stat Assoc* 1970; **65:** 1350–61.

[28] Green M S. Use of predictive value to adjust relative risk estimates biased by misclassification of outcome status. *Am J Epidemiol* 1983; **117:** 98–105.

[29] Marshall R J. Validation study methods for estimating exposure proportions and odds ratios with misclassified data. *J Clin Epidemiol* 1990; **43:** 941–47.

[30] Brenner H, Gefeller O. Use of positive predictive value to correct for disease misclassification in epidemiologic studies. *Am J Epidemiol* 1993; **138:** 1007–15.

[31] Morrison A S. *Screening in Chronic Disease*. New York: Oxford University Press, 1985.

[32] Begg C B. Biases in the assessment of diagnostic tests. *Stat Med* 1987; **6:** 411–23.

[33] Rothman K J. *Modern Epidemiology*. Boston: Little, Brown, 1986.

[34] Walker A M. Anamorphic analysis: Sampling and estimation for covariate effect when both exposure and disease are known. *Biometrics* 1982; **38:** 1025–32.

[35] White E. A two-stage design from the study of the relationship between a rare disease and a rare exposure. *Am J Epidemiol* 1982; **115:** 119–28.

[36] Greenland S, Neutra R R. An analysis of detection bias and proposed correction in the study of estrogens and endometrial cancer. *J Chron Dis* 1981; **34:** 433–38.

[37] Weinberg C R, Sandler D P. Randomized recruitment in case-control studies. *Am J Epidemiol* 1991; **134:** 421–32.

[38] Criqui M H, Austin M, Barrett-Connor E. The effect of nonresponse on risk ratios in a cardiovascular disease study. *J Chron Dis* 1979; **32:** 633–38.

[39] Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990; **1:** 421–29.

[40] Greenland S, Robins J M. Confounding and misclassification. *Am J Epidemiol* 1985; **122:** 495–506.

[41] Gelman A, Carlin J B, Stern H S, Robin D B. *Bayesian Data Analysis*. New York: Chapman and Hall, 1995.

(Revised version received May 1996)