STATISTICS IN MEDICINE, VOL. 8, 1031–1040 (1989)

AN OVERVIEW OF ISSUES RELATED TO THE CORRECTION OF NON-DIFFERENTIAL EXPOSURE MEASUREMENT ERROR IN EPIDEMIOLOGIC STUDIES

WALTER WILLETT

Departments of Epidemiology and Nutrition, Harvard School of Public Health, and the Department of Medicine, Brigham and Women's Hospital, and the Channing Laboratory, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115, U.S.A.

SUMMARY

Procedures to correct estimates of association in epidemiologic studies for the effects of exposure measurement error have rarely been employed in practice. The application of correction procedures would be enhanced by methods that allow the inclusion of covariates, provide corrected confidence intervals, are compatible with commonly employed analytic methods, and that are clearly communicated to potential users. Before using such a procedure, it is important to clearly specify the conceptual 'true' exposure, determine the nature of the measurement error, and decide whether a reproducibility study or validity study is required to quantify the error. The careful use of correction procedures promises to improve our knowledge of the quantitative relationships between many exposures and disease since it is likely that we have substantially underestimated the effects of many exposures and overstated our confidence in null results.

KEY WORDS Epidemiology Methods Measurement Error Diet Nutrition

INTRODUCTION

All biological and physical measurements in any branch of science have error; to a large extent, increments in knowledge depend on reducing this inexactness. Progress is therefore critically dependent on continued improvement in the technical aspects of exposure measurement, whether based on questionnaires, biochemical assays, or physical assessments. At some level, however, further reduction in measurement error is difficult or impractical. It is then important to determine the magnitude of the error and evaluate its effect on relationships under investigation. If the effect of measurement error is appreciable, then it may be appropriate to consider a statistical correction using information on the magnitude of the error to better approximate the relationship that would have been observed if no measurement error had been present. Until this time, however, the correction of observed estimates of association for measurement error has rarely been employed in the epidemiologic literature.

This overview will examine the effects of error in the measurement of independent variables, commonly referred to as exposures. Errors also occur in the measurement of disease, the usual dependent variable in epidemiologic studies, but are typically of lesser magnitude and are discussed elsewhere.[1] Reasons why error correction procedures have not been commonly employed in epidemiologic studies will be considered. Next, the general types of errors frequently seen in epidemiologic studies will be described since they will determine the appropriateness of

0277–6715/89/091031–10$05.00

statistical assumptions employed in error correction procedures. Finally, objectives for procedures to correct for measurement error will be discussed.

EFFECTS OF 'NON-DIFFERENTIAL' MEASUREMENT ERROR IN EPIDEMIOLOGIC STUDIES

Quite appropriately, epidemiologists have focused their primary attention on exposure measurement errors that are differential with respect to disease status, since these errors will distort associations in an unpredictable manner. Differential errors can seldom be corrected except by improvement in study design and will not be discussed further in this paper. Exposure measurement error that is independent of disease status, whether in a cohort or case-control study, is referred to as 'non-differential'. These errors are also described as 'random misclassification' when categorical exposures are employed. It is widely appreciated that the effect of this non-differential error is to bias measures of association such as relative risks, rate differences, correlation coefficients and regression coefficients toward the null values (see Kleinbaum *et al.* [2] for a historical review of this issue). Rothman [3] has pointed out that non-differential misclassification has 'historically not been a great source of concern to epidemiologists, who have generally considered it more acceptable to underestimate effects than to overestimate effects'.

Epidemiologists generally appreciate that nearly all exposures are measured with error; this is usually recognized by noting that an observed positive association is likely to underestimate the true relationship. In a general way, it is also recognized that null findings could be the result of severe non-differential misclassification. This possibility is usually either ignored, acknowledged, or addressed qualitatively by statements regarding the validity of the exposure measurement.

The effect of measurement error on statistical power in epidemiologic studies and necessary sample sizes has been discussed widely over the last decade.[4-6] In general, it has become clear that rather moderate degrees of measurement error can have a substantial impact on statistical power and thus sample size requirements.

The effects of measurement error in covariates have also been considered, both for continuous [7] and dichotomous dependent variables.[8] Greenland has shown that non-differential error in the measurement of a variable that confounds the association between the primary exposure and disease will result in incomplete control of confounding, and thus distort the true association. This concept appears to have been widely recognized and statements recognizing this possibility are increasingly being made in the discussion sections of papers. The error in a confounding variable may, of course, bias the primary association in any direction, depending on the nature of the confounding. Furthermore, Greenland demonstrated that error in the measurement of a covariate can result in artifactual appearance of heterogeneity in the odds ratios among categories of the covariate.

Despite the appearance of many articles in the biometric literature on methods to correct for measurement error, both by simple back calculation[9,10] or by more complex statistical methods,[11-15] the epidemiologic literature is essentially devoid of actual applications of these methods. An exception to this is the the correction of correlation coefficients for attenuation due to within-person variation in one of the variables.[16,17]

WHY IS CORRECTION FOR MEASUREMENT ERROR NOT PRACTISED BY EPIDEMIOLOGISTS?

One can only speculate about reasons for the lack of application of procedures to correct for the effects of measurement error in epidemiologic studies; it may be helpful to consider some

possibilities. These may also be viewed as friendly challenges to those who are involved in the development of such procedures.

First, some of the methods for correction, such as the back calculation methods described by Copeland[9] and Barron,[10] have only been presented for bivariate associations. In actual practice, epidemiologic data are almost always multivariate in nature. At a minimum, age must be controlled in nearly any analysis, and often the list of covariates that must be simultaneously considered is long. Second, confidence intervals are now a minimum requirement for most epidemiologic presentations and a number of the proposed methods have not incorporated an estimate of the variance of the corrected parameter that can be used to calculate the corrected intervals. Third, some of the more complex correction procedures have been developed for models that are rarely used in epidemiologic studies, such as probit analysis. While it might be argued that epidemiologists should convert to the use of these models, it seems unrealistic that this will occur rapidly when the models already in use are otherwise adequate. The reality is that the vast majority of analyses with dichotomous outcomes are conducted using stratified analysis, logistic regression, and proportional hazards models. For continuous outcomes, correlation and least-squares regression analysis meet most needs. Fourth, many of the published methods for error correction are written in a highly technical manner for a small statistical audience and are not easily read by most potential users. Finally, the data on reproducibility and validity that are needed for error correction procedures are frequently not available. This limitation, however, is likely to be self-correcting; if acceptable methods become available, epidemiologists will probably be motivated to collect the necessary information.

Types of errors in epidemiologic studies

Since assumptions regarding the nature of errors are commonly made in correction procedures, the types of errors frequently seen in epidemiologic studies will be briefly discussed. Our understanding of the dominant type of error associated with any measurement is crucial in selecting the type of data that are necessary for correction procedures. Because my own interest in the effects of error evolved from studies of diet, examples from this field will be used.

The specific sources of error are innumerable, however, they can be thought of as two general types: random and systematic. The distinction is that for random error, the average value of many repeated measures will approach the true value. For systematic errors, the mean of repeated measurements does not approach the true value. In some instances, the behaviour of repeated measurements can only be considered conceptually since realistic replicates cannot be made. In epidemiologic studies, random or systematic errors or both can occur at two different levels: within a person, and between persons. Thus, at least four general types of error can exist; these are depicted in Figure 1.

Random within-person error is present if repeated measurements, which are usually not actually obtained in epidemiologic studies, fluctuate at random about an individual's true exposure, which is frequently considered as that person's long-term average level. This random error may represent both true biological variation and/or technical errors in measurement. This distinction is often difficult to make and is usually not critical from the standpoint of error correction since the effects on associations are the same if the long-term average exposure is of interest. One or a small number of measurements for a subject will provide an unbiased but imprecise estimate of that person's true exposure. As an example, the day-to-day variation in dietary intake of specific nutrients has frequently been considered to be random within-person error.[18]

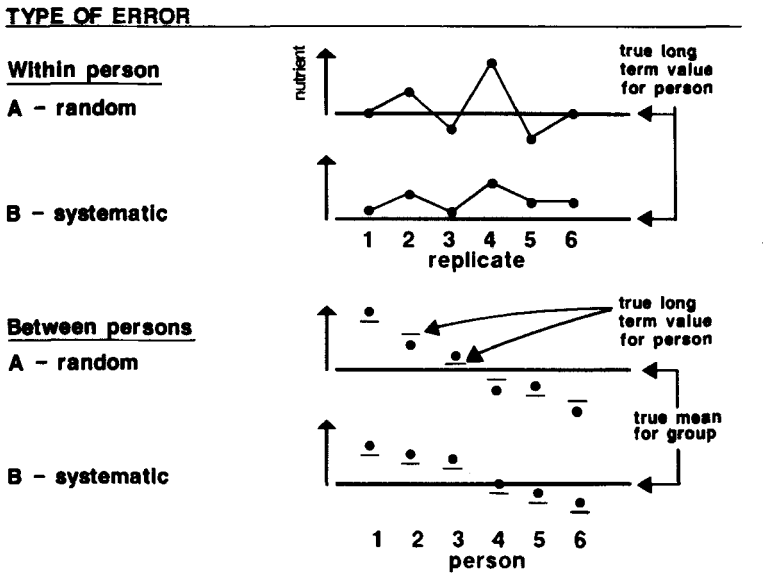

Figure 1. Types of error commonly seen in epidemiologic studies

Within-person errors may also be systematic, such that the average of many repeated measurements does not converge to the subject's true value. For example, when repeated 24-hour recalls of food intake made on different days are used to represent a person's long-term average diet, some subjects may consistently either underestimate or exaggerate their food intake. Systematic within-person error is particularly likely to occur when standardized questionnaires are used; an important food item for a subject (but not necessarily for all subjects) may have been omitted from a questionnaire or misinterpreted by a subject. If such a questionnaire is repeated, this same error is likely to recur; thus the mean of many replicate measurements for an individual will not approach that person's true mean.

Much of the literature addressing the issue of measurement error is based on the assumption that within-person error is strictly random. This may be the result of statistical convenience as well as the usually greater practical difficulty in measuring systematic within-person error. Random within-person error can be measured simply with a single replicate measure for a sample of subjects, that is, a reproducibility study. The measurement of systematic error requires a second, 'gold standard', measure of exposure, that is, a validation study. Unfortunately, no perfect measure exists for any exposure and that is particularly true for dietary intake. The adequacy of the gold standard thus deserves serious consideration in any specific application.

When measuring dietary factors or other exposures among a *group* of persons, errors can also be either random or systematic. Random between-person error can be the result of either using only one or a few replicate measurements per subject in the presence of random within-person error, or the consequence of systematic within-person errors that vary in direction and degree among subjects. Random between-person error implies that an over-estimation for some individuals is counter-balanced by an under-estimation for others so that the mean for a large group of subjects will be the true mean for the group. The standard deviation for the group will, however, be exaggerated.

Systematic between-person error results from systematic within-person error that affects subjects non-randomly. The mean value for a group of persons will thus be incorrect. If the

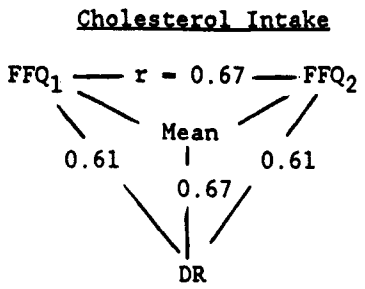

Figure 2. Correlations between cholesterol intake measured by two food frequency questionnaires FFQ1 and FFQ2 administered at an interval of one year and by the mean of our one-week diet records (DR)

systematic error applies equally to all subjects and is simply additive, the observed standard deviation for the group will be correct. However, if individuals are affected to various degrees or the error is multiplicative (for example, proportional to an individual's true level), the standard deviation will be incorrect. Systematic between-person errors are likely to be frequent and can have many causes, for example, a miscalibrated instrument or a defective weighing scale. The omission of a commonly eaten food on a standardized questionnaire or the use of an incorrect nutrient composition value for a common food will affect all individuals in the same direction, but not to the same degree since the use of these foods will differ among subjects. In this last example, it is likely that systematic within-person error will commonly affect individuals unequally. Thus, random and systematic between-person errors are likely to exist in combination. Systematic error that affects all persons equally will not affect most measures of association such as correlation or linear regression coefficients, and relative risks using categories based on ranking. However, random between-person error will generally bias measures of association toward null values, even if it is the consequence of systematic within-person error that differs in direction and magnitude among different persons.

An example of systematic within-person error is provided by a semiquantitative food frequency questionnaire designed to measure long-term diet that was intensively evaluated among 173 women participating in a large prospective study of diet and cancer.[19] Briefly, the questionnaire was administered twice at an interval of about one year, and during that interval participants weighed and recorded all foods and beverages for four one-week periods. These weighed diet records are assumed to reasonably represent true dietary intake over an extended period, being sufficiently extensive to dampen most within-person day-to-day variation and encompassing all seasons. We calculated cholesterol intake, among many other nutrients, from the foods specified on the self-administered questionnaire and from the open-ended data provided by the diet records. Comparing cholesterol intake computed from the two questionnaires, we observed a reasonably strong correlation coefficient ($r=0{\cdot}67$, see Figure 2). Had the error of the questionnaire been strictly random within-person error, we would have expected that the correlation of one questionnaire with truth would have been the square root of the intra-class correlation (0·82, Reference 20). However, the correlation (0·61) between either of the questionnaires and our gold standard (which admittedly is not perfect truth) was actually less than the intra-class correlation. Moreover, the average of the two questionnaires was only slightly more strongly correlated with the diet record cholesterol measurement (0·67). Not surprisingly, for the reasons already cited, these data suggest an element of systematic within-person error associated with the questionnaire such that some of the correlation seen with repeated administrations is simply the result of correlated error.

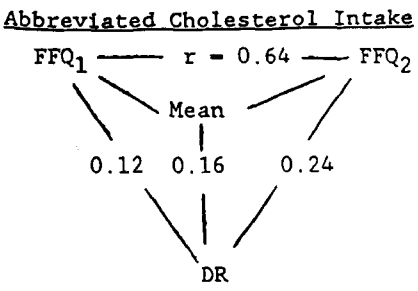

Figure 3. Correlations between cholesterol intake measured by a very abbreviated food frequency questionnaire (FFQ1 and FFQ2, based only on data for whole milk, beef, and cheese intake) and by the mean of four one-week diet records (DR)

The consequences of systematic within-person error can be seen even more strikingly using a subset of the same validation data. Say, for example, our dietary questionnaire had been much less comprehensive and we had asked only about whole milk, beef, and cheese intake and used the responses for these three foods to compute cholesterol intake (see Figure 3). As assessed by reproducibility, the performance of the questionnaire was virtually unchanged ($r = 0{\cdot}64$). However, the validity of the questionnaire assessed by comparison with the diet record was far lower ($r = 0{\cdot}12$ for the first questionnaire and $r = 0{\cdot}24$ for the second questionnaire). Using the average of the two questionnaires provided no improvement in validity ($r = 0{\cdot}16$).

The example provided by the dietary questionnaire is likely to be typical of many questionnaires, physical, and biochemical measures employed in epidemiologic studies; thus, the assessment of performance based only on reproducibility measurements may be very misleading. To provide other examples, the measures of obesity based on weight and height commonly used in epidemiology are highly reproducible (r values typically around 0·95), but correlations with obesity as measured by more exact methods, such as under-water weighing, are appreciably lower (r values around 0·6 after accounting for age, Reference 21). Epidemiologists have commonly assumed that within-person errors in the measurement of blood pressure are random;[22] while this may be reasonably correct, it is likely that an element of systematic error also exists. This could happen, for example, if some individuals are consistently anxious when their blood pressure is measured, or if the anatomy of the arm (which would be constant for any one person) affects blood pressure measurements. True long-term average blood pressure, as might be measured by an indwelling arterial catheter, is difficult to assess in practice and will generally be unknown. While time does not permit the discussion of other examples, it is likely that most epidemiologic measurements are subject to elements of both random and systematic within-person errors, although one form of error or the other may predominate.

OBJECTIVES FOR ERROR CORRECTION PROCEDURES

Before conducting an epidemiologic analysis that incorporates the effects of measurement error, the objectives of this procedure should be considered. Since the focus of contemporary epidemiology is on the quantitative relationships between exposures and disease,[23] and unbiased estimate of effect, rather than a maximally efficient test of the null hypothesis, will nearly always be the first priority of most epidemiologists. At the same time, it is important to realize that epidemiology is likely to remain an approximate rather than an exact science; thus perfect unbiasedness is not critical. In reality, excess relative risks (relative risk − 1) frequently have

95 per cent confidence limits extending beyond 50 per cent of the point estimate. Thus biases in relative risks or regression coefficients of 10 or 20 per cent will often be quite tolerable. For this reason, correction procedures may be unnecessary unless measurement error is substantial, except, perhaps, to provide reassurance that the effect of measurement error was inconsequential. Furthermore, when measurement error is substantial, a correction procedure that provides a reasonable approximation of the unbiased effect may be adequate even if not perfect.

Unbiased confidence intervals are an important second objective for error correction procedures. As previously mentioned, lack of methods to compute corrected confidence intervals has limited the use of some procedures. While epidemiologists generally appreciate that measurement error will bias estimates of association toward null values, it appears less well recognized that such error also causes uncorrected confidence limits to be artifactually narrow. Correction of confidence limits is particularly useful when the point estimate of association in a study is near the null value or not statistically significant because the primary interest then becomes the range of values that are still compatible with the data. Given even a moderate degree of measurement error, appropriate correction of the confidence intervals for each error will yield results consistent with a wider range of non-null values than would be suggested by the uncorrected intervals.

When calculating corrected confidence intervals, the error in estimating the validity of the surrogate or operational measure should also be considered. In some cases the degree of validity will have been well quantified and can be considered to be known without error, however, in other instances the validity of the surrogate measure is itself estimated with a substantial degree of uncertainty. This will occur, for example, if the size of a validation study has been small. In this case, this source of error should also be included in the correction procedure; the effect will be to widen the confidence interval further. For many types of measurements, particularly those based on questionnaires, the validity may vary with the study population, so that estimates of validity from external sources must be used with extreme care. Uncertainty in the estimation of validity has not been incorporated in most of the formal correction procedures published thus far.

It may be tempting to use the fact that measurement error will reduce statistical power to interpret studies that have already been completed. This has been done by calculating an attenuated relative risk that might be expected, given a realistic degree of measurement error, and then computing the statistical power to detect the attenuated relative risk. While this is a logical procedure for planning a study, such *post hoc* calculations are inappropriate for interpreting the results of studies that have already been completed since they do not take into account the observed data. It is even possible, for example, to determine that a study has insufficient power to detect a hypothetical positive association, while the data are actually statistically significant in the opposite direction. Confidence intervals that are adjusted for the effects of measurement will be far superior to *post hoc* power calculations for hypothetical associations since they will incorporate the observed data, the uncertainty due to the size of the study, and the effects of measurement error.

CONSIDERATIONS IN THE APPLICATION OF CORRECTION PROCEDURES

The first task facing an epidemiologist contemplating the use of a correction procedure is to specify the conceptual 'true' exposure. The answer will often be less than obvious since, in one sense, every measure is a surrogate for a more proximal cause of disease. From the ultimate deterministic viewpoint, if the true exposure immediately proximal to the inception of disease were really known, the probability of disease would be one for those exposed and zero for those not exposed. For example, one can envision a chain of events starting with dietary intake of a nutrient,

which influences its blood level and, in turn, the tissue level in the target organ. With newly developed techniques, it may be possible to extend this cascade of events through several additional steps leading to the damage (or protection) of the DNA controlling the expression of a specific tumour growth factor.

The selection of a point in the sequence of causal events that is chosen to represent 'true' exposure should be clearly specified, even if it is somewhat arbitrary. If the object of analysis is to estimate the effect of an intervention, such as would be done in a trial, then the 'true' exposure would be the factor that is potentially to be modified. In the example described above, the 'true' exposure might be the actual nutrient intake if the point of the analysis was to determine the amount of disease that might be avoided by dietary alteration. If, however, the object of the analysis is not related to a specific potential intervention, a more proximal point in the causal chain, such as the tissue level of the nutrient, which could be influenced by metabolic as well as dietary factors, might be chosen as the 'true' exposure.

Even when a specific step in the causal sequence of events is selected as the 'true' exposure, the dimension of time will usually need to be considered since most epidemiologic exposures fluctuate and/or drift temporally within persons. Typically, our understanding of disease etiology is so poor that the time specification can be made only crudely. Still, it will be important to distinguish between differences in exposure on the previous day, over the past year, past ten years, or, say, between ages 10 to 20 years, which may have been decades earlier. Consideration should also be given as to whether average or peak levels represent the conceptually true exposure.

When a measurement is to be used for prediction, as might be done with a screening test, error correction may be inappropriate since the point of the analysis is to evaluate the predictiveness of a measurement, which will be a function of both the underlying biological relationship as well as the validity of the measurement. For example, if we wish to determine the utility of a measure of serum cholesterol to predict coronary heart disease, as it might be used in a screening programme, correction for measurement error would not be appropriate. However, if we are interested in the magnitude of the biological relationship between serum cholesterol and heart disease, for example, to quantify the potential benefit of a long-term reduction in serum cholesterol levels, then correction for the random within-person error in serum cholesterol measurement would be appropriate.

A second major consideration before error correction is undertaken is whether a reproducibility or a validity study is necessary to evaluate the magnitude of measurement error. If the conceptual 'true' exposure has been clearly specified, then this may be obvious in some instances. If the 'true' exposure of interest is by definition the average of many replicates of a single surrogate measure, then the assumption of purely random within-person error is appropriate and a reproducibility study will provide the necessary information. For example, if the object of the analysis is to determine the relation between long-term average serum cholesterol level and risk of heart disease, then a reproducibility study of serum cholesterol, with the time interval between replicates being of similar magnitude as the specified period of interest, will provide the needed data. On the other hand, if we have used a dietary questionnaire to measure food intake, and we wish to know the relation between cholesterol intake and risk of disease, the average of many repeated measurements would not provide the information of interest since the measurement error associated with the questionnaire is likely to have a major systematic within-person component. Therefore, a validation study using an independent, superior measure of intake would be necessary.

Once it has been decided to conduct either a reproducibility or validity study, a number of practical design questions will need to be addressed. The generalizability of this substudy to the main study will be of less concern if it is conducted among a random sample of the total study population. If previously collected data are to be used instead, serious consideration will need to

be given to their appropriateness. The size of the study, the sequence of measurements, and the number of replicate measurements per subject in the case of a reproducibility study deserve careful attention; it is clear that no single solution will be optimal for all applications. Finally, careful consideration should be given to the covariates that will be controlled in the substudy, particularly if correlation coefficients are to be used as a measure of reproducibility or validity. Heterogeneity in variables that contribute to between-person variation in exposure levels, such as age, sex, and body size, will tend to make correlations appear high. However, these variables will usually be controlled in the analysis of the main epidemiologic data, so that correlation coefficients from validation or reproducibility analyses not controlling for these factors will be misleadingly optimistic. Regression coefficients are, in principle, not sensitive to the degree of variation in exposure and will therefore provide a more generalizeable, although frequently less interpretable, measure of validity.

In summary, the quantitative assessment of exposure measurement error and correction for its effects has rarely been practised in epidemiologic studies as of this time. Methodological work on this topic is likely to be one of the most fruitful areas of development in epidemiology during the next several years. However, since many of the issues involved are both conceptual and mathematical, the application of these methods to specific data sets is likely to generate many debates regarding the appropriateness of making any error correction, the assumptions about the nature of the measurement error, the choice of the 'gold standard' for validation studies, and the generalizability of the reproducibility or validity estimate to the main study population. Nevertheless, the application of correction methods now being developed promises to improve our knowledge of the quantitative relationships between many exposures and disease since it is likely that we have substantially under-estimated the effects of many exposures and over-estimated our confidence in null results.

ACKNOWLEDGEMENTS

The author thanks for Tosteson and Donna Spiegelman for their helpful comments.

REFERENCES

1. White, E. 'The effect of misclassification of disease status in follow-up studies: implications for selecting disease classification criteria', *American Journal of Epidemiology*, **124**, 816–825 (1986).
2. Kleinbaum, D. G., Kupper, L. L. and Morganstern, H. *Epidemiologic Research*, Lifetime Learning Publications, Belmont, 1982.
3. Rothman, K. J. *Modern Epidemiology*, Little, Brown, Boston, 1986.
4. Quade, D., Lachenbruch, P. A., Whaley, F. S., McClish, D. K. and Haley, R. W. 'Effects of misclassifications on statistical inferences in epidemiology', *American Journal of Epidemiology*, **111**, 503–515 (1980).
5. Walker, A. M. and Blettner, M. 'Comparing imperfect measures of exposure', *American Journal of Epidemiology*, **121**, 783–790 (1985).
6. McKeown-Eyssen, G. E. and Thomas, D. C. 'Sample size determination in case-control studies, the influence of distribution of exposure', *Journal of Chronic Disease*, **38**, 559–568 (1985).
7. Kupper, L. 'Effects of the use of unreliable surrogate variables on the validity of epidemiologic research studies', *American Journal of Epidemiology*, **120**, 643–648 (1984).
8. Greenland, S. 'The effect of misclassification in the presence of covariates', *American Journal of Epidemiology*, **112**, 564–569 (1980).
9. Copeland, K. T., Checkoway, H., McMichael, A. J. and Holbrook, R. H. 'Bias due to misclassification in the estimation of relative risk', *American Journal of Epidemiology*, **105**, 488–495 (1977).
10. Barron, B. A. 'The effects of misclassification on the estimation on relative risk', *Biometrics*, **3**, 414–418 (1977).

11. Carroll, R. J., Spiegelman, C. H., Gordon, K. K., Bailey, K. T. and Abbott, R. D. 'On errors-in-variables for binary regression models', *Biometrika*, **71**, 19–25 (1984).
12. Stefanski, L. A. and Carroll, R. J. 'Covariate measurement error in logistic regression', *Annals of Statistics*, **13**, 1335–1351 (1985).
13. Armstrong, B. 'Measurement error in the generalized linear model', *Communications in Statistics–Simulation and Computation*, **14**, 529–544 (1985).
14. Kaldor, J. and Clayton, D. 'Latent class analysis in chronic disease epidemiology', *Statistics in Medicine*, **4**, 327–335 (1985).
15. Prentice, R. L. 'Covariate measurement errors and parameter estimation in a failure time regression model', *Biometrika*, **69**, 331–342 (1982).
16. Van Staveren, W. A., Deurenberg, P., Katan, M. B., Burema, J., de Groot, L. O. and Hoffmans, M. D. 'Validity of the fatty acid composition of subcutaneous fat tissue microbiopsies as an estimate of the long-term average fatty acid composition of the diet of separate individuals', *American Journal of Epidemiology*, **123**, 455–463 (1986).
17. Rosner, B. and Willett, W. C. 'Interval estimates for corrected correlation coefficients corrected for within-person variation: implications for study design and hypothesis testing', *American Journal of Epidemiology*, **127**, 377–386 (1988).
18. Beaton, G. H., Milner, J., Corey, P., McGuire, V., Cousins, M., Stewart, E., de Ramos, M., Hewitt, D., Grambsch, P. V., Kassim, N. and Little, J. A. 'Sources of variance in 24-hour dietary recall data: implications for nutrition study design and interpretation', *American Journal of Clinical Nutrition*, **32**, 2546–2549 (1979).
19. Willett, W. C., Sampson, L., Stampfer, M. J., Rosner, B., Bain, C., Witschi, J., Hennekew, C. H. and Speizer, F. E. 'Reproducibility and validity of a semi-quantitative food frequency questionnaire', *American Journal of Epidemiology*, **122**, 51–65 (1985).
20. Nunnally, J. C. *Psychometric Theory*, second edition, McGraw-Hill, New York, 1978, p. 197.
21. Revicki, D. and Israel, R. G. 'Relationship between body mass indices and measures of body adiposity', *American Journal of Public Health*, **76**, 992–994 (1986).
22. Rosner B., Hennenkens, C. H., Kass, E. H., and Miale, W. E. 'Age-specific correlation analysis of longitudinal blood pressure data', *American Journal of Epidemiology*, **106**, 306–313 (1977).
23. Rothman, K. J. 'A show of confidence', *New England Journal of Medicine*, **299**, 1362–1363 (1978).