Proceedings of the Nutrition Society (1994), **53**, 37–42

Measurement error: effects and remedies in nutritional epidemiology

BY DAVID CLAYTON
MRC Biostatistics Unit, Cambridge

Nutritional epidemiology is concerned to elucidate the relationship between intakes of specific foods and nutrients, and specified health outcomes. Usually the outcome of interest is the incidence of a disease. Typically epidemiological evidence for such a relationship exists at two levels: (1) the macro level, in which each data-point refers to an aggregation of subjects for example, a country, town or small area; (2) the micro level, in which the relationship is observed at the level of the individual subject. The ultimate challenge is the resolution of these two levels of evidence so that the observed differences in disease patterns between different communities can be fully explained in terms of relationships demonstrated at the level of individual subjects. That this is a difficult task is due in no small measure to the problem of measurement error; we are unable to obtain perfectly accurate assessments of dietary intakes either for individuals or for communities.

MEASUREMENT ERROR IN MACRO-EPIDEMIOLOGY

An example of the macro level of epidemiological evidence is the relationship between breast cancer incidence and dietary fat. Prentice & Sheppard (1990) reviewed the evidence obtained from aggregated data-points defined both geographically and by time-period. The problems of such studies are well known (for example, see Greenland, 1992). They are primarily: the poor quality of dietary data (often based on 'disappearance' data); non-measurement of important confounders; the use of inappropriate summary measures of population intake (for example, the mean). The last of these points refers to the case where the dose–response relationship is markedly non-linear and is probably less serious than those which precede it. To some extent at least, the first two failings of 'ecological' studies can be offset by enriching them with nested sample surveys in which data enhanced in both quality and quantity are collected in representative samples. I shall call such studies calibration studies.

Calibration studies could be carried out within a setting in which the primary data collection is by a routine information system such as cancer registration. Alternatively they may be carried out as part of the data collection for special purpose cohort studies. Perhaps the most familiar example of this latter situation is Keys' (1980) famous studies of coronary heart disease in seven countries, in which fifteen cohorts were followed up and detailed dietary studies were carried out in subsamples.

My current interest in the problem of calibration stems from the European Prospective Investigation into Cancer and Nutrition (EPIC) studies (Riboli, 1992), a collaborative series of cohort studies of diet and cancer to be carried out across the European Community. Unlike Keys' (1980) study, these studies aim to investigate the relationship at both macro and micro level. Nutritional data will be available for individuals, thus allowing analysis of disease rates in relation to nutritional status within the cohorts as well as between them. Calibration studies are necessary because, owing to logistic

considerations, several different main study methods for dietary assessment are in use in the different cohorts. It is proposed that a common calibration method is employed in subsamples drawn from these cohorts so as to correct for any bias in the macro-level analysis due to between-method differences. A more ambitious aim is the resolution of any conflict of between- and within-cohort evidence. This will require some attempt to confront the difficult problem of regression dilution. This is, as we shall see, a closely related problem.

REGRESSION DILUTION IN MICRO-EPIDEMIOLOGICAL STUDIES

Regression dilution is the name given to the attenuation of the dose–response relationship which occurs when the exposure dose (here nutritional intake) is measured with error. It arises because some of the variability of measured intakes is due to errors of measurement and will not be reflected in the risk gradient; a group defined by having the highest recorded intake of a given food or nutrient will contain disproportionately more subjects with over-recorded intakes while the group with the lowest recorded intakes will contain disproportionately more subjects with under-recorded intakes. Thus, the true difference in intakes between these groups will be less than it appears and, as a result, the dose–response relationship will be wrongly estimated.

Correction of this effect may be seen as a problem in calibration in which a second measurement in a subset of subjects is used to provide a more accurate estimate of the true difference of intakes of groups defined, as described previously, by intakes recorded in the main study. For example, when investigating the relationship between blood pressure and stroke, Macmahon *et al.* (1990) used a 'non-parametric' adjustment which involved grouping subjects into six bands according to one measure of blood pressure but using a second measure of blood pressure taken 2 years later in subsamples to provide an improved estimate of the long-term average value of blood pressure in these six groups. The relationship between blood pressure and these calibrated values provides an improved estimate of the true relationship between the habitual blood pressure and the risk of stroke. A closely related method has been proposed by Rosner *et al.* (1989) in which a calibration curve is obtained by regression analysis of a calibration method *v.* the main study method. This line may again be obtained from a calibration study carried out in a subsample.

It is important to stress that such analyses rely on relatively strong assumptions. Clearly it is essential that a calibration method be unbiased and this requirement holds whether we are calibrating between-cohort or within-cohort relationships. However, when within-cohort calibration is attempted another assumption is necessary, namely that errors of measurement in the main study and calibration study methods are unrelated to each other. This is evident from thinking about the method of Macmahon *et al.* (1990); the groups formed by an initial stratification on the main study will not be calibrated by a second measurement which makes exactly the same mistakes as the first.

These assumptions seem very strong in the context of nutritional epidemiology. Indeed, in the absence of accurate biomarkers it may be very difficult to check their validity. Such biomarker data have typically identified serious bias in questionnaire measurements and, more recently, Plummer & Clayton (1993*a*,*b*) used covariance structure models to show that biomarker data throw serious doubt onto the assumption of independence of errors of different questionnaire methods (24 h recalls, food-

frequency questionnaires, and diary records). However, biomarkers themselves perhaps provide the ideal method for calibration of large-scale studies. As they become available for more aspects of diet, the design of calibration studies and their incorporation into overall analysis will become of increasing importance.

To these required statistical properties of a method suitable for calibration must be added the very important condition that a very high rate of compliance is essential in calibration studies. Otherwise any potential benefit will be more than offset by 'volunteer bias'. Against this must be set the fact that calibration methods need not have particularly high reliability since any lack of reproducibility of the measurement may be offset by increased sample size in the calibration substudy. This point will be amplified in the next section. It again points to a strong role for biomarkers, which often reflect time average intakes over a relatively short time window. The inherent lack of reliability of such measurements may make them unattractive as a main study method, but their relative freedom from bias and from correlation of errors makes them very attractive for calibration measurements.

DESIGN OF BETWEEN-COHORT CALIBRATION STUDIES

Once a calibration method has been selected, the remaining design decisions for calibration studies concern sample size and stratification of the subsample in which it is to be carried out. These matters have been discussed by Plummer *et al.* (1994).

An essential point to emphasize is that calibration studies cannot increase the power of the main investigation(s). This can only be done by improved methodology in these main studies. Indeed, the effect of calibration studies is to introduce a new source of random error into estimates of the relationship of interest, namely that due to errors incurred in the calibration process itself. We accept these in order to correct for bias in the uncalibrated studies, and aim to design them in such a way that the increase in random error is kept within acceptable bounds.

The contribution of calibration error is illustrated in Fig. 1. This indicates a linear relationship between (log) disease rate and mean intake observed at the macro (between-cohort) level. Fig. 1 shows a measurement for a cohort in which the mean intake is known exactly. The deviation, e_1, from the expected regression line is due entirely to Poisson variability of the observed rate. The point after further displacement by a calibration error, e_2, of the mean intake is also shown. It can be seen from Fig. 1 that the joint effect of both types of error is to displace the point from the regression line by the total error:

$$e = e_1 + \beta e_2.$$

The standard deviation of the Poisson error, e_1, may be estimated by

$$\sqrt{1/D},$$

where D is the number of disease events (cases) observed. In the design of cohort studies this component of the error cannot be reduced except by increasing the total size of the cohort or the follow-up time. The standard deviation of the calibration error e_2 is:

$$\frac{\sigma}{\sqrt{N}},$$

https://doi.org/10.1079/PNS19940007 Published online by Cambridge University Press

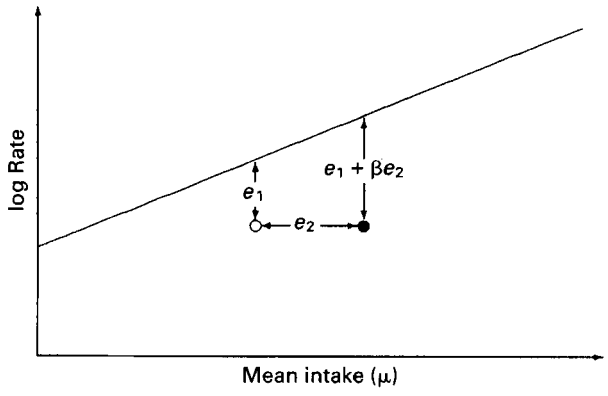

Fig. 1. Errors from a regression line with slope β for a linear relationship between (log) disease rate and mean intake observed at the macro (between-cohort) level. (○), A measurement for a cohort in which the mean intake is known exactly. The deviation, e_1, from the expected regression line is due entirely to Poisson variability of the observed rate. (●), The point after further displacement by a calibration error, e_2, of the mean intake.

where σ is the standard deviation of the calibration measurements and N is the size of the calibration subsample. The standard error of the total error is, therefore:

$$\sqrt{\frac{1}{D} + \frac{(\beta\sigma)^2}{N}}.$$

The second term within the square root sign represents the contribution of calibration error. For the purpose of significance testing for the existence of a relationship we are interested in the precision of estimation around the null hypothesis $\beta = 0$ and the calibration error is irrelevant. However, for estimating the extent of such a relationship the calibration errors become progressively more serious with increased strength of relationship. This simple expression shows clearly that: (1) the size of the calibration study for each data-point should be related to the number of cases of disease which will be observed; a point based on few cases will be very imprecise in the y-direction and expensive calibration to accurately locate it in the x-direction is pointless; (2) lack of reliability in the calibration measurement will result in a large value for σ, but this can be offset by using a larger calibration sample size, N.

Plummer *et al.* (1994) took the breast cancer–fat hypothesis as an example to investigate the implications of these statistical considerations for study design. With plausible assumptions and aiming for at most a 10% reduction in the precision of a study, they showed that a calibration study using a perfect measuring instrument needs a sample size of only about half the number of cases of disease expected to be observed throughout the study. However, the required sample size is six to seven times the number of cases if the calibration measurement correlates with true long-term intake with correlation coefficient about r 0·25 and soars to more than forty times the number of cases if r 0·1. Somewhere between these latter two scenarios would seem to be a realistic estimate of the sample size likely to be required in practice.

The design of calibration studies must also take account of the fact that the analysis

https://doi.org/10.1079/PNS19940007 Published online by Cambridge University Press

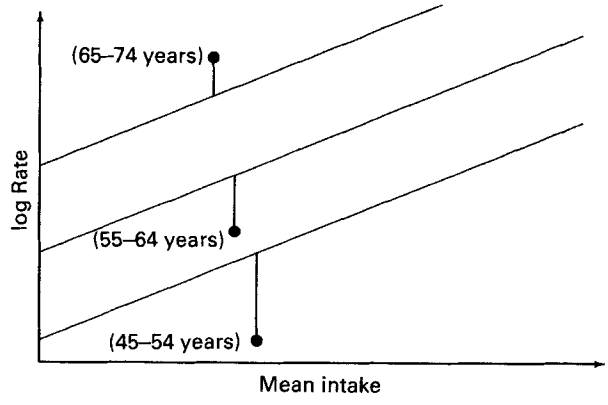

Fig. 2. Stratification by age. The points represent observed rates and estimated intakes for each age-group in one cohort.

must allow for important confounding factors such as age and sex. The analysis at cohort level is now as illustrated in Fig. 2. The points represent observed rates and estimated mean intakes for each age–sex group in one cohort, and the lines represent the relationship between cohort rates and mean intakes within age–sex categories. The considerations of the previous section continue to apply, now within age–sex groups. The aim of calibration studies is to remove systematic error in the x-coordinate of each data-point by using a substudy, but random errors of calibration will have the effect of reducing the precision of each point. The sample size for calibration substudies should be related to the number of events on which the corresponding rate estimates will be based. As before, it is not efficient to spend a lot of time and money on every calibration of the x-coordinate when the y-coordinate is subject to considerable Poisson error.

Plummer *et al.* (1993) show that, with the recruitment age range proposed in the EPIC studies (Riboli, 1992), nearly half the total number of cases of colon cancer which will occur in 20 years follow-up would have been in subjects over 65 years old at recruitment and only 5% of the cases would have been less than 45 years old at recruitment. Thus, if colon cancer were the only end-point of interest, the calibration sample should be very heavily weighted towards the older members of the cohort, since it is they who will provide most of the important outcome data. The position is not, of course, quite so extreme if it is breast cancer which is of primary interest.

The last section drew attention to the similarity between calibration of between-cohort comparisons and 'correction' for regression dilution, which is also a form of calibration. The same simple mathematics applies, but σ now represents the residual standard deviation of the regression of the calibration method on the main study method. The consequences of this mathematics for study design are identical.

DISCUSSION

The present paper has indicated very informally the ways in which calibration substudies can in principle be used to correct for biases occurring as a result of imperfections in main study methodology. The required properties of calibration measurements and the requirements for efficient design of calibration studies have been indicated.

https://doi.org/10.1079/PNS19940007 Published online by Cambridge University Press

However, many problems have been swept under the carpet! First and foremost is the real doubt whether good calibration methods exist for more than a few aspects of diet. The effect of using poor calibration methods may be to do more harm than good. Another difficulty concerns the assumption that it is the mean or expected value of the true intake which is the relevant determinant of disease rate of a group of subjects. If dose–response relationships are markedly non-linear, the variance of true intakes may also be relevant and Plummer & Clayton (1993*a*,*b*) have shown this to be very much more difficult to estimate. Finally, this discussion has assumed that only a single aspect of diet is relevant to disease risk. Application of these ideas to such difficult problems as the adjustment of the effect of fat intake for the confounding effect of total energy intake involves extension of these ideas to multivariate calibration. Extension of the linear regression approach to calibration to the multivariate case is discussed by Rosner *et al.* (1990), but it is not clear how plausible are the necessary assumptions when using currently available measuring instruments.

REFERENCES

Greenland, S. (1992). Divergent biases in ecological and individual-level studies. *Statistics in Medicine* **11,** 1209–1223.

Keys, A. (1980). *Seven Countries: A Multivariate Analysis of Death and Coronary Heart Disease.* Cambridge, Massachusetts: Harvard University Press.

Macmahon, S., Peto, R., Cutler, J., Collins, R., Sorlie, P., Nechan, J., Abbott, R., Godwin, J., Dyer, A. & Stamler, J. (1990). Blood pressure, stroke, and coronary heart disease. *Lancet* **335,** 765–774.

Plummer, M. & Clayton, D. (1993*a*). Measurement error in diet: an investigation using covariance structure models, part I. *Statistics in Medicine* **12,** 925–936.

Plummer, M. & Clayton, D. (1993*b*). Measurement error in diet: an investigation using covariance structure models, part II. *Statistics in Medicine* **12,** 937–948.

Plummer, M., Clayton, D. & Kaaks, R. (1994). Calibration in multi-centre cohort studies. *International Journal of Epidemiology* (In the Press).

Prentice, R. & Sheppard, L. (1990). Dietary fat and cancer: consistency of the epidemiologic data, and disease prevention that may follow from a practical reduction in fat consumption. *Cancer Causes and Control* **1,** 81–97.

Riboli, E. (1992). Nutrition and cancer: background and rationale of the European prospective investigation into cancer and nutrition (EPIC). *Annals of Oncology* **3,** 783–791.

Rosner, B., Spiegelman, D. & Willett, W. C. (1990). Correction of logistic regression relative risk estimates and confidence intervals for measurement error: the case of multiple covariates measured with error. *American Journal of Epidemiology* **132,** 734–745.

Rosner, B., Willett, W. & Spiegelman, D. (1989). Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. *Statistics in Medicine* **8,** 1051–1069.

Printed in Great Britain