Estimation of Population Exposure in Ecological Studies
Author(s): Martyn Plummer and David Clayton
Source: *Journal of the Royal Statistical Society. Series B (Methodological)*, Vol. 58, No. 1 (1996), pp. 113-126
Published by: Wiley for the Royal Statistical Society
Stable URL: https://www.jstor.org/stable/2346168
Accessed: 01-06-2020 23:08 UTC

J. R. Statist. Soc. B (1996)
58, *No.* 1, *pp.* 113–126

Estimation of Population Exposure in Ecological Studies

By MARTYN PLUMMER† and DAVID CLAYTON

Medical Research Council Biostatistics Unit, Cambridge, UK

[*Read before* The Royal Statistical Society *at a meeting on 'Statistical aspects of design' organized by the* Research Section *on Wednesday, April 12th, 1995,* Professor V. S. Isham *in the Chair*]

SUMMARY

This paper discusses design issues in 'ecological studies'—epidemiological studies in which the relationship between disease and behavioural and environmental determinants is studied at the population rather than the individual level. The number of study populations has little relevance beyond a certain point, the power and precision being limited by the total number of disease events and by the size of the sample surveys used to estimate the distributions of determinants within populations. In most circumstances, optimal design requires the size of the sample surveys in each population to be related to the number of disease events which will occur in it, and for sampling to be stratified by age and/or sex.

Keywords: AGGREGATE DATA; CONFOUNDING; EPIDEMIOLOGY; MULTIPLE POPULATIONS; OBSERVATIONAL STUDIES; SAMPLE SURVEY

1. INTRODUCTION

From its earliest beginnings, the science of epidemiology has drawn on evidence at two different levels:

(a) aggregated measurements of disease experience in population groups, usually geographically defined, and
(b) measurements made at the level of the individual subject.

To a large extent these different levels of measurement define the two main branches of the modern discipline, *descriptive epidemiology* and *analytical epidemiology*, but to identify these splits too closely would be mistaken as it would neglect the role of aggregate level evidence in suggesting and helping to establish causal relationships between environment, behaviour and disease. Although much of the early history of the subject from Graunt to Farr was concerned with the development of better data sources and of statistical methods for characterizing the health of populations, establishing correlations with potential causes was always a main concern. Attempts to establish such correlations were not always informal. For example, William Farr carried out an extremely detailed study of the relationship between cholera mortality in the London boroughs and their height above sea-level—an investigation which involved detailed curve fitting (Farr, 1852).

In the current century, the use of 'ecological' evidence, as it has come to be called, has continued to make an important contribution. A notable success is the establishment of strong relationships between atmospheric smoke pollution and mortality from chronic bronchitis, work which was most influential in the eventual

†*Address for correspondence*: International Agency for Research on Cancer, 150 cours Albert Thomas, 69372 Lyon Cedex 08, France.
E-mail: plummer@iarc.fr

 0035–9246/96/58113

passing of smoke control legislation (Morris, 1975). More recent continuation of this work seeks to relate the more modern problem of pollution from car exhausts to a range of health problems, from mental retardation to asthma. In establishing reasons for the 'modern epidemic' of coronary heart disease (CHD), two pieces of ecological evidence have been important. The relationship between the level of physical activity in occupations and mortality from CHD was crucial to the development of the exercise hypothesis (Morris, 1975), and the geographical variations in CHD mortality from northern Europe and the USA on the one hand to the Mediterranean basin on the other gave rise to the hypothesis that some aspect of the fat content of the diet is important.

Ecological evidence has a particular place in nutritional epidemiology. It has been estimated that dietary factors may account for as much as 70% of cancers currently occurring in the USA (Doll and Peto, 1981), but the establishment of relationships at the level of the individual has remained difficult. This difficulty must be due in large part to the twin problems of the relative *homogeneity* of many aspects of diet within cultures and the serious *measurement errors* which attend all methods of measuring diets in free-living populations. In contrast, the diversity of diets *between* different cultures offers considerable evidence at the ecological level.

Despite its past usefulness and the current opportunities which it offers, ecological evidence has fallen into disfavour with the attempts which have been made during the last 25 years to establish epidemiology as a more rigorous scientific discipline. Indeed, in one prominent text-book of modern epidemiology (Rothman, 1986), ecological evidence is listed in the index only under the heading *ecological fallacy*. This term draws attention to the fact, first noted by Robinson (1950), that relationships observed at the ecological level cannot be simply equated with relationships observed in individuals and, it is argued, it is the latter which are the concern of epidemiology. The difference between the relationships at the two levels follows in part from the simple mathematical results which will be discussed in the next section. Perhaps a more serious problem, however, is the quality of much ecological evidence. Specifically,

(a) many of the measurements used to assess the potential causes of disease (which we call *exposures*) in populations are crude and based on inappropriate sampling frames, and
(b) there may be the possibility of *confounding* by unmeasured causal factors whose distributions differ between communities.

While recognizing these criticisms, a role for ecological studies remains (Schwartz, 1994; Susser, 1994a, b). In particular it may be that the only 'natural experiments' available for study lead to variation of exposure between communities, but little variation within communities. In such circumstances ecological evidence could give a truer picture of the causal relationship than studies at the individual level, although such situations will inevitably invite controversy. For example, although evidence relating the incidence of breast cancer to dietary fat consumption is lacking at the individual level, Prentice and Sheppard (1990) have pointed out that there is some consistency of the ecological evidence.

A constructive response to the admitted deficiencies of ecological studies is to seek to improve them, both by extending the measurements made on populations to exclude unmeasured confounding and in improving the quality of population

exposure measurements. However, the design of studies which seek to obtain evidence at this level has received little discussion in the epidemiological literature, and some new problems must be faced. In particular, whereas classical ecological studies are based on very large scale data sources such as national mortality registration, the census, etc., an improvement of the quality of ecological evidence requires

(a) the study of smaller, more homogeneous, population groups and
(b) better estimation of a wider range of disease determinants via special purpose sample surveys.

These improvements necessarily come at the expense of sampling errors which are negligible in studies based on routine sources. This paper is concerned with the design issues raised.

Similar problems arise in the design of studies which combine evidence at both the individual and the ecological level. An important example is the 'seven countries' studies into the causes of CHD (Keys, 1980). International comparisons were of limited usefulness because of the lack of routine disease *incidence* data, which describe the number of new cases of disease in the population in a specified time interval, the variable quality of death certification and the poor quality of data concerning national diets. There is also a strong possibility of confounding. The seven countries study tackled these problems by recruiting 16 cohorts from regions with widely varying CHD mortality. All established risk factors were measured at the individual level, but the diets of the 16 groups were established by surveys in subsamples. Thus, the evidence concerning a diet–disease relationship was at the ecological level but was based on the best available dietary measurement methods and, in principle at least, it should be possible to exclude the possibility of confounding by other known risk factors.

This study was influential in the genesis of the British regional heart study (Shaper *et al.*, 1982), set up to address the observation of large differences in CHD mortality between British towns in earlier ecological studies of water hardness and mortality (Gardner, 1973). Again there is a strong possibility that the CHD–water hardness relationship is distorted by differences in death certification, by difficulties of measuring population exposure to the constituents of drinking water and by unmeasured confounding. In common with the seven countries study, the regional heart study combined aspects of ecological and subject-based research strategies.

Similar issues involving the resolution of ecological and individual evidence are an increasing feature of modern multicentre studies. A recent example is the INTERSALT study into the relationship of mineral intakes (as measured by 24-hour excretion) and blood pressure (Elliott, 1992), and our interest in this problem stems from the European prospective investigation into cancer (EPIC) studies of diet and cancer (Riboli, 1992), in which the diets of a series of cohorts from throughout Europe will be related to subsequent cancer incidence.

In the four studies described above, the aggregated study unit is an identified *cohort*, recruited explicitly for research, and much of the evidence generated is at the individual level. In part the need to recruit cohorts and to follow them up is driven by the lack of population data concerning incidence (rather than mortality) of such diseases as CHD. However, for cancer, a disease which has high quality registration in many regions of the world, there is the possibility of designed ecological studies in

which population exposure is assessed by sample survey methods and set alongside reliable cancer registration statistics. Two recent examples are the wide-ranging studies of cancer in China (Chen *et al.*, 1990) and the study reported by Peers *et al.* (1987) into the relationship between aflatoxin exposure, hepatitis B infection and liver cancer. In this last study, data were assembled for 10 regions of Swaziland by using three main sources—analysis of crop samples (for aflatoxin), analysis of serum samples from blood donors (for hepatitis B infection) and from cancer registration. The statistical analysis of such studies has recently been discussed by Prentice and Sheppard (1995), but the implications for study design are not clear from that work. In particular,

(a) how large should sample surveys of population exposure be and
(b) how should they be targeted on different sections of the study population?

This paper addresses these questions. Similar questions have arisen in the design of the EPIC studies (Plummer *et al.*, 1994). Although diet has been measured at the individual level in these studies so that, unlike the seven countries study, it will be possible to study diet–disease relationships at the individual level, different dietary survey methods have been used in different cohorts. To recapture the ecological evidence, the study will incorporate *calibration* studies which will measure diet in subsamples of each cohort, using a standardized 24-hour recall method.

2. THE ECOLOGICAL PROBLEM

Suppose that we are analysing the effect of a single binary risk factor on a given disease, and the subjects can be divided naturally into *clusters*. These clusters may be different geographical areas, or different time periods in the same area. If all the subject-level data were available, it could be summarized in a *cluster* × *exposure* × *disease* contingency table. An ecological study is defined by lack of knowledge of the complete data. Only the cluster × exposure margin and the cluster × disease margin are available—in epidemiological terms, the disease rate and the exposure prevalence in each cluster—and the problem is to infer the exposure–disease relationship from these marginal data.

If the disease rate in cluster i is π_i in unexposed subjects and $\theta\pi_i$ in exposed subjects, and if the exposure prevalence is p_i, then the marginal disease rate in cluster i is

$$\lambda_i = \pi_i + p_i\pi_i(\theta - 1).$$

We assume that p_i varies between clusters. The model is identifiable when π_i is constant across all clusters, and in this case there is a simple linear relationship between disease rate and exposure prevalence. This model is represented graphically in Fig. 1. In a graphical model, variables are represented by nodes on a graph and the relationships between them by lines. Specifically, there is no line directly connecting two variables if they are conditionally independent, given all other variables in the graph.

In Fig. 1 there is no relationship between cluster and disease once exposure is known. This can be interpreted as an absence of unmeasured confounding factors. One reason why ecological studies are unpopular is the general assumption in the

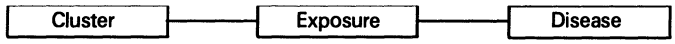

Fig. 1. Conditional independence model for an ecological study

epidemiological literature that unmeasured confounders will always exist (Piantadosi *et al.*, 1988). The serious consequences of this assumption have been explored by Greenland and Morgenstern (1989), Greenland (1992) and Greenland and Robins (1994). The argument may be summarized as follows.

Firstly, increased heterogeneity of exposure, which is one of the motivations behind ecological studies, may bring with it increased heterogeneity of confounding variables. So a confounder which is tightly controlled within clusters may vary widely between clusters and cause confounding when it is ignored in an ecological analysis but not when it is ignored in an individual level analysis.

More generally, a confounder is ignorable at the individual level if its distribution is independent of the distribution of exposure. This is also true for aggregate data if the model which relates exposure to disease outcome is linear. This favourable case is emphasized by Cohen (1994). However, the interpretation of independence of exposure and confounder is different for the two levels of evidence. A confounder may be uncorrelated with exposure within clusters but correlated across clusters or vice versa. So an absence of confounding at the individual level does not guarantee absence of confounding in ecological studies.

When we consider non-linear dose–response models, it is no longer true that independence of exposure and confounder guarantees absence of bias at the aggregate level. This important case includes linear models with interaction terms (referred to as *effect modification* by epidemiologists). So, if there is an unmeasured effect modifier with a distribution that is independent of the exposure of interest, then it may be safely ignored in individual level analyses but may cause bias in ecological studies.

These arguments suggest that, except in the simplest cases, any ecological relationship may be explained away as the effect of unmeasured confounders if one assumes that such unmeasured confounders will always exist. According to this argument, ecological studies may be regarded at best as hypothesis-generating studies and at worst as totally useless. Although this argument is effective against ecological studies which ignore known or suspected confounders, such as age or smoking, we do not believe that it justifies a wholesale rejection of ecological studies. Instead, the design of ecological studies should be improved by collecting data on all potential confounders and these data should be used in the analysis.

The model is extended to include a confounding factor in Fig. 2. Now the relationship between cluster and disease outcome is explained by two variables — the exposure of interest and the confounder.

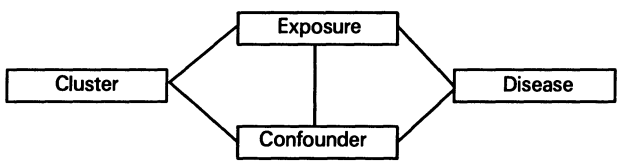

Fig. 2. Conditional independence model including a confounder

The complete data can be summarized in a *cluster* × *exposure* × *confounder* × *disease* contingency table. The disease odds ratios for exposure and confounder, and their interaction, can be estimated by combining information from disease registration sources, which give the cluster × disease margin (or, possibly, the cluster × confounder × disease margin), and information from other sources about the cluster × exposure × confounder margin. Unfortunately, this latter information is often not available from aggregate data sources, which give only the prevalences of each exposure and confounder by cluster. This is insufficient — we also need to know the relationships between exposures and confounders within clusters, and this will not be available when the data are drawn from diverse sources. This information can only be gained by drawing a random sample from each cluster and measuring the exposure and confounder status of all subjects in these samples.

A general model, which includes both categorical and continuous risk factors, is the log-linear model, in which the disease rate is $\exp(\alpha + \beta^{\mathrm{T}} x)$ where x is a vector of covariates. The marginal disease rate in a given cluster is then

$$E\{\exp(\alpha + \beta^{\mathrm{T}} X)\} = \exp\{\alpha + K(\beta)\} \tag{1}$$

where K is the cumulant-generating function of X (Richardson *et al.*, 1987). If precise covariate measurements are available then we can calculate the empirical estimate of the disease rate by using the observed values of x in the subsample (Prentice and Sheppard, 1995). If precise measurements are not available, however, then we need an approximate model for the disease rate which is unaffected by measurement errors. Such a model will be derived, for continuous exposures, in the next section.

We shall assume that information about the distribution of X in cluster i is available from a subsample of size N_i. The important design questions are as follows.

(a) How big should N_i be?
(b) For some confounders, notably age and sex, the disease registration scheme gives us the disease × cluster × confounder margin. How does this affect the sampling strategy?

3. CONTINUOUS EXPOSURES

If we assume that X is normally distributed within clusters and the mean and variance within cluster i are μ_i and Ω_i respectively, then the disease rate in cluster i is, from equation (1),

$$\lambda_i = \exp(\alpha + \beta^{\mathrm{T}} \mu_i + \Omega_i \beta/2).$$

As in the binary risk factor case, the marginal rate depends not only on the mean exposure but also on the association between exposure variables within the cluster, which in this case is determined by the covariance matrix. This expression for λ_i can be justified, even when exposure is not normally distributed, by expanding the cumulant-generating function as far as the quadratic term. In principle, we could include further terms in the expansion, involving the skewness, kurtosis and so on. This would decrease the bias of the estimates but increase the variance. Since higher order moments are extremely difficult to estimate, it seems plausible that the result would be a net increase in mean-square error.

The analysis is considerably simplified if the quadratic term $\beta^T \Omega_i \beta$ can be ignored. The ecological relationship between mean exposure μ_i and disease rate is then log-linear with the same vector of slope parameters as the subject level relationship between exposure X and disease rate. This simplification is possible when

(a) β is very small, or
(b) Ω_i is very small compared with the between-cluster variance of exposure or
(c) Ω_i is constant across clusters; the quadratic term can then be absorbed into the intercept.

For the present, we shall assume that one of these conditions holds.

Let D_i be the number of cases of disease in cluster i for the whole cluster (not the subsample). We assume that this has a Poisson distribution with mean $\lambda_i Y_i$, where Y_i is the total person-years at risk in cluster i. Instead of working directly with D_i we shall work with an empirical log-transformation and use the approximations

$$E\{\log(D_i + \tfrac{1}{2})\} \approx \log(\lambda_i Y_i), \tag{2}$$

$$\operatorname{var}\{\log(D_i + \tfrac{1}{2})\} \approx E\left(\frac{1}{D_i + \frac{1}{2}}\right) \approx \frac{1}{\lambda_i Y_i - \frac{1}{2}} \tag{3}$$

which hold when the expected value of D_i is at least 5. The use of the empirical log-transformation leads to a considerable simplification. Although it is not adequate when there is a large number of clusters, each with a very small number of disease events, it should suffice for most real situations.

The phenomenon of *extra-Poisson variation*—the variation of the disease rate between clusters over that given by equation (3)—has important implications for the interpretation of ecological studies. The presence of a large amount of extra-Poisson variation indicates unmeasured confounding factors which would seriously compromise the findings of the study. Hence it is not enough to show a correlation between disease rate and exposure. We must also ensure that the amount of extra-Poisson variation is small. Any analysis of an ecological study should therefore include, as a diagnostic check, an estimate of the extent of extra-Poisson variation. For this estimate the number of clusters should not be too small—perhaps 10–20 clusters would be sufficient.

3.1. *Precision of Estimation*

The parameters can be estimated by empirically weighted least squares. Let

$$U_i = \log(D_i + \tfrac{1}{2}) - \log Y_i - \alpha - \beta^T \bar{x}_i.$$

Then $E(U_i) = 0$, from approximation (2) assuming that the variance term $\beta^T \Omega_i \beta$ can be ignored. The empirically weighted least squares estimates of α and β are the solutions to the estimating equations

$$\sum_i U_i W_i \begin{pmatrix} 1 \\ \bar{x}_i \end{pmatrix} = 0 \tag{4}$$

where the weighted W_i is given by

$$\frac{1}{W_i} = \frac{1}{D_i + \frac{1}{2}} + \frac{\beta^{\mathrm{T}} S_i \beta}{N_i}$$

and S_i is the sample variance of X within cluster i. The reciprocal of W_i is an estimate of the variance of U_i. It has two components—one due to Poisson variation of the number of cases and the other due to sampling error in the estimation of μ_i (the above expression assumes that N_i is small in comparison with the total population of the cluster). With these weights, the variance matrix of the estimate of β is given approximately by the inverse of

$$\sum_i W_i(\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^{\mathrm{T}},$$

where $\bar{\bar{x}}$ is a weighted mean of the $\bar{x}_i$.

These approximate results must be treated with some caution, because the estimating function on the left-hand side of equation (4) is biased. This bias, which is due to correlation between U_i and $\bar{x}_i$ or $W_i\bar{x}_i$, can be ignored when it is small relative to the variance of the estimating function. If the number of clusters is fixed and $N_i \uparrow \infty$ in all clusters, then the bias tends to 0. However, if we take the other limiting case, in which the number of clusters increases to ∞ and the average sample size in each cluster tends to a finite limit, then the bias is not ignorable, and in particular the estimate of β is not consistent. It follows that the results used here will not hold when there is a large number of clusters each with a small sample. This problem is not limited to our empirical least squares approach, and also afflicts the estimating equation method of Prentice and Sheppard (1995).

Our criterion for choosing the sample size N_i is based on a target for precision of estimation, relative to a study in which all subjects in each cluster are included in the sample. In this limiting situation, the problem reduces to a Poisson regression problem and the variance of the estimate of β is the inverse of $\Sigma_i D_i(\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^{\mathrm{T}}$. A target of 90% efficiency, relative to this ideal case, can be achieved if $W_i = 0.9D_i$. This implies that the sample size is proportional to the expected number of cases in each cluster. The constant of proportionality is determined by $\beta^{\mathrm{T}}\Omega_i\beta$ which is the variance of the log-disease-rate within the cluster. Table 1 shows the required ratio of sample size to number of cases for a single exposure variable, when the target value for W_i is $0.9D_i$ and the rate ratio between the top and bottom quintile groups within a cluster takes values from 1.50 to 3.50, assuming that exposure is normally distributed within the cluster. On the basis of Table 1 it seems that a sample size which is twice the number of cases would be sufficient for most instances.

These results seem counter-intuitive because they suggest that the required sample size increases as the size of the dose–response effect increases, whereas intuition suggests that dose–response effects would be easier to find. This intuition is based on the power of a hypothesis test of no association between exposure and disease while the sample sizes are based on a criterion of relative precision of estimation. In fact Table 1 reflects the equally clear intuitive result that the dose–response effect becomes more difficult to estimate as the clusters become more heterogeneous.

TABLE 1
Ratio of sample size to number of cases required for 90% efficiency, for a given rate ratio between top and bottom quintile groups

Rate ratio	1.50	2.00	2.50	3.00	3.50
Sampling ratio	0.21	0.63	1.07	1.53	2.00

This discussion suggests that a criterion of good power for a test of the hypothesis $\beta = 0$ may give different sample size recommendations from those in Table 1. We shall investigate this alternative criterion in the next section.

3.2. *Power of Hypothesis Test*

The test statistic for testing the hypothesis $\beta = 0$, based on the estimating function on the left-hand side of equation (4), is

$$U = \sum_i Y_i \log\left(\frac{D_i + \frac{1}{2}}{Y_i}\right)(\bar{x}_i - \bar{\bar{x}})$$

where $\bar{\bar{x}}$ is the mean of the $\bar{x}_i$ weighted by person-years at risk. The expectation and variance of U, given $\{\bar{x}_1, \bar{x}_2 \ldots\}$, are approximately

$$E(U|\bar{x}_1, \bar{x}_2 \ldots) = \beta \sum_i Y_i \mu_i (\bar{x}_i - \bar{\bar{x}}),$$

$$\operatorname{var}(U|\bar{x}_1, \bar{x}_2 \ldots) = \sum_i \frac{Y_i}{\lambda_i}(\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^{\mathrm{T}}$$

where we have used the simpler approximation $\log(D_i + \frac{1}{2}) \approx 1/\lambda_i Y_i$ instead of equation (3).

Suppose that x is scalar. Then the power of the test depends on

$$E(U|x)^2/\operatorname{var}(U|x).$$

For small β this can be written

$$\frac{E(U|\bar{x}_1, \bar{x}_2 \ldots)^2}{\operatorname{var}(U|\bar{x}_1, \bar{x}_2 \ldots)} \approx E(D_+) \operatorname{var}(\beta\mu) \operatorname{cor}(x, \mu)^2$$

where $E(D_+) = \Sigma_i Y_i \lambda_i$ is the expected total number of cases, and where $\operatorname{var}(\mu)$ and $\operatorname{cor}(x, \mu)$ are defined by

$$\operatorname{var}(\mu) = \frac{\sum_i Y_i(\mu_i - \bar{\mu})^2}{\sum_i Y_i},$$

$$\text{cor}(x, \mu) = \frac{\sum_i Y_i \bar{x}_i(\mu_i - \bar{\mu})}{\left\{\sum_i Y_i(\bar{x}_i - \bar{\bar{x}})^2 \sum_j Y_j(\mu_j - \bar{\mu})^2\right\}^{1/2}}.$$

var $(\beta\mu)$ can be interpreted as the variance of the log-disease-rate across clusters and cor (μ, x) as the correlation between $\bar{x}_i$ and μ_i across clusters. Hence the power of the test depends on the total number of disease events, the heterogeneity of the disease rate between clusters and the quality of the mean exposure measurements.

So far we have argued conditionally on the observed value of $\bar{x}_i$. But, since $\bar{x}_i$ is a random variable with variance depending on the sample size N_i, the expected value of cor (μ, x) depends on the sample size in each cluster. If equal numbers are sampled in each cluster, so that $N_i = N$ for all i, then the expected correlation between $\bar{x}$ and μ can be written

$$\left(1 + \frac{\sum_i Y_i \Omega_i}{N \sum_i Y_i(\mu_i - \bar{\mu})^2}\right)^{-1/2}.$$

The correlation is high if subjects within each cluster have homogeneous exposures, relative to between-cluster differences, and low if subjects within each cluster are relatively heterogeneous.

To achieve a target value for the correlation, N should be a multiple of the ratio of within-cluster to between-cluster variance

$$\sum_i Y_i \Omega_i \Big/ \sum_j Y_j(\mu_i - \bar{\mu})^2.$$

Table 2 shows the required sample size, expressed as a multiple of the variance ratio, for a range of target values for the correlation. Very little power is gained by increasing the sample size beyond four times the variance ratio. It follows that modest sample sizes will give satisfactory power unless the ratio of between-cluster to within-cluster variance is small. In this situation, most of the information about the exposure–disease relationship comes from within-cluster comparisons and it might be better to conduct subject level studies in each cluster than to attempt an ecological analysis.

This argument does not apply to the case in which X is a vector which includes confounding factors as well as the exposure of interest. Suppose that we have a single

TABLE 2

Sample size required to achieve a given target for the correlation between $\bar{x}$ and μ, expressed as a multiple of between-cluster to within-cluster variance of exposure

Correlation	0.5	0.6	0.7	0.8	0.9	0.95	1.0
Sample size:variance ratio	0.33	0.56	0.96	1.78	4.26	9.25	∞

confounding variable and the log-rate ratio is β_e for a unit increase in exposure and β_c for a unit increase in confounder. The power of the score test in this situation depends on the variance of the estimating function of equation (4) at $\beta_e = 0$ (but $\beta_c \neq 0$). The variance is given approximately by

$$\sum_i W_i(\bar{x}_i - \bar{\bar{x}})(\bar{x}_i - \bar{\bar{x}})^T.$$

So the power of the score test, relative to a study in which N_i is very large, depends on

$$\frac{1}{W_i} = \frac{1}{D_i + \frac{1}{2}} + \frac{\beta_c^2 \omega_i^2}{N_i}$$

where ω_i^2 is the variance of the confounder within cluster i. To ensure that the power of the test is not compromised by a small measurement subsample, we should choose N_i so that W_i/D_i is close to 1. This implies that the sample size should be proportional to the expected number of cases in each cluster. This relative power criterion resembles the relative efficiency criterion of Section 3.1, but the constant of proportionality is determined by the effect of the confounder, rather than the exposure of interest. This result is intuitively appealing because eliminating the effect of the confounder implicitly involves an estimation of β_c.

3.3. *Measurement Error*

So far we have assumed that the vector of exposure variables X can be measured precisely at the individual level. Many exposures of interest are difficult to measure at the subject level but are relatively easy to characterize at the population level. Habitual diet is an important example. International differences in dietary patterns are not difficult to detect, but to rank subjects within a single population in order of intake of a given nutrient is a much more difficult task. Consequently an ecological study which relates aggregate intake measurements to international differences in disease rates may be considerably more powerful than a subject level analysis within any one country.

If the measurement errors are unbiased then, as before, $\bar{x}_i$ is an unbiased estimate of μ_i, and the analysis is unchanged. The only important difference is that the required sample size is increased because the variance of X increases from Ω to $\Omega + \Theta$, where Θ is the variance of the measurement errors. If X is a scalar then, for both the relative efficiency criterion of Section 3.1 and the power criterion of Section 3.2, the sample size must be increased by a factor

$$(\Omega + \Theta)/\Omega$$

to achieve the same target.

The disease rate in cluster i is given by $\lambda_i = \exp(\alpha + \beta^T \mu_i + \beta^T \Omega_i \beta)$ and we have assumed up to now that the quadratic term $\beta^T \Omega_i \beta$ can be ignored. If this is not so and the exposure vector X is measured with error then the analysis becomes much more difficult. As well as using sample surveys to estimate the μ_i, we must also estimate Ω_i and estimation of this variance component is much more demanding both of sample size and of strength of assumptions. The simplest design would be to

take repeat measurements on subjects in the subsample and to assume that errors of measurement are additive and independent. However, our experience with dietary data (Plummer and Clayton, 1993) suggests that such assumptions may not be justified and that more elaborate designs together with covariance structure models may be required. Typically such approaches require very large sample sizes.

4. STRATIFICATION BY AGE

So far we have assumed that covariate information is only available for subjects in the sample selected for exposure measurement. If this is so then we are forced to study the marginal disease rate in each cluster. If, however, information concerning some covariates is available in the disease registration scheme then the analysis can be improved by stratification. For example, if the age of all disease cases is known together with the age distribution of the cluster, then the cases and person-years at risk can be divided into age bands to give estimates of the age-specific disease rates. The sample selected for measurement should also be stratified and comparisons between clusters should be made within age bands. This approach is implied by the graphical model of Fig. 2: within a given age band in which the confounder is held constant, the model reduces to the model in Fig. 1.

An alternative to a stratified analysis would be to use age-standardized disease rates. Rosenbaum and Rubin (1984) have observed that, in a linear model for disease risk, the use of age-standardized rates is not correct unless all the covariates in the model are also age standardized.

Assuming that the effect of exposure is constant across strata, the analysis of Section 3 can be extended to cover the stratified case by fitting a different intercept term in each stratum and adding the estimating functions for β. The sample size calculations now apply *within* strata. So, for efficient sampling, the number of subjects in each stratum of the sample should be related to the expected number of cases within the corresponding population stratum.

Age is a strong confounder for most chronic diseases. Table 3 shows the distribution of cases of colon cancer in males and breast cancer in females from each 10-year age group after 10 years of follow-up, when the age distribution at recruitment is uniform in the range 35–74 years. The incidence of colon cancer increases dramatically with age so most of the colon cancer cases come from the oldest age group. By contrast, the incidence of breast cancer is more stable with age so the age distribution of cases is closer to the uniform distribution.

These results seem to imply that simple random sampling is reasonably efficient for the breast cancer study but inefficient for the colon cancer study, according to the results of Section 3. However, as is often the case, the suboptimal design turns out not to be as bad in practice as might be expected. Table 4 shows the efficiency, relative to a study in which all subjects have exposure measurements, when the measurement sample is drawn by simple random sampling and the sample size is based on an efficiency target for an unstratified analysis. We assume that the analysis is stratified by age even though the sampling mechanism is not. These calculations were made under the simplifying assumption that the distribution of exposure within cluster i and stratum j is $N(\mu_i + \eta_j, \Omega_i)$, so the only difference in the distribution of exposure between strata is a shift in location. In the younger age groups the efficiency is above target and in the oldest age group it is below target. Overall the efficiency is

TABLE 3
Distribution of cases in each age group after 10 years of follow-up

Site	*% for the following age groups:*				
	35–44	*45–54*	*55–64*	*65–74*	*All*
Colon (male)	3	10	26	59	100
Breast (female)	15	23	27	35	100

TABLE 4
Efficiency for the age-stratified study of colon cancer when the sample is drawn by simple random sampling

Target (%)	*Efficiencies (%) for the following age groups:*				
	35–44	*45–54*	*55–64*	*65–74*	*Overall*
90	99	96	93	79	84
70	95	85	72	50	59
50	89	71	51	30	40

about 10% below target. To compensate for the inefficiency of the sampling scheme it would be necessary to take samples about 20% larger, but against this should be weighed the simpler logistics of unstratified sampling.

5. DISCUSSION

In this paper we have discussed the statistical problems involved in the design of sample surveys which, together with suitable disease registration systems, will provide the basis for an ecological study. Under simplifying assumptions of (log-) linearity of disease–covariate relationships, multivariate normality of covariates within clusters and homoscedasticity of covariates between clusters, simple relationships hold which give some guidelines for study design.

Two situations may be identified. In the first we require only to demonstrate a relationship between disease and population exposure, after controlling for variables such as age and sex which are available in the disease registration system. In this case the required sample size in each cluster may be modest being determined only by the ratio of within-cluster to between-cluster variance (see Table 2). In the second situation we require precise estimation of regression coefficients. Perhaps most importantly this includes the case where we need to control for the effects of confounding variables that are not available at disease registration. In such situations, the sample size required in each cluster is a multiple of the expected number of disease events. Ideally, sampling should be stratified by age and/or sex, but the loss of efficiency by not stratifying may be modest except in extreme cases.

The problems of inference from such studies become much more serious if we attempt to relax the simplifying assumptions. In particular it becomes necessary to estimate higher moments of the distribution of covariates within clusters, often in the presence of considerable measurement error. Such estimation must be based on strong assumptions and will be imprecise. We speculate that, in most cases, attempts

to correct the bias introduced by our simplifying assumptions could introduce so much extra variance that the analysis would be worse in a mean-square error sense.

A further issue concerns the design of ecological studies in which covariate measurement is by biochemical analysis of blood or urine (Chen *et al.*, 1990). In this case the estimation of μ_i in the most cost-effective manner may involve *pooling* of samples before biochemical analysis. However, it should be noted that such a strategy further compromises our ability to estimate Ω_i or further moments.

REFERENCES

Chen, J., Campbell, T. C., Li, J. and Peto, R. (1990) *Diet Lifestyle and Mortality in China.* Oxford: Oxford University Press.

Cohen, B. (1994) In defense of ecological studies for testing a linear no-threshold theory. *Am. J. Epidem.*, **139**, 765–768.

Doll, R. and Peto, R. (1981) *The Causes of Cancer.* Oxford: Oxford University Press.

Elliott, P. (1992) Design and analysis of multicentre epidemiological studies: the INTERSALT study. In *Coronary Heart Disease Epidemiology: from Aetiology to Public Health* (eds M. Marmot and P. Elliott), pp. 166–178. Oxford: Oxford University Press.

Farr, W. (1852) Influence of elevation on the fatality of cholera. *J. R. Statist. Soc.*, **15**, 155–183.

Gardner, M. (1973) Using the environment to explain and predict mortality. *J. R. Statist. Soc.* A, **136**, 421–440.

Greenland, S. (1992) Divergent biases in ecologic and individual-level studies. *Statist. Med.*, **11**, 1209–1223.

Greenland, S. and Morgenstern, H. (1989) Ecological bias, confounding and effect modification. *Int. J. Epidem.*, **18**, 269–274.

Greenland, S. and Robins, J. (1994) Ecological studies — biases, misconceptions and counterexamples. *Am. J. Epidem.*, **139**, 747–760.

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

Morris, J. (1975) *Uses of Epidemiology*, 3rd edn. Edinburgh: Churchill Livingstone.

Peers, F., Bosch, X., Kaldor, J., Linsell, A. and Pluijmen, M. (1987) Aflatoxin exposure, Hepatitis B virus infection and liver cancer in Swaziland. *Int. J. Cancer*, **39**, 545–553.

Piantadosi, S., Byar, D. P. and Green, S. B. (1988) The ecological fallacy. *Am. J. Epidem.*, **127**, 893–904.

Plummer, M. and Clayton, D. (1993) Measurement error in diet: an investigation using covariance structure models, part ii. *Statist. Med.*, **12**, 937–948.

Plummer, M., Clayton, D. and Kaaks, R. (1994) Calibration in multi-centre cohort studies. *Int. J. Epidem.*, **23**, 419–426.

Prentice, R. and 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 Control*, **1**, 81–97.

——— (1995) Aggregate data studies of disease risk factors. *Biometrika*, **82**, 113–125.

Riboli, E. (1992) Nutrition and cancer: background and rationale of the european prospective investigation into cancer and nutrition (epic). *Ann. Onc.*, **3**, 783–791.

Richardson, S., Stücker, I. and Hémon, D. (1987) Comparisons of relative risks obtained in ecological and individual studies: some methodological considerations. *Int. J. Epidem.*, **16**, 111–120.

Robinson, W. (1950) Ecological correlations and the behaviour of individuals. *Am. Sociol. Rev.*, **15**, 351–357.

Rosenbaum, P. R. and Rubin, D. B. (1984) Difficulties with regression analyses of age-adjusted rates. *Biometrics*, **40**, 437–443.

Rothman, K. (1986) *Modern Epidemiology.* Boston: Little, Brown.

Schwartz, S. (1994) The fallacy of the ecological fallacy: the potential misuse of a concept and the consequences. *Am. J. Publ. Hlth*, **84**, 819–823.

Shaper, A., Pocock, S., Walker, M., Cohen, N., Wale, C. and Thomson, A. (1982) British regional heart study: cardiovascular risk factors in middle aged men in 24 towns. *Br. Med. J.*, **283**, 179–186.

Susser, M. (1994a) The logic in ecological: I, The logic of analysis. *Am. J. Publ. Hlth*, **84**, 825–829.

——— (1994b) The logic in ecological: II, The logic of design. *Am. J. Publ. Hlth*, **84**, 830–835.