Published in final edited form as:
J R Stat Soc Series B Stat Methodol. 2008 February 1; 70(1): 73–93.

The Combination of Ecological and Case-Control Data

Sebastien J-P.A. Haneuse
Center for Health Studies, Group Health Cooperative, Seattle, WA, USA

Jonathan C. Wakefield
Departments of Biostatistics and Statistics, University of Washington, Seattle, USA

Summary

Ecological studies, in which data are available at the level of the group, rather than at the level of the individual, are susceptible to a range of biases due to their inability to characterize within-group variability in exposures and confounders. In order to overcome these biases, we propose a hybrid design in which ecological data are supplemented with a sample of individual-level case-control data. We develop the likelihood for this design and illustrate its benefits via simulation, both in bias reduction when compared to an ecological study, and in efficiency gains relative to a conventional case-control study. An interesting special case of the proposed design is the situation where ecological data are supplemented with case-only data. The design is illustrated using a dataset of county-specific lung cancer mortality rates in the state of Ohio from 1988.

Keywords

Ecological bias; Efficiency; Outcome-dependent sampling; Two-phase sampling; Within-area confounding

1. Introduction

In an epidemiological ecological study the association between disease risk and exposure is investigated at the level of the group, rather than at the level of the individual. Such studies are appealing as they offer the possibility of high power due to large population sizes and increased exposure variability across areas (Prentice and Sheppard, 1995). In addition, they are logistically convenient since they may make use of routinely-available data (Morgenstern, 1998). Scientific interest, however, usually lies at the level of the individual and it is well known that ecological studies are susceptible to a range of biases with respect to the estimation of individual-level associations. There is a large epidemiological literature on the topic, in particular the difficulty in controlling for confounding, see for example Greenland (1992), Greenland and Robins (1994) and Richardson and Monfort (2000). Ecological studies are also used extensively in the social sciences, Wakefield (2004) provides a review and critique. The collective impact of these biases, for which an umbrella term is ecological bias, may give rise to a phenomenon referred to as the ecological fallacy. This occurs when conclusions regarding individual-level associations drawn on the basis of a group-level analysis differ from those drawn on the basis of an individual-level analysis.

The fundamental difficulty in using group-level data to assess individual-level associations is that of identifiability. Given an individual-level model, the loss of information associated with

Address for corresponence: Sebastien J-P.A. Haneuse, Center for Health Studies, Group Health Cooperative of Puget Sound, 1730 Minor Ave., Suite 1600, Seattle WA 98101, USA. haneuse.s@ghc.org.

only observing ecological data typically results in an inability to estimate all components of the model. A well-known example of this difficulty is in the estimation of contextual effects, where an individual's response is influenced not only by their own characteristics but also by the characteristics of other individuals in a shared environment. Such effects are of great interest in the social sciences and social epidemiology. Unfortunately ecological data alone do not allow the simultaneous estimation of individual and contextual effects (e.g. Wakefield, 2004). In more general settings, non-identifiability arises from the inability of ecological data alone to characterize within-area variability in exposures and confounders. While ecological data provide information regarding the marginal distributions of exposures and confounders, estimation requires knowledge of their joint distribution. Without further information, additional assumptions are required to induce identifiability. Lasserre et al. (2000), for example, propose approximating the within-area variability in the case of binary risk factors by assuming within-group independence of these factors. Given ecological data alone, however, such assumptions are generally untestable (Greenland, 2001; Wakefield, 2004).

The solution to the ecological inference problem is to collect individual-level information. Prentice and Sheppard (1995) describe an aggregate data design in which exposure/confounder data are collected on surveys of individuals within each area in order to estimate the within-area distribution of exposures and confounders. Individual-level outcome data are not obtained, however, and consequently one cannot distinguish between diseased and non-diseased individuals among those surveyed. Subsequent analyses, therefore, are still viewed as being at the level of the group (Sheppard, 2003). Another approach is to combine ecological data with cohort data; the utility of this approach was demonstrated by Wakefield (2004) in a social science context. However, in the situation of a rare event this strategy is not efficient since a random sample of individuals within an area would produce a small number of cases, indicating a rationale for the aggregate data approach of Prentice and Sheppard (1995).

In this paper, we propose a hybrid design in which ecological data are supplemented with case-control data. The case-control data provide a direct link between individual-level responses and explanatory variables. Analyses are therefore at the level of the individual, which allows the direct assessment of the risk-exposure-confounder model. In epidemiological settings, groupings are often based on geographic location and consequently referred to as areas; this will form the context here. Numerous applications of ecological studies exist, in particular for chronic diseases. For example, Prentice and Sheppard (1990) discuss the association between international differences in cancer rates and dietary fat intake, and Maheswaren et al. (1999) examine the association between ischaemic heart disease mortality and magnesium in areas containing a maximum of 50,000 people in north-west England. We focus on inference for a series of 2×2 tables. Although this scenario will be overly simple for most applications, it provides an extendable framework within which the various issues may be examined and for which there is a large body of existing literature (see Wakefield, 2004, and references therein).

The structure of this paper is as follows. In Section 2 we develop the likelihood for the hybrid design with a single binary exposure. There are connections between the proposed design and two-phase sampling (Breslow and Holubkov, 1997a), and these are explored in Section 3. Section 4 provides a simulation study and in Section 5 we extend the design to the case in which the outcomes are stratified by a binary confounder variable, and Section 6 demonstrates the benefits of the hybrid approach via simulation. Section 7 illustrates the proposed methods using lung cancer mortality data from the state of Ohio. Section 8 contains a concluding discussion, including a number of extensions to the basic design. An appendix provides some technical details.

2. Single binary exposure

We begin by developing the likelihood for the case in which the association between a disease outcome Y and a binary exposure X is to be investigated. Suppose the study area is partitioned into K sub-areas and let $Y = 0/1$ represent non-disease/disease, and $X = 0/1$ unexposed/exposed. For notational convenience, we temporarily omit the area-specific index.

The target of inference is assumed to be the individual-level association between Y and X. Let p_x denote the probability of disease, within some well-defined period, for an individual with exposure status x, $x = 0, 1$. We assume the logistic model

$$\log\left(\frac{p_x}{1-p_x}\right)=\beta_0+\beta_x x \tag{1}$$

so $\theta_0 = \exp(\beta_0)$ is the baseline odds, and $\theta_X = \exp(\beta_X)$ the multiplicative change in odds associated with exposure. For a rare disease θ_X approximates the relative risk.

Table 1 summarises the data that are available for a generic area, N_{yx} represents the number of individuals with disease status y and exposure status x, y, $x = 0, 1$. In an individual level study the number of unexposed cases, N_{10}, and exposed cases, N_{11}, would be observed. If the internal cells N_{10} and N_{11} were observed then, assuming independent outcomes within areas, the likelihood would correspond to the product of:

$$N_{1x}|M_x \text{Bionomial}\,(M_x, P_x)\,, \tag{2}$$

for $x = 0, 1$, and with p_x as given in (1). We refer to (2) as the *individual-level likelihood*, $L^I(\theta, N_{11})$ with $\theta = (\theta_0, \theta_X)$. If N_{01} and N_{11} were observed then estimation and inference would proceed in the usual manner where, assuming independent outcomes across areas, the likelihood consists of the product of contributions from each of the K study areas.

The ecological data consist of the aggregate response $N_1 = N_{10} + N_{11}$, along with the marginal exposure data M_0, M_1. Hence, the internal cells of the ecological 2×2 table are unobserved. In the design we consider a case-control sample is drawn, consisting of n_0 controls randomly selected from the N_0 total non-cases and n_1 cases randomly selected from the N_1 total cases; n_{yx} represents the number of individuals in the case-control sample with disease status y and exposure x (Table 1). We emphasize that n_{yx} are sampled directly from n_{yx}, and so are a subset of the population data. In a conventional case-control study, n_0 and n_1 are treated as being fixed and are conditioned upon. In the present context, however, if the number of cases and controls are fixed in advance then the number of cases, N_1, may exceed the total total number in the ecological data, n_1, which is random with support on the range $[0, N]$. Consequently n_0 and n_1 must be treated as random, conditional upon the ecological data, N_1, and the total case-control sample size, n. Specific schemes for determining n_0 and n_1 are described in Section 2.2. Figure 1 provides a graphical model of the hybrid design. Conditional independencies are displayed using single line arrows, double line arrows indicate deterministic relationships, and circular and square boxes represent unobserved and observed quantities, respectively; N_{10} and N_{11} are unobserved random variables, and N_{00} and N_{01} are deterministic quantities that depend on these variables, along with the ecological exposure totals, which are observed.

To simplify notation let $M_x = (M_0, M_1)$ and $N_y = (N_0, N_1)$ for the ecological data, $N_{yx} = (N_{10}, N_{11})$ for the internal cells, $n_y = (n_0, n_1)$ for case-control sample sizes, and $n_{yx} = (n_{01}, n_{11})$ for the case-control outcome data. The probability distribution of the observed data may be decomposed into three components:

$$\text{pr}\left(N_y, n_y, n_{yx} | M_x, n\right) = \text{pr}\left(N_y | M_x\right) \times \text{pr}\left(n_y | N_y, n\right) \times \text{pr}\left(n_{yx} | M_x, N_y, n_y\right) \tag{3}$$

corresponding to the distributions of the ecological data, the case-control sample sizes, and the case-control outcomes. We now derive the forms of each of these components.

2.1. Ecological data

Given the marginal exposure counts M_x, the induced likelihood based on the ecological data, N_y, is obtained by averaging over the distribution of the unobserved internal cells of the ecological 2×2 table. Conditional on the margins, only a single entry needs to be specified to complete the internal structure. If exposure is less common than non-exposure in a particular area then N_{11} should be chosen since it will have the smallest range over which to average. Under (1) and (2) the ecological data follow a convolution of two binomial distributions:

$$\text{pr}\left(N_y | M_x\right) = \theta_0^{N_1} \frac{1}{(1+\theta_0)^{M_0}} \frac{1}{(1+\theta_0\theta_X)^{M_1}} \sum_{N_{11} \in R_1} \begin{pmatrix} M_0 \\ N_1 - N_{11} \end{pmatrix} \begin{pmatrix} M_1 \\ N_{11} \end{pmatrix} \theta_X^{N_{11}}, \tag{4}$$

for $N_1 = 0, \ldots, N$, and where $R_1 = \{\max(0, N_1 - M_0), \ldots, \min(N_1, M_1)\}$. We define the *ecological likelihood* $L^E(\theta)$ as

$$L^E(\theta) = \sum_{N_{11} \in R_1} w^E(N_{11}) L^I(\theta, N_{11}),$$

where $L^I(\theta, N_{11})$ is the individual likelihood corresponding to $\text{pr}(N_{10}|M_0) \times \text{pr}(N_{11}|M_1)$ with each component given by (2), and $w^E(N_{11}) = 1$ for all $N_{11} \in R_1$.

A number of authors including Plackett (1977), McCullagh and Nelder (1989, Section 9.3.3) and Wakefield (2004) have discussed (4). For estimation there is a clear lack of identifiability if only a single area is considered, since we have a single response, N_1, and two unknown parameters (θ_0, θ_X). The likelihood for (θ_0, θ_X) has a ridge with a saddle point at $(N_1/N_0, 1)$ and attains its maximum on the boundary of the parameter space, either at $\theta_X = 0$ or $\theta_X = \infty$. As $N \to \infty$ the ridge becomes progressively flatter so that in the limit the score equations are satisfied by all values of (θ_0, θ_X) on the ridge, illustrating the lack of identifiability. This ridge is equivalent to the tomography line, defined in terms of the exposure-specific probabilities, which is considered by King (1997, Chapter 5). Figure 2(a) shows the ecological likelihood for a particular 2×2 table with $(N_1, M_0, M_1) = (125, 20000, 20000)$. The corresponding profile log-likelihood for $\beta_X = \log\theta_X$, with minimum at $\beta_X = 0$, is shown in Figure 2(b). Wakefield (2004) describes a variety of approximations to the ecological likelihood in non-rare settings. In the case of a rare disease each of the binomial distributions (2) may be approximated by Poisson distributions. In this case it is natural to replace the logistic model (1) with the log-linear form $\log p_x = \beta_0 + \beta_X x$, to obtain the aggregate distribution $N_1 | M_x \sim \text{Poisson}(M_0\theta_0 + M_1\theta_0\theta_X)$, resulting in a likelihood that is flat along the ridge.

2.2. Case-control sample sizes

Care must be taken when the case-control sample sizes are chosen since we cannot sample more cases than are available in the ecological data; a similar issue arises in two-phase sampling (e.g. Breslow and Holubkov, 1997a, p. 453). Hence the control and case sample sizes, n_0 and n_1 respectively, are random variables. In the econometrics literature random case-control sample sizes are common (e.g., Manski and Lerman, 1977; Scott and Wild, 1997). One

possibility is to fix n, and sample cases and controls with probabilities π and $1-\pi$ respectively; if the cases are exhausted then the remaining individuals are selected as controls. An alternative is to fix a nominal number of cases n_1^*, in addition to n, and then take $n_1 = \min\left(N_1, n_1^*\right)$ and setting $n_0 = n - n_1$. These schemes do not impact point estimation since the distribution of n_0 and n_1 is specified so that it is independent of both (θ_0, θ_X) and the unobserved (n_{10}, N_{11}). Hence n_0, n_1 are ancillary and there is no contribution to the overall likelihood from this component. The decomposition given by (3) may therefore be simplified to

$$\text{pr}\left(N_y, n_{yx} | M_x, n\right) = \text{pr}\left(N_y | M_x\right) \times \text{pr}\left(n_{yx} | M_x, N_y, n_y\right). \tag{5}$$

Calculation of the expected information matrix, which is of particular interest for study design, does depend on the scheme adopted, however.

2.3. Case-control outcomes

Crucial to the development of the likelihood for the case-control data is the recognition that conditional upon the internal cells of the ecological 2×2 table, the number of exposed controls, n_{01}, and the number of exposed cases, n_{11} follow independent hypergeometric distributions. For example, suppose we have a population of cases N_1 of which N_{11} are exposed, and we draw a random sample of size n_1 from this population; in this case the number exposed, n_{11}, follows a hypergeometric distribution. Upon conditioning on the unobserved n_{yx}, the case-control outcomes do not depend on the parameters of the model, as is clear in Figure 1. Unconditionally, the likelihood is found by averaging over the unobserved internal cells of the ecological data; given the margins we only need to consider a single cell and we again choose as auxiliary variable N_{11}. The average is with respect to the distribution of $N_{11}|N_y, M_x$, which is an extended hypergeometric random variable (Johnson and Kotz, 1969, Chapter 6). Note that $\text{pr}(N_{11}|N_y, n_y, M_x) = \text{pr}(N_{11}|N_y, M_x)$, so that N_{11} is conditionally independent of n_y given N_y. This conditional independence can be determined from Figure 1, since every path between, N_{11} and n_1, given N_1, is closed (Pearl, 2000). Said another way, once we know N_y there is no further information concerning N_{11}, contained in n_y. Hence the joint distribution of the number of exposed cases and controls, n_{11} and n_{01}, is given by the product of two hypergeometric distributions averaged over an extended hypergeometric random variable:

$$\begin{aligned}
\text{pr}\left(n_{yx} | n_y, M_x, N_y\right) &= \sum_{N_{11} \in R_1^*} \text{pr}\left(n_{01}, n_{11} | N_{11}, N_y, n_y, M_x\right) \text{pr}\left(N_{11} | N_y, n_y, M_x\right) \\
&= \sum_{N_{11} \in R_1^*} \frac{\binom{N_{01}}{n_{01}}\binom{M_0 - N_1 + N_{11}}{n_0 - n_{01}}}{\binom{N_0}{n_0}} \frac{\binom{N_{11}}{n_{11}}\binom{N_1 - N_{11}}{n_1 - n_{11}}}{\binom{N_1}{n_1}} \frac{\binom{M_0}{N_1 - N_{11}}\binom{M_1}{N_{11}}\theta_X^{N_{11}}}{\sum_{u \in R_1}\binom{M_0}{N_1 - u}\binom{M_1}{u}\theta_X^{u}}
\end{aligned} \tag{6}$$

where the support of N_{11} is given by $R_1^* = \{\max(n_{11}, N_1 - M_0 + n_0 - n_{01}), ..., \min(N_1 - n_1 + n_{11}, M_1 - n_{01})\}$. The latter range reflects the ecological constraints in R_1, together with constraints from the case-control contribution, specifically, $N_{10} \geq n_1 - n_{11}$ and $N_{11} \geq n_{11}$. To emphasize the finite-sample nature of this contribution, due to the conditioning on the ecological data, we refer to this likelihood as the *finite sample case-control likelihood*. Averaging over the unobserved N_{11} hence provides a likelihood, (6), which depends only on θ_X and provides no information regarding θ_0. Similarly, in a traditional case-control study the baseline odds parameters cannot be estimated from the case-control data alone.

In a conventional case-control study M_x is not conditioned upon, and inference proceeds via logistic regression with implicit sampling from a hypothetical super-population in which N, $N_1 \to \infty$ in such a way that the proportions of exposed controls, N_{01}/N_0, and exposed cases, N_{11}/N_1, tend to non-zero constants. Under these conditions each of the exposure-specific hypergeometric distributions tend to a binomial distribution. Prentice and Pyke (1979) showed that, in the semi-parametric setting of a parametric logistic regression model with an unspecified distribution for the covariates, asymptotic inference for non-intercept parameters is identical for prospective or retrospective data collection. In our setting we are adding extra *finite sample* information via the ecological margins, and hence we increase effciency over the unrestricted situation considered by Prentice and Pyke (1979). Chatterjee and Carroll (2006) illustrate that adding additional constraints can provide improved inference over an unconstrained analysis; in their case the additional constraint arose from assuming independence of genetic and environmental factors in the population.

In Section 4 we illustrate that the use of the finite sample case-control likelihood, (6), can provide significant effciency gains over conventional logistic regression analyses, even without the direct contribution of the ecological data contained in (3). The use of the finite sample case-control likelihood does not seem to have been previously considered, perhaps because within the survey sampling literature *design-based* inference, which avoids parametric assumptions such as a logistic risk model, is typically carried out. For discussion of this aspect see Chapters 8 and 12 of the edited volume of Chambers and Skinner (2003).

2.4. Likelihood inference

The *hybrid likelihood*, which we denote $L^H(\theta)$, is the product of (4) and (6), and is a function of $\theta = (\theta_0, \theta_X)$. Following simplification to give a single summation we have

$$\begin{aligned} L^H(\theta) &= \theta_0^{N_1} \frac{1}{(1+\theta_0)^{M_0}} \frac{1}{(1+\theta_0\theta_x)^{M_1}} \sum_{N_{11} \in R_1^*} w^H(N_{11})\, \theta_x^{N_{11}} \\ &= \sum_{N_{11} \in R_1^*} w^H(N_{11})\, L^1(\theta, N_{11}) \end{aligned} \tag{7}$$

where

$$w^H(N_{11}) = \frac{\binom{M_0}{N_1 - N_{11}}\binom{M_1}{N_{11}}\binom{M_1 - N_{11}}{n_{01}}\binom{N_{11}}{n_{11}}\binom{M_0 - N_1 + N_{11}}{n_0 - n_{01}}\binom{N_1 - N_{11}}{n_1 - n_{11}}}{\binom{N - N_1}{n_0}\binom{N_1}{n_1}}$$

so that again we have a representation as a weighted sum of individual-level binomial likelihoods.

So far we have considered a single area only; in practice we will have contributions of the form (7) from K areas. In the simplest case the K areas will share the baseline risk and association parameters. We consider the asymptotic distribution of the maximum likelihood (ML) estimator $\widehat{\theta}$ as $K \to \infty$; with an obvious notation, N_{yk}, n_{yxk}, $k = 1, \ldots, K$, are independently distributed across areas, and consistency of the hybrid ML estimator follows from its representation as an M-estimator and from Wald's conditions for consistency of such estimators (van der Vaart, 1998, Theorem, 5.14) and asymptotic normality from van der Vaart (1999, Theorem 5.39). Hence asymptotic inference for the ML estimator $\widehat{\beta}$ may be based upon

$$\widehat{\theta} N\left(\theta, I^H(\widehat{\theta})^{-1}\right), \tag{8}$$

where $I^H(\theta)$ is the observed information. The asymptotics require $\Sigma_k M_{0k}$ and $\Sigma_k M_{1k}$, and at least one of $\Sigma_k n_{0k}$, $\Sigma_k n_{1k} \to \infty$, with the area-specific components going to infinity if there are area-specific parameters in the risk model. The expected information, which is useful for comparing different designs, is computationally daunting since one must take the expectation with respect to the joint distribution of N_y, n_y, n_{yx}, which is given by (3).

We now consider the form of the observed information. Details are presented in terms of θ, since the forms are simpler to represent; in practice asymptotic interval estimates will be evaluated on the log odds β scale and then transformed to the θ scale. Let $S^I(\theta)$, $S^E(\theta)$ and $S^H(\theta)$ denote the score statistics for the individual, ecological and hybrid likelihoods, respectively, where

$$S^I(\theta) = \begin{bmatrix} \frac{N_1}{\theta_0} - \frac{M_0}{1+\theta_0} - \frac{M_1\theta_X}{1+\theta_0\theta_X} \\ \frac{N_{11}}{\theta_X} - \frac{M_1\theta_0}{1+\theta_0\theta_X} \end{bmatrix}.$$

Similarly $I^I(\theta)$, $I^E(\theta)$ and $I^H(\theta)$ are the corresponding observed information matrices with

$$I^I(\theta) = \begin{bmatrix} \frac{N_1}{\theta_0^2} - \frac{M_0}{(1+\theta_0)^2} - \frac{M_1\theta_x^2}{\left(1+\theta_0\theta_X\right)^2} & \frac{M1}{\left(1+\theta_0\theta_X\right)^2} \\ \frac{M1}{\left(1+\theta_0\theta_X\right)^2} & \frac{N_{11}}{\theta_X^2} - \frac{M1\theta_0^2}{\left(1+\theta_0\theta_X\right)^2} \end{bmatrix}.$$

Convenient forms for the score and information of both the ecological and hybrid designs are obtained by exploiting the missing data representation of the likelihood, e.g. Little and Rubin (2002, Chapter 8) . The score vector and observed information for the ecological data are given by

$$S^E(\theta) = E\left[S^I(\theta) | N_y, M_x\right] \tag{9}$$

$$I^E(\theta) = E\left[I^I(\theta) | N_y, M_x\right] - \mathrm{var}\left[S^I(\theta) | N_y, M_x\right], \tag{10}$$

where the expectations are with respect to the distribution of $N_{yx}|N_y, M_x$, which is an extended hypergeometric distribution (Section 2.3). These forms were also presented in the context of survey sampling by Breckling et al. (1994), and in an ecological context by Steel et al. (2004). The expression for $I^E(\theta)$, clarifies the loss of information (given by the second term on the right-hand side of (10)) due to the aggregation of individual level data. The score and information for the hybrid likelihood are give by

$$\begin{aligned} S^H(\theta) &= E\left[S^I(\theta) | N_y, M_x, n_y, n_{yx}\right] \\ I^H(\theta) &= E\left[I^I(\theta) | N_y, M_x, n_y, n_{yx}\right] - \mathrm{var}\left[S^I(\theta) | N_y, M_x, n_y, n_{yx}\right] \end{aligned}$$

where the expectations are with respect to the distribution

$$\text{pr}\left(N_{yx}|N_y, M_x, n_y, n_{yx}\right) = \frac{\text{pr}\left(n_{yx}|N_{yx}, N_y, n_y\right)\text{pr}\left(N_{yx}|N_y, M_x\right)}{\text{pr}\left(n_{yx}|N_y, M_x, n_y\right)}. \tag{11}$$

which we call the *supplemented extended hypergeometric distribution*. In (11) the distribution on the right of the numerator is an extended hypergeometric distribution, and is supplemented via the addition of the case-control data, which is the distribution on the left of the numerator.

2.5. Case-only data

In some situations it may be straightforward to determine the exposure status of cases, for example, from a disease registry, while for the controls no such information is directly available. A hybrid design is available for this situation with a likelihood of the same form as (7) but with weights

$$w^H(N_{11}) = \frac{\binom{M_0}{N_1 - N_{11}}\binom{M_1}{N_{11}}\binom{N_{11}}{n_{11}}\binom{N_1 - N_{11}}{n_1 - n_{11}}}{\binom{N_1}{n_1}}$$

for $N_{11} \in \{\max(n_{11}, N_1 - M_0), \ldots, \min(N_1 - n_1 + n_{11}, M_1)\}$. Inference follows in a similar fashion to the full hybrid design, except that the expectations of the score and information matrices are with respect to the above weights. Supplementing ecological data with case-only data, therefore, provides an interesting alternative design that is practically attractive. Identifiability also results from the addition of control-only data but this scenario is less practical and will generally require more samples than a case-only sample.

2.6. Illustrative example

We consider a single 2×2 table in detail, before reporting a more comprehensive simulation study in Section 4. Returning to the example referred to in Figure 2, we supplement the ecological data (N_1, M_0, M_1) = (125, 20000, 20000) with the case-control data (n_0, n_1, n_{01}, n_{11}) = (50, 50, 26, 35). These data result in a range for the unobserved number of exposed cases, N_{11}, of $R_1 = \{0, \ldots, 125\}$, based on the ecological data alone, and $R_1^* = \{35, ..., 100\}$ for the combined data, illustrating how the support is constrained by the addition of the case-control data. The likelihood (7) was maximised using a Newton-Raphson algorithm with analytical derivatives. For a single table, the ecological likelihood does not provide an identifiable estimator since we have two parameters and a single observation, as illustrated in Figures 2(a) and 2(b).

Analyses of the case-control data only using logistic regression yields an estimate (asymptotic 95% confidence interval) for θ_X of 2.15 (0.95,4.89); the likelihood surface is shown in Figure 2(c). Use of the finite sample case-control likelihood, (6), gave an estimate of 2.34 (1.28,4.29) illustrating the reduction in the width of the interval due to the marginal constraints available from the ecological data. In this example identical values resulted from the hybrid design which adds the direct contribution of the ecological outcome data via the likelihood (4). In general, the hybrid analysis also exploits between-area differences in the exposure margin, but for a single area there is no such gain. The likelihood surface in this case is plotted in Figure 2(d), comparison with Figure 2(c) clearly shows the concentration of the likelihood; this is confirmed by the profile likelihood for $\beta_X = \log \theta_X$ shown in Figure 2(e). In this example the case only

estimate and asymptotic standard error were unchanged from the values produced by the case and control data combined; weights $w^H(N_{11})$ are virtually identical under the two schemes. Figure 2(f) shows the likelihood surface in the case-only situation.

3. Connections with two-phase sampling

In two-phase sampling, a large phase I sample is cross-classified by outcome and discrete covariates. At phase II data on additional variables are sampled from within each of the cross-classified cells. Such a design can provide large efficiency gains over a study which stratifies solely on the basis of outcome status (as in a case-control study). There are clear similarities between the hybrid design and two-phase sampling, with the ecological and case-control data being analogous to phase I and phase II data, respectively. There is a large literature on two-phase studies, see for example White (1982), Flanders and Greenland (1991), Breslow and Holubkov (1997a), Scott and Wild (1997) and Breslow and Chatterjee (1999). Lawless et al. (1999) consider more general outcome-dependent sampling schemes. In an ecological context a plausible two-phase scheme would consist of phase I data that are a $2\times K$ cross-classification of disease status by area, with phase II data sampled within each of the $2\times K$ strata. The crucial distinction between this and the hybrid design is that marginal exposure data are not used in the two-phase scheme. For exposure to be incorporated into the phase I stratification, we would require cross-classification of disease counts by exposure status which corresponds to knowing the internal cells of the K 2×2 tables and is therefore not consistent with an ecological study.

For inference, various approaches have been suggested, including pseudo-, weighted, and full ML estimation (Breslow and Holubkov, 1997b). In comparisons with the hybrid design we implement full ML estimation, for details see Scott and Wild (1997) and Breslow and Holubkov (1997a). Briefly, the two-phase likelihood is complex, with the parameter space constrained because the phase II data are a subset of the phase I data; stratum-dependent offsets are specified within an iterative algorithm, in order to acknowledge the phase II outcome-dependent sampling design. Breslow and Chatterjee (1999) provide details of available code for an R/Splus implementation. In one sense, two-phase regression lies between logistic regression and the hybrid design. In contrast to logistic regression, the group-level disease totals across areas (the phase I stratification) are used in a two-phase analysis. However, a two-phase approach does not make use of the information in the exposure margins, as does the hybrid design.

The development of two-phase methods was motivated by potential efficiency gains associated with judicious stratification of an initial sample, from which sub-samples may then be drawn. In contrast, the present development is motivated by the fundamental difficulty of non-identifiability of individual-level models when ecological data alone is collected. Although the hybrid design was proposed to alleviate ecological bias, substantial efficiency gains may also be achieved through its use of the ecological data.

4. Simulation study for a single binary exposure

The exploitation of between-area exposure variation is a primary motivation for carrying out an ecological study. We report a simulation study in which there are $K = 20$ areas with 40,000 individuals in each area (all subsequent simulations use these values). For simplicity we assume constant θ_0 and θ_X across areas. In this setting the logistic regression analysis must include K area-specific intercepts, to acknowledge the design. In addition we report inference from the hybrid and ecological designs (in this situation the ecological data provide an identifiable likelihood since there are 20 observations and two parameters), and the finite sample case-control and two-phase approaches. In the simulations reported below, as we assume a constant baseline risk across strata (area) we note that full two-phase ML estimation is not equivalent to pseudo-ML; the two are equivalent if the model contains stratum-specific intercepts.

We report results based on 1,000 simulated datasets and across all areas we assume a common individual-level model, with $\theta_0 = 0.002$ and $\theta_X = 2$. We examine four different scenarios, with results in Table 2. In the baseline set of simulations the proportion exposed increases deterministically between 0.2 and 0.8 across areas. Conditional on the corresponding exposure totals, and given the disease model, the expected number of cases ranged between 64 and 144. Within each area, the total number of cases and controls sampled is $n = 20$, with $n_0 = n_1 = 10$. In the first set of simulations the relatively large exposure range results in an efficiency of 59% for the ecological analysis, as compared to the hybrid design. For the latter, the standard error of $\widehat{\theta_x}$ is 0.21. The finite-sample case-control analysis, which is highly efficient in the case of a single area, is far less efficient when compared to the hybrid design since it does not utilise the exposure variability across areas. There is finite sample bias in the logistic regression estimator and low efficiency, while for the two-phase design there is some improvement in both bias and efficiency.

In the second scenario we reduce the range of the proportion exposed across areas to (0.4,0.6) and, as expected, the ecological analysis performs poorly. There is an increase in the finite sample bias for all of the estimators, but the relative efficiency is increased for those methods that use individual-level data. In this case, the standard error for the hybrid estimator of $\widehat{\theta_x}$ increases to 0.29 reflecting the loss of information. In the third scenario we return to the original variability in the proportion exposed in each area, but increase the number of case-control samples to $n_0 = n_1 = 25$. As expected this results in reduced bias and increased efficiency for the logistic, two-phase, and finite sample case-control methods, though the finite sample case-control method still only reaches 57% efficiency when compared to the hybrid analysis. In the final scenario the number of case-control samples is decreased to $n_0 = n_1 = 5$, so that there are only 200 individual-level samples in total in the study. While the analyses that use only the individual level data have low efficiencies and exhibit finite sample bias, the hybrid method performs well. In all simulations two-phase regression is more efficient than logistic regression, but less efficient than the finite sample case-control analysis, which conditions on the ecological data in order to reduce the number of possible enumerations of the observed case-control outcome data. In simulations not reported, doubling the number of areas K resulted in a halving of the variance of $\widehat{\theta_x}$ for all methods. Additional results in Haneuse (2004) show that for the sample sizes considered here the coverage probabilities of confidence intervals based on (8) achieve their nominal levels.

5. Stratified outcomes

In almost all epidemiological studies control for confounding is required, and the inability to control within-area confounding is a major drawback of ecological studies. In this section we extend the basic scenario of Section 2 by considering control for a single binary confounder Z. Again, we initially present the development in terms of a single area. At the individual level assume the logistic model

$$\log\left(\frac{p_{xz}}{1-p_{xz}}\right)=\beta_0+\beta_x x+\beta_z z, \tag{12}$$

where p_{xz} is the probability of disease for an individual with exposure x and confounder z, $x = 0, 1, z = 0, 1$. Hence $\theta_X = \exp(\beta_X)$ is the multiplicative change in odds associated with exposure, while controlling for Z, with an analogous interpretation for $\theta_Z = \exp(\beta_Z)$. Model (12) can easily be extended to include an interaction term, but for simplicity of presentation we present the main effects only model. Let M_{xz} denote the number of individuals with exposure x, and confounder z, , in a generic area, with $M_{xz} = (M_{00}, M_{10}, M_{01}, M_{11})$ and N_{yxz} be the number of

individuals with disease status y in exposure/confounder stratum x, z, $x = 0, 1$, $z = 0, 1$. In different settings, various forms of ecological and/or case-control data may be available. Here we consider the semi-ecological design (see for example, Sheppard, 2003) in which individual-level data on outcomes and confounders are obtained, but information on the exposure of interest is only available in the form of an ecological margin; Table 3 summarises notation. The outcomes stratified by the confounder variable, that is $N_{1+z} = (N_{1+0}, N_{1+1})$ are observed, in addition to the marginal counts of $X = 1$ and $Z = 1$, denoted M_{x+} and M_{+z}, respectively. Hence, the joint classification of X and Z is unobserved. In practice this scenario may arise in the context of chronic diseases where incidence is typically recorded by the potential confounders gender, age and race (see Section 7 for a specific example). It is far less likely that incidence will be available by exposure, however. Lasserre et al. (2000) considered a study with two binary risk factors; the response was lung cancer mortality in 82 French departments. The exposure corresponded to the proportion of men employed in the metal industry, and the proportion resident in towns larger than 2000 inhabitants was used as a proxy for confounding variables related to urbanization. In this example, town of residence would be available from the death certificate and so the stratified ecological data just described would be available.

We assume that within each area cases and controls are sampled within each level of Z. Consequently, there are n_{1z} cases sampled in stratum z, of which n_{11z} are exposed, with n_{0z} and n_{01z} being the corresponding numbers amongst the controls, $z = 0, 1$. We assume that the stratified total number of cases and controls, n_{+z}, $z = 0, 1$, are fixed. To simplify notation let $n_{yz} = (n_{00}, n_{10}, n_{01}, n_{11})$ and $n_{yxz} = (n_{010}, n_{110}, n_{011}, n_{111})$. The stratum-specific, case-control sample sizes, n_{yz}, need to be viewed as random variables though they are again ancillary and can be conditioned upon. We decompose the joint distribution into the distributions of the ecological data, and the case-control data conditional upon the ecological data:

$$\text{pr}\left(N_{1+z}, n_{yxz} | M_{1+}, M_{+1}, N, n_{yz}\right) = \text{pr}\left(N_{1+z} | M_{1+}, M_{+1}, N\right) \times \text{pr}\left(n_{yxz} | N_{1+z}, M_{1+}, M_{+1}, N, n_{yz}\right). \tag{13}$$

5.1. Ecological Data

To obtain the likelihood for the ecological data note that if M_{11} were observed, in addition to M_{x+}, M_{+z}, then each of $N_{1+z}|M_{0z}, M_{1z}$, $z = 0, 1$ is the convolution of a pair of binomial distributions as in (4). Unconditionally we average over the unobserved M_{11} to give

$$\text{pr}\left(N_{1+z} | M_{1+}, M_{+1}, N\right) = \sum_{M_{11} \in S_{11}} \left\{ \prod_{z=0}^{1} \text{pr}\left(N_{1+z} | M_{0z}, M_{1z}\right) \right\} \text{pr}\left(M_{11} | M_{1+}, M_{+1}, N\right) \tag{14}$$

where $S_{11} = \{\max(0, M_{+1} - M_{0+}), \ldots, \min(M_{+1}, M_{1+})\}$, and $\text{pr}(M_{11}|M_{1+}, M_{+1}, N)$. The latter is an extended hypergeometric random variable with odds ratio parameter $\phi_{XZ} = q_{11} \times q_{00}/q_{10} \times q_{01}$, where $q_{xz} = \text{pr}(X = x, Z = z)$ and ϕ_{XZ} is the odds ratio describing the association between the exposure and confounder variables. Hence we have three auxiliary variables, N_{110}, N_{111} and M_{11}, in the ecological likelihood. As a likelihood (14) is a function of ϕ_{XZ}, as well as $\theta = (\theta_0, \theta_X, \theta_Z)$.

5.2. Hybrid likelihood

Following a similar argument to that of Section 2 the joint distribution of the ecological and case-control data is

$$\text{pr}\left(N_{1+z}, n_{yxz}|M_{1+}, M_{+1}, N, n_{yz}\right) = \sum_{M_{11}\in S_{11}^*} \Pr\left(N_{1+z}, n_{yxz}|m_{xz}\right) \text{pr}\left(M_{11}|M_{1+}, M_{+1}, N\right) = \sum_{M_{11}\in S_{11}^*} \left\{\prod_{z=0}^{1} \Pr\left(N_{1+z}|M_{0z}, M_{1z}\right) \Pr\left(n_{y1z}|n_{0z}, n_{1z}, N_{1+z}\right)\right\} \text{pr}\left(M_{11}|M_{1+}, M_{+1}, N\right) \quad (15)$$

where each of the terms in curly brackets is in the form of the hybrid likelihood in the single exposure case (as in (7)), and $S_{11}^* = \{\max(0, M_{+1} - M_{0+}), ..., \min(M_{+1}, M_{1+})\}$. Asymptotic inference is again based upon the observed information, details of the calculation of the score vector and observed information matrix are outlined in the Appendix.

We consider three alternative individual-level designs. First, we implement a two-phase study design consisting of phase I data composed of a $2 \times 2 \times K$ stratification of disease status by confounder by area. Phase II data then consist of case-control samples within phase I strata. Second, a conventional logistic regression analysis uses the case-control data only and includes $2K$ area/confounder specific offsets in the model to acknowledge the matching (since sampling is carried out on the basis of the confounder margin, θ_Z cannot be estimated without additional information). Finally, we examine the finite sample case-control approach that conditions on the ecological data. The probability distribution in this situation is, from (13), given by

$$\text{pr}\left(n_{yxz}|N_{1xz}, M_{x+}, M_{+z}\right) = \frac{\text{pr}\left(N_{1xz}, M_{x+}, n_{yxz}|M_{x+}, M_{+z}\right)}{\text{pr}\left(N_{1xz}|M_{x+}, M_{+z}\right)}$$

with denominator given by (14) and numerator by (15), and depends upon ϕ_{XZ}, as well as upon θ.

6. Simulation study for stratified outcomes

6.1. Constant baseline odds

In this section we assume the parameters of the disease model to be constant across all areas, as in (12) while in Section 6.2 we allow between-area heterogeneity in the baseline odds.

We take $(\theta_0, \theta_X, \theta_Z, \phi_{XZ}) = (0.002, 2, 2, 2)$ and assume that the marginal exposure probability $\text{pr}(X = 1) = q_{10} + q_{11}$ ranges uniformly between [0.1, 0.4] and that in each area the probability of $X = Z = 0$ is $q_{00} = 0.25$; this results in the marginal confounder prevalence ranging between 0.64 and 0.73 across areas. We take 5 cases and 5 controls in each confounder stratum, i.e. $n_{yz} = (5, 5, 5, 5)$.

A Newton-Raphson algorithm was used to find the ML estimates for the finite sample case-control and hybrid analyses. Variances were calculated using the observed information, and asymptotic confidence intervals were again found to display their nominal coverage levels. The summation over M_{11} is computationally expensive since the support is large. To reduce the computational burden a strategy was adopted in which the mode of $M_{11}|M_{x+}, M_{+z}$ was found, summing over the values of non-negligible mass to either side of the mode, the remaining terms being ignored, for further details see Wakefield (2004).

The results are presented in Table 4, and are based on 1,000 simulations. Focusing upon the results for the parameter of interest, θ_X, we see that the hybrid design that uses both case and control information is the most efficient, closely followed by the case-only hybrid design. Two-phase regression has negligible bias but low efficiency because the marginal exposure information is not exploited. The variance of $\widehat{\theta}_x$ in the hybrid design is 21% lower than in the finite sample case-control design, which profits from the use of the ecological data to constrain the counts, via the hypergeometric contributions. For all methods there is virtually no bias in the estimation of θ_Z, because the case-control samples are stratified by Z.

In these simulations, with a semi-ecological design and two binary covariates, we see positive bias. Jackson et al. (2006) examine the benefits of adding individual-level data to ecological data, and consider a scenario in which there are two covariates, one binary and one continuous. In their simulations there is negative correlation between the between-area means and the variances of the continuous variables which leads to negative bias, as can be deduced from the formulas presented in Wakefield (2003).

6.2. Fixed effects baseline odds

In this section we consider the extension to the case in which the baseline odds vary by area. Such a model may be used to control for between-area confounding, though the ideal is to collect area-level variables to alleviate the need for such fixed effects.

For the development we need to explicitly introduce area-specific notation and so we let p_{xzk} represent the probability of disease for an individual with exposure x and confounder z in area k, $x = 0, 1$, $z = 0, 1$, $k = 1, ..., K$, . We replace model (12) with

$$\log\left(\frac{p_{xzk}}{1-p_{xzk}}\right)=\beta_{0k}+\beta_x x+\beta_z z, \tag{16}$$

where we treat $\theta_{0k} = \exp(\beta_{0k})$ as fixed effects. Assuming a constant ϕ_{XZ} across all areas, estimation of the $K + 3$ parameters follows in an analogous fashion to that described in Sections 5.1 and 5.2.

For the simulation study, we again have $K = 20$ areas with 40,000 individuals per area and 5 cases and 5 controls in each confounder stratum, to give 20 individual samples per area. The parameter values are taken as $(\theta_X, \theta_Z, \phi_{XZ}) = (2, 2, 2)$, with the proportion exposed varying uniformly across areas between 0.1 and 0.4. The baseline odds, θ_{0k}, $k = 1, \ldots, 20$, were generated as uniform random variables over the range [0.001, 0.004], with the same set retained for all simulations.

The results over 1,000 simulations are reported in Table 5. We first note that two-phase regression has reduced bias when compared to the logistic regression model. The case-only hybrid design is again competitive, with small bias and high efficiency for the parameter of interest, though reduced efficiency for estimation of ϕ_{XZ}. With respect to estimation of θ_X, the finite sample case-control method gave virtually identical inference to the hybrid design in this setting. This result is as expected since, in contrast to the common baseline odds model, we would expect the hybrid design to be less powerful since the benefits of between-area exposure variability are lost when fixed effect baselines are present in the model. The incorporation of the finite sample information can still be exploited, however. Both the hybrid and the finite sample case-control method are more than twice as efficient as the logistic and two-phase approaches.

7. Ohio lung cancer data

We illustrate the hybrid design using cancer mortality data from the state of Ohio, taken from the National Center for Health Statistics (NCHS) Compressed Mortality File. For each of 88 counties population estimates and lung cancer death counts are available by gender, race (white vs non-white), and year of death (1968 to 1988 inclusively). For simplicity, we focus on population estimates and death counts for 1988. Further, although age information is available as 11 five or ten-year age bands, we consider individuals aged between 55 to 84 years collapsed into a single age category. Over the 88 counties the number of cases range between 4 and 922 with a median of 26. An attractive feature of these data are that counts are stratified by outcome status, gender and race jointly, and so we have individual-level information; we may therefore construct a hypothetical ecological study by considering the corresponding area-specific marginal totals only. Having individual-level information further provides a basis for the direct assessment of competing methods that do not use all information since an analysis based on complete individual-level data may be viewed as a gold standard. Hence, the biases that we report are relative to the complete data analysis.

We report results from three analyses, in each case taking the association of interest to be that between lung cancer and race. In the first analysis we examine the unadjusted association, while in the second and third analyses we stratify by gender and consider models with a single intercept and with area-varying intercepts, respectively. For the case of area-varying intercepts we do not consider an ecological analysis, since the model is unidentifiable. In the first analysis we sample 10 cases and 10 controls, apart from a small number of areas in which there are less than 10 cases; in these areas we sampled all cases, with the remainder individual samples being taken as controls. In the two stratified designs we take 100 case-control samples; 25 male cases, 25 male controls, 25 female cases and 25 female controls. If the cases were exhausted in a particular stratum, then additional controls were sampled. For simplicity no interaction between race and gender is considered.

In Table 6 we see that in the race only analyses the ecological analysis is positively biased, relative to the individual level (complete data) analysis. Logistic regression and two-phase regression have large standard errors, with point estimates that are also positively biased. The finite sample case-control analysis produces a low estimate while the two hybrid analyses provide accurate inference, although the estimates are slightly larger than that in the complete data case since we sampled equal numbers of cases and controls (10 of each) from each area. More information could be gained by varying the numbers sampled in each area; design issues will be the subject of a future paper.

In the fixed baseline analyses that were stratified by gender the ecological estimate is again positively biased but the hybrid analyses are accurate. In the analyses stratified by area the patterns are similar though now the results for the finite sample case-control and hybrid full analyses are virtually identical, as in the simulations of Section 6.2.

8. Discussion

The fundamental difficulty of using ecological data to assess individual-level associations is that of identifiability. Standard ecological approaches are susceptible to a range of biases, the collective impacts of which are referred to as ecological bias. The only solution to reducing ecological bias is to supplement ecological data with individual level information. In this paper we have proposed a hybrid design in which ecological and case-control data are combined, and have provided details of likelihood-based inference. The case-control data provide identifiability and control for confounding. The ecological data contribute between-area information on exposure, and by conditioning on the ecological margin increased efficiency

is gained. Strömberg and Björk (2004) have recently described the use of ecological exposure information in case-control studies, but without a formal statistical model.

In the simulations we have demonstrated the gains in efficiency of the hybrid design, and also that the finite sample case-control method can provide large efficiency gains over a conventional logistic regression approach. The availability of individual-level data allow both model checking and the fitting of more sophisticated models. In particular, estimation of contextual effects is important in a number of areas including social epidemiology. For example, the effects on health of area-level average income, as well as individual-level income, are the subject of much debate, e.g. Judge et al. (1998).

Throughout, computation was performed using Newton-Raphson and EM algorithms, but no systematic comparison of the merits of these approaches has been carried out. Asymptotic inference has also been relied upon, and a clearer understanding of the contributions from within-area (case-control, two-phase) information and between-area (ecological) information would be desirable. This will also lead naturally to formal design considerations; in particular the number of ecological areas to select, the choice of areas within which to sample individual level data, and the numbers of individuals within each of these areas to take.

For small samples in which asymptotic inference is inappropriate it is natural to turn to a Bayesian approach, with computation via Markov chain Monte Carlo (MCMC) and the introduction of auxiliary variables. For studies that are based on small areas in particular, allowing for spatial dependence in the baseline odds is also desirable. For example Clayton et al. (1993) use a model with spatially-dependent residuals in an ecological correlation study context. Implementing such a model may be carried out relatively easily using MCMC. A variety of ecological disease, exposure and confounder data may be available, and for increasing numbers of exposures and confounders, and categorical variables with more than two levels, computation will be prohibitive; the methods described in Dobra et al. (2003), building on work of Diaconis and Sturmfels (1998), may be useful in this respect.

We envisage that the hybrid design will be particularly useful for the investigation of environmental pollutants. As with all observational studies, there are a variety of important practical issues which require careful consideration. When case-control data are to be combined with ecological data, identifying an appropriate sampling frame is of vital importance. For the traditional case-control study two common choices are a population-based and a hospital-based sampling frame. In the context of the hybrid design a natural choice would be a hospital-based sampling frame. For example, suppose the case data are obtained from a cancer registry as all cases diagnosed within a well-defined geographical area (the study region) over a specific time period (the study period). For confidentiality reasons the data are available as the number of cases within each of a set of sub-regions that partition the study region. The population (case and non-case data) are obtained from the census as all individuals who were resident in the study region over the study period (and were eligible), and are also available by sub-region. Each of the cases and population will typically be broken down by demographic information such as age, gender and race. Hospital-based population sampling frames are less appealing since a hospital defined population will not exhaust a geographical area, since other hospitals may take patients from that area.

In practice, as with all epidemiological studies, exposure misclassification is an important issue. For the design that we have proposed the ecological exposure margin is likely to be subject to exposure misclassification. For example, in an environmental context, a pollution concentration surface may be modeled and a cut-off may determine a proportion in each area who are exposed. This proportion is likely to be error-prone. We investigate the effect of this exposure misclassification in the ecological data, via a simulation study which, for simplicity,

considers just a single binary exposure. We assume there are 10 cases and 10 controls within each area, with the same parameter values as in the simulations summarized in Table 6. Exposure data were simulated for each individual, and were then corrupted via probabilities $\text{pr}(W = 1|X = 0) = q_0$, $\text{pr}(W = 0|X = 1) = q_1$, where X is the true exposure of a generic individual, and W the error-prone measure. Summing up the number of error-prone "unexposed" and "exposed" individuals provides the ecological exposure margin. In the simulation study we take $q_0 = q_1 = q$ with q being one of 0, 0.05, 0.10, 0.20. Table 7 reports the mean-squared error of a number of methods, evaluated over 10,000 simulations. As expected, logistic regression, two phase and finite sample case-control are una ected by ecological exposure misclassification. For the ecological analysis, the effect of exposure misclassification is drastic, while for the hybrid designs, the individual-level data mitigates the exposure misclassification for levels below $q = 0.20$, while for $q = 0.20$, the finite sample case-control analysis is clearly superior. We are currently working on extending the basic method to correct for this form of measurement error, via the introduction of exposure misclassification probabilities. Finally we note that, as with all outcome-dependent sampling schemes, practical issues of selection bias and compatibility of populations should not be forgotten when implementing the hybrid design that we have proposed.

Acknowledgments

This research was supported by grant R01 CA095994 from the National Institutes of Health. The authors would like to thank Jon Wellner for helpful discussions. The authors are also grateful for detailed and constructive comments from an Associate Editor and three referees.

Appendix: Score and information calculations for stratified outcomes

We briefly outline detailed arguments presented in Haneuse (2004). Suppose first that N_{11} were observed. Then the score for the ecological data is given by

$$S^E(\theta) = S_0^E(\theta) + S_1^E(\theta),$$

where $S_z^E(\theta)$ is the ecological score corresponding to the likelihood contribution in stratum z, $z = 0, 1$. Let $S(\phi_{XZ})$ represent the score for ϕ_{XZ} based on the extended hypergeometric likelihood corresponding to $N_{11}|M_{1+}, M_{+1}, N$. Unconditionally we have

$$S^E(\theta, \phi_{XZ}) = \begin{bmatrix} E\left[S^I(\theta)|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \\ E\left[S(\phi_{XZ})|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \end{bmatrix},$$

where the expectations are with respect to the distribution of

$$\text{pr}(N_{11}|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}) = \frac{\text{pr}(N_{11}|M_{1+}, M_{+1}, N)\,\text{pr}(N_{1+0}, N_{1+1}|N_{11}, N, M_{1+}, M_{+1})}{\text{pr}(N_{1+0}, N_{1+1}|N, M_{1+}, M_{+1})},$$

where each term on the right is available. For the hybrid design we similarly have

$$S^H(\theta, \phi_{XZ}) = \begin{bmatrix} E\left[S^I(\theta)|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}, n_{1+z}\right] \\ E\left[S(\phi_{XZ})|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}, n_{1+z}\right] \end{bmatrix}$$

where the expectations are now with respect to

$$\text{pr}\,(N_{11}|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}, n_{1+z}) = \frac{\text{pr}\,(N_{11}|M_{1+}, M_{+1}, N)\,\text{pr}\,(N_{1+0}, N_{1+1}, n_{1+z}|M_{1+}, M_{+1}, N)}{\text{pr}\left(N_{1+0}, N_{1+1}, n_{y1z}|M_{1+}, M_{+1}, N\right)}.$$

If N_{11} were observed then the observed information associated with the ecological likelihood is given by

$$I^E(\theta) = I_0^E(\theta) + I_1^E(\theta),$$

where $I_z^E(\theta)$ corresponds to the ecological information given stratum z, $z = 0, 1$. Unconditionally we have

$$I^E(\theta, \phi_{xz}) = \begin{bmatrix} E\left[I^E(\theta)|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \\ E\left[I(\phi_{xz})|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \end{bmatrix} - \begin{bmatrix} \text{var}\left[S^E(\theta)|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \\ \text{var}\left[S(\phi_{xz})|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}\right] \end{bmatrix},$$

where the expectation is with respect to $N_{11}|N_{1+0}, N_{1+1}, N, M_{1+}, M_{+1}$, and $I(\phi_{XZ})$ is the observed information associated with the extended hypergeometric likelihood corresponding to $N_{11}|N$, M_{1+}, M_{+1}. Similar derivations follow for $I^H(\theta)$, except now the expectations are with respect to $N_{11}|N_{1+0}$, N_{1+1}, N, M_{1+}, M_{+1}, n_{1+z}.

References

Breckling J, Chambers R, Dorfman A, Tam S, Welsh A. Maximum likelihood inference from survey sample data. International Statistical Review 1994;62:349–363.

Breslow N, Chatterjee N. Design and analysis of two-phase studies with binary outcomes applied to Wilms' tumor prognosis. Applied Statistics 1999;48:457–468.

Breslow NE, Holubkov R. Maximum likelihood estimation of logistic regression parameters under two-phase, outcome-dependent sampling. Journal of the Royal Statistical Society, Series B, Methodological 1997a;59:447–461.

Breslow NE, Holubkov R. Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data. Statistics in Medicine 1997b;16:103–116. [PubMed: 9004386]

Chambers, R.; Skinner, C., editors. Analysis of Survey Data. John Wiley and Sons; New York: 2003.

Chatterjee N, Carroll R. Semiparametric maximum likelihood estimation exploiting gene-environment independence in case-control studies. Biometrika 2006;92:399–418.

Clayton D, Bernardinelli L, Montomoli C. Spatial correlation in ecological analysis. International Journal of Epidemiology 1993;22:1193–1202. [PubMed: 8144305]

Diaconis P, Sturmfels B. Algebraic algorithms for sampling from conditional distributions. Annals of Statistics 1998;26:363–397.

Dobra, A.; Fienberg, S.; Trottini, M. Assessing the risk of disclosure of confidential categorical data. In: Bernardo, J.; Bayarri, M.; Berger, J.; Dawid, A.; Heckerman, D.; Smith, A.; West, M., editors. Bayesian Statistics. Vol. 7. Oxford University Press; 2003. p. 125-144.

Flanders WD, Greenland S. Analytic methods for two-stage case-control studies and other stratified designs. Statistics in Medicine 1991;10:739–747. [PubMed: 2068427]

Greenland S. Divergent biases in ecologic and individual-level studies. Statistics in Medicine 1992;11:1209–1223. [PubMed: 1509221]

Greenland S. Ecological versus individual-level sources of bias in ecological estimates of contextual health effects. International Journal of Epidemiology 2001;30:1343–1350. [PubMed: 11821344]

Greenland S, Robins J. Ecologic studies – Biases, misconceptions, and counterexamples (with discussion). American Journal of Epidemiology 1994;139:747–771. [PubMed: 8178788]

Haneuse, S. Ph. D. thesis. University of Washington; 2004. Ecological Studies using Supplemental Case-control Data.

Jackson C, Best N, Richardson S. Improving ecological inference using individual-level data. Statistics in Medicine 2006;25:2136–2159. [PubMed: 16217847]

Johnson, N.; Kotz, S. Distributions in Statistics: Discrete Distributions. John Wiley and Sons; New York: 1969.

Judge K, Mulligan J, Benzeval M. Income inequality and population health. Social Science and Medicine 1998;46:567–579. [PubMed: 9460836]

King, G. A Solution to the Ecological Inference Problem. Princeton University Press; Princeton, New Jersey: 1997.

Lasserre V, Guihenneuc-Jouyaux C, Richardson S. Biases in ecological studies: Utility of including within-area distribution of confounders. Statistics in Medicine 2000;19:45–59. [PubMed: 10623912]

Lawless J, Kalbfleisch J, Wild C. Semiparametric methods for response-selective and missing data problems in regression. Journal of the Royal Statistical Society, Series B 1999;21:413–438.

Little, R.; Rubin, D. Statistical analysis of missing data. Vol. Second ed.. John Wiley and Sons; New York: 2002.

Maheswaren R, Morris S, Falconer S, Grosshino A, Perry I, Wakefield J, Elliott P. Magnesium in drinking water supplies and mortality from acute myocardial infarction in North West England. Heart 1999;82:455–460. [PubMed: 10490560]

Manski CF, Lerman SR. The estimation of choice probabilities from choice based samples. Econometrica 1977;45:1977–1988.

McCullagh, P.; Nelder, JA. Generalised Linear Models. Vol. 2nd Edn.. Chapman and Hall; London: 1989.

Morgenstern, H. Ecological studies. In: Rothman, K.; Greenland, S., editors. Modern Epidemiology. Vol. Second ed.. Lipincott-Raven; 1998. p. 459-480.

Pearl, J. Causality: Models, Reasoning and Inference. Cambridge University Press; Cambridge: 2000.

Plackett R. The marginal totals of a 2 × 2 table. Biometrika 1977;64:37–42.

Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. Biometrika 1979;66:403–411.

Prentice RL, Sheppard L. 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 1990;1:81–97. [PubMed: 2102280]

Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. Biometrika 1995;82:113–125.

Richardson, S.; Monfort, C. Ecological correlation studies. In: Elliott, P.; Wakefield, J.; Best, N.; Briggs, D., editors. Spatial Epidemiology: Methods and Applications. Oxford University Press; Oxford: 2000. p. 205-220.

Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika 1997;84:57–71.

Sheppard L. Insights on bias and information in group-level studies. Biostatistics 2003;4:265–278. [PubMed: 12925521]

Steel, D.; Beh, E.; Chambers, R. The information in aggregate data. In: King, G.; Rosen, O.; Tanner, M., editors. Ecological Inference: New Methodological Strategies. Cambridge University Press; 2004.

Strömberg U, Björk J. Incorporating group-level exposure information in case-control studies with missing data on dichotomous exposures. Epidemiology 2004;15:494–503. [PubMed: 15232411]

van der Vaart, AW. Asymptotic statistics. Cambridge University Press; 1998.

Wakefield J. Sensitivity analyses for ecological regression. Biometrics 2003;59:9–17. [PubMed: 12762436]

Wakefield J. Ecological inference for 2×2 tables (with discussion). Journal of the Royal Statistical Society, Series A 2004;167:385–445.

White JE. A two stage design for the study of the relationship between a rare exposure and a rare disease. American Journal of Epidemiology 1982;115:119–128. [PubMed: 7055123]

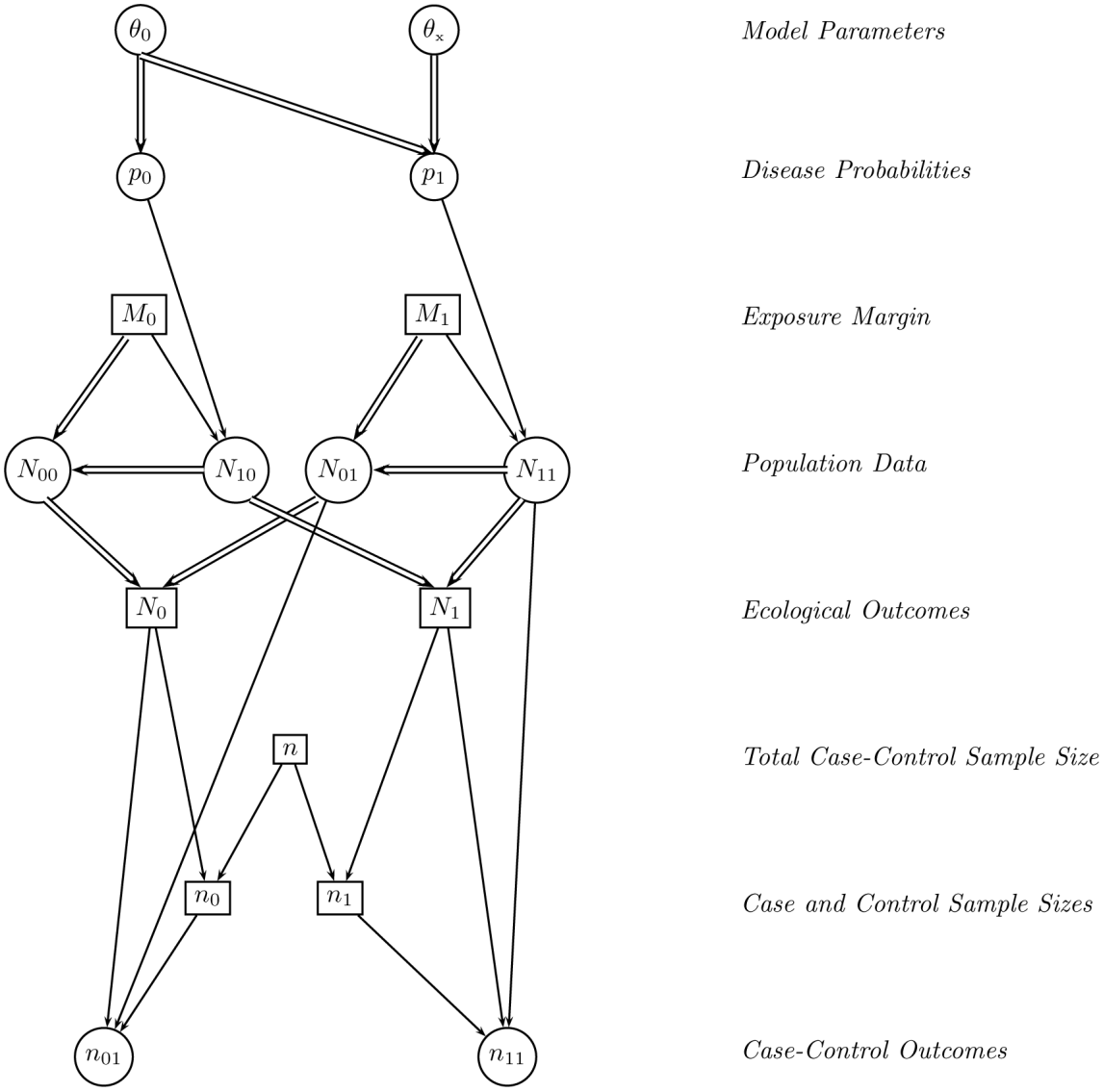
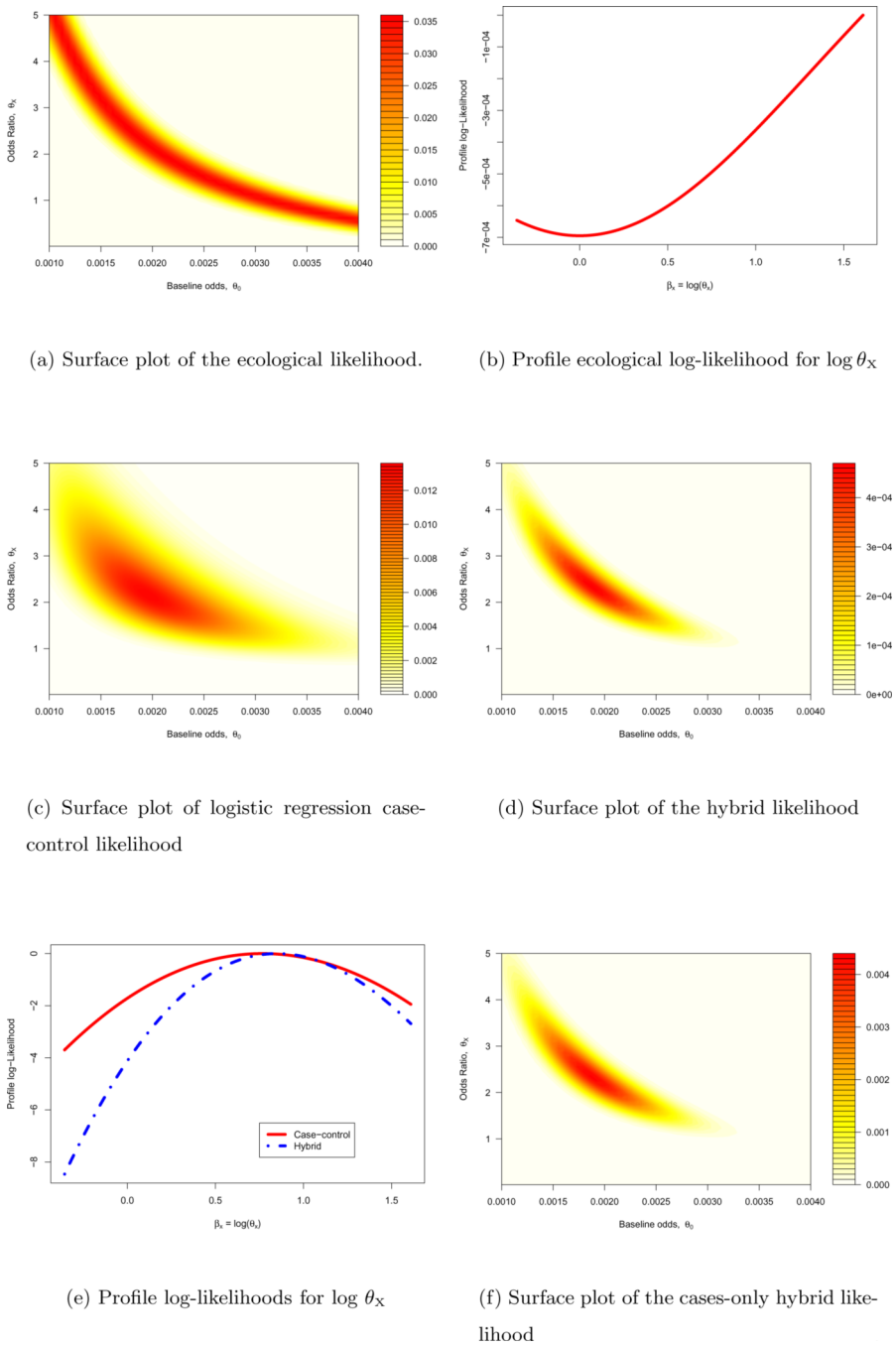


Fig. 1.

Graphical model representation of the hybrid design; conditional independencies are displayed using single line arrows, double line arrows indicate deterministic relationships, and circular and square boxes represent unobserved and observed quantities, respectively.



(a) Surface plot of the ecological likelihood.

(b) Profile ecological log-likelihood for $\log \theta_X$

(c) Surface plot of logistic regression case-control likelihood

(d) Surface plot of the hybrid likelihood

(e) Profile log-likelihoods for $\log \theta_X$

(f) Surface plot of the cases-only hybrid likelihood

Fig. 2.

Likelihood plots for a single area with $(N_1, M_0, M_1) = (125, 20000, 20000)$ and $(n_0, n_1, n_{01}, n_{11}) = (50, 50, 26, 35)$. In (a), (c), (d) and (f) the likelihood surfaces are for θ_0, the baseline odds, and θ_X, the odds ratio; in (b) and (e) the profile log-likelihoods are for $\beta_X = \log \theta_X$.

Table 1

Ecological and case-control data with a binary exposure in a generic area; Y is a disease indicator and X a binary exposure. In an ecological study N_{10} and N_{11} are unobserved.

	Ecological				*Case-control*		
	$Y = 0$	$Y = 1$			$Y = 0$	$Y = 1$	
$X = 0$		N_{10}	N_0	$X = 0$			
$X = 1$		N_{11}	M_1	$X = 1$	n_{01}	n_{11}	
	N_0	N_1	N		n_0	n_1	n

Table 2

Simulation results for θ_X, true value is 2.00; relative efficiencies are calculated with respect to the hybrid design.

	$x \in (0.2,0.8)$ $n = 20$		$x \in (0.4,0.6)$ $n = 20$		$x \in (0.2,0.8)$ $n = 50$		$x \in (0.2,0.8)$ $n = 10$	
	Est. (% Bias)	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**
Ecological	2.02 (1.2)	59.4	2.27 (13.4)	17.1	2.03 (1.6)	38.5	2.03 (1.3)	78.6
LR[1]	2.15 (7.4)	14.5	2.33 (16.6)	22.1	2.07 (3.5)	26.9	2.35 (17.5)	6.7
Two-Phase	2.05 (2.7)	23.8	2.13 (6.5)	35.8	2.04 (1.7)	38.5	2.11 (5.7)	14.3
FSCC[2]	2.02 (1.2)	38.2	2.09 (4.5)	73.2	2.02 (1.0)	57.1	2.07 (3.7)	20.4
Hybrid	2.01 (0.6)	100.0	2.08 (3.8)	100.0	2.02 (0.8)	100.0	2.02 (1.1)	100.0

Common baseline odds assumed.

[1]Logistic Regression,

[2]Finite Sample Case Control.

Table 3

Ecological and case-control data with a binary exposure, and the outcome stratified by a binary confounder in a generic area. In the study design we consider the ecological data consist of N_{1+0}, N_{1+1}, N and M_{1+}, M_{+1}, while the case-control data are n_{010}, n_{110}, n_{011}, n_{111} and n_{00}, n_{10}, n_{01}, n_{11}.

	Ecological $Z = 0$			*Ecological* $Z = 1$		
	$Y = 0$	$Y = 1$		$Y = 0$	$Y = 1$	
$X = 0$		N_{100}	M_{00}		N_{101}	M_{01}
$X = 1$		N_{110}	M_{10}		N_{111}	M_{11}
		N_{1+0}	M_{+0}		N_{1+1}	M_{+1}
	Case-Control $Z = 0$			*Case-Control* $Z = 1$		
	$Y = 0$	$Y = 1$		$Y = 0$	$Y = 1$	
$X = 0$						
$X = 1$	n_{010}	n_{110}	n_{+10}	n_{011}	n_{111}	n_{+11}
	n_{00}	n_{10}	n_{+0}	n_{01}	n_{11}	n_{+1}

Table 4

Simulation results for stratified outcomes, true values are $\theta_X = \theta_Z = \phi_{XZ} = 2.00$. Common baseline odds assumed.

	θ_X		θ_Z		ϕ_{XZ}	
	Est. (% Bias)	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**
Logistic Regression	2.12 (5.8)	26.9	-	-	-	-
Two-Phase Regression	2.03 (1.3)	42.2	2.01 (0.6)	96.6	-	-
Finite Sample Case Control	1.98 (−1.0)	78.7	2.01 (0.5)	100.0	1.88 (−5.8)	134.0
Hybrid: Cases Only	2.01 (0.7)	92.6	2.01 (0.4)	92.5	2.09 (4.4)	48.9
Hybrid: Full Analysis	2.01 (0.6)	100.0	2.01 (0.5)	100.0	2.05 (2.7)	100.0

Table 5

Simulation results for stratified outcomes with baseline odds varying by areas, true values are $\theta_X = \theta_Z = \phi_{XZ} = 2.00$. Area-specific baseline odds assumed.

	θ_X		θ_Z		ϕ_{XZ}	
	Est. (% Bias)	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**	**Est. (% Bias)**	**Rel Eff**
Logistic Regression	2.11 (5.3)	42.2	-	-	-	-
Two-Phase Regression	2.07 (3.5)	46.5	2.05 (2.5)	81.3	-	-
Finite Sample Case Control	2.03 (1.3)	100.4	2.00 (0.1)	104.6	2.05 (2.5)	98.6
Hybrid: Cases Only	2.03 (1.5)	91.9	2.00 (0.1)	86.3	2.08 (4.0)	47.4
Hybrid: Full Analysis	2.03 (1.3)	100.0	2.00 (0.2)	100.0	2.04 (2.7)	100.0

Table 6

Relative risk estimates for blacks versus whites for the Ohio lung cancer data.

	Race only Fixed baseline odds	Stratified by Gender Fixed baseline odds	Stratified by Gender Area-specific baseline odds
Ecological likelihood	1.50 (1.16, 1.93)	1.63 (1.28, 2.06)	-
Logistic regression	1.62 (0.87, 3.01)	1.15 (0.89, 1.48)	1.15 (0.89, 1.48)
Two-phase regression	1.60 (0.89, 2.85)	1.23 (0.97, 1.57)	1.16 (0.90, 1.49)
Finite sample case-control	1.08 (0.74, 1.58)	1.21 (1.01, 1.46)	1.21 (1.01, 1.46)
Hybrid: full analysis	1.34 (1.07, 1.67)	1.33 (1.14, 1.55)	1.20 (1.00, 1.45)
Hybrid: cases-only	1.34 (1.07, 1.67)	1.30 (1.12, 1.52)	1.16 (0.96, 1.39)
Complete data	1.27 (1.17, 1.37)	1.28 (1.18, 1.38)	1.25 (1.15, 1.36)

Table 7

Mean squared error for various designs, and under different levels of ecological exposure misclassification.

	Exposure misclassification			
	$q = 0.00$	$q = 0.05$	$q = 0.10$	$q = 0.20$
Ecological	0.07	0.15	0.42	4.30
Logistic Regression	0.29	0.30	0.27	0.29
Two-Phase	0.16	0.17	0.16	0.16
FSCC	0.11	0.10	0.09	0.09
Hybrid, case only	0.04	0.06	0.09	0.17
Hybrid	0.04	0.06	0.09	0.17