A hybrid model for reducing ecological bias

RUTH SALWAY*

Department of Mathematical Sciences, University of Bath, Bath, UK
R.E.Salway@bath.ac.uk

JON WAKEFIELD

Departments of Statistics and Biostatistics, University of Washington, Seattle, WA, USA

SUMMARY

A major drawback of epidemiological ecological studies, in which the association between area-level summaries of risk and exposure is used to make inference about individual risk, is the difficulty in characterizing within-area variability in exposure and confounder variables. To avoid ecological bias, samples of individual exposure/confounder data within each area are required. Unfortunately, these may be difficult or expensive to obtain, particularly if large samples are required. In this paper, we propose a new approach suitable for use with small samples. We combine a Bayesian nonparametric Dirichlet process prior with an estimating functions’ approach and show that this model gives a compromise between 2 previously described methods. The method is investigated using simulated data, and a practical illustration is provided through an analysis of lung cancer mortality and residential radon exposure in counties of Minnesota. We conclude that we require good quality prior information about the exposure/confounder distributions and a large between- to within-area variability ratio for an ecological study to be feasible using only small samples of individual data.

Keywords: Aggregate data; Dirichlet process prior; Ecological fallacy; Pure specification bias; Within-area variability.

1. INTRODUCTION

Many disciplines make use of aggregate data, including epidemiology, social sciences, and education; Salway and Wakefield (2004) provide a comparison between models and approaches in epidemiology and social sciences. In environmental and social epidemiology, aggregate data may consist of area-level disease rates and summary exposure and confounder values within each area. Usually, the purpose of such ecological, sometimes called geographical correlation, studies is to make inference at the individual level. The major problem with such a study design is the potential for ecological bias, which is due to aggregation. Many authors have documented the sources of ecological bias, see, for example, Greenland and Robins (1994) and Richardson and Monfort (2000), but far fewer have proposed solutions.

The term “ecological bias” is generally used to describe bias that may arise from several different sources in aggregate data (Greenland and Morgenstern, 1989). We will concentrate on bias that arises...
when aggregating a nonlinear individual-level model over the within-area distribution of covariates. Such bias is caused by within-area variability in both the exposure of interest and the confounders; this will be referred to as “within-area variability bias” (Greenland, 1992, uses the term “pure specification bias”).

Existing approaches to correct for within-area variability bias require individual data on exposures in each area. Since we do not require the individual link between covariate data and health outcome, it is feasible for mortality or morbidity data to be taken from one source and individual data from another. For example, disease counts may be obtained from a cancer registry and individual data from a survey. One of the advantages of ecological studies is that in situations where collecting individual data are either difficult or expensive, it may be possible to obtain only very small samples of individual data. Unfortunately, existing approaches require larger samples; simulations have suggested that samples of at least 100 in each area are required, with more needed if the within-area distributions are significantly skewed (Wakefield and Salway, 2001).

The aim of this paper is to describe a new model that reduces within-area variability bias when only small samples of covariate data are available. The method incorporates prior information about the data, which will be crucial when only very small samples of individual data are available.

The paper is organized as follows: In Section 2, we provide a motivating example concerning the association between lung cancer mortality and residential radon and describe the data we will use subsequently. In Section 3, we look in detail at what causes within-area variability bias in order to understand how it arises and how it may be removed. In particular, we consider how existing approaches perform when we have only small samples of individual data. In Section 4, we describe our new method. The new model is compared with existing methods in a range of simulations; full details are available in the supplementary material (available at Biostatistics online, http://www.biostatistics.oxfordjournals.org), and the main findings are summarized in Section 5. Section 6 presents a practical example using the radon and lung cancer data. Section 7 provides a concluding discussion.

2. MOTIVATING EXAMPLE: LUNG CANCER AND RESIDENTIAL RADON EXPOSURE

To motivate the work that follows, we take an example from environmental epidemiology, an analysis of the relationship between lung cancer and residential radon exposure, using ecological data for counties in Minnesota.

Radon is a naturally occurring radioactive gas, present in rocks and soil. Extensive epidemiological studies on highly exposed groups, such as underground miners, consistently indicate a substantially increased risk of lung cancer at high concentrations of radon (National Academy of Sciences, 1999) and form the basis of the widely used biological effects of ionizing radiation (BEIR) models. However, extrapolating to lower doses, typical of those found in residential homes, is more controversial. Currently, the US Environmental Protection Agency (EPA) action safety level for residential radon is 4 pCi/l (pico-Curies per liter). Many studies that address residential radon have been ecological in design but their usefulness is debated in part due to the problems of interpretation in the presence of ecological bias (Stidley and Samet, 1994). As a result, conclusions from such studies are conflicting, with ecological studies often displaying a negative association between radon and lung cancer. Cohen (1994), for example, concludes that the negative association found by his ecological analysis demonstrates that the BEIR models do not hold for small doses. However, other authors (e.g. Greenland and Morgenstern, 1989; Greenland, 1992; Greenland and Robins, 1994) claim that the model does not fully account for potential ecological bias and that this conclusion is invalid. Much of the methodological interest in ecological studies in epidemiology in the last 20 years has been prompted by this work.

To demonstrate the methods in this paper, we present an ecological analysis of lung cancer and residential radon in Minnesota, combining the ecological data with samples of individual radon measurements in each county. An ecological study is often suited to situations such as this as it may exploit the larger
exposure contrasts available from a large study area; Minnesota includes counties with both very low and very high radon concentrations. The data used in this paper are total lung cancer mortality for 1998–2002 and radon measurements in individual homes in Minnesota collected by the EPA and Minnesota Department of Health in 1987 (data available in Nolan and Speed, 2000). Smoking levels are an important confounder in such an analysis; unfortunately, only state-level estimates of the percentage of smokers are available for Minnesota, so we used a predictive model to estimate county-level deviations from the state prevalence. This model was developed using county-level data from the Behavioral Risk Factor Surveillance System 2004 for North Carolina (http://www.schs.state.nc.us/SCHS/brfss/2004/index.html) and variables on age, sex, median income, and race from the 2000 Census.

The data are taken from different sources, and so an individual study is not possible as the link between radon measurements and mortality is not available. The sample sizes associated with the radon measurements, by county, are very small, ranging between 1 and 122 (median 5), with 2 counties having zero individual radon measurements.

Figure 1 shows a plot of the logarithm of lung cancer rate against mean radon concentration; a weighted least-squares regression, using the area sizes as weights, suggests a slight negative relationship between the two (solid line). A simple Poisson regression of mean radon on lung cancer mortality counts gives a relative risk of 0.90 for a radon concentration of 4 pCi/l compared to 0 pCi/l (i.e. a protective effect); this is unchanged by including the proportion of smokers (an ecological confounder) in the model. Figure 2 illustrates the degree of within-area variability in radon measurements; there is considerable within-area variability, which indicates the potential for considerable ecological bias.

3. ECOLOGICAL MODELS AND ECOLOGICAL BIAS

3.1 Notations

Consider a study area partitioned into a disjoint set of $K$ areas, each containing $n_k$ individuals, with $k = 1, \ldots, K$. Using terminology from epidemiology, let $Y_{ki}$ be a Bernoulli random variable representing the disease outcome of individual $i$ in area $k$, over a specific time period, with $Y_{ki} = 1$ representing a case
and $Y_{ki} = 0$ a noncase, $i = 1, \ldots, n_k$. We are interested in how this outcome is related to an exposure variable of interest, $X_{ki}$.

We begin by specifying the disease/covariate relationship for an individual (following Richardson and others, 1987; Prentice and Sheppard, 1995; Wakefield and Salway, 2001), which can be thought of as the model that we would fit if individual data $\{Y_{ki}, X_{ki}\}$ were available. This approach emphasizes that we are interested in estimates of the individual effect. A common individual-level model in epidemiology is

$$Y_{ki} \mid \beta_0, \beta_1, X_{ki} \sim \text{ind Bern} \{p_I(\beta_0, \beta_1, X_{ki})\},$$

$$p_I(\beta_0, \beta_1, X_{ki}) = \exp(\beta_0 + \beta_1 X_{ki}),$$

(3.1)

where the subscript I emphasizes that $p_I(\cdot)$ characterizes the individual relationship. In (3.1), $\exp(\beta_0)$ is the baseline risk and $\exp(\beta_1)$ is the relative risk corresponding to an increase of one unit in the variable of interest. Since most diseases may be considered rare in a statistical sense, a log link is appropriate here.

Model (3.1) is simple and does not take into account other possible causes of ecological bias, such as confounding or contextual effects, and in addition we have considered only a single continuous exposure variable. However, it is straightforward to extend the individual formulation to multivariate exposures and confounders; the radon example in Section 6 incorporates an additional single binary confounder.

In an ecological study, we typically only have total disease counts, $Y_k = \sum_{i=1}^{n_k} Y_{ki}$, and some summary of the exposure distribution, $X_k$. In this paper, we assume also that we have a sample of individual covariate data of size $m_k$ with $2 \leq m_k \leq n_k$, in each area $k$, and we denote these data by $X_k^{m_k} = \{X_{kj}: j = 1, \ldots, m_k\}$. We will assume that $X_k$ is the mean of the sample, that is, $X_k = \sum_{j=1}^{m_k} X_k^{m_k} / m_k$.

### 3.2 Ecological bias

An obvious choice of ecological model is the individual model (3.1) with individual data replaced by ecological data to give the “simple ecological” model

$$Y_k \mid \beta_0^*, \beta_1^*, X_k \sim \text{ind Po} \{n_k p_I(\beta_0^*, \beta_1^*, x_k)\},$$

$$p_I(\beta_0^*, \beta_1^*, X_k) = \exp(\beta_0^* + \beta_1^* X_k),$$

(3.2)
where $\beta^*_1$ is the ecological effect parameter. Ecological bias arises when the simple ecological model (3.2) does not estimate the same parameters as the individual model (3.1); that is, $\beta^*_1 \neq \beta_1$. Model $p_I(\cdot)$ is a convex function, and bias will occur since, by Jensen’s inequality, we have

$$E[p_I(\beta_0, \beta_1, X_{ki})] \leq p_I(\beta_0, \beta_1, \mu_k),$$

(3.3)

where the expectation is over the within-area exposure distribution and the right-hand side is the simple ecological model (3.2) evaluated at the true area mean $\mu_k$. Equality will occur if and only if $X_{ki} = X_k$ for all $i$, that is, if there is no within-area variation in individual exposures. There is also no bias if $p_I(\cdot)$ is linear in $X$, and so within-area variability bias arises only for a nonlinear model.

In specific circumstances, we may still have no bias. We write

$$E[Y_k|\phi_k] = n_k \exp(\beta_0 + \beta_1 \mu_k) \sum_{r=0}^{\infty} \frac{\beta^r_1}{r!} \mu_k (r),$$

(3.5)

where $\mu_k (r) = E\{(X_{ki} - \mu_k)^r\}$ is the $r$th central moment of the within-area exposure distribution $f_k(\cdot)$. The summation term above is exactly the bias component given by Richardson and others (1987) and consists of terms involving higher moments of the within-area exposure distribution. There will be no bias whenever the summation term is independent of the mean, irrespective of the distribution of $X$; this occurs when the second and higher moments do not depend on $\mu_k$. In general, of course, this will not be the case; this result is of more mathematical than practical interest. However, it does suggest that the mean–variance relationship, and in particular the strength of that relationship, is an important factor in the size of ecological bias since this is the dominant term in (3.5) for $\beta_1 < 1$. This corresponds to a relative risk of less than 2.7, which is typical in studies of environmental pollutants, see the examples in Wakefield (2003). Higher moments become increasingly small and so contribute little to the bias.

### 3.3 Ecological models

One solution to ecological bias is to consider explicitly the model obtained by aggregating over the within-area exposure distribution. If all exposures within an area are assumed independent, then

$$Y_k|\phi_k \sim \text{Po}(n_k p_E(\beta_0, \beta_1, \phi_k)),$$

$$p_E(\beta_0, \beta_1, \phi_k) = \exp(\beta_0 + K(\beta_1)),$$

(3.6)

where $K(\beta_1)$ is the cumulant-generating function of the within-area distribution (Richardson and others, 1987) and the parameters $\beta_0, \beta_1$ are the same as in the individual model (3.1). Expression (3.6) may be interpreted as the “average individual risk” within area $k$ and will generally depend on the mean of $X$ and on higher moments of the distribution, as seen in (3.5). If the exposures are dependent, then $Y_k|\phi_k$ will have the same expectation, but the distribution will no longer be Poisson.

We refer to model (3.6) as the “parametric ecological model.” A convenient assumption, and one that may often be suitable in practice, is that the within-area distributions are approximately normal (the “parametric normal ecological model”), with $X_{ki} \sim N(\mu_k, \sigma^2_k)$, in which case (3.6) becomes

$$E[Y_k|\beta_0, \beta_1, \phi_k] = n_k \exp(\beta_0 + \beta_1 \mu_k + \beta^2_1 \sigma^2_k / 2).$$

(3.7)
In practice, we will need to use estimates of the unknown parameters $\mu_k$ and $\sigma_k^2$. While accurate estimation of the mean may often be possible, information about the within-area exposure variances $\sigma_k^2$ is unlikely to be routinely available. So in practice, a sample of exposure data $X_{mk}^k$ in each area is required to estimate $\sigma_k^2$, and, in particular when the size of the subsample is small, estimation of this variance may result in bias due to errors in variables.

When the within-area exposure distribution is not normal, expression (3.7) can be seen as a second-order approximation to the true model (3.6). Thus, in practice (3.7) may be an adequate approximation provided that within-area exposure distributions are not heavily skewed. For very heavily skewed distributions, or larger exposure effects, expression (3.6) may sometimes be available in closed form (e.g. the gamma distribution given in Wakefield and Salway, 2001). The lognormal distribution is often used to model environmental exposures, such as radon (for a theoretical justification for its use, see Ott, 1994), but cannot be used directly here since the moment-generating function does not exist. In this case, we may use the Taylor expansion (3.5) to provide an approximation when $\beta_1$ is small:

\[
E[Y_k|\phi_k] \approx n_k \exp(\beta_0 + \beta_1 \mu_k)(1 + \beta_1 \sigma_k^2/2 + \beta_1^3 \mu_3/6),
\]

\[
\approx n_k \exp(\beta_0 + \beta_1 \mu_k + \beta_1 \sigma_k^2/2 + \beta_1^3 \mu_3/6),
\]

(3.8)

where $\mu_3$ is the third central moment of the within-area exposure distribution. For the lognormal distribution, $\mu_3 = \sigma_k^2/\mu_k (\sigma_k^2/\mu_k^2 + 3)$ and so depends only on the first 2 moments. In general, however, the use of higher-order moments will introduce further inaccuracies due to the increased instability in estimating higher moments such as the skewness.

Prentice and Sheppard (1995) proposed a model that makes no assumption about the within-area distribution but instead explicitly uses the sample of individual-level data $X_{mk}^k$ to empirically estimate (3.4). For a sample of size $m_k$, we have

\[
E[Y_k|\beta_0, \beta_1, X_{mk}^k] = n_k \hat{\theta}_A k,
\]

(3.9)

where

\[
\hat{\theta}_A k = \frac{1}{m_k} \sum_{j=1}^{m_k} \exp(\beta_0 + \beta_1 X_{kj})
\]

(3.10)

is an estimate of the individual average risk, based on the sample data. Following the terminology of Prentice and Sheppard (1995), we will refer to model (3.9) as the “aggregate model.” It may be fitted using an estimating equations approach, see Prentice and Sheppard (1995) for details. When using any sample of the exposure data, $X_{mk}^k$, rather than all the data within the area, $X_{nk}^k$, the estimating equation in the aggregate approach is biased in expectation over all possible choices of sample; this introduces bias in the estimate of $\beta_1$. For large samples this is negligible, but problems arise with small samples since the finite sampling bias in the estimating equations increases as sample sizes decrease.

Prentice and Sheppard (1995) propose an adjusted estimating equation to correct the finite sampling bias (the “corrected aggregate model”). However, this requires estimation of an additional term and they suggest that in practice the increase in variability will outweigh the benefits. The simulations in Section 3.4 (see also those in Wakefield and Salway, 2001; Sheppard and others, 1996) show that while this corrected version can perform very well for moderately sized samples, for very small samples the estimator becomes unstable, which results in convergence problems for the estimation algorithm. We have observed problems of both nonconvergence and convergence to the wrong value, with nonconvergence rates above 50% as sample sizes decrease below 10.
A hybrid model for reducing ecological bias

7

Fig. 3. The bias of estimators arising from a variety of models, as a function of sample size $m_k$ in each area. The within-area distribution is a normal distribution, and the effect parameter is positive; further details appear in the main text. The shaded region is ±10% bias.

3.4 Using small samples of individual data

To explore the bias in each of these methods when only small samples of individual data are available, we simulated 10 data sets with normally distributed within-area distributions for 100 areas containing 10 000 individuals each. The means and variances are linearly related, with a between- to within-area variance ratio of around 1 and a positive effect parameter $\beta_1 = \log(2)$ (full details are as for the baseline scenario for the simulations in the supplementary material, available at Biostatistics online, http://www.biostatistics.oxfordjournals.org). Figure 3 illustrates the bias in the estimate of $\beta_1$ as a function of sample size. The size of bias is shown for the simple ecological model (3.2), which is always biased, the parametric normal model (3.7), and the uncorrected and corrected aggregate models (3.9).

In this scenario, the ecological bias is away from the null and bias from using small samples is toward the null. This results in overestimation for the simple model and negative biases for the parametric and aggregate models as the sample size decreases. Estimation becomes poorer for both parametric and aggregate approaches as the sample size decreases. In particular, the bias in the uncorrected aggregate and the parametric methods is extremely large for small samples of $m_k < 20$. Note also the variability in the corrected aggregate method, reflecting its increased variance, and the instability when using very small sample sizes.

4. Hybrid ecological model

4.1 The hybrid ecological model

As in (3.1), we assume that the individual disease model is given by $Y_{ki} | \beta_0, \beta_1, X_{ki} \sim \text{ind Bern} \{p_t(\beta_0, \beta_1, X_{ki})\}$, where $p_t(\beta_0, \beta_1, X_{ki}) = \exp(\beta_0 + \beta_1 X_{ki})$. We write $X_{ki} | F_k \sim F_k$, where $F_k$ is the unknown distribution function of $X$ in area $k$, and assume that

$$F_k | a_k, F_0 \sim \text{DP}(a_k, F_0),$$

(4.1)

where DP is a Dirichlet Process, $F_0$ is a known baseline distribution function, and $a_k$ is the strength of belief in $F_0$. Both the unknown distribution $F_k$ and the baseline distribution $F_0$ may include dependence between exposures.

We now derive the implied aggregate disease model when we have a sample $X_{mk}^k$ of exposures from area $k$; that is, the model $Y_k | X_{mk}^k, a_k, F_0$. Following results given in Ferguson (1973), it can be shown
that the mean of the disease counts is given by

\[ E[Y_k|X_k^{mk}, \alpha_k, F_{0k}] = n_k \theta_k = n_k \{ w_k \theta_{0k} + (1 - w_k) \hat{\theta}_{Ak} \}, \]  

(4.2)

a weighted combination of the expectation of \( Y_{ki} \) under the prior distribution \( F_{0k} \)

\[ \theta_{0k} = E[\exp(\beta_0 + \beta_1 X_{ki})|F_{0k}], \]  

(4.3)

and under the aggregate model, based on the sample data

\[ \hat{\theta}_{Ak} = \frac{1}{m_k} \sum_{j=1}^{m_k} \exp(\beta_0 + \beta_1 X_{kj}), \]  

(4.4)

with weights given by

\[ w_k = \frac{\alpha_k}{\alpha_k + m_k}. \]  

(4.5)

This model can be fitted via an estimating equations approach in the same way as the aggregate model (Prentice and Sheppard, 1995). While it is possible to derive an expression for the variance if we assume prior independence for the exposures, this is complex and it is more practical to use a constant working variance and sandwich estimation for inference (as in Prentice and Sheppard, 1995). This is the approach we have used throughout this paper.

Model (4.2) is appealing as it represents a compromise between the parametric and the aggregate models. When sample sizes are small, the sample data are combined with a distributional assumption, borrowing strength from the prior and stabilizing the estimator. When samples are larger, they give more accurate information on the within-area distribution and will adjust inadequacies in the prior information, such as incorrectly specified moments.

### 4.2 Adjustment for small samples

The above model will work well when we have moderate sample sizes and fairly accurate prior information. However, when the samples are small it combines the prior data, which may not be directly comparable to the current data (and so cannot be directly combined), with the covariate data, which produces biased estimates when samples are small. Unless the prior information is very accurate and the \( \alpha_k \) are chosen to be correspondingly large (in which case there is little need for the sample of individual data), estimates will be biased due to finite sampling bias in the estimating equations.

Following the corrected aggregate method, we may use a corrected version of the hybrid model to adjust for the finite sampling bias. We use an adjusted estimating equation similar to the corrected aggregate model which requires estimation of extra terms and we estimate these terms with a combination of both the data and the prior distribution; further details are given in Appendix. The corrected aggregate model provides a consistent estimator of \( \beta_1 \) with asymptotic normality as \( m_k \to n_k \). Since the hybrid model tends to the aggregate model as \( m_k \to n_k, n_k \to \infty \), we also have a consistent, asymptotically normal estimator for the hybrid model (provided the prior has the correct support).

The benefit of the corrected version is that it is less biased for smaller samples than the uncorrected aggregate approach by a factor \((1 - w_k)^2\) (see Appendix for details). In addition, when the samples are small, we combine the prior data with the less biased estimation of the corrected aggregate method. Finally, estimation of the adjustment term which is responsible for the instability in the corrected aggregate approach is made less variable by smoothing it with the prior data.
4.3 Choice of prior distribution

The prior distribution function $F_{0k}$ is the prior expectation of the parametric within-area exposure distribution, and the weights (4.5) show that the precision parameter $\alpha_k$ may be viewed as the sample size associated with the specification $F_{0k}$. The distribution $F_{0k}$ may be obtained from historical data, such as previous census data for demographic exposures or previous years’ pollution measurements for environmental exposures. For example, a number of studies (e.g. Field and others, 2000) have shown that radon concentrations typically follow a lognormal distribution.

Obtaining suitable prior information may not be straightforward. In many cases, if we have reasonably sized samples, the prior distributions could be chosen to be independent normal distributions as a reasonable approximation since the observed data will dominate. Previous simulations with the parametric approach (Wakefield and Salway, 2001) suggest that for mildly skewed distributions, assuming a normal distribution will often give reasonable results, particularly when the exposure effect is small. This suggests that when we have a reasonable amount of individual data, specification of a suitable prior distribution for the hybrid approach should concentrate on obtaining good quality prior information about the first 2 moments of the within-area distributions.

The prior data need not be exactly representative of the exposure of interest; they should capture the relevant characteristics through the mean and the mean–variance relationship. Data at a different level of aggregation could be adjusted via expert opinions to give priors at a relevant aggregation level. Historical data could be used, perhaps adjusting for known trends to remove bias in the mean. For environmental pollutants, we could use monitoring site data to provide prior information about individual exposures. In these situations, we are typically using the prior data to inform about the within-area variation, which is often poorly estimated by the sample data. Of course, if appropriate individual data within each area are available, these should be used to form the samples rather than the prior; the advantage in the prior is to make use of information on similar, but not necessarily identical, exposure measurements.

In situations where sample data are sparse, estimation based on the hybrid approach will be sensitive to the choice of prior distribution and the prior moments. While prior information on exposure means may often be readily available, it may be more difficult to obtain good prior values for within-area variances. In Section 3.2, we saw the importance of the within-area variances and in particular we require a characterization of the within-area mean–variance relationship. This suggests that in specifying $F_{0k}$, it is the relationship between the means and the variances that is important, rather than the variances themselves. Where prior data are available, it may be beneficial to use smoothed versions of the variances, based on the mean–variance relationship. If variances are not available, it may sometimes be possible to specify prior data in terms of within-area means and a functional form for the relationship between the means and the variances. At the very least this provides scope for a sensitivity analysis.

5. Simulations

In this section, we summarize the results of a simulation study which considered a range of scenarios; details are available in the supplementary material (available at Biostatistics online, http://www.biostatistics.oxfordjournals.org). We generated $n_k = 10000$ normally distributed exposures within each of $K = 100$ areas, with area-level variances that increase with the means. This gives a between- to within-area variability ratio in exposure of 1. Corresponding mortality outcomes were based on the individual relationship in (3.1) with a positive exposure effect $\beta_1 = \log(2)$ and samples of individuals of size $m_k = 20$ and $m_k = 5$ taken in each area. To each data set, we fit the simple ecological model, the parametric model, and the corrected and uncorrected aggregate models. We considered 3 hybrid models: Hybridtrue, using the true within-area moments from which the data were generated as the prior values; Hybridgood, using the true moments with a small amount of random normal variation added to represent the situation where we have good, unbiased information; and Hybridoor, with larger random normal variation added to represent the
situation where we have poor information. In each case, we used a normal prior distribution for $F_{0k}$ with a common value of $\alpha_k = 20$.

For a positive effect parameter, the total bias is complex as positive and negative biases from different sources may to some extent cancel each other out. The within-area variability bias is away from the null in this case, whereas the bias introduced from using estimated moments is toward the null. The simulations for $m_k = 20$ showed bias in the parametric model of 11%, and in the uncorrected aggregate model of 6%, with the bias worsening for smaller $m_k$ (28% bias for parametric with $m_k = 5$ and 14% bias for the aggregate). The corrected aggregate model is unstable with the algorithm either failing to converge or converging to the wrong value in 22% of simulations with $m_k = 20$ (for $m_k = 5$, 52%).

In general, the hybrid model performs well, with the bias reduced compared to previous models; the bias is less than 4% for $m_k = 20$. The hybrid model corrected for finite sampling bias has reduction in bias at least as good as the uncorrected version, but with much higher variance, around three times as great. It seems that these increased standard errors better reflect the variability in the data as 95% confidence intervals achieve 95% coverage of the true value. Unfortunately, the corrected hybrid model experiences some convergence problems, although not as serious as the aggregate model (it never converges to the wrong value).

The performance of the hybrid model depends primarily on 3 factors; the between- to within-area variability ratio, the quality of the prior information, and the size of the samples. The extent of the bias depends most strongly on the between- to within-area variance ratio; as this decreases, the bias increases for all models. Previous authors (e.g. Richardson and others, 1987; Prentice and Sheppard, 1995) have also noted that ecological analyses work best when this ratio is large in order to exploit between-area mean contrasts. None of the models perform well when the ratio is less than 0.5, and our simulations suggest that attempting ecological inference for extremely small ratios is not advisable. Ideally, the ratio should be 1 or higher; that is, more variability between areas than within. In this case, even if samples are small we can remove nearly all the bias provided we have good quality prior information.

If the prior information is “perfect” and takes the form of the true moments (i.e. the values from which the data were generated), we obtain extremely good estimates of $\beta_1$, with bias of less than 1%; using good unbiased estimates of the moments gives results nearly as accurate. For smaller samples, the bias is increased and the model is only feasible when the between- to within-area variability ratio is not too small and the prior information is of reasonable quality.

Sample size also affects the performance of the models, and larger samples will partially compensate for poor quality prior information; for example, a poor quality prior with $m_k = 20$ is comparable to a good prior with $m_k = 5$. As sample sizes become much larger ($m_k > 100$), there is nearly no bias, unless the within-area distribution is extremely skewed. We conclude that in the situation of interest, when we only have very small samples, we require a good quality prior and a reasonably large variability ratio for an ecological study to be feasible. Figure 4 illustrates how the bias depends on the sample size; the aggregate models from Figure 3 are shown for comparison.

6. LUNG CANCER AND RESIDENTIAL RADON

In this section, we compare the hybrid model to previous models using the data described in Section 2. The relationship between the within-area means and variances is plotted in Figure 5; the variance increases with the mean, so we will have ecological bias away from the null and measurement error bias from using small samples toward the null. The between- to within-area variability ratio is 0.33, and the simulations summarized in Section 5 suggest that we will require good quality prior information to remove the bias.

We extend the simple individual model in (3.1) to include smoking status as a confounder, with $Z_{ki} = 1$ for a current smoker and $Z_{ki} = 0$ for a nonsmoker:

$$\log p_I(\beta_0, \beta_1, X_{ki}) = \beta_0 + \beta_1 X_{ki} + \beta_2 Z_{ki}.$$
Fig. 4. The bias in estimators from the hybrid and aggregate models as a function of the sample size; this figure is based on the average of 10 simulations at each value of $m_k$. See the main text for details of the 3 hybrid models. The shaded region is $\pm 10\%$ bias.

Fig. 5. Relationship between within-area means and variances, smoother imposed.

Aggregating to the county level and using the lognormal approximation from (3.8) give the ecological model

$$E[Y_k|\beta_0, \beta_1, \beta_2, \phi_k] \approx n_k \{1 + p_{zk} (e^{\beta_2} - 1)\} \exp(\beta_0 + \beta_1 \mu_k + \beta_2 \sigma_k^2/2 + \beta_3 \mu_3/6), \quad (6.1)$$

with $p_{zk}$ the proportion of smokers in county $k$. We have assumed that the ecological variables $\mu_{zk}$ and $\sigma_{zk}^2$ are independent of $p_{zk}$; this can be checked using the data and is plausible in this case. We apply the hybrid model to the radon exposure only, assuming that the confounder part of the ecological model is adequately specified; this is partly because the smoking proportions are only estimated and we do not have individual data available but also since it is poor estimation of the within-area variances that is the
main source of bias, and for a binary confounder the variance is entirely specified by the mean. Model (6.1) assumes that the exposure distribution is the same for smokers and nonsmokers, an assumption for which we do not have data to check.

We fit the simple ecological model, which is a Poisson regression on mean radon and proportion smoking, the parametric lognormal (3.8), the aggregate and the corrected aggregate models, all including smoking as a confounder. For these models, we must exclude the 2 counties which have no individual sample data. For the simple and parametric models, we allowed for overdispersion in the estimates of the standard errors, estimating the overdispersion parameter by dividing the Pearson statistic by the degrees of freedom. As suggested by the simulations in Section 5, we use the corrected hybrid model where it converges and specify a lognormal prior distribution $F_{0k}$ with prior means and variances derived from modeled estimates from the High-Radon Project (work by the Lawrence Berkeley National Laboratory; data are available from http://eetd.lbl.gov/IEP/high-radon/). This gives county-level estimates of the geometric mean and an estimate of the proportion of homes with radon levels greater than 4 pCi/l; we used the lognormal assumption to transform these into estimates of the within-area means and variances. These data are of good quality, although they specifically model long-term concentrations, rather than the short-term concentrations that form our individual data. Note that the hybrid model includes those counties with no sample data; the exposure distribution of these counties will depend entirely on the prior data.

We choose a prior sample size of $\alpha_k = 30$, with corresponding weights $w_k$ on the prior data between 0.2 and 1. This choice of $\alpha_k$ reflects our prior belief that the modeled radon estimates from the High-Radon Project are of good quality and reasonably accurate; a convenient interpretation is that we would regard the prior data to have equal weight with a subsample of size 30. Alternatively, we could choose $\alpha_k$ by specifying that, for example, a sample of 150–200 (an ideal size for the aggregate method with skewed data) should correspond to a weight of 0.85 relative to the prior data; this gives $\alpha_k$ between 26 and 35. To explore the sensitivity of results to different choices, we also try values of $\alpha_k = 20$ and $\alpha_k = 50$.

Table 1 summarizes the results of the different models; for these data the corrected aggregate model fails to converge due to the extremely small sample sizes. The relative risks estimated for the effect of smoking are of the correct order of magnitude. We note that the simple, parametric, and aggregate models all estimate a small protective effect of radon exposure on lung cancer, as has been observed in other ecological studies, although of these only the simple ecological model is significant. Fitting the parametric lognormal model with the prior data (ignoring the sample data entirely) also gives a nonsignificant protective relative risk of 0.71. By contrast, the 3 hybrid models all estimate a small positive relationship between low levels of residential radon and lung cancer, with an excess relative risk of 50–54% at an exposure of 4 pCi/l. The 3 choices for the prior sample size $\alpha_k$ have very little effect. The main difference

<table>
<thead>
<tr>
<th>Model</th>
<th>$\beta_1$</th>
<th>95% CI</th>
<th>$\text{RR}_{\text{radon}}$</th>
<th>$\beta_2$</th>
<th>95% CI</th>
<th>$\text{RR}_{\text{smoking}}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>-0.027</td>
<td>(-0.050, -0.004)</td>
<td>0.90</td>
<td>2.792</td>
<td>(2.522, 3.061)</td>
<td>16.31</td>
</tr>
<tr>
<td>Parametric</td>
<td>-0.029</td>
<td>(-0.074, 0.016)</td>
<td>0.89</td>
<td>1.853</td>
<td>(1.185, 2.520)</td>
<td>6.38</td>
</tr>
<tr>
<td>Aggregate</td>
<td>-0.014</td>
<td>(-0.082, 0.055)</td>
<td>0.95</td>
<td>1.571</td>
<td>(0.951, 2.192)</td>
<td>4.81</td>
</tr>
<tr>
<td>Hybrid ($\alpha_k = 20$)</td>
<td>0.108</td>
<td>(0.042, 0.174)</td>
<td>1.54</td>
<td>1.638</td>
<td>(0.695, 2.582)</td>
<td>5.15</td>
</tr>
<tr>
<td>Hybrid ($\alpha_k = 30$)</td>
<td>0.106</td>
<td>(0.051, 0.160)</td>
<td>1.53</td>
<td>1.779</td>
<td>(1.049, 2.510)</td>
<td>5.93</td>
</tr>
<tr>
<td>Hybrid ($\alpha_k = 50$)</td>
<td>0.101</td>
<td>(0.052, 0.149)</td>
<td>1.50</td>
<td>1.912</td>
<td>(1.296, 2.528)</td>
<td>6.77</td>
</tr>
</tbody>
</table>

CI, confidence interval.
is the slightly smaller standard errors when we place a stronger belief in the prior data, all 3 estimate a significant positive effect.

Although the evidence from the hybrid models is not strong, it does support a small increased risk of lung cancer for low levels of residential radon; the estimated excess risk is of a similar order of magnitude to those estimated by case–control studies (e.g. the Iowa Radon Lung Cancer Study, Field and others, 2000, estimate excess odds of 0.50 for an exposure approximately equal to an average concentration of 4 pCi/l). The most striking result here is the change in sign from a small negative effect for the previous approaches to a small positive effect for the hybrid model. There are several reasons for this. First, the effect estimate is close to zero, with the parametric and aggregate models both including zero in the confidence interval; it is highly plausible that just a small amount of extra information might cause the estimate to fall on the positive rather than the negative side. Second, the model for best comparison with the hybrid is the corrected aggregate, but unfortunately it does not converge so we cannot compare the two. Third, there is a large range in the sample sizes, resulting in a large range for the weights in the hybrid model, which gives rise to a complex interaction between the aggregate and the prior parametric models. Finally, the hybrid model uses the weights and the smoothing at every step of the iterative fitting procedure, stabilizing the algorithm itself as well as the estimate.

This analysis demonstrates that issues such as correct aggregation of the individual model, accounting for the measurement error introduced when using small sample sizes, and the need to make informed judgments about the data (both directly, in the form of good quality prior information, and indirectly, through knowledge about the within-area distributions) are all crucial in ecological analyses and can have substantial effects on resulting estimates. This analysis supports a small positive effect at low doses, and we can conclude that the ecological data supplemented with individual within-area radon measurements are consistent with a detrimental radon–lung cancer effect.

7. Discussion

The key to using ecological data for individual inference is the availability of individual data. In this paper, we have presented a new model that combines ecological and individual data and reduces within-area variability bias in situations when individual data are difficult to obtain and only small samples are available.

The hybrid model can be seen as a compromise between the existing parametric and the corrected aggregate models which both perform poorly for small sample sizes; the former produces biased estimates and the latter is unreliable and unstable. With suitable prior information, the hybrid model can reduce within-area variability bias to less than 3%. A key factor is the between- to within-area variability ratio which ideally should be higher than 1; when this ratio is small, ecological analysis is inadvisable without strong prior information and/or larger samples of individual data. If it is known in advance that the ratio is likely to be small, we can compensate to some extent by collecting more data.

The sample size required for accurate estimation with the hybrid model depends not only on the exposure variability ratio but also on the quality of the prior information and the within-area distribution. If resources are limited, we should choose larger samples for those areas where the prior information is known to be poor, those with larger within-area variability, and those where the exposure distribution is likely to be heavily skewed. In contrast, a nearly homogenous area with good prior information requires little individual data. For example, the radon data used in the example deliberately oversample counties with high radon concentrations, which results in larger samples when the variability is likely to be higher. Figure 4 suggests that a good strategy when data are not heavily skewed is to use the aggregate model for large samples ($m_k > 100$), the corrected aggregate or uncorrected hybrid models for moderate samples ($60 < m_k < 100$), the corrected hybrid model for smaller samples ($20 < m_k < 60$), and the corrected hybrid model with good prior information for small samples ($5 < m_k < 20$). Samples of less than 5 in
every area will rarely be feasible, unless within-area variability is small. The hybrid model can be adjusted for these different strategies, on an area-by-area basis, by altering the prior sample size \( \alpha_k \); setting \( \alpha_k = 0 \) in some areas, for example, will have the effect of discarding prior information and using the sample data exclusively for those areas. Whatever strategy is employed, assessing the sensitivity to the choice of \( \alpha_k \) is recommended.

The hybrid model relies on the presence of good quality prior information about the within-area distributions. The most important aspects of the prior are to have fairly accurate information on the area means and to accurately capture the relationship between the mean and the variance. For very small samples, the model may exhibit problems if the within-area distribution is misspecified. Additional simulations not presented here explored the use of a normal prior assumption when the true distribution was lognormal; the results showed extreme instability in the algorithm. This is due primarily to the correction for finite sampling bias, where the extra term required is now estimated from a combination of the highly variable data and the incorrect prior distribution. This suggests that when little is known about the true form of the prior distribution, larger samples will be required in order to use the uncorrected hybrid model.

The use of individual data is essential in overcoming ecological bias, both in the hybrid model and in other recently proposed models (Wakefield, 2004; Jackson and others, 2006). Unfortunately, it can be difficult to obtain suitable data, although nationwide surveys are available in a number of countries. While aggregating to higher levels can increase the within-area samples of individual data, this can have the adverse effect of decreasing the between-to-within-area variability ratio. For example, if the radon–lung cancer analysis is conducted at the state level, the ratio reduces to 0.08, which our simulations suggest is far too small for any ecological analysis to be feasible. One possibility in environmental epidemiology, where samples of individual data can be extremely difficult to obtain, is to make use of the large body of work on modeled exposure surfaces (see, e.g. Zidek and others, 1998; Carlin and others, 1999). More work is required to determine the suitability of using modeled individual exposures and to investigate issues such as model uncertainty, but this may prove to be a valuable source of data, either for individual samples or for providing prior information.

We have not discussed how this model may be extended to deal with multiple continuous exposures and confounders. While it is straightforward to write down a suitable model, in practice fitting this model will be more complex and potentially problematic. Both the individual data and the prior will need to characterize the within-area joint exposure–confounder distribution, capturing the mean–variance–covariance structure for all covariates. One consequence is that individual covariate data may be required on the same set of individuals; for example, combining data from several surveys will not be possible without assuming independence between covariates. Further research is required to investigate the use of the hybrid model in these more complex situations. However, the hybrid model is ideally suited to the semi-ecological study, where individual data are available on the outcome and confounders, with exposure information coming from ecological data.

The model presented here does not account for any residual spatial dependence. However, this could be implemented by embedding the mean model in a Poisson likelihood and adding the usual spatial random effects. The distribution of the disease counts is not strictly Poisson because of dependence introduced by the covariate information; however, for large areas and small covariate subsamples the approximation should prove adequate. This could naturally lead to hierarchical Bayesian versions of the model, where we could additionally incorporate hyperpriors on the prior mean and variances. This might be useful, for example, if prior means and variances are from different sources and one is more reliable than the other. Alternatively, it could be used to formally model a mean–variance relationship when variances are not known. As models become more complex, however, it is likely that larger samples of data will be required to estimate all the components.

Using ecological data for individual inference is problematic. The hybrid model is designed to reduce ecological bias in a situation where otherwise no reliable analysis is possible; it should never be used as
A hybrid model for reducing ecological bias

a substitute for collecting large samples of individual data where this is possible or in place of a well-designed individual study.

ACKNOWLEDGMENTS

The work of the second author was supported by Grant R01 CA095994 from the National Institute of Health. Conflict of Interest: None declared.

APPENDIX

We give details of the hybrid model corrected for finite sampling bias, assuming constant variance of $Y_k$. The approach follows closely that of Prentice and Sheppard (1995) for the corrected aggregate method.

The expectation of the estimating equation for the uncorrected hybrid model is given by

$$E \left[ \sum_{k=1}^{K} D_k^T (y_k - n_k \tilde{\theta}_k) \right] = - \sum_{k=1}^{K} (1 - w_k)^2 \frac{(n_k - m_k)}{m_k(n_k - 1)} (n_k S_k^T - D_k^T \theta_k), \quad (A.1)$$

where $\theta_k$ is the average risk and $\tilde{\theta}_k = w_k \theta_{0k} + (1 - w_k) \hat{\theta}_A k$ is the estimate under the hybrid model given by (4.2). $D_k$ is the $p \times 1$ vector of derivatives, $D_k = \frac{\partial}{\partial \beta} \theta_k$, and $S_k = D_k^T \theta_k$. The summation term in this expression is the same as in the aggregate estimating equation but is given less weight since there is finite sampling bias only in the data part of the model and not from the prior information. So we expect the hybrid approach to generally suffer less finite sampling bias than the aggregate for similar-sized samples (overall, the bias depends on the accuracy of the prior). The bias will be larger for smaller $n_k$.

The correction factor is based on an estimated version of (A.1). However, unlike the corrected aggregate approach where the estimation of this term is highly variable, we can now estimate these terms from a combination of both the data and the prior distribution:

$$\tilde{D}_k = w_k D_{0k} + (1 - w_k) \tilde{D}_A k,$$

$$\tilde{S}_k = w_k S_{0k} + (1 - w_k) \tilde{S}_A k, \quad (A.2)$$

where $D_{0k}$ and $S_{0k}$ are the expected values of $D_k$ and $S_k$ under the prior distribution $F_{0k}$, and $D_A k$ and $S_A k$ are the expected values under the aggregate model using $X_{k}^{mk}$:

$$\tilde{D}_A k = \frac{1}{m_k} \sum_{i=1}^{m_k} \left( \frac{1}{X_{ki}} \right) e^{\beta^T X_{ki}},$$

$$\tilde{S}_A k = \frac{1}{m_k} \sum_{i=1}^{m_k} \left( \frac{1}{X_{ki}} \right) e^{2 \beta^T X_{ki}}. \quad (A.3)$$

For example, for a normal prior distribution $X_{ki} | F_{0k} \sim N(\mu_{0k}, \sigma_{0k}^2)$, we have

$$\theta_{0k} = E[\exp(\beta_0 + \beta_1 X_{ki}) | F_{0k}]$$

$$= \exp(\beta_0 + \beta_1 \mu_{0k} + \beta_1^2 \sigma_{0k}^2 / 2),$$

$$D_{0k} = E \left[ \left( \frac{1}{X_{ki}} \right) \exp(\beta_0 + \beta_1 X_{ki}) | F_{0k} \right]$$
\[ S_{0k} = E \left[ \left( \begin{array}{c} 1 \\ X_{ki} \end{array} \right) \exp \left( 2\beta_0 + 2\beta_1 X_{ki} \right) \bigg| F_{0k} \right] \]

\[ = \left( \mu_{0k} + 2\beta_1 \sigma^2_{0k} \right) \exp \left( 2\beta_0 + 2\beta_1 \mu_{0k} + 2\beta_1 \sigma^2_{0k} \right). \tag{A.4} \]

So the corrected hybrid model is the solution to the corrected estimating equation

\[ \sum_{k=1}^{K} \tilde{D}^r_{Ak} (y_k - n_k \tilde{\theta}_k) + \tilde{m}_k^{-1} (n_k \tilde{S}^r_{Ak} - \tilde{D}^r_{Ak} \tilde{\theta}_{Ak}), \tag{A.5} \]

with

\[ \tilde{m}_k = m_k (n_k - 1) - (1 - w_k)^2 (n_k - m_k) \]

\[ (1 - w_k)^2 (n_k - m_k) \]

which is unbiased in expectation over all possible choices of subsample. This form may be used with a constant working variance matrix and sandwich estimation for empirical standard errors.

REFERENCES


A hybrid model for reducing ecological bias


[Received December 22, 2005; first revision August 17, 2006; second revision January 25, 2007; third revision March 2, 2007; accepted for publication April 24, 2007]