

# *Sources of bias in ecological studies of non-rare events*

RUTH SALWAY<sup>1</sup> and JONATHAN WAKEFIELD<sup>2</sup>


<sup>1</sup>*Department of Mathematical Sciences, University of Bath, Bath, UK*  
*E-mail: RESalway@both.ac.uk*

<sup>2</sup>*Departments of Statistics and Biostatistics, University of Washington, Seattle, USA*

---

Ecological studies investigate relationships at the level of the group, rather than at the level of the individual. Although such studies are a common design in epidemiology, it is well-known that estimates may be subject to ecological bias. Most discussion of ecological bias has focused on rare disease events, where the tractability of the loglinear model allows some characterization of the nature of different biases. This paper concentrates on non-rare events, where the Poisson approximation to the binomial distribution is not appropriate. We limit the discussion to bias that arises from within-area variability in exposures and confounders. Our aims are to investigate the likely sizes and directions of bias and, where possible, to suggest methods for controlling the bias or for addressing the sensitivity of inference to assumptions on the nature of the bias. We illustrate that for non-rare events it is much more difficult to characterize the direction of bias than in the rare case. A series of simple numerical examples based on a chronic study of respiratory health illustrate the ideas of the paper.


*Keywords:* aggregate data, air pollution, confounding, ecological fallacy, within-area variability

1352-8505 © 2005  Springer Science + Business Media, Inc.

---

## 1. Introduction

Ecological studies are studies in which data are available for groups only, rather than for the individuals within the groups. Such studies arise in many disciplines, including epidemiology (Morgenstern, 1995), political science (King, 1997), sociology (Goodman, 1953; Goodman, 1959), geography (Openshaw, 1984), economics and education; see Salway and Wakefield (2004a) for a discussion of the links between the aims and methods of analysis in epidemiology and the social sciences. In this paper we will concentrate on the use of ecological studies in environmental epidemiology, where typically the emphasis is on making individual inference about causal relationships between a disease outcome and an environmental exposure, in the presence of confounding. Ecological studies are particularly appealing in this context since individual exposures to environmental factors, such as pollutants in air

1352-8505 © 2005  Springer Science + Business Media, Inc.

or water, are difficult and expensive to obtain; pollution measures taken from monitoring sites may be routinely available, however. It is well-known that the interpretation of ecological studies is problematic. Fundamental to the problem is that relationships at an ecological level cannot be assumed to hold for the individuals within the groups. The difference between the expectation of an estimator from an ecological study and the individual-level parameter of interest is known as *ecological* or *cross-level bias* (with incorrect inference being referred to as the ecological fallacy), and has been discussed in detail by a number of authors including Piantadosi *et al.* (1988), Greenland and Morgenstern (1989), Greenland (1992) and Greenland and Robins (1994).

Previous work (for example, Richardson *et al.*, 1987, Richardson, 1992; Plummer and Clayton, 1996; Richardson and Montfort, 2000; Wakefield and Salway, 2001; Wakefield, 2003) has focused on investigating ecological bias for studies involving rare disease events, with a loglinear Poisson model being used for inference. This has led to an improved understanding of the underlying causes of ecological bias, and a number of approaches have been suggested to help compensate for the bias. One approach is to assume a particular individual-level model, make assumptions about the within-area exposure distributions and derive the aggregate form; this may then be fitted to the aggregate data (Richardson *et al.*, 1987; Wakefield and Salway, 2001) if the required data summaries are available. An alternative is to use the derived aggregate model to carry out a sensitivity analysis (Wakefield, 2003).

The purpose of this paper is to extend this work to consider non-rare disease events. Although this situation is less common in epidemiology, it may occur when studying high-risk subgroups; an important example is the prevalence of asthma and wheeze, where childhood prevalence can be greater than 10%. In such cases a logistic model may be appropriate. In particular we wish to answer the following questions:

- How may ecological bias affect estimators of disease risk? For example, can we characterize situations in which ecological estimators over- and underestimate the true exposure effect?
- How does the behavior of estimators in a non-rare setting differ from the behavior when dealing with rare disease events?
- Finally, can we improve estimation from ecological studies of non-rare disease events?

There are a variety of different ways in which ecological bias may arise. These include confounding, including the consideration of within-area variability in exposures and confounders, parameters that vary between areas (effect modification by area), contextual effects, measurement error and mutual standardization. In this paper we will consider only the first of these, since it occurs most frequently in epidemiology, and in practice will often give rise to the largest bias.

We will begin by describing an example of a typical ecological study in Section 2.1, and this will be used to illustrate the main ideas of the paper. Although, as just stated, this paper will concentrate on ecological bias that arises as a result of within-area variability in exposures and confounders when dealing with non-rare disease events, we will briefly describe other sources of potential bias (Section 2.2) and

summarize the main results that apply to rare disease events (Section 2.3). In Section 3 we describe a framework for ecological studies and introduce notation.

The remainder of the paper falls naturally into two halves. Firstly, in Section 4 we will characterize the bias in terms of the size and direction. We are interested particularly in when there is little or no bias, whether we have attenuation of the estimator and how this compares to the results for rare diseases. In Section 4 we will consider a range of distributions for within-area variability. Secondly, in Section 5 we will investigate the extent to which approaches that help compensate for bias in rare disease models may be extended to the non-rare case. These will then be applied to simulated data in Section 6. Section 7 contains a concluding discussion.

## 2. Motivation

### 2.1 *Motivating example*

A typical example of a non-rare disease and the association with an environmental exposure is the study of the chronic effects of air pollution on respiratory health. Such studies tend to be ecological or semi-ecological in design, see for example Gardiner and Crawford (1969); Chinn *et al.* (1981) in the UK, and the Harvard Six Cities Study (Dockery *et al.*, 1993) and American Cancer Society Study (Pope *et al.*, 1995) in the US. In a semi-ecological study individual outcomes and confounders are available, along with an ecological exposure measure. Alternatively, all of the data may be ecological with disease counts stratified by age and gender and population counts available by area. The exposure data are often collected from monitoring sites, or perhaps from a modeled exposure surface (Colville and Briggs, 2000; Zhu *et al.*, 2003). Interpretation is further complicated due to unmeasured confounders, which may typically exhibit much larger effects than the exposure.

Throughout this paper we will use as an example a hypothetical study of the chronic effects of air pollution on respiratory health in children. In this case the outcome is a binary indicator of respiratory problems; such data may be available at various geographical scales, for example, from hospital admissions data. Among children, the incidence of respiratory disease is high, with a prevalence greater than 10% so a loglinear Poisson model is not appropriate. The predictor variables of interest are exposures to airborne pollutants such as sulphur dioxide, ozone, nitrogen dioxide, particulate matter and black smoke. While it is extremely difficult, if not impossible, to accurately measure long-term exposure to pollutants at the level of the individual, monitoring stations may provide data from a number of sites across the study area. In this paper we will assume that such monitoring sites provide an accurate estimate of the average level of pollution in an area. If this is not the case, for example if the monitoring site is placed at a location with high or low pollution level, then estimates will be biased still further.

In practice a study such as this would also need to control for confounding. Standard census-based measures of deprivation, such as the Carstairs Index (Carstairs and Morris, 1991) in the UK, are often used as a proxy for behavioral variables, such as diet, alcohol and smoking and to capture other unmeasured confounders. The use of such ecological control is a serious deficiency of such

studies, since a single area-level variable is attempting to characterize a complex joint distribution of within- and between-area confounders; in particular it will be insufficient to capture within-area variation in confounders.

There are a number of reasons why we use simulated data based on this example rather than genuine data. Real data will be more complex with issues such as measurement error, missing data and random effects which require more sophisticated modeling; the intention in this paper is to keep the examples as simple as possible so it is clear what bias is ecological in nature and what is caused by other factors. Secondly, to determine the success of methods to remove bias, we need to know the 'true' exposure effect, for which individual data are required, and few datasets have such information available.

## 2.2 Sources of ecological bias

Sources of ecological bias include within-area variability, unmeasured confounding, effect modification, contextual effects and measurement error (Greenland, 1992; Wakefield and Salway, 2001).

As with any epidemiological observational study where causality is of interest, unmeasured confounding is a major source of potential bias. In an ecological study, bias due to confounding arises from omitting either within-area (variables measured on individuals) or between-area (variables measured at the group level) confounders. Although identifying and obtaining appropriate data in studies at the level of the individual is not a trivial problem, it is at least well-understood; as with individual studies the solution in the ecological context is to identify such variables and either include them in the model or to control for them at the design stage. In an ecological setting the issues are more complex. A key point is that if we have ecological data on a within-area confounder, for example, the area mean, there will still be bias due to the within-area variability, since the ecological data are not sufficient to fully characterize the within-area distribution.

In general, we would expect effect modification, when the exposure effect varies between areas, to be present. Ecological data do not contain enough information to estimate each parameter separately without additional uncheckable assumptions. Hence it is usual to assume that the variability in effects is small, and include a single effect for all areas.

Contextual effects are area-level variables that affect the individual disease risk in addition to the individual exposure; for example the contextual effect may be the average exposure. While this occurs often in infectious disease epidemiology or social epidemiology (for example, an individual's health may depend both on their own poverty and the average poverty level of those around them), contextual effects in the exposure variable are less common in chronic non-infectious disease epidemiology, although they may be induced by unmeasured confounding (see Sheppard and Wakefield, 2004).

Finally, measurement error may be present. For continuous exposures, problems of measurement error in individual exposures can be reduced by an ecological study (Prentice and Sheppard, 1995). Aggregate measures, such as the average, can be more robust to measurement error and so area-level measurements are more stable and reliable, although systematic measurement error will still cause bias in ecological

estimates. For discrete exposures, non-differential exposure misclassification can cause bias away from the null (Brenner *et al.*, 1992; Greenland and Brenner, 1993), in contrast to individual studies.

We will concentrate on ecological bias that arises as a result of within-area variation in exposures and confounders. Unlike other sources of ecological bias, this is unique to ecological studies and arises when we incorrectly assume that all individuals in an area have the same exposure (or value of the confounder). When the relationship between individual disease risk and exposure is nonlinear, it is tempting to assume that the same relationship holds between the disease rate and the average exposure; however, this is in general not the case. Unless either the underlying risk-exposure model is linear, or all individuals truly have the same exposure, the within-area variability will cause bias in ecological estimates of disease risk. The mean exposures and mean confounders alone are not sufficient to control for this misspecification and the size and direction of the bias will depend on the underlying risk-exposure relationship, the parameters in the model and the within-area joint exposure-confounder distribution.

### 2.3 Ecological bias for rare disease events

We summarize some of the main results derived for loglinear models (Wakefield and Salway, 2001) when the disease events may be considered rare. In further sections we will explore the extent to which these results hold for non-rare disease events.

Firstly there are a number of situations in which there is no pure specification bias. As we have already described, there is no bias when there is no within-area variability; that is when exposures and confounders are the same for all individuals within an area. Secondly, if we suppose exposure only varies within areas, there is no bias if within-area means are independent of all higher moments; in particular this means there is no bias for a normal within-area distribution if the within-area variances are all constant. In practice bias will therefore be small when these conditions hold approximately; that is, when the within-area variation is small, approximately constant or approximately independent of the mean.

If these conditions do not hold, the bias depends on the form of the risk-exposure model, the size of the exposure effect, the amount of within-area variability and the within-area exposure distribution. Previous work with a loglinear disease model has looked at normal, gamma, and uniform within-area exposure distributions, for which the ecological model may be derived explicitly. For all these cases, if within-area variances increase with the means, then the true exposure effect is over-estimated. Thus in a typical application bias from this source is *away from the null*.

For the distributions listed above, one may derive explicitly the ecological model that corresponds to the true underlying individual model. In these cases, it is possible to either collect additional exposure data and fit the derived aggregate model directly (although it will be impossible to test the assumed form of the model when only area-level exposure data are available), or use the derived model as a the basis of a sensitivity analysis. Unfortunately, an exact model for lognormal exposures is not

available (Wakefield and Salway, 2001); however, a gamma distribution may be used and will mimic the mean-variance relationship of the lognormal.

### 3. Statistical framework

#### 3.1 Model

We will use the same statistical framework as that presented in Wakefield and Salway (2001). The general approach is to begin with the underlying individual-level model and then consider how this aggregates to produce an ecological-level model. This is beneficial when trying to link ecological parameters to individual parameters, and in attempting to identify causal relationships.

Suppose we have a study area  $A$  partitioned into a set of  $N$  disjoint areas, with area  $A_k$  containing  $n_k$  individuals,  $k = 1, \dots, N$ . The Bernoulli random variable  $Y_{ki}$  represents the response of individual  $i$  in area  $k$ ,  $k = 1, \dots, N$ ,  $i = 1, \dots, n_k$ ; so  $Y_{ki} = 1$  indicates an individual with respiratory disease, and  $Y_{ki} = 0$  without. In this paper, for notational convenience, we will consider a single exposure variable  $X_{ki1}$  and a single confounder  $X_{ki2}$ ; we also write  $\mathbf{X}_{ki} = (X_{ki1}, X_{ki2})^T$ .

We assume that individual disease status depends on  $\mathbf{X}_{ki}$  through the relationship

$$\begin{aligned} Y_{ki} &\sim_{\text{indep}} \text{Bernoulli}(p_{ki}) \\ E[Y_{ki} | \mathbf{X}_{ki}] &= p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki}), \end{aligned} \quad (1)$$

where  $\beta_0$  and  $\boldsymbol{\beta} = (\beta_1, \beta_2)^T$  are unknown parameters;  $\beta_0$  is a baseline parameter,  $\beta_1$  the parameter of interest associated with the exposure  $X_{ki1}$ , and  $\beta_2$  the nuisance parameter associated with the confounder  $X_{ki2}$ . It is convenient to consider  $\beta_0$  separately from the effect parameters  $\boldsymbol{\beta}$ . For simplicity we have assumed no effect modification by area or by confounder stratum.

For rare disease events a common choice for  $p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki})$  is the loglinear model  $p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki}) = \exp(\beta_0 + \beta_1 X_{ki1} + \beta_2 X_{ki2})$  with emphasis on estimation of the relative risk of disease,  $\exp(\beta_1)$ . In this paper we consider non-rare diseases, in which case a plausible model at the individual level is the logistic model

$$p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki}) = \text{expit}(\beta_0 + \beta_1 X_{ki1} + \beta_2 X_{ki2}), \quad (2)$$

where the function  $\text{expit}(x) = e^x / (1 + e^x)$  is the inverse of the logit function. The parameter of interest is the odds ratio  $\exp(\beta_1)$ . A probit link function is also a possible choice; the probit function is given by

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-\frac{1}{2}u^2} du,$$

and is related to the expit function through the relationship

$$\text{expit}(x) \approx \Phi\left(\frac{16\sqrt{3}}{15\pi}x\right) = \Phi(cx) \quad (3)$$

(Johnson and Kotz, 1970), with  $c = 16\sqrt{3}/(15\pi)$ . For non-rare diseases, logistic and probit link functions behave similarly, and as we will see the probit function can be a useful analytically tractable approximation to the logistic function.

In an ecological study the data consist of the total disease counts in each area,  $Y_k = \sum_{i=1}^{n_k} Y_{ki}$ , and limited information on the exposures and confounders, denoted by  $\phi_k$ . For example, we may have estimates of the means in each area  $\bar{\mathbf{X}}_k = (\bar{X}_{k1}, \bar{X}_{k2})^T$  with  $\bar{X}_{kj} = \sum_{i=1}^{m_k} X_{kij}/m_k$ . If a full census of exposure values is obtained then  $m_k = n_k$ ; alternatively the mean may be evaluated from a subset of the population or, in the context of air pollution, from a set of  $m_k$  monitoring sites. For the ecological response  $Y_k$  we have

$$E_Y[Y_k|\beta_0, \boldsymbol{\beta}, \phi_k] = n_k E_X[p(\beta_0, \boldsymbol{\beta}, \mathbf{x}_{ki})|\phi_k]. \quad (4)$$

Throughout this paper expectations are with respect to  $Y$  unless otherwise stated. In general, when there is non-constant within-area exposure, expression (4) will not be equal to

$$n_k \times p(\beta_0, \boldsymbol{\beta}, \bar{\mathbf{X}}_k). \quad (5)$$

We will refer to this as the *naive ecological model*, where it is assumed that the individual relationship is the same (that is, has the same functional form) as the ecological relationship:

$$E[Y_k|\beta_0^*, \boldsymbol{\beta}^*, \bar{\mathbf{X}}_k] = n_k \text{expit}(\beta_0^* + \beta_1^* \bar{X}_{k1} + \beta_2^* \bar{X}_{k2}). \quad (6)$$

The ecological parameter that is estimated in the naive ecological model is  $\boldsymbol{\beta}^* = (\beta_1^*, \beta_2^*)^T$  and ecological bias corresponds to the difference between  $E[\hat{\boldsymbol{\beta}}^*]$  and  $\boldsymbol{\beta}$ . We are interested in how  $\boldsymbol{\beta}^*$  relates to  $\boldsymbol{\beta}$ ; in particular how  $\beta_1^*$  relates to  $\beta_1$ , and when the naive ecological estimate  $\beta_1^*$  may be used as an unbiased estimate of  $\beta_1$ , the individual parameter of interest. If there is no within-area variability in exposures or confounders, then  $\beta_1^* = \beta_1$  and the naive ecological model may be used to estimate the individual exposure effect. In general we are interested in the size and direction of the bias and how this depends on other factors. In some cases, other parameters may also be of interest, but in this paper we restrict ourselves to the situation where we are interested in estimating the individual exposure effect (odds ratio), controlling for confounders.

To derive the true distribution for the ecological data  $Y_k|\phi_k$  in the absence of the individual exposures, we need to specify the within-area exposure distribution. If we assume that  $\mathbf{X}_{ki}$  are continuous *independent* random variables from the distribution  $f(\cdot|\phi_k)$ , then the induced ecological model is

$$Y_k|\beta_0, \boldsymbol{\beta}, \phi_k \sim_{\text{ind}} \text{Bin}\{n_k, p^*(\beta_0, \boldsymbol{\beta}, \phi_k)\},$$

where

$$\begin{aligned} p^*(\beta_0, \boldsymbol{\beta}, \phi_k) &= E_Y[Y_{ki}|\beta_0, \boldsymbol{\beta}, \phi_k] \\ &= E_X\{E_Y[Y_{ki}|\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki}]\} \\ &= \int p(\beta_0, \boldsymbol{\beta}, \mathbf{x})f(\mathbf{x}|\phi_k)d\mathbf{x} \end{aligned} \quad (7)$$

This can be interpreted as the *average individual risk*. For discrete exposures the integral sign in (7) is replaced by a summation. We will refer to model (7) as the *corresponding ecological model*; that is the aggregated ecological model which corresponds to the specified underlying individual model. To emphasize, ecological bias arises because  $p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki})$  in (1) is not of the same functional form as  $p^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\phi}_k)$  in (7). If the exposures are not independent then (7) is still the average risk but the aggregate outcome is no longer Binomial; for example, for discrete  $\mathbf{X}_{ki}$  and conditioning on the exposure margin the distribution is a convolution of Binomial distributions (see Wakefield, 2004b, and the accompanying discussion). The mean and variance may in general be derived, however, and an estimating functions approach followed (see Wakefield and Salway, 2001), if an appropriate dependence structure can be determined.

An important point to note is that we have derived the ecological model in terms of the underlying parameters of the within-area exposure/confounder distribution,  $\boldsymbol{\phi}_k$ , rather than in terms of the available ecological data, which will generally take the form of estimates,  $\hat{\boldsymbol{\phi}}_k$ ; for example, the average pollution over a number of monitoring sites in an area, rather than the true underlying mean pollution. Furthermore, in a typical application the available data will often consist of estimates of only a subset of the parameters of  $\boldsymbol{\phi}_k$ ; for example, estimates of pollution means may be available, but not within-area variances. In the following discussion we will derive equations in terms of  $\boldsymbol{\phi}_k$ , that is the true parameters. If the ecological data are estimates of *all* the parameters additional bias may be present due to measurement error. In the following discussion we will not address this problem; we will assume that all estimates where available are sufficiently accurate to be interchangeable. If the available data consist of estimates of only a subset of the parameters  $\boldsymbol{\phi}_k$ , then bias will result. We consider explicitly the situation where  $\hat{\boldsymbol{\phi}}_k = \{\bar{\mathbf{X}}_k\}$ , so only estimates of the means are available. This corresponds to the naive regression model (6).

## 4. Ecological bias for non-rare disease events

This section extends the work summarized in Section 2.3 to consider non-rare disease events. We are firstly interested in when there will be no ecological bias present, and how this relates to the conditions for rare disease events. If bias is present, we wish to characterize the bias in terms of its size and direction in different situations. In particular, these situations will depend on possible within-area exposure and confounder distributions, and the size of the true exposure effect,  $\beta_1$ . We wish to know when bias will be small, and in what circumstances it will over- or under-estimate the true effect.

### 4.1 Conditions for no bias

As for rare disease events, there will be no bias if we have a linear risk-exposure model (that is, a model of the form  $p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_{ki}) = \beta_0 + \beta_1 X_{ki1} + \beta_2 X_{ki2}$ ) or if there is no within-area variability in exposures or confounders. These are the only situations

in which we may avoid ecological bias due to within-area variability for non-rare disease events. In particular, even if the within-area variances are constant across areas there will still be bias in the ecological estimates of disease risk, as we illustrate in Sections 4.2–4.4.

In terms of the air pollution example, there will be no bias if all individuals within an area are exposed to the same pollution level (this may occur in the context of other exposures, for example, fluoride, in community intervention studies). In this case there is no missing data and the individual exposure effect may be estimated without bias. Alternatively, if disease risk increases linearly with pollution levels, then the mean pollution in each area will be sufficient to characterize the within-area exposure distribution, whatever that may be. However, in practice both these situations are implausible, although if the relative risk is small a nonlinear form may be adequately approximated by a linear model. Unfortunately, if the relative risk is small the results are likely to be viewed with caution because of the possibility of unmeasured confounding.

#### 4.2 Normal within-area distributions

We have seen that ecological bias will result in the presence of within-area variability. From the integral (7) we can see that the magnitude of such bias will in general depend on the within-area distribution. We will now consider the case where we have a normally distributed exposure and confounder, with  $\mathbf{X}_{ki}|\phi_{k1},\phi_{k2} \sim N(\phi_{k1},\phi_{k2})$ , with  $\phi_{k1}$  the  $2 \times 1$  vector of means, and  $\phi_{k2}$  the  $2 \times 2$  covariance matrix of the within-area exposure-confounder distribution for area  $k$ . The integral in equation (7) is not available in closed form but using the probit approximation to the logistic function (equation (3)) gives

$$p^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\phi}_k) = E[\text{expit}(\beta_0 + \mathbf{X}_{ki}^T \boldsymbol{\beta})] \\ \approx \text{expit} \left\{ (1 + c^2 \boldsymbol{\beta}^T \boldsymbol{\phi}_{k2} \boldsymbol{\beta})^{-1} (\beta_0 + \boldsymbol{\beta}^T \boldsymbol{\phi}_{k1}) \right\} \quad (8)$$

where  $c$  is the constant  $c = \frac{16\sqrt{3}}{15\pi}$ . In the case with a single exposure variable, and no confounder so that  $X_{ki1} \sim N(\phi_{k1}, \phi_{k2})$ , this simplifies to

$$p^*(\beta_0, \beta_1, \boldsymbol{\phi}_k) \approx \text{expit} \left\{ (1 + c^2 \beta_1^2 \phi_{k2})^{-1/2} (\beta_0 + \beta_1 \phi_{k1}) \right\}. \quad (9)$$

As for the rare case the equivalent ecological model depends on the within-area variances as well as the means. The naive ecological model is given by  $\text{expit}(\beta_0^* + \beta_1^* \phi_{k1})$ ; that is, the same form as equation (9) without the factor  $(1 + c^2 \beta_1^2 \phi_{k2})^{-1/2}$ . Ecological bias will be negligible when there is very little within-area variation or the exposure effects are small, since in each of these cases, the term  $\beta_1^2 \phi_{k2}$  will be small (or in the two-dimensional case, when the term  $\boldsymbol{\beta}^T \boldsymbol{\phi}_{k2} \boldsymbol{\beta}$  is close to 0).

These are the same circumstances under which the simple ecological Poisson model performs adequately. However, for rare events there is also no bias when the within-area variances are constant; this is not true for non-rare events, when even constant variances will result in ecological bias. For example, if  $\phi_{k2} = \phi_2$  in equation (9) we have

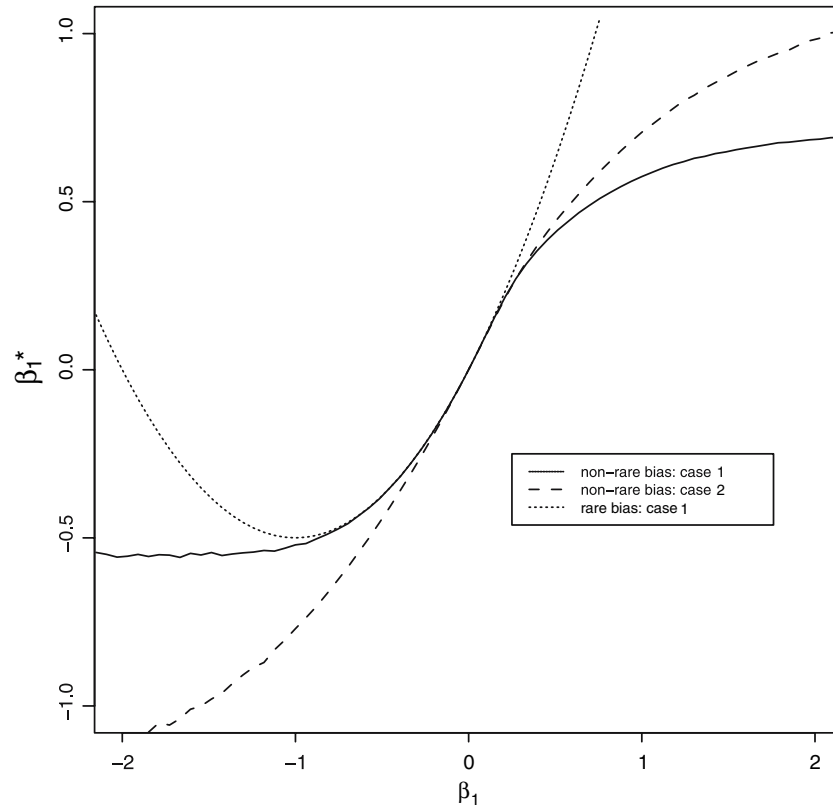
$$\beta_1^* \approx \frac{\beta_1}{(1 + c^2 \beta_1^2 \phi_2)^{1/2}} \leq \beta_1, \quad (10)$$

and there is attenuation.

If the aggregated ecological model involves the within-area variances, as is the case for both loglinear and logistic models, then the variances may be viewed as acting as additional area-level covariates. So fitting the simple ecological model without the variance term may be viewed as analogous to omitting a between-area covariate. If the variances are associated with the means, they are acting as between-area confounders, and omitting them from the model will introduce bias (this is analogous to the usual problem of confounding in individual studies, since the area is the level of analysis). However, if they are constant (or more generally, unrelated to the mean) then they behave as independent covariates. For rare diseases with a loglinear Poisson model, as in individual studies, there is no bias in omitting such a covariate since the variance-effect is absorbed into the intercept and the model does not change its form, and so there is no ecological bias in fitting the simple ecological regression model. However, a non-rare disease with either a logistic or Probit link function behaves like a standard Binomial model with a missing covariate and there will be attenuation (Neuhauser and Jewell, 1993). Greenland *et al.* (1999) describe this type of behavior as *noncollapsibility without confounding*, and note that no bias will result if an estimate of a *population-averaged effect* is required. In terms of ecological analysis, this means that omitting the variances will not result in bias in the marginal effect, that is to say the ecological effect; however, in this paper we are concerned with the situation in which individual effects are of interest, and in this case constant variances will introduce bias into the estimation of such effects.

We are interested in the potential size and direction of the bias in the general case. For simplicity, consider the case of one exposure only, and no confounding, so the corresponding ecological model is approximated by (9). Suppose that there is a linear relationship between the exposure means and variances, so  $E[\phi_{k2}|\phi_{k1}] = a + b \phi_{k1}$ ; we will consider the case where  $b > 0$  so that variances increase with the means, as will typically be the case in an environmental epidemiology scenario. For rare disease events this leads to  $\beta_1^* = \beta_1 + b\beta_1^2/2$  where  $b > 0$  so that for positive  $\beta_1$  we will always have overestimation of a detrimental exposure. Intuitively this overestimation occurs because the variance is acting like an unmeasured confounder that is positively associated with the exposure.

For non-rare disease events the relationship is more complicated. It is difficult to derive even an approximate expression for the relationship between the exposure effect estimate,  $\beta_1^*$ , from the simple ecological regression model and the individual effect parameter,  $\beta_1$ . Figure 1 is based on simulation and shows the bias for two situations. For each value of  $\beta_1$  on the graph, exposures are generated from a normal distribution with means between 1 and 10. In the first case, variances are between 1 and 10 (with a between-area to average within-area variance ratio of 1) and in the second case, variances are between 1 and 5 (with a between-area to average within-area variance ratio of 2); in both cases means and variances increase linearly across areas. Disease incidences are generated via the individual model (2) and the simple ecological model is then fitted to the aggregated ecological data to give an estimate for  $\beta_1^*$ . The two cases are represented by the solid curve and dashed curve respec-



**Figure 1.** Illustration of the bias in the ecological estimate  $\beta_1^*$  versus the individual estimate  $\beta_1$ , for a normally distributed exposure. The solid and dashed lines show the bias in a logistic model with normal within-area exposures for two cases: in the first case, variances are between 1 and 10 (with a between-area to within-area variance ratio of 1) and in the second case, variances are between 1 and 5 (with a between-area to within-area variance ratio of 2); in both cases means and variances increase linearly across areas. The dotted line is the bias for the rare model.

tively, and show the average of these estimates over 10,000 replications. Also shown for comparison is the bias for rare disease events using the Poisson model for the first scenario, given by the dotted curve.

For values of  $\beta_1$  close to zero, the bias is not substantially different to the bias in the Poisson model. However, as  $|\beta_1|$  increases the curves separate and the bias will act very differently. For small positive  $\beta_1$  it is possible to see very slight overestimation of the exposure effect. However the bias is very small in this case, and for larger  $\beta_1$  the bias is towards the null, increasing with the size of  $\beta_1$ . This is the opposite of the behavior when the disease is rare. The dashed line, with the larger between to within variance ratio, generally has smaller bias than the solid line. For large negative values of  $\beta_1$  both cases result in positive estimates for  $\beta_1^*$ ; although it is not shown on the figure, for large positive values of  $\beta_1$  both cases give negative

estimates for  $\beta_1^*$ . So for very large  $\beta_1$  ecological estimates will be in the opposite direction to the true effect.

We will briefly consider two other within-area exposure distributions; in both cases we will assume there is no confounder. First we will look at the bias for uniform within-area exposures, and secondly we will consider lognormal within-area exposures; such a skewed distribution is likely to occur in practice for many environmental exposures, especially air pollution variables.

### 4.3 Uniform within-area distribution

The assumption of uniform exposures has been considered by Greenland (1992) and Wakefield (2003) in the rare case. We assume that  $\mathbf{X}_{ki}|\phi_k \sim_{\text{indep}} U(\mu_k - s_k, \mu_k + s_k)$  so that  $\phi_k = (\mu_k, s_k)$ . For this distribution, evaluation of (7) yields

$$p^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\phi}_k) = \frac{1}{2\beta_1 s_k} \log \left[ \frac{1 + \exp\{\beta_0 + \beta_1(\mu_k + s_k)\}}{1 + \exp\{\beta_0 + \beta_1(\mu_k - s_k)\}} \right] \quad (11)$$

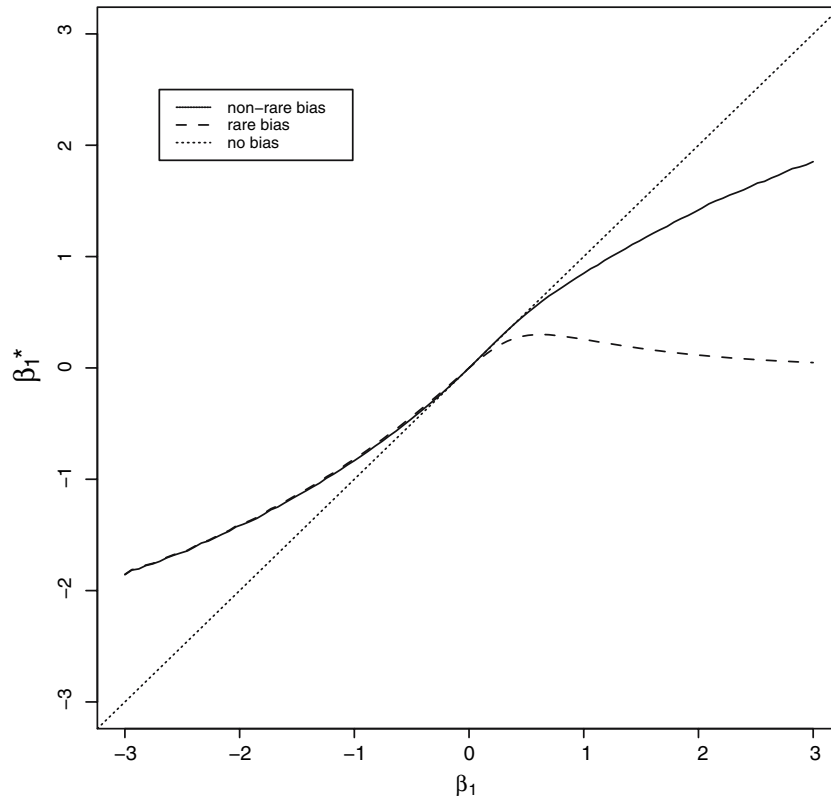
Again, bias will occur even when  $s_k = s$ , that is when the range is constant across areas; in this case there is always attenuation, and the size of the bias increases with the size of  $\beta_1$ , and with  $s$ . Table 1 gives the average estimates of  $\beta_1^*$  (and % bias) for a range of values of  $\beta_1$  and  $s$  in the air pollution context, based on 100 simulations. The baseline parameter was  $\beta_0 = -2$ , and the within-area means range between 20 and 40 with a common within-area variance of  $s^2/3$ . The ratio of the between-area variability to the within-area variability is 33/2, 33/5, 33/10, 33/15 which gives a plausible range for environmental exposures. The bias increases as a function of both the effect size and the variance of the uniform distribution. Such a table may be used to assess the sensitivity of inference for  $\beta_1$  to ecological bias using plausible values of  $s$ . For example, if the odds ratio associated with  $\text{NO}_2$  is truly 2 then if the within-area variance of  $\text{NO}_2$  is 9 ( $s \approx 3$ ), the observed estimate will be 1.6, that is, downwardly biased by 28%. If within-area samples are available then model (11) may be fitted if a uniform distribution is plausible.

The situation is more complicated when  $s_k$  varies between areas. Fig. 2 shows the size of bias in the naive ecological regression model, estimated via simulation as in Section 4.2, for 100 areas with means between 1 and 5, and  $s_k$  varying between 0.2 and 2. The means and variances increase linearly across areas. The dashed line shows

**Table 1.** The entries in the table contain the bias for uniform within-area variability in exposure of  $\text{NO}_2$ .

True $\beta_1$	Variance parameter $s$							
	2	(bias)	5	(bias)	10	(bias)	15	(bias)
$\log(1.2) = 0.182$	0.180	(-1%)	0.172	(-6%)	0.148	(-19%)	0.121	(-33%)
$\log(1.5) = 0.405$	0.392	(-3%)	0.341	(-16%)	0.247	(-39%)	0.174	(-57%)
$\log(2.0) = 0.693$	0.644	(-7%)	0.500	(-28%)	0.320	(-54%)	0.215	(-69%)

There is attenuation in every case.



**Figure 2.** Illustration of the bias in  $\beta_1^*$ , for a uniformly distributed exposure. The solid line shows the bias in the simple ecological model for non-rare diseases, and the dashed line is the bias for rare disease events.

the bias in the rare case for comparison. When exposure variances differ between areas, for rare events the naive model overestimates the effect parameter. For non-rare events there is very slight overestimation for small  $\beta_1$ . For larger exposure effects however, the naive estimate underestimates the true effect, with the bias becoming larger as the true effect parameter  $\beta_1$  becomes larger; for very large values of  $\beta_1$  we will underestimate to such an extent that  $\beta_1^*$  is negative.

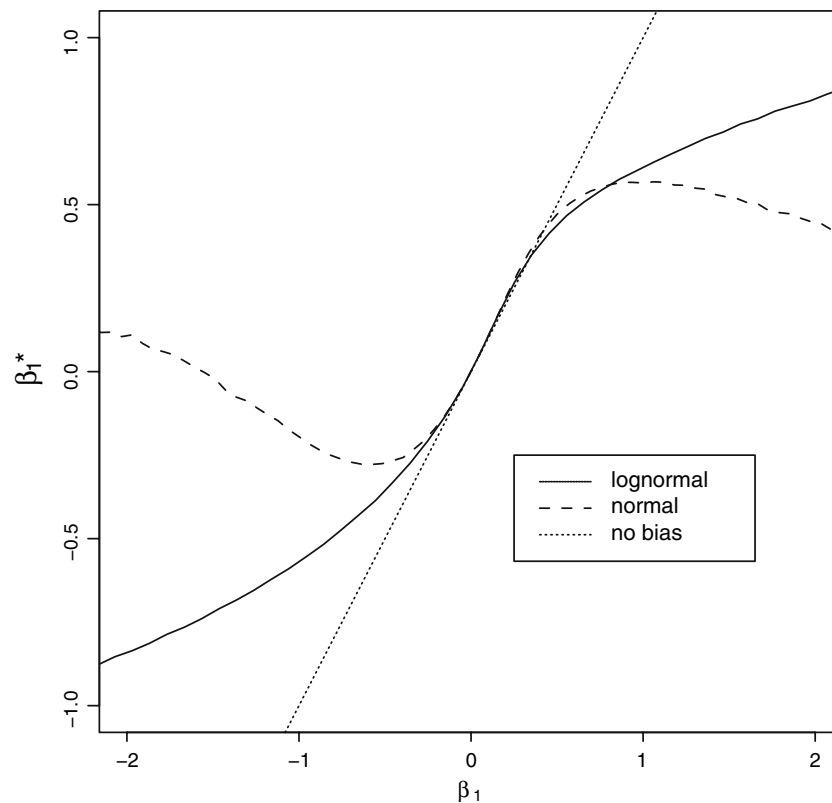
#### 4.4 Lognormal within-area distribution

A common situation for environmental exposures is for within-area distributions to be heavily skewed. We will consider a lognormal within-area exposure distribution;  $X_{ki} \sim \text{LogNormal}(\phi_{k1}, \phi_{k2})$ . For rare diseases the integral (7) takes the form of the moment generating function for the within-area distribution, which does not exist for the lognormal distribution. It can be shown that for the non-rare case equation (7) does converge, although evaluating the integral is problematic.

Figure 3 illustrates the bias by direct simulation, using the same values as in Section 4.2. The solid line is the bias in fitting the simple model when the exposures are lognormal within areas, and the dashed line shows the bias from Section 4.2 when they are normally distributed. For small values of  $\beta_1$ , which would be typical in environmental studies, the bias is larger for the skewed distribution. As the skewness increases, estimates become more biased.

In summary ecological bias for non-rare disease events is more complicated than for rare disease events, where we will over-estimate the true effect in common situations. For all three within-area distributions considered, the exposure effect is slightly over-estimated for small values of  $\beta_1$ , whereas for larger values we will see attenuation to the null. In all cases studied, for very large effect parameters the ecological effect will be estimated in the opposite direction; that is, for a positive individual relationship we may observe a slight protective effect. The point at which this occurs depended on the within-area distribution and on the parameters of the within-area distribution.

Unlike the rare case in which one expects overestimation when the variance increases with the mean, it is more difficult to predict the direction of the bias in the



**Figure 3.** Illustration of the bias in  $\beta_1^*$ , for a lognormal distributed exposure. The solid line shows lognormal distributed exposures, and the dashed line is the bias for normal exposures for comparison. The dotted line is  $\beta_1^* = \beta_1$  corresponding to non bias.

non-rare case. In the rare case, constant spread in the within-area distributions will usually imply no bias; in the non-rare case we obtain attenuation. Only for very small values of  $\beta_1$  is ecological bias likely to be small, and in this situation unmeasured confounding will cast doubt on the validity of the estimate.

#### 4.5 Confounding

In the previous sections we considered a simple situation with a single exposure only and no confounding. Extending the results to consider an exposure and a confounder is not straightforward since the bias in the exposure effect will depend also on the size of the confounder effect, the within-area variation in confounders and the correlation between exposure and confounder. To illustrate this, we will consider the case of a normal distribution for the joint within-area exposure-confounder distribution. Suppose

$$\begin{bmatrix} X_{ki1} \\ X_{ki2} \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_{k1} \\ \mu_{k2} \end{bmatrix}, \begin{bmatrix} \sigma_{k1}^2 & \sigma_{k12} \\ \sigma_{k12} & \sigma_{k2}^2 \end{bmatrix}\right)$$

so  $\phi_k = (\mu_{k1}, \mu_{k2}, \sigma_{k1}^2, \sigma_{k2}^2, \sigma_{k12})$ , then the full model corresponding to equation (8) is

$$\begin{aligned} p^*(\beta_0, \boldsymbol{\beta}, \boldsymbol{\phi}_k) \\ \approx \text{expit}\left\{ [1 + c^2(\sigma_{k1}^2\beta_1^2 + 2\sigma_{k12}\beta_1\beta_2 + \sigma_{k2}^2\beta_2^2)]^{-1}(\beta_0 + \beta_1\mu_{k1} + \beta_2\mu_{k2}) \right\} \end{aligned} \quad (12)$$

where  $c = \frac{16\sqrt{3}}{15\pi}$  as before. So the bias in the estimate  $\beta_1^*$  will depend on the size of the exposure and confounder effects, the within-variation in both exposure and confounder, and the within-area covariances. Note that there will be bias even if  $X_{ki2}$  is *not* a within-area confounder (so  $\sigma_{k12} = 0$ ). This is because although it is not an individual confounder, the within-area variability in  $X_{ki2}$  is still acting as a between-area confounder.

In general, determining the size and direction of the bias becomes much more complex; in particular we may observe over- or under-estimation of the exposure effect at any value of the true effect  $\beta_1$ , even if the average  $\mu_{k2}$  is included. It will be difficult to predict the bias that will be introduced by ignoring within-area variability in exposures and confounders.

## 5. Adjusting for bias

### 5.1 Extended models

If we are able to make an assumption about the within-area exposure distribution, then in some cases it may be possible to directly fit the corresponding ecological model, as derived in Section 4, to obtain unbiased estimates. We need to be able to derive the model explicitly, and have appropriate data available, since these models require additional information about higher moments of the within-area distribution

(so we need good estimates of all the parameters  $\phi_k$ ). For example, if it is reasonable to expect exposures to be normally distributed within areas, then if the within-area exposure variances are available, the corresponding ecological model (9) may be fitted directly from Section 4.2. This may be used as an approximation for small exposure effects and weakly skewed distributions, since Fig. 3 shows that the bias curves are close within this range. Unfortunately it is not possible to derive the corresponding ecological model for lognormal or gamma exposures in closed form, so fitting an appropriate model in cases in which the within-area distributions are heavily skewed is more troublesome, though one may resort to numerical integration.

In terms of inference, the parametric approach uses a full likelihood; for non-rare events the binomial distribution for disease counts is used. The log-likelihood for the disease counts,  $Y_k$ , is

$$l(\boldsymbol{\beta}; Y_k) = Y_k \log\{p(\boldsymbol{\beta}, \boldsymbol{\phi}_k)\} + (n_k - Y_k) \log\{1 - p(\boldsymbol{\beta}, \boldsymbol{\phi}_k)\}$$

and can be maximized using standard maximization techniques, with the estimated Fisher's information matrix being used to obtain standard errors. In practice excess-binomial variation is likely to be encountered (due to unmeasured variables, within-areas variability in exposures/confounders, model misspecification, see Wakefield (2004a) for further discussion), and so a quasi-likelihood or random effects model would be preferred. We have also assumed a binomial distribution which follows if the exposure sampling is independent within areas; strictly speaking the binomial is inappropriate when we have estimated moments but this should not be a big problem if the areas contain a large number of individuals. For a discrete exposure the true likelihood is a convolution, see Wakefield (2004b) and the accompanying discussion and response.

As we mentioned at the end of Section 3.1, we have derived the models, such as equation (9) in terms of the true parameters of the within-area distributions,  $\phi_k$ . In practice, we will have only estimates,  $\hat{\phi}_k$ . This may result in extra bias due to imprecise estimation.

## 5.2 Including individual data

Suppose that in addition to the ecological data,  $Y_k$ , we have observed the covariates  $\mathbf{X}_{ki}$  on a subset of individuals  $m_k$ ;  $2 \leq m_k \leq n_k$  in each area; we write  $\mathbf{X}_k^{m_k} = (\mathbf{X}_{k1}, \dots, \mathbf{X}_{km_k})$ . Prentice and Sheppard (1995) use an estimating functions approach and derive expressions for the mean and variance of  $Y_k | \beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k}$  for a rare disease. We may extend this in the obvious fashion for non-rare diseases.

We define

$$p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k}) = E[Y_{ki} | \beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k}] = \frac{1}{m_k} \sum_{i=1}^{m_k} \text{expit}(\beta_0 + \beta_1 X_{ki1} + \beta_2 X_{ki2}) \quad (13)$$

The mean of  $Y_k$  is given by  $\mu_k = m_k \times p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k})$ , and the variance by

$$\text{var}(Y_k | \beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k}) = n_k p(\beta_0, \boldsymbol{\beta}, \mathbf{X}_k^{m_k}) - \sum_{i=1}^{m_k} \text{expit}(2\beta_0 + 2\beta_1 X_{ki1} + 2\beta_2 X_{ki2})$$

We may then use the estimating function

$$\sum_{k=1}^N \mathbf{D}_k^T(\boldsymbol{\beta}) V_k^{-1}(\boldsymbol{\beta}) \{y_k - \mu_k\} = \mathbf{0}_3,$$

where  $\mathbf{0}_3$  is a  $3 \times 1$  vector of zeroes and  $\mathbf{D}_k^T(\boldsymbol{\beta})$  is the  $3 \times 1$  vector of derivatives. For rare diseases, Prentice and Sheppard (1995) show this estimating equation is asymptotically biased (due to the use of subsamples rather than a complete census) and give an adjustment when  $V_k = 1$ . However, the bias correction requires estimates of the third moment and the instability in the latter means that in some situations the adjustment is not worthwhile (Sheppard *et al.*, 1996).

Unlike the parametric approach, the aggregate approach does not encounter any additional difficulties in implementation when dealing with non-rare rather than rare events. Since it does not require specification of the within-area distribution, if subsample data can be obtained then it is a viable model particularly when the exposures follow non-normal distributions. However, as with the parametric approach, there are problems if the subsample data are subject to non-differential measurement error. For rare diseases with a loglinear model, Prentice and Sheppard (1995) show that the aggregate estimator can help to alleviate bias due to measurement error. However, Carroll (1997) shows that for non-rare diseases measurement error causes the aggregate estimator to overestimate (although a probit link function is used in their paper, we would expect that the same would be true for a logistic link function).

One of the appeals of the aggregate approach for rare diseases is that it lies between truly ecological studies on the one hand, and individual studies on the other. When dealing with rare diseases, it is unlikely that a random sample within an area would record many disease events, so an individual study is still not possible. The aggregate approach allows individual data on exposures to be incorporated whilst still using aggregate data on health outcomes. However, for non-rare disease events, it may be worthwhile to collect both exposure and health data on the sample; if possible this should be incorporated fully into the analysis instead of using the aggregate approach (which does not include the link between individual exposures and health). The likelihood in this case will have two contributions for each area, one from the aggregate and one from the individual-level data.

### 5.3 Sensitivity analysis

When insufficient data are available (such as within-area variances) the models in the previous two sections cannot be fitted, but a sensitivity analysis similar to that described in Wakefield (2003) for a loglinear model can be employed. The approach is to make assumptions about the within-area exposure distribution and derive expressions for the bias in the estimate  $\hat{\boldsymbol{\beta}}_1^*$  as in Section 4. These can then be used to investigate the uncertainty of results.

Unfortunately, this approach is less straightforward for logistic models than for loglinear models. As was seen in Section 4 it is much more difficult to derive useful expressions for the bias in estimates, even using approximations, and for several

important distributions that may occur frequently in practice this is not possible. However, figures similar to those in Section 4 may be plotted by direct simulation to give some idea of the size and direction of bias. Of those distributions considered in Section 4, only normal within-area exposures can be used in a sensitivity analysis with a closed form expression, using equation (9). However, this may be used as an approximation for weakly skewed distributions, since Fig. 3 shows that the bias curves are close within this range. A sensitivity analysis without a closed-form solution will be much more laborious since some form of approximate integration will be needed to evaluate the mean function.

An alternative approach to a sensitivity analysis is to make use of the aggregate approach described in Section 5.2, which uses within-area samples of individual exposure data. To use this idea to explore the sensitivity to within-area distributional assumptions, we may use simulated exposure data rather than real samples. Suppose we have only the average exposure in each area. We fit the naive ecological regression model and obtain a biased estimate of the relative risk. We can assess the extent of the bias by considering samples of simulated individual data. For example, suppose we believe that the within-area exposure distribution is highly skewed. We may generate a sample of lognormal data for each area, with mean given by the actual data, fit the model (13), and compare estimates from this model with estimates from the naive model. By considering different within-area assumptions we may see how estimates are affected by these factors. This approach works because the Prentice and Sheppard model does not require the individual link between exposure and disease status.

## 6. Simulation studies

In this section we describe the use of different estimation techniques/models in three sets of simulations. The first two consider the bias when we have a single exposure only; in the first case exposures are assumed to be normal within areas, and in the second they are lognormal. The third study looks at the bias when there is an exposure and confounder, which follow a bivariate normal within-area joint distribution.

In the context of the respiratory health scenario described in Section 2, we generate data for  $N=100$  areas, with each area containing  $n_k=400$ ,  $k=1,\dots,100$ , boys aged 7–9. In the United Kingdom this corresponds to areas that are larger than electoral wards, but smaller than counties. The response,  $Y_{ki}$ , will represent the presence/absence of wheeze in child  $i$  in area  $k$ . We consider only the subgroup of boys between the ages of 7 and 9 to simplify the analysis, since we may then discount potential confounding due to age and sex. We assume we have a single exposure variable,  $X_{ki1}$ , representing the level of  $\text{NO}_2$  concentration, and a single confounder,  $X_{ki2}$ , nominally representing some measure of deprivation or poverty (for example a measure of diet). The confounder is a continuous within-area confounder and may be viewed as acting as a surrogate for parental smoking, diet and other factors related to socio-economic factors. As discussed in the context of deprivation measures in Section 2.1, such proxy confounders are not ideal for ecological studies; however, unlike deprivation measures such as the Carstairs Index, our  $X_{ki2}$  is a

variable measured at the individual level, and so it is reasonable to think about within-area variation.

### 6.1 Simulation: normal exposures

In the first simulation study we ignore the confounder, and concentrate on the exposure only; we assume a normal distribution for the within-area variability in NO<sub>2</sub> with the variance a linear function of the mean. Individual normal NO<sub>2</sub> levels were generated with means  $15 \leq \phi_{k1} \leq 35$ , and variances  $10 \leq \phi_{k2} \leq 100$ ,  $k = 1, \dots, 100$ , both increasing linearly across areas so that the first area has mean 15 and variance 10. The within-area coefficients of variation range between 20% and 30% which is not unreasonable in environmental studies.

We consider two specific examples with different values of  $\beta_1$ . In the first we assume a small effect of NO<sub>2</sub> with relative risk  $\exp(\beta_1) = 1.1$ , with individual level data generated using this value and  $\beta_0 = -3.5$ . This produces area-level average risks in the range 0.14 to 0.35. In the second scenario we assume a much larger NO<sub>2</sub> effect, with a relative risk of  $\exp(\beta_1) = 1.5$  (and  $\beta_0 = -11$ ).

In each case we compare four different analyses: an ‘‘Individual’’ level analysis that analyses the individual-level data with a logistic regression model; a ‘‘Naive ecological’’ model that takes a binomial logistic regression of the number of cases as a function of the mean exposure; a ‘‘Normal parametric’’ model that fits the approximate induced model given by (8), with first two moments taken as known (the best possible scenario); an aggregate data model based on the mean function (13). For the aggregate model, subsamples of size 100 were used. This represents 25% of the individuals in each area and so will be unrealistically large in many instances. The results, based on 1000 simulations, are presented in Table 2.

As predicted by Figure 1, when the true effect is small the naive ecological model performs reasonably well, with only slight underestimation of the effect of NO<sub>2</sub> and confidence interval coverage (with 91% of intervals containing the true value). In the

**Table 2.** Parameter estimates and standard errors ( $\times 10^{-2}$ ), for various models/estimation procedures for normal within-area exposures.

Assumed model	NO <sub>2</sub> Effect $\beta_1$			NO <sub>2</sub> Odds Ratio	95% CI Coverage
	Estimate	% Bias	Std. Error*		
Truth	0.10			1.10	0.95
Individual	0.10	(0%)	0.14	1.10	0.94
Naive ecological	0.09	(-2%)	0.21	1.10	0.91
Normal parametric	0.10	(0%)	0.23	1.10	0.95
Aggregate	0.09	(-1%)	0.24	1.10	0.91
Truth	0.41			1.50	0.95
Individual	0.41	(0%)	0.42	1.50	0.95
Naive ecological	0.21	(-48%)	0.24	1.24	0.00
Normal parametric	0.39	(-3%)	1.37	1.48	0.94
Aggregate	0.39	(-3%)	1.52	1.48	0.83

second example, the exposure effect  $\beta_1$  is much larger and as a result the bias is much larger when the naive model is used, showing severe attenuation as predicted. The ecological estimate gives an odds ratio for NO<sub>2</sub> of 1.24, far below the true value of 1.5, and the nominal 95% confidence intervals have zero coverage.

The parametric and aggregate approaches both improve the estimate in the case of  $\exp(\beta_1) = 1.5$ . A larger value of  $\beta_1$  causes more variation in the counts,  $Y_k$ , which is more difficult to model at the ecological level, and we see a substantial increase in the standard error of the estimate. This seems to cause additional problems in the parametric model, with only 71% of simulations converging for the large exposure effect. This may be helped with the use of an additional overdispersion parameter in the disease model.

## 6.2 Simulation: lognormal exposures

In the second simulation study, we use the same parameter values and exposure means and variances as above, but generate exposures from a lognormal distribution. These distributions are moderately skewed (with skewness between 0.6 and 0.9), and the within-area coefficients of variation range between 20% and 30%. The results are shown in Table 3.

Results for the naive ecological model are similar to those in the previous section for normal exposures: there is very little bias for the small exposure effect, but significant bias for the larger exposure effect. Of most interest here, however, is the effect of incorrectly assuming a normal within-area exposure distribution in the parametric model. The extent of bias due to the incorrect distributional assumption depends on the exposure effect. For very small exposure effects, there is hardly any bias, although the standard errors do not take into account the extra variability over and above the normal distribution, and consequently the coverage of confidence intervals is reduced. This suggests that for small exposure effects the normal model may be a reasonable approximation.

**Table 3.** Parameter estimates and standard errors ( $\times 10^{-2}$ ), for various models/estimation procedures for lognormal within-area exposures.

Assumed model	NO <sub>2</sub> Effect $\beta_1$			NO <sub>2</sub> Odds Ratio	95% CI Coverage
	Estimate	% Bias	Std. Error*		
Truth	0.10			1.10	0.95
Individual	0.10	(0%)	0.14	1.10	0.95
Naive ecological	0.09	(-4%)	0.21	1.10	0.60
Normal parametric	0.09	(-2%)	0.23	1.10	0.86
Aggregate	0.09	(-1%)	0.26	1.10	0.90
Truth	0.41			1.50	0.95
Individual	0.41	(0%)	0.42	1.50	0.95
Naive ecological	0.21	(-49%)	0.24	1.23	0.00
Normal parametric	0.34	(-16%)	1.00	1.40	0.00
Aggregate	0.39	(-5%)	1.80	1.47	0.76

For the larger exposure effect the additional skewness in the within-area distributions is unaccounted for, and the parametric model assuming normal exposures gives biased results; although estimates are closer to the truth than for the naive ecological regression model, in both cases there is zero coverage of 95% confidence intervals, which indicates that the parametric method should not be used if the exposure distribution is skewed and the exposure effects are not small. Once again the parametric model is somewhat unstable, converging on only 62% of the simulations.

A noticeable feature of the aggregate model is the poor coverage of the confidence intervals, containing the true value only 76% of the time. This is possibly due to the subsample sizes, with a size of 100 being insufficient to fully capture the distribution, particularly in the tails. Consequently, the within-area distributions are estimated to be narrower than is really the case, and this is reflected in a slightly biased estimate and poor estimates of precision. Note that we have already chosen an unfeasibly large subsample size; these results suggest that for highly skewed data and non-rare diseases even larger samples will be required. This is a shortcoming of the aggregate approach, since typically such data will not be available. The need for larger subsamples has been reported in other simulation studies, see for example Guthrie and Sheppard (2001).

### 6.3 Simulation: normal exposure and confounder

Finally, in the third simulation study, we included data on confounders, to see how this might affect the bias. Individual  $\text{NO}_2$  exposures were generated as before, and confounders, with means between 15 and 35, and variances between 3 and 8; all increasing linearly across areas. The true parameters were  $\beta_0 = -9$ ,  $\beta_1 = \log(1.1) = 0.095$  and  $\beta_2 = \log(1.22) = 0.20$ . Hence we see that the confounder effect is greater than the pollution effect. Three sets of data were generated; in the first, exposure and confounder were independent within areas (with a correlation of 0) while in the second and third we introduce positive and negative dependence respectively, by choosing correlations of  $\pm 0.8$ . Positive correlation is obviously the more sensible choice for a confounder representing some measure of poverty, but we include the negative correlation to mimic other situations. When fitting the normal parametric model, based on equation (12), we used three assumptions about the within-area correlation between exposure and confounder: we used values of 0 (so they are assumed to be independent), 0.8 (high positive relationship) and  $-0.8$  (high negative relationship), to see how sensitive inference is to this choice. The results are presented in Table 4.

We are interested specifically in any extra bias that is introduced in the estimate of the exposure effect,  $\beta_1$ , due to within-area variability in confounders. We have chosen a small exposure effect of  $\log(1.1)$ , which showed bias of around 2% in the naive ecological regression when there were no confounders. There are two main questions of interest. Firstly, how much additional bias is introduced in considering the simple ecological model with both exposures and confounder means? Secondly, can we remove this bias, as we did in Section 6.1, by fitting an appropriate parametric model? Also of interest is the importance of the assumption about within-area correlations, corresponding to the three different parametric models fitted.

The naive ecological regression model underestimates the exposure effect in all three cases, despite controlling for exposure and confounder means. In all cases there is more

**Table 4.** Parameter estimates and standard errors ( $\times 10^{-2}$ ), for various models/estimation procedures for normal within-area exposures and confounders.

Assumed model	NO <sub>2</sub> effect $\beta_1$			Odds Ratio	Income effect $\beta_2$		NO <sub>2</sub> 95% CI	Conv. Rate
	Est.	% Bias	Std. Err.*		Est.	Std. Err.*		
Truth	0.10			1.10	0.2		0.95	
True within-area correlation: $\rho = 0$								
Naive ecological	0.09	(-11%)	3.15	1.09	0.18	3.15	0.89	0.99
Aggregate	0.09	(-7%)	1.55	1.09	0.21	1.18	0.89	1.00
Parametric ( $\rho = 0$ )	0.09	(-3%)	1.90	1.10	0.21	1.89	0.05	0.58
Parametric ( $\rho = 0.8$ )	0.10	(8%)	2.17	1.11	0.20	1.67	0.73	0.55
Parametric ( $\rho = -0.8$ )	0.08	(-11%)	2.22	1.09	0.23	1.45	0.95	0.58
True within-area correlation: $\rho = 0.8$								
Naive ecological	0.08	(-14%)	3.17	1.09	0.18	3.17	0.90	1.00
Aggregate	0.09	(-8%)	1.58	1.09	0.21	1.20	0.91	1.00
Parametric ( $\rho = 0$ )	0.07	(-28%)	1.93	1.08	0.23	2.22	0.05	0.58
Parametric ( $\rho = 0.8$ )	0.09	(-5%)	2.23	1.10	0.21	1.85	0.56	0.62
Parametric ( $\rho = -0.8$ )	0.08	(-13%)	2.29	1.09	0.23	1.52	0.92	0.64
True within-area correlation: $\rho = -0.8$								
Naive ecological	0.08	(-13%)	3.12	1.09	0.18	3.13	0.91	0.99
Aggregate	0.09	(-6%)	1.55	1.09	0.21	1.18	0.92	0.99
Parametric ( $\rho = 0$ )	0.08	(-12%)	1.98	1.09	0.22	2.12	0.10	0.58
Parametric ( $\rho = 0.8$ )	0.10	(3%)	2.24	1.11	0.21	1.80	0.69	0.62
Parametric ( $\rho = -0.8$ )	0.09	(-3%)	2.17	1.10	0.22	1.34	0.99	0.53

bias, typically around 12%, than in Section 6.1 where there was no confounding. For such a small exposure effect this translates into an estimate of 0.08 instead of 0.1, which is still reasonable, but with multiple confounders the estimate is likely to be biased still further. Standard errors are all noticeably larger than before, reflecting the increased uncertainty involved with two covariates rather than one. The aggregate data approach produces slightly attenuated estimates in each scenario considered.

The parametric models are an improvement over the naive model, although all three are more biased than in the exposure only case. In all cases the bias is least when the correct assumption is made about within-area correlations, but there is little consistency in the size of bias compared to the extent of correlation misspecification. In particular, assuming no correlation (a convenient assumption when no data are available) results in bias of between 10% and 20% and so in practice does not seem to be a useful assumption, despite its convenience. Finally, these simulations highlight a serious problem with convergence. The parametric models seem highly unstable, and fail to converge just under half the time.

## 7. Discussion

In this paper we have investigated bias due to within-area variability in exposures and confounders when dealing with non-rare disease events. As with non-ecological

data, estimates from logistic models exhibit more complex behavior than linear and loglinear models. It is also more difficult to obtain analytic expressions to aid in bias characterization.

The following is a list of the key differences between loglinear models for rare events and logistic models for non-rare events:

- There is no ecological bias for rare events when the means are independent of higher moments. For non-rare events this is no longer the case. In particular, there is no ecological bias for rare events when the exposure variances are constant. For non-rare events there was attenuation in the cases we examined.
- When exposure variances increase with the means, the naive model for rare events overestimates a true positive effect, for the within-area exposure distributions that we have considered. For non-rare events it either overestimates very slightly for small  $\beta_1$  or underestimates the true effect. When it overestimates the bias is very small and negligible; otherwise there is attenuation.
- The parametric approach for non-rare events requires the use of approximations to derive an ecological model, and this may introduce additional uncertainty not accounted for by the standard errors.
- Standard errors are also likely to be increased when we aggregate, though we have not investigated this in detail.

For the within-area distributions that we have discussed, we have typically seen very little bias for small positive exposure effects, and larger bias, in the form of attenuation towards the null, for larger effect sizes. Unless the effects are very large indeed, which is unlikely in an environmental epidemiology setting, we will not observe a negative effect when the true effect is positive, or vice versa. Although these general trends can be seen from the figures, we have been unable to derive any more useful statements that could be used to quantify possible bias in practice. While these general statements are true when there is a single exposure, the situation becomes increasingly more complex the greater the number of exposures and confounders that we consider. With more covariates in the model there is more within-area variation in the explanatory variables which causes increased standard errors in the estimates and a decrease in power. In particular, there is likely to be increased bias, even for small exposure effects, and this bias could be positive or negative depending on the behavior of the confounders. This is problematic for a typical epidemiology study, such as the air pollution and health example used throughout this paper, where any exposure effect is likely to be very small and there are likely to be many confounders. As we have seen, even if the mean confounders are controlled for in the ecological model, the bias introduced in the exposure estimate may be substantial. For example, exposure effects may be heavily over-estimated, diluted to the extent that they are no longer significant, or even estimated as negative in the presence of within-area variability in known confounders.

We have described two possible approaches that may help to alleviate ecological bias, but neither seem promising for practical applications. The aggregate approach is appealing since it does not require any distributional assumption. However, it is important to have subsamples of a reasonable size; in the simulations we deliberately

used a large sample size of 100. Although in previous simulations for rare diseases (Sheppard *et al.*, 1996; Wakefield and Salway, 2001), this has been shown to be reasonable, our results suggest that, certainly for non-rare diseases, still larger samples will be required for highly skewed distributions. Such sample sizes are clearly impractical in the situation we described with areas of size 400; smaller subsamples will introduce additional bias. Even with our samples, representing 25% of the population, the accuracy of the aggregate approach is reduced still further when including just one extra confounder. With larger geographical areas that contain more individuals we may have individual-level data on a greater number of individuals and the situation may improve.

The parametric approach does not fare any better for practical use. If exposure effects are very small, as is typical, then we may ignore bias introduced by the assumption of normal distributions even when applications such as air pollution suggest exposure distributions are highly skewed. However, even to fit the parametric model assuming normal distributions requires data that will rarely, if ever, be available, in the form of the within-area variances. The models as derived in Section 4.2 are in terms of the *true underlying within-area variances*; even if we have access to samples of data from which to estimate the variances, then unless they are of a reasonable size this will introduce additional measurement error. If large subsamples of data are available to estimate variances, then it would seem better to use the subsamples of data directly in the aggregate approach.

We have not discussed the details of implementing these models, in part because the conclusions of this paper suggest that they will not be useful in practice. In fitting the parametric models we have assumed that the distribution of disease counts is binomial, and used standard likelihood-based methods to fit the models; such models could also be fitted in a Bayesian framework. One reason for not doing so here is to concentrate on ecological bias and not to complicate the discussion with other factors, such as the choice of prior distributions. However, while the distributional assumption is appropriate in this artificial situation when the true underlying within-area means and variances are known, in practice this will not be the case, and the distribution will no longer be binomial. More complicated fitting techniques will be required, for example via estimating equations which only required the first two moments, and it is difficult to see immediately how this would be achieved in a Bayesian setting where the complete likelihood is required. The latter will be complex when the dependence between individual-level outcomes is considered.

Ecological inference when dealing with non-rare events is more complex and unpredictable than for rare events. In a typical environmental application the problems arise not so much from within-area variability in the exposure, where potential exposures effects will be small, but from within-area variability in confounders. This is due to the often larger effect sizes, increased variability and potentially larger numbers of such confounders. Although the general conclusion of this paper is that it will not be possible to correct for ecological bias when dealing with non-rare disease events, a notable exception is in semi-ecological studies. In this case all confounders are measured at the individual level and only a single exposure with a small effect is measured at the ecological level. So estimates will be subject only to bias due to within-area variability in exposures, and not the extra bias that

arises from within-area variability in confounders. As we have seen, for small effects this bias is likely to be small.

We end with the obvious conclusion that the solution to the ecological inference problem is to collect individual-level data, and this is especially true in the non-rare situation. The one advantage here is that since disease events are not rare, *small* samples will yield disease events, and hence information, unlike the rare situation.

## Acknowledgments

The work of the second author was supported, in part, by a grant from the University of Washington Royalty Research Fund. The authors would like to thank Sander Greenland for comments on an earlier draft.

## References

- Brenner, H., Savitz, D.A., Jockel, K.-H., and Greenland, S. (1992) Effects of nondifferential exposure misclassification in ecologic studies. *American Journal of Epidemiology*, **135**, 85–95.
- Carroll, R.J. (1997) Surprising effects of measurement error on an aggregate data estimator. *Biometrika*, **84**, 231–4.
- Carstairs, V. and Morris, R. (1991) *Deprivation and Health in Scotland*, Aberdeen University Press, Aberdeen.
- Chinn, S., Florey, C.d., Baldwin, I.G., and Gorgol, M. (1981) The relation of mortality in England and Wales 1969–73 to measurements of air pollution. *Journal of Epidemiology and Community Health*, **35**, 174–9.
- Colville, R. and Briggs, D. (2000) Dispersion modelling in *Spatial Epidemiology Methods and Applications*, P. Elliott, J.C. Wakefield, N.G. Best and D. Briggs (eds.), Oxford University Press, Oxford, pp. 375–92.
- Dockery, D., Pope, C.A., Xiping, X., Spengler, J., Ware, J., Fay, M., Ferris, B., and Speizer, F. (1993) An association between air pollution and mortality in six U.S. cities. *New England Journal of Medicine*, **329**, 1753–9.
- Gardiner, M.J. and Crawford, M.D. (1969) Patterns of mortality in middle and early old age in the county boroughs of England and Wales. *British Journal of Preventive and Social Medicine*, **23**, 133–40.
- Goodman, L.A. (1953) Ecological regressions and the behavior of individuals. *American Sociological Review*, **18**, 663–4.
- Goodman, L.A. (1959) Some alternatives to ecological correlation. *American Journal of Sociology*, **64**, 610–25.
- Greenland, S. (1992) Divergent biases in ecologic and individual-level studies. *Statistics in Medicine*, **11**, 1209–23.
- Greenland, S. and Brenner, H. (1993) Correcting for non-differential misclassification in ecologic analyses. *Applied Statistics*, **42**, 117–26.
- Greenland, S. and Morgenstern, H. (1989) Ecological bias, confounding and effect modification. *International Journal of Epidemiology*, **18**, 269–74.

- Greenland, S. and Robins, J. (1994) Invited commentary: Ecologic studies – biases misconceptions and counterexamples. *American Journal of Epidemiology*, **139**, 747–64.
- Greenland, S., Robins, J., and Pearl, J. (1999) Confounding and collapsibility in causal inference. *Statistical Science*, **14**, 29–46.
- Guthrie, K. and Sheppard, L. (2001) Overcoming biases and misconceptions in ecological studies. *Journal of the Royal Statistical Society Series A*, **164**, 141–54.
- Johnson, N.L. and Kotz, S. (1970) *Distributions in Statistics Continuous Univariate Distributions. Vol. 2 chapter 22*, Wiley and Sons Ltd., New York.
- King, G. (1997) *A Solution to the Ecological Inference Problem*, Princeton University Press, Princeton, New Jersey.
- Morgenstern, H. (1995) Ecological studies in epidemiology Concepts, principles and methods. *Annual Review of Public Health*, **16**, 61–81.
- Neuhauser, J.M. and Jewell, N.P. (1993) A geometric approach to assess bias due to omitted covariates in generalized linear models. *Biometrika*, **80**, 807–15.
- Openshaw, S. (1984) *Concepts and Techniques in Modern Geography Number 38. The Modifiable Areal Unit Problem*, Geo Books, Norwich.
- Piantadosi, S., Byar, D.P., and Green, S.B. (1988) The ecological fallacy. *American Journal of Epidemiology*, **127**, 893–904.
- Plummer, M. and Clayton, D. (1996) Estimation of population exposure. *Journal of the Royal Statistical Society Series B*, **58**, 113–26.
- Pope, C.A., Thun, M.J., Namboodiri, M.M., Dockery, D.W., Evans, J.S., Speizer, F.E., and Heath, C.W. Jr. (1995) Particulate air pollution as a predictor of mortality in a prospective study of U.S. adults. *American Journal of Respiratory and Critical Care Medicine*, **151**, 669–74.
- Prentice, R.L. and Sheppard, L. (1995) Aggregate data studies of disease risk factors. *Biometrika*, **82**, 113–25.
- Richardson, S. (1992). Statistical methods for geographical correlation studies, in Geographical and Environmental Epidemiology, chapter 17, P. Elliott, J. Cuzick, D. English and R. Stern (eds), Oxford University Press, Oxford.
- Richardson, S. and Montfort, C. (2000). Ecological correlation studies, in Spatial Epidemiology: Methods and Application, P. Elliott, J.C. Wakefield, N.G. Best and D.J. Briggs (eds), chapter 11. Oxford University Press, Oxford.
- Richardson, S., Stucker, I., and Hémon, D. (1987) Comparison of relative risks obtained in ecological and individual studies Some methodological considerations. *International Journal of Epidemiology*, **16**, 111–20.
- Salway, R. and Wakefield, J. (2004a). A comparison of approaches to ecological inference in epidemiology, political science and sociology. in Ecological Inference: New Methodological Strategies, G. King, O. Rosen, and M. Tanner (eds), chapter 14. Cambridge University Press, New York.
- Sheppard, L., Prentice, R.L., and Rossing, M.A. (1996) Design considerations for estimation of exposure effects on disease risk, using aggregate data studies. *Statistics in Medicine*, **15**, 1849–58.
- Sheppard, L. and Wakefield, J. (2005). Discussion of: Statistical issues in studies of the long-term effects of air pollution: The Southern California Children’s Health Study, by Berhane, K., Gauderman, W.J., Stram, D.O. and Thomas, D.C. *Statistical Science*. To appear.
- Wakefield, J. (2003) Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9–17.
- Wakefield, J.C. (2004a) A critique of statistical aspects of ecological studies in spatial epidemiology. *Environmental and Ecological Statistics*, **11**, 31–54.
- Wakefield, J.C. (2004b) Ecological inference for  $2 \times 2$  tables. *Journal of the Royal Statistical Society Series A*, **167**, 385–425.

- Wakefield, J.C. and Salway, R.E. (2001) A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society Series A*, **164**, 119–137.
- Zhu, L., Carlin, B., and Gelfand, A. (2003) Hierarchical regression with misaligned spatial data Relating ambient ozone and pediatric asthma ER visits in Atlanta. *Environmetrics*, **14**, 537–57.

## Biographical sketches

Dr. Jon Wakefield is Professor in the Departments of Statistics and Biostatistics at the University of Washington. He received his bachelor's degree in 1985 and his Ph.D in 1992, both from the University of Nottingham in the United Kingdom. His research interests are in spatial epidemiology, ecological inference and generally, in the modeling of medical data.

Dr. Ruth Salway is a Lecturer in Statistics at the University of Bath in the United Kingdom. She received her Ph.D from Imperial College, University of London (UK) in 2003. Her research interests are in the analysis of ecological studies, with particular interest in their application in epidemiology.