Improving ecological inference using individual-level data

Christopher Jackson[*,†], Nicky Best[‡] and Sylvia Richardson[§]

Department of Epidemiology and Public Health, Imperial College School of Medicine, London, U.K.

SUMMARY

In typical small-area studies of health and environment we wish to make inference on the relationship between individual-level quantities using aggregate, or ecological, data. Such ecological inference is often subject to bias and imprecision, due to the lack of individual-level information in the data. Conversely, individual-level survey data often have insufficient power to study small-area variations in health. Such problems can be reduced by supplementing the aggregate-level data with small samples of data from individuals within the areas, which directly link exposures and outcomes. We outline a hierarchical model framework for estimating individual-level associations using a combination of aggregate and individual data. We perform a comprehensive simulation study, under a variety of realistic conditions, to determine when aggregate data are sufficient for accurate inference, and when we also require individual-level information. Finally, we illustrate the methods in a case study investigating the relationship between limiting long-term illness, ethnicity and income in London. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: ecological inference; ecological bias; aggregate data; survey data; Bayesian hierarchical model

1. INTRODUCTION

Ecological studies analyse data defined at a group level but aim to make inferences about the individuals within the groups. This study design is common in geographical epidemiology and the social sciences, due to the ready availability of data collected at the area level. To make reliable individual-level inferences from these studies, a number of problems

---
[*]Correspondence to: C. H. Jackson, Department of Epidemiology and Public Health, Imperial College School of Medicine, London, U.K.
[†]E-mail: chris.jackson@imperial.ac.uk
[‡]E-mail: n.best@imperial.ac.uk
[§]E-mail: sylvia.richardson@imperial.ac.uk

must be overcome. One crucial difficulty is that the group-level exposure-response relationship may not reflect the individual-level relationship, a problem known as *ecological bias*, or the *ecological fallacy*. See, for example, References [1–4] for discussion of these issues. In particular, ecological inference is theoretically biased when there is a non-linear relationship between the exposure and risk of outcome, and there is within-area variability in the exposure. This is typically the case for the Poisson log-linear models commonly used in epidemiology. As in all observational studies, unmeasured factors may confound the baseline disease risk for groups or the effect of the risk factor under study. Confounders in ecological studies may be unmeasured area-level variables or factors which vary between individuals. Observed exposures are often measured inaccurately, however it has been argued that aggregating exposure data to the group level can help to absorb measurement error [4]. Additionally, as discussed by Greenland [5] and Sheppard [6], using ecological data alone it is difficult to distinguish between relative risks specific to individuals, and *contextual* effects. Contextual effects arise when individual responses are influenced not only by their individual characteristics and behaviours, but by characteristics of other individuals in their area or of the area itself.

Recently there have been many methods proposed to improve ecological inference. Bias can be overcome by accounting for the within-group distribution of exposure data [1, 7, 8]. Wakefield [9] discusses various sources of bias in ecological studies and methods to assess sensitivity to omitted confounders. Lasserre *et al.* [10] show that ecological bias in models for two binary exposures can be reduced by approximating the joint distribution of the two exposures by the product of their marginal distributions. An obvious way to improve ecological inference is to supplement the aggregate data with samples of data at the individual level, collected within the areas. Methods have been devised for incorporating information from individual-level data on the exposure of interest. Prentice and Sheppard [11] describe a semi-parametric model for supplementing ecological data with within-area samples of covariate measurements. Wakefield and Salway [7] present a statistical framework for ecological inference, describing parametric models for supplementing group-level data with individual-level samples of covariates. Best *et al.* [12] present an application of these ideas to geographical incidence of childhood leukaemia, using within-area measurements of environmental benzene exposure.

Ideally, the sample of individual-level data should include both exposures *and response* for selected individuals. This is becoming possible with the increasing availability of sample survey and cohort data on health and demographics. Some U.K.-based examples include the Health Survey for England [13], the Millennium Cohort Study [14] or the Samples of Anonymized Records from the U.K. census (Office for National Statistics). Such data provide direct information about the individual-level link between exposures and response, but little research has been done into statistical models which combine such samples with the group-level information available from ecological data. Detailed geographical information is often unavailable from individual-level data sets for confidentiality reasons, but even when it is available, the data may still lack power to study small-area variations. By combining with aggregate data, power may be increased, while conversely, ecological bias is reduced. In a discussion paper, Wakefield [15] presents a framework for ecological inference on a single binary outcome related to a single binary exposure. It was demonstrated how ecological bias in these situations can be reduced by using subsamples of data from selected individuals consisting of both outcomes and exposures.

The purpose of this paper is to describe how individual-level data can be systematically incorporated in ecological studies and to quantify their benefit under various circumstances. Our motivation is epidemiology, in which there is a need to account for several explanatory or confounding variables when analysing the risk of disease. In epidemiology, there are two main contexts where ecological studies have been used. In social health, the multiple aspects of deprivation and their influence on disease risk and mortality are frequently investigated (e.g. Reference [16]). In environmental epidemiology, 'physical' risk factors such as air pollution, chemical contamination of water or soil, and background radiations are of interest. See, for example, the study of Best *et al.* [12] on benzene exposure and childhood leukaemia.

In Section 2, we describe a general framework for ecological inference. Our hypothetical example contains one continuous covariate, which could represent, for example, pollution or income, and one binary covariate representing an additional confounder such as smoking status. This model accounts for within-area variability of binary and continuous exposures. We use Bayesian hierarchical models, which form a convenient framework for modelling geographically grouped health and environmental data [17].

Section 3 describes an extensive simulation study in which sets of individual responses are generated for several groups conditionally on one binary and one continuous exposure. The simulation is performed under various conditions, and we assess how the accuracy of the inference varies as different models are fitted. We show when ecological data are sufficient for individual-level inference, and when estimates of associations can be improved by including full exposure and outcome data from a small sample of individuals within each area. The improvements in bias and precision are compared under various conditions. We also show that individual data alone lack power, and that combining them with appropriate ecological data can improve mean square error of the estimates. In the simulation, for simplicity, we assume a situation in which there are no contextual effects, but the methods could easily be extended to model area-level baseline risks in terms of area-level variables. In Section 4, we present an example studying the prevalence of limiting long-term illness in London, U.K., in terms of income and ethnicity. The paper is concluded with a short discussion.

2. MODELLING FRAMEWORK

We consider I groups, commonly defined by geographical areas. For each group i, we have the number of disease cases y_i and the total number of individuals N_i. We wish to model the number of disease cases in terms of explanatory variables, or exposures. These can either be continuous, binary or categorical. For simplicity the general model is presented for one binary and one continuous exposure, but the same principles can be used to extend the model straightforwardly to any number of exposures.

2.1. Underlying individual-level model

We assume that the data are generated through the following underlying model. Individual j in group i, with exposure $x_{ij} = (x_{ij}^{(1)}, x_{ij}^{(2)})$, experiences the binary outcome y_{ij} with probability $p_{ij} = p_{ij}(x_{ij})$, where

$$\text{logit}(p_{ij}) = \mu_i + \alpha x_{ij}^{(1)} + \beta x_{ij}^{(2)} \tag{1}$$

$x_{ij}^{(1)}$ is a binary variable, such as smoking status or low social class, and $x_{ij}^{(2)}$ represents a continuous variable, such as pollution exposure or income. μ_i represents the group-specific logit baseline risk, which could be structured in terms of group-level covariates, or exchangeable or spatially-correlated random effects. However, our ecological data only consist of the case count y_i, the population size N_i, the number of individuals $x_i^{(1)}$ exposed to the binary covariate, an estimate of the within-group mean of $x_{ij}^{(2)}$, and possibly an estimate of the corresponding within-group variance. Model (1) can only be directly fitted if we have a sample of individuals from which outcomes y_{ij} and exposures x_{ij} can be linked.

2.2. Model for aggregate data

Our model to estimate α and β from the aggregate data is as follows. Assume that individual-level exposures are unknown. In many cases, ecological covariates are estimated from surveys, rather than a census of the whole population. For example, small-area smoking rates are typically estimated using a combination of sales figures and survey data, and pollution averages are estimated using interpolation from monitoring stations. From this perspective, before individual-level exposures are measured and conditioned on, each individual in group i has an identical marginal probability p_i of outcome. This probability is the integral of the individual's conditional outcome probability $p_{ij}(\mathbf{x})$ over all exposures $\mathbf{x}$ with joint distribution $f_i(\mathbf{x})$ in group i.

$$p_i = \int p_{ij}(\mathbf{x}) f_i(\mathbf{x})\,\mathrm{d}x = E_{\mathbf{x}}(p_{ij}(\mathbf{x})\,|\,i) \tag{2}$$

Then, we can model the number of cases in group i by

$$y_i \sim \mathrm{Bin}(N_i, p_i) \tag{3}$$

If the individual-level exposures were assumed to be fixed and known, for example socio-economic indicators from a census of the full population, then an extension of the convolution likelihood suggested by Wakefield [15] for 2×2 tables would be more appropriate. This is constructed by conditioning on individual-level exposures.

We do not assume that the disease is rare. Although the incidence of many chronic diseases is statistically rare, the exceptions are particularly important, such as cardiovascular disease and childhood respiratory illness. The prevalence of limiting long-term illness, as recorded in the 1991 U.K. census, was approximately 12 per cent. In the case of a rare disease, the binomial sampling model is usually approximated by a Poisson distribution, with a log-linear model for the individual-level risk.

2.2.1. Basic case ecological model.
To obtain an explicit expression for p_i, we perform the integral (2) over the distribution of each exposure. We initially assume that the two exposures are independent, but will later relax this assumption. Firstly, assume that the probability of binary exposure within area i is a constant ϕ_i. Information about ϕ_i is provided by the total number exposed $x_i^{(1)}$ in the area, and the area population N_i, giving the model $x_i^{(1)} \sim \mathrm{Bin}(N_i, \phi_i)$. Secondly, assume that the continuous exposure has distribution $g_i(x)$ in area i. Summing over the two unknown values $x = 0, 1$ of the binary exposure, we obtain

$$p_i = q_{0i}(1 - \phi_i) + q_{1i}\phi_i \tag{4}$$

where q_{0i} is the marginal probability of outcome for an individual who is not exposed to the binary covariate, but whose continuous covariate has not been measured. Similarly, q_{1i} is the equivalent for an exposed individual. Integrating with respect to the continuous exposure,

$$q_{0i} = \int \mathrm{expit}(\mu_i + \beta x) g_i(x) \, \mathrm{d}x \tag{5}$$

$$q_{1i} = \int \mathrm{expit}(\mu_i + \alpha + \beta x) g_i(x) \, \mathrm{d}x \tag{6}$$

where $\mathrm{expit}(x) = \exp(x)/(1 + \exp(x))$. To calculate these, we need information on the distribution g_i of the continuous exposure.

A marginal proportion exposed is sufficient to estimate the full within-area distribution of a binary covariate. But ecological data are often not sufficient to estimate a within-area continuous covariate distribution. In many cases we only observe its mean $\bar{x}_i^{(2)}$ in each area. If we assume that the covariate is constant within areas at its expectation m_i, we obtain

$$q_{0i} = \mathrm{expit}(\mu_i + \beta m_i) \tag{7}$$

$$q_{1i} = \mathrm{expit}(\mu_i + \alpha + \beta m_i) \tag{8}$$

We estimate m_i by the empirical mean $\bar{x}_i^{(2)}$. This model assumes, in effect, that the relationship between the group-level outcome y_i and the group-level mean $\bar{x}_i^{(2)}$ of $x_{ij}^{(2)}$ is the same as the individual-level relationship between the outcome y_{ij} and $x_{ij}^{(2)}$. We refer to the above model as the 'basic case' model for ecological data, as it only makes use of the within-area sum or mean of each covariate. It is well-known that ignoring the within-area variability as in (7) and (8) leads to ecological bias if the variability is large enough. The case of one continuous covariate has been discussed by Richardson *et al.* [1] and Wakefield and Salway [7]. We now discuss various extensions to the basic case that may be necessary in practice, to alleviate the ecological bias of the basic case.

2.2.2. Ecological model including variance of the continuous exposure. In some cases, as well as the within-area mean, we may also have an estimate $\hat{s}_i^2$ of the within-area variance of $x_{ij}^{(2)}$, for example, from geographical modelling of an environmental exposure surface (e.g. Reference [12]). Then we suppose that these exposures are normally distributed, with $x_{ij}^{(2)} \sim N(m_i, s_i^2)$. If an exposure is not naturally normally distributed, it can often be transformed to normality. We can then calculate the area-specific risks by integrating (5) and (6) over g_i, here the density function of the normal distribution. Our assumed underlying model for $p_{ij}(x)$ is a logit-linear model on the exposures (1). In this case, the integral is not available in closed form. However, if we approximate the logit by a probit link function, then (5) and (6) evaluate to

$$q_{0i} = \mathrm{expit}\{(1 + c^2 \beta^2 s_i^2)^{-1/2}(\mu_i + \beta m_i)\} \tag{9}$$

$$q_{1i} = \mathrm{expit}\{(1 + c^2 \beta^2 s_i^2)^{-1/2}(\mu_i + \alpha + \beta m_i)\} \tag{10}$$

where $c = 16\sqrt{3}/(15\pi)$ [18].

If, instead, we were using a Poisson model for a rare outcome, $y_i \sim \text{Pois}(N_i p_i)$, and a log-linear individual-level model $\log(p_{ij}) = \mu_i + \alpha x_{ij}^{(1)} + \beta x_{ij}^{(2)}$, then the integrals can be evaluated explicitly without an approximation. Instead of (9) and (10), we would have (see, for example, Reference [1])

$$q_{0i} = \exp\left\{\mu_i + \beta m_i + \frac{\beta^2 s_i^2}{2}\right\}, \quad q_{1i} = \exp\left\{\mu_i + \alpha + \beta m_i + \frac{\beta^2 s_i^2}{2}\right\}$$

If a series of within-area exposure measurements are available, then these could be modelled, as part of the hierarchical model, with true mean and variance m_i and s_i^2, giving the required information to estimate the ecological relationships (9) and (10). In our application we save computational cost by replacing m_i and s_i^2 by their sample-based estimates $\bar{x}_i^{(2)}$ and $\hat{s}_i^2$, and we assess the sensitivity to this approximation.

We have described a fully parametric model for ecological inference. Prentice and Sheppard [11] described an alternative semi-parametric approach based on estimating functions. This is often simply called the *aggregate data* method. Suppose a sample of covariates (binary or continuous) are available from a subset of n_i individuals, but not the corresponding outcomes on the same individuals. Broadly, the mean and variance of the total disease count y_i are calculated in terms of an *aggregate* risk $(1/n_i) \sum_j p_{ij}$. p_{ij} is the risk for individual j in area i, conditionally on their covariate values. This approach does not require a within-area distribution to be specified for the covariates. It requires samples of covariate data as an explicit part of the model. On the other hand, the parametric approaches described above require individual covariate data implicitly to estimate an appropriate within-area distribution.

2.2.3. Accounting for correlated exposures. It is common that the two exposures are correlated. These might correspond to the risk factor under study and a confounding factor. For example, both smoking and exposure to air pollution are expected to be more common in socio-economically deprived areas. We previously assumed that they are independent in performing integral (2). Now we suppose that $x_{ij}^{(1)}$ and $x_{ij}^{(2)}$ are not independent within groups, and that there is a constant association across groups between the exposure and the confounder. One way to model the dependence is to assume that the within-group distribution of $x_{ij}^{(2)}$ is different conditionally on the two levels of the binary covariate:

$$[x_{ij}^{(2)} \mid x_{ij}^{(1)} = 0] \sim \text{N}(m_{i0}, s_i^2) \tag{11}$$

$$[x_{ij}^{(2)} \mid x_{ij}^{(1)} = 1] \sim \text{N}(m_{i1}, s_i^2) \tag{12}$$

$$m_{i0} \sim \text{N}(M_{x0}, S_x^2), \quad m_{i1} \sim \text{N}(M_{x1}, S_x^2) \tag{13}$$

To account for this, we replace m_i by m_{i0} and m_{i1} in equations (9) and (10), respectively. Then the aggregate data for the continuous exposure, the within-group sample mean $\bar{x}_i^{(2)}$ of $x_{ij}^{(2)}$, can be modelled in terms of m_{i0} and m_{i1}, as follows:

$$\bar{x}_i^{(2)} = \frac{1}{N_i} \sum_{j=1}^{N_i} x_{ij}^{(2)} \sim \text{N}(((N_i - x_i^{(1)})m_{i0} + x_i^{(1)} m_{i1})/N_i, s_i^2/N_i) \tag{14}$$

Ordinarily, there will be little information available in the aggregate data $(x_i^{(1)}, \bar{x}_i^{(2)})$ concerning the means of distributions (11)–(12), as without individual-level data we do not have paired

observations of the two exposures. However, if there are a large number of groups with the majority of individuals unexposed, and a large number of groups with the majority of individuals exposed, then these will contribute information about m_{i0} and m_{i1}, respectively. If we have additional individual-level data, we can also directly model m_{i0} and m_{i1} using (11)–(12).

If we have observed the within-area variance $\hat{s}_i^2$, this can also be modelled in terms of underlying variance parameters, in a similar way to (14), but for simplicity we assume that s_i^2 is the same for the two levels of the binary covariate.

2.2.4. Accounting for interaction. If there is a suspected interaction between the two covariates, that is, the association between $x_{ij}^{(2)}$ and the outcome is different for the two levels of $x_{ij}^{(1)}$, then an additional refinement to the model may be necessary. At an individual level, this would mean replacing equation (1) by (15) to model the interaction:

$$\text{logit}(p_{ij}) = \mu_i + \alpha x_{ij}^{(1)} + \beta x_{ij}^{(2)} + \gamma x_{ij}^{(1)} x_{ij}^{(2)} \tag{15}$$

Interaction can be accounted for in the ecological model by replacing β by $\beta + \gamma$ in equation (8) or (10).

2.2.5. Stratification. The baseline risk of disease in each area usually varies according to the demographic balance of the areas. For example, limiting long-term illness is more common among the elderly, so that it is necessary to stratify the model by age. At the individual level, we would be able to include age-stratum as an extra explanatory variable. At the aggregate level, we may be able to obtain population totals, and totals of individuals reporting limiting long-term illness, for each stratum. Then we could simply use a separate binomial model, in place of (3), for the risk of outcome in each stratum. The models would share the same exposure coefficients across strata. With no stratified outcome data available, a single binomial model must be used, extending equation (4) by expressing p_i as a sum of unobserved strata-specific outcome probabilities. However, it may be necessary to have large variations in the strata balance between areas, in order to determine the true age-outcome relationship using aggregate data alone, which may not be realistic. When stratifying the model, it is also important to check whether the exposures are likely to be correlated with strata. If not accounted for, this can lead to 'mutual standardization bias' [19].

2.3. Combining individual and aggregate data

To summarize the model for the ecological data, we have a binomial model (3) for the area-level outcome y_i. The corresponding area-level risk p_i is calculated explicitly in terms of the transformed group baseline risk μ_i, the individual-level covariate effects α and β, and the within-area distributions of the two covariates (4)–(6). The basic data are y_i, the within-area totals x_i of individuals exposed to the binary covariate, and an estimate $\bar{x}_i^{(2)}$ of the within-area mean of the continuous covariate. If an estimate $\hat{s}_i$ of the within-area variance is not available then we use model (7)–(8). If $\hat{s}_i$ is available, then this is modelled using (9)–(10).

It is easy to extend this model to include information from a small sample of individuals in each area whose outcome and exposures are known. We simply assume the binary outcome y_{ij} for such an individual j in area i is Bernoulli(p_{ij}), with a logit linear model for p_{ij} (1). Then the covariate effects α and β and the intercept μ_i are shared by the models for both

the aggregate and individual-level data. Thus, we can fit a joint model which combines the information between the two sources of data. Note that it is not necessary to have individual data within all of the areas i. In practice, sample survey data will be available from varying numbers of individuals between areas.

Full likelihoods for the models for individual data alone, aggregate data alone, and the combination of individual and aggregate data are given in Appendix A.

2.4. Prior distributions

To borrow estimation power and synthesize information across areas, we use a Bayesian hierarchical model, with area-specific parameters modelled as random effects. This model will be assessed, under various circumstances, with a simulation study, described in Section 3. We assume that the exposure probabilities are independent between groups with $\phi_i \sim U(0,1)$. We model $\mu_i \sim N(\mu, \sigma^2)$, representing exchangeable random baseline risks of disease for each area. In applications to geographical epidemiology, this part of the model might also include spatially-correlated random effects.

Specification of the prior distributions for μ, σ, α, β can be problematic. The obvious choice of a flat prior for μ with a large variance, combined with (1), leads to a marginal prior distribution for p_{ij} (for fixed covariate values) that is heavily biased towards 0 and 1 [15]. Instead we choose a logistic prior for μ with location 0 and scale 1, and $1/\sigma^2 \sim Gamma(1, 0.01)$, which lead to an approximately uniform marginal prior for p_{ij}. In the same way, high variances on covariate effects also lead to biased marginal priors for p_{ij}. Thus, we use weakly informative priors for α and β. To incorporate a 95 per cent prior belief that the odds ratio is between $\frac{1}{5}$ and 5, we choose normal priors with mean 0 and variance 0.68 for both α and β.

This hierarchical model can be fitted by Markov chain Monte Carlo simulation. Samples from full-conditional distributions can be generated using conjugacy or from a Metropolis algorithm with a normal proposal density, using the WinBUGS software [20].

3. SIMULATION STUDY

We perform a simulation study to assess the adequacy of estimation from ecological data using the basic model (7)–(8) described in Section 2, and its extensions. Where estimation is not adequate from ecological data alone, we assess how inference can be improved by including individual-level information from sample surveys. Firstly, we describe the basic conditions under which the data are simulated. Subsequently, some conditions will be varied from the basic case, to assess the resulting changes in the bias and precision of estimation.

3.1. Simulation set up for the basic case

- Suppose there are $I = 100$ groups of $N_i = 1000$ individuals, $i = 1, \ldots, I$. Small-area epidemiological studies in the U.K. are often carried out at the scale of the electoral ward, which contain around 6000 individuals on average. If we were studying disease risks for a subset of the population, such as individuals over the age of 45, then sample sizes of 1000 would be typical.
- The disease status of individual j in group i is generated taking the value 1 with probability p_{ij}, given by equation (1).

- The logit-baseline risks μ_i of disease for the I groups are chosen as 10 equally spaced quantiles of a normal distribution, with mean logit(0.1) and standard deviation 0.2, giving a 95 per cent sampling interval for the baseline disease risk of (0.07, 0.14). This is realistic for a non-rare condition such as cardiovascular disease.

- We take $\alpha = \log(2) = 0.69$ as the effect of the binary covariate $x_{ij}^{(1)}$, a typical value of a moderate odds ratio associated with a lifestyle factor such as smoking. The coefficient of the continuous covariate, $x_{ij}^{(2)}$, is fixed as $\beta = \log(2.3) = 0.8$, an effect size which was reported by Best *et al.* [12] for childhood leukaemia in relation to benzene exposure. Although relative risks associated with environmental exposures are usually smaller, we choose a large value of β to demonstrate a situation in which there is likely to be ecological bias. When β is small, the basic ecological model (7)–(8), which ignores the within-area variance, will be a better approximation to (9)–(10), which incorporates the variance.

- We vary the area-level proportion $\bar{x}_i^{(1)}$ of individuals exposed to the binary covariate between groups, as $\bar{x}_i^{(1)} = C \times l/10, l = 1 \ldots 10$. As the basic case we take $C = 0.2$, so that $\bar{x}_i^{(1)}$ are equally spaced between 0 and 0.2. Then the exposed individuals within each area are always in a minority. Such a narrow range is typical for a covariate such as non-white ethnicity in the U.K., or a behavioural variable such as smoking. Ecological data, consisting of aggregate counts of exposures and outcomes, will be more representative of the true exposure-outcome relationship in the majority group than in the minority group. Thus the 100 groups are identified by the 100 unique combinations of the 10 distinct exposed proportions $\bar{x}_i$ and the 10 distinct baseline risk parameters μ_i used.

- $x_{ij}^{(2)}$ within group i are fixed to be 100 equally spaced quantiles of $N(m_i, s_i^2)$. These group-specific parameters are generated as I equally spaced quantiles of

$$m_i \sim N(M_x, S_x^2) \qquad \log(1/s_i^2) \sim N(a_x, b_x)$$

with $M_x = 1.244, S_x = 0.2796, a_x = 1.95, b_x = 0.7154$. We also induce a between-area correlation of -0.3 between the m_i and the s_i. The choice of these quantities was based on the benzene exposure data from the study of the relationship of environmental benzene and childhood leukaemia by Best *et al.* [12]. The data for $x_{ij}^{(2)}$ are illustrated in Figure 1(a). The ratio of the between-area standard deviation (the standard deviation of m_i) to the within-area standard deviation (the mean of s_i) is $R = 0.74$. This quantity describes the amount of information in the ecological data concerning the true individual-level variability of $x_{ij}^{(2)}$.

From the full simulated individual-level data we form ecological data. These consist of the total number of individuals y_i with the disease in each area, the corresponding number $x_i^{(1)}$ exposed to $x_{ij}^{(1)}$, and the mean of $x_{ij}^{(2)}$ over the area, calculated as the within-area mean $\bar{x}_i^{(2)}$ of a subsample of 10 per cent of the full data.

Initially, we fit the basic ecological model (7)–(8) to the simulated data. Hundred simulation replicates are taken, and the model fitted repeatedly. For each replicate, we calculate the posterior summary statistics for the log odds ratios α and β, and the transformed mean $p_\mu = \text{expit}(\mu)$ and untransformed standard deviation σ of the group-specific logit baseline risks μ_i. To determine the estimation bias, the overall mean over all replicates of the posterior mean is compared to the true underlying value from which the data were simulated. The percentage
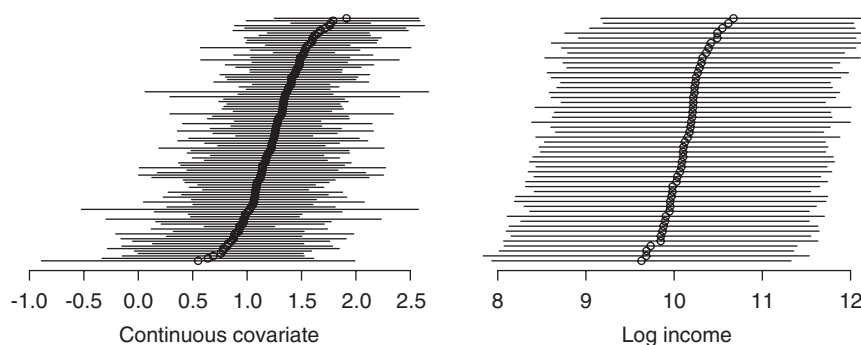
Figure 1. (a) Left: continuous covariate data used in the simulation study. Mean and 2.5–97.5 per cent quantiles of $x_{ij}^{(2)}$ for 100 areas. Ratio of between-area to mean within-area standard deviation is $R = 0.74$; and (b) Right: distribution of log household income for a randomly-chosen 50 of the London wards studied in Section 4. Mean $\pm 1.96 \times$ standard deviation of log income. Ratio of between-area to mean within-area standard deviation is $R = 0.25$.

bias, the coverage of nominal 95 per cent credible intervals, and the root mean square error of the posterior means relative to the true values, are calculated.

We repeat the simulation study under a variety of conditions to determine the resulting changes in bias and precision. There are a total of 34 cases described below, which present a comprehensive range of ecological modelling scenarios. The results for each of these cases are presented in Table I. Case 1 is the basic case described above.

3.1.1. Incorporating individual-level data. For each set of simulated data, we fit the model firstly to the ecological data, and secondly to the ecological data plus full data from a sample of 10 individuals from each group (1 per cent of the area population). This is a typical sample size seen in survey studies such as the Health Survey for England, where the areas are small geographical units such as electoral wards, and data are pooled across several consecutive survey years. If the survey data are too sparse, then larger geographical areas might be studied.

To determine a rough minimum number of individuals required to reduce bias to an acceptable level, we fit the basic case model including coupled exposure and response data from 5, 10 and 15 individuals (cases 2–4). Under each of the other model conditions, we fit each model firstly with only ecological data, and secondly also including a sample of 10 individuals. Next, we apply the basic case model using individual data only (5, 10 and 15 individuals), to assess the utility of combining individual and aggregate data, compared with individual data alone (cases 5–7).

3.1.2. Between-group exposure distributions. The amount of individual-level information in ecological data depends on the relationship between the between-area variability in the group mean exposure, and the within-area variability in individual exposures.

For a binary exposure, as discussed by Wakefield [15], the accuracy of ecological inference depends on the between-group variability of the proportions of individuals exposed in each group. For example, if only 10 per cent of every group are smokers, then data

Table I. Results of the simulation study.

	Sample	Binary α			Continuous β		
		%Bias	%Cove	%RMSE	%Bias	%Cove	%RMSE
Basic case							
1	0	−15.7	98	25.3	−6.61	90	9.19
2	5	−7.9	100	20.5	−5.94	88	8.46
3	10	−5.47	93	21	−4.94	85	8.29
4	15	−4.29	98	17.1	−5.17	84	8.63
Basic case, individual data only							
5	5	−15	97	34.1	−8.66	93	23.4
6	10	−7.3	96	24.4	−3.24	94	17.3
7	15	−5.87	97	21.6	−3.22	91	15.4
$C = 1$							
8	0	−2.67	96	6.07	−4.82	87	8.69
9	10	−2.04	95	6.4	−6.28	81	9.28
$R = 2.96$							
10	0	−16.9	96	27.8	−2.78	78	3.01
11	10	−3.11	96	19	−2.25	76	3.06
Estimated within-area variance							
12	0	−9.83	99	24	−3.01	92	8.69
13	10	−5.83	94	20.6	−2.45	89	7.97
Misspecified $x_{ij}^{(2)}$ distribution							
14	0	−14.8	99	25.2	−2.78	95	7.28
15	10	−6.59	98	17.9	−2.02	93	6.84
True within-area parameters							
16	0	−16.2	99	20.4	−0.305	88	9.13
17	10	−2.95	99	17.8	−0.82	91	7.39
Correlation, ignored							
18	0	20.2	87	40.6	−9.83	80	10.7
19	10	12.2	93	23.4	−9.36	75	10.3
Correlation, modelled							
20	0	−22.4	88	48.1	3.28	93	8.94
21	10	−8.3	94	21.4	0.75	96	7.49
Interaction, ignored							
22	0	26.6	95	30.3	−2.38	91	7.51
23	10	39.1	70	36.9	−2.59	93	7.5
Interaction, modelled							
24	0	5.96	100	62.1	−0.236	100	8.25
25	10	−11.5	98	56.6	−1.39	97	7.59
Small measurement error							
26	0	−15.8	98	24.9	−6.96	82	9.9
27	10	−6.69	98	18.3	−7.98	78	9.64
Large measurement error							
28	0	−13.5	100	24.7	−9.1	76	11.3
29	10	−9.42	96	20	−29.8	0	27.7

Table I. *Continued.*

	Sample	Binary α			Continuous β		
		%Bias	%Cove	%RMSE	%Bias	%Cove	%RMSE
Large ME, individual data only							
30	10	−14.4	93	30.1	−55.5	13	55.2
Omitted confounder, OR 2							
31	0	37.5	95	37.7	−7.19	81	10.1
32	10	44.5	57	41.1	−7.27	82	9.4
Omitted confounder, OR 1.5							
33	0	15.8	98	26.9	−6.94	82	9.47
34	10	31.2	79	30.7	−5.75	88	8.65

Mean percentage bias, coverage and root mean square error, over simulation replicates, of posterior means. True values of α, β are 0.69, 0.83 respectively. The basic case has $C = 0.2/R = 0.74$ (lower contrasts between areas of the mean binary/continuous exposure, respectively), the within-area continuous exposure variance is unknown, and there is no correlation or interaction between the covariates, measurement error or confounding.

consisting of only the total number of smokers in the group and the corresponding number of outcomes contain a lot of information about the risk of outcome in the non-smoking population, and very little information about the minority smoking group. Hierarchical models aim to borrow strength from groups with a lot of information on the exposed population to aid with inference for groups with little information on the exposed population, and vice versa.

In a similar way, the between-group distribution of the continuous covariate is important. If we have a wide spread of high and low m_i, and a relatively low within-group variance, the data are expected to contain more information about the true individual-level exposure-response relationship.

Therefore, we simulate two further sets of data with greater between-area variability in the exposure mean. Firstly, we suppose that the proportion of smokers is not uniform in the range 0–0.2, but uniform on 0–1 (cases 8–9). Secondly, we simulate data where the between-group standard deviation S_x of $x_{ij}^{(2)}$ is increased from 0.28 to $0.28 \times 4 = 1.12$ (cases 10–11). This increases the ratio of the between-area to mean within-area standard deviation from $R = 0.74$ to 2.96.

3.1.3. Modelling the within-group variability. Next we suppose that as well as an estimate $\bar{x}_i^{(2)}$ of the within-area mean m_i of $x_{ij}^{(2)}$, we have an estimate $\hat{s}_i^2$ of its variance, obtained as the empirical variance of a sample of 10 per cent of the simulated data in each area. Firstly, we include this variance in the ecological model, assuming a normal exposure distribution (9)–(10), (cases 12–13). This is expected to alleviate the ecological bias in the basic case. Secondly, we assess the sensitivity of this model to misspecification. New values of $x_{ij}^{(2)}$ are generated from an underlying Gamma distribution with the same mean and variance, and the normal model is fitted (cases 14–15). Thirdly, we take m_i and s_i^2 to be the true mean and variance from which the data are generated (cases 16–17), to assess the adequacy of substituting these true within-area parameters with empirical estimates.

3.1.4. Correlated exposures.

We also simulate data in which the distribution of the continuous covariate has a different mean for each value of the binary covariate, but an identical variance (equations (11)–(13)). Thus there is a within-group correlation between the covariates. We take $M_{x0} = M_{x1} + 4S_x$, so that $M_{x0} = 2.362, M_{x1} = 1.244$. We assess the bias when the covariates are truly correlated but we assume they are independent (cases 18–19). We then assess whether it is sufficient to model the correlation using the ecological data and the model of equation (14) (cases 20–21). For this model we have an estimate $\hat{s}_i^2$ of the within-area variance of the continuous covariate.

3.1.5. Interactions.

We simulate data with an interaction effect $\gamma = \log(1.3)$ between the binary and the continuous exposure. This can be modelled from the ecological data by replacing β by $\beta + \gamma$ in equation (8). If individual-level data are available, then we can use equation (15) directly to model the interaction. We fit four further combinations of models, which either ignore or model this interaction, and include either no individual data or a sample of 10 individuals (cases 22–25). This is an extension of the basic case (7–8) where we do not have an estimate of the within-area variance of the continuous covariate.

3.1.6. Measurement error.

We examine another case in which the continuous covariate $x_{ij}^{(2)}$ is measured with error, as is typical for a pollution exposure. We simulate individual-level covariate data $x_{ij}^{(2)*}$ conditionally on the fixed true values of $x_{ij}^{(2)}$, using a normal measurement error model, $x_{ij}^{(2)*} \sim N(x_{ij}^{(2)}, \sigma_\varepsilon^2)$. Ecological covariate data $\bar{x}_i^{(2)*}$ are formed by aggregating the erroneous measurements $x_{ij}^{(2)*}$. The outcomes are generated using the true measurements $x_{ij}^{(2)}$. Specifically, we simulate two situations of measurement error, where the reliability coefficient $\rho = s_i^2/(\sigma_\varepsilon^2 + s_i^2)$, the proportion of variance of $x_{ij}^{(2)*}$ explained by the variance of the true measurements, is 0.7 (small measurement error, cases 26–27) and a second case where $\rho = 0.4$ (large measurement error, cases 28–29). When the measurement error is large, we fit the individual-level model to the individual data alone (case 30), as well as fitting the basic case ecological model (7)–(8) to the aggregate data alone and combined aggregate and individual data. This is to investigate whether aggregation alleviates the problem of measurement error to which the individual-level data may be susceptible.

3.1.7. Unobserved confounding.

Finally, we investigate the impact of unobserved individual-level risk factors. We consider a situation in which the true individual-level model for outcomes (16) contains a unobserved binary covariate $x_{ij}^{(3)}$, and this is correlated with $x_{ij}^{(1)}$:

$$\text{logit}(p_{ij}) = \mu_i + \alpha x_{ij}^{(1)} + \beta x_{ij}^{(2)} + \alpha_c x_{ij}^{(3)} \tag{16}$$

We use this as the basis of simulations, and investigate two situations, where the coefficient of the confounder is $\alpha_c = \log(2)$ (cases 31–32) and a second case where $\alpha_c = \log(1.5)$ (cases 33–34). The correlation between $x_{ij}^{(1)}$ and $x_{ij}^{(3)}$ is fixed as 0.5. We base the fitted model around the basic case (1) and (7)–(8), with only two covariates, which is misspecified in this case.

3.2. Results

The MCMC chains converge within approximately 5000–10 000 iterations, but are slowly mixing. To obtain quantiles of the posterior distribution that are accurate within 2 decimal

places, all chains were run for 30 000 iterations. This takes about 5 min on a 2.4 GHz computer. Hundred replicates were used for each of the cases investigated, taking several days of run time. This was sufficient to give average posterior mean odds ratios that are stable to within 0.01, as illustrated by Figure 2.

The percentage biases, actual coverages of the nominal 95 per cent credible intervals, and percentage root mean square errors of the estimates are presented in Table I. Coverages of credible intervals are reasonable in most cases. The biases and precisions are illustrated by the 'caterpillar plots' in Figures 3 and 4. These show the mean and the interval between the $0.025-0.975$ quantiles, over the 100 simulation replicates, of the posterior means of the log odds ratios α, β, the mean baseline risk p_μ and the standard deviation of the logit baseline risks σ. Each figure compares the bias and precision for each of the 34 cases. For each case, the solid line indicates the results using the ecological data only, and a dotted line indicates the results incorporating a small sample of individual-level data. The figures and table are discussed from the top (basic case) to the bottom (confounding case).

3.2.1. Basic case results. The basic case simulation gives consistent underestimates of the coefficients for both covariates. Firstly, we are ignoring the within-area variability of $x_{ij}^{(2)}$, so the model is misspecified. Secondly, the between-area variability of mean $x_{ij}^{(1)}$ is narrow $(0-0.2)$, and the ratio of between to within-group standard deviation of $x_{ij}^{(2)}$ is less than 1, so there is little information available in the ecological data to estimate their coefficients α and β accurately. We can improve the estimation simply by incorporating individual-level data. For nearly every case, the mean square error of the estimates of both α and β is decreased by including samples of 10 individuals. A sample of 5 individuals reduces the bias for α from -16 to -8 per cent, while 10 and 15 individuals reduce the bias to -5 and -4 per cent, respectively.
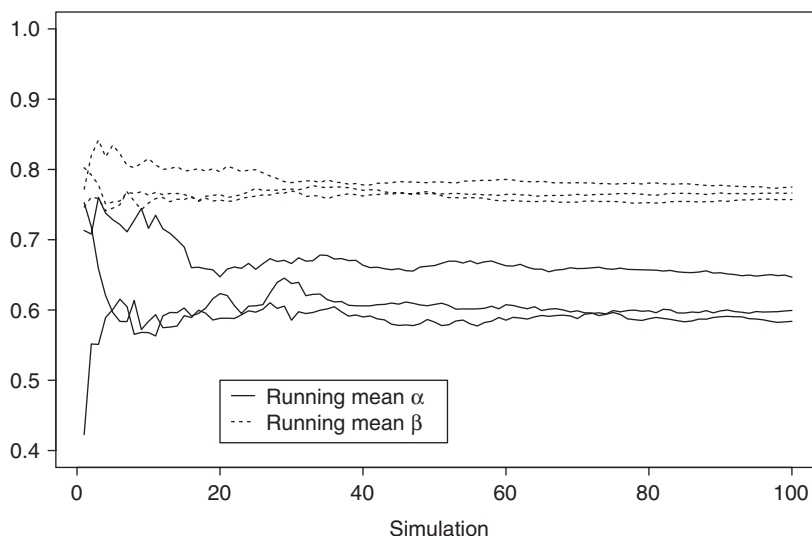

Figure 2. Convergence of the simulations. Running means, over simulation replicates, of posterior mean α and β are illustrated for the first three simulation cases. These and all other cases converged to stable running means.
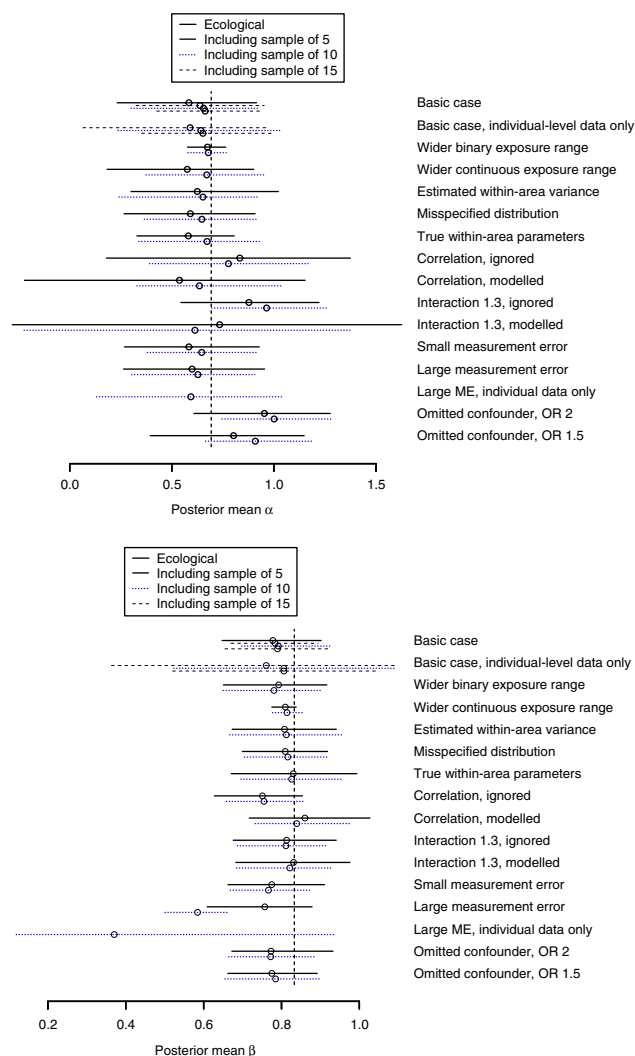
Figure 3. Means and 95 per cent quantile intervals of the simulation replicates of posterior means: α and β. True values shown by vertical dotted lines.

It might be suggested that if the individual data are substantial enough, then there is no point in combining them with ecological data, risking the methodological and interpretative problems of ecological studies. Indeed, if we analyse the sample survey data alone (cases 5–7, Table I) then the biases of the estimates are little different from the estimates from the combined data. The model for the individual-level data is asymptotically unbiased, but some bias remains for the small samples we have studied. However, the *mean square error* of these estimates is substantially increased by incorporating the ecological data (Table I), confirming the benefit of combining the data sets.
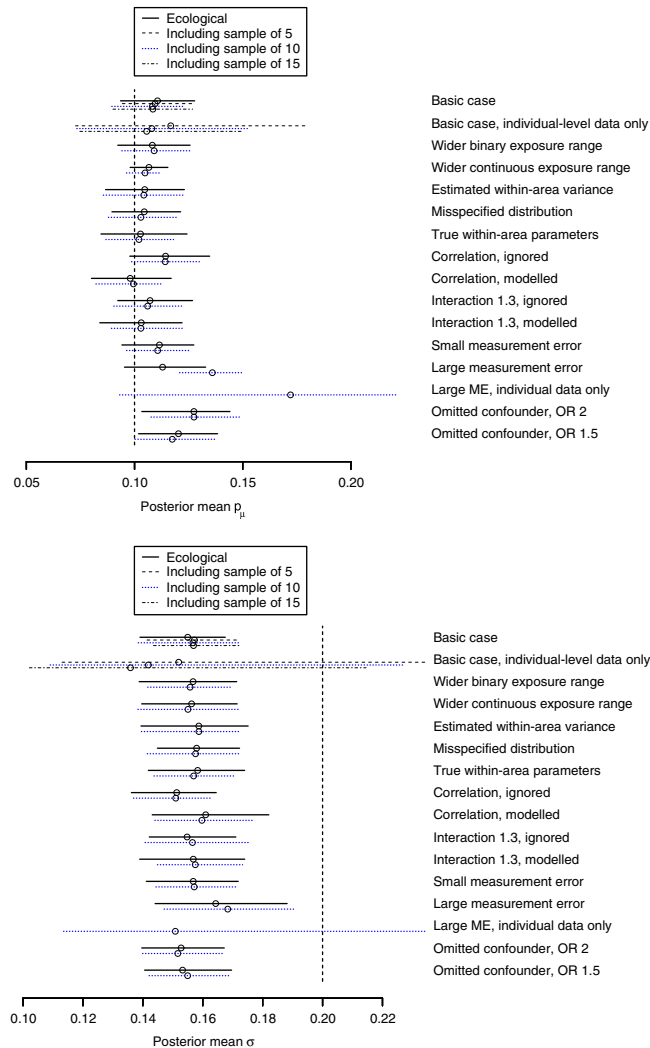
Figure 4. Means and 95 per cent quantile intervals of the simulation replicates of posterior means: p_μ and σ. True values shown by vertical dotted lines.

While the mean random logit baseline risk p_μ is generally estimated accurately, the corresponding variance is generally estimated with bias. We used a Gamma(1, 0.01) hyperprior for $1/\sigma^2$, which has a mean of 0.1 on the scale of σ, but gives a uniform distribution on the scale of p_i. This leads to estimates of σ that are heavily biased towards 0.1, away from the true value 0.2 (Figure 4). The simulations were repeated with a Gamma(1, 0.2) prior, which has a mean of 0.44 on the scale of σ (results not shown). This more informative prior produced substantially different estimates of the posterior mean σ of around 0.22. However, the estimates of α and β are not substantially affected, thus these are robust to the specification

of this hyper-prior. As these are the main parameters of interest, this lack of sensitivity is welcome.

3.2.2. Between and within-group exposure variability.

As the range of proportions of individuals in each group with the binary exposure increases from 0–0.2 to 0–1, the bias in the estimation of the log odds ratio α for the binary exposure decreases substantially, with a hugely increased precision (cases 8–9, Table I). With a minority of exposed individuals in every group, it is difficult to estimate the probabilities for this minority group using only the total exposed and the total number of cases. By increasing the variability of this exposed proportion between groups so that there is no minority group, we improve the estimation of this odds ratio. Similarly when the between-area variance S_x of $x_{ij}^{(2)}$ is quadrupled from its base case value, the simulations consistently provide nearly unbiased estimates of its coefficient β (cases 10–11, Table I).

When the within-group variance is modelled, estimating s_i^2 using the empirical variance of 10 per cent samples within each area, then the bias of estimating β is reduced from -7 to -3 per cent (case 12). There is only a little extra benefit in including additional individual-level coupled exposure-outcome data (case 13). Misspecifying the distribution of the continuous covariate as normal, when it is really gamma (cases 14–15), makes little difference in this case. As expected, using the true values of the mean and variance of the within-area distribution leads to unbiased estimates of β (cases 16–17, Table I). In this case the ecological model is theoretically unbiased. This suggests that if samples of within-area covariate data are available, then these should be incorporated in the hierarchical model to estimate the true mean and variance.

These results demonstrate that ecological data consisting only of within-area means can be adequate for inference about the individual-level relationship, even if the within-area variance is not modelled, *provided that there is sufficient between-group variability in the exposures.* With high between-area compared to within-area variability, the aggregate data will be more representative of the individual-level data. But typically the range of exposures between areas is not very large, in which case additional individual-level data on the exposure, and ideally also the response, can indeed be valuable.

3.2.3. Correlation and interaction.

When the covariates are correlated and this correlation is ignored, inference is biased for both α and β, even when individual-level data are included (cases 18–19). This bias is alleviated for β when the correct model, which takes account of this correlation, is used (case 20). But the inference for both parameters is only adequate when both the correct model is used, and individual-level data are also included (case 21).

If the outcomes are simulated with an interaction effect $\gamma = \log(1.3)$ between the effects of the two covariates, and this interaction is ignored, then α is estimated very poorly (case 22, bias 27 per cent). Additional individual-level data do not help (case 23, bias 39 per cent), since then both the individual and aggregate-level models are misspecified. When the interaction $\gamma = \log(1.3)$ is modelled using the aggregate data, then the bias for α is markedly reduced to 6 per cent. Additional individual-level data increase the precision of the simulation estimates for α, but do not improve the mean bias. The interaction effect γ itself (not presented in the table) is poorly estimated with the ecological data alone (bias -42 per cent), but more reasonably estimated (bias 10 per cent) if the individual-level data are incorporated.

We conclude that the more complex the true individual-level model, the more we need data which directly link the exposures and the outcomes. However, it is of principal importance to specify the individual-level model correctly.

3.2.4. Measurement error and confounding. When the continuous covariate is subject to a modest, unmodelled, measurement error (reliability coefficient $\rho = 0.7$, cases 26–27) the bias for its coefficient β is not substantially different from the basic case. However, with a more substantial measurement error ($\rho = 0.4$), the bias for β is very large (-55 per cent, case 30) from the individual data alone. When the individual and aggregate data are modelled together, the bias caused by the measurement error in the individual data is still noticeable (case 29, bias -30 per cent). When studying aggregated data alone, the bias is substantially reduced (case 28, bias -9 per cent). As expected, aggregation can alleviate bias due to individual-level measurement error.

Omitting a variable, confounded with $x_{ij}^{(1)}$, with a large odds ratio of 2, leads to substantially increased bias (37 per cent) in the estimation of α. This is not alleviated by using individual-level data, since the individual-level model is still misspecified. With a smaller odds ratio of 1.5, the bias for the model with ecological data alone is not affected, but when incorporating individual data under the wrong model, the bias is substantial (31 per cent).

4. APPLICATION TO LIMITING LONG-TERM ILLNESS

To illustrate the models discussed above in a real application, we study the prevalence of limiting long-term illness (LLTI) among men aged between 45 and 59 years of age, inclusive, in London, U.K. It is the only health-related outcome that was systematically recorded at the 1991 U.K. census. Its interpretation has been discussed by Cohen *et al.* [21], who found it to be correlated with several illnesses such as arthritis, asthma, chronic bronchitis, heart disease or diabetes.

We have small samples of individual-level data available from the Health Survey for England (Department of Health, U.K.), including limiting long-term illness, age, sex, ethnicity, and income. Individual ward identifiers were made available to us under a special arrangement with the data providers. Individual-level data are available from 255 electoral wards in London, with 1–9 observations per ward (median 1.6). Thus, in this case, the individual data are very sparse. Aggregate data on limiting long-term illness, age, sex and ethnicity for the corresponding 255 wards are taken from the U.K. census in 1991. From the census, characterization of the socio-economic deprivation of each ward can also be obtained. We use the classical Carstairs deprivation index (Carstairs and Morris, [22]) based on rates of adult male unemployment, car ownership, low social class and household overcrowding. Estimates of the mean and variance of household income, in U.K. pounds, for each ward are obtained from the PayCheck household income model (CACI, Limited).

An exploratory analysis of the between-group variability of the aggregate covariates indicates the existence of a similar pattern between the wards, with low average income correlated with proportion of non-white residents, both being also linked to the prevalence of LLTI and potential contextual variables such as the deprivation index (see Figure 5). Log household income has a small between-area standard deviation compared to its within-area standard deviation (Figure 1, ratio $R = 0.25$). Note that the variance is approximately constant between
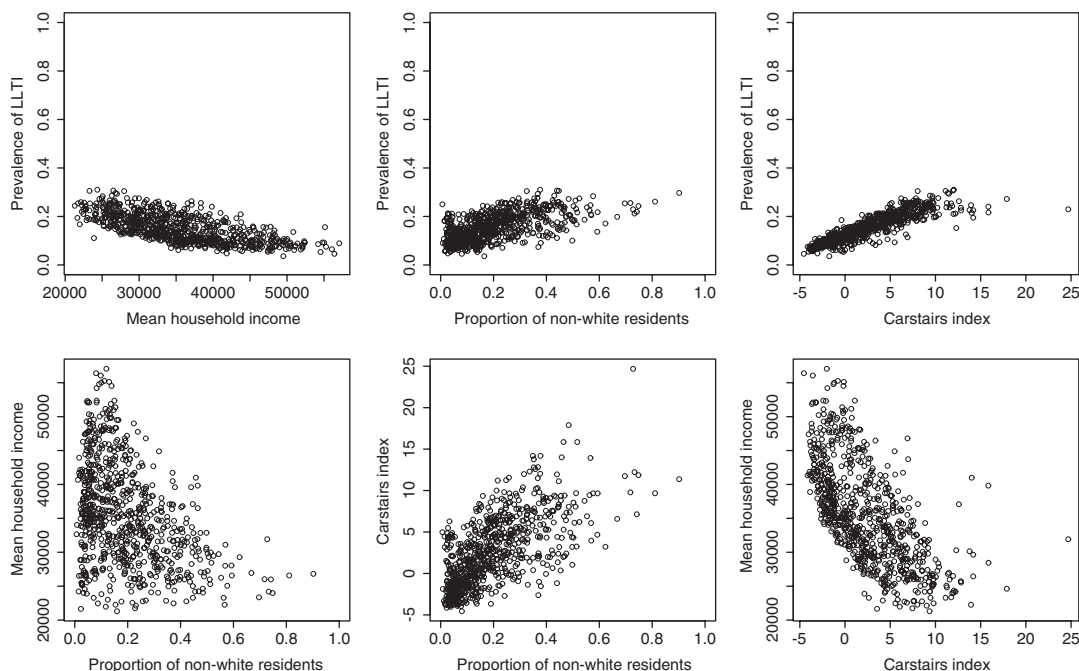
Figure 5. Relationships between ward-level variables.

areas, which may prevent ecological bias when within-area variability is ignored, as discussed by Wakefield [9].

We are interested in characterizing the effect of ethnicity and income on LLTI, and to investigate whether there is a residual contextual effect of area-level deprivation. As the basic individual-level model underlying these data, we consider both (1), and an extension (17) which accounts for area-level deprivation Z_i:

$$\text{logit}(p_{ij}) = \mu_i + \alpha x_{ij}^{(1)} + \beta x_{ij}^{(2)} + \gamma Z_i \tag{17}$$

In equation (17), p_{ij} is the probability of limiting long-term illness, $x_{ij}^{(1)}$ is ethnicity (dichotomous white/non-white) and $x_{ij}^{(2)}$ is log-transformed household income. Since the analysis is restricted to one sex and age class, it seems reasonable to assume in this case that the baseline μ_i is the same for all individuals in the group. The variability of μ_i will quantify the remaining contextual or environmental sources of heterogeneity of LLTI prevalence between London wards. The basic ecological model (7)–(8), which ignores the within-area variability of the continuous exposure, and the ecological model which incorporates this variability (9)–(10) are then derived. The within-area variance of household income was estimated using data on numbers of individuals in a series of income bands within each area. The models are applied to the ecological data alone, followed by the ecological data combined with the individual-level data. Exchangeable normal random effects were used for μ_i, and the priors

from Section 2.4 were used. Finally, any within-area correlation between income and ethnicity is accounted for, using models (11)–(14). One additional modification to the combined aggregate and individual model was rendered necessary in view of the discrepancy between the overall prevalence of LLTI as reported by the census and the HSE: 15 and 26 per cent, respectively, among men aged 45–59 years. Cohen *et al.* [21] found a similar discrepancy between census and survey data on LLTI in Scotland. A constant increment of 0.7 is thus added to the logit baseline of the individual-level component when combining the individual and ecological data.

Estimates from a variety of models are summarized in Table II. The individual analysis indicates a negative effect of non-white ethnicity and a negative effect of income on the risk of LLTI (first line of Table II), but the power is low and the interval estimates are wide and inconclusive for the effect of ethnicity. The coefficients α and β are not changed when area-level deprivation is included in the analysis of the individual data, but this is to be expected as the data are very sparse in each ward. On the other hand, at the aggregate level, the inclusion of deprivation has a marked influence on the estimates of the coefficients α and β, indicating substantial unmeasured confounding if contextual variables are not included. Once deprivation is included in the model of the baseline, the estimates for α and β become more compatible between the individual and ecological model, in the sense that interval estimates overlap. Non-white ethnicity and low household income are significantly associated with LLTI. After accounting for the within-area variance of income, both the coefficients are further away from the null effect, but accounting for the within-area correlation between income and ethnicity makes little difference to the estimates.

When incorporating the individual data alongside the ecological data, the estimates are changed by only a small amount from the ecological data alone, due to the sparsity of the individual-level data. Future studies should address design issues for the optimal choice of individual-level samples. Our simulation study shows the potential improvements to be gained by incorporating samples as small as 5–10 individuals per area.

Table II. Results from the application to limiting long-term illness.

Model	Non-white ethnicity	Log income	Deprivation	σ
Individual	−0.29 (−0.88, 0.28)	−0.56 (−0.80, −0.33)	—	0.17 (0.053, 0.56)
Individual	−0.36 (−0.98, 0.23)	−0.55 (−0.80, −0.32)	−0.022 (−0.032, 0.074)	0.18 (0.052, 0.64)
Ecological(without variance)	1.02 (0.88, 1.16)	−1.35 (−1.45, −1.25)	—	0.24 (0.23, 0.26)
Ecological(without variance)	0.27 (0.15, 0.39)	−0.57 (−0.67, −0.47)	0.068 (0.063, 0.074)	0.17 (0.16, 0.18)
Ecological(with variance)	1.56 (1.34, 1.85)	−1.84 (−2.12, −1.52)	—	0.31 (0.27, 0.36)
Ecological(with variance)	0.50 (0.27, 0.72)	−0.72 (−0.93, −0.51)	0.063 (0.054, 0.073)	0.19 (0.17, 0.21)
Combined(with variance)	1.58 (1.34, 1.83)	−1.78 (−2.17, −1.47)	—	0.31 (0.27, 0.35)
Combined(with variance)	0.48 (0.23, 0.72)	−0.70 (−0.91, −0.50)	0.064 (0.054, 0.074)	0.19 (0.17, 0.22)
Combined(correlation)	1.57 (1.31, 1.81)	−1.80 (−2.10, −1.46)	—	0.31 (0.27, 0.35)
Combined(correlation)	0.50 (0.24, 0.73)	−0.71 (−0.91, −0.51)	0.064 (0.054, 0.073)	0.19 (0.17, 0.21)

Estimated log-odds ratios for non-white ethnicity and log household income, considered as individual effects, with or without additional contextual effect of deprivation, from the individual level data, ecological data (without and with the within-area variance of income), combined individual and ecological data (with income variance), and combined data accounting for the within-area correlation between income and ethnicity (and income variance). σ is the estimated standard deviation parameter of the exchangeable random effects μ_i.

5. DISCUSSION

In this paper we have presented a general model for ecological inference, and illustrated how to systematically combine ecological data with small samples of individual-level exposures and outcomes. Its performance was investigated in detail for a model of disease risk in terms of one binary and one continuous explanatory variable. As expected, a simple ecological regression of area outcomes in terms of mean exposures, ignoring the within-area exposure variability, was shown to give bias if this variability is large enough. We discovered that small individual-level samples can substantially improve inference, reducing bias to acceptable levels, in cases where ecological modelling is biased. Nevertheless, ecological data alone can often give accurate inference, specifically in cases where the between-group variance in exposure means is large in comparison to within-group variances. Ecological data can also alleviate bias due to measurement error in individual-level covariates.

In the presence of two explanatory variables there are even more complexities. The two covariates may be correlated, as in the common example of socio-economic confounding, or there may be an interaction effect. When the individual-level model becomes more complex, then inferences are only adequate when both the correct model is used, and we also have individually-linked data available to identify the proposed model. The framework described may be extended to account for two or more categorical or continuous exposures and confounders. However, we still need to specify the joint within-area distribution of all covariates, which may become rapidly more complex. Suitable individual-level data will probably be required to estimate this distribution accurately.

The accuracy of ecological regression is known to be influenced by many factors other than the ones we have studied in detail here. Ecological bias increases with the size of the effects to be estimated, and the extent of any correlation between the within-area means and variances [9]. Our simulation was based on 100 areas, representing a small region. Using data from a greater number of areas should help to increase precision, as the information on between-area exposure contrasts increases. Our simulations assumed a constant within-area population of 1000. In practice, ecological studies will be based on areas with varied population sizes. However, we do not expect changes in population size to affect the accuracy of ecological inference. As discussed by Wakefield [15], the amount of information in ecological data does not vary greatly with the underlying population size from which the aggregates were taken, as the fundamentally weak identifiability of the individual-level relationship remains the same.

The model framework is built on a binomial likelihood for the group-specific outcome count. This is based on the assumption that the exposures of the individuals within the group are unknown, independent and identically distributed, thus the individual outcomes are independent Bernoulli. This is a reasonable assumption if the ecological data are an estimate of the true individual-level exposures, such as smoking data based on sales figures and surveys, and pollution data obtained from monitoring stations. In our limiting long-term illness example, the mean and variance of household income are estimated from a survey, and the proportion of men aged 45–59 who are non-white (unavailable from the census) is an estimate, assumed to be equal to the proportion of men aged 30–65 who are non-white (available from the census). Once group-level summaries of covariate data become known for the same individuals from whom the group-level outcomes are measures, Wakefield [15] and Forster (discussion of Reference [15]) argue that these should be conditioned on, leading to a likelihood based on a convolution of binomial distributions.

Our results can be useful for guidance for the design of ecological studies which include sample survey data. Future work might explore the utility of choosing non-random samples of individuals. Wakefield [15] showed that for a set of 2×2 tables, it was more helpful to choose the sample from the minority group in each area, for example, individuals exposed to a rare risk factor. This also raises the possibility of combining ecological data with samples of data from case–control studies. This may be particularly beneficial for improving the power of case–control studies in situations where data collection is very expensive or the outcome is very rare.

Linking ecological and individual data may be problematic in practice, as observational data are often unbalanced and incomplete. In some applications, full aggregate and sampled individual data can be obtained from exactly the same source. For example the U.K. 1991 Census published samples of 2 per cent of individual-level census responses as 'Samples of Anonymized Records', and aggregate data at a variety of geographical scales for the entire population. In other applications, we may wish to enhance aggregate data from censuses or registers, nominally covering the whole population, with individual-level data from survey or cohort studies. Obtaining data on the same population from different sources may lead to inconsistencies. Sample surveys may only be available from a few areas, from varying numbers of individuals. The resulting lack of individual-level information on some areas may then be alleviated by having more information from other areas. There may be different covariates available at the aggregate and individual level. For example, outcomes and demographic covariates might be available from individuals, but not the exact exposure to, say, air pollution. The response or exposure variable could be different at aggregate and individual level, for example individual self-reported incidence of a disease and group-level hospital admission rates. For some quantities, such as air pollution, it is difficult to know the individual's exact exposure. Surveys might not cover exactly the same period of time as the aggregate data. There might also be differently-defined covariate information available at each level, or geographical boundaries may be different.

Careful extensions of our framework will be necessary to deal with these additional features where possible. For each of these additional complications, the Bayesian hierarchical modelling approach simplifies the process of deriving and implementing an elaborated model [17]. As well as accounting for random baseline risks across areas, the framework is amenable to extensions to further complexities, such as multiple levels of aggregation, exposure and outcome measurement error, spatial dependence between baseline risks or random regression coefficients.

We end by summarizing the main conclusions of the paper:

- Inference from ecological data alone can be accurate when there are high exposure contrasts between areas, and can reduce bias due to unmodelled measurement error in individual data.
- Combining ecological data with small samples of individual-level data can reduce ecological bias when the exposure contrasts are low.
- Combining individual with ecological data can decrease the mean-square error of estimates from individual-level data alone.
- When the true model is more complex, for example with interactions or confounding, it is most important to use the correct model. Individual data cannot help if the model is misspecified.

APPENDIX A: LIKELIHOODS

The likelihood l_{agg} for the ecological model for the aggregate data alone $\{y_i, x_i^{(1)}, \bar{x}_i^{(2)}; i=1,\dots,I\}$ (equations (3) and (4)) is

$$l_{agg}(\{\mu_i\},\alpha,\beta|\{y_i\}) = \prod_i (q_{0i}(1-\phi_i)+q_{1i}\phi_i)^{y_i}(1-(q_{0i}(1-\phi_i)+q_{1i}\phi_i))^{N_i-y_i}$$

where q_{0i} and q_{1i} are functions of $(\{\mu_i\},\alpha,\beta)$, given by either (7), (8), or (9), (10), if an estimate of the within-area variance of $x_{ij}^{(2)}$ is unavailable or available, respectively. $\{\mu_i\}$ and $\{y_i\}$ denote $\{\mu_i; i=1,\dots,I\}$ and $\{y_i; i=1,\dots,I\}$. Suppose we have samples of individual data from M_i individuals from each area i. The likelihood l_{indiv} for the individual-level data alone $\{y_{ij}; j=1,\dots,M_i; i=1,\dots,I\}$ (model 1) is

$$l_{indiv}(\{\mu_i\},\alpha,\beta|\{y_{ij}\}) = \prod_{i,j} (\text{expit}(\mu_i+\alpha x_{ij}^{(1)}+\beta x_{ij}^{(2)}))^{y_{ij}}(1-\text{expit}(\mu_i+\alpha x_{ij}^{(1)}+\beta x_{ij}^{(2)})))^{1-y_{ij}}$$

The likelihood l_{comb} for the combined individual and aggregate data is simply the product of the above two likelihoods,

$$l_{comb}(\{\mu_i\},\alpha,\beta|\{y_{ij}\},\{y_i\}) = l_{indiv}(\{\mu_i\},\alpha,\beta|\{y_{ij}\}) \times l_{agg}(\{\mu_i\},\alpha,\beta|\{y_{ij}\})$$

REFERENCES

1. Richardson S, Stucker I, Hémon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* 1987; **16**(1):111–120.
2. Greenland S, Morgenstern H. Ecological bias, confounding and effect modification. *International Journal of Epidemiology* 1989; **18**:269–284.
3. Greenland S, Robins J. Ecological studies—biases, misconceptions and counterexamples. *American Journal of Epidemiology* 1994; **139**:747–760.
4. Richardson S, Monfort C. Ecological correlation studies. *Spatial Epidemiology*, Chapter 11. Oxford University Press: Oxford, 2000.
5. Greenland S. A review of multilevel theory for ecological analyses. *Statistics in Medicine* 2002; **21**:389–395.
6. Sheppard L. Bias and information in group-level studies. *Biostatistics* 2003; **4**(2):265–278.
7. Wakefield J, Salway R. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society*, Series A: Statistics in Society 2001; **164**(1):119–137.
8. Fortunato L, Guihenneuc-Jouyaux C, Laurier D, Tirmarche M, Clavel J, Hémon D. Introduction of within-area risk factor distribution in ecological Poisson regression models. *Statistics in Medicine* 2005, under revision.
9. Wakefield J. Sensitivity analyses for ecological regression. *Biometrics* 2003; **59**:9–17.
10. Lasserre V, Guihenneuc-Jouyaux C, Richardson S. Biases in ecological studies: utility of including within-area distribution of confounders. *Statistics in Medicine* 2000; **19**:45–59.
11. Prentice RL, Sheppard L. Aggregate data studies of disease risk factors. *Biometrika* 1995; **82**:113–125.
12. Best N, Cockings S, Bennett J, Wakefield J, Elliott P. Ecological regression analysis of environmental benzene exposure and childhood leukaemia: sensitivity to data inaccuracies, geographical scale and ecological bias. *Journal of the Royal Statistical Society*, Series A: Statistics in Society 2001; **164**(1):155–174.
13. Department of Health. Health Survey for England, London, 1991–2003. http://www.dh.gov.uk/

14. Centre for Longitudinal Studies. Millennium Cohort Study. Institute of Education, London, 2001. URL: http://www.cls.ioe.ac.uk/Cohort/MCS/mcsmain.htm
15. Wakefield J. Ecological inference for 2×2 tables (with discussion). *Journal of the Royal Statistical Society, Series A*: *Statistics in Society* 2004; **167**(3):385–445.
16. Ben-Shlomo Y, White I, Marmot M. Does the variation in the socioeconomic characteristics of an area affect mortality? *British Medical Journal* 1996; **312**:1013–1014.
17. Richardson S, Best N. Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics* 2003; **14**:129–147.
18. Salway R, Wakefield J. Sources of bias in ecological studies of non-rare events. *Environmental and Ecological Statistics* 2005; **12**(3):321–347.
19. Rosenbaum PR, Rubin DB. Difficulties with regression analyses of age-adjusted rates. *Biometrics* 1984; **40**: 437–443.
20. Spiegelhalter DJ, Thomas A, Best NG. *WinBUGS Version 1.4*, *User Manual*. MRC Biostatistics Unit: Cambridge/Imperial College School of Medicine: London, 2001. URL: http://www.mrc-bsu.cam.ac.uk/bug
21. Cohen G, Forbes J, Garraway M. Interpreting self-reported limiting long-term illness. *British Medical Journal* 1995; **311**(7007):722–724.
22. Carstairs V, Morris R. *Deprivation and Health in Scotland*. Aberdeen University Press: Aberdeen, 1991.