

Controlling for provider of treatment in the modelling of respiratory disease risk near cokeworks

Alex Bottle^{1,*},[†] and Jon Wakefield²

¹*Department of Epidemiology and Public Health, Imperial College School of Medicine, St Mary's Campus, Norfolk Place, London W2 1PG, U.K.*

²*Departments of Statistics and Biostatistics, University of Washington, Box 357232, Seattle, WA 98195, U.S.A.*

SUMMARY

The improved quality of hospital admissions data makes them a valuable resource for researchers. However, we show that, when multiple providers of health care are considered, the provider (in this case hospital) may act like other confounders such as socio-economic status, and hence must be controlled for. Such control is not as straightforward as for conventional confounders, however, but we describe a method that is appropriate under certain assumptions. We also describe a number of other statistical issues, such as the modelling of spatial and non-spatial overdispersion, that arose during the use of hospital data in a study to investigate the possible adverse health effects of living in proximity to six cokeworks groups in England and Wales. The outcome data that we consider consist of hospital admissions for all respiratory disease in the under-5s. The ecological level of the analysis is the census-defined enumeration district, and the main (proxy) exposure measure utilised is the spatial location of the enumeration district population-weighted centroid in relation to the cokeworks. We focus on the Teesside cokeworks group, for which we also had sulphur dioxide measurements from dispersion modelling as an alternative exposure measure. The major local providers varied appreciably in their standardized admission ratios for respiratory disease, and when provider was controlled for, the size of the observed excess risk found close to the cokeworks was decreased, making control for the provider of health care vital. However, the presence of multiple pollution sources, in addition to the usual shortcomings of ecological studies, makes interpretation difficult. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: ecological studies; hospital admissions; provider effects; Simpson's paradox; spatial epidemiology

*Correspondence to: Alex Bottle, Department of Epidemiology and Public Health, Imperial College London, St. Mary's Campus, Norfolk Place, London W2 1PG, U.K.

[†]E-mail: robert.bottle@imperial.ac.uk

Contract/grant sponsor: Department of Health

Contract/grant sponsor: Department of the Environment, Food and Rural Affairs

Contract/grant sponsor: Environmental Agency

Contract/grant sponsor: Health and Safety

Contract/grant sponsor: Scottish Executive

Contract/grant sponsor: National Assembly for Wales

Contract/grant sponsor: Northern Ireland Assembly

1. INTRODUCTION

The past 20 years have seen growing interest in spatial epidemiology. Reasons for this include an increased awareness of the potentially detrimental effects of environmental pollutants; improvements in data availability and quality; growth in computing power and software sophistication such as geographical information systems; statistical methodological development; and a number of well publicised examples linking environmental pollutants with health risks, such as a reported excess number of cases of childhood leukaemia near Sellafield nuclear power station. Controlling for provider (in this case hospital) is not as straightforward as for conventional confounders such as age and sex, since there is no well-defined population. Several different types of spatial epidemiological studies are possible [1], including disease mapping, clustering and cluster detection, and spatial regression in which risk is examined with respect to spatially indexed covariates of interest. We will concentrate on the last of these in the context of a study that investigated the risk of respiratory disease in relation to cokeworks in England and Wales. Although coke and steel industries in the United Kingdom have contracted to a fraction of their size a generation ago, about 750 000 people live within 7.5 km of one of the nine sites that still operate in England and Wales. These nine sites may be divided geographically into six groups evenly split between North East England and South Wales. The main point of pollution release is when the coke oven doors are opened to allow the blended coal to be fed into the 1300°F furnace. Although emissions are much lower than they were a generation ago, they remain a cause of local public concern.

There have been several studies of the possible health hazards of living near operating cokeworks, covering a spectrum of end-points. Bhopal *et al.* [2] found elevated rates of symptom reporting for respiratory and ear, nose and throat complaints among residents (especially children) living near Monkton cokeworks, with elevated all-cause mortality among children, but not adults. Dolk *et al.* [3] found a small excess of mortality within 2 km of cokeworks and a decline in mortality with distance from cokeworks, but could not exclude residual socio-economic confounding as an explanation. In a second study, Dolk *et al.* [4] found no evidence of an increased risk of perinatal and infant mortality and low birthweight among births to mothers living near cokeworks. In the view of Bhopal *et al.*, the health effects of point sources of pollution such as cokeworks may be subtle, in that they are not obviously manifested by analyses of mortality or cancer data. Although they have fallen considerably since the 1960s, it appears that emissions from cokeworks contribute to respiratory morbidity and that an investigation using hospital admissions, which had not been considered in detail before, was felt to be warranted.

As highlighted by Wakefield and Elliott [5] there are a number of difficulties inherent in these investigations, especially pertaining to data quality. These include problems due to data anomalies such as double- or under-counting of cancer registrations and inter-rater disagreement over clinical diagnosis, causing concern over the accuracy of numerator data, uncertainties in the denominator data due to census underenumeration or migration, and the frequent lack of a measure of cumulative exposure, requiring the use of some proxy which may have considerable measurement error. Other issues include adjustment for confounders, form of the exposure-risk relationship and accommodation of overdispersion. Another important consideration in spatial studies with aggregate data is the potential for ecological or cross-level bias when relationships at the level of the group (which correspond to areas) are assumed to hold for the individuals within those groups. In this paper we use the cokeworks study to motivate

the derivation of a statistical model that may be used for general point-source studies, being explicit about the assumptions underlying the parametric models. A novel contribution of this paper is to describe a method for controlling for the provider of health care (hospital Trust), which, as we show, may act as a confounder.

The structure of this paper is as follows. In Section 2.1, we describe the study design, and in Section 2.2 illustrate how the provider can act as a confounder in this type of study. In Section 3, we consider a statistical framework that may be used for point-source studies, and address adjustment for confounders, the form of the risk/exposure relationship and overdispersion. Adjustment for the provider is specifically considered in Section 4. In Section 5, we provide a substantive analysis of the cokeworks example in order to illustrate the statistical issues and the control for provider. The full results of the study are given elsewhere [6]. In Section 6, we provide a concluding discussion.

2. COKEWORKS STUDY

2.1. Data

For the current study, Aylin *et al.* [6] considered a number of cardiorespiratory endpoints: respiratory disease as a whole (ICD-9, 460–519), asthma (493), ischaemic heart disease (IHD, 410–414), cerebrovascular disease (CVD, 413–418) and chronic obstructive airways disease (COAD, 491–492). Two Welsh cokeworks groups were excluded from the study due to highly incomplete diagnostic coding (>25 per cent of recorded admissions had missing codes) in the local hospitals taking most of the admissions from residents living near the plants. The missing-data mechanism may depend on spatial location (and therefore exposure), hence introducing the potential for bias. This demonstrated the importance of data quality assessment as the first step before formal statistical analysis, for which four cokeworks groups remained. We present results for the Teesside group only, which had the largest nearby population of the four; SO₂ estimates were also available for this area. The outcome that we examine here is emergency admission for any respiratory disease in the under-5s.

The study had an ecological design with outcome data at the postcode level but exposure and deprivation data at enumeration district (ED) level. Postcodes contain on average 14 households, and EDs on average 400 people. Specific groupings of EDs constitute an electoral *ward*, collections of which in turn make up a *district*. Districts form counties and counties form administrative regions, of which there were ten in Britain at the 1991 census (Wales and Scotland form one region each).

The deprivation data consisted of the Carstairs score [7], which has four components (unemployment, overcrowding, social class of head of household and access to a car) and is at ED level. Each component is standardized across Britain to have zero mean and unit variance, and the Carstairs score is simply the sum of the four component scores. It has been found to be well correlated with hospital admission rates [8] and more closely associated with hospital admission rates than either the Jarman or Department of Environment indices [9]. EDs were assigned to one of five quintiles based on their Carstairs score. Such an index may be used as a proxy for individual-level confounders such as smoking [10] or it may represent a genuine area-level effect. For instance, high crime rate, poor community facilities etc. may have negative influences on health. In either case, inference is subject to the possibility

of ecological bias, which arises when individual-level inference is inferred from aggregate data (see Section 3.1). Ideally, we would like to control for the within-area distribution of confounder variables, and a single measure, such as the Carstairs score, will not generally achieve such control. High proportions of the population living within 2 km of either the Teesside or the Newport sites are in EDs in the fifth (most deprived) quintile, which makes control for confounding important.

2.2. Provider as a confounder

A hospital's admission rates are determined not only by the morbidity in the population it serves (the *catchment area*) but also by myriad other factors such as clinical coding and admissions policy [11]. There are few conditions for which nationally agreed standard treatment protocols exist, and the decision whether or not to admit a patient will be influenced by the opinion of the clinical team and bed availability. As about half of the emergency admissions in the U.K. are general practitioner (GP) referrals, GP preferences for one hospital rather than another to which to send their patients are also important [12]. Area-level admission rates will therefore exhibit a degree of spatial variation due to these factors and this may bias small-area studies of health and exposure to a putative factor. For example, a hospital very near a cokeworks may mask or exaggerate the estimated effect of living near the works, depending on the difference in rates between it and more distant providers; if it has high rates, then admission in near areas will also be higher than in not-near areas, even if there were no effect of cokeworks pollution on health. Adjustment for its effect is not trivial due to the lack of provider-specific population figures. We derive a method for estimating these populations in Section 4.

For the population under 5 living near the Teesside cokeworks, three hospitals received 99 per cent of all respiratory admissions. Using the administrative region as the referent and estimating catchment areas for each hospital using the method described in Section 4.1, standardized admission ratios (SARs) were obtained. These were 1.63, 1.33 and 0.92 (adjusting for age, sex and deprivation). The relation between the distance to nearest cokeworks and distance to nearest provider is shown in Figure 1 for EDs used with pollution estimates in Section 5.3.

The relation is not linear, but the two distances do appear to be positively correlated. EDs further from the cokeworks are in general also further from the provider with the highest admission rates and are therefore less likely to use that provider, thus suggesting that we could see a decline in risk with distance because of the provider effect even if the cokeworks pollutants themselves have no effect on the risk of admission. Admission rates vary by provider and there is an association between provider and exposure, suggesting that provider is a confounder.

3. MODELS FOR POINT SOURCE STUDIES

In this section, we set out the assumptions of the statistical models that we use so that these can be assessed before interpreting observed associations. From a modelling perspective we identify three issues: control for confounders, the form of the exposure-risk relationship and accommodation of overdispersion. We will consider each of these issues but begin by

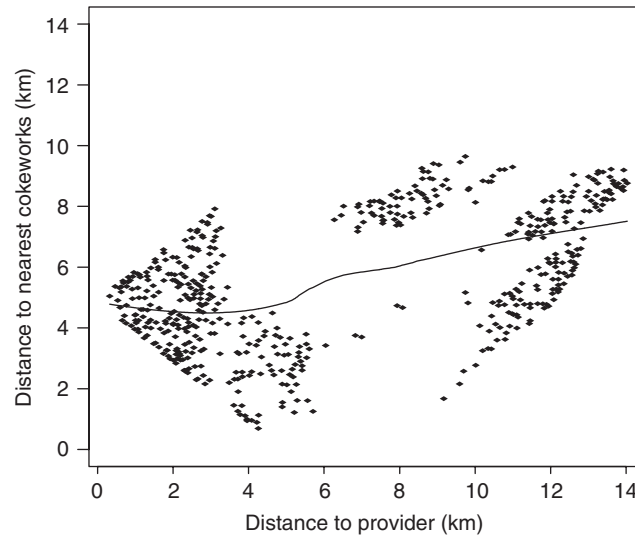


Figure 1. Relation between distance to nearest cokerworks and distance to nearest provider for enumeration districts near the Teesside cokerworks, with local smoother superimposed.

establishing the basic framework. We build the model at the individual level to allow issues of ecological bias to be explicitly considered.

3.1. Statistical framework

Let Y_{ijl} represent the binary response (0 = not admitted, 1 = admitted for respiratory disease) of individual l in area i , stratum j , $i = 1, \dots, I$, $j = 1, \dots, J$, $l = 1, \dots, n_{ij}$. Stratification variables may include some or all of age, sex, deprivation index (e.g. Carstairs), and provider. The value of J will depend on the complexity of the confounder model, and in particular whether interactions are included. We also define Z_{ijl} to be a generic exposure measure, e.g. personal exposure or distance from a putative pollution source. Then $Y_{ijl} \sim \text{Bernoulli}(p_{ijl})$, where p_{ijl} is the probability of admission in area i and stratum j for individual l . A reasonably general model is given by

$$\text{logit } p_{ijl} = X_{ijl}\beta + f(\gamma, Z_{ijl}) \quad (1)$$

where $\exp\{f(\gamma, Z_{ijl})\}$ is a functional form, with parameters γ , representing the effect on odds of admission of exposure given that individual l has exposure Z_{ijl} , X_{ijl} represents a $1 \times J$ vector with a 1 in the j th position, corresponding to the stratum of individual l , and zeroes elsewhere, and $\beta = (\beta_1, \dots, \beta_J)^T$; $\exp(\beta_j)$ gives the odds of admission in stratum j , $j = 1, \dots, J$. This model assumes that the effect of the exposure is constant across strata, a common assumption in ecological studies that is necessary because only a single exposure measure per area is typically available.

In general this model will be far too complex to be fitted given the quality and abundance of data, and so simplifications are required. The exposure measures Z_{ijl} of study participants will rarely be known and in this study individuals are assigned the distance between

population-weighted centroid of the area in which they reside and the nearest cokeworks. This 'aggregation' leads to the possibility of *ecological bias* (e.g. References [13–20]), an umbrella term describing a range of biases that lead to differences between area-level and individual-level inferences. The *ecological fallacy*, in which incorrect inference is provided, does not occur if exposures/confounders are constant within areas, but this is clearly not true here since exposure varies within an area (ED), as does deprivation, a major confounder. We assume that the exposure is constant with respect to stratum within an area and write Z_i to denote the common assigned exposure of all individuals in area i .

We now consider the number of admissions within area i , stratum j , $Y_{ij} = \sum_l Y_{ijl}$. In the rare case, the logistic model (1) becomes:

$$\log p_{ij} = X_{ij}\beta + f(\gamma, Z_i) \quad (2)$$

where $\exp(\beta_j)$ now represent risks. The log-linear model is, in general, more stable for estimation than the logistic alternative. If each of the Bernoulli random variables Y_{ijl} are independent, we have $Y_{ij} \sim \text{Binomial}(n_{ij}, p_{ij})$, which may be approximated by

$$Y_{ij} \sim \text{Poisson}(n_{ij} \times p_{ij}) \quad (3)$$

if the disease is rare (the admission rate for all respiratory disease was 5.8 per cent in the under-5s on Teesside). The advantage of the latter is that we may sum across strata.

3.2. Control for confounders

To control for confounders a number of possibilities are available. Summing across strata via equations (2) and (3) gives

$$Y_i \sim \text{Poisson}\left(\sum_j n_{ij} \exp\{X_{ij}\beta + f(\gamma, Z_i)\}\right)$$

We may estimate the risks $\exp(\beta_j)$ associated with the confounders either in advance via standardization, or simultaneously when the parameters γ are estimated. Taking the former route via indirect standardization gives the model,

$$Y_i \sim \text{Poisson}(E_i \times \exp\{\alpha + f(\gamma, Z_i)\})$$

where $E_i = \sum_j n_{ij}\hat{q}_j$ and $\hat{q}_j = \exp(\hat{\beta}_j)$ represent the risks associated with stratum j . The risks \hat{q}_j may be estimated from the study individuals via $\hat{q}_j = Y_{+j}/n_{+j} = \exp(\hat{\beta}_j)$ to give *internal* standardization, or estimated from other areas that do not contain the study area, but are as close to the study population in other characteristics, to give *external* standardization. In the cokeworks study we used external standardization with the administrative region as the standard population. In the latter case, the parameter e^α represents the overall relative risk between the standard (referent) region and the study region. External standardization is preferable since if internal standardization is carried out *a priori*, the effect of exposure may be distorted [21]; if data from the study region are used to obtain reference rates then the parameters β and γ should be estimated simultaneously. The standardized admission ratios are the maximum likelihood estimates (MLEs) from the saturated model $Y_i \sim \text{Poisson}(E_i \times \theta_i)$ and are given by $\hat{\theta}_i = Y_i/E_i$.

3.3. Form of risk/exposure relationship

Various approaches to modelling disease risk in relation to a point source have been suggested: see Reference [22] for a review. A very simple approach is to dichotomize the areas in the study region to those that are 'near' to the cokeworks, and those that are 'not near', where these categories were defined as $\leq 2 / > 2$ km in our study. This cut-off value was chosen to represent a plausible distance over which health effects would be seen [23]. In relation to the notation of Sections 3.1 and 3.2 we have, for area i ,

$$\exp\{f(\gamma, Z_i)\} = \begin{cases} \exp(\gamma) & \text{if } Z_i \leq 2 \\ 1 & \text{if } Z_i > 2 \end{cases}$$

where Z_i is the distance from the centroid of ED i to the cokeworks and $\exp(\gamma)$ is the relative risk associated with areas 'near' to a cokeworks. In the Poisson framework and using MLE this relative risk is equivalent to the ratio of standardized admission ratios. There are a number of obvious difficulties with this form, in particular the often arbitrary choice of the 'near' definition. Also, the appropriateness of this measure as a summary depends on proportionality of risk between EDs in the near and not-near areas.

We now describe parametric approaches to modelling risk. Various other approaches, including score tests and the use of linear risk scores, are described in Reference [22]. With reference to (2) the log-linear form $f(\gamma, Z_i) = \gamma Z_i$ may be assumed. There are many possible choices of Z_i , for example d_i or d_i^{-1} , where d_i denotes distance from the source of pollution. The choice of $Z_i = d_i$ has been criticised by Diggle and Elliott [24] since, with $\beta < 0$, $d_i \rightarrow \infty$ implies that $\theta_i \rightarrow 0$ and not to a background level of risk. The choice $Z_i = d_i^{-1}$ has the desired behaviour as $d \rightarrow \infty$ but gives infinite risk at source (when $\beta > 0$). This is not a problem for fitting unless the centroid of the closest area lies at the point source and the results may be very sensitive to the closeness of this area's centroid to the location of the point source, and so estimates of risk close to source may be unstable. There is also no estimate of the risk at source, which would be a useful summary.

Diggle *et al.* [25] assume the form,

$$\exp\{f(\gamma, Z_i)\} = 1 + g(d_i, \gamma) \quad (4)$$

with a variety of choices for $g(\cdot, \cdot)$ with the property that $g(d, \cdot) \rightarrow 0$ as $d \rightarrow \infty$. Lawson [26] and Lawson and Williams [27] consider a multivariate Z_i with non-monotonic and directional effects with respect to the point source.

If an exposure measure (for example SO_2 in the cokeworks study) is available then one may assume a log-linear (or more flexible) relationship in this measure.

3.4. Overdispersion

Extra-Poisson variability, in which the variance exceeds the mean, is often observed in spatial epidemiological studies using count data. This excess is due to a variety of sources including unmeasured covariates (confounders or otherwise), errors in the numerator and denominator data, within-area variability in exposures and covariates, and model misspecification. A simple method of accommodating non-spatial extra-Poisson variability is to adopt a quasi-likelihood (QL) approach (see Reference [28]) in which, for example, it may be assumed that $\text{var}(Y_i) = \kappa \times E[Y_i]$. Estimation is straightforward since the QL estimates are given by

the MLEs under the original formulation with the standard errors multiplied by $\sqrt{\hat{\kappa}}$ (with a similar adjustment to the deviance available if hypothesis tests are required). Under the QL approach the observations on different areas are assumed to be independent and so no adjustment is made for spatial dependence. This adjustment is carried out most easily using a random effects approach. To motivate this approach we now extend the model given by (2) to

$$\log p_{ij} = X_{ij}\beta + f(\gamma, Z_i) + U_i + V_i$$

where U_i and V_i represent random effects with and without spatial structure, respectively. These random effects may be interpreted as an attempt to control for the sources of excess-Poisson variability; more detailed explanations may be found in Reference [29], though we emphasize that random effects cannot in general control for confounders. Besag *et al.* [30] suggested a model with two random effects in a disease mapping context, and Clayton *et al.* [20] such a model in an ecological regression setting. In our context we obtain:

$$Y_i \sim \text{Poisson}(E_i \times \exp\{\alpha + f(\gamma, Z_i) + U_i + V_i\}) \quad (5)$$

Including V_i only in model (5) gives a parametric version of the QL approach. A typical choice is $V_i \sim_{\text{i.i.d.}} \text{N}(0, \sigma_v^2)$. Following Besag *et al.* [30] we use a Gaussian intrinsic conditional autoregressive spatial model for U_i in which $U_i | U_j, j \in \partial i, \sigma^2 \sim \text{N}(\bar{U}_i, \sigma_u^2/m_i)$, where ∂i represents the indices of a set of ‘neighbouring’ areas, m_i is the number of such neighbours, and \bar{U}_i is the mean of these neighbours. In the cokeworks study we took EDs to be neighbours if they have a common boundary.

The motivation for the introduction of spatial random effects by Clayton *et al.* [20] in an ecological regression context was to prevent ‘confounding by location’ though the introduction could also distort a true effect. Hence it is important to consider analyses with and without spatial random effects to examine the sensitivity of the exposure effect. The introduction of random effects also allows the possibility of obtaining more appropriate standard errors that reflect the non-spatial excess-Poisson variability and spatial dependence in the residuals. The exact form of dependence will often be of secondary importance when compared with confounding and within-area variability in exposure, and paucity of data will also limit the sophistication of any assumed spatial model (see Reference [31] for a fuller discussion).

4. MODELLING THE PROVIDER EFFECT

For notational simplicity we assume in this section that the only confounder is provider and that we have J providers. If within area i (e.g. ward) we could determine the population at risk by provider j , n_{ij} (the numbers of people in the catchment area of each provider), then provider adjustment would proceed exactly as for other discretized confounders such as age or deprivation quintile, for which population estimates may be obtained from the decennial census. In the case of provider, however, these populations are not available because people living in a typical ward will be admitted to a variety of hospitals and not just one, unless the ward is very remote. However, the numbers of admissions to each particular provider j from ward i , Y_{ij} , are available (from Hospital Episode Statistics data). It is clearly not sensible to allocate all the ward’s population to every hospital used by that population since we require

Table I. Notation for a generic area i , with J providers.

Provider	Admitted	Not admitted	Total
1	Y_{i1}	m_{i1}	n_{i1}
2	Y_{i2}	m_{i2}	n_{i2}
...
J	Y_{iJ}	m_{iJ}	n_{iJ}
Total	Y_i	m_i	n_i

The counts $Y_{i1}, \dots, Y_{iJ}, m_i$ and n_i are observed, the catchment populations n_{i1}, \dots, n_{iJ} are unobserved.

$\sum_j n_{ij} = n_i$. Some form of proportional allocation must therefore be used. In order to obtain rates and expected numbers, populations were estimated by methods we now describe. In Section 4.1, we describe how the denominators may be estimated in the case when the set of providers is known, and in Sections 4.2 and 4.3 we discuss, respectively, readmissions and how we decide upon a set of providers. In Section 4.4, we describe how control is carried out with the statistical model.

4.1. Estimation of catchment area population

We wish to estimate n_{ij} , the denominator for provider j in area i , where it is assumed there are J providers. We have $n_{ij} = Y_{ij} + m_{ij}$, where m_{ij} is the number of people not admitted, and $m_i = \sum_{j=1}^J m_{ij}$ is the total number in the area not admitted to any provider in area i (which is known). Table I summarizes the notation for a generic area i with J providers.

For all J providers, the number of people admitted (Y_{i1}, \dots, Y_{iJ}) follows a multinomial distribution $M_J(y_i, r_i)$ where $r_i = (r_{i1}, \dots, r_{iJ})$, denote the probabilities of admission to each provider in area i , given admission. We estimate r_{ij} by the observable quantities (and MLEs) Y_{ij}/Y_i , which we call the 'provider fractions'. Similarly, the number of people not admitted (m_{i1}, \dots, m_{iJ}) follows the multinomial distribution $M_J(m_i, p_i)$ where $p_i = (p_{i1}, \dots, p_{iJ})$ is unknown. We make the assumption that $p_{ij} = r_{ij}$ so that the people who are admitted are representative of those not admitted. Since $\hat{p}_{ij} = Y_{ij}/Y_i$ we have $\hat{m}_{ij} = Y_{ij}m_i/Y_i$. Hence we obtain

$$\hat{n}_{ij} = Y_{ij} + \frac{Y_{ij}m_i}{Y_i} = n_i \times \frac{Y_{ij}}{Y_i}$$

so that we have effectively made the simple and obvious assumption that the size of the provider population is proportional to the number of provider admissions. Implicit in this assumption is that there is no subgroup within each area, such as subarea, age or ethnic group, for whom admission to one provider is more likely than admission to another. If the assumption is incorrect, then the catchment area population estimates will be inaccurate and control for provider will be poor. In an extreme case, a variant on Simpson's paradox may occur, whereby the relation observed for subgroups is reversed by summing across them. This problem would arise, for example, if an area were physically divided by a large river, so that people on each side would only attend a hospital on the same side. Such a scenario is illustrated in Table II, in which 60 and 40 individuals are admitted to providers 1 and 2, respectively, and we would impute values of 1200 and 800, when in fact the true totals are

Table II. Illustration of Simpson's paradox.

Provider	Subarea 1			Subarea 2		
	Admitted	Not admitted	Total	Admitted	Not admitted	Total
1	60	440	500	0	0	0
2	0	0	0	40	1460	1500
Total	60	440	500	40	1460	1500
Provider	Admitted	Not admitted	Total			
1	60	1140	1200			
2	40	760	800			
Total	100	1900	2000			

The top two tables show the true data and the bottom those imputed under proportionality. From the observed data, which consist of 60 and 40 admissions to providers 1 and 2, respectively, and a total population of 2000, based on the assumption that non-admittance to each provider was in the same proportion as admittance, we would estimate the provider populations for providers 1 and 2 as (1200, 800). In truth they are (500, 1500) since, as seen in the top two tables, the admission rates in each provider depend on a third (confounding) variable, subarea.

500 and 1500, i.e. the provider fractions are switched. This is yet another illustration of the ecological fallacy. To overcome this in such an example it may be possible to examine the provider admission totals within each geographical subarea, e.g. EDs within a ward.

To assess this assumption we assumed that we wished to calculate provider populations at the district level. We calculated the percentage of emergency admissions in the under-5s to South Cleveland Hospital for each ward within two Teesside districts. This showed that there were several wards with much lower use of this provider than the district as a whole. Such heterogeneity of ward-level provider fractions within districts is likely to be more common in large towns and cities, where the greater population density leads to greater choice of hospitals. EDs will be more homogenous, at least geographically, in their hospitalization patterns, but the greater numerical instability at that level of aggregation led us to choose wards as the base unit of the catchment area (extensive investigations were also carried out by Bottle [32]). This trade-off, between large areas with more susceptibility to poor control for confounding and small areas with their associated instability, is very similar to that encountered when continuous confounders such as age are discretized and one must decide on the number of categories.

Recall that to control for provider we require both the number in each area who would attend each of the providers, and the rate of admission to each of these providers. Two remaining difficulties are to define the provider catchment area, in order to obtain rates of admission to that provider, and to decide on the number of providers to include for a particular study region. At first sight the first task would appear to be straightforward since we could simply take all the wards that provide at least one admission. Unfortunately such a definition includes areas that are geographically far from the provider and hence may differ from the study region in risk in respects other than provider alone. At the other extreme, including only the ward containing the provider would yield few admissions and very imprecise rates. To decide on

the number of wards to consider we examined *scree* plots in which the cumulative total admissions by ward (with the ward with the largest number of admissions being added first, and the second largest next, etc.) were plotted against the number of wards. As in principal component analysis, the 'elbow' was sought, but there were many possible choices of cut-off points from the scree plot for the examples considered here. We assessed instead the impact on the stratum-specific rates of choosing different sets of wards to form the provider's catchment area. For the under-5s, who form a single age group, there were two stratification variables other than provider: sex and Carstairs quintile. This gave ten stratum rates per provider. For every ward in immediately neighbouring counties (recall that these are geographically large), admissions and census populations were extracted. Of each ward's total admissions, the percentage going to the provider in question was calculated. The provider's rates were then derived from wards sending at least p per cent of their admissions to the provider, with $p = 1, 2, \dots, 30$. Stipulating immediately neighbouring counties was intended to exclude distant wards that the hospital was not built to serve, as such wards could not sensibly be included in its catchment area (the local hospitals were not national specialist centres). The p per cent criterion was designed to reveal any influence that low-use wards (such as those with $p \leq 5$) might have on stratum rates.

This approach was compared for the major providers for, firstly, the Teesside cokeworks group area and, secondly, for a randomly chosen district in Inner London (Wandsworth) to give a contrast between rural and urban settings. The influence of p is shown for St George's Hospital in London in Figure 2; rates remain stable only after about 6 per cent. In contrast, there was very little variation across the range of p for the Teesside providers (not shown). A minimum requirement for inclusion into the catchment area was therefore set at sending at least 10 per cent of the ward's admissions to the provider in question.

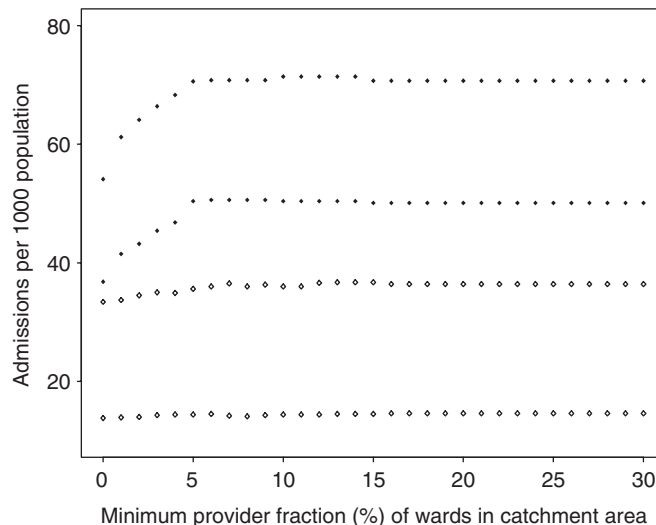


Figure 2. Emergency admission rates in the under-5s to St George's Hospital from wards with minimum provider fraction of between 0 and 30 per cent, most deprived areas (top lines, filled diamonds) and most affluent areas (lower two lines, open diamonds), boys above girls in each case.

Having obtained a list of wards forming the catchment area, the sets y_{ij} , and subsequently \hat{n}_{ij} , were obtained from EDs within each ward by applying the ward-level provider fraction to every ED population within the ward. In this way, the Carstairs quintile (CQ) could be measured at ED level rather than ward level, allowing for each ward to be heterogeneous in terms of deprivation, which is the usual case (in general, the smaller the areas chosen for control for confounders, the lesser the chance of ecological bias). External rates for use in the provider-adjusted standardized admission ratios (see Section 4.4) were then obtained by further subdividing stratum j by sex, in addition to CQ and provider, and summing across all wards in the catchment area. Hence we have assumed that the probability of admission to a provider does not depend on sex or deprivation.

4.2. Allowing for readmissions

The foregoing approach to estimating n_{ij} assumes that people are admitted just once, whereas Y_i and Y_{ij} are in fact unknown and estimated by the known quantities T_i and T_{ij} , respectively, which denote the total number of admissions during the study period, including all readmissions. By matching records on date of birth, sex and postcode, 'pseudo-patients' can be created and readmission rates c_{ij} (for ward i and provider j) and c_i (for the whole of ward i) calculated. We then have,

$$\hat{Y}_{ij} = T_{ij}/c_{ij}$$

and

$$\hat{Y}_i = T_i/c_i$$

The new (adjusted) estimate of n_{ij} is then,

$$\hat{n}_{ij} = \frac{T_{ij}n_i/c_{ij}}{T_i/c_i}$$

The catchment area populations using the unadjusted and adjusted methods were compared for two large London hospitals, and the adjustment made less than 3 per cent difference to the total. Consequently we did not make this adjustment for the cokeworks areas.

4.3. Determining the major providers for the study area

An initial look at how many hospitals are important to any area may be obtained from another scree plot, in which we plot cumulative admissions by provider (with the provider with the most admissions being plotted first, and the second most added next, etc.) against the number of providers. For example, for the near versus not-near analysis for the 0–2 km ring around the Teesside cokeworks, 90 per cent of emergency admissions in the under-5s went to South Cleveland Hospital, and 8 per cent to the South Tees Trust, with the remainder spread between several dozen other providers, which is typical. After the first few most important providers, the contribution to the total admissions of further providers falls rapidly. This example of the 'law of diminishing returns' leads to the concept of the *major provider*. Major providers for each study area included one or more named hospital Trusts (Trusts may have more than one hospital site, but these are not identifiable in routine hospital data) plus an artificial 'other provider', which accounted for remaining admissions (and population) not accounted for by

the named Trusts. An alternative would be to exclude these other admissions, but this would have led to the loss of over 10 per cent of cases, which was felt to be too large a proportion. The population of each ring was distributed amongst the major providers in proportion to their use of those providers (including 'other'), using all emergency admissions to do so. It was, however, checked that each provider recorded respiratory admissions—this was not the case for one Teesside hospital and it was therefore excluded. We now discuss how the major providers can be included as an extra stratification variable in indirect standardization and as an extra covariate in Poisson regression.

4.4. Including provider via standardization and regression

In this section we describe how we control for the provider effect. Here we need to adopt different indices for providers and stratum and let $p = 1, \dots, P$ index providers, and $j = 1, \dots, J$ stratum.

Following the procedure followed in Section 4.1 we have populations by providers, n_{ip} , and also populations by stratum, n_{ij} , by ED i .

Assuming that the provider fractions are constant across strata within area i we may obtain counts

$$n_{ijp} = n_{ij} \times \frac{n_{ip}}{n_i} \quad (6)$$

as the populations by stratum and provider. We then assume the model,

$$Y_{ijp} \sim \text{Poisson}(n_{ijp} \exp\{X_{ij}\beta + W_{ip}\delta_p + f(\gamma, Z_i)\}) \quad (7)$$

where X_{ij} and β are defined as previously, and W_{ip} is the $P \times 1$ vector with a 1 in position p , and $\exp(\delta_p)$ is the relative risk associated with provider p .

Summing (7) over stratum we obtain,

$$Y_{ip} \sim \text{Poisson}\left(\exp\{W_{ip}\delta_p + f(\gamma, Z_i)\} \sum_j n_{ijp} \exp(X_{ij}\beta)\right)$$

to give, using (6)

$$Y_{ip} \sim \text{Poisson}\left(E_i \frac{n_{ip}}{n_i} \exp\{W_{ip}\delta_p + f(\gamma, Z_i)\}\right)$$

where $E_i = \exp(X_{ij}\hat{\beta})$. Hence within the Poisson log-linear regression we simultaneously estimate $\delta_p, \dots, \delta_P$, and γ . As discussed in Section 3.2 this is essential in ecological regression settings since to control for provider using internal standardization would cause distortion of the exposure effect (if provider is a confounder). If provider rates could be determined from data that do not contain the study region, then *a priori* control can be carried out.

For disease mapping internal standardization can be performed, in which case we obtain expected numbers that control for both conventional confounders (e.g. age and sex) and provider.

5. RESULTS

In this section we present results from the cokeworks study that illustrate the issues described in Sections 3 and 4. Teesside is an area of considerable socio-economic deprivation, particularly near the cokeworks and nearby petrochemical industries. About 80 per cent of the population living within 2 km of the Redcar cokeworks fall within the fifth (most deprived) Carstairs quintile. Using distance as our initial measure of exposure, we first examined the possibility of confounding due to deprivation. The Carstairs score falls with distance from the Redcar site (as illustrated in Figure 3, bottom left). Deprived areas have higher admission rates for respiratory disease, particularly for those aged under 65 [8]. Figure 3 shows that respiratory admission rates in the under-5s decline with increasing distance from the Teesside cokeworks (top right), but the earlier discussions indicate that this observed effect may be confounded by deprivation (and provider effects), which is also seen to decline with distance. In the following, unless otherwise stated, the results have been adjusted for sex and CQ, with the reference rates being obtained from the administrative region; the bottom-right plot in Figure 3 shows that the sex- and CQ-adjusted standardized admission ratio (SAR) shows no clear relation with Carstairs score, as we would expect.

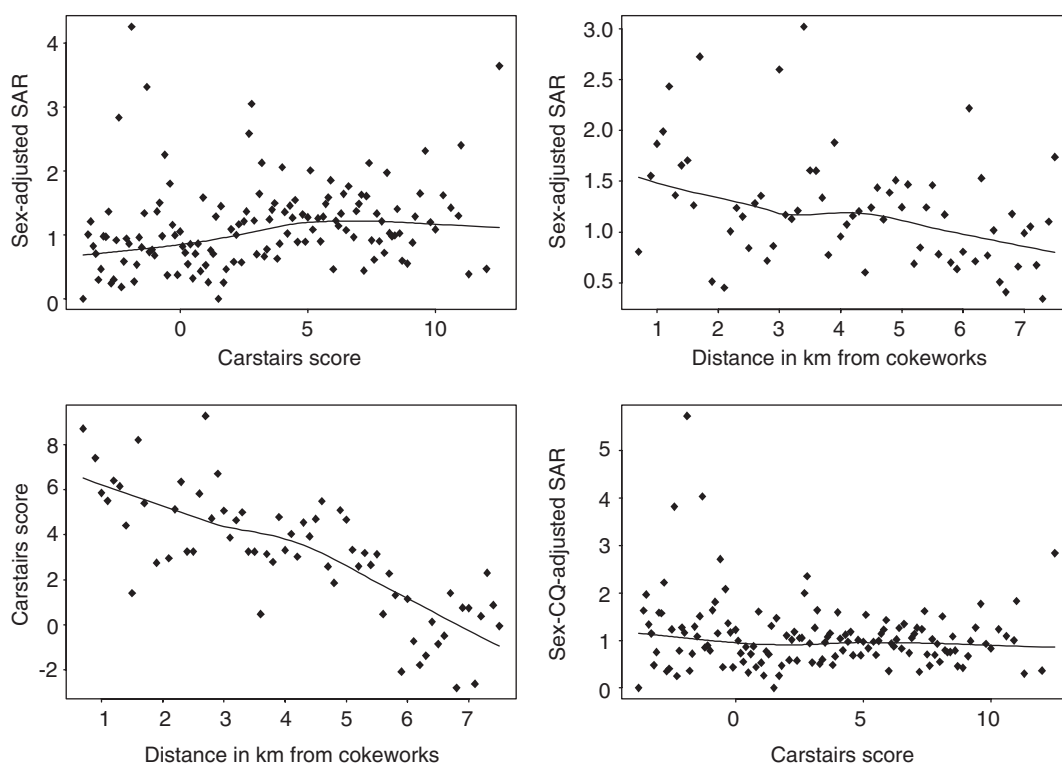


Figure 3. Plots showing the confounding effect of deprivation (Carstairs score) on all respiratory rates in the under-5s living near the Teesside cokeworks (SAR = standardized admission ratio).

Table III. Exponentiated parameter estimates and their 95 per cent confidence intervals for various geographical exposure proxies from models of respiratory admissions in the under-5s near Teesside cokeworks.

Model	Exposure proxy term	Parameter (e^{β}) estimate	95 per cent CI	Change in quasi-deviance from no-effect model
Distance-only:	Distance (km) only	0.92	0.90–0.95	23.3
	Inverse distance only	1.92	1.55–2.38	32.1
	Inverse square of distance only	1.66	1.39–1.98	26.9

The parameter estimates are for a 1 km increase in distance. Deviances and confidence intervals have been adjusted for overdispersion.

5.1. 'Near' versus 'not near'

There were 1351 children under 5 living with 2 km of the Teesside cokeworks (the 'near' distance ring), and 15 756 living between 2 and 7.5 km away (the 'not-near' ring). The near ring had 387 respiratory admissions, with an SAR of 2.14; the not-near ring had 2661 respiratory admissions, with an SAR of 1.44. The ratio of the near ring SAR to the not-near ring SAR was 1.49, with a 95 per cent confidence interval of 1.29–1.72. As described in Section 3.3, this model has a number of drawbacks but as an exploratory first step it is useful.

5.2. Parametric risk models

Table III gives parameter estimates and standard errors for various functions of spatial location in relation to the Teesside cokeworks for all respiratory admissions in the under-5s. Forms fitted were distance, inverse distance and inverse squared distance. Standard errors were calculated using the quasi-likelihood method, with the overdispersion parameter $\hat{\kappa}$ being between 1.7 and 1.8 for all models shown. All distance models were statistically significant (judged either through the confidence intervals excluding unity or via quasi-likelihood ratio tests), suggesting that there is an increased risk near the Teesside cokeworks, with an 8 per cent drop when moving a distance of 1 km away.

5.3. Modelled SO_2 values

Figure 4 shows the SO_2 estimates produced from the Atmospheric Dispersion Modelling System (ADMS [33]) package, which incorporates factors such as chimney output and weather conditions. ED centroids in the study area are shown as dots. The ranges shown do not represent actual levels because background SO_2 levels were not included in the ADMS (these present further difficulties beyond the scope of this article). We used a log-linear model in SO_2 and, as long as the background level is constant across the study region, our relative risks will be valid. We chose to discretize the SO_2 measure into four groups to allow non-linearities. As well as SO_2 estimates, we controlled for sex, ED-level Carstairs quintile (CQ) and major provider of admission, of which there were four, including the 'other' provider.

Table IV gives the relative risk of admission for respiratory illness for children under 5 living amid the large Teesside industrial complex, with the lowest modelled (ED-level) SO_2 range taken as the comparator level. The numbers in each SO_2 exposure group drop off as

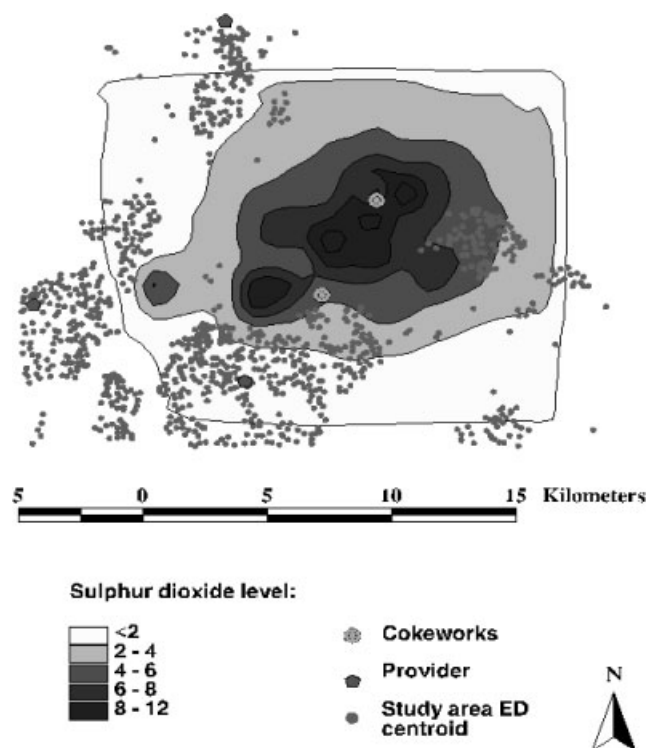


Figure 4. Sulphur dioxide estimates in $\mu\text{g}/\text{m}^3$ from dispersion modelling for Teesside enumeration districts, also showing major providers.

Table IV. Relative risks and standard errors for respiratory admissions at different SO_2 levels in the Teesside industrial complex region using a quasi-likelihood model for overdispersion, a non-spatial random effects only model (heterogeneity model) and a spatial and non-spatial random effects model (convolution model).

SO_2 level ($\mu\text{g}/\text{m}^3$)	Number of cases	Population	Quasi-likelihood approach		Heterogeneity model		Convolution model	
			RR	SE	RR	SE	RR	SE
<2	2834	53 341	1	—	1	—	1	—
2-4	1629	21 715	1.22	0.04	1.16	0.06	1.09	0.08
4-6	499	8701	0.95	0.07	0.94	0.09	0.93	0.11
6-8	94	682	1.94	0.14	2.03	0.33	1.94	0.33

the levels increase, which is reflected in the standard errors of the associated risk estimates. There was a significant approximate doubling of risk at the highest estimated pollution level compared with the lowest with adjustment, though a monotonic pattern across levels was not observed in the estimates.

Using the quasi-likelihood (QL) approach, the dispersion parameter was 1.8. Assuming the

Table V. The effect of omitting deprivation and/or provider from a model of all respiratory admissions in the under-5s in the Teesside industrial complex; RR = relative risk.

SO ₂ (µg/m ³)	RR with full adjustment	RR without CQ	RR without provider	RR without CQ or provider
<2	—	—	—	—
2–4	1.22	1.24	1.37	1.41
4–6	0.95	0.91	1.11	1.08
6–8	1.94	2.19	2.19	2.58

extra-Poisson variation to be spatially unstructured, the heterogeneity random effects model was fitted, giving similar estimates to, but higher standard errors than, those from the QL method. When spatial random effects were included (in addition to non-spatial effects, giving the convolution model) they accounted for about a quarter of the extra-Poisson variability and the standard errors were much larger than for the two simpler models, reflecting the loss of information in this case.

The residual deviance with sex, CQ and provider included in the model was 3291.6 on 3447 degrees of freedom, falling to 3219.8 (a statistically significant change of -72.8 on 3 degrees of freedom) when adding SO₂ as a factor. The addition of SO₂ as a linear term (giving a relative risk of 1.08 per unit increase in pollutant level) rather than as a factor reduced the residual deviance by 15.8. We prefer the model with the factored pollutant since it allows more flexibility for detecting non-linearities and thresholds in the exposure-risk relationship. Unfortunately we were restricted in our ability to detect the form of the relationship because of paucity of information at higher levels of SO₂.

To illustrate the effect of deprivation and provider on the observed relation between pollution and risk of hospitalization, Table V gives the relative risks for each SO₂ level estimated from the QL model when CQ and/or provider were omitted from the model. Omission of provider increased the estimated relative risk for all three exposure groups. Omission of CQ appreciably increased all three estimates. We note that the inclusion of a provider effect makes a considerable difference to the risk estimates and makes interpretation difficult because we may not have adequately controlled for confounding by provider.

We conclude that there is evidence of an association between respiratory risk and SO₂ which warrants further investigation, but the lack of a monotonic relationship, problems with background SO₂ levels and confounding and the usual potential for ecological bias means that it is not possible to apply a causal interpretation or reliably quantify the strength of any effect. The analyses that we have performed indicate that it is important to consider (in descending order of priority) confounding by provider and deprivation, overdispersion, and the possibility of residual spatial effects.

6. DISCUSSION

Standardized admission ratios (SARs) and Poisson regression suggested a higher risk of hospitalization near the Teesside group than further away. The original study had cokeworks as the putative cause of any observed increased risk, but dispersion modelling found that their

emissions amounted to only about a tenth of the total produced by heavy industry in the area; distance from cokeworks had been found to correlate well with distance from Teesside petrochemical plants, emphasizing the need for better exposure data. This article, however, is primarily concerned with the analysis and control for the provider effect.

As a first look at the event rates, SARs are often obtained. They are simple to calculate but, even if proportionality holds, the reference region may not be appropriate, e.g. local variations in hospital use suggest that catchment areas allow better comparisons. Catchment areas, however, are smaller than administrative regions and will have less precise stratum rates. Another difficulty is that for some providers the study region population represents a non-negligible proportion of the provider catchment area. To avoid this, one may compare the study area with the non-overlapping part of the catchment area, though this will of course further reduce the precision of the provider's rates. Imprecision of the standard rates leads some to use direct standardization instead, although use of this will be hindered by imprecise study area rates. Our results were often very similar for both unless the number of cases was small, e.g. < 10 ; this is an example of the well-known trade-off between precision and bias.

There was a reasonable amount of spatially structured variation in residual respiratory admission rates in the under-5s, which inflated the standard errors compared with the simpler QL model. We used common boundaries but could have taken as neighbours all EDs within a certain distance. However, that approach would not have allowed for the strong effect of the river Tees demarcating the catchment areas of South Cleveland Hospital and the nearby South Tees Acute Hospitals Trust.

Residual deviances were not too different between the models involving distance, with the smaller deviance for $1/d$ offset by the greater robustness of the estimate for d . We did not use the non-linear forms of Diggle *et al.* [25] due to problems of instability and sensitivity to choice of initial parameter values [22]. Each parametric model suggests increasing risk with distance from cokeworks for the Teesside site, though the data were not of sufficient quality to make any firm conclusions of the exact form.

Table V shows that provider is a potentially important confounder in small area studies and that provider effects were appreciable when modelling respiratory admission rates on Teesside. The method described here involved first the identification of enough hospitals to account for local utilization patterns, while at the same time minimizing the stratification to preserve stability. This trade-off may be assessed by sensitivity analyses. Second, the estimation of provider-specific denominators by ward was required. With provider-specific population estimates, one may then estimate a provider's admission rates to obtain expected numbers via internal standardization for disease mapping studies, or use provider as a covariate in regression models for studies such as that described here. To derive our estimates we calculated the provider fraction for each ward. The assumption is that there is no subgroup within each area for whom admission to one provider is more likely than admission to another. In theory, this will be invalid if a ward is split by, for example, a large river, but in practice ward definitions tend to make use of such natural boundaries and this is unlikely to bias the results greatly, particularly when compared with other sources of bias (data anomalies, exposure misclassification, ecological bias, confounding, etc.). However, it is quite possible that some unknown factor is operating to reduce the validity of this approach, and, as ever with observational studies, this cannot be determined by routine data. It is more likely that differences in the quality of coding between nearby hospitals, especially under-recording by one and double-counting by the other, will occur. This may bias an individual hospital's estimated rate either

up or down. Figures from the Office for National Statistics suggest that the system captures at least 99 per cent of admissions in recent years, with any one provider rarely dropping below 95 per cent coverage.

Analyses were performed at an aggregate level. As we have noted earlier, the interpretation of such ecological studies is difficult due to within-area variability in exposure measures and confounders. Interpretation is only more straightforward when large relative risks (for example, greater than 3) are observed after control for known confounders. A small but epidemiologically significant relative risk (such as 1.1) is unlikely to be elicited by an ecological study. An alternative study design would be to perform a case-control analysis, since postcodes have known co-ordinates (here the postcode would approximate to an individual's house or, moreover, to where an individual is assumed to receive their exposure), but this is not possible from routinely available data, and case-control studies have their own difficulties that include selection bias.

An exposure proxy such as distance is more frequently available than real exposure data, but the validity of its use should be assessed, which of course may not always be possible. The use of modelled pollutant estimates as the exposure measure is in its infancy but is likely to increase. We used ED-level SO₂ estimates because values are not available at finer resolution, but this inevitably leads to misclassification of an individual's exposure, as it does when using distance; this is before problems of daily movement patterns are considered (although for the under-5s the use of residential location may be less of a problem). Advances in ADMS should allow the estimation of SO₂ levels at greater geographical resolution, but the sophistication of statistical methods still exceeds the quality and availability of exposure data in small area studies.

ACKNOWLEDGEMENTS

The Small Area Health Statistics Unit is funded by a Grant from the Department of Health; Department of the Environment, Food and Rural Affairs; Environment Agency; Health and Safety Executive; Scottish Executive; National Assembly for Wales; and Northern Ireland Assembly. The views expressed in this publication are those of the authors and not necessarily those of the funding departments, data providers, or of the Office for National Statistics.

REFERENCES

1. Elliott P, Wakefield JC, Best NG, Briggs D (eds). *Spatial Epidemiology: Methods and Applications*. Oxford University Press: Oxford, 2000.
2. Bhopal RS, Phillimore P, Moffat S, Foy C. Is living near a coking works harmful to health? *Journal of Epidemiology and Community Health* 1994; **48**:237–247.
3. Dolk H, Thakrar B, Walls P, Landon M, Grundy C, Saez Lloret I, Wilkinson P, Elliott P. Mortality among residents near cokeworks in Great Britain. *Occupational and Environmental Medicine* 1999; **56**:34–40.
4. Dolk H, Pattenden S, Vrijheid M, Thakrar B, Armstrong B. Perinatal and infant mortality and low birth weight among residents near cokeworks in Great Britain. *Archives of Environmental Health* 2000; **55**(1):26–30.
5. Wakefield JC, Elliott P. Issues in the statistical analysis of small-area health data. *Statistics in Medicine* 1999; **18**:2377–2399.
6. Aylin P, Bottle A, Wakefield J, Jarup L, Elliott P. Proximity to coke works and hospital admissions for respiratory and cardiovascular disease in England and Wales. *Thorax* 2001; **56**(3):228–233.
7. Carstairs V, Morris R. *Deprivation and Health in Scotland*. Aberdeen University Press: Aberdeen, 1991.
8. Dolk H, Mertens B, Kleinschmidt I, Walls P, Shaddick G, Elliott P. A standardisation approach to the control of socio-economic confounding in small area studies of environment and health. *Journal of Epidemiology and Community Health* 1995; **49**(Suppl 2):S9–S14.

9. Campbell DA, Radford JMC, Burton P. Unemployment rates: an alternative to the Jarman index? *British Medical Journal* 1991; **303**:750–755.
10. Kleinschmidt I, Hills M, Elliott P. Smoking behaviour can be predicted by neighbourhood deprivation measures. *Journal of Epidemiology and Community Health* 1995; **49**(Suppl 2):S72–S77.
11. Walters S, Phupinyokul M, Ayres J. Hospital admission rates for asthma and respiratory disease in the West Midlands: their relationship to air pollution levels. *Thorax* 1995; **50**:948–954.
12. Clemence L. To whom do you refer? *Health Service Journal* 1998; **23**:26–27.
13. Piantadosi S, Byar DP, Green SB. The ecological fallacy. *American Journal of Epidemiology* 1998; **127**:893–904.
14. Greenland S, Morgenstern H. Ecological bias, confounding, and effect modification. *International Journal of Epidemiology* 1989; **18**:269–274.
15. Richardson S, Stucker I, Hemon D. Comparison of relative risks obtained in ecological and individual studies: some methodological considerations. *International Journal of Epidemiology* 1987; **16**:111–120.
16. Greenland S. Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* 1992; **11**:1209–1223.
17. Greenland S, Robins J. Ecological studies—biases, misconceptions and counterexamples. *American Journal of Epidemiology* 1994; **139**:747–760.
18. Richardson S, Montfort C. Ecological correlation studies. In *Spatial Epidemiology: Methods and Applications*, Elliott P, Wakefield JC, Best NG, Briggs DB (eds). Oxford University Press: Oxford, 2000; 205–220.
19. Wakefield JC, Salway R. A statistical framework for ecological and aggregate studies. *Journal of the Royal Statistical Society, Series A* 2001; **164**:119–137.
20. Clayton D, Bernardinelli L, Montomoli C. Spatial correlation in ecological analysis. *International Journal of Epidemiology* 1993; **22**:1193–1202.
21. Breslow NE, Day NE. *Statistical Methods in Cancer Research II*, Chapter 4. International Agency for Research on Cancer: Lyon, 1987.
22. Morris SE, Wakefield JC. Assessing of disease risk in relation to a pre-specified source. In *Spatial Epidemiology: Methods and Applications*, Elliott P, Wakefield JC, Best NG, Briggs D (eds). Oxford University Press: Oxford, 2000; 152–184.
23. Thomas B, *et al.* Pollution at cokeworks: joint report of investigations into the measurement of polycyclic aromatic hydrocarbons and benzene, toluene and xylene in and around cokeworks. Commission of the European Communities, Luxembourg, Belgium, EUR 13196 EN, 1991.
24. Diggle PJ, Elliott P. Statistical issues in the analysis of disease risk near point sources using individual or spatially aggregated data. *Journal of Epidemiology and Community Health* 1995; **49**:S20–S27.
25. Diggle PJ, Morris SE, Elliott P, Shaddick G. Regression modelling of disease risk in relation to point sources. *Journal of the Royal Statistical Society, Series A* 1997; **160**:491–505.
26. Lawson AB. On the analysis of mortality events associated with a prespecified fixed point. *Journal of the Royal Statistical Society, Series A* 1993; **156**:363–377.
27. Lawson AB, Williams FLR. Armadale: a case study in environmental epidemiology. *Journal of the Royal Statistical Society, Series A* 1994; **157**:285–298.
28. McCullagh P, Nelder PA. *Monographs on Statistics and Probability. Generalized Linear Models* (2nd edn). Chapman & Hall: London, 1989; 168–180.
29. Wakefield JC, Best NG, Waller L. Bayesian approaches to disease mapping. In *Spatial Epidemiology: Methods and Applications*, Elliott P, Wakefield JC, Best NG, Briggs DB (eds). Oxford University Press: Oxford, 2000; 104–127.
30. Besag J, York J, Mollie A. Bayesian image restoration, with application in spatial statistics with discussion. *Annals of the Institute of Statistical Mathematics* 1991; **43**:1–59.
31. Wakefield JC. Sensitivity analyses for ecological regression. *Biometrics* 2002; **59**:9–17.
32. Bottle RA. Adjustments for the provider effect using hospital data in small area studies. *Ph.D. Thesis*, Imperial College, London, 2001.
33. ADMS-Urban version 1.53. An urban air quality management system. *User Guide*. Cambridge Environmental Research Consultants Ltd., November 1999.