Modeling the Impact of Traffic-Related Air Pollution on Childhood Respiratory Illness

Nicola G. Best[*], Katja Ickstadt[†], Robert L. Wolpert[‡],
Samantha Cockings[*], Paul Elliott[*], James Bennett[*],
Alex Bottle[*] and Sean Reed[*]

Revised November 9, 1999, 12:09

ABSTRACT Epidemiological studies of the health effects of outdoor air pollution have suffered from a number of methodological difficulties. These include major problems of estimating exposure; confounding due to socioeconomic deprivation and other factors; failure to account for possible spatial dependence induced by unmeasured covariates with smooth spatial variation; and the potential for bias arising from approximation and aggregation in the data. We present a flexible modeling framework using recently developed Bayesian spatial regression methods to address some of these issues. We apply this approach to a study of the relation of London traffic pollution to the incidence of respiratory ailments in infants using the recently available Hospital Episodes Statistics (HES) dataset.

Key words: Ecological bias, Environmental exposure, Hospital admissions, Marked point process, Poisson regression, Spatial correlation.

1 Introduction

The possible harmful health effects of outdoor air pollution from industrial sources and road traffic are the focus of great public concern and scientific attention in the United Kingdom. Indeed, the goal of reducing the harmful health impact associated with outdoor air quality is a principal issue raised in the UK government's "National

[*] Department of Epidemiology and Public Health, Imperial College School of Medicine at St. Marys, London, UK

[†] Department of Statistics, University of North Carolina, Chapel Hill, NC, USA

[‡] Institute of Statistics and Decision Sciences, Duke University, Durham, NC, USA

Environmental Health Action Plan", drawn up under the auspices of the World Health Organization. Key scientific questions affecting UK policy decisions are whether and how recent apparent increases in the rates of childhood respiratory illness are related to atmospheric pollution arising from motor vehicle emissions.

The evidence for such a link remains equivocal, due in large part to a host of methodological difficulties that have beset scientific investigations of the relationship. These include major problems of estimating exposure; confounding due to possibly unknown or unmeasured risk factors; and the potential for bias arising from approximation and aggregation in the data. These methodological obstacles must be addressed to enable quantitative assessment of the effects of traffic pollution on health can be made available to inform public health policy, transport planning and environmental legislation regulating vehicle emissions.

In this Case Study we present an epidemiological investigation of hospital admission rates for respiratory illnesses in children under one year old living in Greater London, and relate these rates to measures of exposure to traffic-related air pollution. We address the scientific question of whether childhood respiratory illness rates are associated with motor vehicle emissions by studying four specific issues:

1. Do infants living in areas of high traffic density have higher rates of respiratory illness?

2. Which pollutants or vehicle types are associated with the greatest risk?

3. Do factors such as socioeconomic deprivation of the area of residence, or individual characteristics such as gender increase or modify this risk?

4. Can we account for residual spatial confounding due to unknown or unmeasured risk factors?

To answer these questions we construct a hierarchical Bayesian semi-parametric spatial point-process regression model supporting inference about the relationship of children's respiratory illness rates to traffic pollution and other risk factors. We address the exposure estimation problem by employing sophisticated geographical information system (GIS) methods to build exposure extrapolation models; we approach the confounding problem by including in our model a latent spatially varying risk factor; and we reduce or avoid the problem of ecological bias by using each source of data at its natural level of spatial resolution, without any unnecessary approximation or aggregation.

In the next section (2) we provide a detailed discussion of the epidemiological background to this Case Study. The available data, including our use of GIS methods to develop suitable measures of exposure based on the raw data (traffic flows and vehicle emission factors) are described in Section 3–4, and an exploratory analysis of the data is presented in Section 5. In Section 6 we review briefly existing methods for the epidemiological analysis of exposure-response relationships and discuss some of the associated methodological problems. This will motivate the modeling approach described in this section and in Section 7. Section 8 contains a a final discussion.

2 Epidemiological Background

There is evidence of marked geographical variation in the occurrence of many diseases, including cancers and respiratory illnesses (Gardner et al., 1984; Parkin et al., 1997). In particular, urban-rural differentials have been noted, with higher rates of disease in urban areas (Greenberg, 1983; Howe et al., 1993), and a strong association typically observed with measures of socioeconomic deprivation (Jolley et al., 1992). One possible explanation is differences in smoking patterns. Another is the possible adverse health effects of urban air pollution derived from road traffic emissions. Indeed, the search for possible environmental causes of ill health, including a putative effect of traffic related air pollution, is currently a major international focus of epidemiological research.

Several factors have contributed to these developments. One is the apparent increase in prevalence of respiratory illnesses, especially in children, which has characterized almost all countries in the western world, and many developing countries, over the last two to three decades (Anderson et al., 1994; Burney, 1988). Another is the inexorable rise in traffic volumes and associated emissions of air pollutants, which has led to serious declines in air quality in many cities

(HMSO, 1994). A third is the increasing scientific evidence for a link between exposure to traffic pollution and a range of health effects, including cardiorespiratory mortality and morbidity and reduction of lung function (Committee of the Environmental and Occupational Health Assembly of the American Thoracic Society, 1996).

Much of the growing literature on the health effects of ambient air pollution has focused on the acute effects of exposure; recent reviews include those by Schwartz (1994) and Brunekreef et al. (1995); Anderson et al. (1998) provide a summary of time series studies of air pollution and hospitalization for asthma. These findings provide relatively clear evidence of an association between respiratory illness and a range of pollutants (Committee of Medical Effects of Air Pollution, 1995a,b). Specific attention has focused on the effects of fine particulates (Schwartz, 1994; Dockery et al., 1993; Pope et al., 1995).

While, on balance, it is considered that these acute associations with air pollution represent real effects, the underlying mechanisms remain unclear; they may reflect exacerbations of pre-existing disease rather than the occurrence of new events (Committee of Medical Effects of Air Pollution, 1995a,b). To what extent, if any, the findings of the acute studies can be extrapolated to the possible effects of chronic exposure to air pollution is unknown. From a public health perspective, these potential chronic effects are vitally important because of the large numbers of people potentially exposed, *i.e.*, there may be large population attributable risk due to low levels of traffic-related outdoor air pollutants, even if the excess relative risk of such exposures is small. The limited evidence from chronic lung function studies suggests a small but significant negative association with particulate air pollution (Schwartz, 1989; Chestnut et al., 1989). There is also some evidence of a link between PM_{10} exposure and prevalence of chronic respiratory symptoms, especially in children (Schwartz, 1993; Dockery et al., 1993; Roemer et al., 1993). However, there are contradictory findings from a number of recent studies examining relationships between exposure to markers of traffic-related pollution and chronic respiratory health. For example, while several have reported positive associations between, variously, respiratory symptoms, hospitalization or lung function of children living close to roads compared to those living further away, or living in streets with high levels of traffic, self-reported truck-traffic or high predicted NO_2 levels (Murakami et al., 1990; Brunekreef et al., 1997; Osterlee et al., 1996; Wjst et al., 1993; Weiland et al., 1994; Duhme et al., 1996; Edwards et al., 1994; Pershagen et al., 1995), others have reported no association (Livingstone et al., 1996; Waldron et al., 1995; Briggs et al., 1997).

Unfortunately, most population studies of the health effects of chronic exposure to traffic-related pollution suffer from major problems of exposure assessment. Of the studies cited above, many used non-specific indicators such as proximity to main roads. Other studies which have focused on specific pollutants, such as particulates, nitrogen dioxide or ozone, have generally used levels of pollutants recorded at local monitoring stations to classify subjects into exposure groups. The key assumption underlying these approaches is that cumulative personal exposure correlates with a static measure of average ambient pollution concentrations or source activity (road traffic) in the vicinity of the individual's home location. The validity of this assumption clearly depends on the spatial resolution of the exposure measurements and the time-activity pattern of the subjects concerned. A number of studies have found poor correlation between measurements of ambient air quality from outdoor monitoring stations and indoor air pollution and/or personal exposure (for example, see Urell and Samet (1993) and Ozkaynak and Spengler (1996) and references therein). However, there is some evidence to suggest that residential indoor air quality and personal exposure of children both show a reasonable correlation with measures of ambient air pollution (nitrogen dioxide concentration) and road traffic density close to the home (Shima and Adachi, 1998; Hoek et al., 1989). Compared with working adults, children are likely to spend a significantly greater proportion of their time in the microenvironment where they live. It is therefore reasonable to assume that exposure estimates based on home locations are more likely to reflect personal exposure to air pollution in children than adults. There is also evidence that it is possible to reliably estimate either community average or individual multi-day (chronic) personal exposures using available outdoor air quality measurements, even when the cross-sectional correlation between acute personal exposure and ambient air quality is poor (Mage and Buckley, 1995; Wallace, 1996).

Other recent approaches to air pollution exposure assessment include time-activity studies (Noy et al., 1990), which typically lead to more reliable estimates of personal exposure. However, the expense and time demands of such studies makes the design infeasible for a population-based analysis. Vostal (1994) discusses the use of probabilistic models to estimate the chance of individual exposure to high pollution concentrations. This approach requires data on specific activity patterns for population sub-groups and is more relevant for estimating acute rather than chronic exposure to high levels of air pollution.

Other difficulties faced by epidemiologists studying the effects of chronic exposure to traffic-related pollution on respiratory health include the potential for confounding by risk factors which are unrelated to pollution but exhibit similar spatial patterns (*i.e.*, are correlated with pollution exposure). Such risk factors may include smoking, home environment (indoor air quality, presence of allergens etc.), and socioeconomic status. Seasonal effects, weather variables and infectious disease epidemics may also represent important risk factors for respiratory illness. However, to confound studies of the *chronic* effects of exposure to traffic pollution, the average or typical values for these variables would have to show spatial variation in a similar way to the the pollution measures, which is unlikely. By contrast, weather and seasonal variable are major confounders for acute studies since they will tend to be correlated with short-term temporal variations in air quality.

2.1 Work leading to the present study

Given the above discussion, it is clear that great care is required when designing, analyzing and interpreting studies of the health effects of chronic exposure to traffic-related air pollution. Young children represent a particularly suitable group to study in this context for a number of reasons. Their respiratory and immune systems are immature, making them particularly vulnerable to the potentially harmful effects of air pollution; their activity patterns tend to be localized so that the bulk of their exposure to traffic-related pollution may be reasonably assumed to occur within the neighborhood of their home; and by definition, their cumulative lifetime exposure takes place over a relatively short period, thus eliminating concerns about long latency times and the need for historical exposure data.

This Case Study therefore aims to investigate the effects of exposure to vehicle emissions on chronic, non-malignant respiratory health in young children. Our motivation arises from a previous analysis of hospital admissions for childhood respiratory illnesses in relation to traffic density in London, UK, which was carried out by some of the present authors (Best et al., 1998a). Hospital admissions represent a valuable and as yet under-exploited source of routine health data with which to investigate population morbidity in the UK. These data have only recently been made available to us (see Section 3 for further details), and our original analysis considered small area variations in admission rates by electoral wards in the Thames region (which includes London). Four health outcomes were considered: admissions for all respiratory illnesses in children aged less than one year and 1–4 years, bronchiolitis admissions in children under one year old and asthma admissions in 1–4 year old children. Information on exposure to traffic pollution was obtained from measured and modeled estimates of vehicle flows on all major and minor roads in Greater London (see Section 4). These were averaged over electoral wards to provide a measure of average annual traffic density in each small area. These data were analyzed using ecological regression models to relate ward-level counts of hospital admissions to the ward level summaries of traffic density and other area-level covariates such as socioeconomic deprivation. A Bayesian hierarchical modeling approach was adopted, using log-linear Poisson regression with Gaussian conditional autoregressive (CAR) smoothing priors (Besag et al., 1991) to capture the effects of unmeasured confounders and allow for spatial dependence in residual disease risk between neighboring areas. Our findings provided some evidence that, for infants under one year old, risks of respiratory illness, and specifically bronchiolitis, were related to high density of medium and heavy goods vehicles in the ward of residence. This is in accord with previous findings of reduced lung function in children related to diesel traffic (*e.g.*, Brunekreef et al. (1997)).

2.2 The need for further analysis

The Bayesian ecological regression approach used above is becoming increasingly popular as a method for investigating the relationship between geographical variations in disease and suspected environmental and other risk factors (Richardson et al., 1995; Bernardinelli et al., 1997; Knorr-Held and Besag, 1998). This approach offers the advantages of requiring only environmental exposure averages over geographical areas (which are often available even when individual level exposure data are not) and of exploiting spatial correlation in the data. However, inference may be sensitive to the choice of geographical units selected for analysis, and aggregation may distort or mask the true exposure-response relationship for individuals - a phenomenon known as ecological bias (see Section 6.1). By averaging traffic density over electoral wards (which range in area from 0.3 km^2 to 23.8 km^2 in London), it is possible that important local variations in exposure may be obscured in such a way as to distort the analysis. Since we are in the fortunate (and unusual) position of having exposure data available at a high level of spatial resolution, one of our aims in the present Case Study is to develop methods which are better able to exploit the raw data at their natural level of aggregation.

One advantage of using traffic density rather than measures of individual pollutants as the exposure estimate in our original study is that the former may be viewed as a proxy for the synergistic effect of the complex mix of pollutants emitted by vehicle exhausts. Nonetheless, investigation of the effect of specific substances such as particulates and acid gases like nitrogen dioxide is important if we are to better understand the mechanisms underlying the putative relationship between traffic-related pollution and respiratory health. Emissions data for a range of compounds associated with vehicle fuel combustion have recently been linked to our London road traffic database and these additional variables will also be included in the present Case Study.

The models we develop in this Case Study therefore aim to address some of the methodological limitations outlined above and extend our original analysis to provide a more accurate assessment of the effects of traffic pollution on children's health. For practical reasons, we choose to focus on just one of the respiratory outcomes investigated in our original analysis, that is, hospital admission for bronchiolitis in infants under one year old. A brief description of this illness is given below, together with a discussion of the possible factors influencing the likelihood of hospitalization for a child with bronchiolitis.

2.3 Hospital admissions for bronchiolitis

Bronchiolitis is a common disease of the lower respiratory tract of infants, resulting from inflammatory obstruction of the small airways.

It occurs during the first 2 years of life, with a peak incidence at approximately 6 months of age. In many areas, it is the most frequent cause of hospitalization of infants: in our original study, 14% of all emergency admissions, and 37% of respiratory-related emergency admissions for children under one year old in London were diagnosed as bronchiolitis. The illness is viral and occurs both sporadically and epidemically; incidence is highest during winter and early spring.

In order to model the relationship between exposure to traffic pollution and incidence of bronchiolitis, we need to consider the sequence of events which must occur before a child is admitted to hospital with such a diagnosis. Fig. 1 shows a schematic representation of this pathway. The first stage represents potential risk factors which may provoke an episode of bronchiolitis in an exposed child. An infant with bronchiolitis will first exhibit symptoms of mild upper respiratory tract infection such as runny nose and sneezing, which may be accompanied by fever. After several days, wheezy cough and dyspnea may develop. In mild cases, symptoms usually disappear in 1 to 3 days, but in more severe cases the child will be in severe respiratory distress and the parents may decide to seek medical help. This decision is represented by stage 2 in Fig. 1. The parents may take the child directly to the Accident and Emergency (A&E) department of a local hospital, or may first seek advice from their family doctor (called a General Practitioner or GP). This choice is represented by stage 3 in Fig. 1, and may be influenced by factors unrelated to the health status of the child. These include the relative accessibility of hospital and GP services, the parents' perception of the quality of care offered in each case and the parents' relationship with their GP. If the child is first taken to the GP, there are two subsequent pathways: either the GP will decide to treat the infant, in which case no record of the episode will appear in our hospital admissions database. Alternatively, the GP will refer the child to hospital (stage 4). Again, the factors governing the GP's decision may be unrelated to the health status of the child, but typically reflect the degree of diagnostic uncertainty, the facilities and treatments available to the GP, and the GP's training and experience and relations with hospital consultants. However, infants with respiratory distress should usually be hospitalized, hence most children with bronchiolitis who are first taken to their GP will end up in hospital. Given that the child presents at hospital (by whichever route), a decision will then be made as to whether or not to admit him/her for treatment (stage 5). At this stage, a formal diagnosis should be made and recorded in the patient's notes. In most cases, the diagnosis will be bronchiolitis, although the illness may be confused with bronchial asthma or bronchopneumonias. It is also possible that infants presenting with other respiratory symptoms will be mistakenly diagnosed as having bronchiolitis. The final stage (6) in our pathway from exposure to recorded diagnosis of bronchiolitis in our hospital database is the data entry process, whereby the diagnosis entered in the patient's notes is transcribed onto the computerized database. The potential for errors at this stage is discussed in more detail in Section 3.2.

The complexity of factors influencing the pathway shown in Fig. 1 offers clear potential to distort our investigation of the relationship between exposure to traffic-related pollution and risk of hospitalization for bronchiolitis in young children. While our routine data sources do not provide sufficient information to learn about every stage shown in the figure, we attempt to account for as many aspects as possible in the statistical models described in Section 6–7. The possible mechanisms underlying this pathway should also be bourne in mind when attempting to interpret the results of our analyses.

3 Health, Population and Sociodemographic Data

The data used in this study were obtained from several different sources and were recorded at a variety of different geographic scales. In this section we define the study region and describe the health event data, population data and sociodemographic data and method of linkage; the traffic-related exposure data is described later in Section 4.

3.1 Study region geography

The study region chosen for analysis comprises the part of North West Thames Regional Health Authority bounded by the River Thames on the south and by the M25 London orbital motorway on the north and west. This subset of the Thames region studied earlier (Best et al., 1998a) was chosen to maintain a geographically diverse study area so that spatial features might emerge, but reduce the number of health events to a manageable enough size to permit

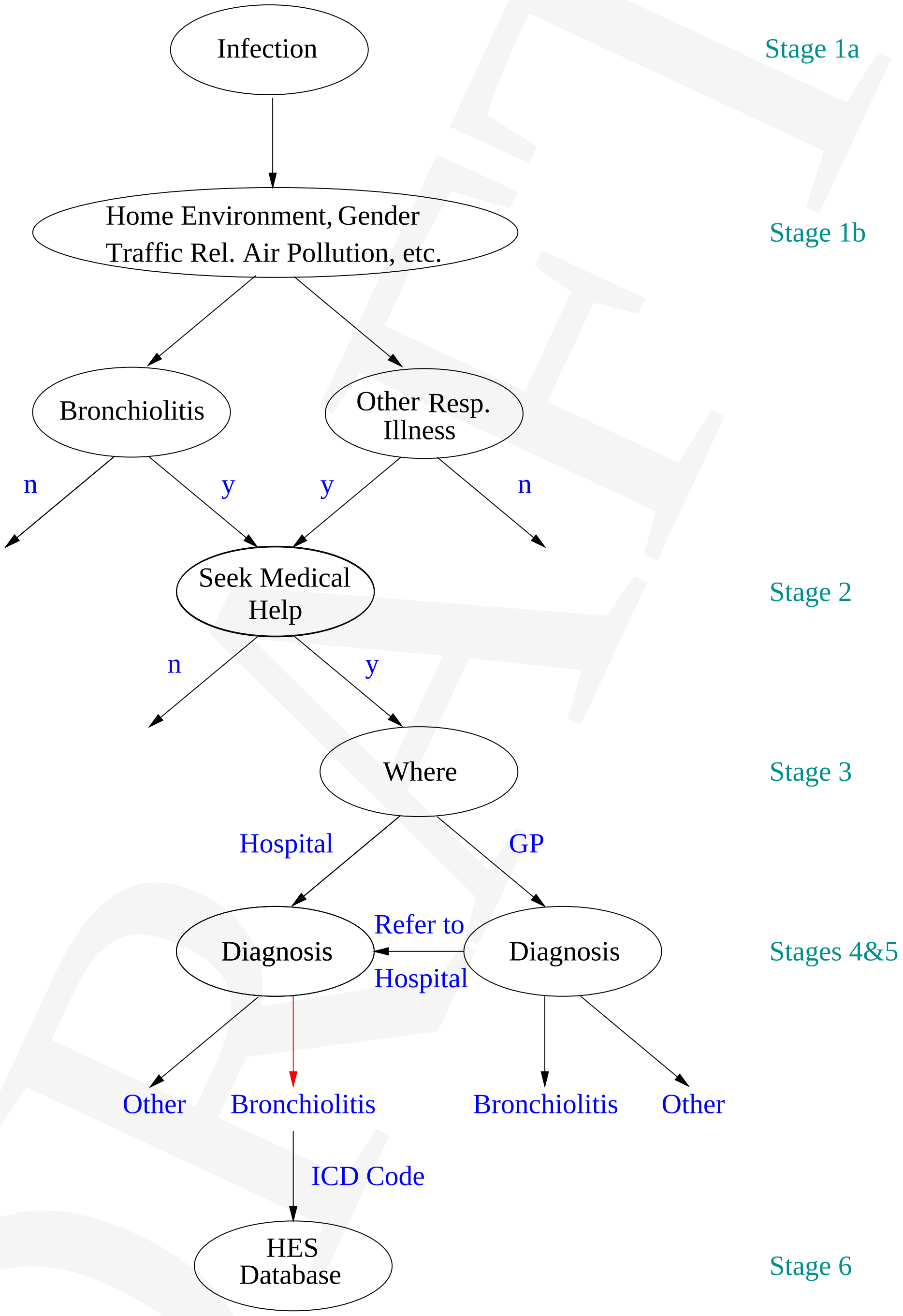

FIGURE 1. Schematic representation of the stages involved between a child developing bronchiolitis, being admitted to hospital with such a diagnosis, and the diagnosis being recorded in the Hospital Episode Statistics (HES) database used in our study.

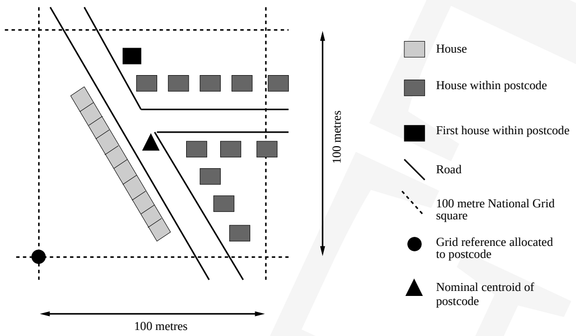

FIGURE 2. Example of how the postcode coordinates and nominal centroid are defined for one postcode.

a more detailed spatial analysis. Additionally, traffic exposure data were only available to us on roads within the M25.

Each form of spatial data we use has its own level of spatial aggregation. At the finest resolution we have the postcode of residence for each child. Postcodes are used for the delivery of mail within the UK (Raper et al., 1992). Each building or residence in the UK is assigned to some postcode. Approximately 1.4 million of the approximately 1.6 million UK postcodes are comprised entirely of residential addresses, with a median of twelve residences per postcode. Areas covered by a single postcode have no formal boundaries and can vary widely in size: in urban areas they are typically a few hundred meters across, although rural postcodes often cover a much larger area. The easting and northing (x and y) coordinates of the south-west corner of the 100m National Grid square containing the "first" residence (this is not systematically defined) of each postcode are available from the Central Postcode Directory (`http://census.ac.uk/cdu/Datasets/Lookup_tables/Postal/Central_Postcode_Directory.htm`); we convert these to a nominal centroid by adding 50m to both coordinates, as suggested in Gatrell et al. (1991). Fig. 2 shows an example of how these postcode coordinates and centroid are defined for one postcode.

The Census and many other socioeconomic data in the UK are available only at aggregated levels of geography. The smallest geographical unit for which Census data are available is termed the *Enumeration District*, or ED, of which there are 151,719 in the UK. The geography of EDs does not reflect the underlying socioeconomic characteristics of the population; rather, they are administrative boundaries arbitrarily imposed for ease of collection of the Census data. The geographical extent of EDs therefore varies considerably, with rural EDs often much larger than urban ones. On average an ED contains around 10–15 postcodes. A postcode can be linked uniquely to the ED within which its centroid is located through the use of annually updated Postcode to Enumeration District lookup tables (`http://census.ac.uk/cdu/Datasets/Lookup_tables/Postal/Postcode_Enumeration_District_Directory.htm`), available from the data archive at the Manchester Information and Associated Services (MIMAS) (`http://mimas.ac.uk`).

The next larger unit of geographical aggregation in the Census hierarchy are termed electoral wards, of which there are 10,933 in the UK. EDs nest exactly within wards and it is therefore straightforward to identify which ward an ED lies within, and also to link postcodes to wards. Within the study area defined for this Case Study, there are 269 wards, 5,060 EDs and 74,622 postcodes.

3.2 Health events

Hospital activity data have been collected in the UK since 1949 from all National Health Service (NHS) hospitals, originally as paper records and more recently in digital format. The current computerized reporting system, known as the Hospital Episode Statistics (HES), came into operation in 1987. The HES database is maintained by the UK Department of Health with a copy held by the Small Area Health Statistics Unit (SAHSU) in the Department of Epidemiology and Public Health at Imperial College, UK. We use the HES database to obtain information on hospital admissions for infants living in the study region.

It is generally recognized that the HES transition period of 1987–91 was disruptive, compromising data coverage, completeness and accuracy (Williams et al., 1994). The improvement in accuracy in recent years probably reflects the fact that, since 1991, HES data have been used for contracting in the internal NHS market. In the UK, hospitals act as 'providers' who supply medical services to 'purchasers' such as other hospitals and general practitioners, using information recorded in the HES database to bill the purchasers.

Background to HES data

The basic unit of the HES database is the *finished consultant episode*, or FCE, covering the period during which a patient is under the care of one consultant. Every NHS hospital in England has to submit electronically the forty or so HES data items for each FCE in each patient's hospital stay. The data items are entered to the hospital's own digital Patient Administration System by trained clinical coders working with either patient notes from the ward trolleys or with discharge summaries. The required data items include some demographic data, such as date of birth, sex and postcode of residence, but to ensure confidentiality they do not include the patient's name, NHS number, or any other unique identifier. Clinical data such as primary and secondary diagnoses and dates and details of any operations performed within the patient's stay are also recorded on the HES system. Diagnoses are coded using the International Classification of Diseases (ICD) codes, often with the help of computer software. Clinical coding at this crucial stage is somewhat subjective and is not subject to any external audit, so clearly there is some scope for inter-coder variation (Walshe et al., 1993; Dixon et al., 1998). The HES database grows by some ten million records per year.

HES data quality

There are three principal aspects of data quality:

- Coverage: the proportion of total hospital inpatient activity recorded in the database;
- Completeness: the proportion of records that have an entry in any specified field;
- Accuracy: how well the database entries reflect the true value of a particular field.

Coverage: In addition to electronic entry of inpatient episode data for HES, hospitals are obliged to submit paper counts (called KP70s) of patient episodes, collected independently of the HES counts. The Department of Health use KP70 totals as the 'gold standard' against which HES is compared. Discrepancies are investigated by the Department. Williams (1997) found that HES was nationally within

1.2% of the KP70 figure for the period 1989–95, and Sheldon et al. (1994) generally found HES and KP70 totals to be in close agreement.

The HES database includes only NHS (*i.e.*, publicly supported) admissions, systematically excluding all private hospitals and private patients treated in NHS hospitals. Unlike in the USA however, the majority of patients in the UK still receive NHS treatment and so will be captured on the HES database if admitted to hospital. The proportion of emergency admissions in young children that are treated privately is also likely to be small.

Completeness: HES data are subject to a large number of automated validation checks during data cleaning (Department of Health, 1998). A recent descriptive analysis of the HES data held by SAHSU found close to 100% completeness for the sex, date of birth, postcode and admission date fields, and over 98% completeness for the primary diagnosis field. However, some hospital Patient Administration Systems insert the vaguest ICD code ('799', in version 9 of the ICD) by default in the absence of any specified condition, thereby having fewer missing primary diagnoses but commensurately more completely ambiguous codes (McKee and Petticrew, 1993).

Accuracy: Most studies that have investigated the accuracy of clinical coding relate to data collected before the current HES system was introduced. One recent study focused on the reproducibility of clinical codes at two large acute hospitals within the former North West Thames region between 1991 and 1993, comparing local, external (with at least four years' experience) and the most senior coding manager (Dixon et al., 1998). Exact (four-digit ICD code) agreements on the main diagnosis were 41% at one hospital and 59% at another hospital; the figures for approximate (three-digit ICD code) agreement were 52% and 69% respectively.

Diagnostic ICD codes are not the only fields subject to error. In particular, the HES database demands that the patient's postcode field not be left blank. This leads to the use of "dump" codes, such as the postcode of the patient's general practitioner (GP) or the hospital's own postcode, when the patient's postcode is unknown. Such a practice may lead to false local 'clusters' of disease if the postcode is used to locate cases in a spatial analysis of HES data.

In summary, the HES database provides a relatively complete record of all NHS hospital admissions in England. In particular it includes essentially all emergency admissions for 0–1 year old children in our study region. The fraction of missing fields within each record is also negligible. However, the (in)accuracy of the data values entered on the HES database is of much greater concern, especially for the primary diagnosis code. Our modeling approach will therefore attempt to address the problem of inter-hospital variation in diagnostic coding and the use of vague ICD codes.

Data used for the Case Study

HES data are collected by fiscal year, which runs from April to March in the UK. Due to concerns over data quality for early years of HES data, we restrict our analysis to the three year period from April 1992 to March 1995. This period also coincides with the time frame over which the road traffic exposure data were collected (see Section 4). Data for more recent years are still being checked and cleaned by the data providers and are not yet available to us.

The subset of data considered for this study therefore comprises all emergency hospital admissions in the period April 1st 1992 to March 31st 1995 for children aged under one year on the day of admission and whose postcode of residence lay within the northwestern part of London that defines our study region. We defined an admission to be the first episode of each hospital stay which ended within the study period. Thus a child admitted on two separate occasions (stays) and having two separate first episode entries would constitute two admissions in our analysis. Conversely, a child who was admitted just once during the study period but was referred to a number of different consultants would acquire more than one episode but would only constitute one admission. Ideally, we would have preferred to carry out a person-based analysis, but in the absence of a unique patient identifier in the HES database it is not possible to link multiple admissions for the same patient. We considered trying to link such admissions by matching records (either exactly or probabilistically) according to the sex, date of birth and postcode fields. However, potential problems with coding errors and changes of address (postcode) were deemed to make this approach unreliable. Our data are thus admission-based, and so are not strictly independent (the same child may be admitted multiple times during our study period). Nonetheless, we argue that it is a reasonable working hypothesis to assume independence for modeling purposes, since a rough estimate based on exact matching of sex, date of birth and postcode suggested that repeat admissions comprised less than 20% of our data.

Data items extracted from the HES database for our study subset included postcode, sex, age, primary and secondary diagnosis codes and hospital. Before classifying the subset into groups for analysis, the following exclusion criteria were applied:

- All admissions for accidents and poisonings —this criterion was imposed to minimize possible confounding between the putative effect of exposure to traffic-related air pollution associated with living close to busy roads, and the increased risk of accidents on or near such roads;
- All birth-related admissions —this criterion was imposed since birth-related admissions were deemed to be qualitatively different from other types of infant emergency admissions.

The remaining admissions were then classified into three groups according to the primary diagnosis field:

Bronchiolitis cases (B):	ICD-9 code 4661
Non-respiratory controls (N):	Any ICD-9 code except codes 460–519 (respiratory diseases) and those covered by the exclusion criteria above.
Unspecified diagnosis (U):	ICD-9 code 799

This gave a total of 2,107 cases of bronchiolitis, 9,904 non-respiratory controls and 1,220 admissions with an unspecified primary diagnosis. As a crude check on the use of "dump" postcodes in these data we calculated the number of admissions with a postcode corresponding to that of the hospital to which the child was admitted: only 5 such admissions were identified, suggesting that there is no systematic tendency for hospital postcodes to be used when the patient's postcode is unknown.

3.3 Population denominator data

In order to make inference about population disease rates we require an estimate of the spatial distribution of the population at risk, *i.e.*, the distribution of 0–1 year old children living in our study region.

There are several problems with the obvious approach of using the most recent UK census counts. In particular, the population counts vary from year to year, and census data are only available for 1991 while our study period covers 1992–1995; furthermore, the smallest aggregation unit by which census counts are available is the ED, whereas our health event data are geo-referenced to the much finer postcode level. We therefore elected to use the national register of births as our source of population data. Registration of births is mandatory in the UK, and all records, including the sex, date of birth and home postcode of the baby, are held on a database by the UK Office for National Statistics. A copy of this database is also held by SAHSU. These data are thus available at the same spatial resolution and cover the same time period as the health events; they are also likely to be more complete than the census counts.

For the purposes of this Case Study, we therefore employ as a population estimate the number of babies born in each postcode between October 1st 1991 and September 30th 1994. This gave a total of 159,706 births. If no child moved house during this period, this would represent the cumulative number of children aged 0–1 year living in each postcode on the mid-point of each financial year (September 30th) of our study period. Clearly some migration of families with newborn babies will occur, but we expect this to represent only a small fraction of the total. Often the family will only move house within the same residential area and so interpolation or aggregation of the birth totals over localized areas (as is done in Section 5) should reduce errors in our population estimates due to migration.

3.4 Sociodemographic and other covariate data

In addition to the putative effect of traffic-related air pollution on respiratory health in young children, there are a number of other variables which may represent important risk factors for respiratory illness and which may confound any relationship with exposure to traffic pollution. In particular, the indoor environment of the child's home may represent an important source of exposure to air pollution: combustion of fossil fuels from any unvented source (*e.g.*, gas stove or water heater) and parental cigarette smoking may raise concentrations of pollutants such as nitrogen dioxide (NO_2) and fine particulates (Dockery et al., 1981; Quackenboss et al., 1986, 1991).

Mould/mildew associated with damp housing conditions may also exacerbate risk of respiratory illness (Strachan, 1991). Boys have generally been found to be more susceptible to respiratory illnesses than girls (Strachan et al., 1994), and illness rates may also vary by race. Seasonal variations in temperature and pollen counts may also affect the risk of respiratory illness (Anderson et al., 1998); however, the present study is concerned with the effects of *chronic* exposure to traffic-related air pollution, and so short-term temporal variations in risk are less relevant.

Unfortunately, the routine datasets used to obtain the health events and population data for this study contain little individual-level covariate information. For the present purposes, the only relevant individual attribute recorded with each birth or hospital admission was gender. For the health event data, the name and postcode of the hospital to which the child was admitted were also recorded as individual attributes for possible study of hospital effects.

In the absence of individual-level covariates, we must attempt to use aggregate measures of the relevant risk factors to reflect the 'typical' value of that covariate among individuals in the group. For this purpose, we group individuals according to their area of residence. Census data, available at the ED level, may then be used to provide information on the ethnic mix and socioeconomic characteristics of each area (ED). Specifically, we calculate the proportion of 0–4 year old children (the finest age strata recorded on the Census) in each ED from each of the following ethnic groups: White, Black (including Black African and Caribbean), Asian, Indian (including Pakistani and Bangladeshi) and Other. Specific information on smoking rates and housing conditions are not available from the Census. However, the Carstairs measure of socioeconomic deprivation (Carstairs and Morris, 1991) has been shown to correlate strongly with cigarette consumption (Kleinschmidt et al., 1995) and may also be viewed as a proxy for poor housing conditions and lifestyle factors which are often correlated with ill health (Jolley et al., 1992).

Accessibility to hospital and to primary care (GP) facilities is likely to influence a parent's decision on choice of care for a sick child. The University of York carried out a small area analysis of hospital utilization rates (Carr-Hill et al., 1994) and derived measures of hospital and GP accessibility for electoral wards in England. These scores are distance-weighted counts of the numbers per capita of hospital beds and GPs, respectively. Maps showing the geographical distribution of these covariates are given in Section 5.

4 Traffic Exposure Data

In this section we describe the raw data and methods used to derive appropriate measures of exposure to traffic-related pollution at two levels of geographical resolution (lines and grids).

Geographical Information Systems (GIS) are tools which enable the integration, manipulation, analysis and display of datasets which are geographically referenced. GIS techniques are increasingly being used within health-related disciplines and applications due to the inherently spatial nature of many health issues ((Gatrell and Senior, 1999); Smith and Jarvis (1998); Cummins and Rathwell (1991)).

A GIS database containing the spatial datasets for this Case Study was developed using the GIS software packages ARC/INFO and ArcView (`http://www.esri.com/`). The raw datasets were obtained at varying levels of resolution, generalization and accuracy and thus various GIS techniques were required to assemble, integrate, overlay and process them. The road traffic GIS database involved three main datasets: a *stick network* of modeled flows and emissions on (mostly) major roads; a set of $1km\times1km$ *squares* containing modeled flows and emissions on minor roads; and a *vector dataset* of roads in London.

4.1 Traffic flow data

Major roads (stick network)

The London Research Centre (LRC) produces modeled estimates of traffic flows from road traffic in major urban areas, including London. Major roads are loosely defined as roads for which traffic flow data are available from the London Transportation Studies (LTS) model (Buckingham et al., 1997). In practice, this includes most roads with a traffic flow in excess of 400 vehicles per hour (during a morning weekday peak period). In central London this gives intense coverage of roads, while in the outer areas it is mostly only main roads (*i.e.*, A roads and motorways) which are included. The data are modeled for a series of *links* or 'sticks' *i.e.*, straight lines connecting specified start and end nodes.

The LRC uses output from the LTS transport model as their basic activity data. The LTS model is a well-established tool, widely used for transport planning in London, and maintained on behalf of the Government Office for London (GOL) by the MVA Consultancy. The LTS model is the only source of comprehensive traffic flow data available on a London-wide scale and on an internally-consistent basis. The model gives coordinates for the start and end points of the links in the stick network; figures for the flow of cars, light vehicles and goods vehicles per hour per period; and link length and transit time. Validation of such models is non-trivial, however the LTS model has been validated against traffic surveys and appears to perform reasonably well (MVA Limited, 1999).

LRC he The LRC processes this basic activity data in a number of ways to build up a detailed picture of traffic flow on major roads in London. Initially the network is clipped to include only those links within and including the M25 orbital motorway. Average link speed is then calculated using the distance/transit time data. The basic vehicle-type figures from the LTS model (*e.g.*, motorcycle, car, heavy goods vehicle) are then refined to reflect the likely distribution of more specific vehicle types (*e.g.*, petrol engined car of < 1.4 liters and without catalyst). This is done with reference to Department of Environment Transport and Regions (DETR) traffic counts averaged across parts of the study area. Vehicle fleet statistics are also used to determine the proportion of petrol/diesel vehicles and the proportion of cars with catalytic converters.

The above data on modeled traffic flows for major roads on the stick network were supplied to us by the LRC for use in this Case Study. Further processing was required to produce estimates of the total annual traffic flows on each line segment for use in our statistical models. The data were provided as typical hourly bi-directional flows for weekdays during morning-peak (7am to 10am), evening-peak (4pm to 7pm) and off-peak (10am to 4pm) periods. These were converted to annual flows for each major road using factors provided b for nightly flows (weighted by whether the road was in Central, Inner or Outer London) and for Saturday and Sunday flows. Bi-directionality was removed simply by summing the total flow for each link.

These estimates gave the annual flows for 1991 but they may also be used for more recent years since most of the major roads in London (except perhaps the M25) were considered to have reached capacity by 1991 and could not carry any more traffic.

Minor roads (grid squares)

LRC also produces modeled traffic flows on minor roads for 1km $\times$ 1km squares (based on the National Grid) approximately covering the area within the M25 motorway. A minor road is simply defined as one for which link-based traffic flow data is not available *i.e.*, a road which is not part of the assigned highway network used in the LTS model.

Detailed traffic flow data are rarely available for minor roads, therefore traffic flows usually have to be estimated on an aggregated basis. This means that the minor roads are essentially treated as area sources. The LRC estimates the total vehicle-kilometers driven on minor roads by taking an estimate of the total vehicle-kilometers driven on the entire road network and subtracting the kilometers already accounted for by the major roads in the LTS model. The estimate for the entire network is obtained from the DETR Rotating Traffic Census. This gives a total traffic flow for all minor roads in the network which must then be apportioned across the grid squares. The LRC achieves this by estimating road network density using a combination of the Ordnance Survey's OSCAR database of road center-lines and manual estimates from the Geographer's A-Z . The total kilometers for each grid square is then further broken down into the various vehicle categories as per the major roads.

For this Case Study, modeled traffic flow data were obtained from the LRC for the following vehicle categories: motorcycles, petrol cars, diesel cars, buses, petrol light goods vehicles, diesel goods vehicles, and medium and heavy goods vehicles. The data provided were yearly totals for 1995.

4.2 Emissions data

Traffic flows are a useful proxy for exposure to traffic-related pollution. However they do not provide information about the loads of specific pollutants to which the population is exposed. Ideally, this sort of information would be obtained either by air quality monitoring of ambient pollution levels or by monitoring of the pollutant load being received by individuals. In practice, however, air quality monitoring is expensive and only gives point-based observations, whereas emissions from traffic are along lines (*i.e.*, roads). Individual monitoring is also costly and can therefore usually only be carried out on a small number of subjects. In addition, monitoring does not identify the sources of the pollutants. An alternative method to using monitored pollution concentration data is to employ atmospheric emissions inventories. An air pollution emissions inventory is a schedule of the sources of pollutants within a particular geographical area (Buckingham et al., 1997). A comprehensive emissions inventory will usually include not only pollutants from traffic but also emissions from other mobile sources (such as rail, rivers and air) and emissions from stationary sources such as industrial processes, and waste disposal.

In the UK, a National Atmospheric Emissions Inventory (NAEI) is maintained by the National Environmental Technology Centre (NETC) on behalf of DETR. The NAEI contains data on a wide range of pollutants and is published for 10km × 10km grid squares for the UK. The NAEI, however, is a *top down* inventory whereby national data is allocated to smaller areas on the basis of the resident population and other indicators of activity levels. By comparison, the LRC produces detailed atmospheric emissions inventories for specific urban areas. These are *bottom up* inventories which compile data at the local level and are generally available at a finer level of geographical resolution than the national inventory.

In producing local emissions inventories, measures of activity relating to the emissions are collected and emissions factors are then applied to the activity data in order to estimate the likely emissions:

$$\text{Activity rate} \times \text{Emission factor} = \text{Emission rate}\,.$$

The LRC has produced an atmospheric emissions inventory for London (the London Atmospheric Emissions Inventory, LAEI, (Buckingham et al., 1997) which is available at the 1km × 1km square level. The activity data used in the inventory are derived from a wide range of sources and different time periods, but the nominal year of the inventory is 1995.

For this Case Study we focus on traffic-related emissions which comprise the major source of air pollution in London. Emissions from traffic on major and minor roads and also cold starts and hot soaks (see Section 4.2) were obtained from the LAEI from the LRC for the following pollutants: sulphur dioxide (SO_2), nitrogen oxides (NO_x), carbon monoxide (CO), carbon dioxide (CO_2), non-methane volatile organic compounds (NMVOC), benzene, 1,3-butadiene, black smoke, total suspended particulates (TSP), particulate matter of less than $10\mu m$ aerodynamic diameter (PM_{10}) and methane (CH_4).

Major roads

The LRC calculates modeled levels of emissions for each link in the stick network described in Section 4.1. Using the traffic flows described above, emissions factors from the UK Emissions Factors Database (`http://www.london- research.gov.uk`) are applied to each link depending on the vehicle flow, speed and road type of the link. This gives the total amount (in metric tonnes per annum) of each pollutant type emitted from road traffic for each individual link in the stick network.

Using GIS line-in-polygon techniques the LRC also re-apportions the emissions data from the stick network to the $1\text{km} \times 1\text{km}$ squares, based on the length of each link contained within each grid square. For each pollutant, this gives the total emissions from traffic on major roads per grid square in tonnes per annum per km^2.

Minor roads

The LRC uses a different technique to calculate emission levels for the minor roads because data on average speeds are not available for these roads and so speed-related emissions factors cannot be used. Instead, typical (urban) driving-cycle based emissions factors (from the UK Emissions Factors Database) are employed to estimate pollutant specific emissions (in tonnes per annum per km^2) on minor roads in each $1\text{km} \times 1\text{km}$ square.

Cold starts/Hot soaks

When vehicles are first started they emit higher levels of pollutants than when they are at their normal operating temperature. These emissions occur at the start of trips (called *cold starts*) and need to be estimated separately from the major and minor roads emission levels. This is done by calculating the number of trip-starts in each grid square using information about trip origins and destinations from the LTS model. The LTS trip-matrix gives trip-ends data for a series of polygons which the LRC then re-allocates to grid squares based on the proportion of each polygon's area contained within each grid square. The number of trip-starts are then multiplied by relevant emissions factors to give total pollutant loads arising from cold starts for each grid square. These calculations are based on the assumptions that all cold start emissions occur within 1km of the trip start and that average ambient temperatures, catalyst performances and times since vehicle was last operated apply.

Similar methods are used to estimate emissions from *hot soaks* which occur as evaporative emissions when the engine is switched off. Benzene and NMVOCs are the only pollutants relevant to hot soaks.

Total emissions from all road traffic by 1km $\times$ 1km square are obtained by adding the estimated cold start and hot soak pollutant loads to the estimates for minor and major roads.

4.3 Exposure measures used for the Case Study

Major roads

One option for the traffic-related exposure measures used in this Case Study is to take the traffic flow and pollutant loads on the stick network and 1km $\times$ 1km squares as supplied by LRC. However, we were concerned that, at the spatial resolution of our intended analyses, there may be important inaccuracies in these data. Specifically, the stick network from which the LRC data are derived is only a straight line approximation to the major roads network. The stick network represents a simplified model of the true location of roads and, as such, may contain inaccuracies which are non-negligible when the data are used for detailed small-area spatial analyses.

Simple examination of the degree of error suggests that some roads may be as much as 1km away from their true location; the average discrepancy is approximately 100–200m. We therefore obtained the Bartholomew digital road data (`http://www.bartholomewmaps.com`) for London. This is a very detailed dataset taken from original 1:5000 source data for London roads, and the degree of resolution and accuracy make it a much more realistic representation of the true road network than the LRC stick network.

Fig. 3 shows the LRC stick network and the Bartholomew roads. The inset presents an example of the discrepancies between the two networks in terms of the generalization and accuracy of road locations.

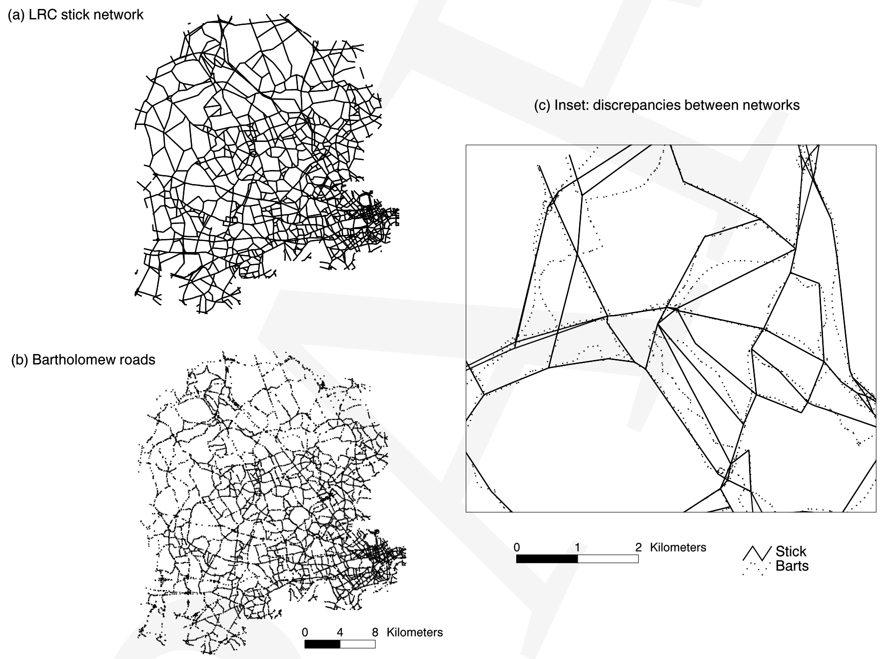

FIGURE 3. Road networks: (a) LRC stick network, (b) Bartholomew roads. Inset (c) demonstrates differences in networks in terms of generalization and accuracy of road locations.

Our goal was to link the LRC stick network with the Bartholomew road data and to then transfer the traffic flow and emissions data from the stick network onto the Bartholomew roads. The straight line links of the stick network did not match the Bartholomew roads sufficiently well to allow fully automated matching. A combination of automated and manual editing facilities within ARC/INFO was employed to manually link the sticks to the roads. This linkage of the stick and road networks was a large undertaking, but it was deemed to be worthwhile because of the potential usefulness of the dataset for the current study and future health-related research within the London area.

The matching work was shared between the Department of Epidemiology and Public Health at Imperial College (DEPH) and the London School of Hygiene and Tropical Medicine (LSHTM). The area was segmented into 5km grids (following the Bartholomew tiles) and a common strip down the middle was defined for both institutions to link. This allowed comparison of the links created by the two institutions, as a measure of quality control. Visual comparisons of the linkages for the whole area were also made at DEPH.

Using this link file, the attribute information on the stick network was transferred onto the Bartholomew road coverage, to produce annual traffic flow data for all major roads in London. For the subsequent analyses, the traffic flows were grouped into two vehicle categories: i) cars and light vehicles *(CLGV)*, and ii) medium and heavy goods vehicles *(MHGV)* (where buses were included in medium and heavy goods vehicles).

Using GIS *line-in-polygon* techniques, we also transferred the attribute data from the Bartholomew roads onto the set of 1km $\times$ 1km squares.

Minor roads

The minor roads data were supplied pre-processed by the LRC on the 1km $\times$ 1km grid squares. For the purposes of the subsequent analyses the traffic flows were aggregated into the same two vehicle categories as the major roads, namely: *cars and light vehicles* (comprising motorcycles, petrol cars, diesel cars, petrol light goods vehicles, diesel light goods vehicles) and *medium and heavy goods vehicles* (comprising buses and medium and heavy goods vehicles).

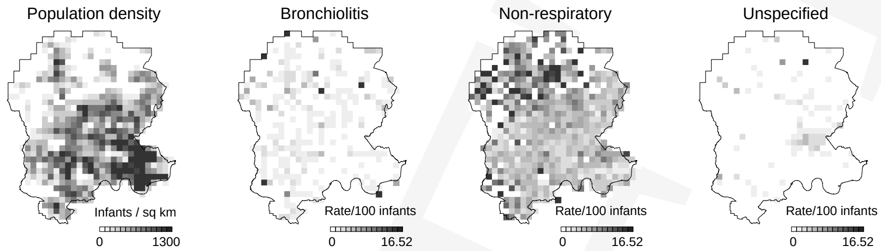

FIGURE 4. Population density (infants per square kilometer) and raw admission rates (per 100 population) in the training database for bronchiolitis, non-respiratory illness and unspecified causes.

Emissions data

Eventually we aim to link the emissions data from the stick network to the Bartholomew roads in the same way as for the traffic flows. However, this work has yet to be completed and for the purposes of this Case Study, we use the line segment and grid square datasets derived from the stick network obtained directly from LRC.

5 Exploratory Analyses

For later model validation we divide our dataset of hospital admissions into separate training and prediction samples. Approximately 10% of the admissions in each diagnostic group (B, N and U) were flagged using uniform random numbers and set aside for later use as a validation sample. All exploratory analyses and model estimation reported in this and subsequent sections were carried out on the remaining 90% subset.

For a continuous estimate of population density across the study region we used the count of births within each kilometer grid cell, scaled by a factor of 0.90 for use with the training sample and by 0.10 for use with the validation subset.

Fig. 4 presents maps showing the population density estimates and raw hospital admission rates per 100 population for the training sample on the kilometer grid.

Our data include admissions to the twenty-seven hospitals within the study region and to a small number outside the region: the latter were assigned a single hospital code (28) for the purposes of this analysis. Table 1 shows the number (and percentage) of admissions in each diagnostic group in our training database for each hospital. This indicates marked variations between hospitals, particularly in the proportion of unspecified ICD codes recorded by each hospital.

Table 2 shows the number of admissions stratified by method of inpatient admission: 'A&E' indicates referral by the hospital's Accident and Emergency department, 'GP' indicates referral by a general practioner, 'Consultant' indicates referral by another hospital consultant and 'Other' indicates other methods of admission (often a referral from the A&E department of *another* hospital). Hospital admission rates by each method seems consistent across diagnostic groups, although somewhat fewer GP-referred admissions end up with an unspecified diagnosis compared to admissions by the other three methods.

The remaining entries in Table 2 summarize the number of admissions in the training database by various covariate categories. For this purpose we have arbitrarily categorized the continuous covariates (deprivation, grouped by quintile, and access scores, group by tertile) and assigned to each area the ethnic group of the majority of children living there. This table highlights a number of differences between the covariate distribution for unspecified admissions compared to bronchiolitis or non-respiratory admissions: the former tend to live in more deprived areas with a high proportion of children from ethnic minorities and high access to hospital and GP services. Admissions for bronchiolitis and non-respiratory diagnoses show quite similar covariate profiles, although a slightly higher proportion of boys are admitted for bronchiolitis.

For the continuous models introduced in Section 7 the hospital identifier may be treated as an individual attribute for each subject, but for the discrete area-level analyses of Section 5.1–6 we must aggregate all covariates to a common level of aggregation, for which we chose the kilometer grid. To achieve this, we first smoothed the hospital frequencies by averaging over a coarser 3km$\times$3km, calculating for each such 9km^2 cell the proportion of all admissions with a postcode centroid in that cell who attended each of the 28 hospital groups. This 28-dimensional vector was then assigned to each of the nine

TABLE 1. Number of admissions (percentage of hospital total) for each diagnostic group in the training database, split by hospital.

Hospital	**Primary Diagnosis**						**Total**
	Bronchiolitis		**Non-Respiratory**		**Unspecified**		
1	165	(11.7%)	1132	(79.9%)	119	(8.4%)	1416
2	250	(17.0%)	1078	(73.4%)	141	(9.6%)	1469
3	212	(16.6%)	1049	(82.2%)	15	(1.2%)	1276
4	213	(17.5%)	916	(75.2%)	89	(7.3%)	1218
5	147	(15.1%)	792	(91.5%)	33	(3.4%)	972
6	226	(21.9%)	780	(75.6%)	26	(2.5%)	1032
7	205	(21.4%)	714	(74.5%)	40	(4.2%)	959
8	112	(19.0%)	468	(79.2%)	11	(1.9%)	591
9	91	(15.5%)	412	(70.3%)	83	(14.2%)	586
10	74	(9.9%)	289	(38.8%)	381	(51.2%)	744
11	60	(18.0%)	265	(79.3%)	9	(2.7%)	334
12	43	(15.1%)	186	(65.5%)	55	(19.4%)	284
13	17	(8.3%)	189	(91.7%)	0	(0.0%)	206
14	6	(4.9%)	115	(93.5%)	6	(4.9%)	123
15	1	(1.1%)	59	(65.6%)	30	(33.3%)	90
16	15	(22.7%)	48	(72.7%)	3	(4.5%)	66
17	6	(11.3%)	44	(83.0%)	3	(5.7%)	53
18	8	(15.4%)	44	(84.6%)	0	(0.0%)	52
19	4	(8.7%)	29	(63.0%)	13	(28.3%)	46
20	2	(8.0%)	21	(84.0%)	2	(8.0%)	25
21	0	(0.0%)	19	(86.4%)	3	(13.6%)	22
22	2	(12.5%)	14	(87.5%)	4	(25.0%)	16
23	4	(18.2%)	14	(63.6%)	3	(13.6%)	22
24	0	(0.0%)	15	(83.3%)	3	(16.7%)	18
25	7	(36.8%)	11	(57.9%)	1	(5.3%)	19
26	1	(11.1%)	5	(55.6%)	3	(33.3%)	9
27	2	(25.0%)	4	(50.0%)	2	(25.0%)	8
28	33	(12.3%)	214	(79.6%)	22	(8.2%)	269
Total	1902	(15.9%)	8926	(74.9%)	1097	(9.2%)	11925

TABLE 2. Number of admissions for bronchiolitis (B), non-respiratory illness (N) and unspecified diagnoses (U) in the training database by method of admission and other covariate categories. Numbers in parentheses refer to percentage of all admissions for a given diagnosis in each covariate category.

	Primary Diagnosis						**Total**	
	B		***N***		***U***			
Admission Method								
A&E	1162	(61.1%)	4868	(54.5%)	706	(64.4%)	6736	(56.5%)
GP	577	(30.3%)	2417	(27.1%)	155	(14.1%)	3149	(26.4%)
Consultant	30	(1.6%)	282	(3.2%)	39	(3.6%)	351	(2.9%)
Other	133	(7.0%)	1359	(15.2%)	197	(18.0%)	1689	(14.2%)
Hospital Access Score								
Low	566	(29.8%)	2605	(29.2%)	134	(12.2%)	3305	(27.7%)
Medium	889	(46.7%)	3928	(44.0%)	535	(48.8%)	5352	(44.9%)
High	447	(23.5%)	2393	(26.8%)	428	(39.0%)	3268	(27.4%)
GP Access Score								
Low	575	(30.2%)	2654	(29.7%)	158	(14.4%)	3387	(28.4%)
Medium	880	(46.3%)	3890	(43.6%)	493	(44.9%)	5263	(44.1%)
High	447	(23.5%)	2382	(26.7%)	446	(40.7%)	3275	(27.5%)
Gender								
Girls	722	(38.0%)	3715	(41.6%)	455	(40.5%)	4882	(41.0%)
Boys	1180	(62.0%)	5199	(58.3%)	648	(59.1%)	7027	(58.9%)
Missing	0	(0.0%)	12	(0.1%)	4	(0.4%)	16	(0.1%)
Deprivation								
Affluent	224	(11.8%)	1098	(12.3%)	87	(7.9%)	1409	(11.8%)
Quintile 2	294	(15.5%)	1309	(14.7%)	113	(10.3%)	1716	(14.4%)
Quintile 3	440	(23.1%)	1986	(22.2%)	183	(16.7%)	2609	(21.9%)
Quintile 4	516	(27.1%)	2495	(28.0%)	316	(28.8%)	3327	(27.9%)
Deprived	428	(22.5%)	2038	(22.8%)	298	(36.2%)	2864	(24.0%)
Ethnic group								
White	1628	(85.5%)	7602	(85.3%)	800	(67.3%)	10030	(84.1%)
Black	61	(3.4%)	332	(3.7%)	141	(11.9%)	534	(4.5%)
Indian	200	(10.4%)	914	(10.2%)	139	(11.7%)	1253	(10.5%)
Asian	6	(0.3%)	32	(0.3%)	5	(3.6%)	43	(0.4%)
Other	6	(0.3%)	45	(0.5%)	12	(5.3%)	63	(0.5%)
Missing	1	(0.05%)	1	(0.01%)	0	(0.0%)	2	(0.02%)

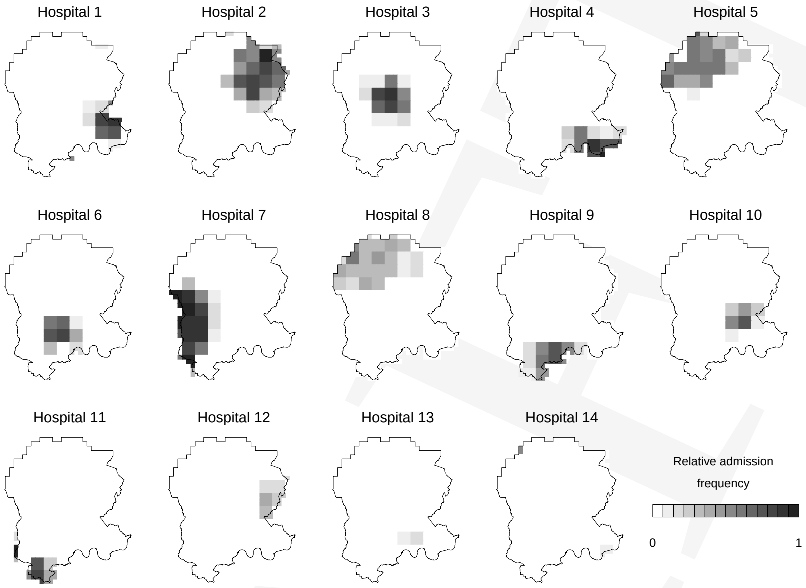

FIGURE 5. Relative admission frequency to each hospital according to place of residence.

one-kilometer grid cells comprising a particular 3km $\times$ 3km cell. The geographical distribution of these *relative admission frequencies* for the 14 hospitals receiving most admissions are displayed graphically in Fig. 5; not surprisingly, these maps reveal marked geographical variation.

A similar procedure was used to calculate area-level relative frequencies of hospital admission by each method ('A&E', 'GP' or 'Other'), after merging the rather small 'Consultant' and 'Other' groups. Maps showing the geographical distribution of these relative frequencies are presented in Fig. 6. Again, a very strong spatial pattern emerges.

Maps showing the geographical distribution of the other covariates listed in Table 2 are shown in Fig. 7; covariates originally recorded at either ED or ward level have been reapportioned to the kilometer grid using area-weighted averages.

Fig. 8 shows maps of the density of cars and light goods vehicles (CLGV), the density of medium and heavy goods vehicles (MHGV) and total vehicle emissions for various pollutants per km^2 across our

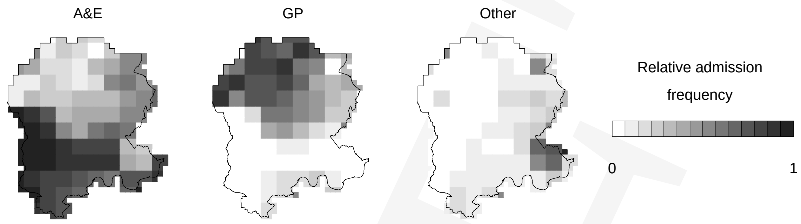

FIGURE 6. Relative frequency of hospital admission via referral from the A&E department, referral by GP or by other methods, according to place of residence.

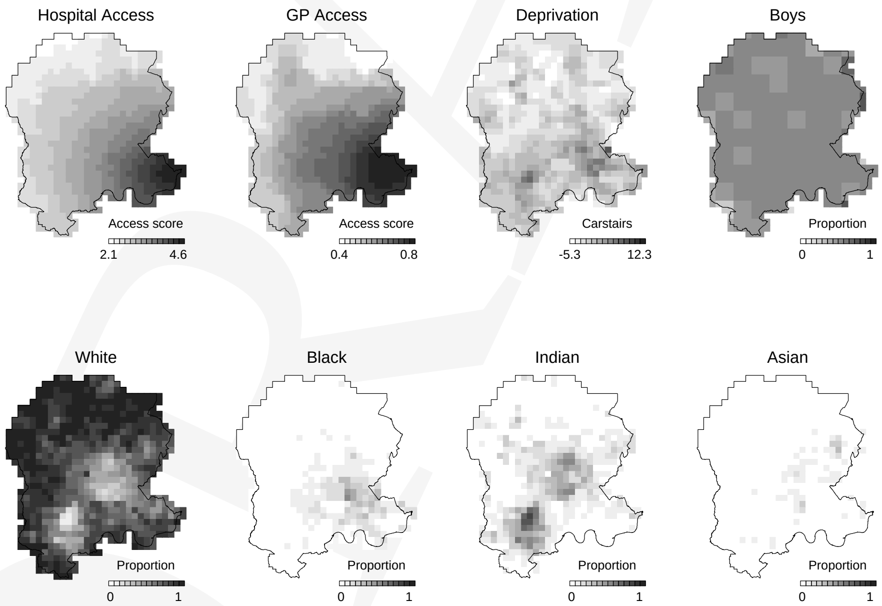

FIGURE 7. Geographical distribution of potential explanatory variables.

study region. We focus on particulates (TSP, black smoke and PM_{10}) and acid gases (CO, NO_x and SO_2) since these are the pollutants most commonly associated with adverse respiratory health effects (Committee of the Environmental and Occupational Health Assembly of the American Thoracic Society, 1996), and not on the organic compounds (NMVOCs, benzene, 1,3-butadine, and methane) more often associated with longer-term adverse health effects such as cancers.

It is clear from these maps that the traffic variables are highly correlated; NO_x, SO_2, TSP, PM_{10} and black smoke are nearly co-linear, with pairwise correlations ranging from 0.97 to 0.99. We therefore decided to select just one of these five variables for use in our subsequent analyses: we chose black smoke, since this had the smallest correlation with the three other traffic exposure variables still considered *i.e.*, CLGV (correlation = 0.6), MHGV (correlation = 0.74), and CO (correlation = 0.72). Note that correlations between the traffic variables and the other potential explanatory variables shown in Fig. 7 were much weaker (typically less than 0.4 in absolute value).

By assigning each child in our database the value of the traffic density and pollutant load for the kilometer grid cell containing their home postcode centroid, we then calculated the empirical distribution of each of the traffic exposure variables (CLGV, MHGV, CO and smoke) for the three diagnostic groups and for the population of 0–1 year olds as a whole. These are shown in Fig. 9. There is considerable variation in exposure within each group, with some suggestion that children admitted to hospital with bronchiolitis typically live in *less* polluted areas than children admitted with other diagnoses.

Finally, we plot raw admission rates for bronchiolitis against selected covariate values on the kilometer grid to explore possible associations between the health outcomes and putative risk factors (Fig. 10). To aid visual assessment of any trends in these relationships, we split each covariate into deciles and calculate the median and 5^{th} and 95^{th} percentiles of the empirical distribution of observed admissions rates for infants in each exposure decile; these are shown as the solid and lower and upper dashed lines on the plots respectively.

There is some suggestion that the admission rate for bronchiolitis in deprived areas (high Carstairs score) is higher than in affluent areas (low Carstairs score). Otherwise, these plots show little evidence

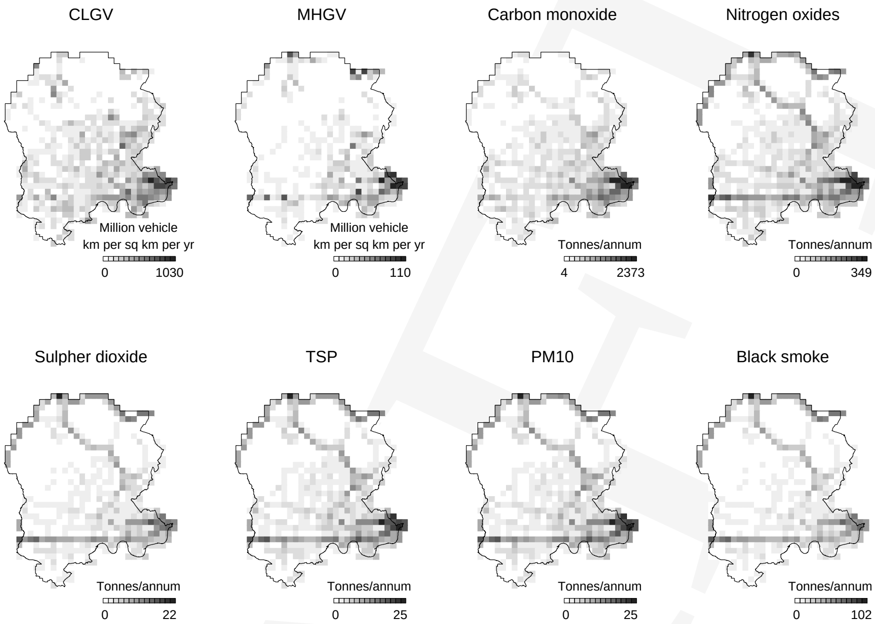

FIGURE 8. Traffic density and pollutant loads on the kilometer grid.

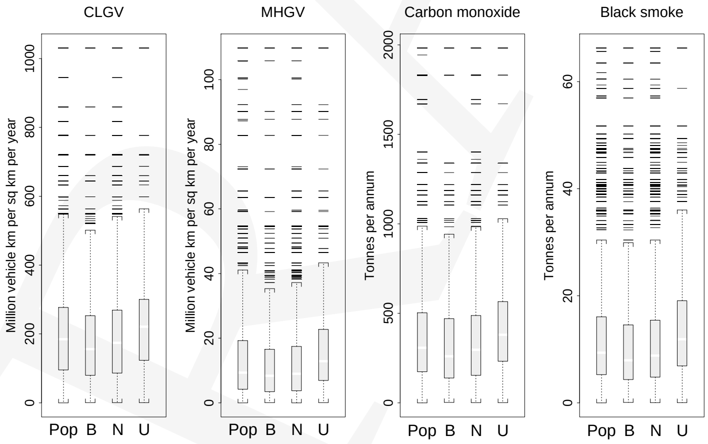

FIGURE 9. Empirical distribution of traffic density and pollutant loads for infants admitted to hospital with with bronchiolitis (B), non-respiratory illnesses (N) and unspecified diagnoses (U) and for the whole population of 0–1 year olds.

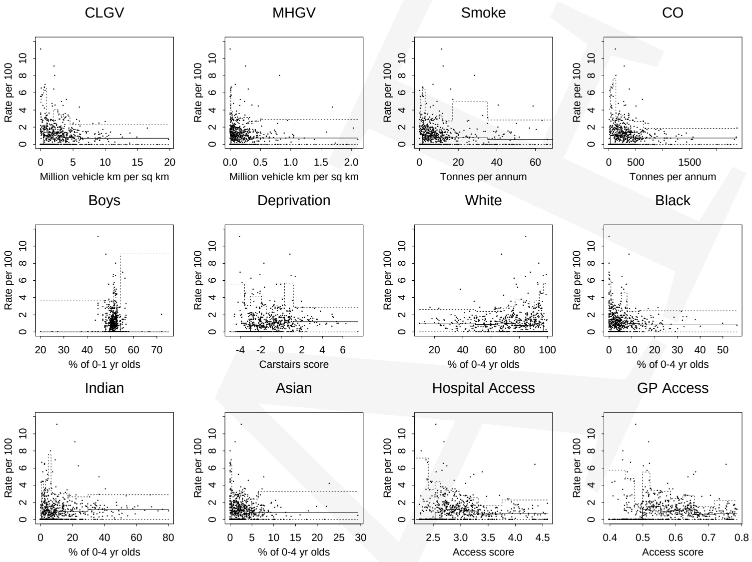

FIGURE 10. Raw admission rates for bronchiolitis versus covariates on the kilometer grid. Solid lines show the median admission rate, and dashed lines show the 5^{th} and 95^{th} percentiles, associated with each decile of exposure. For clarity, the vertical scale is truncated at a rate of 12 per 100; six areas had observed rates exceeding this threshold and are included in the calculations of the empirical rate percentiles shown by the dashed and solid lines.

of strong associations between bronchiolitis admissions and most of the covariates. Similar plots were produced for non-respiratory and unspecified diagnoses (not shown). These showed some evidence that areas with high proportions of Indian and black children have relatively high admission rates for unspecified diagnoses, but otherwise revealed no clear patterns. However, interpretation is difficult or misleading due to the relatively sparse data in many grid cells leading to spuriously high or low estimates of disease risk in these areas. Discussions of the limitations and problems associated with mapping and interpreting raw disease rates in small geographic areas can be found elsewhere (Clayton and Bernardinelli, 1992; Mollié, 1996), and it is well recognized that such data typically exhibit overdispersion with respect to the assumed binomial or Poisson sampling distribution. To overcome this, the usual approach is to adopt some form of Bayesian hierarchical model to allow 'borrowing of strength' between areas and so smooth the raw disease rates and improve interpretation. As mentioned in Section 2.2 this is frequently done using class of log-linear Markov random field models with CAR priors proposed by Besag et al. (1991). However, there are a number of potential limitations to this approach, some of which are fundamental to the model itself, and some of which arise within the context in which the models are usually applied. We expand on these in the next section in order to motivate the modeling approach we adopt for this Case Study.

5.1 Poisson regression analysis

To explore more closely the relationship of explanatory variables (gender, emissions data, area deprivation, *etc.*) to health effects, we use log-linear Poisson regression to model the hospital admissions numbers Y_i^B of bronchiolitis cases, Y_i^N of non-respiratory cases, and Y_i^U of unspecific diagnoses in the i^{th} kilometer grid cell in the form

$$Y_i^* \quad \sim \quad \mathsf{Po}\big(w_i \exp(\textstyle\sum X_{ij}\phi_j^*)\big) \qquad (1)$$

where $* = B, N$ or U, w_i is the estimated population of 0–1 year old children (in hundreds) and X_{ij} is the j^{th} covariate (proportion of male births, smoke emission, Carstairs deprivation score, *etc.*) in the i^{th} kilometer grid cell. (Note that $X_{i0} \equiv 1$, a vector of 1s, so that ϕ_0^* is an overall intercept.) We use a quasi likelihood approach (McCullagh and Nelder, 1989) to adjust for overdispersion and specify the variance function for the observed admission counts to be $\mathsf{V}(Y_i^*) = \kappa^* \mathsf{E}(Y_i^*)$. The unknown parameter $\kappa^* \geq 1$ measures the amount of overdispersion, and is equal to one in the conventional Poisson model. The effect on inference is simply to inflate the regression coefficient standard errors to account for the extra-Poisson variability; note that this approach assumes no parametric model for the form of the overdispersion.

We use the schematic representation of the hospital admission process in Fig. 1 as the basis for building our statistical model. Primary interest concerns how the exposures in stage 1 influence the outcome at stage 6. However there are a number of factors that may influence intermediate stages which are unrelated to the true health status of the infant. Our approach is to first model these 'nuisance' factors to at least partially adjust for the possible biases they may introduce, and to then examine how the explanatory variables of interest (gender, traffic data, area deprivation, *etc.*) influence the adjusted admission rates.

Nuisance factors

We model the complex mix of factors influencing stages 4, 5 and 6 in Fig. 1 by introducing separate 'hospital effects' for each of the hospital groups. These are defined by a (initially) 28-level factor X_{ij}, $j = 1, \ldots, 28$ taking values equal to the relative frequency of admission to hospital group j from grid cell i (see Fig. 5); since the model includes an overall intercept, we alias X_{i1} during model fitting to maintain parameter identifiability. The effect of method of referral on admission rates (stage 4 in Fig. 1) is modeled by a 3-level factor giving the relative frequency of referral from the hospital's A&E department, by a GP, or by other methods, for each grid cell; for this factor we alias A&E referral to ensure identifiability. The parents' decision to seek medical help and their choice of care (stages 2 and 3 in Fig. 1) will be partly influenced by the accessibility of GP and hospital services in their area. This is modeled by separately including either GP or hospital access scores (Carr-Hill et al., 1994) as covariates in the model.

Explanatory variables

We regard the remaining covariates, *i.e.*, gender (proportion of male births in each grid cell), ethnicity (census proportions of 0-4 year old children in the following ethnic groups: white, black, Indian, Asian, other), area deprivation (Carstairs index), traffic density (MHGV and CLGV) and emission loads (CO and black smoke), as potential influences at stage 1b of Fig. 1. These were entered simultaneously in the model and then the traffic density and pollutant emission variables were included one at a time (the latter were entered singly in the models to avoid problems with multicollinearity).

Results

The Poisson regression analysis described here may be seen as a preliminary screening analysis to eliminate any covariates which were clearly unimportant in explaining the risk of hospital admission for any of the outcomes (B, N and U). Our aim was to identify a reduced subset of potential nuisance factors and explanatory variables to explore in more detail using the Bayesian models presented in subsequent sections. We therefore chose to retain the same set of nuisance factors and explanatory variables in the model for each outcome (B, N and U). These were chosen to be any covariate whose 90% confidence interval for the relative risk of admission excluded one for at least one of the three outcomes (see Table 3). Note that, having identified the subset of variables presented in this table, we then checked whether any of the remaining covariates met the 90% confidence interval criteria when added back into this model separately: none did. We then investigated interactions between the covariates, and threshold models for the traffic and emissions variables, but these did not significantly improve the model. One noteworthy finding is the significant but *negative* association between each of the traffic/emissions variables and risk of bronchiolitis. This is consistent with the exploratory plots in Fig. 9; however it is unclear whether this is a genuine association, an artifact of data aggregation (ecological bias) or simply a result of residual confounding due to factors such as socioeconomic deprivation, or of systematic differences in the diagnostic coding policies of hospitals.

The results for κ shown at the bottom of Table 3 suggest that there is overdispersion in these data for all 3 outcomes. The residual deviance for each model is also larger than the degrees of freedom,

TABLE 3. Relative risk estimates $e^{\phi_j^*}$ (90% confidence intervals) from the exploratory Poisson regression models.

Covariate	Bronchiolitis		Non-Respiratory		Unspecified	
Hospital effects (relative to remaining 18 hospitals pooled together)						
H_3	1.01	(0.79,1.29)	0.93	(0.83,1.04)	0.22	(0.15,0.32)
H_6	1.04	(0.78,1.39)	1.25	(1.09,1.44)	0.38	(0.24,0.60)
H_7	0.86	(0.65,1.13)	1.22	(1.06,1.40)	0.46	(0.30,0.71)
H_8	1.75	(0.81,3.80)	1.30	(0.91,1.84)	0.03	(0.01,0.09)
H_{10}	0.61	(0.40,0.95)	0.71	(0.58,0.87)	6.05	(3.94,9.28)
H_{11}	0.94	(0.62,1.43)	2.10	(1.73,2.54)	0.95	(0.51,1.77)
H_{12}	0.50	(0.25,0.98)	0.83	(0.62,1.12)	2.72	(1.41,5.22)
H_{13}	0.04	(0.01,0.13)	0.56	(0.31,1.00)	0.00	(0.00,0.04)
H_{14}	0.01	(0.00,0.20)	0.19	(0.04,0.80)	0.07	(0.00,4.07)
H_{16}	8.91	(2.93,27.12)	0.82	(0.42,1.60)	0.12	(0.01,1.08)
Admission method (relative to referral from A&E)						
GP referral	1.18	(0.71,1.96)	2.75	(2.17,3.49)	2.65	(1.35,5.22)
Other	0.49	(0.34,0.72)	1.36	(1.15,1.62)	0.94	(0.59,1.49)
Ethnicity (relative to white/Asian/other)						
Black	0.46	(0.17,1.20)	1.41	(0.91,2.18)	4.15	(1.62,10.61)
Indian	0.38	(0.27,0.54)	0.63	(0.53,0.74)	1.43	(0.88,2.34)
Deprivation (most deprived vs least deprived)						
Carstairs	5.62	(3.78,8.35)	2.64	(2.20,3.17)	2.50	(1.50,4.16)
Gender						
Boys *vs* girls	0.93	(0.08,10.92)	5.23	(1.59,17.19)	2.42	(0.03,186.30)
Traffic pollution† *(most polluted vs least polluted)*						
MHGV	0.50	(0.35,0.71)	0.85	(0.73,0.98)	0.95	(0.64,1.41)
CLGV	0.54	(0.40,0.75)	0.96	(0.83,1.10)	0.95	(0.65,1.40)
CO	0.49	(0.33,0.71)	0.90	(0.77,1.06)	1.07	(0.69,1.67)
Smoke	0.58	(0.39,0.84)	1.00	(0.85,1.19)	1.05	(0.64,1.71)
Overdispersion and Residual Deviance						
κ	1.38		1.79		1.43	
Deviance	744.43		1045.12		661.28	
df	623		623		623	

† Pollution variables were included one at a time in the multiple regression model. Results for the non-pollution variables are reported for the model including MHGV; differences are negligible when MHGV is replaced by each of CLGV, CO or smoke.

suggesting that the Poisson regression models do not fit particularly well. However, the sparseness of our data in each grid cell casts doubt on the asymptotic chi-squared distribution of the deviance (Bithell et al., 1995). One reason for the apparent poor fit is that our problem is *hierarchical*, with individual, hospital, area and population levels. This structure should be acknowledged by explicit modeling of dependencies (including spatial dependence induced by geographic proximity) between variables at each level. In Section 6–7 we discuss Bayesian hierarchical-spatial modeling approaches to address this issue.

6 Area-Level Hierarchical Models

The Poisson log-linear models described in Section 5.1 do not reflect the hierarchical structure or spatial dependence in the health or the traffic exposure data. The most common Bayesian hierarchical approach to ecological regression offering the opportunity to exploit spatial correlation in the data is a Markov Random Field (MRF) approach.

6.1 Markov Random Field models

Following Besag et al. (1991), we consider the hierarchical model given below for counts of hospital admissions Y_i^* in a set of geographical areas $R_i, i \in I$,

$$\begin{aligned} Y_i^* &\sim \mathsf{Po}\big(w_i \exp(\sum_j X_{ij}\phi_j^* + V_i^* + U_i^*)\big) && (2)\\ V_i^* &\sim \mathsf{No}\big(0, \sigma_v^*\big) \\ U_i^* | U_{k,k\neq i}^* &\sim \mathsf{No}\left(\frac{\sum_{k\in I} q_{ik}u_k}{\sum_{k\in I} q_{ik}}, \frac{\sigma_u^*}{\sum_{k\in I} q_{ik}}\right). && (3) \end{aligned}$$

As before, $* = B, N$ or U, and w_i and X_{ij} are the estimated population (in hundreds of 0–1 year old children) and covariate values in the i^{th} area respectively. The terms V_i^* and U_i^* are area-specific random effects representing unobserved or unknown risk factors which display random and spatially correlated variation respectively. Information about the structure of the spatial variation of the U_i is expressed through the 'weights' q_{ik} in the MRF conditional autoregressive model (3): a common choice is $q_{ik} = 1$ if areas R_i and R_k share a common boundary, and $q_{ik} = 0$ otherwise. Within a Bayesian paradigm, prior distributions on the regression coefficients ϕ_j^* and the variance parameters σ_v^* and σ_u^* are also specified.

The usual choice of areas R_i for use with the above model are political or administrative boundaries such as EDs or wards in the UK, on which most official data are available. A preliminary analysis of our data at the ward level was described in Section 2.1. However, these areas tend to be highly irregular in size and shape and there is some concern over the appropriateness of using adjacency-based MRF conditional autoregressive models, which were developed originally for regular lattices in image analysis (Besag, 1974). In our present study the spatial resolution of the data permit us to overcome this problem by reapportioning our data onto a kilometer grid based on the National Grid. This grid, with 710 cells, is finer and far more regular than the partition into 269 wards, and matches the partition on which the minor roads traffic data are given.

We first fit model (2) with the intercept ϕ_0^* and random effects terms V_i^* and U_i^* only, and then repeat the analysis including the subset of potential explanatory covariates identified from the exploratory Poisson regression analysis (Table 3). We assume gamma prior distributions with shape parameters 0.5 and precisions 0.0005 for the inverses of the random effects variances $1/\sigma_v^*$ and $1/\sigma_u^*$; this choice represents a prior belief that the random effects variance is centered around 0.002, with about 1% chance each that it is less than 0.0002 or greater than 6.4 (Kelsall and Wakefield, 1999). For the vectors of coefficients ϕ^* we choose independent normal prior distributions with means equal to zero to reflect prior indifference to whether the relative risk factor X_{ij} increases or decreases the log-risk of hospital admission for bronchiolitis (or non-respiratory or unspecified conditions); prior variances are chosen such that the ratio between the 90^{th} and the 10^{th} percentile of the prior distribution for the relative risk associated with the maximum versus minimum exposure to X_{ij}, *i.e.*, $\exp\left(\phi_j^*(\max_i X_{ij} - \min_i X_{ij})\right)$, is 100. (We use the same prior distributions for the parameters ϕ^* described in the next section — see Table 6 for further details.) We model the vector of hospital effects as random rather than fixed as in the Poisson regression models of Section 5.1. This reflects the belief that the influences captured by the hospital term (*e.g.*, availability of beds, clincial staff, diagnostic coding) are likely to be similar across hospitals. We assume prior exchangeability, but allow for possibly outlying hospitals by choosing Student-t distributions with zero means, unknown scale parameters σ_h^* and 4 degrees of freedom for the population distributions of the hospital effects. Gamma prior distributions with shape and inverse scale parameters equal to 0.01 are chosen for the inverse scale parameters $1/\sigma_h^*$, reflecting the prior belief that between-hospital variation may be quite large.

Results

All results reported here and in subsequent sections are based on a single long run of the Markov Chain Monte Carlo (MCMC) sampler for each model. The models reported in Section 6 were estimated using the WinBUGS software (Spiegelhalter et al., 1999); programs to implement the models in Section 7 were written in ISO C. Convergence was assessed by visual inspection of the trace and autocorrelation plots: a burn-in of 1000–10000 samples was typically required. Posterior sample sizes were determined by running sufficient iterations to ensure that the Monte Carlo standard error of the mean (Roberts, 1996) was at least one order of magnitude smaller than the posterior standard deviation for each parameter of interest: 5000–20000 iterations were typically required to achieve this level of precision.

The first column of Fig. 11 shows maps of the posterior mean admission rates $\exp(\phi_0^* + V_i^* + U_i^*)$ for each of the three outcomes $* = B, N$ or U before adjustment for observed covariates; the second column shows the same quantity (which we now term the residual or adjusted admission rate) after accounting for nuisance factors and potential explanatory variables. The strong spatial patterns of admission rates for each outcome are largely removed following covariate adjustment, although an area of high residual admission rates for unspecified diagnoses in the east of the study region remains unexplained.

We consider proportional admission rates for bronchiolitis relative to total admissions for bronchiolitis, non-respiratory and unspecified diagnoses by mapping the posterior mean of the ratio

$$\exp(\phi_0^B + V_i^B + U_i^B) / \sum_{*\in(B,N,U)} \exp(\phi_0^* + V_i^* + U_i^*)\,,$$

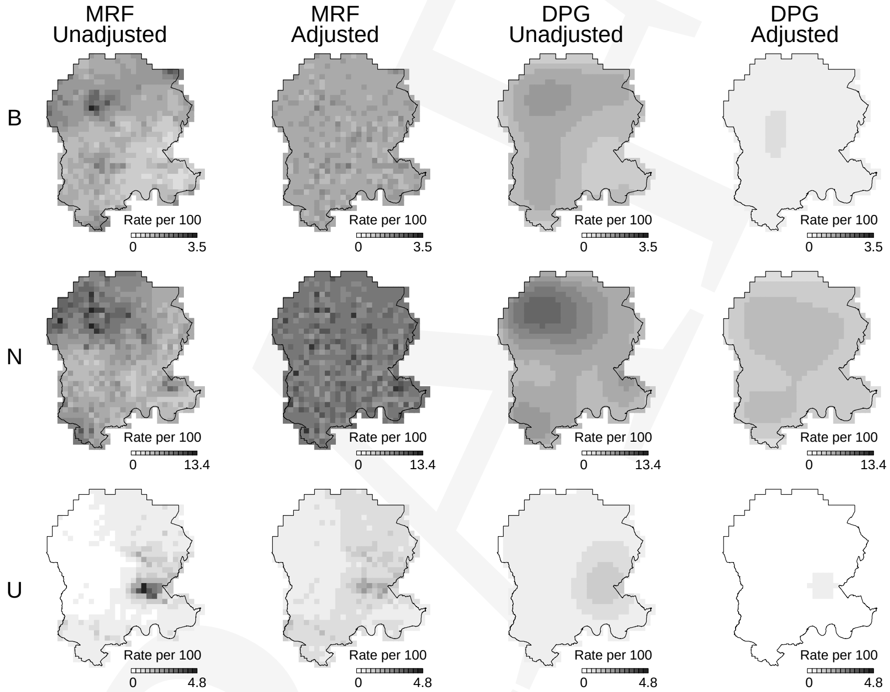

FIGURE 11. Posterior mean admission rates for bronchiolitis (B), non-respiratory illness (N) and unspecified diagnoses (U), before and after adjustment for covariates. First two columns relate to Markov Random Field (MRF) models of Section 6.1; last two columns relate to Discrete Poisson/gamma (DPG) models of Section 6.2.

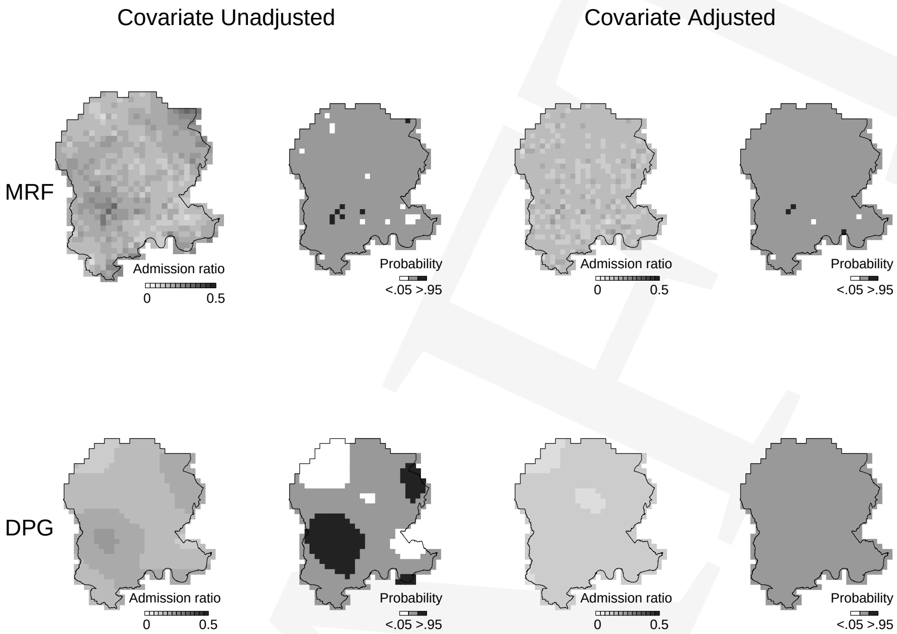

FIGURE 12. Posterior mean proportional admission rates for bronchiolitis before and after adjustment for covariates, and posterior probabilities of exceeding the corresponding overall proportional admission rate in the study region. Top row relates to Markov Random Field (MRF) models of Section 6.1; bottom row relates to Discrete Poisson/gamma (DPG) models of Section 6.2.

both before and after adjustment for covariates (see Fig. 12, top row). Also shown are corresponding maps of the posterior probabilities that the proportional admission rate in each grid cell exceeds the average proportional admission rate across the study region. These maps suggest that the high admission rates for bronchiolitis seen in the northwest of the study region in Fig. 11 are consistent with higher overall admission rates in this area; there is a small cluster of cells with significantly elevated admission ratios for bronchiolitis in the southwest, some of which may be explained by the available covariates.

Table 4 shows the posterior variances of the random effects from the covariate-unadjusted and adjusted MRF models. We note that

TABLE 4. Posterior random effects variances (95% credible intervals) from MRF models before and after adjustment for covariates.

Variance	Bronchiolitis	Non-Respiratory	Unspecified
Model without covariate adjustment			
σ_v^*	0.004	0.011	0.002
	(0.0002,0.023)	(0.002,0.032)	(0.0001,0.008)
σ_u^*	0.541	0.293	1.673
	(0.314,0.796)	(0.170,0.419)	(1.210,2.278)
$s_u^{*\dagger}$	0.162	0.105	0.630
	(0.109,0.221)	(0.075,0.133)	(0.480,0.815)
Model after covariate adjustment			
σ_v^*	0.074	0.046	0.009
	(0.034,0.122)	(0.11,0.46)	(0.0003,0.063)
σ_u^*	0.001	0.002	0.867
	(0.0001,0.004)	(0.0002,0.006)	(0.024,1.384)
$s_u^{*\dagger}$	0.0004	0.0006	0.285
	(0.00004,0.0013)	(0.00007,0.0023)	(0.106,0.478)
σ_h^*	0.078	0.055	1.082
	(0.005,0.613)	(0.012,0.208)	(0.142,4.635)

† s_u = empirical variance of the spatial random effects, $\sum_i U_i^2/(I-1)$; this may be interpreted as a *marginal* variance (whereas σ_u is a conditional variance) and so is directly comparable with the variance of the unstructured random effects, σ_v^*.

inclusion of covariates leads to a large reduction in total between-area variation ($\sigma_v^* + \sigma_u^*$): prior to covariate adjustment, the spatially-structured component of variation (σ_u^*) dominates; this is largely removed by inclusion of nuisance factors and explanatory covariates, although there continues to be unstructured heterogeneity. There is also considerably more spatially structured between-cell variation in admissions for unspecified diagnoses compared to bronchiolitis or non-respiratory illnesses.

Posterior means and 95% credible intervals (CI) for the relative risks (RR) associated with each explanatory covariate are shown in the top half of Table 5. The final two columns of this table give the ratio of relative risks of admission for bronchiolitis versus non-respiratory and unspecified diagnoses respectively associated with exposure to each covariate, *i.e.*, $\exp(\phi_j^B - \phi_j^N)$ and $\exp(\phi_j^B - \phi_j^U)$.

There was no evidence of significantly (*i.e.*, the 95% credible inter-

TABLE 5. Parameter estimates (95% credible intervals) from the MRF (top half) and $\mathrm{DPG}_{8\times 9;3}$ (bottom half) regression models.

Covariate	***B***	***N***	***U***	***B vs N***	***B vs U***
Relative risk estimates from MRF model, and odds ratios for B versus N or U					
Ethnicity (relative to white/asian/other)					
Black	0.51 (0.25,0.99)	1.08 (0.72,1.58)	3.98 (1.19,8.94)	0.49 (0.21,1.03)	0.17 (0.04,0.49)
Indian	0.57 (0.40,0.81)	0.72 (0.57,0.89)	1.50 (0.67,2.67)	0.81 (0.52,1.21)	0.43 (0.19,0.90)
Gender					
Boys *vs* girls	2.70 (0.66,7.44)	1.23 (0.60,2.23)	0.71 (0.08,2.84)	2.46 (0.49,7.65)	8.81 (0.55,41.81)
Deprivation (most deprived vs least deprived)					
Carstairs	5.81 (3.32,9.55)	2.82 (2.05,3.79)	2.16 (0.89,4.66)	2.12 (1.09,3.68)	3.20 (1.06,7.44)
Traffic pollution (most polluted vs least polluted)					
MHGV	0.16 (0.01,0.82)	0.32 (0.06,0.97)	1.22 (0.04,7.22)	0.87 (0.02,5.44)	0.72 (0.01,4.90)
Relative risk estimates from DPG model					
Ethnicity (relative to white/asian/other)					
Black	0.62 (0.35,1.13)	1.06 (0.79,1.38)	3.18 (1.75,5.75)	0.59 (0.31,1.14)	0.20 (0.08,0.47)
Indian	0.56 (0.41,0.76)	0.66 (0.56,0.77)	1.61 (0.98,2.72)	0.86 (0.60,1.22)	0.35 (0.19,0.63)
Gender					
Boys *vs* girls	2.50 (0.42,13.88)	1.73 (0.77,4.13)	0.89 (0.12,8.80)	1.42 (0.18,10.10)	2.73 (0.16,43.05)
% Admissions attributed to each excess risk factor from DPG model, and % excess admissions for B vs N or U					
Baseline	12.8% (0.7%,19.8%)	23.4% (10.0%,31.5%)	31.9% (7.7%,50.4%)	−14.5% (−47.6%, 8.4%)	−22.6% (−43.1%, −8.9%)
Carstairs	53.8% (42.3%,69.0%)	33.1% (28.9%,39.0%)	28.1% (11.8%,57.1%)	11.9% (5.6%,17.1%)	15.6% (−2.3%, 31.6%)
MHGV	0.1% (0.0%,1.5%)	0.1% (0.0%,0.5%)	0.7% (0.0%,5.9%)	13.7% (−49.0%, 49.9%)	−31.4% (−50.0%, 48.8%)
Latent Spatial Risk Factors	34.0% (16.3%,49.3%)	44.9% (30.8%,53.1%)	37.8% (6.4%,74.8%)	−5.9% (−25.4%, 6.9%)	−4.6% (−20.1%, 33.4%)

B = bronchiolitis, N = non-respiratory illness and U = unspecified diagnoses val excludes one) elevated or reduced risk of bronchiolitis admission to any of the 11 hospital groups relative to the overall bronchiolitis admission rate. Non-respiratory admission rates to hospital 11 were significantly higher than average (RR = 1.69, 95% CI = 1.15–2.35), while admissions with unspecified diagnoses were higher than average to hospital 10 (RR = 6.28, 95% CI = 1.67–15.05) and lower than average to hospitals 3 (RR = 0.29, 95% CI = 0.10–0.65) and 13 (RR = 0.09, 95% CI = 0.00–0.14).

GP-initiated hospital referrals were associated with a significantly higher risk of admission for bronchiolitis (RR = 1.91, 95% CI = 1.22–2.85) and non-respiratory diagnoses (RR = 2.86, 95% CI = 2.04–3.72), but not unspecified diagnoses, relative to referrals from the hospital's A&E department. Referrals by other methods were associated with a significantly lower risk of admission for bronchiolitis compared with A&E referrals (RR = 0.59, 95% CI = 0.33–0.96).

The MRF models have a number of advantages over the Poisson regression models reported in Section 5.1. In particular, they allow for 'borrowing of strength' across the units of analysis (1km^{2}cells) and explicit modeling of spatial dependencies in the data. This leads to more reliable estimates of disease rates and of the magnitude of the risk associated with exposure to relevant covariates. However, as seen above, the MRF model implicitly requires that all data be aggregated to a common spatial scale for analysis; this leads to a number of problems. First, the analysis will depend on the area units chosen. Second, the process of aggregation introduces approximation errors. Third, and most serious, the relationship between disease prevalence and average exposure at the area level may be quite different from the disease-exposure relationship for individuals. Different aspects of this problem have been described by a number of authors, under a number of different names: the terms *ecological bias*, *ecological fallacy*, *aggregation bias*, and the *modifiable areal unit problem* (MAUP) have all been used (Robinson, 1950; Greenland, 1992; Richardson, 1992; King, 1997). Carrying out a MRF analysis on kilometer grid cells rather than wards can only offer limited help in reducing the effects of the ecological fallacy.

6.2 Discrete Poisson/Gamma models

A principal reason why the MRF models may lead to ecological bias lies in the log-linear exposure/response relationship. The logarithmic link function leads to *products* rather than *sums* for exposure variables under aggregation of areal units, and thus leads to statistical models which are inconsistent under different levels of aggregation. In order to minimise the deleterious effects of ecological bias we now introduce a hierarchical Bayesian Poisson regression model with an *identity* link function leading to a linear exposure/response relationship (King, 1997, pp 249–255).

In this section we concentrate on a discrete version of the models employing our data at the 1km^2 grid cell level. As in the MRF approach we allow for modeling unmeasured confounders through a spatially varying random effect (latent) term. However, due to the identity link function we are not restricted to modeling these unmeasured or unobserved covariates on the same spatial scale as our data, but can choose an appropriate partition of the study region to reflect their spatial structure (although we still need to specify this partition *a priori*). In Section 7 we will then extend this model class to a continuous random field model to allow for the finest level of aggregation for all parts of our data, including individual attributes, and eliminate the need to specify the structure of the latent spatial term in advance. These models should then further reduce the problem of the ecological fallacy.

Let $\mathcal{Y}$ denote our study region in the northwestern part of London, and, as before, let $\{y_i^B\}_{i\in I}$, $\{y_i^N\}_{i\in I}$, and $\{y_i^U\}_{i\in I}$ be the numbers of hospital admissions with a diagnosis of bronchiolitis, non-respiratory illness, and unspecified, respectively, in a region $R_i \subset \mathcal{Y}$, $i \in I$. The regions R_i could be EDs, wards, grid cells, or any other arbitrary partition of the study region. As for the MRF analysis, we take $\{R_i\}_{i\in I}$ to be part of the National Grid, *i.e.*, a collection of 710 grid cells of size 1km^2.

Again, we model the numbers of admissions in area R_i as independent random variables with Poisson distributions $\{Y_i^*\} \sim \mathsf{Po}(\Lambda_i^* \times w_i)$, $* = B, N$ or U, where, as before, w_i is an estimate for the population of 0–1 year old children (in hundreds) in region R_i.

On the next stage of the hierarchy we model the rates $\Lambda_i^*, * = B, N$ or U, in area R_i (in counts of admissions per 100 births) as a combination of additive (excess) and multiplicative (relative) risk factors. We discuss the rationale and interpretation of this model in Section 6.2 in the context of the specific risk factors considered for this Case Study; see also Breslow and Day (1980, pp 55–7). In this approach, we represent the effects of excess risk factors by the term $\sum_{j\in J_E} X_{ij}\beta_j^*$ of possibly spatially varying covariates X_{ij} in region R_i and regression coefficients β_j^* (with X_{i0} equal to a vector of 1s and β_0^* corresponding to the intercept term). The relative risk factors enter the regression model through the term $\exp(\sum_{j\in J_R} X_{ij}\phi_j^*)$ of possibly spatially varying covariates X_{ij} in region R_i and regression coefficients ϕ_j^*.

Thus, the intensity (w.r.t. to the population weights w_i) of our discrete Poisson regression models with identity link can generally be presented in the following form:

$$\Lambda_i^* = c(\phi^*) \times e^{\sum_{j\in J_R} X_{ij}\phi_j^*} \times \sum_{j\in J_E} X_{ij}\beta_j^*\,, \quad i \in I\,.$$

Here ϕ^* is the vector of regression coefficients ϕ_j^* and $c(\phi^*)$ is a normalising factor introduced both to maintain a mean relative risk factor term $c(\phi^*)\times\exp(\sum_{j\in J_R} X_{ij}\phi_j^*)$ of unity and to reduce correlations between the ϕ_j^* and β_j^* and thus to speed up convergence in our underlying MCMC scheme. We choose the inverse of this normalising factor to be

$$c(\phi^*)^{-1} = \exp(\sum_{j\in J_R} \bar{X}_j\phi_j^*) \quad \text{with} \quad \bar{X}_j = \sum_{i\in I} X_{ij}w_i / \sum_{i\in I} w_i\,,$$

or equivalently use centered relative risk covariates.

This class of Poisson regression models with identity link is consistent under different levels of aggregation if the models include only excess (*i.e.*, additive) risk factors; for an application see Best et al. (1998b). Allowing for both excess and relative risk factors within the Poisson regression with identity link makes this class of models more flexible than Poisson models with log-linear link. Choosing which risk factors enter additively and which multiplicatively is then an additional step in our modeling task, and is usually based on *a priori* considerations (Breslow and Day, 1980, pp 122–4).

Our regression models (like the MRF models described earlier) include a spatially varying latent term which we model as additive (rather than multiplicative as in the MRF models). We represent this term by

$$X_{il}\beta_l^* = \big(\sum_{m\in M} k_{im}^*\gamma_m^*\big) \times \beta_l^*$$

where l is the last index of J_E. Each of the quantities $k_{im}^*\gamma_m^* \times \beta_l^*$, $i \in I$, $m \in M$, can be viewed as the influence that unmeasured or unobserved risk factors in area S_m, or at a fixed location s_m in S_m, exert in area R_i. Here the regions S_m are in northwestern London or nearby; they might again be EDs, wards, 1km^{2}grid cells or an arbitrary partition of Northwest London. In the following we take them to be small rectangles covering all of northwestern London plus a 4km surrounding buffer zone. The kernel matrix elements k^*_{im} are chosen to be Gaussian, *i.e.*,

$$k^*_{im} = \frac{1}{2\pi\rho^{*2}} e^{-|x_i - s_m|^2/2\rho^{*2}} ,$$

with distance scale $\rho^* > 0$ which determines how rapidly the influence k^*_{im} declines with increasing distance $|x_i - s_m|$ from the center x_i of R_i to the center s_m of S_m. The latent risk factors γ^*_m are also taken to be uncertain.

Prior distributions

Inference about the parameters $\phi^* = \{\phi^*_j\}_{j \in J_R}$, $\beta^* = \{\beta^*_j\}_{j \in J_E}$, $\gamma^* = \{\gamma^*_m\}_{m \in M}$ and possibly the distance scale ρ^* of the Gaussian kernel and derived quantities such as the intensities Λ^*_i, is based on a Bayesian analysis using independent prior distributions obtained as follows.

Adopting the same approach as for the MRF models in Section 6.1, we choose independent normal prior distributions $\mathsf{No}(\mu^*_j, \sigma^*_j)$ for the vectors of (log) relative risk regression coefficients ϕ^*, where $\mu^*_j = 0$ and σ^*_j is chosen such that the ratio between the 90^{th} and the 10^{th} percentile of the prior distribution for the relative risk associated with the maximum versus minimum exposure to X_{ij} is 100. Hospital effects are again treated as random and are assumed to follow Student-t distributions with zero means, unknown scale parameters σ^*_h and 4 degrees of freedom. For the vectors of coefficients β^* of the excess risk factors we choose independent gamma distibutions $\mathsf{Ga}(\alpha^*_j, \tau^*_j)$ with means α^*_j/τ^*_j and variances $\alpha^*_j/(\tau^*_j)^2$ in order to maintain non-negativity. The parameters of these gamma distributions are taken such that with 80% prior probability the number of bronchiolitis admissions (or non-respiratory admissions or admissions with unspecified diagnoses) associated with each excess risk factor lies between one-tenth and ten times a nominal equal distribution. This is achieved by setting $\alpha^*_j = 0.575$, $j \in J_E$, $* \in \{B, N, U\}$, giving a factor of 100 for the ratio of the 90^{th} to the 10^{th} percentile of a gamma distribution, and by setting the prior means α_j^*/τ_j^* for each coefficient β_j^* such that *a priori* all excess additive risk factors contribute equally.

The prior distributions for the latent risk factors γ_m^*, $m \in M$, are also taken from the gamma family with shape parameters $\alpha_{\gamma_m}^*$ and precisions τ_γ^*. Their prior means are set to be $\alpha_{\gamma_m}^*/\tau_\gamma^* = |S_m|$, the area of region S_m in km^2, which leads to a spatially extensible model for the sum $\sum_{m\in M} k_{im}^*\gamma_m^*$ over any partition $\{S_m\}$. The prior uncertainty about the spatial distribution of the latent effects is expressed through τ_γ^*: our uncertainty is about the same as if we had to infer *a priori* the value of the latent spatially-varying function by observing locations of points randomly scattered with the same distribution, at a density of about τ_γ^* points per km^2. Thus, the smaller the values for τ_γ^*, the higher the spatial variation of the latent risk factor under our prior.

The distance scale ρ^* for the kernel we treat as certain but run our models for several fixed values (see below). Putting a discrete prior concentrated on a few values for ρ^* and calculating a posterior distribution would be more elegant but is computationally feasible only for a very small number M of rectangles in the latent partition $\{S_m\}$.

Specific discrete Poisson/gamma models

As mentioned above, the greater flexibility of the Poisson/gamma models allows us to choose whether covariate terms should enter multiplicatively or additively. However, there is often little scientific evidence to guide this choice; in the present application, our decision is based on the assumption that the following model represents a plausible mechanism by which the various risk factors for bronchiolitis may operate.

Bronchiolitis is a viral illness where the source of infection is often a family member with an apparently minor respiratory illness. Our model therefore assumes a baseline rate of bronchiolitis admissions which may be attributed to such infections. One possible route by which chronic exposure to high levels of traffic pollution may affect risk of bronchiolitis is by lowering an infant's 'tolerance threshold' to the extent that they develop clinical symptoms of bronchiolitis if exposed to an infection which would otherwise have led to only mild respiratory illness. Such a mechanism would result in a number of extra hospital admissions, attributable to chronic exposure to ambient air pollution, in addition to those 'baseline' cases of bronchiolitis which would have occurred anyway. We may envisage a similar mechanism for the effects of socioeconomic deprivation on bronchiolitis risk, and there may be other unknown or unmeasured risk factors which also operate by this route. We therefore treat the intercept (baseline), traffic related pollution variables, Carstairs deprivation index and the spatially varying random effects term representing latent unobserved covariates as additive excess risk factors in our model. A different mechanism is envisaged for risk factors such as gender and ethnicity, which we assume affect an infant's overall susceptibility to bronchiolitis. For example, exposure to the same underlying 'attributable' risk factor (baseline infection, traffic pollution *etc*) may be more likely to lead to clinical symptoms of bronchiolitis in boys compared to girls. We therefore regard gender and ethnicity as relative risk factors which enter our model multiplicatively. We also model the hospital effects and method of admission as multiplicative relative risk factors, since, although these do not directly affect the risk of an infant developing bronchiolitis, they are likely modify the overall rate of admission to hospital.

Fig. 13 shows a schematic representation of the various mechanisms described above, and can be thought of as an expanded version of stages 1a and 1b shown in Fig. 1. For comparability, we assume the same model for non-respiratory and unspecified admissions.

Before fitting the full model with additive and multiplicative risk factors as described above, we first consider a set of models that include only the intercept (representing the baseline admission rate) and spatially varying latent risk factors. As already noted, the latent term is expressed as a mixture of gamma-distributed random variables representing the influence of unmeasured risk factors in areas S_m. These areas are arbitrary, but must be pre-specified for these discrete Poisson/gamma models. Our choice is governed by a trade-off between the resolution required to detect local variations in disease risk and the computation time involved. We consider two alternatives: a partition of northwestern London plus a 4km surrounding buffer zone into (i) an 11×13 grid of approximately 11km^2 cells, and (ii) an 8×9 grid of approximately 22km^2 cells. For each latent partition, we consider three distance scales $\rho^* = 1$km $\rho^* = 3$km and $\rho^* = 5$km for $* = B, N$ or U. In what follows, we will use the no-

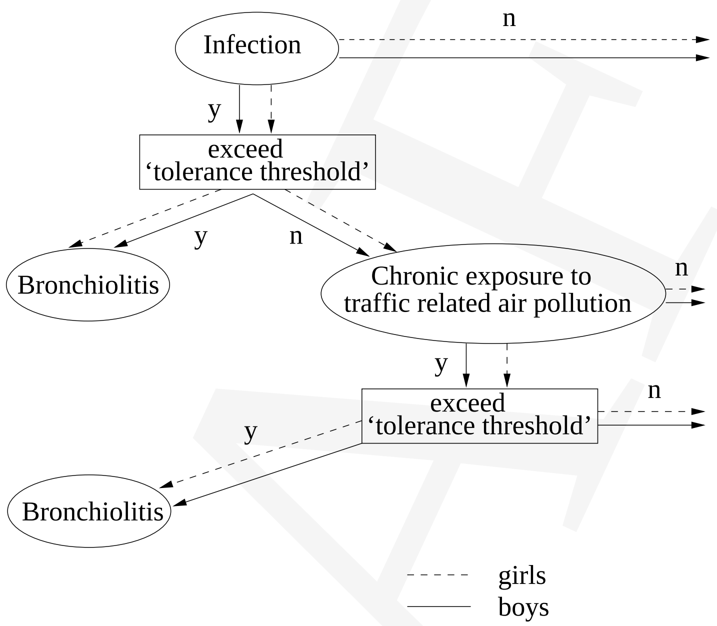

FIGURE 13. Schematic representation of the various mechanisms by which an infant may develop bronchiolitis. An underlying infectious ætiology is assumed for all cases. Excess cases may be generated if an infected infant has also been exposed to some other relevant risk factor, *e.g.*, ambient air pollution (other such exposures not shown include living in a deprived area or other unknown or unmeasured risk factors). Links between boxes have associated weights representing the probability that that particular pathway occurs. Risk factors such as gender (or ethnicity, which is not shown) may modify these probabilities, and so separate links are shown for boys and girls.

tation $\mathrm{DPG}_{x\times y;z}$ to denote the discrete Poisson/gamma model with latent partition on an $x \times y$ grid and kernel distance scale $z$km.

We then include socioeconomic deprivation and traffic related pollution as additional excess risk factors in the model, and introduce multiplicative terms for gender, ethnicity and the nuisance factors (hospital effects and admission method). Table 6 shows the specific prior distributions chosen for this model. Note that the original scale for the Carstairs deprivation score is centered about zero, but to ensure non-negative contributions from each excess risk factor term we add a positive constant (equal to the absolute value of the minimum) to the score for each area; this does not affect the interpretation of the deprivation index.

Results

Fig. 14 shows the posterior mean of the admission rates $\beta_0^B + \left(\sum_{m\in M} k_{im}^B \gamma_m^B\right) \times \beta_l^B$ and latent influences γ_m^B for the six combinations of latent partition and kernel distance scale for the covariate-unadjusted DPG models for bronchiolitis. Not surprisingly, the largest distance scale (ρ^B=5km) produces the smoothest rate estimates for both the 8×9 and 11×13 latent partition; for a given value of ρ^B, both partitions lead to similar maps for the posterior mean bronchiolitis admissions rates. Results for non-respiratory and unspecified admissions followed a similar pattern for the six models. Partly for computational reasons, we restrict subsequent analyses to the $\mathrm{DPG}_{8\times 9;3}$ model, *i.e.*, with latent influences defined on an 8×9 grid of cells S_m and kernel distance scale $\rho^* = 3$km. The latter choice was based on results from the MRF models which suggest that much of the spatial variation in admission rates is removed after inclusion of the covariate effects. Hence the greater spatial variation expressed by the 1km kernel distance scale compared to larger scales may be unnecessary in the covariate-adjusted model. The 3km and 5km distance scales appear to result in quite similar amounts of spatial smoothing, and so the choice between these two values was arbitrary.

The third column of Fig. 11 shows maps of the posterior mean admission rates $\beta_0^* + \left(\sum_{m\in M} k_{im}^* \gamma_m^*\right) \times \beta_l^*$ for each of the three outcomes estimated from the $\mathrm{DPG}_{8\times 9;3}$ model without adjustment for known covariates; the final column shows the same quantity after accounting for observed nuisance factors and explanatory covariates. These maps show similar spatial patterns of admission rates for each out-

TABLE 6. Prior distributions for the covariate-adjusted $\mathrm{DPG}_{8\times 9;3}$ regression model for bronchiolitis.

Regression coefficients			
Baseline	$\beta_0^B \sim \mathsf{Ga}(\alpha_0^B, \tau_0^B)$	$\tau_0^B = 4\alpha_0^B/\overline{Y}^B = 1.93$	$\alpha_0^B = 0.575$
Carstairs	$\beta_1^B \sim \mathsf{Ga}(\alpha_1^B, \tau_1^B)$	$\tau_1^B = 4\alpha_1^B\overline{X}_{1_E}/\overline{Y}^B = 9.489$	$\alpha_1^B = 0.575$
MHGV	$\beta_2^B \sim \mathsf{Ga}(\alpha_2^B, \tau_2^B)$	$\tau_2^B = 4\alpha_2^B\overline{X}_{2_E}/\overline{Y}^B = 4.188$	$\alpha_2^B = 0.575$
Latent	$\beta_3^B \sim \mathsf{Ga}(\alpha_3^B, \tau_3^B)$	$\tau_3^B = 4\alpha_3^B/\overline{Y}^B = 1.93$	$\alpha_3^B = 0.575$
GP ref.	$\phi_1^B \sim \mathsf{No}(0, \tau_1^B)$	$\tau_1^B = \frac{\left(2.56(\max_i X_{i1_R} - \min_i X_{i1_R})\right)^2}{\log 100} = 0.31$	
Other ref.	$\phi_2^B \sim \mathsf{No}(0, \tau_2^B)$	$\tau_2^B = \frac{\left(2.56(\max_i X_{i2_R} - \min_i X_{i2_R})\right)^2}{\log 100} = 0.31$	
Black	$\phi_3^B \sim \mathsf{No}(0, \tau_3^B)$	$\tau_3^B = \frac{\left(2.56(\max_i X_{i3_R} - \min_i X_{i3_R})\right)^2}{\log 100} = 0.10$	
Indian	$\phi_4^B \sim \mathsf{No}(0, \tau_4^B)$	$\tau_4^B = \frac{\left(2.56(\max_i X_{i4_R} - \min_i X_{i4_R})\right)^2}{\log 100} = 0.20$	
Male	$\phi_4^B \sim \mathsf{No}(0, \tau_4^B)$	$\tau_5^B = \frac{\left(2.56(\max_i X_{i5_R} - \min_i X_{i5_R})\right)^2}{\log 100} = 0.10$	
Other parameters			
Latent risk factors	$\gamma_m^B \sim \mathsf{Ga}(\alpha_{\gamma_m}^B, \tau_\gamma^B)$	$\alpha_{\gamma_m}^B = \tau_\gamma^B \lvert S_m \rvert = 22$	$\tau_\gamma^B = 1.0/\mathrm{km}^2$
Distance scale	$\rho^B \equiv 3\mathrm{km}$		
Hospital precision	$1/\sigma_h^B \sim \mathsf{Ga}(0.5, 0.0005)$		

Notes:
$\overline{Y}^B$ = overall observed admission rate for bronchiolitis in the study region.
$\overline{X}_{j_E}$ = population weighted average value of j^{th} excess risk factor.
X_{ij_R} = j^{th} relative risk factor.

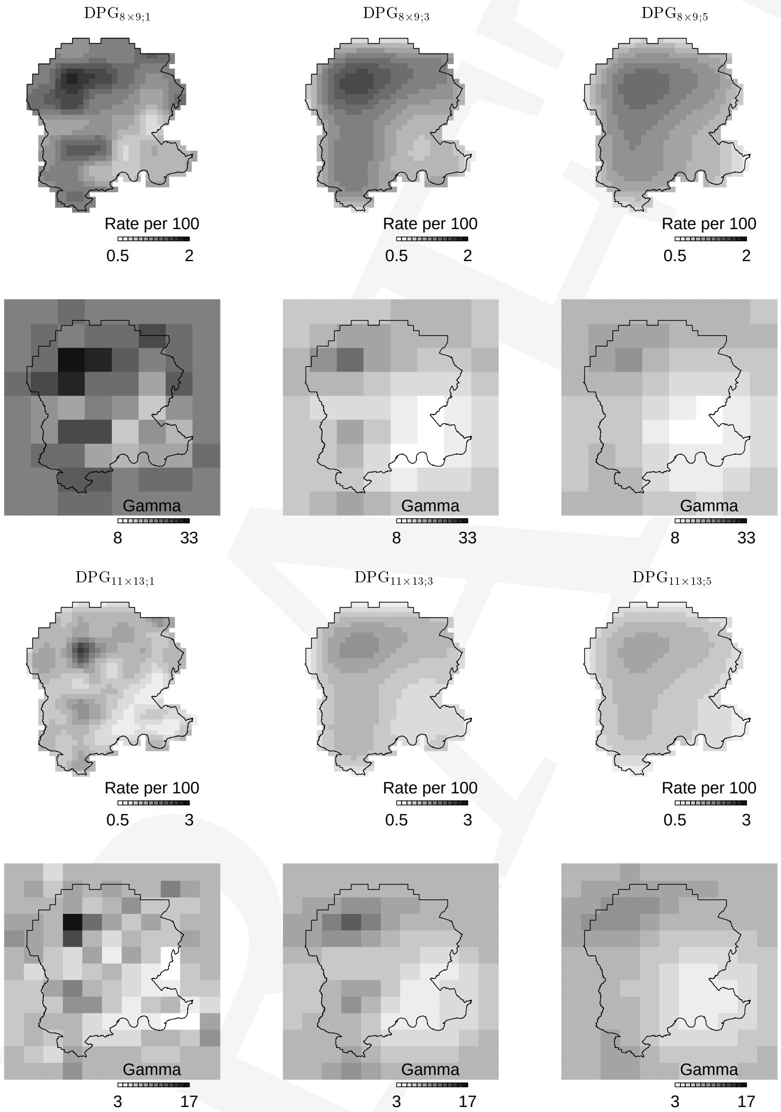

FIGURE 14. Posterior mean admission rates (rows 1 and 3) and latent influences (rows 2 and 4) for the covariate-unadjusted DPG models for bronchiolitis.

come compared to the MRF models, but are considerably smoother.

Maps of the posterior mean proportional admission rates for bronchiolitis relative to all admissions, given by the ratio

$$\frac{\beta_0^B + \left(\sum_{m\in M} k_{im}^B \gamma_m^B\right) \times \beta_l^B}{\sum_{*\in\{B,N,U\}} \beta_0^* + \left(\sum_{m\in M} k_{im}^* \gamma_m^*\right) \times \beta_l^*},$$

are shown in Fig. 12 (bottom row) both before and after adjustment for observed covariates. Also shown are corresponding maps of the posterior probabilities that the proportional admission rate in each grid cell exceeds the average proportional admission rate across the study region. Again, the spatial patterns are similar to those estimated from the MRF model, although the grid cells with significantly elevated or reduced admission ratios for bronchiolitis cover a much larger geographical area due to the greater degree of smoothing induced by the DPG models.

Posterior means and 95% credible intervals for the relative or percentage attributable risks associated with each multiplicative or additive explanatory covariate respectively are shown in the bottom half of Table 5. For multiplicative risk factors, the final two columns give the posterior odds of admission for bronchiolitis versus non-respiratory and unspecified diagnoses respectively associated with exposure to the relevant covariate, *i.e.*, $\exp(\phi_j^B - \phi_j^N)$ and $\exp(\phi_j^B - \phi_j^U)$; for additive risk factors, the final two columns give the excess percentage of admissions for bronchiolitis versus non-respiratory and unspecified diagnoses respectively that are attributed to exposure to the relevant covariate, *i.e.*,

$$\frac{\sum_i C_i(\phi^B) X_{ij} \beta_j^B}{\sum_i C_i(\phi^B) X_{ij} \beta_j^B + C_i(\phi^*) X_{ij} \beta_j^*} - \frac{\Lambda_i^B}{\Lambda_i^B + \Lambda_i^*}, \quad j \in J_E, * = N \text{ or } U$$

where $C_i(\phi^*) = c(\phi^*) \times e^{\sum_{j\in J_R} X_{ij}\phi_j^*}$.

We note that 96.71% (95% CI = 82.93%–99.98%) of bronchiolitis admissions were attributed to the latent spatial risk factors in the $\mathrm{DPG}_{8\times 9;3}$ model *before* adjustment for observed covariates (the remaining admissions were attributed to the baseline intercept term). The corresponding estimates for non-respiratory and unspecified admissions are 98.41% (95% CI = 92.67%–100.0%) and 99.5% (95% CI = 97.73%–100.0%) respectively. Comparison with the results in Table 5 shows that the risk attribution associated with the latent spatial covariate is considerably reduced for all three outcomes after including the nuisance factors and explanatory covariates in the model.

Compared with the MRF model, the DGP model tends to yield slightly more extreme estimates of the hospital effects. Specifically, there was evidence of elevated risk of bronchiolitis admission to hospital 7 relative to the overall bronchiolitis admission rate (RR = 1.50, 95% CI = 1.03–2.38). Non-respiratory admission rates were higher than average to hospitals 6 (RR = 1.30, 95% CI = 1.05–1.63), 7 (RR = 1.26, 95% CI = 1.03–1.59) and 11 (RR = 2.07, 95% CI = 1.56–2.75), while admissions with unspecified diagnoses were higher than average to hospitals 10 (RR = 6.70, 95% CI = 2.78–15.65) and 12 (RR = 5.54, 95% CI = 1.95–15.71) and lower than average to hospitals 3 (RR = 0.27, 95% CI = 0.13–0.55), 6 (RR = 0.37, 95% CI = 0.16–0.80), 8 (RR = 0.08, 95% CI = 0.01–0.40) and 13 (RR = 0.01, 95% CI = 0.00–0.06).

By contrast, the DPG model yields less extreme estimates for the effect of method of referral on admission rates compared to the MRF model. This is probably due to the fact that some hospitals tend to admit most patients via direct referral from their A&E department, while others admit a large proportion following GP referral, and so both the hospital effects and method of referral may offer similar explanations for the spatial pattern of admission rates seen in our data. The only effect of admission method for the DPG models was a higher risk of admission for bronchiolitis for GP-initiated hospital referrals compared to A&E referrals (RR = 1.98, 95% CI = 1.15–3.25).

6.3 Model validation

Model validation for complex hierarchical models is a topic of continuing research. Here we carry out a simple assessment of model adequacy by using each of the Poisson regression, MRF and DPG models reported in preceding sections to predict the number of admissions per grid cell for the validation dataset. Comparison of the distribution of deviance residuals for observed versus predicted counts for the various models suggests that the Poisson regression models lead to more extreme residuals than either the MRF or DPG models; the latter two modeling approaches show similar residual patterns.

7 Marked Point Process Regression Models

The Poisson regression models with identity link introduced in Section 6.2 already show some advantages over the MRF models: they allow for modeling the latent random effects term on a different spatial resolution than the cases and other covariates, and they are capable of modeling excess (additive) as well as relative risk factors. However, we still have to fix a partition $\{S_m\}_{m\in M}$ for modeling the unobserved covariates prior to our analysis (when we do not know on which spatial scale a latent spatial pattern might emerge) and we are still unable to include covariate information on an individual level. In this section we will extend the models of the previous section to a continuous random field version, eliminating all unnecessary aggregation by using all data at their finest spatial scale.

This is achieved by further refining both the areas $\{R_i\}_{i\in I}$ of $\mathcal{Y}$ on which the admission data are given and the areas $\{S_m\}_{m\in M}$ on which the latent gamma terms are given until in the limit both the admissions' locations and the random effects may be represented by continuous random fields. We can then introduce individual attributes to our model by going over to a marked random field.

Thus we view the locations $\{y_n^B\}_{n\in N_B}$, $\{y_n^N\}_{n\in N_N}$, and $\{y_n^U\}_{n\in N_U}$ of children admitted to hospitals for bronchiolitis, of the admissions for non-respiratory illnesses and of the admissions with an unspecified diagnosis, respectively, and their associated vectors $\{a_n^*\}_{n\in N_*}$, $* = B, N$ or U, of individual attributes (*e.g.*, gender) as *marked* points $\{x_n^*\}_{n\in N_*}$, $* = B, N$ or U in the product space $\mathcal{X} = \mathcal{Y} \times \mathcal{A}$ of pairs $x^* = (y^*, a^*)$ of locations $y^* \in \mathcal{Y}$ and attribute vector (marks) $a^* \in \mathcal{A}$. We model the marked points $\{x_n^*\}_{n\in N_*}$, $* = B, N$ or U, as Poisson random fields $N^*(dx) \sim \mathsf{Po}(\Lambda^*(x)w^*(dx))$ on $\mathcal{X}$ with uncertain intensity $\Lambda^*(x)w^*(dx)$. Thus, $N^*(dx)$, $* = B, N$ or U, is a doubly-stochastic Poisson (or Cox) process on the product space. We take the intensity $\Lambda^*(x)w^*(dx)$ to have a density function $\Lambda^*(x)$ with respect to a reference weight measure $w^*(dx)$ intended to reflect the known distribution of the population at risk and their attributes. We model $w^*(dx) = w_Y^*(dy)w_A^*(da|y)$, $* = B, N$ or U, as a product of a non-negative spatial weight measure $w_Y^*(dy)$ and a location-specific conditional probability measure $w_A^*(da|y)$ of attributes.

On the next stage of the hierarchy we model the density $\Lambda^*(x)$, $* = B, N$ or U, as a combination of excess and relative risk factors. The excess (additive) risk factors, which here may include an intercept, traffic related pollution covariates, social deprivation and a spatially varying random effects term, are represented by the term $\sum_{j \in J_E} X_j(x)\beta_j$ with risk factors $X_j(x)$ which may depend on either or both the location y or the attribute vector a and regression coefficients β_j. The relative (multiplicative) risk factors, which here may include gender, ethnicity, admission method, and hospital effects, now enter the model as $\exp(\sum_{j \in J_R} X_j(x)\phi_j)$ with regression coefficients ϕ_j. In all $\Lambda^*(x)$, $* = B, N$ or U, is given by

$$\Lambda(x)^* = c(\phi^*) \times e^{\sum_{j \in J_R} X_j(x)\phi_j^*} \times \sum_{j \in J_E} X_j(x)\beta_j^* \, .$$

Again, $c(\phi^*)$ is a normalizing constant.

7.1 A gamma random field model for the latent term

Infinitely refining the areas $\{S_m\}_{m \in M}$ on which the latent gamma terms are given leads to a gamma random field model for the latent random effects. The gamma random field has a representation as a discrete sum of infinitely many point masses at random locations $s_m \in \mathcal{S}$ and corresponding magnitudes $\gamma_m > 0$. For a gamma random field $\Gamma(ds) \sim \mathsf{Ga}(\alpha(ds), \tau(s))$ with shape measure $\alpha(ds)$ and inverse scale function $\tau(s)$ the sum $\Gamma(A) = \sum_m \{\gamma_m : s_m \in A\}$ of magnitudes for points in any small set A have approximately the gamma probability distribution with shape parameter $\alpha(A)$ and mean $\alpha(A)/\tau(a)$ for any $a \in A$ (if $\tau(a)$ is constant in A this is exact). Furthermore, $\Gamma(A)$ and $\Gamma(B)$ for non-overlapping sets A and B are independent. For a discussion on gamma random fields including their construction see Wolpert and Ickstadt (1998).

The gamma random field can therefore be interpreted as a sum of point sources at random locations s_m with positive magnitudes γ_m. We now regard our unmeasured or unobserved covariates as arising from these point sources, and falling off with increasing distance from these sources at rates described by a kernel $k(x, s)$. The aggregate effect at a location y for an individual with attribute vector a is then given by the kernel mixture $\sum k(x, s_m)\gamma_m$ (where $x = (y, a)$) or in integral form by $\int_{\mathcal{S}} k(x, s)\Gamma(ds)$. The latent random effects term which we again treat to be excess additive can thus be written as

$$X_l(x)\beta_l^* = \big(\sum k^*(x, s_m)\gamma_m^*\big) \times \beta_l^* = \beta_l^* \times \int_{\mathcal{S}} k^*(x, s)\Gamma^*(ds) \, ,$$

where l is the last index of set J_E and $* = B, N$ or U; as in the discrete versions of the models we use a Gaussian kernel $k^*(x,s) = 1/(2\pi\rho^{*2}) \times \exp(-|y-s|^2/(2\rho^{*2}))$.

7.2 Specific point process regression models

We illustrate the continuous Poisson/gamma approach with a model that includes as excess risk factors an intercept, a traffic pollution measure (MHGV), and a spatially varying latent random field. It includes a single relative risk factor, gender, which enters as an individual attribute. We only consider admissions for bronchiolitis. Prior distributions for the regression coefficients β and ϕ were chosen in the same manner as described earlier for the DPG model (Table 6). Prior distributions for the shape measure $\alpha(ds)$ and inverse scale function $\tau(s)$ of the latent gamma random field were chosen to ensure that the each of the three excess risks (intercept, MHGV, and spatial random effect) would have equal prior expectations (so that about 1/3 of the bronchiolitis cases might be associated with each risk factor, *a priori*), with about 90% prior probability that the effect would be between 1/10 and 10 times this amount. In contrast to the DPG models, we treat the distance scale ρ here as uncertain, with a lognormal prior distribution chosen to ensure a prior median interaction distance of 400m with prior 95% interval of 50-3000m.

7.3 Results

Fig. 15 shows the posterior mean of the admission rate $\beta_0^B + \beta_l^B \times \int_S k(x,s)\Gamma(ds)$ attributed to the 'unmeasured' sources (*i.e.*, the baseline and latent covariates). We present these rates integrated over the 1km grid for comparability with the results from the discrete MRF and DPG models; however, the continuous random field model allows risk associated with the latent spatial covariates to be presented consistently at arbitrary spatial scales, and so other partitions could be chosen for presentation if desired. The overall spatial pattern is similar to that estimated using the covariate-unadjusted MRF and DPG models (Fig. 11), although it is less smooth than either of the discrete models. This is consistent with the fact that the posterior mean of the distance scale ρ was 728m with 95% credible interval (489m 1070m), which suggests that spatial correlation in the unexplained (*i.e.*, latent) risk extends over a smaller neighborhood than

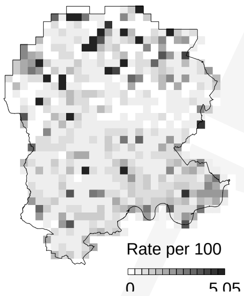

FIGURE 15. Posterior mean admission rates for bronchiolitis attributed to the 'unmeasured' sources (*i.e.*, the baseline and latent covariates), estimated using the point process regression models.

was assumed for either the MRF or DPG models.

The posterior percentage attributed risks for each excess risk factor are summarized as follows: Baseline 12.02% (95% CI 7.49%–16.22%); Latent spatial covariates 87.46% (95% CI 83.17%–92.05%); MHGV 0.52% (95% CI 0.03%–2.00%). These are broadly consistent with our findings for the DPG model. In particular, only a very small fraction of cases are attributed to traffic related pollution exposure, with the majority of admissions being attributed to the latent spatial covariate. Note that since we do not include social deprivation in this model, the results are not directly comparable with those reported for the covariate-adjusted DPG model in Table 5; however, since social deprivation shows a strong spatial pattern, the effects of this covariate are likely to be captured by the latent spatial random field.

The posterior mean relative risk of bronchiolitis admission for boys compared to girls was 1.61 (95% CI = 1.07–2.58). This agrees with the findings from the discrete MRF and DPG models that areas with a higher proportion of boys tend to have higher admission rates for bronchiolitis; however, the relative risk parameter in the present model is directly interpretable in terms of the *individual* level risk associated with being a boy, rather than the *ecological* level risk associated with living in areas with a high proportion of boys. The risk estimate is also somewhat smaller than we obtained from the MRF or DPG models (1.61 compared to 2.70 and 2.50 respectively). This suggests that the discrete model estimates based on aggregate covariates may be subject to ecological bias, although the models are not adjusted for the same set of covariates and so cannot be compared directly.

8 Discussion

This Case Study attempts to address an important public health question using a variety of routinely collected data sources, many of which have only recently been made available for epidemiological research. As in any observational study one must be careful in making inference from these non-experimental data. The specific concerns inherent in the HES data are not yet well understood, but some valuable insights into their quality, interpretation and usefulness have emerged from this study. For example, it is now clear that London hospitals differ widely in their diagnostic coding policies and in the referral process by which their patients are admitted. These factors may lead to large apparent geographical variations in the ICD-specific hospital admission rates that are unrelated to variations in the incidence and severity of bronchiolitis across the study region. Careful interpretation of the substantive findings of this study is therefore imperative in order to avoid misleading conclusions.

We now return to the four questions originally posed in Section 1.

1. Do infants living in areas of high traffic density have higher rates of respiratory illness?

Analysis based on our Poisson regression and MRF models does not reveal higher respiratory illness rates in areas of high traffic density; if anything, it suggests that rates might be lower there. This probably reflects confounding (due to multicollinearity) and bias (due to systematic differences in diagnostic coding between inner city and outer city hospitals) in the HES data. This interpretation is supported by the discovery of a similar negative association between MHGV and risk of admission for non-respiratory diagnoses, suggesting that the effect of traffic exposure is non-specific. Furthermore, since admission rates for all three diagnoses (B, N and U) are expected to be affected similarly by any biases in the HES data, the *ratios* of admission rates for the different diagnoses should be less affected. We found no 'significant' difference between the relative risk estimates for the effect of MHGV on admission for bronchiolitis compared with either non-respiratory or unspecified diagnoses, in the sense that the 95% credible intervals for the relative risk ratios include one (see final two columns of Table 5). This again suggests that exposure to high traffic density does not have a specific effect on admissions for bronchiolitis in infants.

To check whether the true association between traffic density and bronchiolitis might be masked by correlation between MHGV and other covariates such as deprivation or hospital effects, we re-ran the MRF model including only MHGV and a spatial random effect as linear predictors. The estimated association with MHGV remained negative, although there was considerable posterior uncertainty and the 95% credible interval for the relative risk included one (RR = 0.24; 95% CI = 0.01–1.22).

The discrete and continuous Poisson/gamma (DPG and CPG) models were proposed as more flexible alternatives to the MRF analysis. They allow both excess and relative risk factors to be included in the same model (the MRF models allow only relative risks). Our choices between additive and multiplicative forms for the covariate effects in these models were based on *a priori* considerations rather than being data driven, since the data rarely offer conclusive support in favor of one or other model (Breslow and Day, 1980). However, we may view the contrast between the log linear MRF models on the one hand, and the linear DPG/CPG models on the other, as a form of sensitivity analysis to the assumed form of the relationship between traffic pollution and risk of bronchiolitis. Despite different interpretations for the regression coefficients associated with the relative versus excess risk factor models, and with different prior assumptions, both approaches lead us to the same *qualitative* conclusion: these data provide no evidence to support concerns that chronic exposure to traffic pollution poses a significant risk of hospitalization for bronchiolitis in infants living in northwest London. To this extent, the methodologies appeared robust, although other applications may show more sensitivity (particularly in applications where the data show a stronger positive association between exposure and response).

This conclusion appears to conflict with the results of our preliminary ward-level analysis of bronchiolitis admissions in the whole of Greater London (Section 2.1). However, the present study used much finer levels of spatial aggregation (indeed the CPG model requires no aggregation at all), and so may suffer less from the possible effects of ecological bias. Our handling of admissions with unspecified diagnoses also differs between studies: in the original analysis, we treated these cases as missing at random, and reassigned them as bronchiolitis, other respiratory or non-respiratory admissions according to the observed admission rates for the three diagnoses. This would lead to more apparent cases of bronchiolitis in central London (the southeast corner of the present study region) where many of the unspecified cases are clustered and where traffic density is also highest. However, it is not clear that the true diagnosis for these admissions is really missing at random (it may be that cases presenting with complicated conditions, or infants with non-specific symptoms, are more likely to be assigned an ICD 799 code), and so we chose to model unspecified diagnoses as a separate outcome in the present study. This analysis also focuses on a subset of the original study region (*i.e.*, the northwest quadrant). Preliminary examination of the association between bronchiolitis admission rates and traffic density in other quadrants shows inconsistencies across London: there is some suggestion that the northeast and southeast quadrants show a slight positive relationship, while the southwest shows no relationship. However, these findings may well be driven by a few outlying areas and further detailed follow-up analysis is required.

2. Which pollutants or vehicle types are associated with the greatest risk?

It proved impossible to distinguish among the effects of different pollutants on risk of bronchiolitis admission due to the extremely high correlations between the emissions data available for this study. Since we found no evidence of an increase in risk of bronchiolitis admissions associated with exposure to high traffic density, we did not pursue this question further.

3. Do factors such as socioeconomic deprivation of the area of residence, or individual characteristics such as gender increase or modify this risk?

All our models showed a consistent and large positive relationship between deprivation and admission rates for all three diagnoses, suggesting that infants living in socioeconomically disadvantaged areas are more likely to be admitted to hospital, irrespective of the cause.

The Carstairs index used to measure deprivation in this study is a relatively crude proxy for a range of factors such as poverty, smoking, poor diet, poor housing conditions and overcrowding, which may explain this finding, although more detailed individual level data on such covariates would be needed in order to further identify the specific risk factors involved. It may also reflect such non-ætiological factors as access to health care, use of A&E (emergency room) services and thresholds for admission to hospital. Comparison of relative admission rates for bronchiolitis versus non-respiratory or unspecified diagnoses revealed that deprivation had a significantly greater impact on the former. Under the MRF model, infants living in the most deprived area were over 2.5 times more likely to be admitted to hospital for bronchiolitis than for either non-respiratory or unspecified diagnoses compared with infants living in the most affluent area; under the DPG model, approximately 12%–15% more admissions for bronchiolitis were attributed to living in a deprived rather than affluent area compared to non-respiratory or unspecified diagnoses.

Investigation of *individual* covariate effects is not feasible using the area-level MRF and DPG models. Instead it was necessary to treat attributes such as gender as aggregated covariates, which complicates interpretation. This problem is avoided by using the CPG model, which has the great advantage of allowing all data to enter at their natural level of (dis)aggregation and so allows regression coefficients for individual attributes to be interpreted directly as *individual* level risks rather than as *ecological* level risks. The gender effect in this model was consistent with that in the area-level models (all show that boys experience elevated risk), but much better-defined (the 95% credible intervals were more than five times narrower).

4. Can we account for residual confounding due to unknown or unmeasured risk factors?

Both the MRF and Poisson-gamma models addressed this issue through the inclusion of spatial random effects, although the distributional form varied between models. One difficulty with the area-level MRF and DPG models is that the spatial structure for these random effects must be specified *a priori*; however, there is generally very little scientific evidence to inform this choice. By contrast, the location and shape of the latent spatial random field in the CPG model is estimated from the data and so avoids the need to impose arbitrary prior assumptions about this aspect of the model.

In summary, this Case Study revealed no association between chronic exposure to traffic pollution and hospital admissions for viral bronchiolitis strong enough to overcome other sources of variation in the observational data. It revealed the need for a systematic strengthening of diagnostic coding standards across hospitals before the HES database will support detailed epidemiological analyses using sophisticated modern statistical modeling. However, without the detailed statistical modeling carried out for this study it would remain unclear whether the unexpected negative results for the association between traffic density and bronchiolitis admissions were indicative of fundamental inadequacies in the data or to the failure of simple models to account fully for the multiple sources of variation. In this respect the Bayesian models presented in this paper have allowed us to gain considerable insight into the structure and nature of the available data and the potential sources of bias, have raised a number of questions which require further study, and have pointed the way to needed improvements in HES data collection.

Acknowledgments

The authors gratefully acknowledge the support of the following grants for this work: US NSF grant DMS-9626829; MRC Career Establishment Grant No. G9803841; Pan-Thames Environmental R&D Initiative Grant No. 339. The authors are grateful to the Office for National Statistics and the Department of Health and the London Research Centre for providing data. The authors would also like to thank Lucy Sadler from the London Research Centre for helpful discussions concerning the road traffic data used in this project, the London School of Hygiene and Tropical Medicine for the joint linking of the road networks, and Richard Arnold for helpful discussions. The work is based on data provided with the support of the ESRC and JISC and uses census and boundary material which is copyright of the Crown, the Post Office and the ED-LINE consortium. The computing facilities used for the work were partially supported by a Wellcome equipment grant.

9 References

Anderson, H., Butland, B., and Strachan, D. (1994). Trends in prevalence and severity of childhood asthma. *British Medical Journal* **308**, 1600–4.

Anderson, H. R., Ponce de Leon, A., Bland, J. M., Bower, J. S., Emberlin, J., and Strachan, D. P. (1998). Air pollution, pollens, and daily admissions for asthma in London 1987–92. *Thorax* **53**, 842–8.

Bernardinelli, L., Pascutto, C., Best, N., and Gilks, W. (1997). Disease mapping with errors in covariates. *Statistics in Medicine* **16**, 741–52.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B (Methodological)* **36**, 192–236. (With discussion).

Besag, J., York, J., and Mollié, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics* **43**, 1–59. (With discussion).

Best, N., Bennett, J., Cockings, S., Falconer, S., Maheswaran, R., Wakefield, J., Smith, A., and Elliot, P. (1998a). Small area mapping and statistical analyses of variations in disease risk and possible explanatory factors in the Thames Regions. Tech. rep., Department of Epidemiology & Public Health, Imperial College, London.

Best, N. G., Ickstadt, K., and Wolpert, R. L. (1998b). Spatial Poisson regression for health and exposure data measured at disparate resolutions. Discussion Paper 98-36, Duke University ISDS, USA. URL `ftp://ftp.isds.duke.edu/pub/WorkingPapers/98-36.ps`.

Bithell, J. F., Dutton, S. J., Neary, N. M., and Vincent, T. J. (1995). Controlling for socio-economic confounding using regression methods. *Journal of Epidemiology and Community Health, Supplement 2* **49**, S15–S19.

Breslow, N. E. and Day, N. E. (1980). *Statistical methods in cancer research. Volume 1 - The analysis of case-control studies.* No. 32 in IARC Scientific Publications. International Agency for Research on Cancer, Lyon.

Briggs, D. J., Collins, S., Elliott, P., Fischer, P., Kingham, S., Lebret, E., Pryl, K., van Reeuwijk, H., Smallbone, K., and van der Veen, A. (1997). Mapping urban air pollution using GIS: A regression-based approach. *International Journal of Geographic Information Science* **11**, 699–718.

Brunekreef, B., Dockery, D. W., and Krzyanowski, M. (1995). Epidemiologic studies on short-term effects of low levels of major ambient air pollution components. *Environmental Health Perspectives* **103 (Supplement 2)**, 3–13.

Brunekreef, B., Janssen, N. A. H., de Hartog, J., Harssema, H., and Knape, M. (1997). Air pollution from truck traffic and lung function in children living near motorways. *Epidemiology* **8**, 298–303.

Buckingham, C., Clewley, L., Hutchinson, D., Sadler, L., and Shah, S. (1997). London Atmospheric Emissions Inventory. Tech. rep., London Research Centre.

Burney, P. (1988). Asthma deaths in England and Wales 1931-85: evidence for a true increase in asthma mortality. *Journal of Epidemiology and Community Health* **42**, 316–20.

Carr-Hill, R., Hardman, G., Martin, S., Peacock, S., Sheldon, T., and Smith, P. (1994). A formula for distributing NHS revenues based on small area use of hospital beds. Tech. rep., Occasional Paper Series, Centre for Health Economics, University of York.

Carstairs, V. and Morris, R. (1991). *Deprivation and health in Scotland.* Aberdeen University Press, Aberdeen.

Chestnut, L. G., Schwartz, J., Savitz, D. A., and Burchfiel, C. M. (1989). Pulmonary function and ambient particulate matter: epidemiological evidence from NHANES I. *Arch Environ Health* **46**, 135–44.

Clayton, D. G. and Bernardinelli, L. (1992). Bayesian methods for mapping disease risk. In Elliott et al. (1992), chap. 18, 205–20. Paperback edition published 1996.

Committee of Medical Effects of Air Pollution (1995a). *Asthma and outdoor air pollution.* Her Majesty's Stationary Office, London.

Committee of Medical Effects of Air Pollution (1995b). *Non-biological particles and health.* Her Majesty's Stationary Office, London.

Committee of the Environmental and Occupational Health Assembly of the American Thoracic Society (1996). Health effects of outdoor pollution. *American Journal of Respiratory Critical Care Medicine* **153**, 3–50.

Cummins, P. and Rathwell, T. (1991). Geographical Information Systems and the National Health Service. Tech. rep., Association for Geographic Information Education, Training and Research Publication No.3.

Department of Health (1998). *How HES Data is Processed 1996/97 Datayear v.3.0. SD2HES.*

Dixon, J., Sanderson, C., Elliott, P., Walls, P., Jones, J., and Petticrew, M. (1998). Assessment of the reproducibility of clinical coding in routinely collected hospital activity data: a study of two hospitals. *J Pub Health Med* **20**, 63–69.

Dockery, D., Pope, III, C., Xiping, X., Spengler, J., Ware, J., Fay, M., Ferris, B., and Speizer, F. (1993). An association between air pollution and mortality in six US cities. *New England Journal of Medicine* **329**, 1753–1759.

Dockery, D. W., Spengler, J. D., Reed, M. P., and Ware, J. (1981). Relationships among personal, indoor and outdoor NO_2 measurements. *Environment International* **5**, 101–7.

Duhme, H., Weiland, S. K., Keil, U., Kraemer, B., Schmid, M., and Stender, M. (1996). The association between self-reported symptoms of asthma and allergic rhinitis and self-reported traffic density on street of residence in adolescents. *Epidemiology* **7**, 578–82.

Edwards, J., Walters, S., and Griffiths, R. (1994). Hospital admissions for asthma in preschool children: relationship to major roads in Birmingham, United Kingdom. *Archives of Environmental Health* **49**, 223–7.

Elliott, P., Cuzick, J., English, D., and Stern, R., eds. (1992). *Geographical and environmental epidemiology: Methods for small-area studies.* Oxford University Press, Oxford. Paperback edition published 1996.

Gardner, M. T., Winter, P. D., and Barker, D. J. P. (1984). *Atlas of mortality from selected diseases in England and Wales 1968–1978.* Wiley, Chichester.

Gatrell, A. C., Dunn, C. E., and Boyle, P. J. (1991). The relative utility of the Central Postcode Directory and Pinpoint Address Code in applications of Geographical Information Systems. *Environment and Planning A* **23**, 1447–1458.

Gatrell, A. C. and Senior, M. (1999). Health and healthcare applications. In *Geographical Information Systems: Principles, Techniques, Applications and Management*, eds. P. Longley, M. Goodchild, D. Maguire, and D. Rhind, vol. 2, 925–938. John Wiley & Sons.

Greenberg, M. R. (1983). *Urbanization and cancer mortality.* Oxford University Press, Oxford.

Greenland, S. (1992). Divergent biases in ecologic and individual-level studies. *Statistics in Medicine* **11**, 1209–1223.

HMSO (1994). *Royal Commission on Environmental Pollution. Eighteenth Report. Transport and the Environment.* HMSO, London.

Hoek, G., Brunekreef, B., and Hofschreuder, P. (1989). Indoor exposure to airborne particles and nitrogen dioxide during an air pollution episode. *J Air Poll Contr Assoc* **39**, 1348–9.

Howe, H. L., Keller, J. E., and Lehnherr, M. (1993). Relation between population density and cancer incidence, Illinois, 1986–90. *Am J Epidemiol* **138**, 29–36.

Jolley, D., Jarman, B., and Elliott, P. (1992). Socio-economic confounding. In Elliott et al. (1992), chap. 11, 115–124. Paperback edition published 1996.

Kelsall, J. E. and Wakefield, J. C. (1999). Discussion of 'Bayesian models for spatially correlated disease and exposure data' by Best et al. In *Bayesian Statistics 6*, eds. J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith, 131–56. Oxford University Press, Oxford.

King, G. (1997). *A solution to the ecological inference problem.* Princeton University Press, Princeton.

Kleinschmidt, I., Hills, M., and Elliott, P. (1995). Smoking behaviour can be predicted by neighbourhood deprivation measures. *J Epidemiol Community Health* **49 (Supplement 2)**, S72–S77.

Knorr-Held, L. and Besag, J. (1998). Modelling risk from a disease in time and space. *Statistics in Medicine* **17**, 2045–60.

Livingstone, A. E., Shaddick, G., Grundy, C., and Elliott, P. (1996). Do people living near inner city main roads have more asthma needing treatment? Case-control study. *British Medical Journal* **312**, 676–7.

Mage, D. T. and Buckley, T. J. (1995). The relationship between personal exposures and ambient concentrations of particulate matter. Presented at the 88th Annual Meeting of the Air and Waste Management Association, San Antonio, TX.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models.* Chapman and Hall, London.

McKee, M. and Petticrew, M. (1993). Disease staging — a casemix system for purchasers? *J Pub Health Med* **15**, 25–36.

Mollié, A. (1996). Bayesian mapping of disease. In *Markov Chain Monte Carlo in Practice*, eds. W. R. Gilks, S. Richardson, and D. J. Spiegelhalter, 359–379. Chapman & Hall, New York, NY, USA.

Murakami, M., Ono, M., and Tamura, K. (1990). Health problems of residents along heavy-traffic roads. *Journal of Human Ergology* **19**, 101–6.

MVA Limited (1999). London Transportation Studies Technical Note 25: B1.10 Base Forcast Validation. Tech. rep., MVA Limited, 115 Shaftesbury Avenue, London.

Noy, D., Brunekreef, B., Boleij, J. S. M., Houthuijs, D., and De Koning, R. (1990). The assessment of personal exposure to nitrogen dioxide in epidemiological studies. *Environment International* **12**, 407–11.

Osterlee, A., Drijver, M., Lebret, E., and Brunekreef, B. (1996). Chronic respiratory symptoms in children and adults living along streets with high traffic density. *Occupational Environment Medicine* **53**, 241–7.

Ozkaynak, H. and Spengler, J. (1996). The role of outdoor particulate matter in assessing total human exposure. In *Particles in Outdoor Air: Concentrations and Health Effects*, eds. R. Wilson and J. Spengler, 63–84. Harvard University Press, Cambridge, MA.

Parkin, D. M., Whelan, S. L., Ferlay, J., Raymond, L., and Young, J. (1997). *Cancer incidence in Five Continents.* IARC Scientific Publications No. 143, Lyon.

Pershagen, G., Rylander, E., Norberg, S., Eriksson, M., and Nordell, S. (1995). Air pollution involving nitrogen dioxide exposure and wheezing bronchitis in children. *International Journal of Epidemiology* **24**, 1147–53.

Pope, C., Thun, M., Namboodiri, M., Dockery, D., Evans, J., Speizer, F., and Heath, C. (1995). Particulate air pollution as a predictor of mortality in a prospective study of US adults. *American Journal of Respiratory Critical Care Medicine* **151**, 669–674.

Quackenboss, J. J., Krzyanowski, M., and Lebowitz, M. D. (1991). Exposure assessment approaches to evaluate respiratory health effects of particulate matter and nitrogen dioxide.

Journal of Exposure Analysis and Environmental Epidemiology **1**, 83–106.

Quackenboss, J. J., Spengler, J., Kanarek, M. S., Letz, R., and Duffy, C. P. (1986). Personal exposure to nitorgen dioxide: relationship to indoor/outdoor air quality and activity patterns. *Environ Sci Tecnol* **20**, 775–83.

Raper, J. F., Rhind, D. W., and Shepherd, J. W. (1992). *Postcodes: the new geography.* Longman Group UK Limited, Essex.

Richardson, S. (1992). Statistical methods for geographical correlation studies. In Elliott et al. (1992), chap. 17, 181–204. Paperback edition published 1996.

Richardson, S., Monfort, C., Green, M., Draper, G., and Muirhead, C. (1995). Spatial variation of natural radiation and childhood leukaemia incidence in Great Britain. *Statistics in Medicine* **14**, 2487–2501.

Roberts, G. O. (1996). Markov chain concepts related to sampling algorithms. In *Markov chain Monte Carlo Methods in practice*, 45–58. Chapman & Hall, New York, NY, USA.

Robinson, W. S. (1950). Ecological correlations and the behaviour of individuals. *American Sociological Review* **15**, 351–357.

Roemer, W., Hoek, G., and Brunekreef, B. (1993). Effect of ambient winter air pollution on respiratory health of children with chronic respiratory symptoms. *Am Rev Respir Dis* **147**, 118–24.

Schwartz, J. (1989). Lung function and chronic exposure to air pollution: a cross-sectional analysis of NHANES II. *Environ Research* **50**, 309–21.

Schwartz, J. (1993). Particulate air pollution and chronic respiratory disease. *Environmental Research* **62**, 7–13.

Schwartz, J. (1994). Air pollution and daily mortality: a review and meta-analysis. *Environmental Research* **64**, 36–52.

Sheldon, T. A., Smith, P., Borowitz, B., Martin, S., and Carr-Hill, R. (1994). Attempt at deriving a formula for setting general practitioner fundholding budgets. *British Medical Journal* **309**, 1059–64.

Shima, M. and Adachi, M. (1998). Indoor nitrogen dioxide in homes along trunk roads with heavy traffic. *Occup Environ Med* **55**, 428–33.

Smith, R. and Jarvis, C. (1998). Just the medicine. *Mapping Awareness* **12**, 30–33.

Spiegelhalter, D., Thomas, A., and Best, N. (1999). WinBUGS user manual, version 1.2. URL `http://www.mrc-bsu.cam.ac.uk/bugs`. On-line User Manual.

Strachan, D. P. (1991). Damp housing, mould allergy and childhood asthma. *Proc R Coll Physicians Edinb* **21**, 140–6.

Strachan, D. P., Anderson, H. R., Limb, E. S., O'Neill, A., and Wells, N. (1994). A national survey of asthma prevalence, severity, and treatment in Great Britain. *Arch Dis Child* **70**, 174–8.

Urell, M. and Samet, J. (1993). Particulate air pollution and health — new evidence on an old problem. *Am Rev Respir Dis* **147**, 1334–5.

Vostal, J. J. (1994). Physiologically based assessment of human exposure to urban air pollutions and its significance for public health risk. *Environmental Health Perspectives* **102 (Supplement 4)**, 101–6.

Waldron, G., Pottle, B., and Dod, J. (1995). Asthma and motorways — one district's experience. *Journal of Public Health Medicine* **17**, 85–9.

Wallace, L. (1996). Indoor particles: a review. *J Air Waste Manage Assoc* **46**, 98–126.

Walshe, K., Harrison, N., and Renshaw, M. (1993). Comparison of the quality of patient data collected by hospital and departmental computer systems. *Health Trends* **25**, 105–8.

Weiland, S., Mundt, K., Ruckmann, A., and Keil, U. (1994). Self-reported wheezing and allergic rhinitis in children and traffic density on street of residence. *AEP* **4**, 243–7.

Williams, B. (1997). Utilisation of national health service hospital in England by private patients 1989-95. *Health Trends* **29**, 21–5.

Williams, D. R. R., Anthony, P., Young, R. J., and Tomlinson, S. (1994). Interpreting hospital admissions data across the Korner divide: the example of diabetes in the North Western region. *Diabetic Med* **11**, 166–9.

Wjst, M., Reitmeir, P., Dold, S., Wulff, A., Nicolai, T., Freifrau von Loeffelholz-Colberg, E., and von Mutius, E. (1993). Road traffic and adverse effects on respiratory health in children. *British Medical Journal* **307**, 596–600.

Wolpert, R. L. and Ickstadt, K. (1998). Simulation of Lévy random fields. In *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. D. Dey, P. Müller, and D. Sinha, vol. 133 of *Lecture Notes in Statistics*, 227–242. Springer-Verlag, New York, NY, USA.