

Using Bayesian graphical models to model biases in  
observational studies and to combine multiple data  
sources: Application to low birth-weight and water  
disinfection by-products

Nuoo-Ting (Jassy) Molitor, Nicky Best, Chris Jackson and  
Sylvia Richardson  
Imperial College UK

September 30, 2008

## Abstract

Data in the social, behavioral and health sciences frequently come from observational studies instead of controlled experiments. In addition to random errors, observational data typically contain additional sources of uncertainty such as missing values, unmeasured confounders, and selection biases. Also, the research question is often different to that which a particular data source was designed to answer, and so not all relevant variables are measured. As a result, multiple data sources are often necessary to identify the biases and inform about different aspects of the research question. Bayesian graphical models provide a coherent way to connect a series of local sub-models, based on different data sets, into a global unified analysis. In this manuscript, we present a unified modeling framework that will account for multiple biases simultaneously and give more accurate parameter estimates than standard approaches. We illustrate our approach by analyzing data from a study of water disinfection by-products and adverse birth outcomes in the U.K.

Key words: Bayesian graphical model; multiple biases; sensitivity analysis; water disinfection by-product; adverse birth outcomes.

## 1 Introduction

Many social, behavioral, and health science studies are observational in nature and do not arise from carefully-controlled experimental designs. Observational data typically suffer from problems such as missing values, unmeasured confounders, measurement errors, and selection effects. Each of these

factors represents a source of uncertainty in the data, but when analysing such data, these uncertainties, other than simple random errors, are often ignored or assumed to be random within the level of controlled covariates. Although these assumptions allow us to apply standard analysis techniques, the findings of such studies based on these assumptions tend to be biased and lead to inaccurate conclusions (Greenland, 2003, 2005). Recently, Greenland (2005) proposed a so-called multiple bias model which accounts for several uncertainties at the same time, and allows us to understand how much of an impact each of these uncertainties has in terms of biasing and / or affecting the precision of the results. Also, sensitivity analyses have been conducted by several researchers (Langholz, 2001; Lin *et al.*, 1998; Wakefield, 2003; Rosenbaum, 2004; McCandless *et al.*, 2007) to examine hidden biases in observational studies.

Each of the above approaches to handling bias in observational studies involves making plausible assumptions about the nature of potential biases and examining sensitivity to these assumptions. This is because, in general, it is hard to detect or quantify biases within a single dataset and other sources of information are needed. In addition, combining different data sources allows all available evidence to be incorporated into the modelling process, so that the uncertainty in the parameter estimates fully reflects the evidence base. Further, different types of observational data have different strengths and limitations. For example, administrative data sources such as disease registers are usually population based but contain only limited information on each subject, while other administrative data such as census output contain mainly aggregate (area) level information

rather than personal-level data. Hence, if one is interested in addressing a research question related to individual-level effects with potentially complex patterns of confounding and interaction, then administrative data can not provide enough detailed information. On the other hand, survey data often contain rich personal-level information, but the sample size is usually small and problems of selection biases and missing data are common. As a result, findings based on survey data lack statistical power and can be difficult to generalise to an entire population. Therefore, in order to answer different aspects of the research question of interest properly, as well as being able to identify different biases which are hard to detect within a single data set, one needs to combine multiple data sources.

In this paper, we are interested in a setting where multiple sources of data need to be combined due to two reasons. The first is to gain benefit from the increase in power associated with analyzing a large combined data set, and the second is to use different sources of information to reflect uncertainty due to missing outcomes and covariates. The estimation of missing values is achieved through consideration of aggregate-level information such as census output, together with detailed survey data on a subset of individuals, to impute individual-level covariate and outcome data in a population-based health register. Our approach builds on our previous work in this area. In particular, Molitor *et al.* (2006) used aggregate community-level central site air pollution data to help impute missing long-term and personal-level air pollution exposure, while Jackson *et al.* (2008) considered a problem closely related to the present paper, and developed a two-stage model to impute missing confounders in a population register using additional aggregate and

survey data.

In order to combine multiple data sources and draw inference from them, we need to utilize complex modelling techniques. We adopt a Bayesian graphical modelling framework which allows us to build up a series of local sub-models based on different data sources and link them together into a coherent global analysis (Spiegelhalter, 1998; Richardson and Best, 2003). We use this approach to analyze data obtained from an environmental epidemiological study which was undertaken in the authors' department. The goal of this study is to examine the association between adverse birth outcomes such as low birthweight (birth weight  $< 2.5$  kg) and mothers' exposure to water disinfection byproducts (DBPs) such as trihalomethanes (THMs). Water disinfection byproducts occur when chlorine (which is routinely added to tap water supplies in the UK for disinfection purposes) reacts with natural organic matter (e.g. humic and fulvic acid) in the water and forms a range of halogenated organic compounds (Rook, 1974). Any relationship between adverse birth outcomes and DBP's is likely to be small; adverse birth outcomes are also relatively rare (approximately 6% of babies weigh less than 2.5kg at birth in the UK), hence large sample sizes are required to obtain the necessary statistical power to detect potential associations. In previous studies the evidence for an association has been inconclusive (Toledano *et al.*, 2005; Nieuwenhuijsen *et al.*, 2000a,b). In the largest study to date, Toledano *et al.* (2005) used population-based administrative data from the National Birth Registry (NBR)([www.gro.gov.uk/gro/content/research/index.asp](http://www.gro.gov.uk/gro/content/research/index.asp)) for three water supply regions in the UK and found a small excess risk of low birthweight associated with exposure to high levels of THMs, which was sta-

tistically significant in one of the three regions. However, the NBR suffers from two important deficiencies. Firstly, gestational age is not recorded in the NBR. There are two main mechanisms leading to low birthweight: premature birth (baby born at  $< 37$  weeks gestation) or intra-uterine growth retardation (which may result in a baby born at full term, i.e.  $\geq 37$  weeks gestation, being of low birthweight). Since the causal determinants involved in these two processes are different (Kramer, 1987; Barros *et al.*, 1992), it is important to distinguish between pre- and full-term low birthweight babies when examining the association with THM exposure. Secondly, the NBR records very limited information about the mother, and established risk factors for low birthweight, such as maternal smoking during pregnancy and maternal ethnicity, are not available. The distribution of both these variables tends to vary geographically, so they could potentially confound any association between environmental exposures such as THMs and low birthweight. We therefore draw on two additional data sources to help impute the partially missing outcome (pre- or full-term low birthweight) and missing confounders (maternal smoking and ethnicity) in the NBR: (1) survey data from the first sweep of the Millennium Cohort Study (MCS)([www.cls.ioe.ac.uk/studies.asp?section=000100020001](http://www.cls.ioe.ac.uk/studies.asp?section=000100020001)); (2) aggregate information on ethnicity from the 2001 census ([www.statistics.gov.uk](http://www.statistics.gov.uk)) and small-area estimates of tobacco expenditure derived from consumer surveys ([www.caci.co.uk](http://www.caci.co.uk)).

The remainder of this paper is organised as follows. In Section 2, we introduce details of each data source used to investigate the association between low birthweight babies and mother's exposure to THMs. In Sec-

tion 3, we explain how we build up the unified model for these multiple data sources by building and linking different sub-models through the use of Bayesian graphical models, and provide mathematical details of our model. In Section 4, we describe a simulation study to check model performance and in Section 5 we present the results from the real data analysis. In Section 6, we discuss our results and modelling approach and suggest future work.

## 2 Data sources

Here we detail the various data sources used in this manuscript.

- National Birth Registry (NBR)

The UK National Birth Registry provides routinely collected data related to all births in the country, such as sex, birth weight, postcode of residence at birth, date of birth, and mother's age. In our study, the analysis was restricted to singleton births who were born between September 2000 to August 2001 in an area serviced by a water supply company in northern England. This data set has been linked to modelled estimates of THM concentrations for water supply zones in the study region using a postcode-to-water supply zone link file developed by the Small Area Health Statistics Unit (SAHSU) (Toledano *et al.*, 2005; Whitaker *et al.*, 2005). Water zones are defined by the water supply company and cover a population of around 50,000 who are all supplied with water from the same source.

- Millennium Cohort Study (MCS)

Millennium Cohort Study has been set up in order to understand the impact of social conditions surrounding birth and early childhood on health over the life course of subjects involved (Smith and Joshi, 2002). The resulting data include rich personal-level variables relevant to the present study, including mother's age, ethnicity, smoking status during pregnancy, income, education, and baby's sex, birth weight and gestational age. In England, subjects were recruited between September 2000 and August 2001 using cluster randomised sampling based on stratifying residential wards (administrative areas containing around 5000 individuals) into three categories: 'advantaged', 'disadvantaged' and 'high ethnic minority'. Hence, adjustment must be made for these strata in any analysis to ensure inference is representative of the general population. The postcode of birth for all MCS subjects was made available to us under special license, and after using this to match MCS subjects to those whose residence is served by the water company in northern England, we end up with 1333 singleton births.

- Aggregate data

The ratio of non-white to white residents in each census output area (COA, which contain around 200-300 individuals) was obtained from the 2001 Census, and the average annual income and average weekly expenditure on tobacco as a proportion of total expenditure on tobacco, wine, beer, fruit, vegetables and saturated fat in each COA were obtained from Consumer survey data compiled by CACI Information Solutions Limited. These were linked to each individual in the MCS

and NBR using postcode-to-COA link files developed by SAHSU. The ward-level Carstairs deprivation index (Carstairs and Morris, 1991), a general indicator of area socio-economic status based on rates of car ownership, low social class, household overcrowding and male unemployment from the 2001 Census, was also linked to each individual using their postcode.

### 3 Model Setup

We begin by presenting an overview of our model using a graphical representation, and then provide mathematical details of the model setup.

#### 3.1 Graphical representation of our model

Figure 1 shows a graphical representation of our model for imputing missing outcomes and missing covariates by combining administrative, survey and aggregate data. Following standard notation for graphical models (e.g. Spiegelhalter (1998)), the ovals or ‘nodes’ in the graph represent the variables (data, missing values and parameters) and arrows between nodes indicate direct dependencies between variables. This graph can be thought of as a schematic representation of our model to convey the essential structure, with a view to demonstrating in a fairly generic way how models for combining multiple data sources can be built. Since our final model is a Bayesian full probability model, it can also be represented by a more elaborate version of this graph corresponding formally to a Directed Acyclic Graph (DAG), (Lauritzen and Spiegelhalter, 1988; Spiegelhalter *et al.*, 1996) although we

omit the details here. An important property of such graphical models is that conditional independence assumptions that hold between variables can easily be deduced from their assumed structure. This greatly facilitates model building since a complex joint model can be broken down into a series of simpler sub-models that are conditionally independent given any shared nodes. See Spiegelhalter (1998), Richardson and Best (2003), Best and Green (2005) for further discussion of this approach.

Our full model is built from two main sub-models focusing respectively on the outcome and the covariates. The first sub-model in Figures 1(a) and 1(b) is the outcome model which is applied to both the survey data (MCS) and administrative data (NBR). This sub-model represents the association between the outcome,  $y_{rik}$ , for subject  $i$ , who lives in area  $r$  and belongs to the data source,  $k$  (where we use  $k = 1$  as the index for the MCS data and  $k = 2$  as the index for the NBR data), and the exposure,  $E_{rik}$ , after adjusting for important covariates,  $\mathbf{c}_{rik}$  and  $\mathbf{x}_{rik}$  (where  $\mathbf{c}_{rik}$  represents a partially observed covariate vector and  $\mathbf{x}_{rik}$  represents a fully observed covariate vector). In our example, we classify the birthweight outcome into three categories: 1, 2, or 3 if birthweight is normal, pre-term low birthweight or full-term low birthweight, respectively. In the MCS data, all categories for birth outcomes,  $y_{ri1} = 1, 2, 3$ , and covariates,  $\mathbf{c}_{ri1}$ , were fully observed. In the NBR, categories 2 and 3 of the birth outcome,  $y_{ri2} = 2, 3$ , are missing as well as the covariates,  $\mathbf{c}_{ri2}$ . In Figure 1, we distinguish between observed and missing quantities by shading nodes with unobserved values. Note that the model parameters are also shaded since these are unknown quantities to be estimated. In order to impute the missing outcomes in the NBR, we need

to modify the functional form of the outcome model used for the observed outcomes to account for the fact that only categories 2 and 3 of the birth outcomes are missing (see section 3.2.1 for details). The basic structure of the graph remains the same for both cases however.

The second sub-model in Figure 1(c) is the missing covariate model. Due to the fact that the model using only the survey (MCS) data lacks statistical power, we still need information from other sources to help us impute the missing covariates in the administrative (NBR) data. Aggregate data such as census output provides readily available population-based area-level information that can be used for this purpose. In order to use the area-specific data to impute the missing covariates values for subjects in the NBR, we require that: (1) both the survey (MCS) and administrative (NBR) data contain geographical identifiers such as postcodes or census output area codes that enable records to be linked to the aggregate areas; (2) the individual-level covariates we wish to impute tend to cluster geographically, such that aggregate level characteristics are likely to be predictive of the individual-level variables. This is reasonable in the present application, where the missing covariates of interest, maternal smoking and ethnicity, both exhibit strong geographical clustering at small-area level. The missing covariate sub-model thus uses information about the relationship between the vector of area variables,  $\mathbf{A}_r$ , in the aggregate data and individual covariates,  $\mathbf{c}_{ri1}$ , in the survey (MCS) data to help impute the missing individual covariates,  $\mathbf{c}_{ri2}$ , in area  $r$  for the administrative (NBR) data.

The two sub-models can be linked to form a unified global model by conditioning on any common variables — in this case the covariates  $\mathbf{c}_{rik}$ .

There are several nice features about this unified model. First, estimation of parameters, including imputation of missing covariates and outcomes, is done simultaneously within one model, thus allowing uncertainty in all parts of the model to be correctly propagated to the main parameters of interest. Second, the model incorporates data from multiple sources in a coherent way. Third, one can explain the somewhat complex unified model as a combination of relatively easy to understand sub-models. In principle, it is also straightforward to link further sub-models in the same way. For example, if the aggregate data,  $\mathbf{A}_r$ , used in the covariate sub-model, were only available for a sub-set of areas, an additional small-area estimation sub-model (see, for example, Rao (2003), for a review of small area estimation methods) could be added to impute  $\mathbf{A}_r$  in areas with missing values. This sub-model would typically require additional geographically-indexed individual and/or aggregate data sources containing variables predictive of  $\mathbf{A}_r$ , but could also make use of the sample survey data,  $\mathbf{c}_{ri1}$ , already in the model.

One consequence of the unified model is that information flows *both ways* between sub-models. In particular, information about the estimated association between the covariates and the outcome (from the outcome sub-model) is used in addition to the aggregate data (in the covariate sub-model) to help impute the missing covariates. This is sometimes termed ‘feedback’ in Bayesian full probability models (see section “use of cut function” in Spiegelhalter *et al.* (2003) and Lunn *et al.* (2008)). A parallel can be drawn with the standard multiple imputation approach for missing covariate data, where the response variable is sometimes included as a predictor in the imputation model (Little and Rubin, 1987).

### 3.2 Mathematical presentation of the model as applied to the low birthweight study

In addition to the graphical presentation, each sub-model can also be expressed mathematically. Recall that the outcome,  $y_{rik}$  is a categorical variable with 3 categories (see section 3.1). The partially observed covariates of interest are binary indicators of maternal smoking during pregnancy,  $c_{1,rik}$ , and non-white ethnicity,  $c_{2,rik}$ . We also consider two further fully observed covariates, maternal age,  $x_{1,rik}$ , and baby's sex,  $x_{2,rik}$ , which are established predictors of low birthweight and we define  $x_{0,rik} = 1$  to be an intercept term. For notational convenience, we denote the Carstairs index of the ward of residence of each individual as a third fully observed covariate,  $x_{3,rik}$ , although we emphasize that this is actually an area-level rather than an individual-level covariate. The exposure,  $E_{rik}$ , is a binary classification (low/medium =  $\leq 30\mu g/l$  and high =  $> 30\mu g/l$ ) of the total THM concentration during the final trimester of pregnancy in the water zone of the mother's residence at the time of the baby's birth.  $\mathbf{A}_r, r = 1, \dots, R$ , is a vector of three aggregate variables representing, (a) the proportion of the resident population who are of non-white ethnicity, (b) the mean estimated annual income and (c) the mean estimated weekly expenditure on tobacco taken as a proportion of total expenditure on tobacco, wine, beer, fruit, vegetables and saturated fat for each census output area.

### 3.2.1 Mathematical details of outcome sub-model

Our outcome sub-model is a multinomial logistic regression. We specify the same categorical distribution but with different probabilities for the observed and missing outcomes,  $y_{rik}^{obs}$  and  $y_{rik}^{miss}$ , respectively. The reason for needing different probabilities is that the missing outcomes are not random (if they were, the same probabilities could be used for both groups), but are known to be restricted to categories  $y_{ri2} = 2, 3$  in the NBR data. For subjects with observed outcomes (all MCS subjects, and those with normal birthweight in NBR), we specify

$$\begin{cases} y_{rik}^{obs} \sim \text{Categorical}(\mathbf{p}_{u=1:3,rik}), \\ p_{u,rik} = \frac{\zeta_{u,rik}}{\sum_u \zeta_{u,rik}} \\ \log(\zeta_{u,rik}) = \boldsymbol{\alpha}_u^T \mathbf{c}_{rik} + \beta_u E_{rik} + \boldsymbol{\gamma}_u^T \mathbf{x}_{rik} \quad \text{for } u = 1, 2, 3, \end{cases} \quad (1)$$

where the notation  $\mathbf{p}_{u=1:3,rik}$  denotes a vector of probability for categories 1 to 3 of the outcome ( $y_{rik}$ ), for the subject  $i$  who belongs to the data source  $k$  and lives in area  $r$ . The first category of the outcome,  $y_{rik} = 1$ , is treated as the reference group and the associated parameters  $\boldsymbol{\alpha}_1 = [\alpha_{11}, \alpha_{21}]$ ,  $\beta_1$  and  $\boldsymbol{\gamma}_1 = [\gamma_{11}, \gamma_{21}]$  are set to zero for identifiability reasons. We can then interpret  $e^{\beta_2}$  and  $e^{\beta_3}$ , respectively, as the odds of pre-term low birthweight (compared to normal) and full-term low birthweight (compared to normal) associated with living in a water zone with high levels of total THMs relative to the odds associated with living in a water zone with low levels of total THMs. Independent normal priors with mean of zero and a large variance

(variance=1000) were specified for each coefficient  $\alpha_{1u}$ ,  $\alpha_{2u}$ ,  $\beta_u$ ,  $\gamma_{1u}$ ,  $\gamma_{2u}$  ( $u = 2, 3$ ).

For subjects in NBR with missing outcomes, we need to impose the condition that the true outcome is either category 2 or 3 (*i.e.*  $p_{1,rik} = 0$ ) and renormalise the remaining probabilities, which is done as follows

$$\left\{ \begin{array}{l} y_{rik}^{miss} \sim \text{Categorical}(\mathbf{p}_{u=1:3,rik}^*), \\ p_{u=1,rik}^* = 0, \\ p_{u,rik}^* = \frac{p_{u,rik}}{\sum_{u=2}^3 p_{u,rik}} \end{array} \right. \quad \text{for } u = 2, 3, \quad (2)$$

where  $p_{u,rik}$  is given in (1).

### 3.2.2 Mathematical details of missing covariate sub-model

In order to impute the missing binary indicators of maternal race ( $c_{1,rik} = 0$  for white and 1 for nonwhite) and maternal smoking during pregnancy ( $c_{2,rik} = 0$  for non-smoker and 1 for smoker) in a manner that properly accounts for their within-person correlation, we use the multivariate probit model (Chib and Greenberg, 1998) and our covariate sub-model has the following form:

$$\left\{ \begin{array}{l} c_{q,rik} = I(c_{q,rik}^* > 0) \quad \text{for } q = 1, 2, \\ \mathbf{c}_{rik}^* \sim \text{Multivariate Normal}(\boldsymbol{\mu}_{rik}, \Omega), \\ \mu_{q,rik} = \delta_{0,q,s[r]} + \boldsymbol{\delta}_{1,q}^T \mathbf{A}_r + \delta_{2,q} E_{rik} \\ \Omega = \begin{bmatrix} 1 & b \\ b & 1 \end{bmatrix} \quad -1 \leq b \leq 1. \end{array} \right. \quad (3)$$

$I()$  is the indicator function which indicates  $c_{q,rik}$  has value of 1 when  $\mathbf{c}_{q,rik}^*$  is larger than zero. The basic idea behind our implementation is to link the binary variables to underlying continuous normal variables,  $\mathbf{c}_{rik}^* = [c_{1,rik}^*, c_{2,rik}^*]$ , and then model the continuous variables jointly by assuming that they arose from a bivariate normal distribution with mean  $\boldsymbol{\mu}_{rik} = [\mu_{1,rik}, \mu_{2,rik}]$  and common variance-covariance matrix,  $\Omega$ . We model the means as a function of the aggregate variables  $\mathbf{A}_r$  and exposure variable,  $E_{rik}$ , and also include strata-specific intercepts  $\delta_{0,s[r]}$ , where  $s[r]$  denotes which of the three MCS sampling strata ('advantaged', 'disadvantaged' or 'high ethnic minority') area  $r$  belongs to. Without these stratum-specific baselines, the estimated association between the aggregate variables and the individual covariates would be biased due to the non-random sampling mechanism used in the MCS (Brick and Kalton, 1996). Independent normal priors with mean of zero and a large variance (variance=1000) were specified for the various  $\delta$  parameters. For identifiability reasons, we fix the diagonal elements of the variance-covariance matrix to one, and leave the off-diagonal element,  $b$ , to be estimated. A uniform (-1, 1) prior is specified for  $b$  since it is necessary to restrict this parameter to be between -1 and 1 to ensure the entire covariance matrix is positive definite. Note that the model has been implemented in the freely available Bayesian modeling software WinBUGS (Spiegelhalter *et al.*, 2003) and is available in the software section of the BIAS web page ([www.bias-project.org.uk](http://www.bias-project.org.uk)).

### 3.2.3 Multilevel models

Our data are multilevel in nature, in the sense that we have individual-level outcomes and covariates, and these are also linked to various area-level variables, defined at census output area (aggregate covariates), electoral ward (MCS sampling strata and Carstairs index) and water zone (TTHM exposure estimates). In principle, this multilevel structure could be accounted for in our models, by including various area-level random effects in either or both of the outcome and covariate sub-models. The role of such random effects is to account for residual clustering or correlation in the outcomes that is not explained by the covariates in the model. For example, we know that rates of low birthweight do exhibit small area variation due to socioeconomic differences between areas, although we would expect the inclusion of individual smoking, ethnicity, maternal age and the Carstairs deprivation index in our outcome model to largely account for this variability. Likewise, in the covariate sub-model, there may be clustering of smoking and ethnicity at output area or ward level due to socioeconomic factors. We note that inclusion of output area measures of ethnicity, income and tobacco expenditure, and the ward level MCS sampling strata (which reflect ethnicity and deprivation at ward level) is likely to explain most of this variation. In section 5, we report the results of sensitivity analyses where we include ward level random effects in the covariate sub-model (3), and in the outcome model (1)-(2), and examine whether there is strong evidence of residual area-level clustering in either of these models.

## 4 Simulation Study

In our Bayesian graphical model, missing values are imputed in two ways. First, missing covariates,  $\mathbf{c}$ , are generated from aggregate information,  $\mathbf{A}$  (covariate sub-model). Second, missing outcomes,  $y$ , are imputed from observed and missing covariates,  $\mathbf{c}$  (outcome sub-model). In the unified model, there is also feedback between the two imputation sub-models, such that the observed and missing outcomes also influence the imputation of the missing covariates (see section 3.1). We examined the performance of both imputation processes separately and then simultaneously via the use of a simulation study. In particular, our goal was to explore the importance of the relative strengths of association between variables in different parts of the model, and the influence of the feedback mechanism, and how these impact on the overall accuracy of the imputation processes.

### 4.1 Simulation study design

Each of our synthetic datasets was designed to include the following variables: aggregate data representing tobacco expenditure and proportion of non-whites in each area; two binary individual level covariates representing maternal smoking and non-white ethnicity; and a 3-category outcome indicator representing birthweight. Note that, for simplicity, the exposure and fully observed covariates were not included in the simulation set-up since the main focus of the simulation study was on the variables with missing values and how well these were imputed by our model. We anticipate that exclusion of additional fully observed variables in the simulation set-up cre-

ates a more stringent test of the imputation models since there is then less information available to help estimate the missing values. Also, in order to avoid long computational time, the total sample size for each synthetic dataset was chosen to be much smaller than the combined NBR and MCS data sets, at  $K = 1333$ . (Note that this is actually the size of the MCS sample in the real data, but here it represents the full population under study). This sample size was found to be sufficient for demonstrating the utility of the proposed methods under different scenarios, and using a larger sample size is unlikely to materially alter the conclusions.

We then considered four different scenarios based on varying the strength of association between the variable in each of the covariate ( $C$ ) and outcome ( $Y$ ) sub-models. These scenarios are summarised in Table 1 and described in more detail as follows:

- Strong  $Y$ -strong  $C$ : The first scenario was based entirely on real data, namely the real aggregate census and CACI data,  $A$ , on tobacco expenditure and proportion of non-whites in each area, and the real MCS data on maternal smoking and ethnicity,  $C$ , and baby's birthweight,  $Y$ . The coefficients and odds ratio reported for the strong  $A \rightarrow C$  and strong  $C \rightarrow Y$  associations in the lower part of Table 1 are the point estimates obtained fitting multivariate probit ( $A \rightarrow C$ , see equation 3) and multinomial ( $C \rightarrow Y$ , see equation 1) regression models to these real data.

The remaining three scenarios were based on a combination of real and simulated data, depending on which association we wished to weaken. The

real aggregate data,  $A$ , was used for all scenarios.

- Weak  $Y$ -strong  $C$  scenario: Here we used the real MCS covariate data  $C$ , and then generated a new categorical birthweight indicator  $Y$  from the multinomial regression model (1) with ‘weak’  $C \rightarrow Y$  odds ratio given in the lower part of Table 1.
- Strong  $Y$ -weak  $C$  and Weak  $Y$ -weak  $C$  scenarios: For the two ‘weak  $C$ ’ scenarios, it was necessary to simulate covariate data as well as outcome data. This was done by first generating covariates using model (3) conditional on the real aggregate data and the ‘weak’  $A \rightarrow C$  coefficients reported in Table 1. Categorical outcome  $Y$  were then generated from model (1) using the simulated covariates  $C$  and either the strong  $C \rightarrow Y$  or weak  $C \rightarrow Y$  odds ratios as appropriate. Note that all the ‘weak’ log odds ratios and coefficients were chosen to be approximately half the corresponding strong log odds ratios or coefficients.

We will refer to the four complete datasets described above as the *index datasets* for each of the four scenarios. For each index dataset, we then generated twenty replicate datasets containing missing values. For each replicate dataset, we randomly assigned outcomes to be missing with probability 0.1 for those subjects whose outcomes were in category 2 or 3. This percentage is lower than for the real data, but due to the smaller sample size used for the simulation study, it proved difficult to generate a higher percentage of missing outcomes without ending up with empty cells in the cross-classification of observed outcome and covariate categories. For those subjects who have missing outcomes, we assigned missing values to their

smoking and race covariates as well. We randomly chose extra subjects to have missing covariate values, so that the total percentage of individuals with missing covariate values was 80% (similar to the real data), subject to the restriction that there was at least one observed value of each covariate combination in every dataset to ensure stable estimates. The percentage of missing values was the same as the actual percentage of missing covariate information in the combination of the real MCS and NBR data.

## 4.2 Analysis of the synthetic data

Each of the  $4 \times 20$  partially observed synthetic datasets was analysed in four different ways:

1. Covariate sub-model: The outcome variable was ignored and just the missing covariate sub-model (3) was fitted. Note that the ‘Weak Y - Weak C’ and ‘Weak Y - Strong C’ datasets were not analysed using this model, since the relevant portion of the data (aggregate data and covariates) is the same as for the ‘Strong Y - Weak C’ and ‘Strong Y - Strong C’ scenarios respectively. For each subject, we calculated the predicted distribution of  $\mathbf{c}_{rik}^*$  based on the fitted values of  $\boldsymbol{\mu}_{rik}$  (see equation (3)) and then calculated the proportion of this  $\mathbf{c}_{rik}^*$  distribution falling above or below zero in each dimension to obtain predicted probabilities of each of the four possible covariate patterns (smoking=0 and non-white=0, smoking=0 and non-white=1, smoking=1 and non-white=0, smoking=1 and non-white=1).
2. Outcome sub-model: The covariate values were fixed at their known

values in the index dataset for each scenario, with missing values only in the outcome, and just the outcome sub-model, defined by (1) and (2), was fitted. For each dataset, we obtained coefficients (log odds ratios) for pre- and full-term low birthweight (compared with normal birthweight) associated with maternal smoking and with non-white ethnicity.

3. Unified model: The unified model, (1), (2) and (3), was fitted to the entire dataset. Predicted probabilities for each covariate pattern, and estimated log odds ratios in the outcome model were calculated as described in points 1 and 2 above.
4. Unified model but cutting the feedback between outcome and covariate sub-models: This allows the imputed covariates to be used in the outcome model, but ignores the information from the outcome when imputing the covariates (See section “use of the cut function” in Spiegelhalter *et al.* (2003) and Lunn *et al.* (2008)). By comparing the results of the unified model (above) with this cut model, we can explore the relative influence of the two different sources of information (i.e. the relationship between the aggregate data and covariates, and the relationship between the covariates and outcome) on the resulting imputations and parameter estimates.

For comparison purposes, we also used the covariate and outcome sub-models to analyse each of the four complete (index) datasets. It was not necessary to analyse the fully observed data using the unified model since there were no missing values to impute, so the two sub-models become inde-

pendent conditional on the observed data. Both fully and partially observed data were analyzed using WinBUGS (Spiegelhalter *et al.*, 2003). We specified the same priors for each model as described in section 3.2.1 and 3.2.2, and all the following simulation results were based on 10000 iteration burn-in period and another 10000 samples for inference.

### 4.3 Results of the simulation study

#### 4.3.1 Examining the imputation of missing covariates

##### *Missing data at one level (Covariate sub-model)*

The ability of the covariate sub-model to impute the missing covariates is illustrated in Figure 2. This shows box plots of the distribution of the average subject-specific predicted probabilities of each covariate pattern obtained from fitting the covariate sub-model to the partially observed covariate data (shaded boxes) versus the predictions from fitting the same model to the fully observed covariate data (white boxes). Note that the predicted probabilities for plotting shaded boxes are the average values for each subject across the twenty replicate partially observed datasets whereas the predicted probabilities for plotting white boxes are from the analysis of the one fully observed index dataset. For each covariate pattern, the box plots are split according to whether or not the subjects' true covariate values correspond to that pattern. When the true association between the aggregate predictors and the covariates was strong (top panel), we see that the two sets of predictions (partial and complete data) are almost identical, and that they both discriminate the true covariate patterns quite well

(i.e. assign higher probabilities of each covariate pattern to subjects whose true covariates correspond to that pattern than to those whose true pattern is different). This indicates that, even in the presence of 80% missing values, our Bayesian missing covariate model produces quite accurate imputations and yields very similar inferences to an equivalent model with fully observed data. When the true association between the aggregate predictors and the covariates was weak (bottom panel), not surprisingly, the ability of both the complete data model and the imputation model to discriminate the true covariate patterns deteriorates. Nevertheless, the agreement between the predictions from the fully and partially observed data is still quite high, indicating that the missing covariate sub-model is performing as well as can be expected in the presence of a weak underlying signal. Note that the predictions for the weak  $C$  and strong  $C$  scenarios are not directly comparable since the underlying true distribution of the covariate patterns is different for the two scenarios.

*Missing data at two levels (Unified model)*

We now consider the situation in which imputation is required for both missing covariates and outcomes, and its impact on the covariate predictions. Figure 3 compares box plots of the distribution of predictions of the covariate patterns from the fully and partially observed data, as in Figure 2, but this time the predictions are obtained from the unified model (1) - (3). Also, for clarity, we just focus on predictions for one of the covariate patterns (white non-smokers,  $C = 0, 0$ ); results for the other three covariate patterns are similar. There are now four scenarios to consider: either a strong (top two panels) or weak (bottom two panels) association between the aggregate

data and covariates, combined with either a strong (first and third panels) or weak (second and fourth panels) association between the covariates and the outcome. This time, the labels at the bottom of each graph distinguish between the true values of the outcome *and* the true covariate pattern for each subject. Comparing the pairs of white and shaded distributions for each combination of True  $C$  and  $Y$ , we see that the covariate predictions in the face of missing data are now different from those obtained with full data, particularly for those with outcomes  $Y = 2$  and  $3$ . This is because, when fitting the unified model to the partial data, the covariate predictions are influenced by feedback from the outcome model as well as by their estimated association with the aggregate data. For the fully observed data, it would be necessary to explicitly include the outcome as an additional predictor in the covariate sub-model in order to allow a similar feedback. In general, we see that the feedback from the outcomes in the partial data models is beneficial in that, relative to the full data model, the predicted probabilities of the covariate pattern  $C = 0, 0$  are better able to discriminate between subjects whose true covariates are  $C = 0, 0$  or not (compare the right half of each panel with the corresponding distributions in the left half). This feedback is particularly beneficial in situations where the aggregate data are only weak predictors of the the covariates (weak  $C$  scenarios).

### **4.3.2 Examining impact of the imputation model on the covariate-outcome association**

To examine the impact of the two imputation processes on the estimated covariate-outcome association, we first obtained “reference” estimates of the

log odds ratio ( $\hat{\alpha}_{q,u}$ ;  $u = 2, 3$ ) by fitting the outcome model to the fully observed index dataset for each of the four scenarios. For each scenario, we then analysed the 20 replicate partial datasets using the outcome sub-model fitted to data with missing outcomes only, and the unified model, with and without feedback from outcomes to covariate sub-models, fitted to data with missing outcomes and covariates. In each case we calculated the bias and mean square error (MSE) between the posterior mean estimates of the log odds ratios  $\bar{\alpha}_{q,u}$  and the corresponding reference estimate  $\hat{\alpha}_{q,u}$ . Table 2 summarises the average biases and the MSEs for each scenario and analysis model. For clarity, we focus only on the results for full-term low birthweight versus normal birthweight, since the relationship between the covariates and this outcome is much stronger in the real MCS than for pre-term low birthweight.

When outcomes, but not covariates, are missing, the bias and MSE are small for all four scenarios (‘outcome model’ columns of Table 2). This suggests that the outcome imputation model performs almost as well as the reference model with complete data. When covariates are also missing, the biases and MSEs increase considerably (‘unified model’ columns of Table 2), indicating that uncertainty about the missing covariates has a bigger impact on the outcome model parameter estimates than does uncertainty about the missing outcomes. This is not surprising, given that a much higher percentage of the covariate data is missing compared to the outcome data. Interestingly, the impact of covariate uncertainty on the outcome model is much greater in the two ‘Weak Y’ scenarios (weak true association between outcome and covariates). Some light can be shed on this by examining the

results of cutting feedback between the outcome and covariate sub-models in the unified analysis (final columns of Table 2). While bias and MSE increase when feedback is cut in the two ‘Strong Y’ scenarios, they tend to decrease in the ‘Weak Y’ scenarios. The posterior uncertainty about the outcome model parameter estimates is also consistently lower by about 30-50% for all scenarios when feedback is cut (for example, the posterior standard deviations of  $\alpha_{23}$  with and without feedback, respectively, are 0.58 versus 0.43 for the ‘Strong Y-Strong C’ scenario, and 0.85 versus 0.67 for the ‘Weak Y-Strong C’ scenario). The impact of feedback on the posterior uncertainty probably reflects the fact that feedback propagates uncertainty about the imputed outcomes through to the covariate imputations, which in turn will be reflected in wider interval estimates for the outcome model regression coefficients. Cutting feedback ignores the outcome information, and hence also the outcome uncertainty, in the covariate imputations. In the ‘Strong Y’ scenarios, where the outcome and covariates contain strong information about each other, feedback still tends to improve parameter estimates in the outcome model in terms of reducing bias and MSE, despite the additional uncertainty. However, when the information about the covariates being propagated by feedback is weak (‘Weak Y’ scenarios), the additional uncertainty appears to mainly add noise, resulting in higher bias and MSE in the outcome model parameters in most cases.

## 5 Results of application to the low birthweight study

In this section we detail results from an analysis of low birth-weight babies in the UK. We focus on subjects from the 98 wards in northern England which were covered by a single water company and were also sampled in the MCS. This gives a total of 9278 singleton births, of whom 1333 appeared in the MCS with full data, and the remaining 7945 births appeared only in the NBR. Note that births in the MCS were matched to records in the NBR using postcode, sex and date of birth so that they could be excluded from the latter to avoid duplication. Four of the 7945 NBR births also appeared in the MCS, but had missing values in either the outcome and/or covariates of interest, and so were treated as NBR births.

Table 3 summarises the distributions of the main variables of interest in the two datasets. Overall, approximately 7% of the subjects had low birth-weight that could not be classified as either pre- or full-term, and nearly 86% had missing covariate information on maternal smoking and ethnicity. The main exposure variable of interest was total THMs (TTHMs) which, following Toledano *et al.* (2005), was classified into low ( $\leq 30\mu\text{g}/\text{l}$ ), medium ( $30 - 60\mu\text{g}/\text{l}$ ), and high ( $> 60\mu\text{g}/\text{l}$ ) levels. From Table 3, we observed that only 8% of individuals were exposed to low TTHM concentrations. Therefore, we combined both low and medium categories together in our analyses to ensure there was enough information to obtain stable estimates of the exposure effect. Similarly, we also dichotomised the Carstairs deprivation index with the first three categories placed in the low-medium deprivation

group and categories 4 and 5 placed in the high deprivation group. Comparing the distribution of fully observed variables between the MCS and NBR datasets, we see that they are broadly comparable with the exception of the MCS sampling strata and Carstairs deprivation index. Somewhat surprisingly, there were higher proportions of babies in the ‘disadvantaged’ stratum and high Carstairs category in the NBR than the MCS (given the intention of the MCS to over-sample disadvantaged groups). However, the sampling strata and Carstairs index are adjusted for in all our analyses so this imbalance should not distort our overall findings. Table 3 also summarises the distributions of the aggregate variables across the 2349 census output areas that nest within the 98 wards comprising the study region.

The data were analyzed using the following models:

Model A Multinomial logistic regression model (1) applied only to the 1333 subjects in the MCS with fully observed data.

Model B Bayesian unified model (1)–(3) applied to the combined MCS, NBR and aggregate data, with imputation for both missing outcomes and covariates.

The outcome model in each of the above analyses included the dichotomised TTHM exposure variable, with adjustment for maternal smoking, maternal ethnicity, maternal age, baby’s gender and dichotomised Carstairs deprivation index. Also, all models included a strata-specific intercept to adjust for the sample selection mechanism of the MCS.

We also considered a number of sensitivity analyses:

Model C As for Model B, but cutting feedback from the outcome to the covariate sub-model.

Model D As for Model B, but including ward-level random effects in the covariate model (3).

Model E As for Model B, but including ward-level random effects in the outcome model (1)–(2).

Model F As for Model A, but excluding maternal smoking and ethnicity from the outcome model (1).

The WinBUGS program was used to fit all six models, with a burn-in of 20,000 iterations followed by 30,000 iterations saved for analysis purposes. Table 4 summarises the estimated odds ratios for various covariates of interest under each of the models. For clarity, we focus attention on the full-term low birthweight versus normal birthweight outcome; results for risk of pre-term low birthweight versus normal birthweight did not show any strong or statistically significant relationships with any of the covariates of interest under any of the models, and so are not shown. Considering first the results for the TTHM exposure effect of primary interest, all six models show evidence of a positive association between living in an area with high TTHM levels and risk of full-term low birthweight. However, the magnitude and statistical significance of the estimated odds ratio varies across models. The combined data (Model B) yields a large, statistically significant effect (OR: 2.10, 95% credible interval: (1.03, 3.93)); however, this association became modest and non-significant when we only used the MCS data which

contained fully observed information (Model A). One explanation for the different estimates is that the MCS data alone lack power to detect an effect of the TTHM exposure, whereas combining the MCS and NBR data leads to a more reliable estimate based on a much larger sample size. On the other hand, it is possible that the imputation of the maternal smoking and ethnicity covariates in Model B did not fully succeed in adjusting for confounding and so the Model B TTHM odds ratio estimate may still be biased. Comparing the TTHM odds ratios for Models A and F (MCS only, with and without adjustment for maternal smoking and ethnicity), it is clear that maternal smoking and ethnicity are important confounders of the TTHM-birthweight association, and that failure to adjust for them tends to over-estimate the risk associated with high TTHM exposure. However, as discussed below, the effects of maternal smoking and ethnicity appear to be well estimated in the combined data Model B, and so the TTHM effect estimated by this model should be appropriately adjusted for confounding. The TTHM exposure effect estimated from the combined data was also robust to whether or not feedback from the outcome to the covariate imputation model was allowed (comparing Models B (feedback allowed) and C (feedback cut)). The TTHM exposure effect was more sensitive to the inclusion of ward random effects in the outcome model (Model E), which resulted in a higher odds ratio (2.55) and a somewhat wider credible interval. This suggests that there may be residual heterogeneity in full-term low birthweight at ward level, although part of the increase in uncertainty in the TTHM odds ratio may reflect the fact that, due to the missing values in the NBR, there is quite limited information for estimating area-level random effects in

the data. By comparison, including ward-level random effects in the covariate model had minimal impact on the TTHM odds ratio estimates (Model D).

Turning to the results for maternal smoking and non-white ethnicity, we see that the odds ratios are slightly lower under Model B (combined data) than Model A (fully observed MCS data only), although both sets of results are consistent with the literature indicating strong positive associations between these two variables and the risk of full-term low birthweight. Although the interval estimates are also slightly narrower in Model B, it may seem surprising that we don't gain more precision here from combining the data sources. However, recall that these covariates are completely missing from the NBR, so the additional information in the combined model to help estimate the effects of these variables on the risk of full-term low birthweight is only from indirect sources — namely the aggregate data and TTHM exposure, plus feedback from the birthweight outcome, and so there is considerable uncertainty about the imputed covariates. Cutting the feedback from the outcome to the covariate model in the combined data model (Model C) results in slightly lower odds ratios for both covariates, and a small gain in precision, as expected from the simulation study. However, given that the covariate-outcome relationship is clearly strong in this case, the findings from the simulation study would suggest that the Model B results, which include the feedback mechanism, are likely to be more appropriate. The smoking and non-white odds ratios are fairly robust to the inclusion of ward random effects in either the covariate model (Model D) or outcome model (Model E), with just a small increase in uncertainty.

## 6 Discussion

In this paper, we have developed a Bayesian graphical modelling framework which allows one to formally quantify the impact of different sources of bias — such as unmeasured confounding and missing outcomes — involved in analyzing observational data. Multiple data sources are needed to help identify the model. Simulation studies show that our imputation models performed well when the explanatory variables were highly correlated with the response variables in each sub-model of the graph. Applying our model to an epidemiological study of water disinfection byproducts and the risk of low birthweight yielded improved estimates of the exposure effect of interest, and demonstrated the benefit of combining information from multiple sources as opposed to only relying on one data source. On the other hand, our simulation studies also showed that if the available information in different data sources are only weakly related, the combination of multiple datasets might not provide much of an advantage over using a single data source.

We also demonstrated that missing values can be imputed using our modelling framework in a way that uses all available information contained in the different parts or sub-components of the model. In particular, Bayesian full probability models allow feedback from the outcome model to help inform imputation of the missing covariates. This is difficult to achieve using competing approaches such as multiple imputation. In particular, when both the covariates and outcome of interest contain missing values, it is not possible to set up a standard two-stage multiple imputation model (i.e. where

the missing data are first imputed, and then the model of interest is fitted to the completed datasets) that consistently imputes the covariates taking account of the outcome. Whilst the feedback in a Bayesian full probability model is generally desirable, our simulation studies indicated that there is a complex interplay between the information and uncertainty that is propagated through our unified model. In situations where the true association between the covariates and outcome is weak, the feedback mechanism can result in added noise in the outcome model parameter estimates. In practice, we may not know whether the true association between the outcome and covariates is strong or weak, and so examining the sensitivity of inference to the presence or absence of feedback can be helpful.

When imputing missing values, it is important to consider the mechanism leading to the missing data and model this appropriately. In our study, the missingness process is known by design for both the missing covariates and missing outcomes, so provided we condition on the relevant design factors in our model, we can assume the data are missing at random (MAR; Little and Rubin (1987)). The smoking and ethnicity covariates are missing for all subjects not sampled in the MCS, and so we included the strata variables used to select the MCS subjects as predictors in the covariate imputation model. There may well be some additional missing values due to non-response in the MCS, but these will constitute only a small fraction of the total missing data in our study, and conditional on the strata and aggregate variables used in the imputation model, it is reasonable to also view these as MAR. As noted in Section 3.2.1, the missing outcome mechanism is not random, in the sense that if the outcome is missing, then by

design it must be in category 2 or 3 but not 1. We implicitly condition on this known mechanism by specifying a separate model (2) for the missing outcomes which restricts them to category 2 or 3, and then treat the missing outcomes as MAR by assuming the same model parameters as for the observed category 2 and 3 outcomes in (1).

One disadvantage of the Bayesian graphical modelling approach is that it requires a large amount of computer time to estimate all the parameters of the model, making analysis of very large data sets problematic. The analysis of the birthweight data reported here took approximately 30 hours to run on a 1.60 GHz (3.25 GB of RAM) PC. Two stage approximations to the full Bayesian model, such as the approach used by Jackson *et al.* (2008), may be necessary to handle very large data sets. However, it is worth noting that computing time is cheap relative to the time and resources involved in collecting the various datasets utilized in this analysis.

In order to keep the analysis manageable for illustration purposes, we restricted it to the subset of wards sampled in the MCS in a single water company region. However, these data are part of a larger epidemiological study being carried out by SAHSU covering fourteen water company regions in England and Wales. We are currently applying our unified Bayesian model to analyse data for each water company separately; risk estimates will then be combined across water companies using meta-analysis methods in a similar way to our previous study (Toledano *et al.*, 2005).

The modelling framework used in this paper is closely related to that used by Jackson *et al.* (2008) to combine administrative and survey data, but with two main differences. Firstly, Jackson *et al.* (2008) only consid-

ered imputation of missing covariates in the administrative data, and not missing outcomes. Secondly, they used a different model to account for the correlation between the covariates. We initially investigated using the same approach for the present paper, which involved combining the two binary covariates (maternal smoking and ethnicity) into one categorical variable with four levels (1: smoking=0 non-white=0, 2: smoking=1 non-white=0, 3: smoking=0 non-white=1, 4: smoking=1 non-white=1), and regressing this categorical variable on aggregate information using a multinomial logistic model. This approach allowed us to account for the correlation between the two missing covariates, but large uncertainties would often appear in the parameter estimates since some of the categories would contain few or no individuals. We found that the multivariate probit model resulted in more stable parameter estimates, and was also more computationally efficient since it avoids the need to sample from the multinomial distribution.

In principle, it would also be possible to extend the model to include a larger number of unmeasured covariates/confounders. However, this may prove computationally prohibitive, even using the multivariate probit model. In standard regression modelling, propensity score (Pearl, 2000; Rosenbaum and Rubin, 1983) methods are increasingly being used as an efficient way of dealing with a large number of observed confounders. The basic idea behind such methods is to provide a measure of the probability that a person belongs to a treatment or exposure group using only their covariate values, and to then condition on this propensity score rather than the full set of covariates in the regression model. As an alternative to the approach presented here, we have also developed an extension of the propensity score method to handle

large numbers of *unobserved* confounders and incorporated this within a Bayesian framework (McCandless *et al.*, 2008). Applying this method to the present data to adjust for additional unmeasured socio-economic factors such as income, education and BMI yielded similar results. There may still be potential confounders, such as other environmental factors like air pollution, that we have not adjusted for, but there is no particular reason to believe that these would be strongly correlated with TTHMs.

Our Bayesian modelling framework can be extended in a number of ways. For example, if additional survey or cohort studies containing relevant covariate data were available, these could be incorporated as extra data sources,  $k = 3, 4, \dots$  in the covariate imputation model (3). It would also be straightforward to change the outcome model to, say linear regression, and treat birthweight as a continuous measure. This would have the advantage of increasing statistical power over using a categorical outcome measure, but it would be more difficult to adjust for missing gestational age. Other potential sources of bias could also be modelled by adding appropriate sub-models. In particular, a Bayesian misclassification model (Gustafson, 2004) could be added to account for measurement errors in the THM exposure estimates, and to adjust for factors which affect a mother's true intake of THMs (such as time spent showering, amount of bottled water consumed). This sub-model would need to be informed either by a separate validation data set that included both water-zone THM concentrations and personal exposure measurements on a subset of individuals, and/or by using published misclassification probabilities such as those in Whitaker *et al.* (2003). It would also be possible to make use of other relevant published

information, as well as raw data sources, in our modelling framework by constructing informative priors for model parameters such as the log odds ratios in the outcome model (1) and (2). The systematic review of chlorination disinfection byproducts and adverse birth outcomes by Nieuwenhuijsen *et al.* (2000a) represents one such information source, although translating the qualitative conclusions about risks and biases provided in such a review into quantitative priors is not straightforward, and is best approached as a form of sensitivity analysis. In all cases, care is needed to ensure that the data and information sources being combined within our modelling framework are compatible and can be assumed to be representative of the same underlying population.

## **Acknowledgment**

The authors would like to thank the Small Area Health Statistics Unit (SAHSU) at Imperial College and the Economic and Social Data Service (ESDS) for provision of data, and Heather Joshi and Jon Johnson for their assistance in obtaining the special license access to postcoded Millenium Cohort data. We would also like acknowledge the contribution of Mireille Toledano and Mark Nieuwenhuijsen who are responsible for conceiving the epidemiological project on chlorination and birthweight and for their help in discussing the epidemiological aspects of the analysis. We are also grateful to Mireille Toledano, Mark Nieuwenhuijsen, Daniela Fecht, James Bennett, Kees de Hoogh and Peter Hambly for their help in obtaining and processing the data.

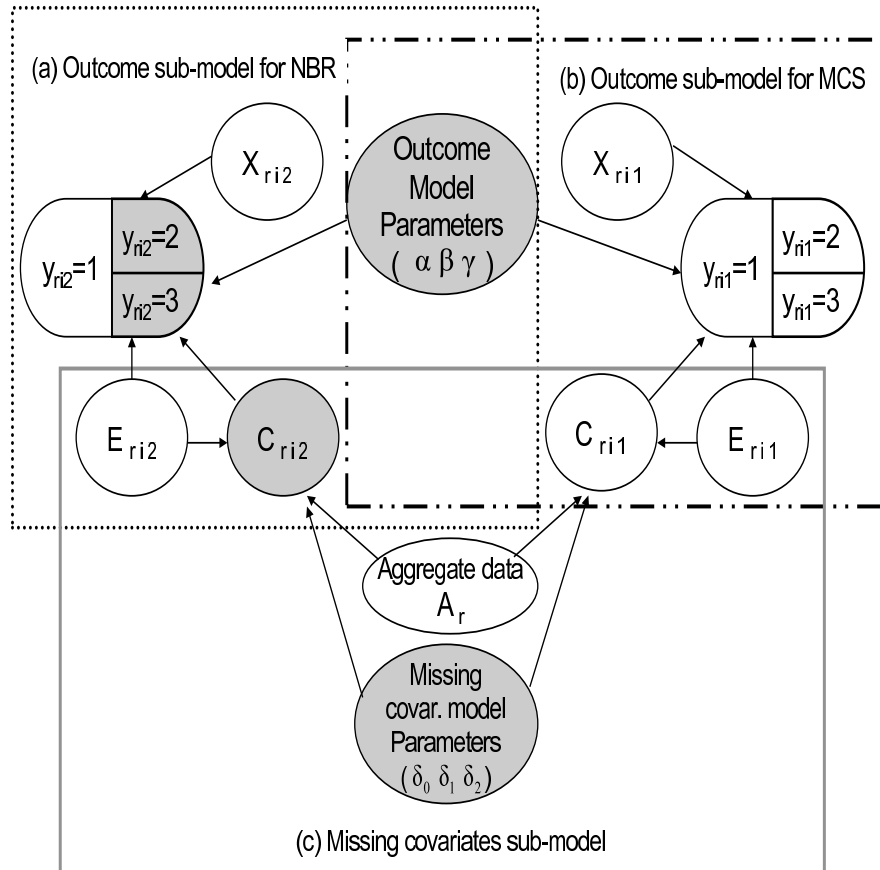


Figure 1: Graphical representation of model. Outcome sub-model represents regression model relating exposure (E), fully observed covariates (X) and partially observed covariates (C) to the 3-category outcome  $y$  (see equations (1)–(2)). Covariate sub-model represents regression model of partially observed covariates (C) on exposure (E) and aggregate information (A) (see equation (3)). Note: Quantities in grey are unobserved

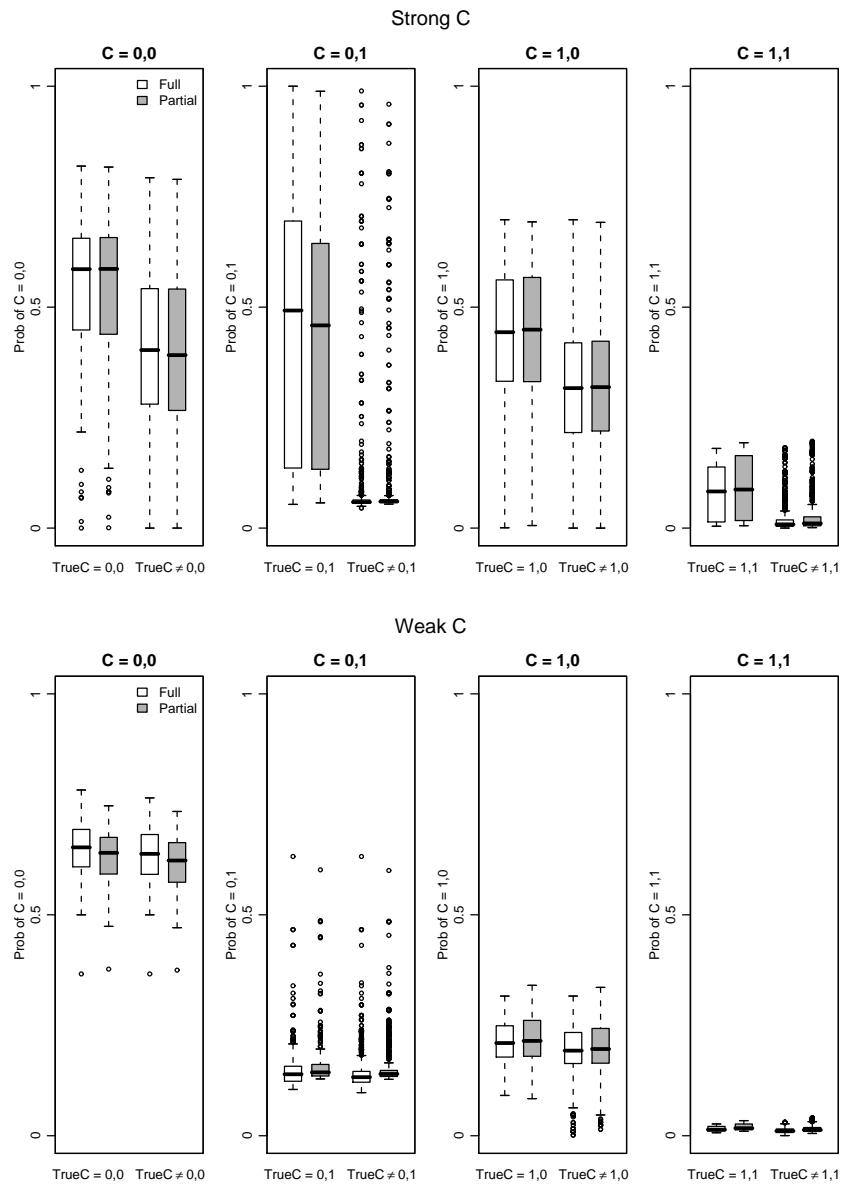


Figure 2: Performance of the covariate sub-model fitted to the fully observed covariate data (white boxes) or partially observed covariate data (shaded boxes) generated under either strong or weak association with the aggregate variables. Box plots show the distribution of average subject specific predicted probabilities for each covariate pattern ( $C$ : (0,0) = non smoker white; (0,1) = non smoker non white; (1,0) = smoker white; (1,1) = smoker non white), split according to the true value of the subjects' covariate pattern.

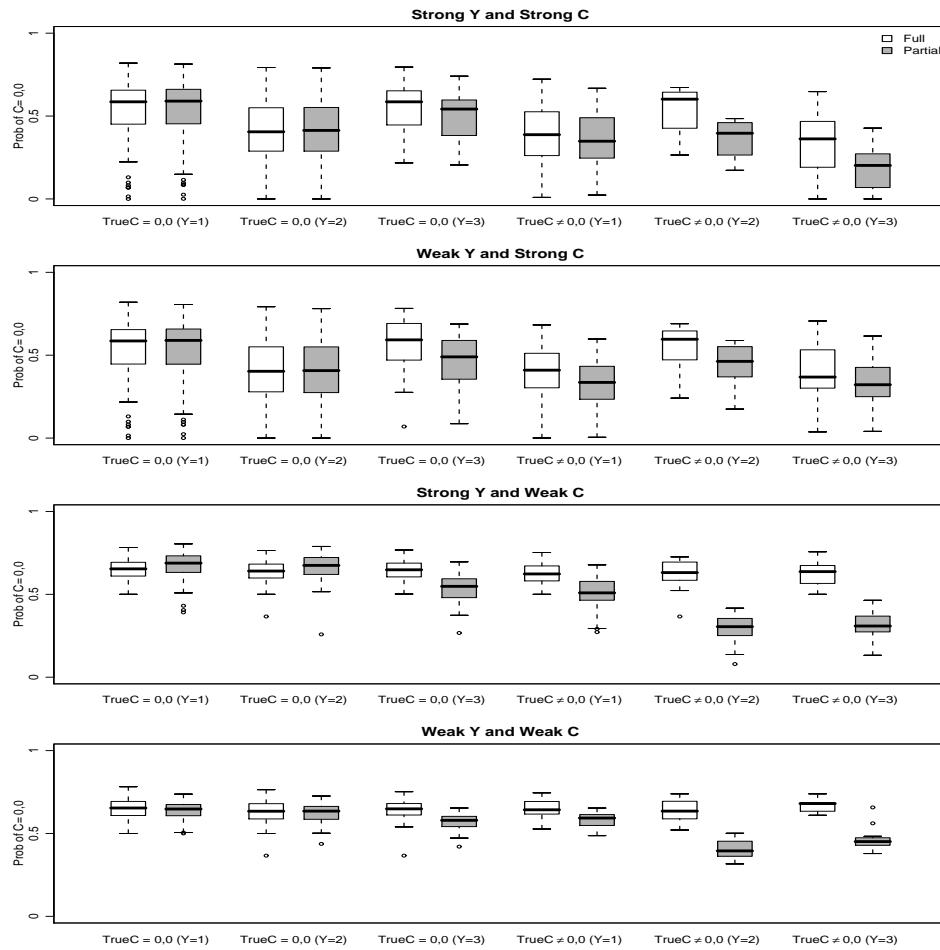


Figure 3: Performance of the unified model fitted to the fully observed data (white boxes) or partially observed covariate and outcome data (shaded boxes) generated under each of the four scenarios. Box plots show the average predicted probabilities of being a white non smoker (covariate pattern  $C=0,0$ ), split according to the true covariate pattern ( $C=0,0$  or not) and outcome status ( $Y=1, 2$  or  $3$ ) of the subjects.

Table 1: Summary of simulation design and parameter values used to generate the data

Data Scenarios	Covariate sub-model				Outcome sub-model			
	A		- - ->		C		- - ->	
<b>Strong Y-Strong C</b>	Real (Census, CACI)	Strong	Real (MCS)	Strong	Real (MCS)	Strong	Real(MCS)	
<b>Weak Y-Strong C</b>	Real (Census, CACI)	Strong	Real (MCS)	Weak	Real (MCS)	Strong	Simulate	
<b>Strong Y-Weak C</b>	Real (Census, CACI)	Weak	Simulate	Strong	Simulate	Weak	Simulate	
<b>Weak Y-Weak C</b>	Real (Census, CACI)	Weak	Simulate	Weak	Simulate	Strong	Simulate	
	<b>Strong <math>A - \rightarrow C^\# (\delta \times \text{IQR}^\dagger)</math></b>				<b>Strong <math>C - \rightarrow Y (\text{OR}^*)</math></b>			
	C				Y			
A	Smoke	Non-white	C	LWBP $^\diamond$	LWBF $^\diamond$			
Ratio of non-white	-0.042	0.125	Smoke	1.083	2.411			
Tobacco expenditure	0.601	0.052	non-white	1.094	5.930			
	<b>Weak <math>A - \rightarrow C^\# (\delta \times \text{IQR}^\dagger)</math></b>				<b>Weak <math>C - \rightarrow Y (\text{OR}^*)</math></b>			
	C				Y			
A	Smoke	Non-white	C	LWBP $^\diamond$	LWBF $^\diamond$			
Ratio of non-white	-0.036	0.062	Smoke	1.041	1.553			
Tobacco expenditure	0.260	0.026	non-white	1.041	2.226			

\*: Odds ratio.

†: Values in the  $A \rightarrow C$  table are equal to  $\delta \times \text{IQR}$ , where  $\delta$ s are coefficients in the equation (3) and  $\text{IQR}$  is the difference between the 75<sup>th</sup> and 25<sup>th</sup> percentiles of distribution of non-white ethnicity and tobacco expenditure. Also, the interquartile range for the ratio between non-white to white and proportion of tobacco expenditure are 0.1 and 0.3, respectively.

‡: Covariates smoke and non-white were generated by using the multivariate probit model with correlation,  $\rho = -0.45$

◊: LWBP: low birthweight pre-term; LWBF: low birthweight full-term.

Table 2: Comparison of average bias and mean square error (MSE) of the log odds ratios,  $\alpha_{qu}$ , in the outcome model (1), obtained by fitting different analysis models to data from the four simulation scenarios

<i>Analysis model</i> <b>Data Scenarios</b> Parameters	<b>Complete Data</b> ( <i>reference analysis*</i> )		<b>Partial Data</b> ( <i>Unified model<sup>†</sup></i> )		<b>Unified model with feedback cut<sup>‡</sup></b>	
	Estimate	Bias (MSE)	Bias (MSE)	Bias (MSE)	Bias (MSE)	Bias (MSE)
<b>Strong Y - Strong C</b>						
smoke ( $\alpha_{13}$ )	0.90	-0.01 (0.01)	-0.17 (0.27)	0.64 (0.43)		
non-white ( $\alpha_{23}$ )	1.79	-0.03 (0.01)	-0.43 (0.25)	0.67 (0.47)		
<b>Strong Y - Weak C</b>						
smoke ( $\alpha_{13}$ )	0.99	0.02 (0.00)	0.02 (0.51)	0.83 (0.71)		
non-white ( $\alpha_{23}$ )	2.56	-0.01 (0.01)	-0.15 (0.49)	1.89 (3.63)		
<b>Weak Y - Strong C</b>						
smoke ( $\alpha_{13}$ )	-0.02	-0.07 (0.01)	-0.59 (1.34)	-0.08(0.07)		
non-white ( $\alpha_{23}$ )	0.32	-0.10 (0.03)	-0.29 (0.41)	0.14 (0.09)		
<b>Weak Y - weak C</b>						
smoke ( $\alpha_{13}$ )	0.35	0.01 (0.03)	-0.56 (0.89)	0.26 (0.11)		
non-white ( $\alpha_{23}$ )	1.00	-0.06 (0.04)	-0.23 (1.32)	0.82 (0.84)		

\* Full data was analyzed by using the multinomial logistic regression model

†:Data with missing in  $Y$ (outcome) was analyzed by using the outcome sub-model

‡:Data with missing in  $Y$ (outcome) and  $C$  (covariates) was analyzed by using the unified model

‡:Data with missing in  $Y$ (outcome) and  $C$  (covariates) was analyzed by using the unified model with cut function

Table 3: Demographic summary of participants in the low birthweight study

	MCS N(%)	NBR N(%)	MCS+NBR N(%)
<b>Birthweight categories</b>			
Normal birthweight	1224 (91.8)	7308 (92)	8532 (92.0)
Low birthweight pre-term	68 (5.1)	–	68 (0.7)
Low birthweight full-term	41 (3.1)	–	41 (0.4)
Low birthweight missing term	–	637 (8)	637(6.9)
<b>Maternal ethnicity</b>			
White	1097 (82.3)	–	1097 (11.8)
Non-white (asian, black, others)	236 (17.7)	–	236 (2.6)
Missing	–	7945 (100)	7945 (85.6)
<b>Maternal smoking</b>			
No	819 (61.4)	–	819 (8.8)
Yes	514 (38.6)	–	514 (5.6)
Missing	–	7945 (100)	7945 (85.6)
<b>Maternal age</b>			
< 20	138 (10.4)	874 (11.0)	1012 (10.9)
20-24	291 (21.8)	1894 (23.8)	2185 (23.6)
25-29	381 (28.6)	2237 (28.2)	2618 (28.2)
30-34	352 (26.4)	1906 (24.0)	2258 (24.3)
>=35	171 (12.8)	1034 (13.0)	1205 (13.0)
<b>Babies' sex</b>			
Females	657 (49.3)	3908 (49.2)	4565 (49.2)
Males	676 (50.7)	4037 (50.8)	4713 (50.8)
<b>Total THMs exposure</b>			
Low (< 31g/l)	156 (11.7)	576 (7.2)	732 (8.0)
Medium (31 – 60g/l)	578 (43.4)	4011 (50.5)	4589 (49.4)
High (> 60g/l)	599 (44.9)	3358 (42.3)	3957 (42.6)
<b>MCS sampling stratum</b>			
Advantaged	442 (33.2)	1623 (20.4)	2065 (22.2)
Disadvantage	707 (53.0)	5190 (65.3)	5897 (63.6)
High ethnic minority	184 (13.8)	1132 (14.3)	1316 (14.2)
<b>Carstairs deprivation index</b>			
Low and Medium (1-3)	565 (42.4)	2627 (33.1)	3192 (34.0)
High (4-5)	768 (57.6)	5318 (66.9)	6086 (66.0)
<b>Aggregate data - census output area</b>			
Mean annual income (pounds)	Median (Interquartile Range)		
	26887 (22356, 33078)		
Ratio of nonwhite to white	0.03 (0.01, 0.08)		
Proportion of expenditure on tobacco (%)	35 (30, 42)		

Table 4: Comparison between models of estimated risk of delivering a low birthweight baby at full-term associated with exposure to TTHMs and maternal smoking and ethnicity

Model <sup>†</sup>	Data (Subjects)	OR (95 % Credible Interval)		
		TTHM > 60 g/l	Smoker	Non-White
Model A	MCS(1333)	1.65 (0.76, 3.20)	2.58 (1.14, 5.12)	5.58 (1.97, 12.55)
Model B	NBR+MCS+Aggregate (9278)	2.10 (1.03, 3.93)	2.38 (1.07, 4.56)	5.36 (2.39, 10.37)
Model C	NBR+MCS+Aggregate (9278)	2.07 (1.01, 3.85)	2.28 (1.08, 4.30)	4.95 (2.26, 9.53)
Model D	NBR+MCS+Aggregate (9278)	2.14 (1.01, 4.09)	2.42 (1.10, 4.71)	5.40 (2.35, 10.76)
Model E	NBR+MCS+Aggregate (9278)	2.55 (1.01, 5.25)	2.57 (1.15, 4.99)	5.02 (2.03, 10.53)
Model F	MCS(1333)	2.06 (1.00, 3.86)	—	—

Model A: Multinomial logistic regression model adjusting for smoker and non-white

Model B: Unified model

Model C: Unified model (cut function)

Model D: Unified model adjusting for ward random effects in the covariate sub-model

Model E: Unified model adjusting for ward random effects in the outcome sub-model

Model F: Multinomial logistic regression model without adjusting for smoker and non-white

<sup>†</sup>: Adjusted for maternal age, baby's sex, and Carstairs deprivation index

## References

- Barros, F. C., Huttly, S. R., Victora, C. G., Kirkwood, B. R. and Vaughan, J. (1992) Comparison of the causes and consequences of prematurity and intrauterine growth retardation: a longitudinal study in southern brazil. *Pediatrics*, **90**, 238–244.
- Best, N. and Green, P. (2005) Structure and uncertainty: Graphical models for understanding complex data. *Significance*, **2**, 177–181.
- Brick, J. M. and Kalton, G. (1996) Handling missing data in survey research. *Statistical Methods in Medical Research*, **5**, 215–238.
- Carstairs, V. and Morris, R. (1991) *Deprivation and health in Scotland*. Aberdeen: Aberdeen University Press.
- Chib, S. and Greenberg, E. (1998) Analysis of multivariate probit models. *Biometrika*, **85**, 347–361.
- Greenland, S. (2003) The impact of prior distributions for uncontrolled confounding and response bias: a case study of the relation of wire codes and magnetic fields to childhood leukaemia. *Journal of the American Statistical Association*, **98**, 47–54.
- (2005) Multiple-bias modelling for analysis of observational data. *J. R. Statist. Soc. A*, **168**, 267–306.
- Gustafson, P. (2004) *Measurement Error and Misclassification in Statistics and Epidemiology: Impacts and Bayesian Adjustments*. New York: Chapman and Hall.

- Jackson, C., Best, N. and Richardson, S. (2008) Bayesian graphical models for regression on multiple datasets with different variables. *Technical report, Bias project, Imperial College London. (Currently under review)*. URL [www.bias-project.org.uk](http://www.bias-project.org.uk).
- Kramer, M. S. (1987) Intrauterine growth and gestational duration determinants. *Pediatrics*, **80**, 502–511.
- Langholz, B. (2001) Factors that explain the power line configuration wiring code-childhood leukaemia association: what would they look like? (with discussion). *Bioelectromagnetic Supplement*, **5**, S19–S31.
- Lauritzen, S. L. and Spiegelhalter, D. J. (1988) Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society, Series B*, **50**, 157–224.
- Lin, D. Y., Psaty, B. M. and A., K. R. (1998) Assessing sensitivity of regression results to unmeasured confounders in observational studies. *Biometrics*, **54**, 948–963.
- Little, R. J. A. and Rubin, D. B. (1987) *Statistical Analysis with Missing Data*. New York: John Wiley.
- Lunn, D., Best, N., Spiegelhalter, D., Graham, G. and Neuenschwander, B. (2008) Combining MCMC with ‘sequential’ PK-PD modelling. *Technical report, Bias project, Imperial College London. (submitted)*. URL [www.bias-project.org.uk](http://www.bias-project.org.uk).

- McCandless, L., Gustafson, P. and Levy, A. (2007) Bayesian sensitivity analysis for unmeasured confounding in observational studies. *Statistics in Medicine*, **26**, 2331–2347.
- McCandless, L., Richardson, S. and Best, N. (2008) Adjustment for unmeasured confounding using propensity scores. *Technical report, Bias project, Imperial College London. (submitted)*. URL [www.bias-project.org.uk](http://www.bias-project.org.uk).
- Molitor, J., Molitor, N.-T., Jerrett, M., McConnell, R., Gauderman, J., Berhane, K. and Thomas, D. (2006) Bayesian Modeling of Air Pollution Health Effects with Missing Exposure Data. *Am. J. Epidemiol.*, **164**, 69–76.
- Nieuwenhuijsen, M. J., Toledano, M. B., Eaton, N. E. and Elliott, P. (2000a) Chlorination disinfection byproducts in water and their association with adverse reproductive outcomes: a review. *Occup Environ Med*, **57**, 73–85.
- Nieuwenhuijsen, M. J., Toledano, M. B. and Elliott, P. (2000b) Uptake of chlorination disinfection byproducts; a review and a discussion of its implications for exposure assessment in epidemiological studies. *Journal of Exposure Analysis and Environmental Epidemiology*, **10**, 586–599.
- Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. New York: Cambridge University Press.
- Rao, J. N. K. (2003) *Small Area Estimation*. New York: Wiley.
- Richardson, S. and Best, N. G. (2003) Bayesian hierarchical models in ecological studies of health-environment effects. *Environmetrics*, **14**, 129–147.

- Rook, J. J. (1974) Formation of haloforms during chlorination natural waters. *J. Soc Water Treat Exam*, **23**, 234–243.
- Rosenbaum, P. R. (2004) Design sensitivity in observational studies. *Biometrika*, **91**, 153–164.
- Rosenbaum, P. R. and Rubin, D. B. (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika*, **70**, 41–55.
- Smith, K. and Joshi, H. (2002) The millennium cohort study. *Popul Trends*, 30–34.
- Spiegelhalter, D. J. (1998) Bayesian graphical modelling: a case study in monitoring health outcomes. *Journal of the Royal Statistical Society, Series C, (Applied Statistics)*, **47**, 115–133.
- Spiegelhalter, D. J., Thomas, A., Best, N. and Lunn, D. (2003) *WinBUGS version 1.4 user manual*. Cambridge: MRC Biostatistics Unit.
- Spiegelhalter, D. J., Thomas, A. and Best, N. G. (1996) Computation on bayesian graphical models. In *Bayesian Statistics 5* (eds. J. Bernardo, J. Berger, A. Dawid and F. Smith), 407–425. Oxford: Oxford University Press.
- Toledano, M. B., Nieuwenhuijsen, M. J., Best, N. G., Whitaker, H., Hambly, P., de Hoogh, C., Fawell, J., Jarup, L. and Elliott, P. (2005) Relation of trihalomethane concentrations in public water supplies to stillbirth and birth weight in three water regions in england. *Environmental Health Perspectives*, **113**, 225–232.

- Wakefield, J. C. (2003) Sensitivity analyses for ecological regression. *Biometrics*, **59**, 9–17.
- Whitaker, H., Best, N., Nieuwenhuijsen, M. J., Wakefield, J., Fawell, J. and Elliott, P. (2005) Modelling exposure to disinfection by-products in drinking water for an epidemiological study of adverse birth outcomes. *Journal of Exposure Analysis and Environmental Epidemiology*, **15**, 138–146.
- Whitaker, H., Nieuwenhuijsen, M. J. and Best, N. (2003) The relationship between water concentrations and individual uptake of chloroform: a simulation study. *Environmental Health Perspective*, **111**, 688–694.