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, Chris Jackson, Sylvia Richardson, and Nicky Best
Imperial College UK

February 8, 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, due to the complicated nature of the research question, a single data set may not provide sufficient information for valid inference. As a result, multiple data sources are often necessary to identify the biases and inform about different aspects of the research question. Standard analyses of each data source separately may fail to capture uncertainty other than simple random errors, thus may produce misleading results. Therefore it becomes necessary to link together different sub-models for each source in a comprehensive way. 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.

1 Introduction

Many social, behavioral, and health science studies are observational in nature and do not arise from carefully-controlled experimental designs. In addition to random errors, observational data typically suffer from other sources of uncertainty such as missing values, unmeasured confounders, measurement errors, and selection biases. Often, when analyzing data obtained from these studies, these indeterminacies, other than simple random errors, are 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 the 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 on the resulting inference. 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 the 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. 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 generalize 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 levels of information to properly estimate 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 (e.g. 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) 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 statistically 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. ≥ 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); (2) aggregate information on ethnicity from the 2001 census and small-area estimates of tobacco expenditure derived from consumer surveys.

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 Section 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 maternal age. This data set has been linked to modelled estimates of THM concentrations for some regions of England and Wales using a postcode-to-water supply zone link file developed by the Small Area Health Statistics Unit.
(SAHSU) to account for the changing locations in both water zone boundaries and postcodes over time (Toledano et al., 2005; Whitaker et al., 2005). 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 was one of the regions studied by Toledano et al. (2005); the time period was chosen to match the selection criterion of MCS subjects who were born in England (see below).

- Millennium Cohort Study (MCS)

The Millennium Cohort Study (MCS) 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). Three sweeps have been undertaken to date. For the first sweep upon which our study focuses, babies were eligible if they were born in the UK (England, Wales, Scotland, and Northern Ireland) during a specified 12-month period in the early 2000’s (for England, this was between September 2000 and August 2001) and lived in selected UK electoral wards at age 9 months. Responses were obtained for 18,819 babies. Both mothers and fathers were interviewed and the collected information included poverty and wealth, and the quality of family life and its support by public policy and the broader community. 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. The MCS sample were selected from a random sample of electoral wards (areas containing around 5000 individuals), disproportionally stratified to ensure adequate representation of all
four UK countries, deprived areas and areas with high concentrations of Black and Asian families. In England, wards were stratified into three categories for sampling purposes: ‘advantaged’, ‘disadvantaged’ and ‘high ethnic minority’, and 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

Census data in the UK provide a complete population count of different ethnicities for a given area (e.g. census output areas, COA, which contain around 200-300 individuals) taken on a specific date. Consumer survey data compiled by CACI Information Solutions Limited can be used to obtain small area estimates of various socioeconomic and lifestyle variables, such as income and expenditure, starting from September 2006. In our study, the ratio of non-white to white residents in each COA 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 was obtained from CACI. These were linked to each individual in the MCS and NBR using postcode-to-COA link files developed by SAHSU.
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 circles or ‘nodes’ in the graph represent the variables (data, missing values and parameters) and arrows between nodes indicate direct dependencies between variables. For clarity, we suppress full details of model parameters, which are specific to the particular application. 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 Figure 1 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, \( y_{ri} \), for subject \( i \) (where we use \( i = m \) as the index for the MCS data and \( i = n \) as the index for the NBR data), who lives in area \( r \), and the exposure, \( E_{ri} \), after adjusting for important covariates, \( C_{ri} \). In our example, we classify the birthweight outcome into three categories: \( y_{ri,1} = 1 \) if birthweight is normal, \( y_{ri,2} = 1 \) if the baby is born pre-term with low birthweight and \( y_{ri,3} = 1 \) if the baby is born full-term with low birthweight. In the MCS data, all levels for birth outcomes, \( y_{rm,u} (u = 1, 2, 3) \), and confounders, \( C_{rm} \), were fully observed. In the NBR, levels 2 and 3 of the birth outcome, \( y_{rn,u} (u = 2, 3) \), are missing as well as the confounders, \( C_{rn} \). 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 level 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 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 region-specific data, \( A_r \), 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 two 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 area variables, $A_r$, in the aggregate data and individual covariates, $C_{mr}$, in the survey (MCS) data to help impute the missing individual covariates, $C_{nr}$, 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 $C_{ri}$. 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, $A_r$, 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 $A_r$ in areas with missing values. This sub-model would typically require additional geographically-indexed individual and/or
aggregate data sources containing variables predictive of $A_r$, but could also make use of the sample survey data, $C_{mr}$, 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 (Spiegelhalter and Best (2003); see section “use of cut function”). 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, $y_{ri,u}$ is a categorical variable with 3 levels representing normal birthweight ($u = 1$), pre-term low birthweight ($u = 2$) and full-term low birthweight ($u = 3$). The covariates of interest include binary indicators of maternal smoking during pregnancy, $C_1$, and non-white ethnicity, $C_2$. We also consider two further covariates, maternal age and baby’s sex, which are established predictors of low birthweight. Since these two variables are fully observed in both the survey (MCS) and administrative (NBR) data, we denote them separately by $X$ (which also includes a dummy variable representing an intercept term) and model them as in standard regression analysis, with associated regression coefficients $\gamma$. The exposure, $E$, is a binary classification (low/medium =
≤ 30µg/l and high = > 30µg/l) of the total THM concentration during the final trimester of pregnancy in the water zone of the mother’s residence at time of the baby’s birth. The aggregate variables, $A_r$, are taken to be proportion of the resident population who are of non-white ethnicity and mean estimated annual income and mean estimated weekly expenditure on tobacco as a proportion of total expenditure on tobacco, wine, beer, fruit, vegetables and saturated fat for each census output area $r = 1, \ldots, R$.

### 3.2.1 Mathematical details of outcome sub-model

Our outcome sub-model is a multinomial logistic regression. We divide our total subjects into observed and missing outcome groups ($O^{obs}$ and $O^{miss}$), and specify the same multinomial distribution but with different probabilities for the observed and missing outcomes, $y_{ri,u}^{obs}$ and $y_{ri,u}^{miss}$, 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 $u = 2, 3$ in the NBR data. For subjects with observed outcomes (all MCS subjects, and those with normal birthweight in NBR), we specify

$$
\begin{cases}
    y_{ri,1:3}^{obs} \sim \text{Mult}(p_{ri,1:3},1) & \text{for } i \in O^{obs} \text{ and } r = 1, \ldots, R, \\
    \log(\frac{p_{ri,u}}{p_{ri,1}}) = \alpha_u^T C_{ri} + \beta_u E_{ri} + \gamma_u^T X_{ri} & \text{for } u = 2, 3,
\end{cases}
$$

where the notation $y_{ir,1:3}$ denotes elements 1 to 3 of the outcome vector for subject $i$ in area $r$. The first level of the outcome, $y_{ri,1}$, is treated as the reference group and the associated parameters $\alpha_1$, $\beta_1$ and $\gamma_1$ 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.

For subjects with missing outcomes, we need to impose the condition that the true outcome is either level 2 or 3 and renormalise the remaining probabilities, which is done as follows

\[
\begin{align*}
  y^\text{miss}_{ri,1:3} &\sim \text{Mult}(\mathbf{p}^*_{ri,1:3}, 1) \quad \text{for } i \in \mathcal{O}^{\text{miss}} \text{ and } r = 1, \ldots, R, \\
p^*_{ri,1} &= 0, \\
p^*_{ri,u} &= \frac{p_{ri,u}}{\sum_{u=2}^{p_{ri,u}}} \quad \text{for } u = 2, 3,
\end{align*}
\]

(2)

where \( p_{ri,u} \) 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 \((C_{1,ri} = 0 \text{ for white and } 1 \text{ for nonwhite})\) and maternal smoking during pregnancy \((C_{2,ri} = 0 \text{ for non-smoker and } 1 \text{ for smoker})\) in a manner that properly accounts for their within-person correlation, we use the multivariate probit model (Chib and Greenberg, 1998). The basic idea behind our implementation is to link the binary variables to underlying continuous normal variables, \( C^*_r = [C^*_{1,r}, C^*_{2,r}] \), and then model the continuous variables jointly by assuming that they arose from a bivariate normal distribution with area-specific mean \( \mathbf{\mu}_r = [\mu_{1,r}, \mu_{2,r}] \) and common variance-covariance matrix, \( \Omega \).

We model the area means as a function of the aggregate variables \( \mathbf{A}_r \), 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 co-
variates would be biased due to the non-random sampling mechanism used in the MCS (Brick and Kalton, 1996). This gives the following form for our covariate sub-model

\[
C_{q,ri} = I(C^*_q,ri > 0) \quad \text{for } q = 1,2,
\]

\[
C^*_r \sim \text{Multivariate Normal } (\mu_r, \Omega),
\]

\[
\mu_{q,r} = \delta_{0,q,s[r]} + \delta^T_{1,q} A_r,
\]

\[
\Omega = \begin{bmatrix}
1 & b \\
b & 1
\end{bmatrix}
\]

where \(I()\) is the indicator function. 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.

### 4 Simulation Study

In our Bayesian graphical model, missing values are imputed in two ways. First, missing covariates, \(C\), are generated from aggregate information, \(A_r\) (covariate sub-model). Second, missing outcomes, \(y_i\), are imputed from observed and missing covariates, \(C\) (disease 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-white 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 the exposure and fully observed covariates were not included in the simulation set-up. For computational reasons, 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).

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. The first scenario, Strong $Y$-strong $C$, was based entirely on real data, namely the real aggregate census and CACI data, $A$, on tobacco expenditure and proportion of non-white 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.
For the weak $Y$-strong $C$ scenario, we also 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. 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.

For each of the four complete datasets (one per scenario) created as described above, we then generated twenty replicate datasets containing missing values. For each replicate dataset, we randomly chose 80% of the smoking and race covariates and assigned them to be missing, 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. We also 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.
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 then calculated the probability 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) predicted by the fitted model, and averaged these across datasets for each scenario.

2. Outcome sub-model: The covariate values were assumed to be known for all subjects, 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 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 Spiegelhalter et al. (2003), section “use of the cut function”).

For comparison purposes, we also used the covariate and outcome sub-models to analyse each of the four complete 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 independent conditional on the observed data. The fully observed data were analyzed using STATA V9 (Stata Corporation, 2005) and the partially observed data were analyzed using WinBugs (Spiegelhalter et al., 2003).

4.3 Results of the simulation study

4.3.1 Examining the imputation of missing covariates

*Missing data at one level (analysis method 1)*

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). 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 (analysis method 3)**

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 latter 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 situation 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 data for each of the four scenarios. For each scenario, we then analysed the 20 replicate partial datasets using analysis methods 2 (outcome model fitted to data with missing outcome only), 3 and 4 (unified model, with and without feedback from outcomes to covariate sub-models, fitted to data with missing outcomes and covariates), and in each case calculated
the mean square error (MSE) between the posterior mean estimates of the log odds ratios $\bar{a}_{q,u}$ and the corresponding reference estimate $\hat{a}_{q,u}$. Table 2 summarised both the average point estimates and the MSEs for each scenarios 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 MSEs are small for all four scenarios (column 2 of Table 2). This suggests that the outcome imputation model performs almost as well as the reference model with complete data, particularly when the true covariate-outcome association is strong (compare top two rows of column 2 to bottom two rows). When covariates are also missing, the MSEs increase considerably (columns 3 and 4 of Table 2), indicating that most of the uncertainty in the outcome model parameter estimates is due to uncertainty about the missing covariates rather than missing outcomes. This is not surprising, given that a much higher percentage of the covariate data is missing compared to the outcome data. Of more interest is the impact of the feedback between the outcome and covariate sub-models in the unified analysis. When the true $Y$-$C$ association is strong, the MSEs are smaller when using the unified model without the cut function (fully Bayesian model) as opposed to the unified model with cut function (compare rows 1 and 2 of column 3 with column 4). This finding illustrates that when outcomes and covariates are strongly correlated, feedback is needed. On the other hand, when the true association between $Y$ and $C$ is weak, the extra feedback provides no strong information for imputing better estimates and actually appears to be detrimental, since the MSEs obtained from the unified model without cut function are much larger.
than those obtained from the unified model with cut function (compare rows 3 and 4 of column 3 with column 4).

5 Results of application to 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 which were covered by the United Utility Company (UU) 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 distribution of the main variables of interest in the two datasets. Overall, approximately 7% of the subjects had low birthweight 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 ($\leq 30\mu g/l$), medium ($30 - 60\mu g/l$), and high ($> 60g/l$) 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. 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. Somewhat surprisingly,
there was a higher proportion of babies in the ‘disadvantaged’ stratum in the NBR than the MCS (given the intention of the MCS to over-sample disadvantaged groups). However, the sampling strata are adjusted for in all our analyses so this imbalance should not distort our overall findings. Table 3 also summarises the distribution 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.

Model C Bayesian unified model (1)–(3) applied to the combined MCS, NBR and aggregate data, with imputation for both missing outcomes and covariates, but cutting feedback from the outcome to the covariate sub-model.

The outcome model in each of the above analyses included the dichotomised TTHM exposure variable, with adjustment for maternal smoking, maternal ethnicity, maternal age and baby’s gender. Also, all models included a strata-specific intercept to adjust for the sample selection mechanism of the MCS. To check whether maternal smoking and ethnicity were important confounders of the TTHM-birthweight association, a fourth model (Model D) was also fitted, which was as the same as Model A except that no adjustment was made for maternal smoking and ethnicity.

For Models B and C, the WinBUGS program was implemented with a
burn-in of 20,000 iterations followed by 20,000 iterations saved for analysis purposes. Non-informative priors were specified for all parameter estimates. For models A and D, STATA V9 was used.

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 four 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: 2.26, 95% credible interval: (1.09, 4.17)); 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. Comparing the TTHM odds ratios for Models A and D (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, as is evident when comparing results obtained from Model B (feedback allowed) to those obtained from Model C (cut function utilized).

Turning to the results for maternal smoking and non-white ethnicity, we see that the odds ratios are slightly higher 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 risk full-term low birthweight. It may seem surprising that we appear to gain little precision here from combining the data sources — indeed the interval estimates for the smoking and ethnicity odds ratios are actually slightly wider in Model B (combined data) than for Model A based only on the MCS data. 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 risk of full-term low birthweight is only from indirect sources — namely the aggregate data, 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 odds ratios for maternal smoking and ethnicity that are slightly closer to those from the MCS only data (Model A). 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 accurate.
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 risk of low birthweight yielded improved estimates of the exposure effect of interest, and demonstrated the benefit of using all available information 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 allowed in a Bayesian full probability model is generally desirable, our simulation studies indicated that it can be
detrimental in situations where the true association between the covariates and outcome is weak. In this case, the outcome contains little information about the covariate values, and the feedback mechanism just adds noise to the imputations. 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.

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.

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 considered 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 effi-
cient since it avoids the need to sample from the multinomial distribution.

Our Bayesian modeling framework can be extended in several directions,
for example by including additional sub-models to account for measurement
errors of exposure, and inclusion of factors which affect a mother’s true in-
take of the exposure variable. 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 mul-
tivariate probit model. In standard regression modelling, propensity score
(Pearl, 2000; Rosenbaum and Rubin, 1983) methods are increasingly being
used as an efficient method for dealing with a large number of observed con-
founders. The basic idea behind the method 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. In future work,
we intend to investigate the extension of propensity score ideas to handle
unobserved confounders and to incorporate these ideas within a Bayesian
framework.

Acknowledgment

The authors would like to acknowledge the contribution of Mireille Toledano
and Mark Nieuwenhuijsen who are responsible for conceiving the epidemi-
ological project on chlorination and birthweight and for their help in dis-
cussing the epidemiological aspects of the analysis. We are also grateful to
Mireille Toledano, Mark Nieuwenhuijsen, Daniela Fecht, James Bennett and Peter Hambly for their help in obtaining and processing the data.
Figure 1: Graphical representation of model. Note: Quantities in grey are unobserved.
Figure 2: Comparisons of the subject-specific conditional probability for all covariates patterns between the strong and weak generating relationship when missing values was observed only in covariates.
Figure 3: Comparisons of the subject-specific conditional probability for covariates pattern for white non-smokers. No cut function was applied.
Table 1: Summary of simulation design and parameter values used to generate the data

<table>
<thead>
<tr>
<th>Covariate sub-model</th>
<th>Outcome sub-model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Scenarios</td>
<td></td>
</tr>
<tr>
<td>Strong Y-Strong C</td>
<td>Real (Census,CACI)</td>
</tr>
<tr>
<td></td>
<td>Strong Real (MCS)</td>
</tr>
<tr>
<td>Weak Y-Strong C</td>
<td>Real (Census,CACI)</td>
</tr>
<tr>
<td></td>
<td>Strong Real (MCS)</td>
</tr>
<tr>
<td>Strong Y-Weak C</td>
<td>Real (Census,CACI)</td>
</tr>
<tr>
<td></td>
<td>Weak Simulate</td>
</tr>
<tr>
<td>Weak Y-Weak C</td>
<td>Real (Census,CACI)</td>
</tr>
<tr>
<td></td>
<td>Weak Simulate</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Strong A→C^♯ (δ×IQT↑)</th>
<th>Strong C→Y (OR*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Ratio of non-white</td>
<td>-0.042 0.125</td>
</tr>
<tr>
<td>Tobacco expenditure</td>
<td>0.601 0.052</td>
</tr>
<tr>
<td>C</td>
<td>LWBP° LWBF°</td>
</tr>
<tr>
<td>Smoke</td>
<td>1.083 2.411</td>
</tr>
<tr>
<td>Non-white</td>
<td>1.094 5.930</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Weak A→C^♯ (δ×IQT↑)</th>
<th>Weak C→Y (OR*)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>C</td>
</tr>
<tr>
<td>Ratio of non-white</td>
<td>-0.036 0.062</td>
</tr>
<tr>
<td>Tobacco expenditure</td>
<td>0.260 0.026</td>
</tr>
<tr>
<td>C</td>
<td>LWBP° LWBF°</td>
</tr>
<tr>
<td>Smoke</td>
<td>1.041 1.553</td>
</tr>
<tr>
<td>Non-white</td>
<td>1.041 2.226</td>
</tr>
</tbody>
</table>

*: Odds ratio.
†: Values in the A → C table are equal to δ×IQT, where δs are coefficients in the equation (3) and IQT is the difference between the 75th and 25th 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, ρ = −0.45
○: LWBP: low birthweight pre-term; LWBF: low birthweight full-term.
Table 2: Comparison of estimates and MSEs 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

<table>
<thead>
<tr>
<th>Data Scenarios</th>
<th>Analysis model</th>
<th>Complete Data</th>
<th>Partial Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(reference analysis(^*))</td>
<td>Outcome model only(^\dagger)</td>
<td>Unified model(^\sharp)</td>
</tr>
<tr>
<td>Parameters</td>
<td>Est</td>
<td>Est(MSE)</td>
<td>Est(MSE)</td>
</tr>
<tr>
<td>----------------</td>
<td>-----</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td><strong>Strong Y - Strong C</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke ($\alpha_{13}$)</td>
<td>0.88</td>
<td>0.91 (0.01)</td>
<td>1.07 (0.27)</td>
</tr>
<tr>
<td>non-white ($\alpha_{23}$)</td>
<td>1.78</td>
<td>1.83 (0.01)</td>
<td>2.22 (0.27)</td>
</tr>
<tr>
<td><strong>Strong Y - Weak C</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke ($\alpha_{13}$)</td>
<td>0.98</td>
<td>0.97 (0.00)</td>
<td>0.97 (0.51)</td>
</tr>
<tr>
<td>non-white ($\alpha_{23}$)</td>
<td>2.53</td>
<td>2.57 (0.01)</td>
<td>2.71 (0.50)</td>
</tr>
<tr>
<td><strong>Weak Y - Strong C</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke ($\alpha_{13}$)</td>
<td>-0.01</td>
<td>0.05 (0.01)</td>
<td>0.57 (1.33)</td>
</tr>
<tr>
<td>non-white ($\alpha_{23}$)</td>
<td>0.36</td>
<td>0.41 (0.02)</td>
<td>0.61 (0.39)</td>
</tr>
<tr>
<td><strong>Weak Y - weak C</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>smoke ($\alpha_{13}$)</td>
<td>0.38</td>
<td>0.34 (0.03)</td>
<td>0.91 (0.86)</td>
</tr>
<tr>
<td>non-white ($\alpha_{23}$)</td>
<td>1.00</td>
<td>1.06 (0.04)</td>
<td>1.23 (1.31)</td>
</tr>
</tbody>
</table>

\(^*\) Full data was analyzed by using the multinomial logistic regression model (STATA V9)
\(^\dagger\): Data with missing in $Y$ (outcome) was analyzed by using the disease sub-model (Winbugs)
\(^\sharp\): Data with missing in $Y$ (outcome) and $C$ (covariates) was analyzed by using the unified model (Winbugs)
\(^\dagger\): Data with missing in $Y$ (outcome) and $C$ (covariates) was analyzed by using the unified model with cut function (Winbugs)
Table 3: Demographic summary of participants in the low birthweight study

<table>
<thead>
<tr>
<th>Birthweight categories</th>
<th>MCS</th>
<th>NBR</th>
<th>MCS+NBR</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N (%)</td>
<td>N (%)</td>
<td>N (%)</td>
</tr>
<tr>
<td>Normal birthweight</td>
<td>1224 (91.8)</td>
<td>7308 (92)</td>
<td>8532 (92.0)</td>
</tr>
<tr>
<td>Low birthweight pre-term</td>
<td>68 (5.1)</td>
<td>NA</td>
<td>68 (0.7)</td>
</tr>
<tr>
<td>Low birthweight full-term</td>
<td>41 (3.1)</td>
<td>NA</td>
<td>41 (0.4)</td>
</tr>
<tr>
<td>Missing</td>
<td>NA</td>
<td>637 (8)</td>
<td>637 (6.9)</td>
</tr>
<tr>
<td>Maternal ethnicity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>White</td>
<td>1097 (82.3)</td>
<td>NA</td>
<td>1097 (11.8)</td>
</tr>
<tr>
<td>Non-white (Asian, black, others)</td>
<td>236 (17.7)</td>
<td>NA</td>
<td>236 (2.6)</td>
</tr>
<tr>
<td>Missing</td>
<td>NA</td>
<td>7945 (100)</td>
<td>7945 (85.6)</td>
</tr>
<tr>
<td>Maternal smoking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>819 (61.4)</td>
<td>NA</td>
<td>819 (8.8)</td>
</tr>
<tr>
<td>Yes</td>
<td>514 (38.6)</td>
<td>NA</td>
<td>514 (5.6)</td>
</tr>
<tr>
<td>Missing</td>
<td>NA</td>
<td>7945 (100)</td>
<td>7945 (85.6)</td>
</tr>
<tr>
<td>Maternal age</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>&lt; 20</td>
<td>138 (10.4)</td>
<td>874 (11.0)</td>
<td>1012 (10.9)</td>
</tr>
<tr>
<td>20-24</td>
<td>291 (21.8)</td>
<td>1894 (23.8)</td>
<td>2185 (23.6)</td>
</tr>
<tr>
<td>25-29</td>
<td>381 (28.6)</td>
<td>2237 (28.2)</td>
<td>2618 (28.2)</td>
</tr>
<tr>
<td>30-34</td>
<td>352 (26.4)</td>
<td>1906 (24.0)</td>
<td>2258 (24.3)</td>
</tr>
<tr>
<td>&gt;=35</td>
<td>171 (12.8)</td>
<td>1034 (13.0)</td>
<td>1205 (13.0)</td>
</tr>
<tr>
<td>Babies’ sex</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Females</td>
<td>657 (49.3)</td>
<td>3908 (49.2)</td>
<td>4565 (49.2)</td>
</tr>
<tr>
<td>Males</td>
<td>676 (50.7)</td>
<td>4037 (50.8)</td>
<td>4713 (50.8)</td>
</tr>
<tr>
<td>Total THMs exposure</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Low (&lt; 31g/l)</td>
<td>156 (11.7)</td>
<td>576 (7.2)</td>
<td>732 (8.0)</td>
</tr>
<tr>
<td>Medium (31 – 60g/l)</td>
<td>578 (43.4)</td>
<td>4011 (50.5)</td>
<td>4589 (49.4)</td>
</tr>
<tr>
<td>High (&gt; 60g/l)</td>
<td>599 (44.9)</td>
<td>3358 (42.3)</td>
<td>3957 (42.6)</td>
</tr>
<tr>
<td>MCS sampling stratum</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advantaged</td>
<td>442 (33.2)</td>
<td>1623 (20.4)</td>
<td>2065 (22.2)</td>
</tr>
<tr>
<td>Disadvantage</td>
<td>707 (53.0)</td>
<td>5190 (65.3)</td>
<td>5897 (63.6)</td>
</tr>
<tr>
<td>High ethnic minority</td>
<td>184 (13.8)</td>
<td>1132 (14.3)</td>
<td>1316 (14.2)</td>
</tr>
<tr>
<td>Aggregates data - census output area</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Median (Interquartile Range)</td>
<td>26887 (22356, 33078)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Income (pound)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ratio between nonwhite to white</td>
<td>0.03 (0.01, 0.08)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proportion of tobacco expenditure†(%)</td>
<td>35 (30, 42)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

†:Proportion of tobacco expenditure among total expenditure on tobacco, wine, beer, fruits, vegetables, and saturated fat
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

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Model A&lt;sup&gt;♮&lt;/sup&gt;</th>
<th>Model B&lt;sup&gt;∗&lt;/sup&gt;</th>
<th>Model C&lt;sup&gt;⋄&lt;/sup&gt;</th>
<th>Model D&lt;sup&gt;✠&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>OR (95% CI&lt;sup&gt;‡&lt;/sup&gt;)</td>
<td>OR (95% CI&lt;sup&gt;†&lt;/sup&gt;)</td>
<td>OR (95% CI&lt;sup&gt;†&lt;/sup&gt;)</td>
<td>OR (95% CI&lt;sup&gt;‡&lt;/sup&gt;)</td>
<td></td>
</tr>
<tr>
<td>TTHM &gt; 60 g/l</td>
<td>1.51 (0.76, 3.01)</td>
<td>2.26 (1.09, 4.17)</td>
<td>2.26 (1.15, 4.18)</td>
<td>1.97 (1.02, 3.83)</td>
</tr>
<tr>
<td>Smoker</td>
<td>2.45 (1.18, 5.07)</td>
<td>2.73 (1.25, 5.31)</td>
<td>2.46 (1.12, 4.64)</td>
<td>NA</td>
</tr>
<tr>
<td>Non-white</td>
<td>5.25 (2.17, 12.72)</td>
<td>7.41 (3.34, 14.49)</td>
<td>6.39 (2.90, 12.69)</td>
<td>NA</td>
</tr>
</tbody>
</table>

<sup>♮</sup>: Model A: Multinomial logistic regression model with adjusting for smoke and non-white
<sup>∗</sup>: Model B: Bayesian unified model with imputing missing covariates and outcomes (no cut function)
<sup>⋄</sup>: Model C: Bayesian Unified model with imputing missing covariates and outcomes (cut function)
<sup>✠</sup>: Model D: Multinomial logistic regression model without adjusting for smoke and non-white

<sup>‡</sup>: Bayesian Credible Interval
<sup>†</sup>: Confidence Interval
<sup>♮</sup>: Adjusted for maternal age and baby’s sex
References


M. J. Nieuwenhuijsen, M. B. Toledano, and P. Elliott. Uptake of chlorination disinfection byproducts; a review and a discussion of its implications


