

Adjustment for Missing Confounders Using External Validation Data and Propensity Scores

Lawrence C. McCandless¹

Sylvia Richardson²

Nicky Best²

¹ Faculty of Health Sciences, Simon Fraser University, Canada. ² Department of Epidemiology
and Public Health, Imperial College London.

October 2009

Corresponding author:

Lawrence McCandless
Assistant Professor of Biostatistics
Faculty of Health Sciences
Simon Fraser University
8888 University Drive
Burnaby BC V5A 1S6
Canada
mccandless@sfu.ca
Tel: 778-782-8651
www.fhs.sfu.ca/portal_memberdata/lmccandless

Abstract

Reducing bias from missing confounders is a challenging problem in the analysis of observational data. Information about missing variables is sometimes available from external validation data, such as surveys or secondary samples drawn from the same source population. In principal, the validation data permits us to recover information about the missing data, but the difficulty is in eliciting a valid model for nuisance distribution of the missing confounders. Motivated by a British study of the effects of trihalomethane exposure on risk of full-term low birthweight, we describe a flexible Bayesian procedure for adjusting for a vector of missing confounders using external validation data. We summarize the missing confounders with a scalar summary score using the propensity score methodology of Rosenbaum and Rubin. The score has the property that it breaks the dependence between the exposure and the missing confounders within levels of covariates. To adjust for bias in a Bayesian analysis, we need only update and adjust for the summary score during Markov chain Monte Carlo simulation. Simulation results illustrate that the proposed method reduces bias from several missing confounders over a range of different sample sizes for the validation data.

Keywords: Bias; Observational studies; Bayesian inference, causal inference

Running title: Adjustment for Missing Confounders.

1. Introduction

A challenge in observational research is how to reduce bias when there are missing confounding variables. A popular approach is use sensitivity analysis techniques that work from the assumption of a single binary missing confounder (e.g. Rosenbaum and Rubin (1983a)). In regression analysis, this involves a parametric model for the observed data, averaging over the distribution of the missing variable. The resulting model is nonidentifiable and indexed by so-called *bias parameters* that characterize the confounding effect of the missing variable. To eliminate confounding, the investigator substitutes values for bias parameters taken from the literature. Alternatively, one can use a Bayesian approach where uncertainty about bias parameters is incorporated into the analysis using prior distributions (McCandless, Gustafson and Levy, 2007).

In practice we may have complicated patterns of missing confounders and the assumption of a single binary missing variable is unrealistic. One estimation strategy is to use Bayesian iterative simulation methods based on data-augmentation (Gelman, Carlin, Stern and Rubin 2004). We model the joint distribution of the data and missing confounders. Inference proceeds via posterior updating of the missing confounders using Markov chain Monte Carlo (MCMC). But the difficulty is in eliciting a satisfactory model for the nuisance distribution of missing confounders. They may be high dimensional, correlated and have continuous or categorical components. Parametric models may give inadequate representations of complex patterns of missing data.

In this article, we consider the setting where supplementary information on missing confounders is available from external validation data. Examples include surveys or secondary samples from the population under study, typically of modest size. We distinguish between the *primary data* which denotes the original dataset and the *validation data* which denotes a second smaller sample of subjects drawn from the same source population and with complete information about missing variables. The motivation for this work is the intuitive idea that

it should be possible to develop a flexible procedure for using the validation data in order to recover information about the missing covariates.

The problem of combining inferences from primary and validation data to control for missing confounders has been studied in the context of two-stage sampling designs. Schill and Drescher (1997) and Breslow and Holubkov (1997) review two-stage sampling methods for control of confounding and other biases in observational studies. Fully parametric methods for adjusting for several missing confounders are available (Wacholder and Weinberg 1994; Wakefield and Salway 2001; Lin, Sundberg, Wang and Rubin 2006; Jackson, Best and Richardson 2008). But they are restricted to the setting of one or two covariates that are categorical or continuous. Alternatively, Chatterjee, Chen and Breslow (2003) describes techniques that use non-parametric density estimates of the distribution of the missing confounders. However high dimensional density estimation is difficult in small samples, and these methods are best suited to the case of a single missing covariate.

In this article we describe a flexible Bayesian method for adjusting for several missing confounders using external validation data. It can be used when the confounders are both continuous and categorical, and it does not require strong parametric assumptions about the distribution of the missing variables. To adjust for bias we use the idea of propensity scores, proposed by Rosenbaum and Rubin (1983b). Propensity scores techniques are a class of statistical methods that alleviate the challenges of specifying a regression model for the outcome variable in the face of multiple confounders. See Lunceford and Davidian (2004) for an overview of propensity score techniques. Little and Rubin review missing data imputation techniques using probabilities of selection (2002).

In the present investigation, the focus is somewhat different from standard applications of propensity score techniques because we distinguish between measured versus missing confounders. To adjust for the *missing* confounders, we summarize them using a scalar summary score, which can be interpreted as the propensity score adjusted for *measured* confounders. The

score has the property that it breaks the association between the exposure and missing confounders. To adjust for bias using the validation data, we first specify a joint model for the primary and validation data. Adjustment for the summary score is then performed naturally as part of fitting the joint model during Markov chain Monte Carlo simulation. Modelling variation in the outcome variable within levels of the missing confounders is not required.

To illustrate the problem of missing confounders, Section 2 begins by describing a study from environmental epidemiology of the effect of trihalomethane exposure, a water disinfection by-product, on risk of full-term low birthweight in England. The primary data are obtained from the Hospital Episode Statistics (HES) database, which benefits from a large sample size and national UK geographic coverage. However, the HES has only limited information on factors influencing birthweight such as maternal smoking and ethnicity. Rich covariate information on seven missing confounders is taken from validation data from the Millennium Cohort Study (MCS), which describes the health of a cohort of UK mothers and children. In Section 3, we describe a method to adjust for missing confounders using propensity score techniques. We outline the model, prior distributions and an algorithm for posterior simulation. We apply the method in Section 4 and show that trihalomethane exposure is associated with increased risk of full-term low birthweight, but that this association is reduced upon adjustment for missing confounders. Simulation results in Section 5 show that the proposed method reduces bias from missing confounders over a range of sample sizes for the validation data.

2. Example: Estimating the Effect of Trihalomethane Exposure on Low Birthweight

To illustrate the problem of missing confounders, we consider the example of an observational study of the relationship between trihalomethanes, a water disinfection by-product, and risk of full-term low birthweight in the United Kingdom (Toledano, Nieuwenhuijsen, Best et al. 2005; Molitor, Jackson, Best and Richardson 2009). Trihalomethanes are formed when chlorine, which

is routinely added to public water supplies in the UK, reacts with natural organic materials in the water. Pregnant mothers are exposed to trihalomethanes through drinking and bathing. Animal studies show that water disinfection by-product compounds cause reproductive and developmental harm at higher doses. But epidemiologic investigations of adverse birth outcomes have yielded contradictory findings with some studies reporting increased risk of low birthweight, while others showing no association (Toledano et al. 2005). It is important to distinguish between low birthweight due to preterm birth versus full-term low birthweight, which may indicate intrauterine growth retardation. Full-term low birthweight is rare and any increase in risk is likely to be small, but with important public health consequences in view of the large numbers of mothers and babies exposed to trihalomethanes in the population (see Table 1). Furthermore, exposure assessment is prone to measurement error and published studies are often missing information on important confounders. These study characteristics are likely to mask any true association.

In the present investigation, we build on the work of Toledano et al. (2005) and Molitor et al. (2009). These studies used the UK National Birth Register to obtain data on birthweight. However we use hospital birth records because they also include information on gestational age. Thus our primary data are taken from the Hospital Episode Statistics (HES), which is a data warehouse with details of all hospital admissions, including births, from National Health Service (NHS) hospitals in the UK. We consider a total of 9793 births occurring between 2000 and 2001 in a region of Northern England serviced by a single water supply company. A small proportion of births occurring in non-NHS hospitals or at home are not included, but there is no reason to believe that these missing births will cause bias in our analysis. Each birth was linked to area-level estimates of trihalomethane water concentrations using a postcode-to-water supply zone link file that was developed by the Small Area Health Statistics Unit at Imperial College London. See Toledano et al. (2005) for further details. The outcome under study was full-term low birthweight, which is defined as gestational age greater than 37 weeks in combination with

a birthweight less than 2.5kg.

Let Y be an indicator variable for the outcome, taking value one if an infant has full-term low birthweight, and zero otherwise. Let X be an indicator variable for trihalomethane exposure, taking value one if the area-level exposure is greater than $60\mu\text{g}/\text{L}$ and zero otherwise. Let C denote a vector of $p = 5$ confounding variables that are contained in the primary data. These include indicator variables for mother’s age (≤ 25 , $25-29$, $30-34$, ≥ 35), an indicator variable if the baby is male, and Carstairs score quintile, which measures neighborhood-level socioeconomic deprivation. Upper quintiles imply greater deprivation.

Table 1 gives demographic details of the exposure groups. Full-term low birthweight is more common in the exposed group occurring in 3.9% of births versus 3.0% for the unexposed group. To explore the association between X and Y in the primary data, we fit a logistic regression of Y on X while adjusting for C . The results are presented in Table 2 under the heading “NAIVE”. We see an odds ratio of 1.39 with 95% interval estimate (1.11, 1.75) indicating that trihalomethane exposure seems to be associated with increased risk of full-term low birthweight.

A difficulty with the NAIVE analysis is that the effect estimate is likely to be biased from missing confounders. The HES data has the advantage of a large sample size and nearly exhaustive coverage. But it contains only limited information on factors that influence birthweight, such as maternal smoking and ethnicity. Trihalomethane water concentrations vary by neighborhood, as do smoking rates and other socioeconomic variables. Without appropriate adjustment for confounding, the risk patterns between exposure groups may be an artifact of systematic differences in characteristics of populations.

In this investigation, information about missing confounders is available from external validation data. The UK Millennium Cohort Study (MCS) contains survey information on mothers and infants born during the period 2000-2001 in the same region where the primary data were collected. The MCS data are a disproportionately stratified sample based on neighborhood income and ethnicity. Hence, inferences from the MCS data must adjust for non-random sampling.

This can be accomplished using sample survey weights supplied in the MCS documentation. Following Molitor et al. (2009), postcode at birth is used to match MCS subjects with birth records in the HES, resulting in a match for 824 births. We thus have information about missing confounders for 824 out of the 9793 births in the same region during 2000-2001. Thus the primary data has sample size $n = 8969$, while the validation data where full information on missing confounders is available has sample size $m = 824$.

Upon consultation with subject area experts, we identify seven variables in the validation data that could potentially confound the exposure-outcome association. Let U denote the $q = 7$ vector of missing confounders, which include lone parent family, number of live children for mother, maternal smoking, alcohol consumption, body mass index prior to pregnancy ≥ 25 kg/m², non-white ethnicity, and an indicator variable for low education. Table 1 gives a breakdown of the covariate distributions for the validation data. We see that non-white ethnicity is imbalanced between exposure groups, whereas smaller imbalances are observed for the other variables.

Other covariates are available in the MCS but are judged to be less important predictors of the exposure or outcome based on subject matter considerations. These include maternal employment, benefits and maternity leave, housing type (ownership versus rent, # rooms) and neighborhood satisfaction. Self-reported income is also available, but is excluded from the analysis because the Carstairs score in the HES is thought to adequately capture socioeconomic status.

Denote the primary data as $\{(Y_i, X_i, C_i, U_i) \mid \text{for } i \in 1 : n = 8969\}$ and the validation data as $\{(Y_j, X_j, C_j, U_j) \mid \text{for } j \in 1 : m = 824\}$. The quantity U_i is completely unobserved. To study potential confounding induced by U , Table 3 presents odds ratios for the association between Y_j and X_j when adjusting for C_j alone, versus adjusting for C_j and U_j in the validation data. In the first column of Table 3, we fit a weighted logistic regression of Y_j on X_j and C_j using the survey weights supplied in the MCS documentation. The odds ratio for the exposure effect

is equal to 2.06 with 95% interval (0.87, 4.89). In the second column we fit the same weighted regression, but adjust for both C_j and U_j and obtain 1.75 (0.70, 4.34). The odds ratio is shifted towards 1 and the interval is slightly wider, indicating some evidence of additional confounding by U_j . This suggests that the missing U may confound the association between exposure and outcome in the HES primary data.

We explored possible interactions between the exposure and either smoking or ethnicity by including product terms in the MCS outcome model. However full-term low birthweight occurs in only 23/824 of births in the MCS and this restricts fitting interaction terms. Stratified analysis of 2×2 tables over exposure, outcome and either smoking or ethnicity identified main effects but no significant interactions. Thus interactions may remain that cannot be identified because of low power. Further details on interactions are given in Appendix I of the supplementary materials.

Table 3 indicates that the components of U play a role in explaining variability in birthweight. Valid inference from the primary data may require adjustment for missing confounders. As discussed in the introduction, a standard analytic approach is to model the distribution of missing covariates. But this is challenging in the trihalomethane data because U is high dimensional with components that are correlated and that may depend on X and C .

3. Bayesian Adjustment for Missing Confounders (BAYES).

We present a Bayesian method to adjust for missing confounders using external validation data and propensity scores, which we henceforth call by the acronym BAYES. In Section 3.1, we describe a model for the missing confounders. We model the joint density function $P(Y, X, U|C)$ and then integrate over the missing U . This gives likelihood functions for the primary data and validation data that can be used for Bayesian inference. A family of prior distributions for model parameters is given in Section 3.2, while Section 3.3 describes an algorithm for posterior simulation.

3.1 Models

3.1.1 A Model for the Complete Data

Suppose that (Y_i, X_i, C_i, U_i) and (Y_j, X_j, C_j, U_j) for $i \in 1 : n$ and $j \in 1 : m$ are identically distributed observations drawn from the same population with probability density function $P(Y, X, C, U)$. Building on the Bayesian propensity score analysis of McCandless, Gustafson and Austin (2009), we model the conditional density $P(Y, X|C, U)$ using a pair of logistic regression models:

$$\text{Logit}[P(Y = 1|X, C, U)] = \beta X + \xi^T C + \tilde{\xi}^T g\{Z(U)\} \quad (1)$$

$$\text{Logit}[P(X = 1|C, U)] = \gamma^T C + Z(U), \quad (2)$$

where $Z(U) = \tilde{\gamma}^T U$.

Equation (1) is a model for the outcome and includes an exposure effect parameter β and a linear term for the covariates C with regression coefficients $\xi = (\xi_0, \dots, \xi_p)$. Equation (2) models the probability of exposure, which depends on the measured and missing confounders via the regression coefficients $\gamma = (\gamma_0, \dots, \gamma_p)$ and $\tilde{\gamma} = (\tilde{\gamma}_1, \dots, \tilde{\gamma}_q)$. To ease modelling of regression intercept terms, we set the first component of C equal to one, so that C is a $(p + 1) \times 1$ vector, which includes the intercept.

In equations (1) and (2), the quantity $Z(U) = \tilde{\gamma}^T U$ is a scalar summary of U , which can be interpreted as the propensity score adjusted for C . If we condition on C , then quantity $\gamma^T C$ is a constant intercept term in equation (2). Variability in X due to U is mediated entirely through the summary score $Z(U)$.

From equation (2), we have

$$X \perp\!\!\!\perp U|C, Z(U). \quad (3)$$

This result is analogous to the conclusion of Theorem 1 of Rosenbaum and Rubin (1983b). It states that within levels of C , conditioning on $Z(U)$ forces independence between X and

U . In Appendix II of the supplementary materials we prove that if there is no unmeasured confounding conditional on (C, U) , then equation (3) implies that there is no unmeasured confounding conditional on $(C, Z(U))$. This means that exposure effect measures computed from the marginal density $P(Y|X, C, Z(U))$ have a causal interpretation. To control for confounding bias in a Bayesian analysis, we can estimate the exposure effect by using models which assume that $Y \perp\!\!\!\perp U|X, C, Z(U)$. Further discussion of Bayesian regression adjustment for the propensity score is given by Rubin (1985).

Accordingly, equation (1) includes the quantity $Z(U)$ as a covariate in a regression model for the outcome. Because $Z(U)$ is a complex scalar quantity with no epidemiological interpretation, its link to Y is modelled in a nonparametric manner via the linear predictor $g\{\cdot\}$. For the trihalomethane data example, we use splines and let $g\{z\} = \sum_{j=0}^l \tilde{\xi}_j g_j\{z\}$, where the quantities $g_j\{\cdot\}$ are natural cubic spline basis functions with l knots and regression coefficients $\tilde{\xi} = (\tilde{\xi}_0, \dots, \tilde{\xi}_l)$. This gives a smooth yet flexible relationship between $Z(U)$ and Y within levels of X and C . See Lunceford and Davidian (2004) for a detailed discussion of regression modelling strategies that use the propensity score as a covariate.

We could alternatively control confounding from C and U by fitting the outcome model

$$\text{Logit}[P(Y = 1|X, C, U)] = \beta X + \tilde{\xi}^T g\{Z(U, C)\} \quad (4)$$

where $Z(U, C) = \text{Logit}[P(X = 1|U, C)] = \gamma^T C + \tilde{\gamma}^T U$. This is a standard approach to adjustment for confounding (e.g. Lunceford and Davidian 2004), and it does not distinguish between measured and missing confounders in calculating the propensity score. However, an advantage of using equation (1) rather than equation (4) is that it allows direct modelling of variability in Y arising from C . We include a linear term “... + $\xi^T C$ + ..” while summarizing U with the summary score $Z(U)$. This is appropriate in the present context because C is measured while U is not.

3.1.2 The Resulting Model when U is Missing

In the primary data, the quantities Y_i, X_i and C_i are observed while U_i is missing. Equations (1) and (2) define the density $P(Y, X|C, U)$, and we can use it to calculate the marginal model for $P(Y, X|C)$ integrating over U . We have

$$\begin{aligned} P(Y, X|C) &= E\{P(Y, U, X|C)\} \\ &= \int P(Y|X, C, U)P(X|C, U)P(U|C)dU, \end{aligned} \quad (5)$$

where $P(Y|X, C, U)$ and $P(X|U, C)$ are given in equations (1) and (2). To complete the specification, we require a model for U given C .

One simplification is to assume that C and U are marginally independent, meaning that $P(U|C) = P(U)$. In this case, we can model $P(U)$ using the empirical distribution $\{U_j|j \in 1 : m\}$ from the validation data alone. We may approximate

$$P(Y, X|C) \approx \frac{1}{m} \sum_{j=1}^m P(Y|X, C, U_j)P(X|U_j, C). \quad (6)$$

Using this approach, we model $P(Y, X|C) = E\{P(Y, U, X|C)\}$ as the empirical average of $P(Y|X, C, U_j)P(X|U_j, C)$ over samples of U_j . An advantage of this representation is that it requires no parametric modelling assumptions for the nuisance distribution of U_j . It can be used regardless of the correlation structure of the components of U_j and with continuous and categorical variables.

The assumption that U and C are marginally independent may be implausible, and it can be examined in the validation data. A different estimation strategy would be to model the distribution of the missing confounders directly. This would be feasible in the trihalomethane data example because the components of U are primarily categorical. For the case of two missing confounders, this approach is explored by Molitor et al. (2009). For the present investigation, we seek to avoid a parametric model for U because it is high dimensional. Furthermore, in the trihalomethane data we find only weak evidence of associations between U and C (see Appendix III). We explore sensitivity of the BAYES procedure to the conditional independence assumption between U and C through simulations in Section 5.

3.1.3 Bias Parameters and Nonidentifiability.

In equations (1) and (2), the parameters $\tilde{\xi}$, $\tilde{\gamma}$ model the relationship between the missing confounder U and the data Y , X , C . We call these quantities *bias parameters* in the sense that they model bias from U . If the parameters $\tilde{\gamma}$ or $\tilde{\xi}$ are large in magnitude then this means that U contains powerful confounders (Greenland 2005).

The model for the primary data given in equation (6) is nonidentifiable. To illustrate, suppose that the validation data has sample size $m = 1$. Then equation (6) becomes

$$P(Y, X|C) \approx \left[\frac{\exp(Y(\beta X + \xi^T C + \tilde{\xi}^T g\{\tilde{\gamma}^T U^*\}))}{1 + \exp(\beta X + \xi^T C + \tilde{\xi}^T g\{Z(U^*)\})} \right] \left[\frac{\exp(X(\gamma^T C + \tilde{\gamma}^T U^*))}{1 + \exp(\gamma^T C + \tilde{\gamma}^T U^*)} \right],$$

where U^* is a known fixed quantity taken from the validation data. The conditional distribution of Y given X and C cannot distinguish between the quantities ξ_0 and $\tilde{\xi}^T g\{Z(U^*)\}$ because they both serve as regression intercept terms. We can only estimate the sum $\xi_0 + \tilde{\xi}^T g\{Z(U^*)\}$. Similarly, we cannot distinguish between γ_0 and $\tilde{\gamma}^T U^*$. Because U is missing, the density $P(Y, X|C)$ can arise from different and indistinguishable patterns of confounding and baseline prevalences of Y and X .

If the bias parameters $\tilde{\xi}$ and $\tilde{\gamma}$ are known a priori, then the model for the primary data in equation (6) is identifiable. The quantities $\tilde{\xi}^T g\{Z(U)\}$ and $\tilde{\gamma}^T U$ are known offsets in the density $P(Y, X|C)$. We can calculate maximum likelihood estimates for (β, ξ, γ) from the likelihood function for the primary data, given by

$$\begin{aligned} L(\beta, \xi, \gamma) &= \prod_{i=1}^n P(Y_i, X_i|C_i). \\ &\approx \prod_{i=1}^n \left\{ \frac{1}{m} \sum_{j=1}^m P(Y_i|X_i, C_i, U_j) P(X_i|U_j, C_i) \right\} \\ &= \prod_{i=1}^n \left\{ \frac{1}{m} \sum_{j=1}^m \left[\frac{\exp(Y_i(\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{Z(U_j)\}))}{1 + \exp(\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{Z(U_j)\})} \right] \right. \\ &\quad \left. \times \left[\frac{\exp(X_i(\gamma^T C_i + \tilde{\gamma}^T U_j))}{1 + \exp(\gamma^T C_i + \tilde{\gamma}^T U_j)} \right] \right\}. \end{aligned} \tag{7}$$

Consequently, an alternative frequentist approach to adjusting for the missing confounders is

to plug in estimates for the bias parameters $\tilde{\xi}$ and $\tilde{\gamma}$ into equation (7) and then maximize the likelihood over (β, ξ, γ) . This approach to external adjustment is conceptually similar to the sensitivity analysis technique of Rosenbaum and Rubin (1983a). We revisit this maximum likelihood procedure in Section 4 and compare it to the BAYES method.

3.2 Prior Distributions

The quantities $\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}$ are regression coefficients, and we assign prior distributions of the form

$$\beta, \xi_0, \dots, \xi_p, \tilde{\xi}_0, \dots, \tilde{\xi}_l, \gamma_0, \dots, \gamma_p, \tilde{\gamma}_1, \dots, \tilde{\gamma}_q \sim N \left\{ 0, \left(\frac{\log(15)}{2} \right)^2 \right\}.$$

This models the belief that the odds ratio for the exposure effect β is not overly large and lies between 1/15 and 15 with probability 95%. We make similar assumptions about the association between Y and C , $Z(U)$ given X , and also the association between C , U and X . Such priors are very plausible and capture the magnitude and direction of effect estimates in typical epidemiologic investigations (Vaillefond, Raftery and Richardson, 2001). In Section 4.1, we study prior sensitivity in the trihalomethane data example.

3.3 Posterior Simulation

Let *data* denote both the primary and validation data. Inferences from BAYES are obtained from the posterior density $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma} | \text{data})$, which we sample from using MCMC simulation

techniques. We have

$$\begin{aligned}
P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma} | data) &\propto \left\{ \prod_{i=1}^n P(Y_i, X_i | C_i) \right\} \times \left\{ \prod_{j=1}^m P(Y_j, X_j | C_j, U_j) \right\} \times P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}) \\
&\approx \prod_{i=1}^n \left\{ \frac{1}{m} \sum_{j=1}^m \left[\frac{\exp(Y_i(\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{Z(U_j)\}))}{1 + \exp(\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{Z(U_j)\})} \right] \right\} \\
&\times \left\{ \frac{\exp(X_i(\gamma^T C_i + \tilde{\gamma}^T U_j))}{1 + \exp(\gamma^T C_i + \tilde{\gamma}^T U_j)} \right\} \\
&\times \prod_{j=1}^m \left\{ \left[\frac{\exp(Y_j(\beta X_j + \xi^T C_j + \tilde{\xi}^T g\{Z(U_j)\}))}{1 + \exp(\beta X_j + \xi^T C_j + \tilde{\xi}^T g\{Z(U_j)\})} \right] \right\} \\
&\times \left\{ \frac{\exp(X_j(\gamma^T C_j + \tilde{\gamma}^T U_j))}{1 + \exp(\gamma^T C_j + \tilde{\gamma}^T U_j)} \right\} \times P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}),
\end{aligned}$$

where the products over i and j are the likelihood functions for the primary and validation data, respectively, and $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma})$ is the prior density for $\beta, \xi, \tilde{\xi}, \gamma$ and $\tilde{\gamma}$.

We sample from $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma} | data)$ by updating from the conditional distributions for $[\beta, \xi, \tilde{\xi} | \gamma, \tilde{\gamma}, data]$ and $[\gamma, \tilde{\gamma} | \beta, \xi, \tilde{\xi}, data]$ using the Metropolis Hastings algorithm. To update from $[\beta, \xi, \tilde{\xi} | \gamma, \tilde{\gamma}, data]$, we use a proposal distribution based on a random walk that updates each component $\beta, \xi_0, \dots, \xi_p, \tilde{\xi}_0, \dots, \tilde{\xi}_l$ one at a time using a mean zero normal disturbance. Multivariate updating from $[\gamma, \tilde{\gamma} | \beta, \xi, \tilde{\xi}, data]$ is accomplished using the proposal distribution described in McCandless et al. (2009).

A difficulty with this sampling scheme is that the likelihood function for the primary data is expensive to compute. It requires calculating $P(Y_i, X_i | C_i, U_j)$ over all combinations of i and j . Alternatively, we can recognize that the expression for $P(Y, X | C)$ in (6) is a sample mean estimate of $E\{P(Y, U, X | C)\}$. We can use a quadrature estimate

$$P(Y, X | C) \approx \sum_{k=1}^M \omega_k \left[\frac{\exp\{Y(\beta X + \xi^T C + \tilde{\xi}^T g\{\hat{Z}_k\})\}}{1 + \exp\{\beta X + \xi^T C + \tilde{\xi}^T g\{\hat{Z}_k\}\}} \right] \left[\frac{\exp\{X(\gamma^T C + \tilde{\gamma}^T U)\}}{1 + \exp\{\gamma^T C + \tilde{\gamma}^T U\}} \right]$$

based on a histogram of $\{\tilde{\gamma}^T U_j | j \in 1 : m\}$ instead of equation (6). Here the index k equals $1, \dots, M$, where M is the number of histogram bins. The quantities \hat{Z}_k are the interval midpoints in the histogram and ω_k are the bin frequencies. In applications we find this summation faster to compute because it requires fewer evaluations of $P(Y, X | C, U)$.

4. Analysis Results for the Trihalomethane Data

4.1 Bayesian Adjustment for Missing Confounders (BAYES).

Before applying the BAYES method to the trihalomethane data, we set a priori values for the knots used to define the linear predictor $g\{\cdot\}$ in equation (1). Following McCandless et al. (2009), we fit the logistic regression model given in equation (2) to the validation using maximum likelihood to estimate the bias parameter $\tilde{\gamma}$. The quantities $Z(U_j)$, computed by evaluating $\{Z(U_j) = \tilde{\gamma}^T U_j | j \in 1 : m\}$, range from -0.3 to 2.0. Two knots are chosen as 0.03, 0.92 to define approximate tertiles for the true distribution of $Z(U)$.

Table 3 gives a preliminary illustration of $Z(U_j)$ as a tool to control confounding. In the rightmost column we fit the model in equation (1) to the validation data all by itself, and using weighted logistic regression with the MCS survey weights. The resulting odds ratios for the exposure effect is 1.77 (0.73, 4.31), which agrees closely with the estimate 1.75 (0.70, 4.34) obtained from the analysis which adjusts for U_j directly. In other words, $Z(U_j)$ succeeds in approximating the vector U_j in the validation data.

We now fit the BAYES method to the primary and validation data combined. As discussed in Section 2, the MCS data are collected through disproportionately stratified sampling based on neighborhood income and ethnicity. Thus we must account for non-random sampling when combining the likelihood of the primary data with that of the validation data. To do this we use the survey weights that are supplied with the MCS documentation. We calculate the likelihood function for the primary data in equation (7) by weighting each of the j^{th} contributions using the survey weight for the j^{th} observation in the MCS. The sample mean in equation (6) becomes a weighted average to account for unequal probability of selection for participants in the MCS.

We then apply BAYES to the trihalomethane data by sampling from the posterior density $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma} | data)$. We obtain a two different MCMC chains with overdispersed starting values and length 100 000 after 40 000 burn-in iterations. To illustrate sampler convergence, Figures 1 and 2 give density plots for the two different MCMC chains. Convergence of the parameters

β, ξ, γ is much better than for the bias parameters $\tilde{\xi}, \tilde{\gamma}$ because the solid curves match well with the broken curves. These results are somewhat expected because the model for the missing confounders is only weakly identifiable. As argued in Section 3.1.3, the variable U is unmeasured in the primary data, and there is little information to distinguish between the intercept ξ_0 and linear predictor $\tilde{\xi}^T g\{Z(U)\}$. Similarly, there are correlations between γ_0 and $\tilde{\gamma}^T U$ in the MCMC sampler. Slow mixing is also reported in other contexts using nonidentifiable models (Gelman et al. 2004; Gustafson 2005).

We argue that poorer mixing of the bias parameters has a modest impact on the overall model fit. To illustrate, the top right corner of Figure 2 gives density plots of the model deviance, given by $-2 \log \left[\prod_{i=1}^n P(Y_i, X_i | C_i) \times \prod_{j=1}^m P(Y_j, X_j | C_j, U_j) \right]$, and calculated at each MCMC iteration. The deviance is a measure of overall model fit with low values correspond to better fitting (Gelman et al. 2004). In the figure, the densities of deviance for the two chains are closely overlapping. Slow mixing of the bias parameters does not greatly affect model fit. Because the convergence of β, ξ and γ is satisfactory, we can compute summaries of the marginal posterior distributions of β, ξ and γ . Ideally, we could use longer MCMC runs, but this is computationally intensive.

Table 2 presents the results of applying BAYES to the trihalomethane data. It contains posterior means and 95% credible intervals for the exposure effect and covariate effects, adjusted for the seven missing confounders using the validation data. We see that the missing confounders have a sizable impact on estimation of the exposure effect. Compared to the NAIVE analysis, the association between trihalomethane exposure and full-term low birthweight is weaker with odds ratio 1.24 (0.93, 1.65). Thus the BAYES point estimate for the exposure effect β is shifted towards zero. This result make sense because in Table 3 when analyzing the validation data alone, we see that adjustment for either U_j or $Z(U_j)$ drives the estimate of β towards zero compared an analysis ignoring U_j .

The interval estimate for the exposure effect calculated from BAYES is wider than for NAIVE

(0.57 versus 0.46 on the log odds scale). This result seems puzzling at first because an analysis of the primary and validation data combined intuitively ought to yield less posterior uncertainty compared to an analysis of the primary data alone. However, the increased sample size of the combined analysis is balanced by the uncertainty in the stochastic imputation of the missing confounders. Furthermore, the NAIVE analysis ignores bias uncertainty from the missing confounders. It assumes that the bias parameters are precisely equal to zero. If there are missing confounders in the primary data, then the NAIVE interval estimates should be falsely precise. We study the frequentist coverage probability of BAYES and NAIVE interval estimates in Section 5.

Prior sensitivity presents possible challenges because of nonidentifiability, and we investigate whether the BAYES analysis results depend heavily on the prior distributions of Section 3.2. We repeat the analysis by fixing the prior variances equal to 10^3 rather than $(\log(15)/2)^2$. The resulting point and interval estimates for the quantities (β, ξ, γ) are almost identical to those in Table 2, with difference ≤ 0.03 on the log odds scale. Greater sensitivity is observed for the bias parameters but is difficult to assess because of the impact of slow MCMC mixing.

4.2 Maximum Likelihood Estimation to Adjust for Missing Confounders

For comparison, we also apply a method for adjusting for missing confounders that does not use Bayesian techniques. We call the method `FREQ`, meaning frequentist adjustment for missing confounders, and we define the method as follows: First, fit the regression models in equations (1) and (2) in the validation data alone and compute maximum likelihood estimates of the bias parameters $\tilde{\xi}$ and $\tilde{\gamma}$. This is straightforward because U_j is observed. Next, we substitute the point estimates into the likelihood function for the primary data, given in equation (7), and maximize it with respect to (β, ξ, γ) using the Newton Raphson algorithm. Standard errors for the estimated (β, ξ, γ) are obtained from the observed information matrix.

FREQ is a fast analogue of BAYES. The validation data is used as a source of information about bias parameters, but there is no joint analysis of multiple datasets. FREQ can also be used as a diagnostic procedure to determine if carrying out BAYES is worthwhile. If FREQ and NAIVE give markedly different inferences, then this indicates that confounding is important and BAYES should be pursued. Both methods use the same knots in the linear predictor $g\{\cdot\}$.

The results of applying FREQ to the data are given in the third column of Table 2. FREQ gives an odds ratio for the trihalomethane effect equal to 1.18 (0.94, 1.50). Comparing FREQ and BAYES results, we can see that inferences are similar in the sense that both estimates of the exposure effect are driven towards zero relative the NAIVE analysis. However, the BAYES interval estimate for the exposure effect β is wider. Furthermore, the FREQ and NAIVE interval estimate for β have roughly the same length on the log odds scale. While FREQ corrects for missing confounders using the same models as BAYES, the method ignores uncertainty in the bias parameters. In contrast, BAYES propagates uncertainty through the analysis. This impacts the length of the interval for β . For BAYES the width is 21% wider with length 0.57 versus 0.47.

5. The Performance of BAYES and FREQ in Synthetic Data

The previous analysis motivates questions about the performance of the BAYES in more general settings. For example, does regression adjustment for $Z(U)$ in lieu of U give unconfounded exposure effect estimates? Does modelling uncertainty in the bias parameters give meaningful improvement in the coverage probability of interval estimates? A further issue is the sample size m of the validation data. If m is small, then we may expect that BAYES and FREQ will break down because they fail to recover the marginal distribution of propensity scores in the source population. We explore these issues using simulations by analyzing synthetic datasets which contain confounding from multiple missing variables.

5.1 Simulation Design

We generate and analyze ensembles of 100 pairs of synthetic datasets, where each pair consists of primary data with $n = 1000$ and validation data with $m = 100, 250, 500$ or 1000 . We consider the case where there are four measured confounders and four additional missing confounders (thus C is a 5×1 vector and U is 4×1). Primary data ($n = 1000$) and validation data ($m = 100, 250, 500$ and 1000) are generated using the following algorithm: Simulate $\{C_i, C_j\}$ for $i \in 1 : n, j \in 1 : m$, and also $\{U_i, U_j\}$ for $i \in 1 : n, j \in 1 : m$, where each component of C_i, C_j, U_i, U_j is independent and identically distributed as a $N(0,1)$ random variable. Next, for fixed $\gamma_0, \dots, \gamma_4 = 0.1$, and $\tilde{\gamma}_1, \dots, \tilde{\gamma}_4 = 0.2$, simulate $\{X_i, X_j\}$ for $i \in 1 : n, j \in 1 : m$ using the logistic regression model of equation (2). Finally, for fixed $\beta = 0$, $\xi_0, \dots, \xi_4 = 0.1$ and $\tilde{\xi}_1, \dots, \tilde{\xi}_4 = 0.2$, simulate $\{Y_i, Y_j\}$ for $i \in 1 : n, j \in 1 : m$ for using the model $\text{Logit}[P(Y = 1|X, C, U)] = \beta X + \xi^T C + \tilde{\xi}^T U$. Note that the first component of C_i and C_j is equal to one so that γ_0 and ξ_0 are regression intercept terms. The choices for $\xi, \tilde{\xi}, \gamma, \tilde{\gamma}$ give odds ratios equal to $\exp(0.1)=1.1$ or $\exp(0.2)=1.2$. The results of Section 5.2 show that this combination of bias parameters produces a large amount of confounding. Fixing $\beta = 0$ models the setting of zero exposure effect.

We analyze the 100 pairs of datasets for each value of m by using BAYES, FREQ and NAIVE to obtain point and 80% interval estimates of the exposure effect β . Sampler convergence is assessed using separate trial MCMC runs.

5.2 Results

Figure 3 summarizes the performance of BAYES, FREQ and NAIVE analyses of the synthetic datasets. The upper panels quantify average bias and efficiency of point estimates, as a function of m the sample size of the validation data. The lower three panels give coverage and length of 80% interval estimates. In other words, for each point on the graph we have analyzed an independent collection of 100 pairs of data using either BAYES, FREQ or NAIVE.

For NAIVE, the estimates of β should perform poorly because the method ignores the missing confounders. In the top left panel, we see that the solid curve lies far from zero indicating that NAIVE estimates are badly biased. The solid curve is flat and does not depend on m because the NAIVE analysis ignores the validation data completely. Similarly, in the lower left panel the solid curve hovers below 60%, indicating that the coverage probability of NAIVE interval estimates is far below the nominal level of 80%.

In contrast, BAYES and FREQ eliminate bias from the missing confounders over a range of values for m . In the upper panels, we see that the solid curves hover near zero bias. BAYES and FREQ are essentially unbiased for all m under consideration. Summarizing the four missing confounders using the summary score $Z(U)$ appears to substantially reduce confounding. We do not consider the case where $m < 100$. The reason is because we found that the validation data contain little information about the bias parameters. Sampler convergence deteriorates and point estimates of $\tilde{\xi}, \tilde{\gamma}$ are highly variable. There is not enough information available to capture the joint distribution of the missing confounders.

The lower panels of Figure 3 summarize the performance of interval estimates for the exposure effect β . BAYES interval estimates have improved coverage probability compared to FREQ or NAIVE. As m tends to zero, the coverage probability of FREQ drops off more sharply compared to BAYES. Ignoring uncertainty in the bias parameters appears to adversely affect interval estimation when the validation data is small.

Interestingly, estimates of β calculated from BAYES are sometimes more efficient than FREQ or NAIVE. For large m , BAYES intervals are shorter than FREQ or NAIVE, despite the fact that they model uncertainty from missing confounders. BAYES point estimates of β also have smaller variance. Intuitively, the validation data contain information about β which is ignored by FREQ. This suggests that if it is reasonable to assume that the models in equations (1) and (2) are correct for both primary and validation data, then estimation may be improved by analysing the datasets together rather than separately.

As a final exercise, we study the performance of BAYES and FREQ when the conditional independence assumption between U and C is violated. Recall from Section 3.1.2 that we use this assumption to average over the distribution of the missing U in the primary data. In Appendix III of the supplementary materials we present further simulation results. We illustrate that the performance of BAYES and FREQ is insensitive to small departures from conditional independence typical of those in trihalomethane data example. But when U and C are strongly correlated, then BAYES and FREQ tend to overadjust for confounding. They become too conservative and their performance deteriorates. The reason is because adjusting for C_i has the effect of indirectly adjusting for bias from U_i because they are correlated with one another. Hence, adjusting for C in this case is sufficient and the NAIVE method does well. Similar results are given by Fewell et al. (2007) who show that when measured and unmeasured confounders are correlated this tends to reduced bias from the unmeasured confounding.

6. Discussion

In this article, we describe a flexible Bayesian procedure for adjusting for several missing confounders using external validation data. We summarize the missing variables using a scalar summary score $Z(U)$, which can be interpreted as the propensity score adjusted for measured confounders. Conditioning on $Z(U)$ breaks the association between X and U , within levels of C . To adjust for missing confounders, we need only update and adjust for $Z(U)$. Simulations illustrate that the method reduces bias from several missing confounders, provided that the sample size for the validation data is not too small ($m \geq 100$) and the missing confounders are not highly correlated with the measured ones. We recommend investigating the correlation between measured and missing confounders in order to identify possible conservativeness.

An alternative approach is to model the missing confounders directly. In the trihalomethane data example the components of U are primarily categorical. Hence one could imagine a general location model being useful for imputations in this context (D'Agostino and Rubin, 2000).

However, our objective is to show that we can avoid modelling the covariates, as this is particularly beneficial in data integration settings of high dimension. One challenge would be to characterize the benefit or harm of using BAYES in low dimensional problems. Lunceford and Davidian (2004) demonstrate that regression adjustment for the propensity score generally reduces confounding at the expense of efficiency when compared to multiple regression. Similar findings may hold in the present context, and in principle this could be explored through further simulations that are beyond the scope of this paper.

We apply BAYES to the trihalomethane data. The MCS was collected through disproportionately stratified sampling, and we used the sample survey weights to obtain approximate exchangeability of the primary and validation data (see Section 4.1). In the primary data, the exposure and outcome are associated, but this association is reduced upon adjustment for missing confounders. The validation data reveal bias from the missing covariates, including smoking and ethnicity. In particular, non-whites are at greater risk of the outcome and also have greater trihalomethane exposure. After adjusting for bias the effect estimate is attenuated towards zero. The FREQ analyses, which ignores uncertainty in the magnitude and direction of bias, gives interval estimates that are too narrow, but a similar bias correction.

In the trihalomethane data, the analysis results may be somewhat conservative because the correlations between U and C are mostly positive (see Appendix III). BAYES and FREQ assume that measured and missing confounders are uncorrelated, when in fact they do have some small positive association. Therefore they may slightly overadjust the exposure effect estimate towards zero.

A limitation of our analysis is that using area-level exposure estimates as substitutes for individual-level exposure measurements can introduce ecological bias. Nonetheless, Whitaker et al. (2005) show that the within-area variance of the exposure in our study population is less than the between-area variance. This suggests that any ecological bias is not overly large. Furthermore, our objective is to study tap water levels of exposures that are under regulatory

control, and therefore heterogeneity of exposure due to personal activities is less of a concern.

Acknowledgements: This work was supported by Economic and Social Research Council awards number RES-576-25-5003 and RES-576-25-0015. Lawrence McCandless started this work while at the Department of Epidemiology and Public Health, Imperial College London, UK. We are grateful to the Small Area Health Statistics Unit at Imperial College for access to the HES data and to the postcoded MCS data. We acknowledge in particular the help of Peter Hambly for processing the data bases. We would also like to thank Jassy Molitor and Mireille Toledano for useful discussion which helped us with the interpretation of the results.

References

- Breslow, N. E. and Hulobkov, R. (1997) “Weighted likelihood, pseudo-likelihood and maximum likelihood methods for logistic regression analysis of two-stage data,” *Statistics in Medicine*, 16, 103–116.
- Chatterjee, N., Chen, Y. H. and Breslow, N. E. (2003) “A pseudoscore estimator for regression problems with two-phase sampling,” *Journal of the American Statistical Association*, 98, 158–169.
- Fewell, Z., Smith, D. and Sterne, J. A. C. (2007) “The impact of residual and unmeasured confounding in epidemiologic studies: A simulation study,” *American Journal of Epidemiology*, 166, 646–656.
- Gelman, A., Carlin, J., Stern, H. and Rubin, D. (2004) *Bayesian Data Analysis, Second Edition*, New York: Chapman Hall/CRC.
- Greenland, S. (2005) “Multiple bias modelling for analysis of observational data (with discussion),” *Journal of the Royal Statistical Society Series A*, 168, 267–306.

- Gustafson, P. (2005) “On model expansion, model contraction, identifiability, and prior information: Two illustrative scenarios involving mismeasured variables,” *Statistical Science*, 20, 111-140.
- Jackson, C., Best, N. B. and Richardson, S. (2008) “Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors,” *Journal of the Royal Statistical Society Series A*, 171,159–178.
- Lin, D. Y., Psaty, B. M. and Kronmal, R. A. (1998) “Assessing the sensitivity of regression results to unmeasured confounders in observational studies,” *Biometrics*, 54, 948–963.
- Little, R. J. A. and Rubin, D. B. (2002) *Statistical Analysis with Missing Data, 2nd edition.*, New York: John Wiley.
- Lunceford, J. K. and Davidian, M. (2004) “Stratification and weighting via the propensity score in estimation of causal treatment effects: A comparative study,” *Statistics Medicine*, 23, 2937–60.
- McCandless, L. C., Gustafson, P. and Levy, A. R. (2007) “Bayesian sensitivity analysis for unmeasured confounding in observational studies,” *Statistics in Medicine*, 26, 2331–2347.
- McCandless, L. C., Gustafson, P. and Austin, P. C. (2009) “Bayesian propensity score analysis for observational data,” *Statistics in Medicine*, 28, 94–112.
- Molitor, N., Jackson, C., Best, N. B. and Richardson, S. (2009) “Using Bayesian graphical models to model biases in observational studies and to combine multiple data sources: Application to low birthweight and water disinfection by-products,” *Journal of the Royal Statistical Society Series A*, 172, 615–37.
- Rosenbaum, P. R. and Rubin, D. B. (1983a) “Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome,” *Journal of the Royal Statistical Society Series B*, 45, 212–218.

- Rosenbaum, P. R. and Rubin, D. B. (1983b) “The central role of the propensity score in observational studies for causal effects,” *Biometrika*, 70, 41–57.
- Rubin, D. B. (1985) “The use of propensity scores in applied Bayesian inference”. In *Bayesian Statistics 2*, eds. Bernardo J. M., De Groot M. H., Lindley D. V., Smith A. F. M. Valencia University Press: Valencia (1985), 463–72.
- Toledano, M. B., Nieuwenhuijsen, M. J., Best, N., 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.
- Schill, W. and Drescher, K. “Logistic analysis of studies with two-stage sampling: A comparison of four approaches,” *Statistics in Medicine*, 16, 117–132.
- Viallefont, V., Raftery, A. E. and Richardson, S. (2001) “Variable selection and Bayesian model averaging in case-control studies,” *Statistics in Medicine*, 20, 3215–3230.
- Wacholder, S. and Weinberg, C. R. (1994) “Flexible Maximum Likelihood Methods for Assessing Joint Effects in Case-Control Studies with Complex Sampling,” *Biometrics*, 50, 350–357.
- Wakefield, J. and Salway, R. (2001) “A statistical framework for ecological and aggregate studies,” *Journal of the Royal Statistical Society Series A*, 164, 119–137.
- Whitaker, H., Nieuwenhuijsen, M. J. and Best, N. (2003) “The relationship between water concentrations and individual uptake of chloroform: a simulation study,” *Environmental Health Perspectives*, 111, 688–694.
- Yin, L., Sundberg, R., Wang, X. and Rubin, D. B. “Control of confounding through secondary samples,” *Statistics in Medicine*, 25, 3814–3825.

Table 1: Characteristics of the primary data and validation data. Rows contain totals (percentages) for dichotomous variables and means \pm standard deviation for ordinal variables.

	<u>HES primary data $n = 8969$</u>		<u>MCS validation data $m = 824$</u>	
	<u>Trihalomethane exposure $\geq 60\mu\text{g}/L$</u>	<u>Trihalomethane exposure $< 60\mu\text{g}/L$</u>	<u>Trihalomethane exposure $\geq 60\mu\text{g}/L$</u>	<u>Trihalomethane exposure $< 60\mu\text{g}/L$</u>
<i>Variables in both HES and MCS</i>				
Full-tem low birthweight	164 (3.9)	142 (3.0)	14 (4.0)	9 (1.9)
Mother's age				
≤ 25	1361 (32)	1722 (36)	111 (32)	148 (31)
25 - 29	1188 (28)	1321 (28)	105 (30)	135 (28)
30 - 34	1063 (25)	1151 (24)	84 (24)	137 (29)
≥ 35	581 (14)	582 (12)	46 (13)	58 (12)
Male Baby	2177 (52)	2392 (50)	176 (51)	254 (53)
Carstairs quintile	4.1 ± 1.3	4.2 ± 1.2	4.0 ± 1.1	4.0 ± 1.2
<i>Variables MCS only</i>				
Lone parent family	.	.	72 (21)	105 (22)
Number of live children ≥ 0	.	.	13 (3.4)	14 (2.9)
Smoking during pregnancy	.	.	126 (36)	181 (38)
Non-white ethnicity	.	.	77 (22)	48 (10)
Alcohol during pregnancy	.	.	110 (32)	163 (34)
Body mass index $\geq 25\text{Kg}/\text{m}^2$.	.	85 (25)	134 (28)
Low education	.	.	138 (40)	210 (44)
<i>Total</i>	4193	4776	346	478

Table 2: Odds ratios (95% interval estimates) for the association between covariates and full-term low birthweight in the primary ($n = 8969$) and validation ($m = 824$) data combined.

Description	Odds ratio (95% interval estimate)		
	NAIVE*	BAYES**	FREQ**
Trihalomethanes > 60 μ g/L	1.39 (1.11, 1.75)	1.24 (0.93, 1.65)	1.18 (0.94, 1.50)
Mother's age			
≤ 25	1.14 (0.86, 1.51)	1.10 (0.82, 1.44)	1.14 (0.86, 1.52)
25 - 29 [†]	1.0	1.0	1.0
30 - 34	0.81 (0.57, 1.14)	0.73 (0.52, 1.02)	0.80 (0.56, 1.13)
≥ 35	1.10 (0.74, 1.64)	1.14 (0.77, 1.66)	1.11 (0.74, 1.66)
Male baby	0.76 (0.60, 0.95)	0.74 (0.59, 0.93)	0.75 (0.59, 0.95)
Carstairs quintile	1.37 (1.21, 1.55)	1.38 (1.22, 1.56)	1.37 (1.21, 1.56)

* Analysis ignores the missing confounders and validation data

** Adjusted for the seven missing confounders

[†] Reference group

Table 3: Odds ratios (95% interval estimates) describing the confounding induced by U in the validation data alone ($m = 824$). The lefthand column gives odds ratios for the association between Y and X adjusting for C . The middle column gives odds ratios adjusting for both C and U , whereas the rightmost column adjusts for C and the summary score $Z(U)$.

Description	Odds ratio (95% interval estimate) adjusting for		
	(X_j, C_j) only	(X_j, C_j) and U_j	(X_j, C_j) and $Z(U_j)$
Trihalomethane > 60 μ g/L	2.06 (0.87, 4.89)	1.75 (0.70, 4.34)	1.77 (0.73, 4.31)
Mother's age			
≤ 25	0.65 (0.24, 1.76)	0.52 (0.18, 1.50)	0.65 (0.24, 1.77)
25 - 29 [†]	1.0	1.0	1.0
30 - 34	0.13 (0.02, 1.06)	0.13 (0.02, 1.09)	0.14 (0.02, 1.11)
≥ 35	1.57 (0.50, 4.97)	1.60 (0.48, 5.29)	1.65 (0.51, 5.34)
Male baby	0.59 (0.25, 1.40)	0.61 (0.25, 1.48)	0.61 (0.26, 1.45)
Carstairs quintile	1.54 (0.79, 2.98)	1.41 (0.72, 2.78)	1.55 (0.80, 3.01)
Lone parent family	.	1.56 (0.59, 4.15)	.
Number of live children	.	1.80 (0.73, 4.43)	.
Smoking during pregnancy	.	2.86 (1.03, 7.93)	.
Non-white ethnicity	.	3.65 (0.87, 15.25)	.
Alcohol during pregnancy	.	1.76 (0.65, 4.74)	.
Body mass index ≥ 25 Kg/m ²	.	1.06 (0.39, 2.89)	.
Low education	.	1.32 (0.54, 3.20)	.

[†] Reference group

Figure 1: Posterior density estimates for the exposure effect β , and the covariate effects ξ and γ , based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

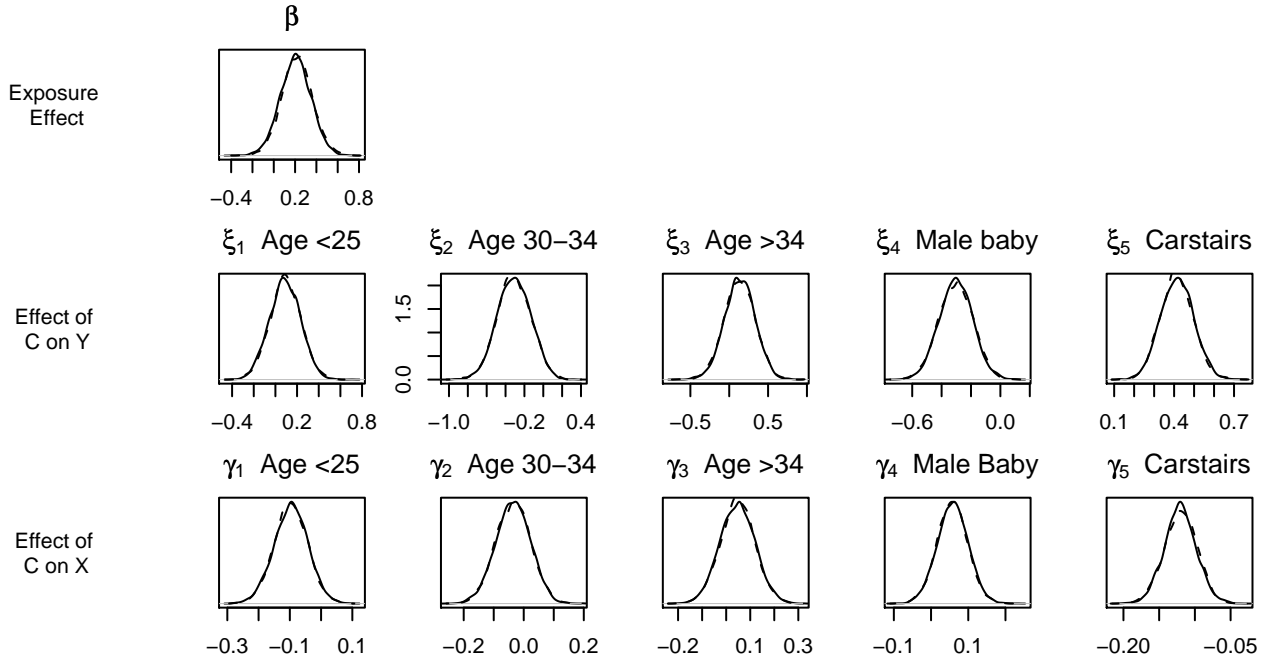


Figure 2: Posterior density estimates for the bias parameters ($\tilde{\xi}, \tilde{\gamma}$) and regression intercept terms (ξ_0, γ_0), based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

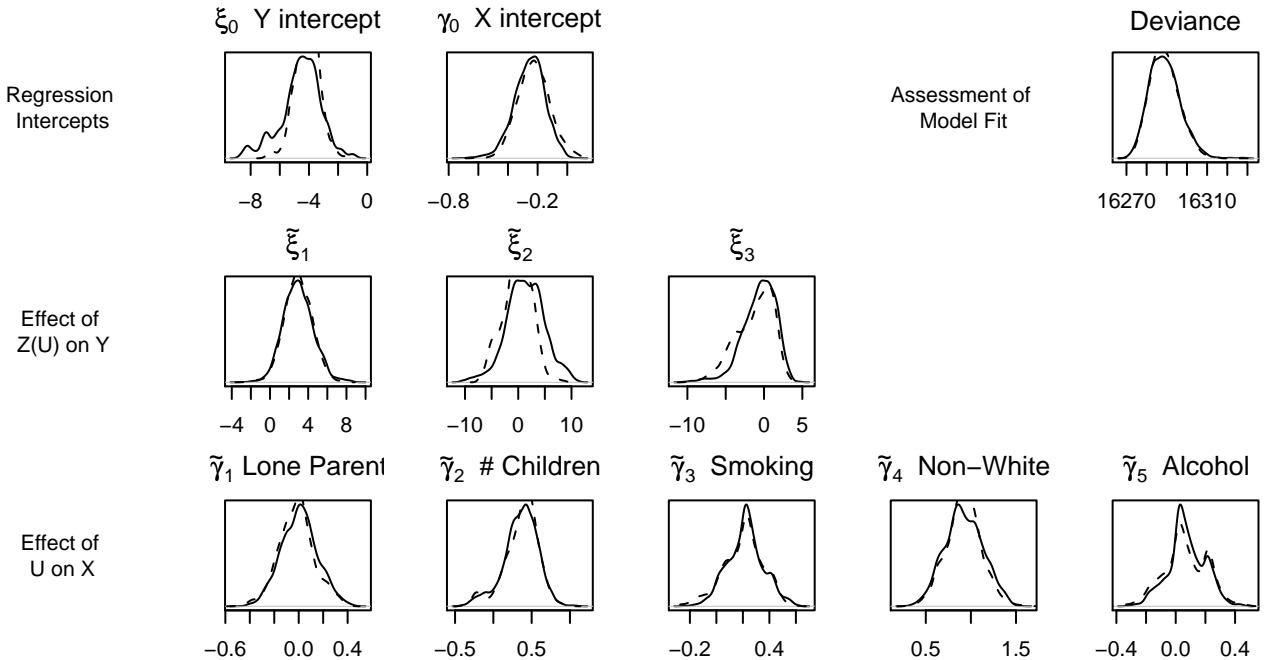


Figure 3: Performance of point and 80% interval estimates for the exposure effect β calculated using either BAYES, FREQ or NAIVE. The top panels describe the bias (solid line with simulation standard error (SE) < 0.02) and variance (broken line, SE < 0.01) of point estimates. The bottom panels describe the coverage probability (solid line, SE < 4%) and length (broken line, SE < 0.01) of 80% interval estimates.

