# Adjustment for Missing Confounders Using External Validation Data and Propensity Scores

Lawrence C. McCandless[1]
Sylvia Richardson[2]
Nicky Best[2]

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

August 2010

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 principle, 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 induces conditional independence between the exposure and the missing confounders given the measured confounders. It balances the unmeasured confounders across exposure groups, within levels of measured covariates. To adjust for bias, we need only model 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 to 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 (Little and Rubin 2002). This approach involves modelling 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 numerous, 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 secondary samples taken from the source population, or alternatively, population surveys such as census datasets. See for example, the paper of Stürmer, Schneeweiss, Avorn and Glynn (2005) from pharmacoepidemiology who use survey data with information on missing confounders to improve inferences in a large healthcare database study. 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 additional 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 confounders.

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.

We describe a novel 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 Rubin and Thomas (1996) and Lunceford and Davidian (2004) for an overview of propensity score techniques. Little and Rubin (2002) review missing data imputation techniques using probabilities of selection.

In the present investigation, the focus is somewhat different from standard applications of

propensity scores. We assume that some but not all confounders are measured. To adjust for the *missing* confounders, we summarize them using a scalar summary score, which can be interpreted as the propensity score conditional on *measured* confounders. The score has the property that it induces conditional independence between the exposure and missing confounders, given measured confounders. It balances the missing confounders across exposure groups. To adjust for missing confounders in the primary data, we need only adjust for the summary score. Our approach is to first specify a joint model for the primary and validation data with the propensity score as a missing covariate. We then integrate the propensity score out of the likelihood function for the primary data during Markov chain Monte Carlo simulation. Modelling the distribution of the missing confounders is not required.

To illustrate the problem of missing confounders, Section 2 describes 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 Bayesian method to adjust for missing confounders using propensity scores. 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. Additionally, we contrast the results with those obtained using propensity score calibration (PSC) (Stürmer, Schneeweiss, Avorn and Glynn 2005). PSC is an alternative method to adjust for several missing confounders using propensity scores and external validation data. In Section 5, we present simulation results that study performance over a range of sample sizes for the

validation data. Section 6 concludes with a discussion.

## 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 show 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). 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 8780 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 is 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 details. The outcome under study is 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 $60\mu$g/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 ($\leq$ 25, 25-29, 30-34, $\geq$ 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 3.8% of births versus 3.1% 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 1.32 with 95% interval estimate (1.04, 1.68) 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 three distinct strata determined by neighborhood income and ethnicity. Hence, inferences from the MCS data must adjust for non-random sampling. In regression analysis, this is accomplished by including strata-specific intercepts in the model that are indicator variables for each of the sampling strata. 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 8780 births in the same region during 2000-2001. Thus the primary data has sample size $n = 7956$, while the validation data where more complete 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 children living outside of the home, maternal smoking, alcohol consumption, body mass index prior to pregnancy $\geq 25$ kg/m$^2$, 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.

Denote the primary data as $\{(Y_i, X_i, C_i, U_i)|$ for $i \in 1 : n = 7956\}$ and the validation data as $\{(Y_j, X_j, C_j, U_j)|$ 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 logistic regression of $Y_j$ on $X_j$ and $C_j$. To account for

the non-random sampling of the MCS data, we include strata-specific intercepts to indicate the sampling strata. 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 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 $U_j$. Together with epidemiological knowledge, this suggests that the missing $U_i$ may confound the association between $X_i$ and $Y_i$ in the primary data. However, it is difficult to hazard a guess of how much confounding $U_i$ might induce.

One way to study the effect of missing confounders in the primary data is to do a sensitivity analysis. In Table 4, we present results using the method of Rosenbaum and Rubin (1983a), which assumes that there is a single binary unmeasured confounder. Table 4 gives odds ratios for the association between $X_i$ and $Y_i$, conditional on $C_i$, under various assumptions about the magnitude of confounding. The analysis assumes that the effect of the missing confounder is homogeneous across levels of $X_i$ and $C_i$. Point estimates and standard errors are computed via maximum likelihood. See Rosenbaum and Rubin (1983a), and McCandless et al. (2007) for computational details. Note that because the confounder is unmeasured, and the labelling is arbitrary, we may assume without loss of generality that it increases the probability of exposure. Table 4 illustrates that a single binary confounder that doubles the odds of the exposure and outcome would suffice to eliminate the association between $X_i$ and $Y_i$ seen in the NAIVE analysis of Table 2.

# 3. Bayesian Adjustment for Missing Confounders using Propensity Scores (BayesPS)

We present a Bayesian method to adjust for missing confounders using external validation data and propensity scores, which we henceforth call by the acronym BayesPS. In Section 3.1, we introduce the propensity score conditional on measured confounders, and we illustrate that it can be used to adjust for missing confounders. We obtain likelihood functions for the primary

and validation data, integrating over the distribution of the missing propensity score. 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 Exposure and Outcome Variables*

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:

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

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

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

Equation (1) is a model for the outcome and includes an exposure effect parameter $\beta$ and a linear term for the covariates $C$ with regression coefficients $\xi = (\xi_0, \ldots, \xi_p)$. Equation (2) models the probability of exposure, which depends on the measured and missing confounders via the regression coefficients $\gamma = (\gamma_0, \ldots, \gamma_p)$ and $\tilde{\gamma} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_q)$. 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 $Z = \tilde{\gamma}^T U$ is a scalar summary of $U$. We define $Z$ as the propensity score conditional on $C$, or for the sake of brevity, *we say that $Z$ is the propensity score*. The quantity $Z$ is not a propensity score in the usual sense because it lies on the log odds scale. However, if we condition on $C$, then the quantity $\gamma^T C$ is a constant intercept term in equation (2), and variability in $X$ due to $U$ is mediated entirely through $Z$.

From equation (2) we have that

$$X \perp\!\!\!\perp U|C, Z. \tag{3}$$

This result is analogous to the conclusion of Theorem 1 of Rosenbaum and Rubin (1983b). It states that within levels of $C$, conditioning on $Z$ forces independence between $X$ and $U$. The quantity $Z$ functions as a standard propensity score, in the sense that it balances the distribution of confounders between exposure groups. However, the key difference is that $Z$ balances the *missing* confounders $U$, conditional on $C$. In Appendix I, 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 $(C, Z)$. This means that exposure effect measures computed from the marginal density $P(Y|X, C, Z)$ 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 $Y \perp\!\!\!\perp U|X, C, Z$. Thus we can assume that the unmeasured confounders enter into the conditional expectation of $Y$ only as a function of $Z = \gamma^T U$. For discussion of Bayesian regression adjustment for the propensity score, see Rubin (1985).

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

The preceding discussion is valid only if $Z$ is known. However the regression coefficients $\tilde{\gamma}$ are not known, and if we multiply the wrong coefficients by $U$ then none of the conditional independence assumptions hold. The parameter $\tilde{\gamma}$ is estimated simultaneously with $(\beta, \gamma, \xi, \tilde{\xi})$ in a full Bayesian analysis. Consequently, when fitting equations (1) and (2), we average over

the posterior distribution of $\tilde{\gamma}$. Provided that all models are correctly specified, then the posterior will tend to concentrate at the true value of $\tilde{\gamma}$ as we collect more data, hence providing increasingly better control of confounding.

We note that we could alternatively control confounding from $C$ and $U$ by fitting the outcome model

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

where $\tilde{Z} = \text{Logit}[P(X=1|U,C)] = \gamma^T C + \tilde{\gamma}^T U$. 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 propensity score $Z$. This is appropriate in the present context because $C$ is measured while $U$ is not. Another approach to control confounding from $U$ would be to include the covariates and propensity score in the outcome model simultaneously, in the spirit of double robustness (Lunceford and Davidian 2004).

### *3.1.2 The Resulting Model when U is Missing*

In the primary data, the quantities $Y, X$ and $C$ are observed while $U$ 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

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

where $P(Y|X,C,U)$ and $P(X|U,C)$ are given in equations (1) and (2). To complete the specification, equation (5) requires a model for $U$ given $C$, which can be difficult to specify if $U$ is high dimensional with complex correlation structure. However, this can be circumvented by

using propensity scores because only $Z$ is needed to fit equations (1) and (2). We do not need to model $U$ directly in order to estimate the exposure effect. Instead we propose to model the missing one-dimensional $Z$.

One simplification is to assume that $C$ and $U$ are marginally independent, meaning that $P(U|C) = P(U)$. In this case, we propose to model the missing $Z$ in the primary data by using the empirical distribution of the propensity scores in the validation data, which are given by $\{Z_j = \tilde{\gamma}^T U_j | j \in 1 : m\}$. This permits us to compute equation (5) using the approximation

$$P(Y, X|C) \approx \frac{1}{m} \sum_{j=1}^{m} \left[ \frac{\exp\{Y(\beta X + \xi^T C + \tilde{\xi}^T g\{Z_j\})\}}{1 + \exp\{\beta X + \xi^T C + \tilde{\xi}^T g\{Z_j\}\}} \right] \left[ \frac{\exp\{X(\gamma^T C + Z_j)\}}{1 + \exp\{\gamma^T C + Z_j\}} \right]. \quad (6)$$

Alternatively, we can use a quadrature estimate

$$P(Y, X|C) \approx \sum_{k=1}^{M} \omega_k \left[ \frac{\exp\{Y(\beta X + \xi^T C + \tilde{\xi}^T g\{\hat{Z}_k\})\}}{1 + \exp\{\beta X + \xi^T C + \tilde{\xi}^T g\{\hat{Z}_k\}\}} \right] \left[ \frac{\exp\{X(\gamma^T C + \hat{Z}_k)\}}{1 + \exp\{\gamma^T C + \hat{Z}_k\}} \right], \quad (7)$$

based on a histogram of the empirical distribution of the propensity scores in the validation data. Here the index $k$ equals $1, \ldots, M$, where $M$ is the number of histogram bins. The quantities $\hat{Z}_k$ are the interval midpoints in the histogram and $\omega_k$ are the bin frequencies. In applications, we find equation (7) faster to compute than equation (6) because it is a summation of size $M$ rather than size $m$ and typically $M << m$.

An advantage of using equation (7) as a model for the primary data is that it requires no parametric assumptions for the nuisance distribution of $U$. It can be used regardless of the correlation structure of the components of $U$ and with continuous and categorical variables. However, a disadvantage is that it assumes that $U$ and $C$ are marginally independent and this may not be plausible in some settings. For example, in the trihalomethane data we might expect that mothers living in more deprived areas may be more likely to smoke or head lone parent families.

A different strategy is to model the distribution of $Z$ given $C$. The quantity $Z$ is a missing continuous covariate in the regression model for the primary data, and we can use a general

location model (Little and Rubin 2002) to assigns a linear regression model

$$Z|C \;\sim\; N\Big(\hat{\theta}^T C, \hat{\sigma}^2\Big). \tag{8}$$

The linear predictor $\hat{\theta}^T C = \hat{\theta}_0 + \hat{\theta}_1 C_1 + \ldots + \hat{\theta}_p C_p$ is the estimated mean propensity score conditional on $C$ with variance $\hat{\sigma}^2$. The estimates $(\hat{\theta}, \hat{\sigma}^2)$ are obtained from a preliminary analysis of the validation data. Specifically, we regress $\hat{\tilde{\gamma}}^T U_j$ onto $C_j$, where $\hat{\tilde{\gamma}}$ is the maximum likelihood estimate of $\tilde{\gamma}$ computed by fitting equation (2) to the validation data alone. By using equation (8), we can then obtain $P(Y, X|C)$ for the primary data by using equation (7) and setting $(\hat{Z}_k, \omega_k)$ for $k = 1, 2, ..., M$ as a histogram approximation to the normal distribution in equation (8). In the trihalomethane data, we set $M = 7$ with $\hat{Z}_1, \ldots, \hat{Z}_7$ equal to $q \times \hat{\sigma} + \hat{\theta}^T C$, and additionally, $\omega_1, \ldots, \omega_7$ equal to $\phi(q) / \sum_{q=-3}^{3} \phi(q)$ for $q = -3, -2, -1, 0, 1, 2, 3$, where $\phi()$ is the probability density function of a standard normal.

The motivation behind equation (8) is that we do not need to know the quantity $U$ in order to control confounding in the primary data. It suffices to use a missing data model for the missing propensity score $Z$. A priori, using a Gaussian model to approximate the distribution of the missing propensity scores is plausible because the quantity $Z$ lies on the log odds scale. Alternatively, we could follow the recommendation of Little and Rubin (2002) and assign a more flexible distribution such as the t-distribution. Equation (8) uses plug-in point estimates and does not incorporate uncertainty in $(\hat{\theta}, \hat{\sigma}^2)$. Although it would be desirable to estimate $(\hat{\theta}, \hat{\sigma}^2)$ jointly with other model parameters during MCMC, this is computationally demanding. The quantity $U$ is unobserved in the primary data, and updating $(\theta, \sigma^2)$ leads to slow convergence and increased computational time. The plug-in estimates of equation (8) are easy to implement, although less rigorous and they may lead to smaller interval estimates for model parameters. See Little and Rubin (2002; Section 10.2.3) for discussion of using plug-in estimates in missing data models.

Alternatively, we could model the distribution of $U$ directly. This is feasible in the trihalomethane data example because the components of $U$ are primarily categorical. For the case

of two missing confounders, this is explored by Molitor et al. (2009). However, only $Z$ is needed to estimate the exposure effect, and our intent is to avoid a model for $U$. Full Bayesian updating and monitoring convergence of $U$ is computationally expensive, even for moderate dimensional $U$.

## *3.2 Prior Distributions*

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

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

This models the belief that the odds ratio for the exposure effect $\beta$ is not overly large and lies between 1/15 and 15 with probability 95%. These priors make similar assumptions about the association between $Y$ and $(C, Z)$ given $X$, and also the association between $X$ and $(C, U)$. Such priors are plausible and capture the magnitude and direction of effect estimates in typical epidemiologic investigations (Greenland, 2005). 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 BayesPS are obtained

from the posterior density $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}|data)$, which we sample from using MCMC. We have

$$\begin{aligned} P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}|data) &\propto \left\{ \prod_{i=1}^{n} P(Y_i, X_i|C_i) \right\} \times \left\{ \prod_{j=1}^{m} P(Y_j, X_j|C_j, U_j) \right\} \times P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}) \\ &\approx \prod_{i=1}^{n} \left\{ \sum_{k=1}^{M} \omega_k \left[ \frac{\exp\{Y_i(\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{\hat{Z}_k\})\}}{1 + \exp\{\beta X_i + \xi^T C_i + \tilde{\xi}^T g\{\hat{Z}_k\}\}} \right] \right. \\ &\quad \left. \times \left[ \frac{\exp\{X_i(\gamma^T C_i + \hat{Z}_k)\}}{1 + \exp\{\gamma^T C_i + \hat{Z}_k\}} \right] \right\} \\ &\quad \times \prod_{j=1}^{m} \left\{ \left[ \frac{\exp(Y_j(\beta X_j + \xi^T C_j + \tilde{\xi}^T g\{\tilde{\gamma}^T U_j)\}))}{1 + \exp(\beta X_j + \xi^T C_j + \tilde{\xi}^T g\{\tilde{\gamma}^T U_j\})} \right] \right. \\ &\quad \left. \times \left[ \frac{\exp(X_j(\gamma^T C_j + \tilde{\gamma}^T U_j))}{1 + \exp(\gamma^T C_j + \tilde{\gamma}^T U_j)} \right] \right\} \times P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}) \end{aligned} \quad (9)$$

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 $\beta, \xi_0, \ldots, \xi_p, \tilde{\xi}_0, \ldots \tilde{\xi}_l$ 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).

# 4. Analysis Results for the Trihalomethane Data

We use BayesPS to adjust for missing confounders in the trihalomethane data. In Section 4.1, we use the empirical distribution approach, described in Section 3.1.2, which assumes that $U$ and $C$ are marginally independent (i.e. $P(U|C) = P(U)$). Then in Section 4.2, we study the independence assumption and apply BayesPS using the Gaussian regression model for the missing propensity scores in equation (8). Finally, in Section 4.3, we contrast BayesPS with propensity score calibration (PSC) (Stürmer et al. 2005). PSC is a method that also uses propensity scores and external validation data in order to adjust for missing confounders. PSC is based on regression calibration. It assumes a linear measurement error model for the relation-

ship between the missing propensity score and an error prone propensity score. PSC proceeds in two stages. First, the measurement error model is estimated in the validation data, and then the fitted model is used to impute the missing propensity score. See Stürmer et al. (2005) for a detailed discussion.

## *4.1 BayesPS assuming U and C are marginally independent*

Before applying BayesPS to the trihalomethane data, we set a priori values for the knots used to define the linear predictor $g\{.\}$ in equation (1). Following McCandless et al. (2009), we fit the logistic regression model given in equation (2) to the validation data using maximum likelihood to estimate the parameter $\tilde{\gamma}$. The quantities $\hat{Z}_j$, computed by evaluating $\{\hat{Z}_j = \hat{\tilde{\gamma}}^T U_j | j \in 1:m\}$, 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 $Z$.

Table 3 gives a preliminary illustration of $Z$ as a tool to adjust for missing confounders in the validation data. In the rightmost column we fit the model in equation (1) using $\hat{Z}_j$ to control for confounding, rather than using $U_j$. 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. This illustrates that we can use $\hat{Z}_j$ as a covariate to control confounding in the validation data, rather than including the entire covariate vector $U_j$ in the model for $Y_j$.

We now fit BayesPS to the primary and validation data combined. We assume that $P(U|C) = P(U)$, meaning that $U$ and $C$ are marginally independent, and we use the empirical distribution of the propensity scores in the validation data as a model for the missing propensity scores in the primary data. 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 functions of the primary and validation data. Following Molitor et al. (2009), we include strata-specific intercepts when fitting equations (1) and (2) to the MCS data. These intercepts enter into equation (9) as indicator variables that

are added to the linear predictors for the likelihood contributions of the validation data. This eliminates bias from non-random sampling of the MCS data, and ensures that inferences for $(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma})$ have valid uncertainty assessments.

We then apply BayesPS to the trihalomethane data by sampling from the posterior density $P(\beta, \xi, \tilde{\xi}, \gamma, \tilde{\gamma}|data)$. We obtain 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 $\beta, \xi, \gamma$ is better than for $\tilde{\xi}, \tilde{\gamma}$ because the solid curves in Figure 1 match well with the broken curves. These results are somewhat expected because the model for the missing confounders is only weakly identifiable. The primary data contain no information with which to inform the parameters $\tilde{\xi}, \tilde{\gamma}$, which determine the magnitude of confounding from $U$. Slow mixing is also reported in other contexts using nonidentifiable models (Little and Rubin 2004).

We argue that poorer mixing 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 corresponding to better fitting. In the figure, the densities of deviance for the two chains are closely overlapping. Slow mixing of $\tilde{\xi}, \tilde{\gamma}$ does not greatly affect model fit. Because the convergence of $\beta$, $\xi$ and $\gamma$ is satisfactory, we can compute summaries of the marginal posterior distributions of $\beta$, $\xi$ and $\gamma$. Ideally, we could use longer MCMC runs, but this is computationally intensive.

The second column in Table 2 presents the results of applying BayesPS to the trihalomethane data in the case where $P(U|C) = P(U)$. 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 1.20 (0.91, 1.63). Thus the

BayesPS point estimate for the exposure effect $\beta$ 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 $Z_j$ drives the estimate of $\beta$ towards zero compared to an analysis ignoring $U_j$.

The interval estimate for the exposure effect calculated from BayesPS is wider than for NAIVE (0.58 versus 0.49 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 magnitude and direction of bias from the missing confounders. Furthermore, the NAIVE analysis ignores bias uncertainty from the missing confounders. 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 BayesPS and NAIVE interval estimates in Section 5.

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

## *4.2 BayesPS assuming $U$ and $C$ are not marginally independent.*

The assumption that $U$ and $C$ are independent seems questionable in the trihalomethane data. For example, mothers with living in deprived areas may be more likely to smoke and drink alcohol during pregnancy. To study the independence assumption, we follow the latter part of Section 3.1.2 and assign a general location model for the missing propensity scores in the primary data using equation (8). First, we estimate $\tilde{\gamma}$ by fitting equation (2) to the validation data alone,

and calculate the estimated propensity scores, given by $\{\hat{Z}_j = \hat{\tilde{\gamma}}^T U_j | j = 1, \ldots, m\}$. Then we fit linear regression of $\hat{Z}_j$ on $C_j$ in order to calculate $(\hat{\theta}, \hat{\sigma}^2)$ used in equation (8). We include strata-specific intercepts in the regression models to account for non-random sampling of the validation data. We obtain $\hat{\theta} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_5) = (0.13, -0.02, -0.04, -0.09, -0.00, -0.00)$, with standard errors for each component roughly equal to 0.03. Additionally, $\hat{\sigma}^2 = 0.33$, and the adjusted $R^2$ is 0.05. To help with interpreting these results, recall that $\theta$ is a vector of length $p+1$, with components that correspond to each of the $p+1$ components of $C$ listed in Table 2. The quantity $\theta_0$ is the y-intercept, and $\theta_1, \ldots, \theta_p$ govern how the mean propensity score depends on each component of $C$.

Using $(\hat{\theta}, \hat{\sigma}^2)$ we apply BayesPS using the methodology described in Section 3.1.2. The results are presented in the third column of Table 2 under the heading "$P(U|C) \neq P(U)$". We see an odds ratio for the exposure effect of 1.23 (0.94, 1.64) that is shifted towards 1.0 compared to the NAIVE analysis, and which is very similar to BayesPS assuming $P(U|C) = P(U)$. Table 2 indicates that incorporating information about the dependence between $U$ and $C$ has only a modest effect on the exposure effect estimate. This is understandable because the adjusted $R^2$ is small and most of the variation in the propensity scores is not attributable to $C$. Using equation (8) to implement BayesPS roughly amounts to using a Gaussian approximation to the empirical distribution of the propensity scores. Consequently, the results in Table 2 are not greatly sensitive to the independence assumption. Note that because we are plugging in point estimates $(\hat{\theta}, \hat{\sigma}^2)$ into the analysis, this means that the interval estimates may be slightly too narrow. See Little and Rubin (2002) for details.

## *4.3 Comparison with Stürmer's Propensity Score Calibration (PSC)*

We contrast our results to those obtained using propensity score calibration (PSC), which is an alternative method to adjust for missing confounders that also uses propensity scores and external validation data. PSC assumes a linear measurement error model for the relationship

between the missing propensity score and the *error prone propensity score*, which Stürmer et al. (2005) define as the quantity $P(X = 1|C)$. To implement PSC, one must first estimate the measurement error model in the validation data. Next, the fitted model is used to impute the missing propensity score for each study unit in the primary data by using the error prone propensity score. If the measurement error model is correct, then the imputed propensity score can be included as a covariate in the regression model for the outcome in order to adjust for the missing confounders. Further details describing PSC are given in Appendix II. See also Stürmer et al. (2005) for complete discussion of PSC.

We implement PSC exactly as described in Stürmer et al. (2005), except that we use a logistic regression model for the outcome of full-term low birthweight, rather than the Cox proportional hazards model described in the paper. Interval estimates are computed using the bootstrap. The results are given in the fourth column of Table 2. We obtain an odds ratio for the exposure effect of 2.36 (1.36, 7.16), which differs substantially from the results of NAIVE and BayesPS. PSC depends on a measurement error surrogacy assumption that requires conditional independence between the outcome variable and the error prone propensity score given the true propensity score. Using the likelihood ratio test described by Stürmer et al. (2007), we obtain a p-value of 0.27 indicating that surrogacy does not hold. Thus PSC does not appear to be suitable in the trihalomethane data. Note however that in the simulations of Section 5 we show that PSC can perform nearly as well as BayesPS in some settings. See also Appendix II for detailed comparisons with PSC. We emphasize that PSC ignores the outcome variable in the validation data when adjusting for missing confounders. Thus even when the model assumptions of PSC are correct, it may still perform less favorably than BayesPS.

# 5. The Performance of BayesPS in Synthetic Data

The trihalomethane analysis motivates questions about the performance of BayesPS in more general settings. For example, does regression adjustment for $Z$ in lieu of $U$ give unconfounded

exposure effect estimates? How might BayesPS compare to a gold standard analysis of the primary data that adjusts for all the missing confounders directly? A further issue is the sample size $m$ of the validation data. If $m$ is small, then we may expect that BayesPS will break down because it fails to recover the distribution of propensity scores in the source population. We explore these issues using simulations by analyzing synthetic datasets which contain confounding from multiple missing covariates.

## *5.1 Simulation Design*

We generate and analyze ensembles of 200 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 $5 \times 1$ vector including intercept, and $U$ is $4 \times 1$). 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 $\gamma_0, \ldots, \gamma_4 = 0.1$, and $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_4 = 0.2$, 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 $\beta = 0$, $\xi_0, \ldots, \xi_4 = 0.1$ and $\tilde{\xi}_1, \ldots, \tilde{\xi}_4 = 0.2$, simulate $\{Y_i, Y_j\}$ for $i \in 1 : n, j \in 1 : m$ using the outcome model

$$\text{Logit}[P(Y = 1|X, C, U)] = \beta X + \xi^T C + \tilde{\xi}^T U. \tag{10}$$

Note that the first component of $C_i$ and $C_j$ is equal to one so that $\gamma_0$ and $\xi_0$ 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 choice of parameters produces a large amount of confounding. Fixing $\beta = 0$ models the setting of zero exposure effect.

For each value of $m$, we analyze the 200 pairs of datasets using BayesPS, NAIVE and PSC to obtain point and 80% interval estimates of the exposure effect $\beta$. Because the components of

$U$ and $C$ are generated as independent unit normals, we apply the BayesPS method assuming that $P(U|C) = P(U)$ and by using the empirical distribution approach described in Section 3.1.2. Sampler convergence is assessed using separate trial MCMC runs. In addition, because the quantities $U_i$ for $i \in 1:n$ are known by construction, we also apply a method called GOLD, which involves fitting equation (10) to the primary data using $(Y_i, X_i, C_i, U_i)$, and ignoring the validation data altogether. Thus GOLD is a gold standard method for the best case scenario when all the confounders are observed.

## *5.2 Results*

Figure 3 summarizes the performance of BayesPS, NAIVE, PSC and GOLD analyses of the synthetic datasets. The top panels quantify bias and variance of point estimates, averaged over the simulation runs, and as a function of $m$ the sample size of the validation data. The lower panels give coverage and average length of 80% interval estimates. In other words, for each data point on the graphs we analyzed an independent collection of 200 pairs of primary/validation data using BayesPS, NAIVE, PSC and GOLD, and then we averaged the results.

For NAIVE, the estimates of $\beta$ should perform poorly because the method ignores the missing confounders. In the top left panel, we see that the dotted curve lies far from zero indicating that NAIVE estimates are badly biased. The dotted 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 dotted curve hovers at 50%, indicating that the coverage probability of NAIVE interval estimates for the exposure effect $\beta$ is far below the nominal level of 80%. In comparison, we see that GOLD estimates, which are denoted by solid curves, give better inferences for $\beta$.

Figure 3 illustrates that BayesPS eliminates bias from the missing confounders over a range of values for $m$. In the upper left panel, we see the dashed curves lies near zero bias. BayesPS estimates of $\beta$ are essentially unbiased for all $m$ under consideration. Summarizing the four missing confounders using the summary score $Z$ appears to substantially reduce confounding.

We do not consider the case where $m < 100$. The reason is because there is not enough information in the validation data to estimate the empirical distribution of the missing propensity scores. Sampler convergence deteriorates and point estimates are highly variable. The lower left panel of Figure 3 summarizes the performance of interval estimates for the exposure effect $\beta$. BayesPS interval estimates have improved coverage probability compared to NAIVE, although the performance deteriorates for small $m$. Note that the simulation standard errors for the coverage probability estimates are approximately $\pm\sqrt{0.8 \times 0.2/200} = \pm 3\%$. Additional simulation runs are desirable but computationally expensive.

One interesting observation is that BayesPS estimates of $\beta$ are sometimes more efficient than either NAIVE or GOLD. Additionally, for large $m$, BayesPS intervals are shorter than NAIVE or GOLD, despite the fact that they acknowledge uncertainty from missing confounders. This is perhaps expected because BayesPS incorporates the validation data into the analysis, while NAIVE and GOLD ignore it. But it nonetheless illustrates that incorporating external information about confounding can actually *increase* precision of the exposure effect estimates. 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 analyzing the datasets together rather than separately.

Unlike in the trihalomethane data example, PSC also reduces bias from missing confounders. In the top left panel, we see it eliminates bias over the range of $m$ under consideration. The point estimates are slightly less efficient than BayesPS, particularly for $m$. However, this is understandable because PSC ignores the outcome variable in the validation data, and is therefore disadvantaged compared to BayesPS. Interval estimates tend to be wider and give greater than nominal-level coverage. It should be noted that the performance of PSC depends on a measurement error surrogacy assumption described in detail by Stürmer et al. (2005). In Appendix II, we show that PSC deteriorates compared to BayesPS through small modifications of the simulation design.

# 6. Discussion

In this article, we describe a Bayesian procedure for adjusting for several missing confounders using external validation data. We summarize the missing variables using a scalar summary score $Z$, which can be interpreted as the propensity score conditional on measured confounders. Conditioning on $Z$ breaks the association between $X$ and $U$, within levels of $C$. To adjust for missing confounders, we need only adjust for $Z$. Simulations illustrate that BayesPS reduces bias from several missing confounders, provided that the sample size for the validation data is not too small ($m \geq 100$).

The primary and validation data should be drawn from the same population to ensure that the regression models are estimating the same quantities. For the trihalomethane data example, the validation data were collected through disproportionately stratified sampling, and we incorporated strata-specific intercepts into the regression models to obtain approximate exchangeability of the primary and validation data. Additionally, both datasets must contain the common set of variables $Y$, $X$ and $C$ that are measured in the same fashion. We note that valid causal inference is contingent on the assumption that there are no *additional* unmeasured confounders beyond those recorded in the validation data.

We propose two strategies for modelling the distribution of the missing propensity scores in the primary data. The first approach can be used if $U$ and $C$ are marginally independent, which can be assessed in the validation data. It models the missing $Z$ by using the empirical distribution of the propensity scores in the validation data. The second approach can be used if the assumption of independence between $U$ and $C$ does not hold. It treats $Z$ as a missing covariate and assigns a general location model for $Z$ (Little and Rubin 2002). In the trihalomethane data, the exposure effect estimate is insensitive to the independence assumption, however we should not expect that this will always be the case. When $U$ and $C$ are strongly correlated then this will tend to reduce the amount of confounding from $U$. The reason is because adjusting for $C$

in the primary data has the effect of adjusting for $U$ since they are correlated with one another. Ignoring the correlations between $U$ and $C$ will cause BayesPS to overadjust for $U$. Similar findings are described by Fewell et al. (2007) who show that when measured and unmeasured confounders are correlated this tends to reduced bias from unmeasured confounding.

An alternative approach is to model the missing $U$ directly. For the trihalomethane data, this is feasible because there are 7 missing confounders that are mostly categorical. See Molitor et al. (2009) for discussion of the case of two missing dichotomous confounders. However, because only $Z$ is needed to estimate the exposure effect, our intent is to build a procedure that does not require a model for $U$. Furthermore, full Bayesian updating of $U$ and monitoring convergence is computationally challenging even for the case of 7 missing confounders as in the present example.

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.

# 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.

Greenland, S. (2005) "Multiple bias modelling for analysis of observational data (with discussion)," *Journal of the Royal Statistical Society Series A*, 168, 267–306.

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.

Little, R. J. A. and An, H. (2004) "Robust likelihood-based analysis of multivariate data with missing values," *Statistica Sinica*, 14, 949–968.

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, 615–637.

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.

Rubin, D. B. and Thomas, N. (1996) "Matching using estimated propensity scores: relating theory to practice," *Biometrics*, 52, 249–264.

Schill, W. and Drescher, K. (1997) "Logistic analysis of studies with two-stage sampling: A comparison of four approaches," *Statistics in Medicine*, 16, 117–132.

Stürmer, T., Schneeweiss, S., Avorn, J. and Glynn, R. J. (2005) "Adjusting effect estimates for unmeasured confounding with validation data using propensity score calibration," *American Journal of Epidemiology*, 162, 279–289.

Stürmer, T., Schneeweiss, S., Rothman, K. J., Avorn, J. and Glynn, R. J. (2007) "Performance of Propensity Score Calibration–A Simulation Study," *American Journal of Epidemiology*, 165, 1110–1118.

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.

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. (2006) "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 ± standard deviation for ordinal variables.

| | HES primary data $n = 7956$ | | MCS validation data $m = 824$ | |
|---|---|---|---|---|
| | Trihalomethane exposure | | Trihalomethane exposure | |
| | $\geq 60\mu g/L$ | $< 60\mu g/L$ | $\geq 60\mu g/L$ | $< 60\mu g/L$ |
| *Variables in both primary and validation data* | | | | |
| Full-tem low birthweight | 144 (3.8) | 130 (3.1) | 14 (4.0) | 9 (1.9) |
| Mother's age | | | | |
| $\leq 25$ | 1216 (32) | 1539 (36) | 111 (32) | 148 (31) |
| 25 - 29 | 1059 (28) | 1157 (28) | 105 (30) | 135 (28) |
| 30 - 34 | 958 (25) | 991 (24) | 84 (24) | 137 (29) |
| $\geq 35$ | 526 (14) | 510 (12) | 46 (13) | 58 (12) |
| Male Baby | 1956 (52) | 2076 (50) | 176 (51) | 254 (53) |
| Carstairs quintile | 4.1 ± 1.3 | 4.3 ± 1.2 | 4.0 ± 1.1 | 4.0 ± 1.2 |
| *Variables in validation data only* | | | | |
| Lone parent family | . | . | 72 (21) | 105 (22) |
| $\geq 0$ children living outside of home | . | . | 13 (3.4) | 14 (2.9) |
| Smoking during pregancy | . | . | 126 (36) | 181 (38) |
| Non-white ethnicity | . | . | 77 (22) | 48 (10) |
| Alcohol during pregnancy | . | . | 110 (32) | 163 (34) |
| Body mass index $\geq 25\text{Kg/m}^2$ | . | . | 85 (25) | 134 (28) |
| Low education | . | . | 138 (40) | 210 (44) |
| *Total* | 3759 | 4197 | 346 | 478 |

Table 2: Odds ratios (95% interval estimates) for the association between exposure, covariates and the outcome in the primary data. The NAIVE analysis ignores the missing confounders and the validation data. The BayesPS and PSC analyses adjust for the seven missing confounders using the validation data.

| Description | NAIVE* | BayesPS** | | PSC** |
|---|---|---|---|---|
| | | $P(U\|C) = P(U)$ | $P(U\|C) \neq P(U)$ | |
| Trihalomethane exposure $> 60\mu g/L$ | 1.32 (1.04, 1.68) | 1.20 (0.91, 1.63) | 1.23 (0.94, 1.64) | 2.36 (1.36, 7.16) |
| Mother's age | | | | |
| $\leq 25$ | 1.17 (0.87, 1.58) | 1.11 (0.81, 1.43) | 1.13 (0.85, 1.46) | NA |
| 25 - 29 | 1.0 | 1.0 | 1.0 | |
| 30 - 34 | 0.90 (0.62, 1.28) | 0.79 (0.55, 1.13) | 0.82 (0.59, 1.13) | NA |
| $\geq 35$ | 1.05 (0.68, 1.61) | 1.09 (0.72, 1.55) | 1.18 (0.79, 1.75) | NA |
| Male baby | 0.77 (0.60, 0.98) | 0.75 (0.59, 0.96) | 0.75 (0.58, 0.94) | NA |
| Carstairs quintile | 1.34 (1.17, 1.52) | 1.35 (1.18, 1.53) | 1.28 (1.07, 1.56) | NA |

* Analysis ignores the missing confounders

** Adjusted for the missing confounders

NA - Not available.

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 $Y_j$ and $X_j$ adjusting for $C_j$ only. The middle column gives odds ratios adjusting for both $(C_j, U_j)$, whereas the rightmost column adjusts for $(C_j, \hat{Z}_j)$.

| 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 $\hat{Z}_j$ |
| Trihalomethane $> 60\mu g/L$ | 2.06 (0.87, 4.89) | 1.75 (0.70, 4.34) | 1.77 (0.73 4.31) |
| Mother's age | | | |
| $\leq 25$ | 0.65 (0.24, 1.76) | 0.52 (0.18, 1.50) | 0.65 (0.24, 1.77) |
| 25 - 29$^\dagger$ | 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) |
| $\geq 35$ | 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) | . |
| # of children living outside of home | . | 1.80 (0.73, 4.43) | . |
| Smoking during pregancy | . | 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 $\geq 25\text{Kg/m}^2$ | . | 1.06 (0.39, 2.89) | . |
| Low education | . | 1.32 (0.54, 3.20) | . |

$^\dagger$ Reference group

Table 4: Odds ratios for the association between the exposure and outcome in the primary data, adjusted for measured confounders, in a sensitivity analysis assuming that there is a single binary unmeasured confounder. Estimates are calculated using the method of Rosenbaum and Rubin (1983a).

| Effect of confounder on odds of trihalomethane exposure | Effect of confounder on odds of low birthweight | Prevalence of confounder among the unexposed subjects | | |
|---|---|---|---|---|
| | | 0.1 | 0.5 | 0.9 |
| Doubles the odds | Reduces by $\frac{2}{3}$ the odds | 1.40 (1.10, 1.78) | 1.59 (1.25, 2.02) | 1.43 (1.12, 1.82) |
| | Reduces by $\frac{1}{2}$ the odds | 1.38 (1.08, 1.75) | 1.48 (1.17, 1.89) | 1.38 (1.08, 1.75) |
| | No Effect | 1.32 (1.04, 1.68) | 1.32 (1.04, 1.68) | 1.32 (1.04, 1.68) |
| | Doubles the odds | 1.23 (0.97, 1.56) | 1.19 (0.93, 1.51) | 1.28 (1.01, 1.63) |
| | Triples the odds | 1.16 (0.92, 1.48) | 1.13 (0.89, 1.43) | 1.27 (1.00, 1.62) |
| Triples the odds | Reduces by $\frac{2}{3}$ the odds | 1.48 (1.16, 1.88) | 1.77 (1.39, 2.25) | 1.47 (1.16, 1.87) |
| | Reduces by $\frac{1}{2}$ the odds | 1.43 (1.13, 1.82) | 1.58 (1.24, 2.01) | 1.40 (1.10, 1.78) |
| | No Effect | 1.32 (1.04, 1.68) | 1.32 (1.04, 1.68) | 1.32 (1.04, 1.68) |
| | Doubles the odds | 1.16 (0.91, 1.48) | 1.13 (0.89, 1.43) | 1.27 (1.00, 1.62) |
| | Triples the odds | 1.06 (0.83, 1.35) | 1.05 (0.83, 1.34) | 1.26 (0.99, 1.60) |

Figure 1: Posterior density estimates for the exposure effect $\beta$, and the covariate effects $\xi$ and $\gamma$, based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

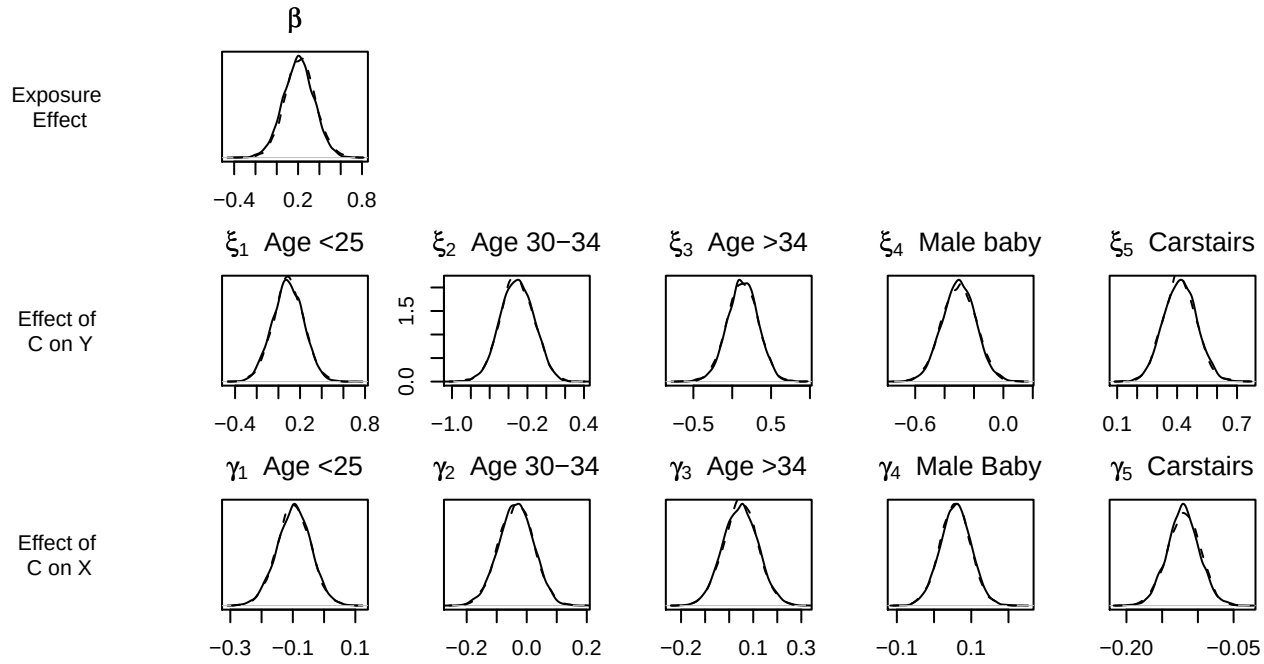


Figure 2: Posterior density estimates for the parameters $(\tilde{\xi}, \tilde{\gamma})$ that govern the magnitude of confounding from $U$, based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

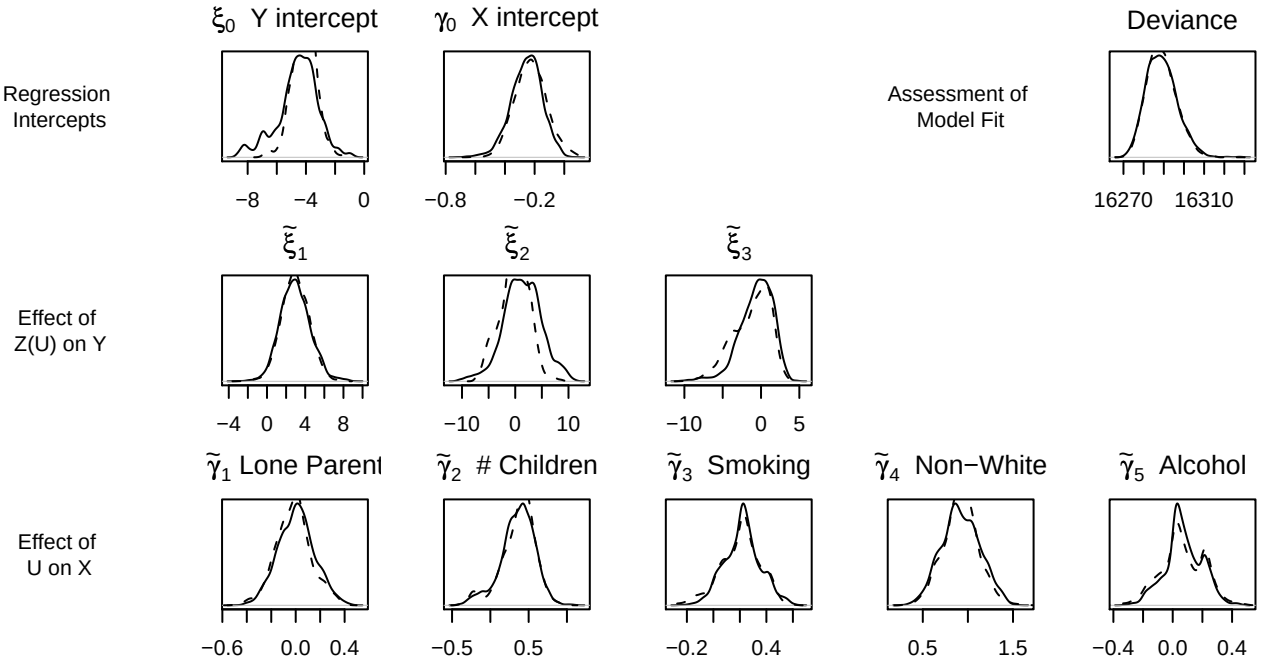

Figure 3: Performance of point and 80% interval estimates for the exposure effect $\beta$ calculated using either GOLD (solid curve), BayesPS (dashed curve), NAIVE (dotted curve), or PSC (dotted-dashed curve).