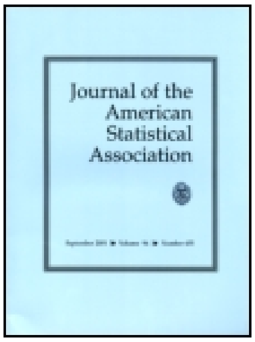

Journal of the
American
Statistical
Association

Taylor & Francis
Taylor & Francis Group

Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://www.tandfonline.com/loi/uasa20

Adjustment for Missing Confounders Using External Validation Data and Propensity Scores

Lawrence C. McCandless , Sylvia Richardson & Nicky Best

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

Adjustment for Missing Confounders Using External Validation Data and Propensity Scores

Lawrence C. McCandless, Sylvia Richardson, and Nicky Best

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 permit us to recover information about the missing data, but the difficulty is in eliciting a valid model for the 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 computation. Simulation results illustrate that the proposed method reduces bias from several missing confounders over a range of different sample sizes for the validation data. Appendices A–C are available as online supplementary material.

KEY WORDS: Bayesian inference; Bias; Causal inference; Observational studies.

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 modeling 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 article of Stürmer et al. (2005) from pharmacoepidemiology, who used 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 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) reviewed 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; Yin et al. 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) described techniques that use nonparametric density

Lawrence C. McCandless is Assistant Professor, Faculty of Health Sciences, Simon Fraser University, Burnaby, BC, Canada (E-mail: *mccandless@sfu.ca*). Sylvia Richardson is a Professor of Statistics and Epidemiology, at the School of Public Health, Department of Epidemiology and Biostatistics, Imperial College London, UK (E-mail: *sylvia.richardson@imperial.ac.uk*). Nicky Best is a Professor of Statistics and Epidemiology, at the School of Public Health, Department of Epidemiology and Biostatistics, Imperial College London, UK (E-mail: *n.best@imperial.ac.uk*). This work was supported by the Economic and Social Research Council award numbers 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. The authors are grateful to the Small Area Health Statistics Unit at Imperial College for access to the Hospital Episode Statistics (HES) data and to the postcoded Millennium Cohort Study (MCS) data. The authors acknowledge in particular the help of Peter Hambly and Margaret Douglass for processing the databases. The authors also thank Alexina Mason, Jassy Molitor, and Mireille Toledano for useful discussion that helped in the interpretation of the results.

Journal of the American Statistical Association
March 2012, Vol. 107, No. 497, Applications and Case Studies
DOI: 10.1080/01621459.2011.643739

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) reviewed 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 of the confounders are measured. To adjust for the *missing* confounders, we summarize them using a scalar summary score Z, which can be interpreted as the propensity score adjusted for the *measured* confounders. The score has the property that it induces conditional independence between the exposure and the missing confounders, given the measured confounders. To adjust for missing confounders in the primary data, we need only adjust for the propensity score Z. Our approach is to first specify a joint model for the primary and validation data. Because the propensity score is missing in the primary data, it is a missing covariate. We assign a model for the missing propensity score and then integrate it out of the likelihood function for the primary data.

To illustrate the problem of missing confounders, Section 2 describes a study from environmental epidemiology of the effect of trihalomethane exposure, a water disinfection byproduct, on the risk of full-term low birthweight in United Kingdom. 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 no information on maternal smoking, ethnicity, and other factors that influence birthweight. 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 an 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 et al. 2005) and multivariate imputation by chained equations (MICE) (van Buuren 2007). 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 byproduct, and the risk of full-term low birthweight in the United Kingdom (Toledano et al. 2005; Molitor et al. 2009). Trihalomethanes are formed when chlorine, which is routinely added to public water supplies in the United Kingdom, reacts with natural organic materials in the water. Pregnant mothers are exposed to trihalomethanes through drinking and bathing. Animal studies show that water disinfection byproduct compounds cause reproductive and developmental harm at higher doses. However, epidemiologic investigations of adverse birth outcomes have yielded contradictory findings, with some studies reporting an increased risk of low birthweight, while others showing no association (Toledano et al. 2005). It is important to distinguish between preterm low birthweight versus full-term low birthweight, which may indicate intrauterine growth retardation. Full-term low birthweight is rare, and any increase in the 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 HES, which is a data warehouse with details of all hospital admissions, including births, from National Health Service (NHS) hospitals in the United Kingdom. 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 2.5 kg.

Let Y be an indicator variable for the outcome, taking value 1 if an infant has full-term low birthweight, and 0 otherwise. Let X be an indicator variable for trihalomethane exposure, taking value 1 if the area-level exposure is greater than 60 μg/L, and 0 otherwise. Let $\mathbf{C}$ 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

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 $n = 7956$ Trihalomethane exposure		MCS validation data $m = 824$ Trihalomethane exposure	
	≥60 μg/L	<60 μg/L	≥60 μg/L	<60 μg/L
Variables in both primary and validation data				
Full-term low birthweight	144 (3.8)	130 (3.1)	14 (4.0)	9 (1.9)
Mother's age				
≤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)
≥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)
> 0 children living outside of home	—	—	13 (3.4)	14 (2.9)
Smoking during pregancy	—	—	126 (36)	181 (38)
Nonwhite ethnicity	—	—	77 (22)	48 (10)
Alcohol during pregnancy	—	—	110 (32)	163 (34)
Body mass index ≥ 25 kg/m^2	—	—	85 (25)	134 (28)
Low education	—	—	138 (40)	210 (44)
Total	3759	4197	346	478

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 $\mathbf{C}$. 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 an 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 have the advantage of a large sample size and nearly exhaustive coverage. However, it is missing information on maternal smoking and ethnicity, which influence birthweight. 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 the characteristics of populations.

In this investigation, information about missing confounders is available from external validation data. The UK 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 were obtained by sampling births in clusters based on residential wards, which are geographic areas of residence containing roughly 5000 individuals. The wards were stratified into three categories: 'advantaged," 'disadvantaged," and 'high ethnic minority." Hence, inferences from the MCS data should account for nonindependent sampling and

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, PSC, and MICE analyses adjust for the seven missing confounders using the validation data

		BayesPS[b]			
Description	NAIVE[a]	$P(\mathbf{U}\mid\mathbf{C}) = P(\mathbf{U})$	$P(\mathbf{U}\mid\mathbf{C}) \neq P(\mathbf{U})$	PSC[b]	MICE[b]
Trihalomethane exposure > 60 μg/L	1.32 (1.04–1.68)	1.20 (0.91–1.63)	1.23 (0.97–1.62)	2.36 (1.36–7.16)	1.22 (0.85–1.74)
Mother's age					
≤25	1.17 (0.87–1.58)	1.11 (0.81–1.43)	1.13 (0.85–1.50)	NA	0.77 (0.50–1.19)
25–29	1.0	1.0	1.0	1.0	
30–34	0.90 (0.62–1.28)	0.79 (0.55–1.13)	0.81 (0.57–1.12)	NA	0.79 (0.52–1.21)
≥35	1.05 (0.68–1.61)	1.09 (0.72–1.55)	1.14 (0.74–1.70)	NA	1.06 (0.57–1.98)
Male baby	0.77 (0.60–0.98)	0.75 (0.59–0.96)	0.76 (0.59–0.94)	NA	0.78 (0.58–1.06)
Carstairs quintile	1.34 (1.17–1.52)	1.35 (1.18–1.53)	1.35 (1.19–1.54)	NA	1.08 (0.80–1.46)

[a] Analysis ignores the missing confounders.
[b] Adjusted for missing confounders using the validation data.
NA, not available.

Table 3. Odds ratios (95% interval estimates) describing the confounding induced by $\mathbf{U}$ in the validation data alone ($m = 824$). The left-hand column gives odds ratios for the association between Y_j and X_j adjusting for $\mathbf{C}_j$ only. The middle column gives odds ratios adjusting for both $(\mathbf{C}_j, \mathbf{U}_j)$, whereas the right-hand column adjusts for $(\mathbf{C}_j, \hat{Z}_j)$

	Odds ratio (95% interval estimate) adjusting for		
Description	$(X_j, \mathbf{C}_j)$ only	$(X_j, \mathbf{C}_j)$ and $\mathbf{U}_j$	$(X_j, \mathbf{C}_j)$ and $\hat{Z}_j$
Trihalomethane > 60 µg/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[a]	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)	—
No. of children living outside of home	—	1.80 (0.73–4.43)	—
Smoking during pregnancy	—	2.86 (1.03–7.93)	—
Nonwhite 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)	—

[a] Reference group.

clustering within wards. 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 have a sample size of $n = 7956$, while the validation data, where more complete information on missing confounders is available, has a sample size of $m = 824$.

Upon consultation with subject area experts, we identify seven variables in the validation data that could potentially confound the exposure–outcome association. Let $\mathbf{U}$ 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 ≥ 25 kg/m^2, nonwhite ethnicity, and an indicator variable for low education. Table 1 gives a breakdown of the covariate distributions for the validation data. We see that nonwhite ethnicity is imbalanced between exposure groups, whereas smaller imbalances are observed for the other variables.

Denote the primary data as $\{(Y_i, X_i, \mathbf{C}_i, \mathbf{U}_i)|$ for $i \in 1 : n = 7956\}$ and the validation data as $\{(Y_j, X_j, \mathbf{C}_j, \mathbf{U}_j)|$ for $j \in 1 : m = 824\}$. The quantity $\mathbf{U}_i$ is completely unobserved. To study potential confounding induced by $\mathbf{U}$, Table 3 presents odds ratios for the association between Y_j and X_j when adjusting for $\mathbf{C}_j$ alone, versus adjusting for $\mathbf{C}_j$ and $\mathbf{U}_j$ in the validation data. In the first column of Table 3, we fit a logistic regression of Y_j on X_j and $\mathbf{C}_j$. 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 $\mathbf{C}_j$ and $\mathbf{U}_j$ and obtain the value 1.75 (0.70–4.34). The odds ratio is shifted toward 1 and the interval is slightly wider, indicating some evidence of additional confounding by $\mathbf{U}_j$. Together with epidemiological knowledge, this suggests that the missing $\mathbf{U}_i$ may confound the association between X_i and Y_i in the primary data.

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 $\mathbf{C}_i$, 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 $\mathbf{C}_i$. 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 labeling 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 the outcome would suffice to eliminate the association between X_i and Y_i seen in the NAIVE analysis. One such confounder is nonwhite ethnicity (see Tables 1 and 3). Nonwhites have increased odds of full-term low birthweight and are simultaneously more likely to have trihalomethane exposure.

3. BAYESIAN ADJUSTMENT FOR MISSING CONFOUNDERS USING PROPENSITY SCORES (BayesPS)

3.1 Models

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. We begin by introducing the propensity score conditional on measured confounders and illustrate that it can be used to adjust for missing confounders. Next, we obtain likelihood functions for the

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)

		Prevalence of confounder among the unexposed subjects		
Effect of confounder on odds of trihalomethane exposure	Effect of confounder on odds of low birthweight	0.1	0.5	0.9
Doubles the odds	Reduces by 2/3 the odds	1.40 (1.10–1.78)	1.59 (1.25–2.02)	1.43 (1.12–1.82)
	Reduces by 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 2/3 the odds	1.48 (1.16–1.88)	1.77 (1.39–2.25)	1.47 (1.16–1.87)
	Reduces by 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)

primary and validation data, integrating over the distribution of the missing propensity score.

To illustrate, suppose that $(Y_i, X_i, \mathbf{C}_i, \mathbf{U}_i)$ and $(Y_j, X_j, \mathbf{C}_j, \mathbf{U}_j)$ for $i \in 1:n$ and $j \in 1:m$ are identically distributed observations drawn from the same population with probability density function $P(Y, X, \mathbf{C}, \mathbf{U})$. Building on the Bayesian propensity score analysis of McCandless, Gustafson, and Austin (2009), we tease out the effect of X on Y, adjusting for $\mathbf{C}$ and $\mathbf{U}$, by using a pair of logistic regression models:

$$\text{logit}[P(X = 1|\mathbf{C}, \mathbf{U})] = \boldsymbol{\gamma}^T\mathbf{C} + \tilde{\boldsymbol{\gamma}}^T\mathbf{U}, \tag{1}$$

$$\text{logit}[P(Y = 1|X, \mathbf{C}, Z)] = \beta X + \boldsymbol{\xi}^T\mathbf{C} + \tilde{\boldsymbol{\xi}}^T\mathbf{g}\{Z\}, \tag{2}$$

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

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

In Equation (1), the quantity $Z = \tilde{\boldsymbol{\gamma}}^T\mathbf{U}$ is a scalar summary of $\mathbf{U}$. We define Z as the propensity score conditional on $\mathbf{C}$, 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 and can take on any value between $+\infty$ and $-\infty$. However, from Equation (1), we have that

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

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

In Appendix A, we prove that if there is no unmeasured confounding conditional on $(\mathbf{C}, \mathbf{U})$, then Equation (3) implies that there is no unmeasured confounding conditional on $(\mathbf{C}, Z)$. This means that to control for confounding from $\mathbf{U}$, we can estimate the exposure effect by modeling the conditional distribution of Y given $(X, \mathbf{C}, Z)$ as in Equation (2). We have no reason to believe that Z captures all information relevant to Y within $\mathbf{U}$. However, there is no need to specify a model for Y given $(X, \mathbf{C}, \mathbf{U})$ to adjust for missing confounders. It suffices to use the pair (1) and (2). For discussion of Bayesian regression adjustment for the propensity scores, see Rubin (1985).

Accordingly, Equation (2) 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 modeled in a nonparametric manner via the linear predictor $\boldsymbol{g}\{.\}$. For the trihalomethane data example, we use splines and let $\tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{z\} = \sum_{j=1}^{l} \tilde{\xi}_j g_j\{z\}$, where the quantities $g_j\{.\}$ are natural cubic spline basis functions with l knots and regression coefficients $\tilde{\boldsymbol{\xi}} = (\tilde{\xi}_0, \dots, \tilde{\xi}_l)$. This gives a smooth yet flexible relationship between Y and Z within levels of X and $\mathbf{C}$. See Little and An (2004) for a detailed discussion of regression modeling strategies that use the propensity score as a covariate.

The preceding discussion is valid only if Z is known. However, the regression coefficients $\tilde{\boldsymbol{\gamma}}$ are not known, and if we multiply the wrong coefficients by $\mathbf{U}$, then the conditional independence of Equation (3) does not hold. Furthermore, a crucial assumption of our method is that Equation (1) correctly captures the true relationship between X and $(\mathbf{C}, \mathbf{U})$. For example, if the true exposure model has an interaction between individual components of $\mathbf{C}$ and $\mathbf{U}$, then our methodology would not be applicable. In Equations (1) and (2), the parameter $\tilde{\boldsymbol{\gamma}}$ is estimated simultaneously with $(\beta, \boldsymbol{\gamma}, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}})$ in a full Bayesian analysis. Provided that all models are correctly specified, the posterior will tend to concentrate at the true value of $\tilde{\boldsymbol{\gamma}}$ as we collect more data, hence providing increasingly better control of confounding.

In the primary data, the quantities Y, X, and $\mathbf{C}$ are observed, while $\mathbf{U}$ is missing. Equations (1) and (2) define the density

$P(Y, X|\mathbf{C}, Z)$, and we can use it to calculate the marginal model for $P(Y, X|\mathbf{C})$ integrating over Z. We have

$$\begin{aligned} P(Y, X|\mathbf{C}) &= E\{P(Y, X, Z|\mathbf{C})\} \\ &= \int P(Y|X, \mathbf{C}, Z)P(X|\mathbf{C}, Z)P(Z|\mathbf{C})dZ, \end{aligned} \tag{4}$$

where $P(Y|X, \mathbf{C}, Z)$ and $P(X|\mathbf{C}, \mathbf{U}) = P(X|\mathbf{C}, Z)$ are given in Equations (1) and (2). To complete the specification, Equation (4) requires a model for Z given $\mathbf{C}$.

One simplification is to assume that $\mathbf{C}$ and $\mathbf{U}$, and hence $\mathbf{C}$ and Z, are marginally independent, meaning that $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$. 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 $\{Z_j = \tilde{\boldsymbol{\gamma}}^T\mathbf{U}_j | j \in 1 : m\}$. This permits us to compute Equation (4) using the approximation

$$\begin{aligned} P(Y, X|\mathbf{C}) \approx \frac{1}{m}\sum_{j=1}^{m}&\left[\frac{\exp\{Y(\beta X + \boldsymbol{\xi}^T\mathbf{C} + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{Z_j\})\}}{1+\exp\{\beta X + \boldsymbol{\xi}^T\mathbf{C} + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{Z_j\}\}}\right] \\ &\times\left[\frac{\exp\{X(\boldsymbol{\gamma}^T\mathbf{C} + Z_j)\}}{1+\exp\{\boldsymbol{\gamma}^T\mathbf{C} + Z_j\}}\right]. \end{aligned} \tag{5}$$

Alternatively, we can use a quadrature estimate,

$$\begin{aligned} P(Y, X|\mathbf{C}) \approx \sum_{k=1}^{M}\omega_k&\left[\frac{\exp\{Y(\beta X + \boldsymbol{\xi}^T\mathbf{C} + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\hat{Z}_k\})\}}{1+\exp\{\beta X + \boldsymbol{\xi}^T\mathbf{C} + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\hat{Z}_k\}\}}\right] \\ &\times\left[\frac{\exp\{X(\boldsymbol{\gamma}^T\mathbf{C} + \hat{Z}_k)\}}{1+\exp\{\boldsymbol{\gamma}^T\mathbf{C} + \hat{Z}_k\}}\right], \end{aligned} \tag{6}$$

based on a histogram of the empirical distribution of the propensity scores in the validation data. Here, the index k equals $1, 2, \ldots, M$, 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 Equation (6) faster to compute than Equation (5) because it is a summation of size M rather than size m and, typically, $M << m$.

An advantage of using Equation (6) to estimate the effect of X on Y is that it requires no parametric assumptions for the nuisance distribution of $\mathbf{U}$. It can be used regardless of the correlation structure of the components of $\mathbf{U}$ and with continuous and categorical variables. However, a disadvantage is that it assumes that $\mathbf{U}$ and $\mathbf{C}$ 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 $\mathbf{C}$. 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 assign a linear regression model

$$Z|\mathbf{C} \sim N(\hat{\boldsymbol{\theta}}^T\mathbf{C}, \hat{\sigma}^2). \tag{7}$$

The linear predictor $\hat{\boldsymbol{\theta}}^T\mathbf{C} = \hat{\theta}_0 + \hat{\theta}_1 C_1 + \cdots + \hat{\theta}_p C_p$ is the estimated mean propensity score conditional on $\mathbf{C}$ with variance $\hat{\sigma}^2$. The estimates $(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ are obtained from a preliminary analysis of the validation data. Specifically, we regress $\hat{\tilde{\boldsymbol{\gamma}}}^T\mathbf{U}_j$ onto $\mathbf{C}_j$, where $\hat{\tilde{\boldsymbol{\gamma}}}$ is the maximum likelihood estimate of $\tilde{\boldsymbol{\gamma}}$ computed by fitting Equation (1) to the validation data alone. By using Equation (7), we can then obtain $P(Y, X|\mathbf{C})$ for the primary data by using Equation (6) and setting $(\hat{Z}_k, \omega_k)$ for $k = 1, 2, \ldots, M$ as a histogram approximation to the normal distribution in Equation (7). In the trihalomethane data, we set $M = 7$, with $\hat{Z}_1, \ldots, \hat{Z}_7$ equal to $q \times \hat{\sigma} + \hat{\boldsymbol{\theta}}^T\mathbf{C}$, 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. See also Section 4.2 for details and an illustration.

The motivation behind Equation (7) is that we do not need to know the quantity $\mathbf{U}$ 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 (7) uses plug-in point estimates and does not incorporate uncertainty in $(\boldsymbol{\theta}, \sigma^2)$. Although it would be desirable to estimate $(\boldsymbol{\theta}, \sigma^2)$ jointly with other model parameters during MCMC, this is computationally demanding. The quantity $\mathbf{U}$ is unobserved in the primary data, and updating $(\boldsymbol{\theta}, \sigma^2)$ leads to slow convergence. In practice, we can do a sensitivity analysis where we replace the plug-in estimates with a few draws from the posterior distribution of $(\boldsymbol{\theta}, \sigma^2)$ corresponding to noninformative priors. See Section 4.2 and Appendix C for an illustration.

One alternative strategy to gain independence between $\mathbf{U}$ and $\mathbf{C}$ would be to replace $\mathbf{U}$ with the residuals from a regression of $\mathbf{U}$ on $\mathbf{C}$. This is feasible in the trihalomethane data example because the components of $\mathbf{U}$ are primarily categorical. However, our propensity score approach is aimed at showing how full modeling or imputation of $\mathbf{U}$ (which is computationally expensive even for moderate-dimensional $\mathbf{U}$) can be avoided if one is only interested in estimating the exposure effect. This is why we do not consider this route further.

As a point of comparison, we consider two alternative methods. In Sections 4.3 and 5.2, we compare BayesPS with PSC, which is another method that uses propensity scores and external validation data to adjust for several missing confounders. Additionally, we contrast BayesPS with MICE (van Buuren 2007), which is a missing data method that uses multivariate imputation on a variable-by-variable basis via a set of conditional densities, one for each missing variable.

3.2 Prior Distributions

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

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

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%. These priors make similar assumptions about the association between Y and $(\mathbf{C}, Z)$ given X, and also the association between X and $(\mathbf{C}, \mathbf{U})$. 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 the validation data. Inferences from BayesPS are obtained from the posterior density $P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\text{data})$, which we sample using MCMC. We have

$$\begin{aligned}
&P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\text{data})\\
&\quad\propto \left\{\prod_{i=1}^{n} P(Y_i, X_i|\mathbf{C}_i)\right\} \times \left\{\prod_{j=1}^{m} P(Y_j, X_j|\mathbf{C}_j, \mathbf{U}_j)\right\}\\
&\qquad \times P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})\\
&\quad\approx \prod_{i=1}^{n}\left\{\sum_{k=1}^{M}\omega_k\left[\frac{\exp\{Y_i(\beta X_i + \boldsymbol{\xi}^T\mathbf{C}_i + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\hat{Z}_k\})\}}{1+\exp\{\beta X_i + \boldsymbol{\xi}^T\mathbf{C}_i + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\hat{Z}_k\}\}}\right]\right.\\
&\qquad \left.\times\left[\frac{\exp\{X_i(\boldsymbol{\gamma}^T\mathbf{C}_i + \hat{Z}_k)\}}{1+\exp\{\boldsymbol{\gamma}^T\mathbf{C}_i + \hat{Z}_k\}}\right]\right\}\\
&\qquad \times\prod_{j=1}^{m}\left\{\left[\frac{\exp(Y_j(\beta X_j + \boldsymbol{\xi}^T\mathbf{C}_j + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\tilde{\boldsymbol{\gamma}}^T\mathbf{U}_j)\}))}{1+\exp(\beta X_j + \boldsymbol{\xi}^T\mathbf{C}_j + \tilde{\boldsymbol{\xi}}^T\boldsymbol{g}\{\tilde{\boldsymbol{\gamma}}^T\mathbf{U}_j\})}\right]\right.\\
&\qquad \left.\times\left[\frac{\exp(X_j(\boldsymbol{\gamma}^T\mathbf{C}_j + \tilde{\boldsymbol{\gamma}}^T\mathbf{U}_j))}{1+\exp(\boldsymbol{\gamma}^T\mathbf{C}_j + \tilde{\boldsymbol{\gamma}}^T\mathbf{U}_j)}\right]\right\} \times P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}),
\end{aligned} \tag{8}$$

where the products over i and j are the likelihood functions for the primary and the validation data, respectively, and $P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}})$ is the prior density for β, $\boldsymbol{\xi}$, $\tilde{\boldsymbol{\xi}}$, $\boldsymbol{\gamma}$, and $\tilde{\boldsymbol{\gamma}}$, and we adopt the quadrature approach of Equation (6).

We sample from $P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\text{data})$ by updating from the conditional distributions for $[\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}|\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}, \text{data}]$ and $[\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \text{data}]$ using the Metropolis–Hastings algorithm. To update from $[\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}|\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}, \text{data}]$, we use a proposal distribution based on a random walk that updates each component $\beta, \xi_0, \ldots, \xi_p, \tilde{\xi}_1, \ldots \tilde{\xi}_l$ one at a time using a mean zero normal disturbance. Multivariate updating from $[\boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \text{data}]$ is accomplished using the proposal distribution described in McCandless et al. (2009).

4. ANALYSIS RESULTS FOR THE TRIHALOMETHANE DATA

4.1 BayesPS Assuming **U** and **C** Are Marginally Independent

We use BayesPS to adjust for missing confounders in the trihalomethane data. We begin by using the empirical distribution approach, described in Section 3.1, which assumes that **U** and **C** are marginally independent (i.e., $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$). Before applying BayesPS to the trihalomethane data, we set a priori values for the knots used to define the spline basis function $\boldsymbol{g}\{.\}$ in Equation (2). Following McCandless et al. (2009), we fit the logistic regression model given in Equation (1) to the validation data using maximum likelihood to estimate the parameter $\tilde{\boldsymbol{\gamma}}$. The quantities $\hat{Z}_j$, computed by evaluating $\{\hat{Z}_j = \hat{\tilde{\boldsymbol{\gamma}}}^T\mathbf{U}_j|j \in 1:m\}$, 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 right-most column, we fit the model in Equation (2) using $\hat{Z}_j$ to control for confounding, rather than using $\mathbf{U}_j$. The resulting odds ratio 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 that adjusts for $\mathbf{U}_j$ 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 $\mathbf{U}_j$ in the model for Y_j.

We now fit BayesPS to the primary and the validation data combined. We assume that $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$, 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 by cluster randomized sampling, stratifying residential wards into three categories based on neighborhood income and ethnicity. Thus, we should account for nonindependent sampling and clustering of individuals when combining the likelihood functions of the primary and the validation data. Following Molitor et al. (2009), we include strata-specific intercepts when fitting Equations (1) and (2). These intercepts enter into Equation (8) as strata indicator variables that are added to the linear predictors for the likelihood contributions of the validation data. This methodology is described by Gelman et al. (2004, section 7.4) for Bayesian inference for survey samples with stratified sampling.

We ignore possible clustering at the ward level, and this is a potentially serious limitation of our analysis. In principle, such clustering could be accounted for by including additional ward-level random effects into the models. However, a sensitivity analysis by Molitor et al. (2009) revealed that residual clustering at the ward level was not overly large for the MCS data. In the interest of simplicity, we ignore the clustering.

We then apply BayesPS to the trihalomethane data by sampling from the posterior density $P(\beta, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}}, \boldsymbol{\gamma}, \tilde{\boldsymbol{\gamma}}|\text{data})$. We obtain two different MCMC chains with overdispersed starting values and length 1,00,000 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, $\boldsymbol{\xi}$, and $\boldsymbol{\gamma}$ is better than for $\tilde{\boldsymbol{\xi}}$ and $\tilde{\boldsymbol{\gamma}}$ 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 $\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\gamma}}$, 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 $-2\log[\prod_{i=1}^{n} P(Y_i, X_i|\mathbf{C}_i) \times \prod_{j=1}^{m} P(Y_j, X_j|\mathbf{C}_j, \mathbf{U}_j)]$, 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 $\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\gamma}}$ does not greatly affect model fit. Because the convergence of β, $\boldsymbol{\xi}$, and $\boldsymbol{\gamma}$ is satisfactory, we can compute summaries of the marginal posterior distributions of β, $\boldsymbol{\xi}$, and $\boldsymbol{\gamma}$. Ideally, we could use longer MCMC runs, but this is computationally intensive.

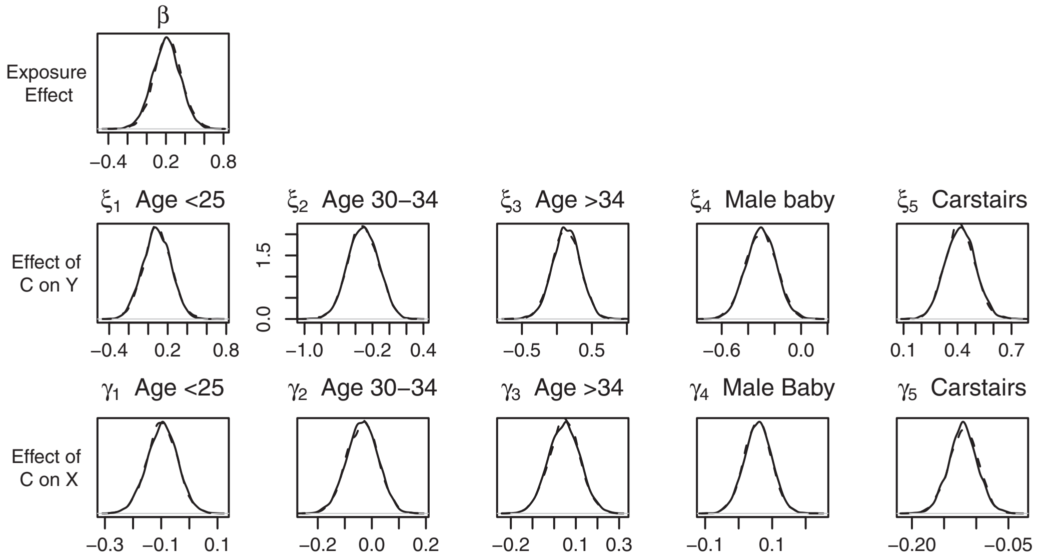

Figure 1. Posterior density estimates for the exposure effect β and the covariate effects $\boldsymbol{\xi}$ and $\boldsymbol{\gamma}$, based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

The second column in Table 2 presents the results of applying BayesPS to the trihalomethane data when we assume that $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$. 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 with 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 β is shifted toward zero. This result makes sense because in Table 3, when analyzing the validation data alone, we see that adjustment for either $\mathbf{U}_j$ or Z_j drives the estimate of β toward zero, as compared with an analysis ignoring $\mathbf{U}_j$. It is worth noting that at least in principle, the bias-corrected exposure effect could be shifted *away* from zero, depending on the direction of confounding observed in the validation data.

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 the validation data combined intuitively ought to yield less posterior uncertainty as compared with 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

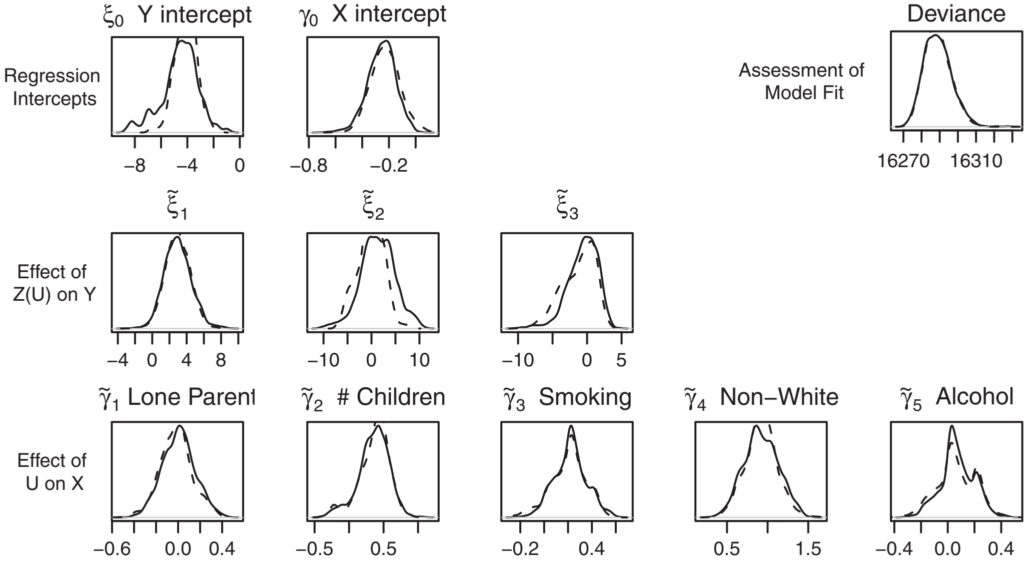

Figure 2. Posterior density estimates for the parameters $(\tilde{\boldsymbol{\xi}}, \tilde{\boldsymbol{\gamma}})$ that govern the magnitude of confounding from $\mathbf{U}$, based on two different MCMC chains with overdispersed starting values (solid curve versus broken curve).

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 $(\log(15)/2)^2$ (results not shown). The point and interval estimates for the quantities β and $\boldsymbol{\xi}$ are almost identical to those in Table 2, with difference ≤ 0.03 on the log odds scale. Greater sensitivity is observed for the parameters $(\tilde{\boldsymbol{\gamma}}, \tilde{\boldsymbol{\xi}})$ 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 marginally independent seems questionable in the trihalomethane data. For example, mothers 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 and assign a general location model for the missing propensity scores in the primary data using Equation (7). First, we estimate $\tilde{\boldsymbol{\gamma}}$ by fitting Equation (1) to the validation data alone and calculate the estimated propensity scores, given by $\{\hat{Z}_j = \hat{\tilde{\boldsymbol{\gamma}}}^T \mathbf{U}_j \mid j = 1, \ldots, m\}$. Then, we fit a linear regression of $\hat{Z}_j$ on $\mathbf{C}_j$ to calculate $(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$ used in Equation (7). We include strata-specific intercepts in the regression models because of the stratified sampling of residential wards based on neighborhood income and ethnicity. We obtain $\hat{\boldsymbol{\theta}} = (\hat{\theta}_0, \hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_5) = (0.13, -0.02, -0.04, -0.09, -0.00, -0.00)$, 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 $\boldsymbol{\theta}$ 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 θ_0 is the y intercept, and $\theta_1, \ldots, \theta_p$ govern how the mean propensity score depends on each component of **C**.

Using $(\hat{\boldsymbol{\theta}}, \hat{\sigma}^2)$, we apply BayesPS using the methodology described in Section 3.1. The results are presented in the third column of Table 2 under the heading "$P(\mathbf{U}|\mathbf{C}) \neq P(\mathbf{U})$." We see an odds ratio for the exposure effect of 1.23 (0.94–1.64) that is shifted toward 1.0 as compared with the NAIVE analysis, and which is very similar to BayesPS assuming $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$. 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 (7) 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.

A limitation of our analysis is that Equation (7) uses plug-in estimates of hyperparameters. To understand the role of parameter uncertainty, we did a sensitivity analysis by repeating the analysis using a few samples from the posterior distribution of $(\boldsymbol{\theta}, \sigma^2)$ under a noninformative prior. The results are given in Table A.1 of Appendix C and show that using different estimates has a modest effect on the results. If a stronger link between Z and **C** were to exist, more sensitivity might be expected.

4.3 Comparison With PSC and MICE

We contrast our results with those obtained using PSC and MICE, which are alternative methods to adjust for missing confounders using validation data. PSC also uses propensity scores and 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) defined as the quantity $P(X = 1|\mathbf{C})$. 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 to adjust for the missing confounders. Further details concerning PSC are given in Appendix B.

We implement PSC exactly as described by 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 their article. 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) to test this assumption, we obtain a p-value of 0.27. Stürmer et al. (2007) use a threshold of $p < 0.30$ as an indication 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 and Appendix B, we show that PSC can perform nearly as well as BayesPS in some settings. We refer the reader to Stürmer et al. (2007) for details of the measurement error surrogacy assumption.

Finally, we apply MICE to the trihalomethane data, which attempts to model the missing confounders directly. MICE is a missing data method (van Buuren 2007) that uses multivariate imputation on a variable-by-variable basis via a set of conditional densities, one for each missing variable. We apply MICE to the primary and the validation data using default settings of the mice package in R (van Buuren and Groothuis-Oudshoorn 2010). This generates multiple imputations of the missing $\mathbf{U}_i$. Standard complete data methods are used to analyze each imputed dataset, and this is accomplished using logistic regression of Y on $(X, \mathbf{C}, \mathbf{U})$ for the primary and the validation data combined. The results from the different imputations are combined to model uncertainty from the missing confounders. The results are given in the right-most column of Table 2. MICE gives an odds ratio for the exposure effect equal to 1.22 (0.85–1.74), which is in close agreement with the results of BayesPS, although with slightly wider interval estimate.

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 $\mathbf{U}$ give unconfounded exposure effect estimates? How might BayesPS compare with 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 that contain confounding from several 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, $\mathbf{C}$ is a 5×1 vector, including intercept, and $\mathbf{U}$ is a 4×1 vector). Primary data ($n = 1000$) and validation data ($m = 100$, 250, 500, and 1000) are generated using the following algorithm: Simulate $\{\mathbf{C}_i, \mathbf{C}_j\}$ for $i \in 1 : n$, $j \in 1 : m$, and also $\{\mathbf{U}_i, \mathbf{U}_j\}$ for $i \in 1 : n$, $j \in 1 : m$, where each component of $\mathbf{C}_i, \mathbf{C}_j, \mathbf{U}_i,$ and $\mathbf{U}_j$ is independent and identically distributed as an 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 (1). 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:

$$\operatorname{logit}[P(Y = 1|X, \mathbf{C}, \mathbf{U})] = \beta X + \boldsymbol{\xi}^T \mathbf{C} + \tilde{\boldsymbol{\xi}}^T \mathbf{U}. \qquad (9)$$

Note that the first component of $\mathbf{C}_i$ and $\mathbf{C}_j$ is equal to 1 so that γ_0 and ξ_0 are regression intercept terms. The choices for $\boldsymbol{\xi}$, $\tilde{\boldsymbol{\xi}}$, $\boldsymbol{\gamma}$, and $\tilde{\boldsymbol{\gamma}}$ 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.

Equation (9), with which we generate data, is different from Equation (2), with which we analyze data using BayesPS. We do not simulate directly from Equation (2) because we have no reason to believe that Z captures the relationship between Y and the missing confounder $\mathbf{U}$. As discussed in Section 3.1, Equation (2) merely approximates the true model for Y by substituting the propensity score in place of the full vector of missing confounders. Thus, we view Equation (9) as a better representation of the true data-generating mechanism for simulation purposes.

For each value of m, we analyze the 200 pairs of datasets using BayesPS, NAIVE, PSC, and MICE to obtain point and 80% interval estimates of the exposure effect β. Because the components of $\mathbf{U}$ and $\mathbf{C}$ are generated as independent unit normals, we apply the BayesPS method assuming that $P(\mathbf{U}|\mathbf{C}) = P(\mathbf{U})$ and by using the empirical distribution approach described in Section 3.1. Sampler convergence is assessed using separate MCMC runs. In addition, because the quantities $\mathbf{U}_i$ for $i \in 1 : n$ are known by construction, we also apply a method called GOLD, which involves fitting Equation (9) to the primary data using $(Y_i, X_i, \mathbf{C}_i, \mathbf{U}_i)$, 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 GOLD, BayesPS, NAIVE, PSC, and MICE 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 bottom 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 GOLD, BayesPS, NAIVE, PSC, and MICE and then averaged the results.

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 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 bottom-left panel, the dotted curve hovers at 50%, indicating that the coverage probability of NAIVE interval estimates for the exposure effect β 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 β.

Figure 3 illustrates that BayesPS (long dashed curve) eliminates bias from the missing confounders over a range of values for m. In the top-left panel, we see that the curve lies near zero bias. BayesPS estimates of β 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 bottom-left panel of Figure 3 summarizes the performance of interval estimates for the exposure effect β. BayesPS interval estimates have improved coverage probability compared with NAIVE estimates, 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 are computationally expensive.

MICE (dotted-dashed curve) is also compared with BayesPS. We see that the performances are very similar. MICE also succeeds in reducing bias from the missing confounders across a range of values of m. One interesting observation is that BayesPS and MICE estimates of β are sometimes more efficient than GOLD estimates. For large m, BayesPS and MICE intervals are shorter than GOLD intervals, despite the fact that they acknowledge uncertainty from missing confounders. This is perhaps expected because BayesPS and MICE incorporate the validation data into the analysis, while GOLD ignores it. However, it nonetheless illustrates that incorporating external information about confounding can actually *increase* the precision of the exposure effect estimates. This suggests that if it is reasonable to assume that the models in Equations (1) and (2)

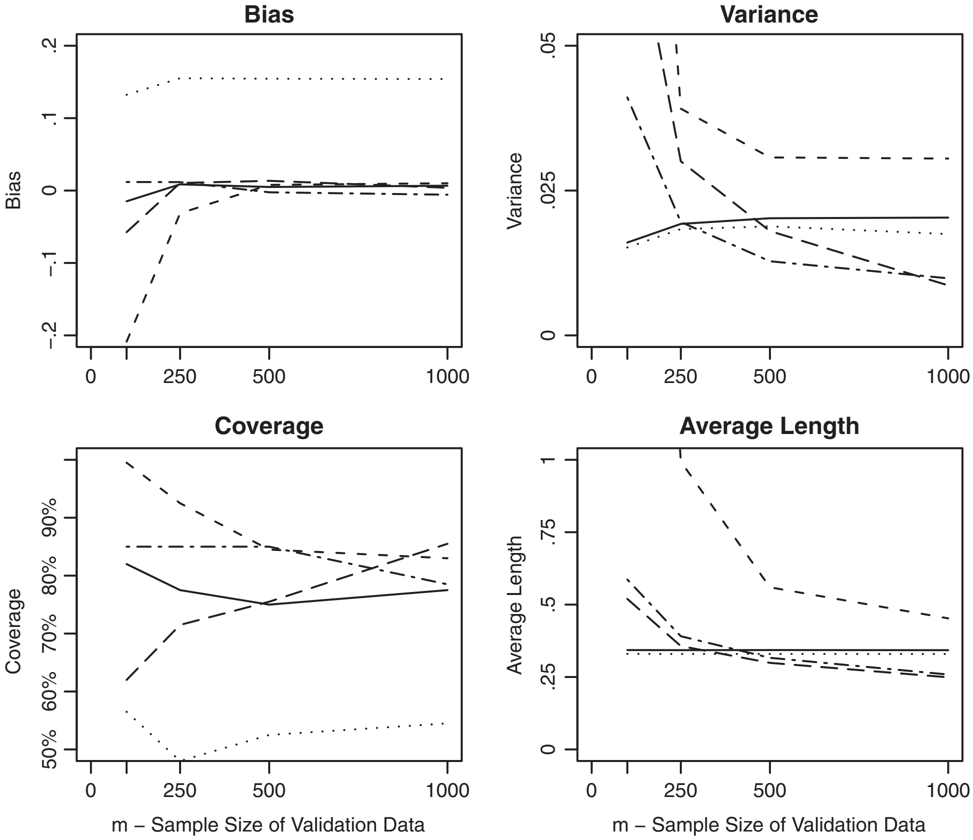

Figure 3. Performance of point and 80% interval estimates for the exposure effect β calculated using GOLD (solid curve), BayesPS (long dashed curve), NAIVE (dotted curve), PSC (short dashed curve), or MICE (dotted-dashed curve).

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 that it eliminates bias over the range of m under consideration. The point estimates are slightly less efficient than BayesPS estimates, particularly for $m < 500$. However, this is understandable because PSC ignores the outcome variable in the validation data, and is therefore disadvantaged compared with 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 (Stürmer et al. 2005). In Appendix B, we show that PSC deteriorates compared with BayesPS through small modifications of the simulation design that violate the measurement error assumption.

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 $\mathbf{U}$, within levels of $\mathbf{C}$. 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 ($m \geq 100$).

In the trihalomethane data example, the validation data were collected in a separate survey that used stratified cluster randomized sampling. It is important to account for nonindependent sampling of the validation data in our BayesPS model to ensure that the validation data can be considered exchangeable with the primary data (i.e., representative of the same underlying population). Adjustments based on survey weights are difficult to incorporate into the Bayesian framework, and this may place some limits on the generality of our method. However, various model-based adjustments for complex sampling designs are available (Gelman et al. 2004, section 7.4). Such approaches can, in principle, be incorporated into our BayesPS model. For example, in our case study, we included stratum indicator variables as additional covariates in the regression model to adjust for the stratified sampling design of the MCS data. In the interest of simplicity, we ignored clustering at the ward level, and this is an important limitation of our work. In principle, we could have included ward-level random effects in the regression to account for clustering. However, a previous analysis of birthweight outcomes in the MCS data (Molitor et al. 2009) indicated that ward-level clustering was not large.

We also note that when using BayesPS, both the primary and the validation datasets must contain the common set of variables Y, X, and $\mathbf{C}$ that are measured in the same fashion. Additionally, 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 modeling 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, Smith, and Sterne (2007), who showed 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 seven missing confounders, which are mostly categorical. In Sections 4.3 and 5.2, we contrast BayesPS with MICE and we see that the performance of both methods is very similar. Nonetheless, one theoretical weakness of MICE is that the specified conditional densities may be incompatible. The stationary distribution to which the Gibbs sampler attempts to converge may not exist (van Buuren 2007). A further alternative would be to attempt to model the joint distribution of **U** given **C** directly using a general location model (Little and Rubin 2002). 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** and consequently will be more robust to potential misspecification of the model for **U**. Furthermore, full Bayesian updating of **U** and monitoring of convergence can be computationally intensive.

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, Nieuwenhuijsen, and Best (2005) showed 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.

SUPPLEMENTARY MATERIAL

Appendices A–C are available as online supplements to this article.

[Received February 2010. Revised August 2011.]

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. [40]

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. [40]

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. [51]

Gelman, A., Carlin, J., Stern, H., and Rubin, D. (2004), *Bayesian Data Analysis* (2nd ed.), New York: Chapman Hall/CRC. [46,50]

Greenland, S. (2005), "Multiple Bias Modelling for Analysis of Observational Data" (with discussion), *Journal of the Royal Statistical Society,* Series A, 168, 267–306. [45]

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. [40]

Little, R. J. A., and An, H. (2004), "Robust Likelihood-Based Analysis of Multivariate Data With Missing Values," *Statistica Sinica*, 14, 949–968. [44]

Little, R. J. A., and Rubin, D. B. (2002), *Statistical Analysis With Missing Data* (2nd ed.), New York: Wiley. [40,41,45,51]

Lunceford, J. K., and Davidian, M. (2004), "Stratification and Weighting via the Propensity Score in Estimation of Causal Treatment Effects: A Comparative Study," *Statistics in Medicine*, 23, 2937–2960. [41]

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009), "Bayesian Propensity Score Analysis for Observational Data," *Statistics in Medicine*, 28, 94–112. [44,46]

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. [40,43]

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. [41,42,46,50]

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. [40,43]

—— (1983b), "The Central Role of the Propensity Score in Observational Studies for Causal Effects," *Biometrika*, 70, 41–57. [41,44]

Rubin, D. B. (1985), "The Use of Propensity Scores in Applied Bayesian Inference," in *Bayesian Statistics 2*, eds. J. M. Bernardo, M. H. De Groot, D. V. Lindley, and A. F. M. Smith, Valencia: Valencia University Press, pp. 463–472. [44]

Rubin, D. B., and Thomas, N. (1996), "Matching Using Estimated Propensity Scores: Relating Theory to Practice," *Biometrics*, 52, 249–264. [41]

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. [40]

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. [40,41,48,50]

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. [48]

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. [41]

van Buuren, S. (2007), "Multiple Imputation of Discrete and Continuous Data by Fully Conditional Specification," *Statistical Methods in Medical Research*, 16, 219–242. [41,45,48,51]

van Buuren, S., and Groothuis-Oudshoorn, K. (2010), "MICE: Multivariate Imputation by Chained Equations in R." *Journal of Statistical Software*, 45. Available at *http://www.jstatsoft.org/v45/i03*. [48]

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. [40]

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. [40]

Whitaker, H., Nieuwenhuijsen, M. J., and Best, N. (2005), "The Relationship Between Water Concentrations and Individual Uptake of Chloroform: A Simulation Study," *Environmental Health Perspectives*, 111, 688–694. [51]

Yin, L., Sundberg, R., Wang, X., and Rubin, D. B. (2006), "Control of Confounding Through Secondary Samples," *Statistics in Medicine*, 25, 3814–3825. [40]