Biometrics 68, 661–686
September 2012
DOI: 10.1111/j.1541-0420.2011.01731.x

Bayesian Effect Estimation Accounting for Adjustment Uncertainty

Chi Wang,[1,2,*] Giovanni Parmigiani,[3,4] and Francesca Dominici[4]

[1]Markey Cancer Center, University of Kentucky, Lexington, Kentucky 40536, U.S.A.
[2]Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, Kentucky 40536, U.S.A.
[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.
[4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[*]*email:* chi.wang@uky.edu

Summary. Model-based estimation of the effect of an exposure on an outcome is generally sensitive to the choice of which confounding factors are included in the model. We propose a new approach, which we call Bayesian adjustment for confounding (BAC), to estimate the effect of an exposure of interest on the outcome, while accounting for the uncertainty in the choice of confounders. Our approach is based on specifying two models: (1) the outcome as a function of the exposure and the potential confounders (the outcome model); and (2) the exposure as a function of the potential confounders (the exposure model). We consider Bayesian variable selection on both models and link the two by introducing a dependence parameter, ω, denoting the prior odds of including a predictor in the outcome model, given that the same predictor is in the exposure model. In the absence of dependence ($\omega = 1$), BAC reduces to traditional Bayesian model averaging (BMA). In simulation studies, we show that BAC, with $\omega > 1$, estimates the exposure effect with smaller bias than traditional BMA, and improved coverage. We, then, compare BAC, a recent approach of Crainiceanu, Dominici, and Parmigiani (2008, *Biometrika* **95,** 635–651), and traditional BMA in a time series data set of hospital admissions, air pollution levels, and weather variables in Nassau, NY for the period 1999–2005. Using each approach, we estimate the short-term effects of $PM_{2.5}$ on emergency admissions for cardiovascular diseases, accounting for confounding. This application illustrates the potentially significant pitfalls of misusing variable selection methods in the context of adjustment uncertainty.

Key words: Adjustment uncertainty; Bayesian model averaging; Exposure effects; Treatment effects.

1. Introduction

Estimating the effect of an exposure on an outcome, while properly adjusting for confounding factors, is a common goal in biomedical research. A prominent and controversial example arises in observational studies of the health effects of environmental contaminants, where the choice of potential confounders is challenging, and major policy decisions can depend on it. The most common practice is currently to select a statistical model for the estimation of the effect, and report effect estimates and confidence intervals (CIs) that are conditional on that model being correct. This does not account for "adjustment uncertainty," that is uncertainty about which variables should be included in the model to properly adjust for confounding.

It is possible to effectively convey this uncertainty by sensitivity analysis, showing the variation of the effect estimate and its interval over a range of plausible choices of confounders (Dominici, McDermott, and Hastie, 2004; Peng, Dominici, and Louis, 2006). Bayesian model averaging (BMA) has been suggested as a more formal tool to account for model uncertainty. Bayesian predictions that account for uncertainty in the selection of predictors (Raftery, Madigan, and Hoeting, 1997; Hoeting et al., 1999) are based on treating the indicators of whether each predictor is included in the model as unknown nuisance parameters. This results in a weighted average of predictions whose weights depend on the support that each selection receives from the data. This principled approach enjoys a number of desirable properties from a frequentist point of view as well, and has performed competitively in out-of-sample prediction comparisons (Chipman, George, and McCulloch, 2002; Yeung, Bumgarner, and Raftery, 2005). The conceptual simplicity and solid logic behind treating the unknown confounder subset as a parameter is attractive in adjustment uncertainty as well. Raftery (1995) and Hoeting et al. (1999) suggested to estimate the exposure effect by a weighted average of model-specific effect estimates, again using the model's posterior probabilities as weights. Viallefont, Raftery, and Richardson (2001) applied this method to estimate an exposure's odds ratio in case-control studies. Other applications include air pollution research (Clyde, 2000; Koop and Tole, 2004).

However, though effective in some cases, traditional implementations of BMA can face severe limitations in effect estimation. Most of these can be traced to the fundamental difficulty arising with the fact that regression coefficients may have a different interpretation across models, a fact only

recently being introduced explicitly in the specification of prior distributions (Consonni and Veronese, 2008). Crainiceanu et al. (2008) noted that model uncertainty methods useful in prediction may not generally perform well in adjustment uncertainty. They introduced a two-step approach (CDP) to estimate an exposure effect accounting for adjustment uncertainty. In the first step, this approach regresses exposure on a large set of potential confounders and selects confounders that are associated with exposure. In the second step, it regresses outcome on exposure, after including the confounders identified in the first step. Compared to this approach, traditional BMA with vague priors on the model space did not perform well. This is because the posterior model probabilities used to weight the model-specific estimates of the exposure effect might not reflect the model's ability to estimate the exposure effect, properly adjusting for confounding. For example, it can be that large weights are assigned to models that do not adequately adjust for confounders, leading to a biased estimate of the exposure effect. This problem may become more serious when limited prior information is available on the effect of interest.

Here, we develop a novel Bayesian approach to adjustment uncertainty, which we call "Bayesian adjustment for confounding" (BAC). We consider the selection of confounders as a random variable, as in BMA, while overcoming the pitfalls described earlier. Our method makes explicit allowance for the fact that the interpretation of the effects can vary across models. BAC addresses this by explicitly focusing on models that are fully adjusted for confounding. Our technique generalizes BMA to simultaneous modeling of the exposure and the outcome. Our approach is based on specifying two models: (1) the outcome as a function of the exposure and the potential confounders (the outcome model); and (2) the exposure as a function of the potential confounders (the exposure model). The key to our approach is the specification of a prior distribution such that, conditional on a predictor's inclusion in the exposure model, the same predictor should also have a higher probability to be included in the outcome model. To this end, our prior specification includes a dependence parameter, ω, representing the odds of including a predictor in the outcome model given that the same predictor is in the exposure model. This leads to a model-weighting strategy for effect estimation accounting for adjustment uncertainty. This strategy assigns high weights to models that are likely to include all the necessary confounders. Our method is explicitly designed to provide competitive results even without strong prior information on the magnitude of the effect.

Although we do not take a causal inference perspective, our method has points of contacts with causal inference methodologies that are based on joint modeling of exposure and outcome as functions of confounders (Rosenbaum and Rubin, 1983; Robins, Mark, and Newey, 1992) and with their Bayesian counterparts (McCandless, Gustafson, and Austin, 2009). This literature strongly emphasizes, as we do, the critical role of model specification and the need for robustness to the choice of confounders (Rubin, 1997; Bang and Robins, 2005; Greenland, 2008). From this perspective, our methodology achieves a combination of three desirable properties: effect estimation efficiency, via the exposure model; variable selection robustness, achieved by allowing the selection to be a random variable; and bias reduction, achieved by including prior information to favor predictors of exposure in the selection of variables for the outcome model.

2. Bayesian Adjustment for Confounding

2.1 *Models*

We build a model for estimating the effect of exposure, or treatment, X on outcome Y. We also have information on a set of M potential confounders $\boldsymbol{U} = \{U_1, \ldots, U_M\}$ identified because they are likely to affect Y, though their effects could be weak. *A priori*, there may be uncertainty about whether potential confounders should be adjusted for in effect estimation.

Although many of our ideas are more general, we discuss our approach in the context of simultaneous linear regression models with two equations, namely, one for exposure and one for outcome. In each equation, potential confounders are either included or excluded, depending on unknown vectors of indicators $\alpha^X \in \{0,1\}^M$ and $\alpha^Y \in \{0,1\}^M$. Here, $\alpha^X_m = 1$ (or $\alpha^Y_m = 1$) whenever U_m is included in the exposure (or outcome) model. For brevity, we refer to the parameters, α's as "models." Conditional on unknown parameters (indicated by Greek letters), and confounders, the regression equations for exposure X_i and outcome Y_i are,

$$E\{X_i\} = \sum_{m=1}^{M} \alpha^X_m \delta^{\alpha^X}_m U_{im}\,, \tag{1}$$

$$E\{Y_i|X_i\} = \beta^{\alpha^Y} X_i + \sum_{m=1}^{M} \alpha^Y_m \delta^{\alpha^Y}_m U_{im}\,, \tag{2}$$

where i indexes the sampling unit. For regression coefficients, β and δ, we use a notation that explicitly keeps track of the fact that those coefficients differ in meaning with the α's. This is especially important when one attempts to make inferences that involve estimates of the exposure effect obtained using different models. Intercept columns can be included among the U's. Some α^Y_m's can be set to one, if confounders are deemed required.

In developing a model for effect estimation, when a true confounder is added or removed from the regression model, the interpretation of the exposure coefficient changes; however, when a model includes all true confounders, and one adds an additional variable that is not associated with X or that is not associated with neither X nor Y, the interpretation of the exposure coefficient does not change. This is in contrast to prediction, where the predicted quantities typically maintain the same interpretation across models.

Thus, when studying confounding adjustment, it is useful to consider the smallest outcome model that includes all the necessary confounders. We denote it by $\boldsymbol{\alpha}^Y_*$, and refer to it as the minimal model. The estimand of interest—the true effect of X on Y, is the coefficient of X in this model, or $\beta_* = \beta^{\alpha^Y_*}$. If there are interactions between exposure and confounders, the estimands are model coefficients of both the main effect and the interaction terms. Without loss of generality, we will focus on the situation where there are no interaction terms. Our goal is estimation of β_* when $\boldsymbol{\alpha}^Y_*$ is unknown. A key observation is that all models that contain at least as many confounders as the minimal model will provide estimates of

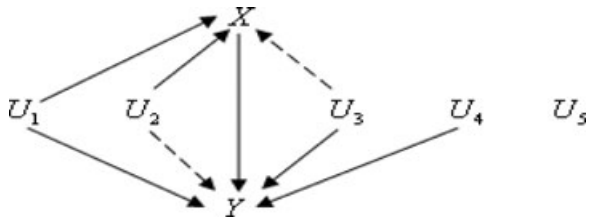

Figure 1. An illustrative example. Solid arrows indicate strong correlation, and dashed arrows indicate weak correlation.

Table 1
The correlation matrix of the simulated data set in Section 2.2

	X	U_1	U_2	U_3	U_4	U_5	Y
X	1.00	0.57	0.58	0.04	0.01	−0.01	0.41
U_1	0.57	1.00	0.00	−0.06	0.03	−0.03	0.51
U_2	0.58	0.00	1.00	−0.02	0.01	0.04	0.09
U_3	0.04	−0.06	−0.02	1.00	0.02	−0.03	0.48
U_4	0.01	0.03	0.01	0.02	1.00	−0.01	0.50
U_5	−0.01	−0.03	0.04	−0.03	−0.01	1.00	−0.02
Y	0.41	0.51	0.09	0.48	0.50	−0.02	1.00

the exposure effect that are also interpretable as estimates of β_*. On the other hand, a model that does not include the minimal model, that is, a model that excludes at least one true confounder, will provide estimates of a parameter that is not the estimand of interest.

2.2 *A Basic Illustration*

It is useful to illustrate our approach using a simple example. Consider the situation depicted in Figure 1 —U_1 is strongly correlated with both exposure and outcome; U_2 is strongly correlated with exposure, but weakly with outcome; U_3 is strongly correlated with outcome and weakly correlated with exposure; U_4 is strongly correlated with outcome and uncorrelated with exposure; and finally, U_5 is uncorrelated with both.

In this example, U_1, U_2, and U_3 are the true confounders of the effect of X on Y and the minimal model that can provide a correctly adjusted effect is $\alpha_*^Y = (1,1,1,0,0)$. The true model is $\boldsymbol{\alpha}^Y = (1,1,1,1,0)$; this "includes" $\boldsymbol{\alpha}_*^Y$, that is, it includes all the variables in α_*^Y. In addition, the true model also includes U_4. Because U_4 is not correlated with X, the interpretation of β^{α^Y} is the same as that of $\beta^{\alpha_*^Y}$. Therefore, the true model also allows for correct adjustment. Because U_4 is correlated with Y, including it can improve overall model fitting, which may yield smaller standard error of the X coefficient estimate. Thus, the true model may potentially lead to greater efficiency than the minimal model, though greater efficiency is not guaranteed in a finite sample. The full model $\alpha^Y = (1,1,1,1,1)$ also contains $\boldsymbol{\alpha}_*^Y$ and a correctly defined coefficient. On the other hand, a model that does not include α_*^Y will estimate a parameter that is not properly adjusted for confounding. For example, the model $\alpha^Y = (1,0,1,1,0)$ will estimate a β^{α^Y} that is not adjusted by U_2, which is an important confounder. Nonetheless, it may still be a useful model for prediction and may receive relatively strong support from the data.

To illustrate, we construct a simulated data set where the variables satisfy the relationships of Figure 1, using the correlations of Table 1, and regressions, as,

$$\begin{aligned} X_i &= \delta_1^X U_{i1} + \delta_2^X U_{i2} + \delta_3^X U_{i3} + \epsilon_i^X \\ Y_i &= \beta X_i + \delta_1^Y U_{i1} + \delta_2^Y U_{i2} + \delta_3^Y U_{i3} + \delta_4^Y U_{i4} + \epsilon_i^Y, \end{aligned} \quad (3)$$

where $i = 1,\ldots,1000$, $\epsilon_i^X, \epsilon_i^Y$ are independent $N(0,\sigma_X^2)$ and $N(0,\sigma_Y^2)$, respectively, and the U_m's are independent $N(0,\sigma_U^2)$. We set $\delta^X = (1,1,0.1)$, $\delta^Y = (1,0.1,1,1)$, $\beta = 0.1$, and $\sigma_X^2 = \sigma_Y^2 = \sigma_U^2 = 1$. Using data so generated, we estimate β by maximum likelihood using two models—one is the true model and the other is the smaller model $\alpha^Y = (1,0,1,1,0)$, which, unlike (3) does not include the true confounder U_2. Results are summarized in Table 2.

The BICs (Schwarz, 1978) for the true model and the smaller model are similar (2882.228 for true model and 2878.249 for smaller model), indicating that they fit the data comparably. The likelihood ratio test for the difference between them has p-value 0.087. However, the two models provide widely different estimates of β. The estimate from true model is 0.121 (95% CI 0.059–0.183), whereas that from smaller model is 0.160 (95% CI 0.116–0.204). In fact, the two estimates have different interpretations. In this case, only the larger and true model provides an estimate of the exposure effect that is properly adjusted for confounding. This simple example illustrates that model selection approaches for adjustment uncertainty in effect estimation should be different from model selection approaches whose goal is prediction of the outcome. In the former, models are valuable to the extent that they correctly estimate a single parameter of interest. In the latter, models are valuable to the extent that they accurately predict the outcome—which can often be achieved even by models that provide systematically biased estimates of the exposure effect.

2.3 *Prior Distributions and Implementation of BAC*

The importance of including in the outcome model all the potential confounders that belong to the minimal model suggests that an approach that acknowledges the fact that only a fraction of the models harbor the coefficient of interest with the correct interpretation, could be successful in addressing adjustment uncertainty from a Bayesian standpoint. We propose to pursue this idea via a novel approach called BAC, which jointly considers the exposure and outcome models, as in equations (1) and (2), and includes unknown model selection parameters, α^X and α^Y. We specify a prior distribution on $\alpha^Y|\alpha^X$, such that

$$\begin{aligned} \frac{P\big(\alpha_m^Y = 1\big|\alpha_m^X = 1\big)}{P\big(\alpha_m^Y = 0\big|\alpha_m^X = 1\big)} &= \omega, \\ \frac{P\big(\alpha_m^Y = 1\big|\alpha_m^X = 0\big)}{P\big(\alpha_m^Y = 0\big|\alpha_m^X = 0\big)} &= 1, \quad m = 1,\ldots,M, \end{aligned} \quad (4)$$

where $\omega \in [1,\infty]$ is a dependence parameter denoting the prior odds of including U_m into the outcome model, when U_m is included in the exposure model. When $\omega = \infty$, the first equation in (4) becomes $P(\alpha_m^Y = 1|\alpha_m^X = 1) = 1$, and requires that any

Table 2

Comparison of model posteriors from BMA, BAC, and TBAC. The estimate of β from BMA is 0.157 with 95% credible interval (0.105, 0.203), that from BAC is 0.121 with 95% credible interval (0.059, 0.182), and that from TBAC is 0.121 with 95% credible interval (0.059, 0.183). BMA is implemented forcing the exposure to always be in the model (FBMA). The dependence parameters, ω, in both BAC and TBAC are set to ∞

Model	$\widehat{\beta}$	95% Confidence interval	BIC	BMA weight	BAC weight	TBAC weight
(1,1,1,1,0; true model (3))	0.121	(0.059, 0.183)	2882.228	0.060	0.985	0.970
(1,0,1,1,0)	0.160	(0.116, 0.204)	2878.249	0.927	0.000	0.000
(1,1,1,1,1)	0.122	(0.060, 0.184)	2888.834	0.001	0.015	0.030
(1,0,1,1,1)	0.160	(0.116, 0.204)	2884.771	0.012	0.000	0.000
(1,1,1,0,0)	0.096	(0.009, 0.183)	3545.253	0.000	0.000	0.000

Note: The weight in each of the three methods is defined as $P(\alpha^Y|D)$, the posterior of α^Y. This posterior is calculated differently in each method. The posterior from BMA is calculated using a uniform prior on α^Y; that from BAC is calculated from the marginal of $P(\alpha^X, \alpha^Y|D)$, where the prior of $P(\alpha^X, \alpha^Y)$ is defined in equation (5); and that from TBAC is calculated by using $P(\alpha^Y|\boldsymbol{X})$ defined in equation (7) as the prior on α^Y.

U_m for which $\alpha_m^X = 1$ is automatically included in the outcome model. When $1 < \omega < \infty$, our prior on $\alpha^Y|\alpha^X$ provides a chance to rule out the predictors that are only associated with X but not associated with Y. To account for the feedback effect of α^Y on α^X, we also set

$$\frac{P\big(\alpha_m^X = 1\big|\alpha_m^Y = 0\big)}{P\big(\alpha_m^X = 0\big|\alpha_m^Y = 0\big)} = \frac{1}{\omega}, \quad \frac{P\big(\alpha_m^X = 1\big|\alpha_m^Y = 1\big)}{P\big(\alpha_m^X = 0\big|\alpha_m^Y = 1\big)} = 1,$$

to assign low probabilities for predictors not selected by the outcome model to be included in the exposure model. The joint prior of (α^X, α^Y) implied by these conditional specifications is,

$$\begin{aligned} P\big(\alpha_m^X = 0, \alpha_m^Y = 0\big) &= P\big(\alpha_m^X = 0, \alpha_m^Y = 1\big) \\ &= P\big(\alpha_m^X = 1, \alpha_m^Y = 1\big) = \omega/(3\omega + 1) \\ P\big(\alpha_m^X = 1, \alpha_m^Y = 0\big) &= 1/(3\omega + 1). \end{aligned} \tag{5}$$

The conditional prior of α^Y given α^X in (4) plays a key role in approximating the marginal posterior distribution of the exposure coefficient under the minimal model, β_*,

$$P(\beta_*|D) = \sum_{\alpha^Y} P(\beta_*|\boldsymbol{\alpha}^Y, D) P(\boldsymbol{\alpha}^Y|D),$$

where $D = (\boldsymbol{X}, \boldsymbol{Y})$ contains vectors of observed data for X and Y. Our analysis is also conditional on observed data for potential confounders U, and they will not be noted in posteriors for simplicity of notation. When ω is large, the conditional prior in (4) greatly increases the chance for predictors strongly correlated with X to be included in the outcome model. These predictors are confounders if they are also correlated with Y. Therefore, the prior leads to a posterior distribution of α^Y ($P(\alpha^Y|D)$) that assigns mass mostly to models that are fully adjusted for confounding, that is, models containing the minimal model. For these models, $\beta^{\alpha^Y} = \beta_*$ so that $P(\beta_*|\boldsymbol{\alpha}^Y, D) = P(\beta^{\alpha^Y}|\boldsymbol{\alpha}^Y, D)$. Therefore, approximately,

$$P(\beta_*|D) \doteq \sum_{\alpha^Y} P(\beta^{\alpha^Y}|\boldsymbol{\alpha}^Y, D) P(\boldsymbol{\alpha}^Y|D), \tag{6}$$

where $P(\beta^{\boldsymbol{\alpha}^Y}|\boldsymbol{\alpha}^Y, D)$ can be directly estimated from observed data. This approximation will be further discussed in Section 3.

Our goal is to calculate the posterior distribution of the parameters of interest $(\alpha^X, \alpha^Y, \beta_*)$ in equations (1) and (2). In our implementation, we assume the following priors for model parameters: $\boldsymbol{\delta}^{\boldsymbol{\alpha}^X}|(\boldsymbol{\alpha}^X, \tau_X) \sim N(\boldsymbol{\mu}_{0\alpha^X}, (\tau_X)^{-1}\phi^2 \boldsymbol{\Sigma}_{0\alpha^X})$, $(\beta^{\alpha^Y}, \boldsymbol{\delta}^{\boldsymbol{\alpha}^Y})|(\boldsymbol{\alpha}^Y, \tau_Y) \sim N(\boldsymbol{\mu}_{0\alpha^Y}, (\tau_Y)^{-1}\phi^2 \boldsymbol{\Sigma}_{0\alpha^Y})$, $\tau_X, \tau_Y \sim$Gamma$(\nu/2, \nu\lambda/2)$, where ν, λ, ϕ, the M-vector $\boldsymbol{\mu}_{0\alpha^X}$, the $(M+1)$-vector $\boldsymbol{\mu}_{0\alpha^Y}$, the $M \times M$-matrix $\boldsymbol{\Sigma}_{0\alpha^X}$, and the $(M+1) \times (M+1)$-matrix $\boldsymbol{\Sigma}_{0\alpha^Y}$ are hyperparameters that are selected as in Raftery et al. (1997). To implement the Markov chain Monte Carlo (MCMC) algorithm, we make the following assumptions:

A1: (β^{α^Y}, X) are independent of α^Y given $(\boldsymbol{\alpha}^X, \widetilde{Y})$, where $\widetilde{Y} = Y - \beta^{\alpha^Y} X$.
A2: X is independent of α^Y given α^X.
A3: (β^{α^Y}, Y) are independent of α^X given (α^Y, X).
A4: $\widetilde{Y}$ is independent of α^X given α^Y.

The assumptions can be interpreted as follows. A1: Given a known $\widetilde{Y}$ and a known exposure model, the selection of the outcome model should no longer depend on the exposure and its effect on Y. A2: Given that we know the covariates that are included in the exposure model (i.e., α^X), the outcome model should not provide additional information on X. The two remaining assumptions can be interpreted similarly, except that they are conditioning on the outcome model instead of the exposure model.

We use a MCMC algorithm to draw posterior samples of $(\boldsymbol{\alpha}^X, \boldsymbol{\alpha}^Y, \beta^{\alpha^Y})$ to approximate $P(\alpha^X, \alpha^Y, \beta_*|D)$. These posterior samples are obtained by iteratively sampling from $P(\boldsymbol{\alpha}^X|\beta^{\boldsymbol{\alpha}^Y}, \boldsymbol{\alpha}^Y, D)$, $P(\boldsymbol{\alpha}^Y|\beta^{\alpha^Y}, \boldsymbol{\alpha}^X, D)$ and $P(\beta^{\alpha^Y}|\boldsymbol{\alpha}^X, \boldsymbol{\alpha}^Y, D)$. Sampling from the first two full conditionals is achieved by the MC^3 method (Madigan and York, 1995). The derivation of these full conditionals is described in Web Appendix A.

2.4 *Two-stage Bayesian Adjustment for Confounding (TBAC)*

In this subsection, we consider a second approach which, when calculating the posterior distribution of $(\beta^{\alpha^Y}, \boldsymbol{\alpha}^X, \boldsymbol{\alpha}^Y)$, cuts the feedback from α^Y to α^X. This approach, called two-stage BAC (TBAC), treats the exposure and outcome models separately in two stages.

TBAC requires Assumption A2 as well as the following assumption:

A1′: β^{α^Y} is independent of α^Y given $\widetilde{Y}$.

Assumption A1′ is similar to Assumption A1 except that X is not taken into account because TBAC will treat X as fixed when considering the outcome model in its second stage.

In stage one of TBAC, we specify a uniform prior on α^X, a conditional prior on $\alpha^Y|\alpha^X$ as defined in equation (4) and use the exposure model only to calculate $P(\alpha^X|\boldsymbol{X})$ and $P(\alpha^Y|\boldsymbol{X})$. These two posterior distributions are calculated as follows:

$$P(\boldsymbol{\alpha}^X|\boldsymbol{X}) \propto P(\boldsymbol{X}|\boldsymbol{\alpha}^X)P(\boldsymbol{\alpha}^X)$$

$$P(\boldsymbol{\alpha}^Y|\boldsymbol{X}) = \sum_{\alpha^X} P(\boldsymbol{\alpha}^Y|\boldsymbol{\alpha}^X, \boldsymbol{X})P(\boldsymbol{\alpha}^X|\boldsymbol{X}) \overset{\text{using A2}}{=\!=} \sum_{\alpha^X} P(\boldsymbol{\alpha}^Y|\boldsymbol{\alpha}^X)P(\boldsymbol{\alpha}^X|\boldsymbol{X}), \tag{7}$$

where the expression of $P(\boldsymbol{X}|\alpha^X)$ is given in Web Appendix A.

In stage two of TBAC, we use $P(\alpha^Y|\boldsymbol{X})$ as prior on α^Y and approximate $P(\alpha^Y, \beta_*|D)$ by $P(\boldsymbol{\alpha}^Y, \beta^{\alpha^Y}|D)$. We assume the same prior distributions for model parameters as in BAC and implement two separate MCMC algorithms for each of the two stages. Details on the sampling algorithms are described in Web Appendix A.

TBAC can be considered as a BMA method on the outcome model with an informative model prior $P(\alpha^Y|\boldsymbol{X})$ obtained from stage one. This prior is the key difference between TBAC and traditional BMA, in which a flat uniform prior on the outcome model is typically assumed. In the following section, we will provide a detailed comparison between BAC/TBAC and BMA.

3. Relation to BMA

In the context of effect estimation, several authors (Raftery, 1995; Hoeting et al., 1999) suggested to calculate the posterior distribution of the effect by taking an average over models, weighted by their posterior probabilities,

$$\sum_{\alpha^Y} P(\beta^{\boldsymbol{\alpha}^Y}|\boldsymbol{\alpha}^Y, \boldsymbol{Y})P(\boldsymbol{\alpha}^Y|\boldsymbol{Y}). \tag{8}$$

This corresponds to marginalization, according to the law of total probabilities, but only if the parameters β^{α^Y} have the same interpretation.

From the perspective of adjustment uncertainty, (8) can be decomposed into two parts, which are, the sum over models that include the correct estimand, and the rest. That is,

$$\sum_{\alpha^Y \supseteq \alpha_*^Y} P(\beta_*|\boldsymbol{\alpha}^Y, \boldsymbol{Y})P(\boldsymbol{\alpha}^Y|\boldsymbol{Y}) + \sum_{\alpha^Y \not\supseteq \alpha_*^Y} P(\beta^{\boldsymbol{\alpha}^Y}|\boldsymbol{\alpha}^Y, \boldsymbol{Y})P(\boldsymbol{\alpha}^Y|\boldsymbol{Y}), \tag{9}$$

where $\alpha \supseteq \alpha'$ indicates that model α contains all the variables that are also contained in model α'. The second term of (9) averages across models that do not include $\boldsymbol{\alpha}_*^Y$, and therefore, do not estimate the same effect.

In BMA, one needs to be careful about not assigning large weights to the models in the second term of equation (9). A common practice in traditional implementations of BMA is to use uniform, or highly dispersed, priors on the α^Ys and often on the effect of interest as well. When the prior is the same for all models, the ratio of the weights given to models α_1 and α_2 is the Bayes Factor $(P(\boldsymbol{Y}|\alpha_1)/P(\boldsymbol{Y}|\alpha_2)$; Kass and Raftery, 1995) and the posterior model probabilities in BMA are driven by a model's predictive ability, which may differ from its ability to properly adjust for confounding in effect estimation.

To illustrate, the fifth column in Table 2 lists model weights used by BMA in the simulated data set in Section 2.2. Most of the weight (92.7%) is assigned to model $(1, 0, 1, 1, 0)$, which does not include all requisite confounders, and estimates the effect at 0.160 (95% CI 0.116–0.204). In contrast, only 6.0% of the weight is assigned to the true model (3) which estimates the correct β_*. Thus, the BMA estimate of β (which is equal to 0.157) is severely biased and its associated 95% credible interval (0.105, 0.203) does not cover the true value of 0.1. We repeated the simulation 1000 times. The coverage rate for the 95% credible interval is only 0.79.

BAC and TBAC are constructed using the same general principles as BMA, but, in our view, offer a far more appropriate prior for the model α^Y. The conditional prior $P(\alpha^Y|\alpha^X)$ defined in equation (4) includes BMA as a special case of $\omega = 1$, where a flat uniform prior is assigned to α^Y. But when ω is larger than one, the prior of $\alpha^Y|\alpha^X$ is informative and incorporates information on which U′s are good predictors of X. TBAC exploits the exposure model to identify confounders highly correlated with X. Some of these confounders, if weakly correlated with Y, may not be identified by the outcome model alone. BAC shares the same property as TBAC, and in addition uses a full Bayesian approach in its implementation, which includes feedback from the outcome model to the exposure model. Therefore, compared to BMA, BAC, and TBAC attempt to place most of the posterior weights $P(\alpha^Y|D)$ on the first term in equation (9) and away from the second. To illustrate, Table 2 lists the model posterior weights based on BAC—98.5% of the weight is assigned to the true model, compared to only 6.0% assigned to the same model as the one selected by BMA. No weight is assigned to models not nesting the minimal model, compared to 93.9% in total assigned by BMA. This result illustrates that linking the two variable selection problems can assign large weights to models including the minimal model, in cases when BMA can fail to do so. This is also the heuristic behind approximation (6).

4. Simulations

In this section, we conduct simulation studies to illustrate and compare the practical properties of BAC, TBAC, CDP (a two-step frequentist approach accounting for adjustment uncertainty by Crainiceanu et al., 2008), traditional BMA (Raftery, 1995; Hoeting et al., 1999), and standard stepwise selection (Mickey and Greenland, 1989). We consider two simulation scenarios. The first shows that BMA can provide a very biased estimate of the exposure effect even under a very simple setting with only two confounders in the true model. In contrast, BAC can fully adjust for confounding and

Table 3

Comparison of estimates of β from six methods, along with the gold standard (true model) in the first simulation scenario. BIAS is the difference between the mean of estimates of β and the true value, SEE is the mean of standard error estimates, SSE is the standard error of the estimates of β, MSE is the mean square error, and CP is the coverage probability of the 95% CI or credible interval

Method		BIAS	SEE	SSE	MSE	CP
True model		0.000	0.044	0.044	0.002	0.95
BAC	$\omega = \infty$	0.000	0.044	0.044	0.002	0.94
	$\omega = 10$	0.018	0.047	0.050	0.003	0.91
	$\omega = 4$	0.027	0.046	0.052	0.003	0.87
	$\omega = 2$	0.034	0.045	0.052	0.004	0.84
TBAC	$\omega = \infty$	0.000	0.044	0.044	0.002	0.95
	$\omega = 10$	0.018	0.047	0.050	0.003	0.92
	$\omega = 4$	0.026	0.046	0.051	0.003	0.89
	$\omega = 2$	0.034	0.045	0.051	0.004	0.84
CDP		0.000	0.044	0.045	0.002	0.95
FBMA		0.041	0.044	0.051	0.004	0.78
NBMA		−0.009	0.050	0.074	0.006	0.72
Stepwise		0.019	0.039	0.058	0.004	0.72

provide unbiased parameter estimates. The second shows similar results in a more complex setting.

In our simulations, we consider both BAC and TBAC with $\omega = 2, 4, 10$ or ∞. For BMA, we consider two different implementations—the first is forcing the exposure to always be in the model (FBMA), whereas the second (NBMA) is not. For the stepwise method, the threshold for adding a variable into the model is taken as 0.20, and the threshold for removing a variable is taken as 0.05 (Mickey and Greenland, 1989).

Our first scenario is similar to the one in Crainiceanu et al. (2008) and considers the true model, $Y_i = \beta X_i + \delta_1^Y U_{1i} + \delta_2^Y U_{2i} + \epsilon_i^Y$, where $i = 1, \ldots, 1000$, and ϵ_i^Y are independent $N(0, 1)$. (X_i, U_{1i}, U_{2i}) are independent normal vectors with mean zero and a covariance matrix, $\mathbf{\Sigma} = (\sigma_{kl})_{3\times 3}$, where $\sigma_{kk} = 1, k = 1, 2, 3$, $\sigma_{12} = \sigma_{21} = \rho$, and $\sigma_{13} = \sigma_{23} = \sigma_{31} = \sigma_{32} = 0$. The set of potential confounders, $\boldsymbol{U}$, includes U_1, U_2 as well as 49 additional independent $N(0, 1)$ random variables. In our simulation, ρ is set to 0.7 and $\beta = \delta_1^Y = \delta_2^Y = 0.1$. We generated 500 data sets. For each, we calculated the maximum likelihood estimate (MLE) of β from the true model and compared it with the estimates from six estimation methods: BAC, TBAC, CDP, FBMA, NBMA, and stepwise selection. The results are summarized in Table 3.

Unless noted, BAC and TBAC will refer to the special case of $\omega = \infty$ in the rest of this section. BAC, TBAC, and CDP produce very similar estimates, both close to the estimates obtained from the true model. All these methods have point estimates around 0.1, the true value of β. Their MSEs are also similar to each other. In contrast, the mean of point estimates based on FBMA are much larger than 0.1, indicating that FBMA systematically overestimates the exposure effect in this example. The MSE of FBMA is also higher. The mean of point estimates based on NBMA is 0.091, which is close to the means from BAC and TBAC. Despite this good average behavior, NBMA produces the worst results. The MSE of NBMA is 0.006, which is much higher than 0.002 for BAC and TBAC. The distribution of the point estimates from NBMA reveals why NBMA has small bias and large MSE: whereas it is centered roughly around the true value, this value falls in a region of low mass. Thus, NBMA rarely provides an estimate close to the true value, even though the average of the point estimates across data sets is close. The point estimates based on the stepwise method are systematically larger than 0.1. The MSE is higher than that of the true model.

The difference between BAC, TBAC, and CDP on one side, and BMA and stepwise approaches on the other is even more pronounced when comparing CIs or credible intervals (both referred to as CI for brevity). The coverage probabilities of 95% CIs based on BAC, TBAC, and CDP are close to 0.95, the desired value. In contrast, the coverage probabilities of FBMA and NMBA are only 0.78 and 0.72, respectively.

It is interesting to investigate the impact of the dependence parameter, ω, on confounding adjustment in BAC and TBAC. As ω decreases from ∞ to 2, the connection between exposure model and outcome model becomes weaker. The estimates, therefore, become closer to those from BMA. The biases increase from 0.000 to 0.034, the MSEs increase from 0.002 to 0.004, and the coverage probabilities drop from 0.94 to 0.84 in BAC and from 0.95 to 0.84 in TBAC. The results show that ω controls the degree of confounding adjustment, with $\omega = \infty$ providing the fullest adjustment in this scenario.

Our second simulation scenario considers a larger number of potential confounders that are correlated with the exposure and also with the outcome. We consider both variables that are strongly and weakly correlated with exposure, and assume the following true outcome model: $Y_i = \beta X_i + \delta_1^Y U_{1i} + \cdots + \delta_{14}^Y U_{14i} + \epsilon_i^Y$, where $i = 1, \ldots, 1000$, ϵ_i^Y are independent $N(0, 1)$, and $(X_i, U_{1i}, \ldots, U_{7i})$ are independent normal vectors with mean zero and a covariance matrix, $\mathbf{\Sigma} = (\sigma_{kl})_{8\times 8}$, where $\sigma_{kl} = 1$ if $k = l$ or $\sigma_{kl} = \rho^{k+l-2}$ if $k \neq l$, $1 \leq k, l \leq 8$. We also assume that the $U_{8i}, \ldots, U_{14i}$ independently follow $N(0, 1)$ distribution and are independent of $(\boldsymbol{X}_i, \boldsymbol{U}_{1i}, \ldots, \boldsymbol{U}_{7i})$. The set of potential confounders $\boldsymbol{U}$ includes $U_1, \ldots, U_{14}$ as well as 43 additional independent $N(0, 1)$ random variables that are independent with both X and Y. In our simulation, β is set to 0.1, $\delta_1 = \cdots = \delta_{14} = 0.1$ and $\rho = 0.7$. Similarly to the first scenario, we generated 500 data sets. For each simulated data set, we calculated the MLE of β from the known true model and compared it to the estimates from the six methods: BAC, TBAC, CDP, FBMA, NBMA, and stepwise. The results are summarized in Table 4.

The differences we noted between BAC, TBAC, and CDP on one side, and BMA and stepwise on the other, are even more pronounced in this more complex example. The point estimate obtained using FBMA is biased and larger than the point estimate based on the true model. The coverage probabilities of 95% CIs are only 0.55 and 0.63 for FBMA and NBMA, respectively. The point estimate using the stepwise method is also biased. The coverage probability is only 0.66. In contrast, the point estimates based on BAC, TBAC, and CDP are close to those based on the true model, and the coverage probabilities are very close to the desired value. The choice of ω in the priors of BAC and TBAC has a pronounced effect on the estimates. When ω decreases from ∞ to 2, the

Table 4
Comparison of estimates of β from six methods, along with the gold standard (true model) in the second simulation scenario. For BAC and TBAC, ϕ is set to 2.85. For FBMA, several different ϕs are considered

Method		BIAS	SEE	SSE	MSE	CP
True model		0.000	0.051	0.049	0.002	0.96
BAC	$\omega = \infty$	0.009	0.051	0.050	0.003	0.96
	$\omega = 10$	0.045	0.055	0.058	0.005	0.84
	$\omega = 4$	0.064	0.055	0.061	0.008	0.75
	$\omega = 2$	0.080	0.055	0.062	0.010	0.64
TBAC	$\omega = \infty$	0.006	0.051	0.050	0.003	0.97
	$\omega = 10$	0.043	0.055	0.058	0.005	0.85
	$\omega = 4$	0.062	0.055	0.060	0.007	0.76
	$\omega = 2$	0.078	0.055	0.061	0.010	0.66
CDP		0.000	0.051	0.048	0.002	0.97
FBMA	$\phi = 2.85$	0.097	0.054	0.061	0.013	0.55
	$\phi = 1.05$	0.070	0.055	0.060	0.009	0.70
	$\phi = 0.30$	0.039	0.053	0.055	0.005	0.87
	$\phi = 0.10$	0.019	0.046	0.039	0.002	0.96
NBMA		0.056	0.064	0.096	0.012	0.63
Stepwise		0.044	0.043	0.067	0.006	0.66

coverage probability drops from 0.96 to 0.64 in BAC and 0.97 to 0.66 in TBAC.

The performance of BMA depends strongly on the spread of prior. For the Normal-Gamma prior we considered, the spread can be controlled by hyperparameter, ϕ. Following the recommendation by Raftery et al. (1997), we chose $\phi = 2.85$ for BAC, TBAC, and BMA in all the examples in this article. This prior is quite spread with 95% of the mass between -5.27 and 5.27. The FBMA estimate under this prior is significantly biased. But the performance of FBMA improves when a more concentrated prior with smaller ϕ is used. Table 4 lists the estimates of FBMA based on different values of ϕ. When $\phi = 0.1$, with 95% of the mass between -0.19 and 0.19, the FBMA estimates are as good as those based on BAC and TBAC. This suggests that strong prior information, concentrating in the region of the true value, is required for FBMA to have good performance. In contrast, BAC and TBAC provide reasonable estimates even under the most spread prior $\phi = 2.85$. This shows that strong prior information is not a requisite for Bayesian approaches for effect estimation as long as appropriate methods are applied.

We also computed the posterior inclusion probability (Barbieri and Berger, 2004) defined, for the mth confounder, as $p_m = \sum_{\boldsymbol{\alpha}^Y : \alpha^Y_m = 1} P(\boldsymbol{\alpha}^Y | D)$, which is estimated by the proportion of appearances of confounder m in the chain of outcome models. Figure 2 shows the estimated posterior inclusion probabilities for all the confounders, in a simulated data set from our second scenario, using TBAC. The first seven confounders have high posterior inclusion probability, indicating that they are important for estimating the exposure effect β. This is consistent with their high correlation with X.

4.1 *Additional Simulations*

In Web Appendix B, we describe simulations designed to evaluate and compare the performance of BAC priors with different ω's in the presence of predictors correlated with X but not

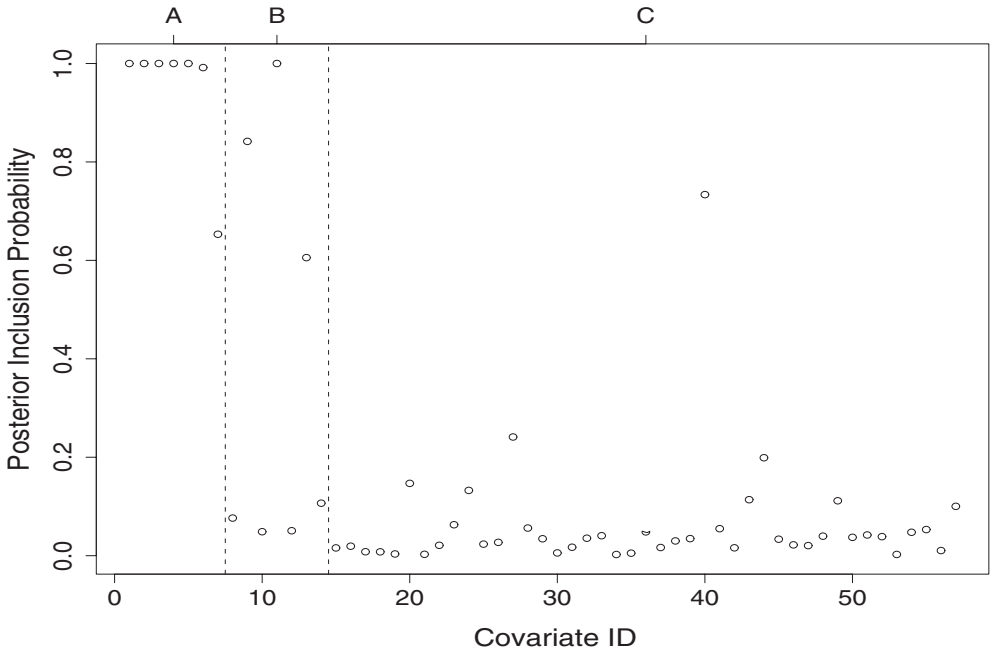

Figure 2. Posterior inclusion probability of potential confounders, separated into three groups by two vertical dashed lines. The first seven (group A) are in the true model and are correlated with X, the next seven (group B) are in the true model but are independent of X, the rest (group C) are not in the true model and are independent of X.

with Y. These predictors are not confounders because they are not associated with Y given X. Including them in the outcome model will not help for confounding adjustment and may decrease the efficiency of effect estimation. We found that using $\omega = 10$ yields smaller MSE compared to $\omega = \infty$. This is because $\omega = 10$ gives a nonzero probability for a predictor included in the exposure model not to be included in the outcome model. In other words, this prior is able to exclude a predictor of X from the outcome model if that predictor is not correlated with Y. Therefore, in the presence of predictors only correlated with X but not with Y, a prior using a finite ω tends to have higher efficiency than $\omega = \infty$.

In Web Appendix C, we describe simulations designed to evaluate the performance of BAC and TBAC when the exposure model is misspecified. A disadvantage with BAC and TBAC is that they require two models, whereas BMA only requires one. However, in our context, this does not come necessarily with an increased risk of model misspecification. Our simulation results show that both BAC and TBAC are robust to misspecification of the exposure model. The key feature in confounding adjustment is to include a sufficient number of confounders. A roughly correct exposure model may often be enough to ensure that this happens.

In Web Appendix D, we describe simulations designed to compare BAC to TBAC when $\omega = \infty$. We found that the two methods behave similarly in the majority of the cases examined here. However, they show some differences when dealing with predictors weakly associated with both X and Y. Compared to TBAC, BAC assigns lower weights to models that include those predictors. As a result, the two methods give somewhat different posterior distributions of $\boldsymbol{\alpha}^X$ and $\boldsymbol{\alpha}^Y$. But because these predictors have limited impact on the estimation of the exposure effect, they still provide very similar exposure effect estimates.

In Web Appendix E, we provided simulation results to compare BAC and TBAC with BMA under the two simulation scenarios described in this section but with a smaller sample size of 100. We found that the MSE from BMA is smaller than that from BAC and TBAC in the first scenario, but is larger in the second scenario. The results indicate that, although BAC and TBAC in general perform better, BMA may sometimes yield smaller MSE when the sample size is small. Combined with results from Web Appendix B, we conclude that there is not a single value of ω that is uniformly optimal in terms of MSE. The choice of ω should depend on sample size, complexity of confounding structure, as well as the bias/variance trade off. And the prior with $\omega = \infty$ is usually conservative, which provides unbiased estimates.

5. Air Pollution Example

In air pollution epidemiology, adjusting for confounding bias is probably the biggest challenge when estimating a small health effect associated with exposure to an environmental contaminant. In addition, because of the heavy policy implications associated with the public health impact of air pollution, most of the epidemiological evidence has been severely challenged by the threat of confounding bias.

In this section, we apply the newly proposed methods (BAC, TBAC) to daily time series data for Nassau County, NY for the period 1999–2005. Although this data analysis is mainly used as an illustration of our newly proposed approach, the results clearly illustrate the potential application and impact of BAC and TBAC in epidemiology studies of observational data. The data include 1532 daily records of emergency hospital admissions, weather variables, and $PM_{2.5}$ levels. A more extensive description of this data set can be found in Dominici et al. (2006). The goal is to estimate the increase in the rate of hospitalizations for cardiovascular disease (CVD) associated with a $10\,\mu g/m^3$ increase in $PM_{2.5}$, while accounting for age-specific longer-term trends, weather and day of the week. The hospitalization rate is calculated separately for each age group ($\geq$75 or not) on each day. In our model, to control for longer-term trends due, for example, to changes in medical practice patterns, seasonality, and influenza epidemics, we include smooth functions of calendar time. We also include a smooth function to allow seasonal variations to be different in the two age groups. To control for the weather effect, we include smooth functions of temperature and dew point. To start, we consider a full model that is large enough to include all the necessary confounders (Dominici et al., 2000, 2004; Peng et al., 2006),

$$\begin{aligned} Y_{at} = {} & \beta\ \mathrm{PM}_{2.5t} + \ \mathrm{DOW}\ + \text{intercept for age group } a \\ & + ns(\mathrm{Temp}_t, df_{\mathrm{Temp}}) + ns(\mathrm{Temp}_{t1-3}, df_{\mathrm{Temp}}) \\ & + ns(\mathrm{Dew}, df_{\mathrm{Dew}}) + ns(\mathrm{Dew}_{t1-3}, df_{\mathrm{Dew}}) + ns(t, df_t) \\ & + ns(t, df_{at}) \times \text{age group}\ + \epsilon_t, \end{aligned}$$

where the outcome $Y_{at} = \sqrt{\text{CVD hospital admissions/size of population at risk}}$ for each age group a ($\geq$75 or not) on day $t(= 1, \ldots, 1532)$. $\mathrm{PM}_{2.5t}$ denotes the level of particulate matter having diameter less than 2.5 μm on day t. DOW are indicator variables for the day of the week. Temp_t and Temp_{t1-3} are the temperature on day t and the three-day running mean, respectively. Dew_t and Dew_{t1-3} are the dew point on day t and the 3-day running mean. The quantity $ns(., df)$ is a natural cubic spline with df degrees of freedom. We include $ns(t, df_t)$, $ns(\mathrm{Temp}_t, df_{\mathrm{Temp}})$, $ns(\mathrm{Temp}_{t1-3}, df_{\mathrm{Temp}})$, $ns(\mathrm{Dew}, df_{\mathrm{Dew}})$, and $ns(\mathrm{Dew}_{t1-3}, df_{\mathrm{Dew}})$ to adjust for the potential nonlinear confounding effects of seasonal variations, temperature and dew point. The quantity $ns(t, df_{at}) \times$ age group is a natural cubic spline of t for the $\geq$ 75 age group to allow its seasonal variation to be different from the other age group. Similar to Crainiceanu et al. (2008), df_{Temp} is set to 12, df_{Dew} is set to 12, df_t is set to 16 per year, and df_{at} is set to 4. These degrees of freedom are considered sufficiently large for the full model to include all the potential confounders (Crainiceanu et al., 2008). The residuals ϵ_t are assumed to be independent and identically distributed with a normal $N(0, \sigma^2)$ distribution. After dropping some potential confounders due to collinearity, we work with a set of 164 potential confounders.

We consider six approaches: BAC, TBAC, CDP, FBMA, NBMA, and stepwise. For BAC and TBAC, we consider priors with $\omega = 2, 4, 10$, or ∞. The estimated $PM_{2.5}$ effect ($\times 10,000$) denoted by $\widehat{\beta}$ is listed in Table 5: BAC, TBAC (with $\omega = \infty$) and CDP provide estimates of the short-term effect of $PM_{2.5}$ on CVD hospital admissions with 95% CIs that do not include

Table 5
Comparison of estimates of $PM_{2.5}$ *effect on CVD hospitalization rate based on BAC, TBAC, CDP, FBMA, NBMA, stepwise, and the full model*

Method		$\widehat{\beta}$	SE($\widehat{\beta}$)	95% CI
Full model		0.291	0.092	(0.110, 0.471)
BAC	$\omega = \infty$	0.226	0.081	(0.067, 0.385)
	$\omega = 10$	0.217	0.079	(0.060, 0.371)
	$\omega = 4$	0.186	0.085	(0.019, 0.351)
	$\omega = 2$	0.155	0.079	(0.007, 0.317)
TBAC	$\omega = \infty$	0.229	0.083	(0.071, 0.403)
	$\omega = 10$	0.216	0.075	(0.071, 0.367)
	$\omega = 4$	0.190	0.080	(0.035, 0.347)
	$\omega = 2$	0.155	0.077	(0.010, 0.313)
CDP		0.221	0.089	(0.045, 0.396)
FBMA		0.140	0.077	(−0.008, 0.298)
NBMA		0.007	0.033	(0.000, 0.131)
Stepwise		0.106	0.066	(−0.023, 0.234)

0. With $\omega = \infty$, both BAC and TBAC provide similar estimates of the exposure effect as CDP. Moreover, all three methods provide smaller standard errors than the one obtained under the full model. In comparison, FBMA and NBMA provide a very different and not statistically significant estimate of the exposure effect. Some confounders known to be important, such as temperature and dew point, are downweighted in BMA. Both temperature and dew point are positively correlated with $PM_{2.5}$ and negatively correlated with hospitalization rate. Failure to include them in the model diminishes the $PM_{2.5}$ effect. This illustrates that in practical applications BMA and BAC can lead to different conclusions. The key difference lies in the linking strength between the exposure model and the outcome model. As the strength decreases, which corresponds to smaller value of ω, the estimates from BAC and TBAC become closer to that from BMA.

6. Discussion

Estimating an exposure effect, while accounting for the uncertainty in the adjustment for confounding, is of essential importance in observational studies. Building upon work by Dominici et al. (2004) and Crainiceanu et al. (2008), in this article, we develop Bayesian solutions to the estimation of the association between X and Y accounting for the uncertainty in the confounding adjustment. Given a set of potential confounders, we simultaneously address model selection for both the outcome and the exposure. Although we discuss our methods in the setting of linear models, BAC and TBAC are general concepts and are not constrained to the linear case. For example, they can be extended to generalized linear models using relatively well understood computational strategies.

Like BMA, BAC, and TBAC take a weighted average over models rather than making inference based on a single model. However, they attempt to provide an estimate of the exposure effect by combining information across regression models that include all the requisite confounders, to ensure that the regression coefficient of interest maintains the same interpretation across models. A nice feature of BMA that is retained by BAC and TBAC is that the importance of confounders can be evaluated based on posterior inclusion probability. This information may reveal underlying connections between exposure and confounders, which may become of interest for future research. BAC and TBAC are more computationally intensive than BMA.

Successful application of BAC and TBAC rely on availability of all confounders. Scientific knowledge is required to ensure that these assumptions are valid. Statistical methods may also help to check whether there is evidence for the existence of unmeasured confounders. For example, one can decompose the association between exposure and outcome into distinct spatio-temporal scales and check for the consistency in the estimation of exposure effect across these spatio-temporal scales (Janes, Dominici, and Zeger, 2007).

If there are no unmeasured confounders, the full model, that is the model including all variables correlated with X and Y, those correlated with Y only, as well as potentially others that are not associated with either, will provide unbiased estimates of the exposure effect. However, using the full model will generally yield wider CIs compared to BAC and TBAC. By combining estimations from different smaller models, especially from models that only include requisite confounders but do not include many unnecessary variables, BAC and TBAC can provide more precise inference than the full model.

TBAC parallels CDP in its two-stage structure, and in the inclusion of variables selected from the exposure model into the outcome model. However, there are also important differences. TBAC provides a model-based solution rather than a partially algorithmic one, and also arguably considers uncertainty more fully in a Bayesian framework. BAC further takes into account the feedback effect and considers a full Bayesian approach. Also, in CDP, models are evaluated based on the change in deviance between sets of increasing dimensionality, a criterion that could lead to different conclusions compared to BAC and TBAC. Large spaces of confounders may potentially be required for CDP users to reliably observe the stabilization of the estimated effect that is required for the method to succeed. However, no restrictions on dimensionality apply to BAC and TBAC. Computationally, CDP is clearly faster, and also offers helpful visualizations. The two methods produce results with similar frequentist properties in our simulation studies.

In the propensity score literature, it is recommended to include variables that are strongly correlated with Y but only weakly correlated with X into the model for calculating the propensity score, as the bias resulting from their exclusion would dominate any loss of efficiency in modest or large studies (Rubin, 1997; Brookhart et al., 2006). One of the strengths of our method, shared by others such as doubly robust estimation (Scharfstein, Rotnitzky, and Robins, 1999), is that we can identify these in a data-based way, rather than having to rely on prior knowledge as required in propensity score adjustment.

An alternative Bayesian variable selection approach is the Bayesian lasso (Park and Casella, 2008), assuming a mixture prior of a point mass at zero and a double exponential distribution for regression coefficients (Hans, 2010). An alternative version of both BAC and TBAC could be constructed using this prior instead. We expect that the use of the Bayesian lasso on the outcome model alone would present similar limitations to traditional BMA, but have not explored this in detail.

In summary, in this article, we have motivated, defined, and evaluated a tool for accounting for uncertainty in the selection of confounders in effect estimation. Our approach adopts the fully probabilistic structure of BMA, without suffering from the pitfalls we highlighted in traditional BMA implementations, and is likely to contribute to a more reasoned and quantitative approach to the specification of models used to determine health effects of common exposures, and the reporting of the associated uncertainty.

7. Supplementary Materials

Web Appendices referenced in Sections 2 and 4 are available under the article Information link at the *Biometrics* website `http://www.biometrics.tibs.org`. An R package implementing BAC and TBAC is available at `http://sweb.uky.edu/`∼cwa236/BEAU/.

Acknowledgements

We thank Ciprian Crainiceanu for his significant input into earlier versions of this manuscript, Eric Tchetgen for a helpful discussion, and David Diez, Roger D. Peng, and David Haws for helpful comments on computer programming. We also thank the Co-Editor, the Associate Editor, and two referees for insightful comments that have substantially improved the article. The work of Francesca Dominici was supported by grants EPA R83622, EPA RD83241701, EPA RD83479801, NIH/NIEHS R01ES012054, NIH/NIEHS R01ES012044, and NIH/NCI P01CA134294.

Conflict of Interest: None declared.

References

Bang, H. and Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. *Biometrics* **61,** 962–973.

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics* **32,** 870–897.

Brookhart, M. A., Schneeweiss, S., Rothman, K. J., Glynn, R. J., Avorn, J., and Sturmer, T. (2006) .Variable selection for propensity score models. *American Journal of Epidemiology* **163,** 1149–1156.

Chipman, H. A., George, E. I., and McCulloch, R. E. (2002). Bayesian treed models. *Machine Learning* **48,** 299–320.

Clyde, M. (2000). Model uncertainty and health effects studies for particulate matter. *Environmetrics* **11,** 745–763.

Consonni, G. and Veronese, P. (2008). Compatibility of prior specifications across linear models. *Statistical Science* **23,** 332–353.

Crainiceanu, C. M., Dominici, F., and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95,** 635–651.

Dominici, F., Samet, J. M., and Zeger, S. L. (2000). Combining evidence on air pollution and daily mortality from the twenty largest U.S. cities: A hierarchical modeling strategy (with discussion). *Journal of the Royal Statistical Society, Series A: Statistics in Society* **163,** 263–302.

Dominici, F., McDermott, A., and Hastie, T. J. (2004). Improved semiparametric time series models of air pollution and mortality. *Journal of the American Statistical Association* **99,** 938–948.

Dominici, F., Peng, R. D., Bell, M., Pham, L., McDermott, A., Zeger, S. L., and Samet, J. M. (2006). Fine particulate air pollution and hospital admission for cardiovascular and respiratory diseases. *The Journal of the American Medical Association* **295,** 1127–1134.

Greenland, S. (2008). Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* **167,** 523–529.

Hans, C. (2010). Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing* **20,** 221–229.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial (with discussion). *Statistical Science* **14,** 382–417.

Janes, H., Dominici, F., and Zeger, S. L. (2007). Trends in air pollution and mortality: An approach to the assessment of unmeasured confounding. *Epidemiology* **18,** 416–423.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90,** 773–795.

Koop, G. and Tole, L. (2004). Measuring the health effects of air pollution: To what extent can we really say that people are dying of bad air. *Journal of Environmental Economics and Management* **47,** 30–54.

Madigan, D. and York, J. (1995). Bayesian graphical models for discrete data. *International Statistical Review* **63,** 215–232.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine* **28,** 94–112.

Mickey, R. M. and Greenland, S. (1989). The impact of confounder selection criteria on effect estimation. *American Journal of Epidemiology* **129,** 125–137.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association* **103,** 681–686.

Peng, R. D., Dominici, F., and Louis, T. A. (2006). Model choice in time series studies of air pollution and mortality. *Journal of the Royal Statistical Society, Series A: Statistics in Society* **169,** 179–203.

Raftery, A. E. (1995). Bayesian model selection in social research. *Sociological Methodology* **25,** 111–163.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association* **92,** 179–191.

Robins, J. M., Mark, S. D., and Newey, W. K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48,** 479–495.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127,** 757–763.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Rejoinder to "Adjusting for nonignorable drop-out using semiparametric nonresponse models". *Journal of the American Statistical Association* **94,** 1135–1146.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6,** 461–464.

Viallefont, V., Raftery, A. E., and Richardson, S. (2001). Variable selection and Bayesian model averaging in case-control studies. *Statistics in Medicine* **20,** 3215–3230.

Yeung, K. Y., Bumgarner, R. E., and Raftery, A. E. (2005). Bayesian model averaging: Development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics* **21,** 2394–2402.

Received April 2010. *Revised March* 2011. *Accepted March* 2011.

Analyses that Inform Policy Decisions

R. Gutman[1,*] and D.B. Rubin[2]

[1]Department of Biostatistics, Brown University, Providence, Rhode Island 02912, U.S.A.
[2]Department of Statistics, Harvard University, Cambridge, Massachusetts 02138, U.S.A.
[*]*email:* rgutman@stat.brown.edu

1. Analyses that Inform Policy Decisions are, de Facto, Causal

We begin by thanking the Co-Editor, David Zucker, for inviting our discussion of Wang, Parmigiani, and Dominici (2012; hereafter WPD), which describes methodology to estimate, using their example, the association between contaminants and hospitalization rates, controlling for "confounders." Our view is that, to inform "major policy-related decisions" (WPD, Section 1), as they describe in their example, it is imperative that the goal is to estimate the effect of intervening to reduce contamination on health-related outcomes; that is, the goal must be to estimate the causal effects of different levels of pollution on outcomes. Yet in Section 1, WPD claim that their analysis does not follow a "causal inference perspective" and that their methodology only estimates the association between contaminant levels and health-related outcomes, "controlling for confounders." However, if the analysis is not causal, and only descriptive, the meaning of "confounders" is baffling to us.

It is our interpretation, because of repeated references to policy implications and uses of the phrase "controlling for confounders," that WPD intends to provide an analysis to estimate causal effects and not merely to estimate associations. As such, our comment will be on the basis of a causal inference perspective, which we summarize in Section 2. Then in Section 3, we embed WPD's approach within this causal inference framework, explicating assumptions that are needed for this embedding. Section 4 presents simple simulations and conclusions.

2. The RCM Framework

A commonly used causal inference framework was partially introduced by Neyman (1923) in the context of randomization-based inference in randomized experiments; Rubin (1974, 1975, 1978) later developed and extended the framework to include observational studies and Bayesian inference. The resulting framework, often called "Rubin's Causal Model" (RCM Holland, 1986), comprises three main parts: the use of potential outcomes to define causal effects, the explication of an Assignment Mechanism (AM—a term coined in Rubin, 1975), and a (Bayesian) model of the "science." For simplicity of exposition, we only deal with situations where the units, indexed by i, $i = 1, \ldots, N$, are modeled as independent given parameters, as in WPD, but we note that this assumption is not innocuous when units are correlated (e.g., represent successive days, as in WPD's example). We try to follow both WPD's notation and example to ease communication.

Suppose there are $D + 1$ possible treatment levels (e.g., levels of pollution) indicated by $X_i \in \{0, \ldots, D\}$. Unit i has $D + 1$ potential outcomes: $Y_i(0), \ldots, Y_i(D)$, representing, in the WPD example, the hospitalization rates when exposed to each of the possible treatment levels, all at the same time after treatment. A common estimand, and one that is related to WPD's estimand, is the average treatment effect, across all N units, obtained by moving from level $X = d_2$ to level $X = d_1$:

$$\beta(d_1, d_2) = \frac{1}{N} \sum_{i=1}^{N} [Y_i(d_1) - Y_i(d_2)]. \tag{1}$$

No $\beta(d_1, d_2)$ can be observed because, in all real life studies, we can only observe one of the potential outcomes for each unit: this is the fundamental problem facing causal inference (Rubin, 1978).

The only way to overcome this problem is to collect data on different units assigned to different treatments. But when comparing observed potential outcomes across units, we must consider the distribution of pretreatment characteristics (i.e., covariates U_i) in groups of units treated differently. The key piece of information that is needed is how each unit received the treatment level it actually received: In the RCM language, we need a model for the AM: $P(X_i \mid U_i, Y_i(0), \ldots, Y_i(D), \phi)$, where ϕ is a vector parameter governing this distribution.

In general, we would like the AM to generate assignments, that given the covariates, U_i, do not depend on the potential outcomes, so that the AM is unconfounded (Rubin, 1990). For example, if we could randomize the pollution levels to different units, perhaps with probabilities that depend on the U_i, the AM would be unconfounded, or more formally:

$$\begin{aligned} &P(X_i = d \mid U_i, Y_i(0), \ldots, Y_i(D), \phi) \\ &\quad = P(X_i = d \mid U_i, \phi) \quad i = 1, \ldots, N;\ d = 1, \ldots, D. \end{aligned} \tag{2}$$

Suppose the AM is unconfounded; also suppose that the number of observed units with covariate value $U_i = u$ that received treatment d, $N_{(u,d)}$, is positive for all d and all u such that

$p(u)$, the proportion of units in the population with covariate value $U_i = u$, is positive. Then $\beta(d_1, d_2)$, can be unbiasedly estimated by

$$\sum_u p(u) \left[\frac{1}{N_{(u,d_1)}} \sum_{U_i=u} Y_i(d_1) - \frac{1}{N_{(u,d_2)}} \sum_{U_i=u} Y_i(d_2) \right]. \quad (3)$$

In the case of completely randomized experiments, we can estimate that $N_{(u,d)} \doteq N_d$, because the distribution of U in all treatment groups is the same in expectation. Without complete randomization, the distribution of U is not necessarily similar in the different treatment groups, and the treatment effect needs to be estimated with appropriate care (for example, by appropriately using propensity score subclassification or matching to balance the distributions of U across treatment groups). Otherwise, the resulting estimate may be badly biased, as illustrated by the simulations in our Section 4.

The third part of the RCM framework incorporates a (Bayesian) model that describes the distribution of the "science," or in other words, a model that attempts to approximate the joint distribution of the potential outcomes and the covariates, $P(Y_i(0), \ldots, Y_i(D), U_i \mid \theta)$, where θ is a vector parameter governing this distribution; commonly, θ and ϕ are chosen to be distinct (Rubin, 1976). This model, together with the AM, generates a model for all possible observed values, $(X_i, U_i, Y_i(0), \ldots, Y_i(D) \mid \phi, \theta)$, given the parameters, where for Bayesian inference the parameters are given a prior distribution.

In our view, the most effective way to generate causal inferences for quantities like $\beta(d_1, d_2)$ is to condition on all observed values and multiply impute the unobserved potential outcomes, as explicated in Section 4.5 of Rubin (2008) and briefly illustrated in our Section 4.

3. Implicit Assumptions Made by WPD for Policy-Relevant Conclusions

The first model described by WPD is:

$$E(X_i \mid U_i, \phi) = \sum_{i=1}^{M} \delta_m^X U_{im}\,, \quad (4)$$

where for notational convenience we suppress their use of α, N is effectively infinite, and we use our notation for parameters on the left-hand side. This model (WPD's "exposure model") describes the expectation of the treatment for unit i given the covariates, or in our terminology, the model for the AM after making the unconfoundedness assumption implicitly: Equation (4) does not depend on the potential outcomes, and therefore conforms to our assumption (2).

The second model (WPD's "outcome model"), our equation (5) below, which again uses our notation for parameters on the left-hand side, describes the expectation of the observed outcome for unit i when receiving treatment level X_i, that is $Y_i(X_i)$, given U_i and parameters

$$E(Y_i(X_i) \mid U_i, \theta, \phi) = \beta X_i + \sum_{i=1}^{M} \delta_m^Y U_{im}. \quad (5)$$

Model (5) in general includes the effect of the AM, because it involves the observed potential outcomes, $Y_i(X_i)$, which depend on the AM; here, the implicit assumption again is unconfoundedness. Also, model (5) explicitly describes only part of the science, because it only specifies how hospitalization rates vary with each level of the treatment (pollution levels) given the covariates; missing from this formulation, and thus implicitly assumed to be irrelevant, is the model for the marginal distribution of the covariates. This latter model is of potential interest, when, for example, the interventions are to be implemented in different areas with different covariate distributions, or when the covariates involve previous pollution levels, as would occur in a time series model (see, e.g., discussion in Section 4.4 of Rubin, 2008).

WPD defines the treatment effect of X on Y as β in equation (5), so that in our causal notation, $\beta(d_1, d_2)$ is assumed to equal β for all levels that differ by one unit of measurement. From a policy-relevant perspective, this definition is only useful when the regression model (5) is approximately true. For example, if we assume that the true model for $Y_i(X)$ given U_i is inverse logistic, then for high and low levels of pollution, small changes in pollution levels would have little effect on the outcomes, whereas for medium levels of pollution, small changes could have large effects. In our experience, we never know the true physical model for the science. Thus, it is difficult to interpret definitions of treatment effects unless they are based on the underlying potential outcomes, because "regression coefficients may have a different interpretation across models" (WPD, Section 1).

In the next section, we summarize a small simulation that illustrates the behavior of WPD's methodology, as well as two other methodologies, when used for estimating causal effects. Our conclusion is that WPD's methodology may result in misleading estimates of policy-relevant effects, because of inappropriate adjustment for confounders, relative to more robust alternatives based on the explicit multiple imputation of missing potential outcomes.

4. Simulation

For simplicity, we assume that the treatment factor has only two levels: $X_i = 0$—low pollution, and $X_i = 1$—high pollution. We also assume that there exists only one confounder, U_i, whose distributions can differ in the two treatment groups: In the control group U_i is Normal(0,1) and in the treatment group U_i is Normal $(B\sqrt{\frac{1+\sigma^2}{2}}, \sigma^2)$, where σ^2 is the ratio of variances in the two distributions, and B is the standardized initial bias. We also assume that we have a sample of n units from the high pollution level, and a sample of n units from the low pollution level, both from a population of size $N = \infty$. The potential outcomes, $Y_i(0), Y_i(1)$, are the hospital admission rates for unit i, with low and high pollution, which have distributions that depend on U_i. The assignment of X_i is unconfounded: U_i is the only relevant covariate and is observed for all $2n$ units. We allow for nonparallel response surfaces at the two levels of pollution:

$$Y_i(0) \mid U_i, \theta_0 = \frac{1}{1 + \exp(-[U_i + U_i^2 - 3])} \quad \text{and}$$
$$Y_i(1) \mid U_i, \theta_1 = \frac{1}{1 + \exp(-[U_i - 1])}, \quad (6)$$

where $Y_i(0)$ and $Y_i(1)$ are conditionally independent given U_i. The estimand is the population average treatment effect from moving from the high pollution level to the low pollution level, $\tau \equiv \beta(0,1)$ defined in our (1).

We assess the frequentist operating characteristics of three estimation procedures at $4 \times 4 \times 3$ different configurations: $\sigma^2 \in \{0.5, 1, 2, 4\} \times B \in \{0, 0.25, 0.5, 1, 2\} \times n \in \{600, 1200, 2400\}$. The first procedure is the simple difference in observed means. The second procedure uses a model, approximating WPD's procedure:

$$E(Y_i(X_i) \mid U_i, \theta) = \beta X_i + ns(U_i, 15), \tag{7}$$

where ns is the natural cubic spline with 15 *df*. The third procedure is based on the ideas presented at the end of our Section 2, and is the main part of a recent Ph.D. thesis (Gutman, 2011), summarized in the following nine steps.

1 Discard units in the treatment and control groups that have no close U_i match in the other group.
2 Partition the remaining units into six subclasses such that there are at least three units from each treatment group in each of the subclasses, and the distributions of U_i in each subclass are similar in the control and treatment groups.
3 From observed data, estimate $E(Y_i(X_i) \mid U_i, \theta) = h_{X_i}(U_i \mid \theta_{X_i})$ independently in each treatment group, $X_i = 0, 1$, using a regression spline with seven knots located at the borders of the subclasses; the location of the knots is selected, by Step 2, so that the distribution of U_i in each subclass is similar in the control and treatment groups.
4 Draw the vector parameters θ_{X_i} ($X_i \in \{0, 1\}$) from independent posterior distributions.
5 Using the drawn value of θ_{X_i} ($X_i \in \{0, 1\}$), independently impute the missing potential outcomes in the treatment group, the $Y_i(0)$ with $X_i = 1$, and in the control group, the $Y_i(1)$ with $X_i = 0$.
6 Repeat steps 4 and 5, $M = 20$ times.
7 Estimate the treatment effect τ and its sampling variance in each of the imputed datasets; from imputed dataset m; let $\hat{\tau}_m$ be the estimated treatment effect, and let $\hat{W}_m = var(\hat{\tau}_m)$ be its estimated sampling variance, $m = 1, \ldots, M$.
8 Estimate the average treatment effect by $\hat{\tau} = \frac{1}{M}\sum_{m=1}^{M} \hat{\tau}_m$ and its standard error by $\sqrt{\frac{1}{M-1}\sum_{m=1}^{M}(\hat{\tau}_m - \hat{\tau})^2 + \frac{1}{M}\sum_{m=1}^{M}\hat{W}_m}$. This combining rule is known as Rubin's Rule for MI (Rubin, 1987).
9 Calculate the interval estimate for τ, using the t approximation for the interval given by Barnard and Rubin (1999).

We drew 1000 replications at each configuration of σ^2, B, n, and recorded whether the resulting 95% interval covered the treatment effect, the size of the bias, and the root mean square error (RMSE).

When $B = 0$, $\sigma^2 = 1$, and for all sample sizes, the observed coverage for the simple difference in means was 95%, which was expected because with this configuration, U_i is independent of the treatment assignment X_i, and thus the situation does not require any form of adjustment, even though U_i influences the outcome; this configuration corresponds to a completely randomized experiment. When $B \neq 0$ or $\sigma^2 \neq 1$, the coverage rates for the observed difference in means were 0% for all sample sizes, which is not surprising because in those configurations, X_i is dependent of U_i, which implies initial bias due to U_i, which requires adjustment to estimate the treatment effect without substantial bias.

Table 1
95% Interval coverage rate, bias and RMSE $n = 600$; cubic spline model (7)

σ^2	B	0	0.25	0.5	1	2
0.5	Coverage	0.05	0.04	0.01	0.00	0.00
	Abs. Bias	14.00	14.35	16.46	30.00	63.14
	RMSE	14.45	14.80	16.88	30.33	63.38
1	Coverage	0.93	0.80	0.28	0.00	0.00
	Abs. Bias	0.84	4.23	12.25	39.58	69.57
	RMSE	6.56	7.54	13.29	39.85	69.78
2	Coverage	0.11	0.25	0.13	0.00	0.00
	Abs. Bias	18.96	15.42	19.18	36.65	59.91
	RMSE	19.89	16.64	20.07	37.04	60.16
4	Coverage	0.00	0.00	0.00	0.01	0.05
	Abs. Bias	57.48	46.27	37.02	36.64	34.73
	RMSE	57.79	46.66	37.53	37.14	35.34

Absolute Bias and RMSE are in 10^{-3}.

Table 2
95% interval coverage rate, bias and RMSE $n = 600$; multiple imputation

σ^2	B	0	0.25	0.5	1	2
0.5	Coverage	0.97	0.96	0.94	0.96	0.98
	Abs. Bias	1.05	0.89	0.92	0.40	0.44
	RMSE	5.00	4.98	5.16	6.11	8.77
1	Coverage	0.94	0.96	0.95	0.96	0.96
	Abs. Bias	0.88	0.64	0.74	0.86	0.93
	RMSE	6.51	6.24	6.31	6.72	8.46
2	Coverage	0.93	0.95	0.96	0.93	0.95
	Abs. Bias	0.89	0.79	0.91	2.21	2.44
	RMSE	8.10	7.60	7.43	7.93	8.47
4	Coverage	0.95	0.96	0.96	0.94	0.92
	Abs. Bias	0.56	0.95	1.04	1.70	2.02
	RMSE	9.37	8.83	8.45	8.57	9.10

Absolute Bias and RMSE are in 10^{-3}.

Table 1 displays the results with $n = 600$ when using the least squares estimate of β from (7). The only configuration for which adjusting for the covariate using (7) results in coverage close to the nominal level occurs when $B = 0$ and $\sigma^2 = 1$, or in other words, when X_i and U_i are independent. Such results are true for larger samples as well. The results are not surprising because of differences in the distributions of the covariate in the control and treated groups, and because we do not know the "true" response surfaces. This phenomenon has been described since Cochran (1968) for a single covariate, and repeatedly in the collection by Rubin (2006) for single and multiple covariates.

Table 2 displays the results with $n = 600$ for our proposed multiple imputation procedure, and shows that the coverage

rates are close to the nominal level; this conclusion is generally maintained for larger sample sizes as well. Moreover, this method obtains close to the correct coverage by having smaller bias and smaller mean square error than model (7), not by increasing the widths of intervals.

A possible criticism of our simulation is that we did not include an interaction of exposure by confounder term in model (7). Our response to this has four parts. First, the number of parameters estimated in model (7) is 18, consisting of an intercept, a treatment indicator, 15 *df* for the spline, and the residuals' variance. With our proposed procedure, we used 16 parameters consisting of an intercept, 6 *df* for the spline, and the residuals' variance, in each of the treatment groups. Thus, the total number of parameters is similar in both procedures.

Second, there are many ways to incorporate an interaction between the exposure and the confounder, for example, by adding the term $X_i \times U_i$, or the term $X_i \times ns(U_i, df)$. Incorporating either of these terms results in models with different numbers of parameters, different estimated treatment effects, and different interpretations of the parameters. The choice of which model to use is not described in WPD. Moreover, if those interactions are considered as additional possible covariates, there is always a positive probability that they will not be selected into the outcome model by WPD's methodology.

Third, the inclusion of an interaction term in WPD's methodology was discussed very briefly, only stating that when an interaction of an exposure by confounder term is included in the model, the resulting estimand comprises a vector parameter, which includes the coefficients of the treatment indicator and the coefficients of the interaction. From this statement, it is unclear whether WPD would intend to estimate quantities such as $\beta(0, 1)$. In any case, they do not provide any methodology to obtain an estimate or a standard error for any such quantity.

Fourth, we added an estimation method based on a model that includes an interaction:

$$E(Y_i(X_i) \mid U_i, \theta) = \beta X_i + ns(U_i, 8) + \beta_{int} X_i ns(U_i, 8), \quad (8)$$

with 8 *df*. Because WPD did not provide a methodology to estimate $\beta(0, 1)$, we chose to use a Bayesian approach that draws samples from the posterior predictive distribution of $\beta(0, 1)$ to obtain a 95% interval. Table 3 shows that model (8) has better coverage rates than model (7), but the former could perform as badly as the latter when σ^2 and B are large. Moreover, the bias and RMSE obtained using method (8) is much larger than using either model (7) or our methodology.

In our simulation, it is possible that the control group and treatment group will have ranges of covariate values that do not overlap. The WPD methodology and the multiple imputation methodology handle such situations in different ways. WPD's approach effectively extrapolates to estimate the average treatment effect for the full range of covariate values observed in either sample, whereas our methodology discards the nonoverlapping observations, and thereby changes the estimand to be the average treatment effect for values of the covariate where there is overlap. The bias, MSE, and coverage of the two procedures are calculated for the estimand that the procedure is trying to estimate. Admittedly, our estimand is generally easier to estimate than WPD's estimand, which we view as generally impossible to estimate without making empirically unassailable assumptions, which if made, should be clearly stated.

Table 3
95% interval coverage rate, bias and RMSE n = 600; model (8)

σ^2	B	0	0.25	0.5	1	2
0.5	Coverage	1	1	1	1	1
	Abs. Bias	0.89	0.53	0.26	0.03	18.48
	RMSE	31.82	27.38	23.06	21.05	1094.7
1	Coverage	1	1	1	1	0.81
	Abs. Bias	24.68	19.91	17.69	20.79	62.12
	RMSE	6.51	6.24	6.31	6.72	8.46
2	Coverage	0.99	1	0.89	0.79	0.63
	Abs. Bias	3.19	3.53	10.20	19.09	2130.34
	RMSE	17.64	15.68	20.54	28.63	2615.60
4	Coverage	0.58	0.51	0.27	0.09	0.01
	Abs. Bias	11.87	32.26	69.43	159.00	19494.30
	RMSE	27.18	41.98	73.99	161.71	22625.18

Absolute Bias and RMSE are in 10^{-3}.

The results in the simulation support the conclusion that methodologies similar to WPD's for the estimation of causal effects can have poor operating characteristics, even in very simple settings with one confounder and a binary treatment. As the number of confounders increases, these methodologies can perform even more poorly because initial bias in any confounder will influence the estimation procedure. In addition, as the number of treatments increases, the probability of overlap between all confounders for all treatments decreases, typically resulting in higher probabilities of larger initial bias across treatment groups.

Methods that do not work for a single covariate with a binary treatment cannot generally work for multiple covariates and multiple treatments. The multiple imputation method presented in this discussion to estimate causal effects is for a single confounder, but this method can be generalized to several covariates using subclassification and matching on propensity scores (Rosenbaum and Rubin, 1983; Rubin and Stuart, 2006; Rubin and Thomas, 1996) with similar conclusions expected.

In summary, when we are interested in reaching conclusions that have major policy implications, it is our view that a causal framework must be used. We believe that typically all covariates that may influence the outcome should be considered when estimating the assignment mechanism, and that only after ensuring that we have similar distributions of the covariates across treatment groups within subclasses, can a statistically valid procedure be reliably implemented.

References

Barnard, J. and Rubin, D. B. (1999). Small-sample degrees of freedom with multiple imputation. *Biometrika* **86,** 948–955.

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics* **24,** 295–313.

Gutman, R. (2011). Topics and applications in missing data and causality. Ph.D. Thesis, Harvard University, Cambridge, MA.

Holland, P. W. (1986). Statistics and causal inference (with discussion). *The Journal of the American Statistical Association* **81,** 945–970.

Neyman, J. (1923). Sur les applications de la thar des probabilities aux experiences agaricales: Essay de principle. english translation of excerpts by Dabrowska, D. and Speed, T. (1990). *Statistical Science* **5,** 465–472.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology* **66,** 688–701.

Bayesian inference for causality: the importance of randomization. In *Proceedings of the Social Statistics Section of the American Statistical Association*, 233–239. Alexandria, VA: American Statistical Association.

Rubin, D. B. (1976). Inference and missing data. *Biometrika* **63,** 581–592.

Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics* **6,** 34–58.

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys*. New York, NY: John Wiley and Sons.

Rubin, D. B. (1990). Formal modes of statistical inference for causal effects. *Journal of Statistical Planning and Inference* **25,** 279–292.

Rubin, D. B. (2006). *Matched Sampling for Causal Effects*. New York: Cambridge University Press.

Rubin, D. B. (2008). Statistical inference for causal effects, with emphasis on applications in epidemiology and medical statistics. In *Handbook of Statistics: Epidemiology and Medical Statistics*, C. R. Rao, J. P. Miller, and D. C. Rao (eds), Chapter 2, 28–63. The Netherlands: Elsevier.

Rubin, D. B. and Stuart, E. A. (2006). Affinely invariant matching methods with discriminant mixtures of proportional ellipsoidally symmetric distributions. *The Annals of Statistics* **34,** 1814–1826.

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

Wang, C., Parmigiani, G., and Dominici, F. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* **68,** 661–671.

Discussions

Stijn Vansteelandt

Department of Applied Mathematics and Computer Sciences,
Ghent University, Krijgslaan 281 S9
9000 Ghent, Belgium
email: stijn.vansteelandt@ugent.be

1. Introduction

The problem of confounder selection is central to the analysis of essentially all observational studies, yet it has received remarkably little attention in the causal inference literature. I congratulate Wang, Parmigiani, and Dominici—herafter referred to as WPD—for addressing this important problem and making a very elegant proposal.

The problem of confounder selection differs from other covariate selection procedures in that confounders are, by definition, simultaneously associated with exposure and outcome. Yet, by factorization of the observed data likelihood into the outcome distribution (conditional on exposure and covariates) and the exposure distribution (conditional on covariates), standard procedures involve building models for these separate distributions one at a time. Methods solely based on outcome regression therefore have a tendency to ignore potentially important confounders. Namely, those covariates that are strongly associated with the exposure, but only moderately with the outcome, risk being dismissed from the analysis model as a result of multicollinearity. This may result in biased outcome regression analyses, which severely understate the actual uncertainty regarding the exposure's effect. In the causal inference literature, this problem has been mitigated through the development of propensity score methods which explicitly incorporate the exposure distribution into the analysis (Rosenbaum and Rubin, 1983). However, with few exceptions (e.g., McCandless, Gustafson, and Austin, 2009), these procedures also separate model building for the propensity score from model building for the outcome. Contrary to outcome regression methods, they thereby prioritize covariates that are strongly associated with the exposure, regardless of their association with the outcome. It has been shown that this can result in inefficient inferences (Hahn, 2004). In addition, any bias in the exposure effect (e.g., because of unmeasured confounding, which is almost always present) gets amplified by the exposure's variance inflation factor, which can be sizable when the model includes strong predictors of the exposure (Pearl, 2010; Vansteelandt, Bekaert, and Claeskens, 2012).

All the above arguments call for simultaneously selecting covariates for inclusion in models for the outcome and exposure distributions. The authors' proposal prevents factorization of the joint outcome and exposure distributions through a prior dependence on the covariate inclusion probabilities. It forms a promising step in this direction.

2. How Does Bayesian Adjustment for Confounding (BAC) Relate to Propensity Score Adjustment?

For $\omega = \infty$, the proposed procedure imposes that covariates appearing in the exposure model will also be included in the outcome model, but allows for the outcome model to include more. This is also characteristic of the frequentist approach of Crainiceanu, Dominici, and Parmigiani (2008). Not surprisingly, a similar performance is thus seen in the reported simulation studies. One important difference is that the use of model averaging may result in more honest descriptions of the overall model uncertainty. A further potential difference comes from the procedure's a priori tendency to exclude covariates from the exposure model, which can be seen because $P(\alpha_m^X = 0) = 2/3$ when $\omega = \infty$. From a subjective Bayesian viewpoint, $P(\alpha_m^X = 0)$ ought to represent one's prior beliefs. Considering that in most practical applications all covariates are—realistically seen—at least somewhat associated with the exposure and that the decision to measure data on a specific covariate in observational studies is often based on prior belief that this covariate is jointly associated with exposure and outcome, lower values of this probability could be desirable. In particular, it seems more prudent to choose a maximum value of $P(\alpha_m^X = 0) = 1/2$, which expresses a priori ignorance. For this choice, one would obtain

$$\frac{P\left(\alpha_m^X = 1 | \alpha_m^Y = 0\right)}{P\left(\alpha_m^X = 0 | \alpha_m^Y = 0\right)} = \frac{2}{\omega + 1}, \frac{P\left(\alpha_m^X = 1 | \alpha_m^Y = 1\right)}{P\left(\alpha_m^X = 0 | \alpha_m^Y = 1\right)} = \frac{2\omega}{\omega + 1}.$$

It would be interesting to see how this (or related) modification(s) might affect the performance of the proposed strategy, especially in settings where various covariates are strongly predictive of the exposure and only moderately predictive of the outcome.

The approach of Crainiceanu et al. (2008), and thereby also BAC at $\omega = \infty$, has close connections with regression adjustment for the propensity score. This is so, even though the propensity score is usually defined for dichotomous exposures, because the propensity score can be redefined to be the conditional expectation of the exposure, given confounders, when the interest lies in linear regression models for the outcome (see e.g., Robins, Mark, and Newey, 1992). A common strategy is then to build a propensity score model using standard model building steps, and to subsequently regress the outcome on the exposure, propensity score, and those covariates that are most strongly associated with the outcome. Like BAC at $\omega = \infty$, this approach tends to adjust for all important predictors of the exposure (by the propensity score) to control for confounding, and in addition adjusts for strong predictors of the outcome to improve efficiency. It could nonetheless be preferable to BAC. First, because all predictors of the exposure are combined into a univariate propensity score, the outcome regression model will likely become more parsimonious. Second, unlike BAC, it entails a doubly robust procedure (Robins et al., 1992), which yields consistent additive exposure effect estimators when either the propensity score model or the outcome regression model is correctly specified. Alternatively, one may regress the outcome on the exposure and a model-averaged propensity score, as obtained by standard frequentist or Bayesian model averaging procedures. It will be worthwhile to formally compare BAC with these and other alternatives.

To gain preliminary insight, I have repeated simulation study 2 of WPD with $n = 100$ to mimic a setting where the information is scarce. The suggested propensity score procedure was implemented in two ways: once by stepwise regression (using the BIC) and once by a model-averaged propensity score using the weights exp(AIC/2) (Buckland et al., 1997), where the AIC relates to the exposure model only. This resulted in an exposure effect estimator with bias 0.029 and 0.007, empirical standard deviation 0.216 and 0.218, and MSE 0.047 and 0.048, respectively. This improves upon fitting a main effects model with all covariates, but is inferior to BAC, which respectively yield a bias of 0.014 and 0.059, empirical standard deviation 0.255 and 0.162, and MSE 0.065 and 0.032. These differences appear at odds with the close relatedness of the different proposals. They call for a more in depth exploration to understand whether the superior performance of BAC is consistent (versus specific to this simulation) and whether it is related to its preferential tendency to exclude covariates from the exposure model, which could be a concern under other data-generating mechanisms with strong correlation between confounders and exposure.

3. Limitations of BAC

The proposals of WPD seem generic, but are essentially limited to linear models. This is because adjustment for a covariate in a nonlinear model typically changes the meaning and magnitude of the exposure effect systematically. Contrary to what the authors claim, this is so even when that covariate is not associated with the exposure and thus not a confounder, as a result of so-called noncollapsibility of many association measures. Adjustment for such covariates in nonlinear models is also typically disadvantageous as it may reduce precision (Robinson and Jewell, 1991). Vansteelandt et al. (2012) overcome these concerns by focussing confounder selection on the quality of population-averaged exposure effects. For instance, in logistic regression models for the outcome,

$$E(Y_i | X_i, U_i) = \text{expit}\left(\beta^{\alpha^Y} X_i + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\alpha^Y} U_{im}\right),$$

they note that

$$E\{Y(x)\} = E\left\{\text{expit}\left(\beta^{\alpha^Y} x + \sum_{m=1}^{M} \alpha_m^Y \delta_m^{\alpha^Y} U_{im}\right)\right\},$$

is the expected outcome that would be observed if all members of the population had exposure level $X = x$, provided that U_i is a set of covariates sufficient to control for confounding. The population-averaged exposure effect can then be defined as a contrast of $E\{Y(x)\}$ for different values of x, e.g., $E\{Y(x+1)\} - E\{Y(x)\}$ or $E\{Y(x+1)\}/E\{Y(x)\}$. Alternatively, it can refer to a contrast with the observed outcome, e.g., $E(Y)/E\{Y(0)\}$, or express the effect of changes in the observed exposure distribution, e.g., $E\{Y(1.1X)\}/E\{Y(X)\} = E\{Y(1.1X)\}/E(Y)$. Extending BAC/TBAC to these effect estimands will be essential.

A possibility would be to adapt the proposed procedure so that it provides model-averaged predictions of exposure and outcome, which may subsequently be used as input for more general (possibly frequentist) causal effect estimators. For instance, for discrete X, a doubly robust estimator of

$E\{Y(x)\}$ is

$$\frac{1}{n}\sum_{i=1}^{n}\frac{I(X_i=x)}{P(X_i|U_i)}\{Y_i - E(Y_i|X_i=x,U_i)\} + E(Y_i|X_i=x,U_i).$$

The proposed procedure could be used to obtain model-averaged estimators of $P(X_i|U_i)$ and $E(Y_i|X_i=x,U_i)$ in a way that primarily targets the inclusion of all confounders. These arguments more generally suggest that the authors' focus on a single regression coefficient in the outcome regression model may be limiting.

4. What is BAC Targeting?

BAC intuitively makes sense, but its heuristic nature leaves some vagueness as to whether it enjoys any specific optimality properties. Vansteelandt et al. (2012) develop model selection strategies which target minimal mean squared error of the exposure effect estimator. This can be a desirable goal, given practitioners' disproportionate focus on point estimates. Low MSE may also indirectly be targeted by regressing the outcome on the exposure and a model-averaged propensity score, with model-averaging weights given by the reciprocal of the variance of the exposure effect estimator in a propensity score adjusted analysis. For illustration, I evaluated this in the simulation study of the authors' Web Table 1 (where the presence of covariates that are solely associated with the exposure is particularly challenging for propensity score analyses). This resulted in an exposure effect estimator with bias 0.007, empirical standard deviation 0.040 and MSE 0.0016. This greatly improves upon both BAC and fitting a model which includes all main effects, which gave bias 0.001 and 0.007, empirical standard deviation 0.072 and 0.078, and MSE 0.006 and 0.006, respectively.

A more prudent strategy may be to seek an honest reflection of the overall uncertainty, and thus to target confidence validity. This is most easily obtained by avoiding model selection to the extent possible, by working under large models. Budtz-Jörgensen et al. (2005) find that this may even lead to tighter confidence intervals than those obtained after model selection when model uncertainty is taken into account. For instance, fitting a full model (i.e., including all covariates) considered in simulations 1 and 2 of WPD, resulted in an exposure effect estimator with bias 0.0005 and 0.002, empirical standard deviation 0.046 and 0.051, and MSE 0.002 and 0.003, respectively; coverages of 95% confidence intervals were 94.9% and 94.7%. These results are almost identical to those obtained by fitting the true model (see Tables 3 and 4, respectively), suggesting that there was actually no need for model selection in those simulation studies. To fully appreciate the performance of BAC, further simulation studies will be needed in settings where model reduction is essential.

Vansteelandt et al. (2012) suggest alternative ways to guarantee confidence validity. Under certain assumptions, they demonstrate that a conservative asymptotic variance of the exposure effect estimator can be obtained, even when imprecision because of estimation and model selection on the propensity score is ignored, provided that the propensity score is efficiently estimated. Application of (such) procedures which exclusively adopt model building on the propensity score, thus guarantees (approximate) confidence validity, provided of course that the candidate list of models includes the true propensity score model.

It seems that BAC might come closer to a procedure that guarantees confidence validity for large ω (by forcing predictors of the exposure in the outcome regression model), although its potential tendency to preferentially exclude covariates from the exposure model leaves doubts as to whether this is generally true. For small ω, BAC might instead come closer to a procedure that favors minimal MSE, although it remains unclear to what extent this is so as BAC does not directly target minimal MSE. Further investigation will be important to understand the properties of BAC in function of ω. This will help practitioners to make an informed choice of ω, and may reduce the risk of selecting ω in a data-driven way.

5. Conclusions and Possible Extensions

In summary, the proposed procedure of confounder selection is attractive for its focus on confounders and for acknowledging model uncertainty. In its current form, the procedure is essentially limited to the analysis of linear models because of noncollapsibility of many association measures, but the approach could be extended to more general models along the lines described in Section 3.

The general principle of preventing factorization of the joint outcome and exposure distributions is attractive for confounder selection. It has also been seen in other contexts. For instance, when the outcome regression model is defined to be one that minimally includes the exposure and propensity score, then the inclusion of the propensity score necessarily prevents factorization (McCandless et al. 2009). Tan (2006) prevents factorization by ignoring part of the information on the joint distribution of outcome and covariates (at fixed exposure levels). It will be of interest to understand better how these different alternatives relate to BAC. For instance, what if standard Bayesian model averaging were applied on the basis of the joint outcome and exposure distribution in settings where the linear outcome regression model includes the propensity score?

The general principle of preventing factorization of the joint outcome and exposure distributions may also be more broadly applicable in other contexts. For instance, Greenland (2008) rightly argues that in practical applications, all available covariates will be somewhat associated with outcome and exposure. He therefore argues in favor of shrinkage estimators under a "large" model for the outcome. A limitation of standard shrinkage procedures is that they do not target the exposure effect of interest. This may potentially be remedied by focussing on exposure and outcome models of the form

$$\begin{aligned}
E(X_i|U_i) &= \sum_{m=1}^{M}\delta_m^X U_{im} \\
E(Y_i|X_i,U_i) &= \beta X + \sum_{m=1}^{M}\delta_m^Y U_{im} \\
\delta_m^X &\sim N\left(0,\sigma_X^2\right) \quad \text{for } m=1,\ldots,M \\
\delta_m^Y\left|\delta_m^X\right. &\sim N\left(0,\sigma_Y^2\left(1+\delta_m^{X\,2}/\sigma_X^2\right)\right) \quad \text{for } m=1,\ldots,M,
\end{aligned}$$

where priors are used to penalize large regression coefficients. Here, the choice of priors ensures less penalization of outcome

regression coefficients that correspond to strong predictors of the exposure. The resulting ridge regression seems close in spirit to BAC with a focus on confounding adjustment, but might be computationally advantageous relative to BAC and have the further advantage of not selecting confounders out of the analysis.

In conclusion, the principle underlying BAC holds much promise. However, as shown, many related—perhaps less computationally challenging—procedures can be conceived, which do not require novel model averaging strategies. Much obscurity remains as to how BAC relates to these and other more common approaches based on propensity scores and as to how to best choose ω in practice. Before widespread use of BAC can be recommended, closer scrutiny of its theoretical properties will be essential, as well as evaluation in realistic simulation studies where dimension reduction is critical (that is where, unlike in the considered simulation studies, fitting a full model which includes all covariates yields poorly performing estimators).

References

Budtz-Jorgensen, E., Keiding, N., Grandjean, P., and Weihe, P. (2007). Confounder selection in environmental epidemiology: Assessment of health effects of prenatal mercury exposure. *Annals of Epidemiology* **17,** 27–35.

Crainiceanu, C.M., Dominici, F., and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95,** 635–651.

Greenland, S. (2008). Variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology* **167,** 523–529.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *Review of Economics Statitics* **86,** 73–76.

McCandless, L.C., Gustafson, P., and Austin, P.C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine* **28,** 94–112.

Pearl, J. (2010). On a class of bias-amplifying covariates that endanger effect estimates. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, Grunwald P and Spirtes P (eds), 417–424 , AOAI, Corvallis, OR.

Robins, J.M., Mark, S.D., and Newey, W.K. (1992). Estimating exposure effects by modelling the expectation of exposure conditional on confounders. *Biometrics* **48,** 479–495.

Robinson, L.D. and Jewell, N.P. (1991). Some surprising results about covariate adjustment in logistic regression models. *International Statistical Review* **59,** 227–240.

Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Tan, Z. (2006). A distributional approach for causal inference using propensity scores. *Journal of the American Statistical Association* **101,** 1619–1637.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research* **21,** 7–30.

Discussion of Adjustment Uncertainty and Propensity Scores

Lawrence C. McCandless

Faculty of Health Sciences, Simon Fraser University, Burnaby BC V5A 1S6, Canada
email: lmccandl@sfu.ca

I am delighted to comment on this provocative and innovative paper. My comments will address the intersection between adjustment uncertainty and propensity score (PS) techniques. Adjustment uncertainty, as defined by Wang, Parmigiani and Dominici (WPD, 2011), refers to the uncertainty about which covariates should be accounted for to adjust properly for confounding. PS techniques are a class of methods for controlling confounding in observational studies. In a PS analysis, we combine information from a vector of covariates for each study participant into a single summary score, which is the probability of exposure given the covariates. The score is then used to adjust for confounding. We can stratify on the PS, use inverse probability weighting and matching, or we can include it as a covariate in a regression model for the outcome. To estimate the PS, we need a model for the exposure. Ideally, the investigator has identified the confounders beforehand and can elicit a realistic model for probability of exposure. But, in practice, researchers are often faced with a large array of covariates, and this demands a data-driven approach to model selection.

Variable selection for PS models is an exciting area of innovation in statistics. Much recent work has been championed by Schneeweiss et al. (2009) in pharmacoepidemiology. Schneeweiss argues that when choosing variables to include in the model for exposure, we must take into consideration the relationship with the outcome. Including variables unrelated to the outcome in the PS increases the variance of the exposure-effect estimate, with no commensurate reduction in bias. Essentially, the message is that we should only include genuine confounders in the PS.

Schneeweiss et al. (2009) describes novel procedures for variable selection. One thing that struck me when reading the Schneeweiss paper was that the methods are similar to those described by Crainiceanu, Dominici, and Parmigiani (2008).

Table 1

A comparison of the performance of four different exposure models that are used to calculate the estimated PS. Models 1, 2, and 3 are missing important confounders. They give biased estimates of the true exposure effect, which is equal to zero

Model	Exposure effect estimate	Prediction error for X	Prediction error for Y	Fitted model for X
1	2.18	0.69	1.10	$\text{logit}[\hat{P}(X=1)] = 0.0$
2	0.86	0.62	0.68	$\text{logit}[\hat{P}(X=1)] = 0.0 + 0.84C_1$
3	1.69	0.62	1.04	$\text{logit}[\hat{P}(X=1)] = 0.0 + 0.84C_2$
4	0.00	0.54	0.64	$\text{logit}[\hat{P}(X=1)] = 0.0 + 1.0C_1 + 1.0C_2$

For example, compare table 3 from Schneeweiss et al. (2009) with figure 1 of Crainiceanu et al. (2008). Both papers attempt to fit joint models for the exposure and outcome to identify confounders. Thus, it appears that an immediate application of adjustment uncertainty is that it provides a principled framework for variable selection in PS models.

How to combine adjustment uncertainty with PS techniques? A simple strategy would be to modify the procedure of Crainiceanu et al. (2008). At Steps 3 and 8, we would first calculate the estimated PS based on each subset of $U = (U_1, \ldots, U_M)$ in the dominant model class. Next, we would plot point and interval estimates for the exposure effect calculated by adjusting for the estimated PS. In other words, we swap out the U's in place of a linear predictor calculated from the PS.

A Bayesian approach has additional advantages. Rather than adjusting for the estimated PS based on a single exposure model, we could average over different models. This incorporates uncertainty in the PS that arises from uncertainty in the PS model. I suspect that such a method would be superior to conventional techniques in the face of model misspecification. In addition, a Bayesian approach would permit the incorporation of standard Bayesian machinery, such as prior information, complex modeling, and MCMC computation.

I was intrigued by the discussion of feedback from the outcome in BAC and TBAC. Feedback describes the situation in which the outcome variable "interferes" with estimation of the exposure model. One consequence of combining the PS with Bayesian model averaging is that feedback could potentially inform the choice of exposure model, and therefore, the estimated PSs. To see how this would work, consider a simple illustration involving a Gaussian response variable Y, a dichotomous 0/1 exposure X, and two Gaussian covariates C_1 and C_2. Suppose that the true data generating mechanism is,

$$\begin{aligned} C_1, C_2 &\sim N(0,1) \\ X|C_1, C_2 &\sim \text{Bernoulli}\left(\frac{\exp(C_1+C_2)}{1+\exp(C_1+C_2)}\right) \\ Y|X, C_1, C_2 &\sim N(2C_1 + C_2, 1). \end{aligned}$$

In this scenario, the true effect of X on Y given (C_1, C_2) is zero. The variables C_1 and C_2 are both confounders, but C_1 is more strongly associated with Y.

Let's suppose that the investigator plans to adjust for confounding from (C_1, C_2) by estimating the PS and then, including it as a covariate in a regression model for the outcome. In other words, the plan is to first estimate the PS for each study participant, denoted $\hat{Z}$, where $\hat{Z}$ is the fitted value from logistic regression with X as the dependent variable and either C_1 or C_2 as independent variables. Next, the investigator will estimate the effect of X on Y while controlling confounding by using the model,

$$Y = \alpha + \beta X + \xi \hat{Z} + \epsilon, \quad \text{where } \epsilon \sim N(0, \sigma^2), \qquad (1)$$

to estimate the unknown parameters $(\alpha, \beta, \xi, \sigma^2)$. Note that using equation (1) for exposure effect estimation differs from WPDs approach. WPD build separate models for the exposure and outcome as a direct function of the original covariates.

Suppose that the true data generating mechanism is unknown to the investigator and a debate centers upon which variables to include in the model for the exposure. We limit the discussion to choosing among four candidate models that include either C_1 or C_2 or both (C_1, C_2) or neither as covariates in the logistic regression model with X as the dependent variable. Table 1 shows the four candidate exposure models, which have been fitted to a very large synthetic dataset that is sampled from the true data generating mechanism. Thus, we can assume that random error is negligible. The estimated exposure effects are given for each model (true value is zero). Obviously Model 4 gives the right answer, whereas Models 1, 2, and 3 are missing important confounders.

In addition, Table 1 shows the estimated average prediction error of the fitted regression models for X and for Y. The average prediction error for X is defined as,

$$-E[X \log \hat{Z} + \{1 - X\} \log\{1 - \hat{Z}\}],$$

where the expectation is over the joint distribution of X and $\hat{Z}$ (see section 5 of McCandless, Gustafson, and Austin (2009) for discussion on how this is estimated. Essentially, we train the model and then estimate the average prediction error in the same large dataset). The prediction error for Y is,

$$E\left(\frac{1}{2\hat{\sigma}^2}[Y - \{\hat{\alpha} + \hat{\beta}X + \hat{\xi}\hat{Z}\}]^2\right).$$

Smaller prediction errors tell us that the model gives a better overall fit for the data.

If we look at prediction error for X, then Models 2 and 3 do equally well, and both give errors equal to 0.62. This is expected because C_1 and C_2 have the exact same association with the probability of exposure. From a conventional PS modeling perspective, the investigator would be

ambivalent in choosing C_1 over C_2 for inclusion into the exposure model. However, the interesting point is that adjusting for $\hat{Z}$ calculated from C_1 does much better in predicting the outcome variable. The prediction error for Y for model 2 is 0.68, which is smaller than 1.04 for model 3. The message is that incorporating outcome risk factors in the PS improves the fit of the outcome model.

If we incorporate feedback from the outcome, then a Bayesian analysis will prefer PS estimates calculated from model 2 rather model 3 because it seeks to optimize the fit of X and Y simultaneously. C_1 is the preferred variable because it is a more powerful confounder. However, the benefits of feedback are not obvious. One the one hand, joint fitting of the exposure and outcome model is appealing because it makes fuller use of the data. But, on the other hand, it need not improve estimation of the exposure effect. McCandless et al. (2009) investigated Bayesian regression adjustment for the PS. They showed that incorporating feedback from the outcome can increase the mean squared error of the exposure effect, even if all models are correctly specified. Furthermore, if equation (1) is misspecified, then feedback can introduce bias in the estimated PS because of contamination between models.

Interestingly, using Y to estimate the PS flies in the face of convention. Rubin (1997) writes "In this prediction of treatment group measurement, it is critically important that the outcome variable (e.g., death) play no role; the prediction of treatment group must involve only the covariates." Later, he acknowledges "A final possible limitation of PS methods is that a covariate related to treatment assignment but not to outcome is handled the same as a covariate with the same relation to treatment assignment but strongly related to outcome." Thus, I view the interplay between exposure and response models as an exciting opportunity for innovation. Note that the magnitude of the feedback will depend strongly on the manner in which $\hat{Z}$ enters into equation (1). In my example, the variable Y is assumed to depend linearly on the PS, but in real-life examples it might make sense to incorporate the PS nonparametrically into the model (Lunceford and Davidian, 2004).

References

Crainiceanu, C. M., Dominici, F., and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika* **95,** 635–651.

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.

McCandless, L. C., Gustafson, P., and Austin, P. C. (2009). Bayesian propensity score analysis for observational data. *Statistics in Medicine* **28,** 94–112.

Rubin, D. B. (1997). Estimating causal effects from large data sets using propensity scores. *Annals of Internal Medicine* **127,** 757–763.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Mogun, H., Avorn, J., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20,** 512–552.

Wang, C., Parmigiani, G., Dominici, F. (2011). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics* 1–25.

Rejoinder: Bayesian Effect Estimation Accounting for Adjustment Uncertainty

Chi Wang,[1,2,*] Giovanni Parmigiani,[3,4] and Francesca Dominici[4]

[1]Markey Cancer Center, University of Kentucky, Lexington, Kentucky 40536, U.S.A.
[2]Department of Biostatistics, College of Public Health, University of Kentucky, Lexington, Kentucky 40536, U.S.A.
[3]Department of Biostatistics and Computational Biology, Dana-Farber Cancer Institute, Boston, Massachusetts 02115, U.S.A.
[4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.
[*]*email:* chi.wang@uky.edu

We are grateful to the Editors for the opportunity to publish our article with discussion and to the discussants for their insightful and stimulating comments.

We proposed an approach to account for uncertainty in variable selection when the goal is estimation of a regression coefficient within a linear model. Despite the variety of more advanced methodologies available for describing relationships between variables, linear regression is still a very widely applied tool across science. The vast majority of linear regressions are run for the explicit purpose of estimating coefficients, often a single coefficient; the near totality of these use inference methods that ignore the uncertainty in the selection of the other variables entering the regression equation. So, although admittedly this is a relatively simple setting, we hope to have addressed a broadly applicable need.

The discussants raise many issues. Most touch on one of two themes. First, to what extent can this methodology be extended successfully beyond linear models? Second, what is its

Table 1

Comparison of estimates of β from BAC, TBAC, and the full model. Data were generated as in the two simulation scenarios in the article, but with sample size 100. *BIAS is the difference between the mean of estimates of β and the true value, SEE is the mean standard error of the estimates, SSE is the sample standard error of the estimates of β across simulations, MSE is the mean squared error, and CP is the coverage probability of the* 95% *confidence interval or credible interval*

Simulation scenario	Method	BIAS	SEE	SSE	MSE	CP
One	BAC	−0.006	0.142	0.152	0.023	0.93
	TBAC	−0.005	0.149	0.155	0.024	0.94
	Full model	−0.009	0.207	0.207	0.043	0.94
Two	BAC	0.059	0.162	0.170	0.032	0.92
	TBAC	0.041	0.175	0.178	0.033	0.94
	Full model	−0.005	0.253	0.257	0.066	0.95

relation to approaches for effect estimation developed in the causal inference literature? These interests are very encouraging, as they suggest that our ideas can provide the foundation for yet more broadly applicable tools. With regard to causal inference, one of our take-home messages from the discussions is that this field would benefit from a more explicit investigation of the implications of model uncertainty. We look forward to future developments in this area.

Because of the scope of the points raised, fully addressing them would require years of work. We do look forward to that work and we started chipping a little bit at it, with additional simulations we present here. Our responses are organized by topic, as there is some overlap in the issues raised.

1. When it is Useful to Account for Model Uncertainty

Vansteelandt points out that our simulations 1 and 2 in Section 4 do not fully illustrate the need for accounting for model uncertainty, because the sample size is large ($n = 1000$) compared to the number of potential confounders ($M < 60$). We agree; model uncertainty becomes more pressing with small to moderate sample sizes, or with a larger number of potential confounders. To illustrate, Table 1 presents additional simulation results supplementing those originally reported in Web Appendix E. The settings are the same as in simulations 1 and 2, except that the sample size is now 100, and we have added results for the full model. Compared to BAC and TBAC, the full model yields larger standard errors and MSEs for the estimate of β.

2. Optimality of BAC and Choice of ω

We also agree with Vansteelandt's comment that further study of the optimality properties of BAC would be important. In general, a Bayesian estimator based on posterior mean minimizes the expected MSE, where the expectation is computed with respect to the prior. Thus, as a general rule of thumb, BAC will perform well on average over possible parameter values, and will perform well for specific parameter choices that are well supported by the prior. It will do less well in scenarios which are considered less likely by the prior. BAC incorporates exposure model information in the specification of the outcome model prior. This is more effective for exposure effect estimation than the "flat" prior implicit in BMA (see Section 3 of the article).

One of the challenges of applying BAC is the choice of ω. Vansteelandt makes interesting observations about the potential connection between the choice of ω and frequentist optimality criteria, including confidence validity and MSE. This issue needs to be investigated further. Generally speaking, a large ω may favor inclusion of confounders in the outcome model and, thus, reduce bias; however, relations to other criteria are less clear. In the situations of simulations 1 and 2, $\omega = \infty$ yields the smallest MSE. In addition, $\omega = 1$ corresponds to the BMA method, which does not utilize the exposure model information and has the largest MSE in most of our simulations. We conjecture that the relation between MSE and value of, ω, may not always be monotone and may be data dependent.

3. BAC for Causal Inference

In causal inference, the estimand is often the average causal effect (ACE) of exposure (or treatment) defined as $\Delta(d_1, d_2) = E\{Y(d_1)\} - E\{Y(d_2)\}$, where $Y(d_1)$ and $Y(d_2)$ are the outcomes, possibly unobserved, that occur if the exposure levels were d_1 and d_2, respectively. Schafer and Kang (2008) studied the connections between $\Delta(d_1, d_2)$ and the exposure coefficient, β, from the true regression model that characterizes the relationships among potential outcomes, exposure, and confounders. If the regression model is linear and there are no interactions between X and the *true* confounders $\boldsymbol{U}^*$, then $\Delta(d_1, d_2) = \beta(d_1 - d_2)$. When there are interactions between X and $\boldsymbol{U}^*$, this equality still holds if the confounders are included after subtracting their population means. In these cases, BAC is directly useful for estimating the ACE.

To understand limitations and potential extensions, it is useful to review the assumptions for this equality. The first is the *stable unit treatment value assumption* (SUTVA; Rubin, 1980), which states that the potential outcomes for one unit are unaffected by the exposure assignments of other units. The second is *strong ignorability* (Rosenbaum and Rubin, 1983) which assumes that $P(X \mid Y(0), \ldots, Y(D), \boldsymbol{U}^*) = P(X \mid \boldsymbol{U}^*)$. This in turn guarantees that $P(Y(d) \mid \boldsymbol{U}^*) = P(Y \mid X = d, \boldsymbol{U}^*)$, where Y is the observed outcome. If these hold,

$$\Delta(d_1, d_2) = E_{U^*}[E\{Y|X = d_1, \boldsymbol{U}^*\} - E\{Y|X = d_2, \boldsymbol{U}^*\}], \quad (1)$$

where E_{U^*} is the expectation with respect to U^*.

If the outcome Y is discrete and we are using a generalized linear model (GLM), then $\beta(d_1 - d_2)$ is different from $\Delta(d_1, d_2)$. Vansteelandt points out that in this case $\Delta(d_1, d_2)$ can still be obtained by contrasting population-averaged exposure effects as in (1). See also Lunceford and Davidian (2004) for the definition of $\Delta(d_1, d_2)$ in the context of a logistic regression. Therefore, if the adjustment for confounding is done by using a regression model, the extension of BAC to a GLM is straightforward, because posterior distributions of population average exposure effects can be obtained as a by-product of the posterior samples of all the unknown parameters, complemented, if required by the design, by modeling of the marginal distribution of $\boldsymbol{U}^*$.

Confounder selection is central to the analysis of essentially all observational studies, yet the problem of uncertainty about this selection has received remarkably little attention in the causal inference literature. Although many papers provide some general advice on this topic (Brookhart et al., 2006; Stuart, 2010), a general statistical approach that selects the variables that must be included, and whose objective is effect estimation rather than prediction, is lacking. Based on jointly modeling the exposure and outcome models, BAC addresses model selection and the associated uncertainty in both the exposure and outcome models and weights more heavily the models that include all the necessary confounders in an automated and data-driven way. Because the posterior distribution is obtained by averaging across different exposure and outcome models, it will provide estimates of ACE that account for model uncertainty in both. Importantly, we only need to assume that $\boldsymbol{U}$, the *full* set of the measured confounders, contains the *true* set $\boldsymbol{U}^*$. We do not need to know $\boldsymbol{U}^*$ exactly.

4. Propensity Scores

BAC could also be further developed to account for model uncertainty in causal inference when propensity score-based methods are used to adjust for confounding (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004; Schafer and Kang, 2008; Stuart, 2010). Propensity score-based methods are usually developed for binary exposures, although generalization to arbitrary type of exposure is possible (Imai and van Dyk, 2004; Hirano and Imbens, 2005). Connections between BAC and methods for causal inferences with continuous exposure that use generalized propensity score methods need to be explored further.

Although the literature on propensity scores is vast, existing approaches do not account for the uncertainty about which confounders should be included in the exposure or the outcome models. Generally, a propensity score model is specified with *a priori* knowledge about which covariates should be included. Then the estimated propensity score can be included as a covariate in the outcome model or used to create strata with similar scores, in which case the ACE is estimated by a weighted average of group-specific ACE estimates weighted by the proportion of subjects in that stratum (Rosenbaum and Rubin, 1983, 1984; Lunceford and Davidian, 2004). Although stratification is expected to balance the covariates, it is often suggested to fit, within each stratum, an outcome regression model including some or all the potential confounders, to reduce residual within-stratum confounding (Lunceford and Davidian, 2004; Stuart, 2010). An alternative approach is to construct inverse-propensity weighted estimators, including doubly robust estimators (Scharfstein et al., 1999; Robins, Rotnitzky, and Zhao, 1994; Tan, 2010).

Confounder selection could be addressed by standard variable selection techniques applied to both the propensity and outcome models. To account for the uncertainty in selection, one could conduct two separate BMA approaches. However, we presented evidence that suggests that there are important differences between *adjustment uncertainty* and *model uncertainty*. In adjustment uncertainty, the goal is to estimate the effect of an exposure X on the outcome Y accounting for the uncertainty about which confounders $\boldsymbol{U}$ need to be included into the model. In model uncertainty, all predictors $(X, \boldsymbol{U})$ are equally important, and their inclusion in the regression model is evaluated based on measures of performance in predicting Y. We, thus, expect that standard approaches for variable selection and for accounting for model uncertainty may not be ideal in this context of propensity score analysis. More specifically, variable selection based on the propensity score model only prioritizes covariates that are strongly associated with X, whereas variable selection based on the outcome model only prioritizes the covariates strongly associated with Y. Both these approaches can result in inefficient and biased inferences because they will likely fail to identify the set of true confounders $\boldsymbol{U}^*$ (Brookhart et al., 2006; Hahn, 2004; Imbens, 2004; Schneeweiss et al., 2009). With BAC, we proposed an approach for conducting a joint BMA for the exposure and the outcome models in the context of a linear regression. Extending this approach to propensity score methods is possible and promising.

In propensity score methods, "standard errors for the treatment effect estimate are usually calculated without acknowledging uncertainty in the estimated propensity scores" (McCandless et al., 2009). BAC jointly models the exposure and outcome models and utilizes model averaging to summarize information across different models. Thus, it provides a Bayesian framework to fully account for uncertainty in variable selection, and can potentially give more accurate estimates of the standard errors. McCandless raises the interesting question of whether feedback from the outcome model should be used for estimation of the exposure model. In the propensity score literature, it is debated whether Y should be used to estimate the propensity score (McCandless et al., 2009). In our article, we discuss two approaches; TBAC allows no feedback. BAC allows for a specific type of feedback: Y can inform which confounder should be included in the exposure model; but conditional on this inclusion, Y does not inform the estimation of the regression coefficients in the exposure model. This is because the exposure model parameters are independent from the outcome model parameters conditional on the α's. Web Appendix D compared BAC to TBAC in simulations. Although there is some difference in covariates' inclusion probabilities based on BAC versus TBAC in the exposure model, there is no major difference in the inclusion probabilities in the outcome model. However, these results could be specific to the simulation scenarios used, and more comprehensive investigation will be required to fully understand the feedback effect.

5. An Alternative Prior

Vansteelandt proposes an alternative prior, which assigns to each potential confounder an equal prior probability of being included or excluded in the exposure model. We have implemented BAC using this prior and applied it to simulation 2, with sample size 100. The results are very close to those obtained from BAC using our original prior: the bias is 0.061, the standard error of the estimates is 0.169 and the MSE is 0.032. This suggests that the difference in results between the suggested propensity score procedure with variable selection (Vansteelandt, Section 2) and BAC may not be due to the prior specifications. Vansteelandt performs model selection on

the propensity score model by selecting the confounders for inclusion in the propensity score model based on their ability to predict X. This assigns lower weights to propensity score models that include covariates that are associated with Y but not associated with X, for example, U_8–U_{14}. These variables, although they are not confounders, are predictors of Y. Including them in the model may reduce the standard error of the estimate (Brookhart et al., 2006).

6. Nonlinearities, Interactions, and Sparse Confounders

Gutman and Rubin compare methods in an example that includes obvious nonlinearities in the effect of X on Y, pronounced interactions involving X and U, and limited overlap between $U \mid X = 0$ and $U \mid X = 1$. Within linear models with no interactions, there is a correspondence between ACE estimation and regression analysis. If the model assumptions fail, then (a) this correspondence will no longer hold, and (b) analysis based on linear models will give poor results whatever the goal. We are aware of this. It is highly unlikely that any of us would have used linear models in the analysis of data with these features just described, which would have been revealed by the most basic model diagnostics. Also, Gutman and Rubin label as WPD a linear model which does not jointly model exposure and outcome, and includes no consideration of model uncertainty—the two main points of our work. This labeling seriously misrepresents our proposal. More meaningful comparisons would require a bona fide extension of BAC to nonlinearity and interactions, and a decision about whether coefficients or ACE are of interest. We think BAC could be generalized in both directions but the ultimate method would differ depending on whether ACE or coefficients are investigated.

BAC was motivated by the estimation of the health effects of air pollution in the context of time series studies. Here, because the treatment assignment, the outcome, and the confounders are all serially correlated, the key assumptions for estimating causal effects are either violated or more complicated to formalize. For example, it is not necessarily clear how to define "ignorable treatment assignment" when the "treatment" is spread across time and can change with previous exposures or outcomes. Steps in the directions of meeting these challenges exist, for example in the context of causal inference for dynamic treatment regimens. See Zhang, Joffe, and Small (2011) and references therein. However, estimating ACE in time series studies would require the development of a whole new causal framework.

Using an assumption of independence between units, Gutman and Rubin propose an estimation approach (GR) that consists of three steps; (1) preprocessing; (2) matching; and (3) semi-parametric estimation of ACE within each stratum obtained by fitting two separate regression models for $Y(1) \mid U$ and $Y(0) \mid U$. This method is well justified in the context of data such as those in their comment, but it is tailored for a class of problems that is completely different from that for which BAC is designed. This point is well illustrated by Gutman and Rubin's simulation scenario, which involves the combination of the following features: (1) there is only one confounder U; (2) the empirical distributions of U in the treated and the control groups have limited overlap; (3) the treatment effect is a highly nonlinear function of U; and (4) the outcome is generated from a deterministic function of X and U without any noise. Figure 1(a) depicts potential outcomes versus U in a data set simulated from their scenario. These are superimposed to histograms of U by group, highlighting the limited overlap of the two empirical distributions. It would be interesting to know what type of real application motivated this construct.

Their setting creates an ill posed problem for any regression model, irrespective of the issue of selection of confounders. In presence of highly nonoverlapping distributions of the confounders between the two groups, a misspecified regression will fail to extrapolate reliably. The importance of overlap for estimating ACE has been widely accepted and extensively studied (King and Zeng, 2005; Schafer and Kang, 2008; Crump et al., 2009; Stuart, 2010). One solution is to preprocess the data and fit the model only for the units whose U's are in the overlapping region of the empirical distributions of $U|X = 0$ and $U|X = 1$.

Gutman and Rubin compare GR to the regression model of their equation (8); they label it WPD, we refer to it as *Model* (8). When they evaluate the MSE, coverage and bias, the estimand for *Model* (8) is the full-population truth, that is the average treatment effect across the entire range of U's. In contrast, the estimand for GR is different: their estimand is a sub-population truth, where the averaging is done only on the region of overlap. Because the exposure effect varies with U, the two target estimands can differ, and they do by more than twofold in some samples, as shown in Figure 1(b). Estimation of the full-population truth is much harder, and not an entirely meaningful endeavor given the extrapolation needed.

To provide a more objective comparison, we separately evaluate the full-population and sub-population estimation problems. We generate data as in Gutman and Rubin for $B = 2$, $\sigma^2 = 0.5, 1, 2, 4$. We consider both $n = 600$, as they do, and $n = 60$. We compare three methods: (1) the *True Model* as defined in equation (6) of Gutman and Rubin; (2) *Model* (8), which allows for a nonlinear interaction between X and U, fit using either the same normal-gamma prior used in BAC, or a noninformative prior $\pi(\theta, \sigma_Y^2) \propto \sigma_Y^{-2}$, where θ is a vector of model coefficients and σ_Y^2 is the variance of error term; and (3) a simple extension of BAC. The extension of BAC is defined as follows. We take *Model* (8) as the full model. We then include random indicators for inclusion of both main effect and interaction terms, and jointly considering exposure and outcome models in a linear form, as in the article. We set the constraints that an interaction term can be selected into the model only if the main effect term has already been selected, and that a main effect can be removed from the model only if the interaction has been removed. We, then, fit these three approaches to both the full data set and to a subset, obtained by discarding units from both the treatment and control groups with U's not in the range of U's in the other group (King and Zeng, 2005).

We first performed this analysis on the full data. In this case, the estimand is the full-population truth. Results are summarized in the top half of Table 2 and indicate that: (1) the true model is best; this is expected because under the true model we are imputing the missing potential outcomes

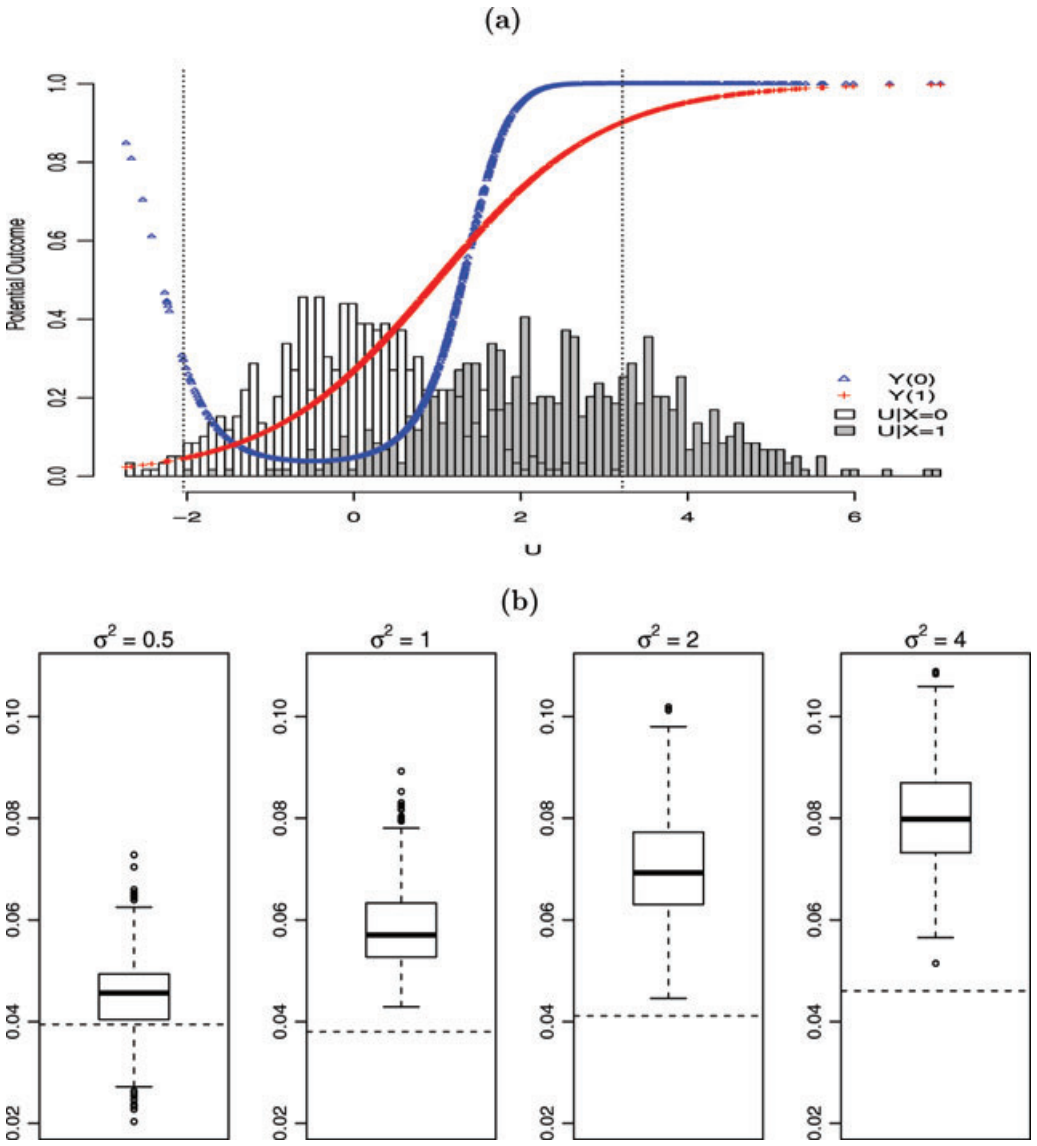

Figure 1. (a) Potential outcomes versus U in a simulated data set with $B = 2$, $\sigma^2 = 2$. Histograms show the distribution of U by group. The two dotted lines indicate the overlapping region of the empirical distributions of U. (b) Subpopulation ACEs from 500 simulated data sets in each simulation scenario. The dashed line is the full-population ACE. Value were calculated based on a simulated data set with extremely large sample size ($n = 3{,}000{,}000$).

perfectly, even outside the range of the observed data; (2) BAC is far superior to *Model* (8) that is, it performs much better than fitting a wrong, but somewhat flexible, model to the full data. This suggests that accounting for model uncertainty robustifies the analysis against model mis-specification. Although a full extension of BAC would require a binary exposure model, in the time allotted for preparation of this rejoinder we could only fit a linear exposure model. Nonetheless, BAC provides a real advantage, likely because even a partly misspecified exposure model can provide useful information on which confounders to include.

Next, we performed an analysis on the overlap region only, changing the estimand to be the sub-population truth. The numerical value of this estimand can differ across simulated data sets. However, unlike in Gutman and Rubin, it is the same for all the methods, for a given data set. It would be more rigorous to define the truth based on a fixed and unknown subpopulation, to generate metrics that also reflect sampling variability in the preprocessing, but for comparability with Gutman and Rubin we did not do so here. Results are in the bottom half of Table 2 and can roughly be compared to those of Gutman and Rubin, though differences may exist in our definitions of subpopulations. Our results indicate that: (1) *Model* (8) and BAC evaluated on the subpopulation perform a lot better than when evaluated on the full data; (2) no clear ranking emerges between BAC and *Model* (8), with the former performing better at the smaller sample size; and (3) the gap with performance reported by Gutman and Rubin

Table 2

Comparison of estimates of ACE from true model, Model (8) *in Gutman and Rubin and BAC. For Model* (8)*, both noninformative (NI) and normal-gamma (NG) priors were considered. Results were based on* 500 *replications. BIAS is the difference between the mean of estimates of ACE and the true value, and RMSE is the root mean square error. Values in the table have been multiplied by* 1000

Estimation of full-population ACE									
		$\sigma^2 = 0.5$		$\sigma^2 = 1$		$\sigma^2 = 2$		$\sigma^2 = 4$	
$n = 600$		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
True model		−0.20	5.03	−0.19	4.88	−0.17	4.63	0.18	4.46
Model (8)	NI prior	122.91	936.68	−35.41	72.13	-5×10^3	4×10^4	1×10^5	5×10^5
	NG prior	18.52	19.69	31.04	37.03	82.28	93.62	174.46	183.97
BAC		12.02	17.45	9.73	13.18	23.65	47.28	−11.20	37.98
		$\sigma^2 = 0.5$		$\sigma^2 = 1$		$\sigma^2 = 2$		$\sigma^2 = 4$	
$n = 60$		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
True model		−0.82	15.03	0.63	14.86	0.08	14.25	0.58	13.23
Model (8)	NI prior	-6×10^4	6×10^5	-8×10^4	6×10^5	-4×10^4	7×10^5	-7×10^4	1×10^6
	NG prior	82.81	109.33	167.60	188.30	248.34	257.58	313.05	318.29
BAC		40.63	68.59	49.63	74.43	64.70	89.52	61.94	93.76
Estimation of sub-population ACE									
		$\sigma^2 = 0.5$		$\sigma^2 = 1$		$\sigma^2 = 2$		$\sigma^2 = 4$	
$n = 600$		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
True model		−0.44	5.55	−0.46	5.12	−0.54	5.24	−0.79	5.75
Model (8)	NI prior	−0.42	5.54	−0.08	5.17	0.28	5.42	−0.80	6.16
	NG prior	3.28	6.65	3.26	6.33	2.99	6.87	0.69	6.73
BAC		7.73	10.28	9.28	13.31	6.16	10.67	0.17	9.39
		$\sigma^2 = 0.5$		$\sigma^2 = 1$		$\sigma^2 = 2$		$\sigma^2 = 4$	
$n = 60$		BIAS	RMSE	BIAS	RMSE	BIAS	RMSE	BIAS	RMSE
True model		−1.66	20.39	−2.81	20.10	−4.70	18.72	−6.83	18.59
Model (8)	NI prior	−90.83	4×10^3	284.11	6×10^3	−264.98	7×10^3	-1×10^5	3×10^6
	NG prior	111.08	162.19	56.58	111.55	12.97	67.59	−11.74	45.44
BAC		62.18	83.88	46.45	71.20	9.23	52.14	−8.97	42.97

for the GR estimates varies with the scenario, but is generally not wide.

In summary, BAC performs well, despite the unrealistic and off-the-topic scenario. Conclusions based on the tables in Gutman and Rubin are not informative for two reasons: (1) what they call WPD is neither BAC, nor close to it; and (2) methods should be compared only when they are estimating the *same* quantity.

Both the Gutman and Rubin and our simulation studies illustrate that estimating a full-population truth in presence of nonlinearities, interactions, and sparse confounders is much harder than estimating a carefully chosen subpopulation truth. Choosing a subpopulation is critical to the success of methods like GR, and relies heavily on the investigators' ability to identify *a priori* the necessary confounders. With a single confounder and large sample size, the preprocessing

and the matching steps are relatively straightforward, and nonparametric estimation is possible. BAC is motivated by the common situation where we have a large number of potential confounders, we don't know which are the key ones, and we don't have an overwhelmingly large sample size. In the presence of many confounders and limited overlap in the distribution of some of the confounders between the treated and untreated, extensions of the ideas in BAC could help in identifying the set of confounders used in preprocessing and matching, and could also potentially lead to significant improvement in these steps. BAC could also help in identifying the important confounders that might need to be included in a regression model if it is deemed necessary to estimate the ACE within each stratum by adding some of the confounders in the regression.

Acknowledgements

We thank Cory Zigler for very helpful comments and Donna Gilbreath for scientific editing. We also thank the Co-Editor and the Associate Editor for valuable comments.

Additional References

Crump, R., Hotz, V., Imbens, G., and Mitnik, O. (2009). Dealing with limited overlap in estimation of average treatment effects. *Biometrika* **96,** 187–199.

Hahn, J. (2004). Functional restriction and efficiency in causal inference. *The Review of Economics and Statistics* **86,** 73–76.

Hirano, K. and Imbens, G. (2005). The propensity score with continuous treatments. In *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives: An Essential Journey with Donald Rubin's Statistical Family*, A. Gelman and X. Meng (eds), chapter 7. Chichester, UK: John Wiley & Sons, Ltd.

Imai, K. and van Dyk, D. A. (2004). Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association* **99,** 854–866.

Imbens, G. W. (2004). Nonparametric estimation of average treatment effects under exogeneity: A review. *Review of Economics and Statistics* **86,** 4–29.

King, G. and Zeng, L. (2005). The dangers of extreme counterfactuals. *Political Analysis* **14,** 131–159.

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.

Robins, J. M., Rotnitzky, A., and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association* **89,** 846–866.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70,** 41–55.

Rosenbaum, P. R. and Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association* **79,** 516–524.

Rubin, D. (1980). Discussion of "Randomization analysis of experimental data: The Fisher randomization test" by Basu. *Journal of the American Statistical Association* **75,** 591–593.

Schafer, J. and Kang, J. (2008). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological Methods* **13,** 279–313.

Schneeweiss, S., Rassen, J. A., Glynn, R. J., Avorn, J., Mogun, H., and Brookhart, M. A. (2009). High-dimensional propensity score adjustment in studies of treatment effects using health care claims data. *Epidemiology* **20,** 512–522.

Stuart, E. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science* **25,** 1–21.

Tan, Z. (2010). Bounded, efficient and doubly robust estimation with inverse weighting. *Biometrika* **97,** 661–682.

Zhang, M., Joffe, M. M., and Small, D. S. (2011). Causal inference for continuous-time processes when covariates are observed only at discrete times. *Annals of Statistics* **39,** 131–173.