Model-Averaged Confounder Adjustment for Estimating Multivariate Exposure Effects

Ander Wilson[1,*], Corwin M. Zigler[1], Chirag J. Patel[2], Francesca Dominici[1]
[1]Department of Biostatistics, Harvard T.H. Chan School of Public Health
[2] Department of Biomedical Informatics, Harvard Medical School

Abstract

In environmental and nutritional epidemiology and in many other fields, there is increasing interest in estimating the effect of simultaneous exposure to several agents (e.g. multiple nutrients, pesticides, or air pollutants) on a health outcome. We consider estimating the effect of a multivariate exposure that includes several continuous agents and their interactions when the number of potential confounding variables is large relatively to the sample size. Our goal is to develop an approach to estimate a multivariate exposure effect with a parsimonious model that is fully adjusted for confounding. We develop a new approach rooted in the ideas of Bayesian model averaging to adjust for confounding in the multivariate exposure setting. We introduce a data-driven, informative prior that assigns to likely confounders a higher probability of being included into a regression model for effect estimation. Our approach can also be formulated as a penalized likelihood that leads to a simple tuning approach. Through a simulation study we demonstrate that the proposed approach identifies parsimonious models that are fully adjusted for observed confounding and estimates the multivariate exposure effect with smaller MSE compared to several alternatives. We apply the method to an Environmental Wide Association Study (EWAS) using National Heath and Nutrition Examination Survey to estimate the effect of mixtures of nutrients and pesticides on lipid levels.

Keywords: Bayesian model averaging, Confounding, Exposome, Model uncertainty, Multiple exposures, Multivariate exposure effects

1 Introduction

With the rapidly increasing availability of environmental exposure data, there is growing interest in studying the multitude of exposures, often referred to as the *exposome*, that may influence complex diseases (Wild, 2005; Louis and Sundaram, 2012). Recent studies have screened large numbers of environmental agents for associations with health outcomes and biological endpoints (Patel et al., 2013; Patel and Ioannidis, 2014; Wilson et al., 2014b). Increasingly,

research is focusing on estimating the health effects associated with simultaneous exposure to a mixture of multiple agents, rather than estimating the effects associated with exposure to a single agent while controlling for others. For example, recent research has estimated the effect of multiple air pollutants and temperature on health outcomes (Wilson et al., 2014a; Zanobetti et al., 2014), the effect of metal mixtures on neurodevelopment outcomes (Bobb et al., 2015), and the effect of mixtures of nutrients on atherosclerosis in mice (Verschuren et al., 2011).

Due to the predominant reliance on observational data, confounding remains a common consideration when estimating the health effects of multiple agents. Methods for confounding adjustment in this context are met with two important challenges. First, the number of measured potential confounders is often large and methods are required to select a more parsimonious set to adjust for. This challenge also exists in the single agent setting and several methods have been proposed to address it when estimating the effect of a single agent (Crainiceanu et al., 2008; Wang et al., 2012a; Wilson and Reich, 2014). The second challenge is unique to the multivariate exposure setting. Rather than estimating the effect of each agent $(X_1, \ldots, X_m)$ individually on an outcome Y, we want to estimate the effect of a multivariate exposure that includes several continuous agents and interactions between agents, e.g. $\mathbf{Z} = (X_1, \ldots, X_m, X_1X_2, \ldots, X_{m-1}X_m)$.

A large body of research addresses confounder adjustment for a single dichotomous treatment using propensity score adjustment (Rosenbaum and Rubin, 1983; Lunceford and Davidian, 2004). In this framework, several model selection approaches have been proposed (e.g. Vansteelandt et al., 2012; Zigler and Dominici, 2014). In the context of multiple bivariate treatments administered simultaneously, Taubman et al. (2009) and Danaei et al. (2013) combined them into a composite measure and then model the composite measure as a single dichotomous treatment. For analysis of failure time data, Hernán et al. (2001) proposed a marginal structural model to estimate the effect of multiple treatments over time when there are time-dependent confounders using inverse probability weighting (Robins et al., 1994). The approach assumes a Cox proportional hazard model to estimate the effect of two time-dependent dichotomous treatments, but does not consider interactions. For a single continuous agent, Imai and van Dyk (2004) proposed a generalized propensity score. This approach models the conditional distribution of the agent given the potential confounders. However, none of these methods fully address the challenge of model selection when estimating the effect of a continuous multivariate exposure that includes interactions on an outcome. Extending propensity score based approaches to this setting would require modeling the conditional distribution of the multivariate exposure $\mathbf{Z}$ given the potential confounders, which is impractical for even a moderate number of agents.

In epidemiologic analyses of the health effect of exposure to a single continuous agent, adjusting for confounding is often addressed with sensitivity analysis or exposure modeling. Sensitivity analysis shows how the estimated exposure effect varies over a range of models that include different sets of potential confounders (Dominici et al., 2004; Peng et al., 2006). Exposure modeling identifies

potential confounders using a model with the exposure as the dependent variable and potential confounders as independent variables (Greenland, 2008). Potential confounders identified with the exposure model are then adjusted for in the outcome model that estimates the exposure effect on the outcome.

When estimating the effect of exposure to a single continuous agent, joint analyses of the exposure and outcome models can identify covariates that are associated with both the exposure and the outcome and therefore are potential confounders. Recent approaches have formalized the framework in model based confounder selection methods for a single continuous agent. Crainiceanu et al. (2008) developed visualization tools and theoretical results for this approach. Bayesian adjustment for confounding (BAC; Wang et al., 2012a, 2015) uses an exposure model, an outcome model, and a joint approach for variable selection and model averaging to adjust for confounding. Bayesian penalized credible region confounder selection (BayesPen; Wilson and Reich, 2014) uses a decision-theoretic approach to find the sparsest solution that contains all important covariates and all confounding variables. Wilson and Reich (2014) also proposed an extension of BayesPen for models with multiple additive exposures. However, all these approaches which rely on exposure modeling have limitations in the context of a multivariate exposure. First, exposure modeling requires the specification of an exposure model for each single agent which is not feasible when the number of agents is large. Second, these approaches identify confounders of the effect of each single agent, not confounders of the effect of a more general multivariate exposure that includes, for example, interactions between agents.

The goal of this paper is to address the challenge of identifying a parsimonious model for confounder adjustment to estimate the multivariate exposure effect. To support this goal, we introduce the notion of confounding of a multivariate exposure effect using the formality of the potential-outcomes framework for causal inference (Rubin, 1978). In doing so, we elucidate the key point that existing approaches developed in the setting of a single exposure cannot be easily adapted to target confounders of the multivariate exposure effect.

While this clarification of confounding of the multivariate exposure effect is accomplished without regard to any particular statistical model, our estimation approach follows the line of research in (Wang et al., 2012a; Wilson and Reich, 2014) that relies on the mechanics of a linear regression model. The proposed method is based on a Bayesian model averaging approach (BMA; Raftery et al., 1997) with a prior probability of including each potential confounder in the outcome model specifically designed to adjust for confounding in the context of a multivariate exposure. The prior has two desirable properties: 1) covariates that are linearly associated with both the multivariate exposure vector $\mathbf{Z}$ and the outcome Y have high probability of being included in the outcome model; and 2) using an easily implemented tuning approach, covariates that are associated with the exposure $\mathbf{Z}$ only, but not with the outcome, are not forced into the model. The approach can also be formulated as a penalized likelihood problem which yields an intuitive tuning approach for the strength of the prior on covariate inclusion. Our proposed method is computationally simple even

when the multivariate exposure becomes larger (e.g. 50 or 100 including interactions). We also make software available in an `R` package called `regimes` (REGression In Multivariate Exposure Settings). We apply the proposed approach to data from the National Health and Nutrition Examination Survey (NHANES) and estimate the effect of 132 agents grouped into 24 multivariate exposures representing clusters of related nutrients and persistent pesticides on lipid levels.

2 Effects of Simultaneous Exposure to Multiple Agents: Notation, Estimand, and Confounding

We begin by formulating notation, defining the estimand of interest, and clarifying notions of confounding in the present context. Each of the quantities defined in this section are assumed available on a sample of $i = 1, \ldots, n$ individuals, but quantities are defined for a single individual for ease of exposition. Let $\mathbf{X} = (X_1, X_2, \ldots, X_m)$ denote the levels of m agents, for example, measures of m persistent pesticides measured in a person's blood. Let Y denote the measured health outcome of interest, for example, a measure of lipid levels in blood. In addition, we observe a large vector of pre-exposure covariates, $\mathbf{C} = (C_1, \ldots, C_k)$.

Formally, interest lies in the change in a health outcome that would result from a change in the simultaneous exposure to multiple environmental agents. Let $\boldsymbol{\delta}$ be an m-dimensional vector used to denote a change in agents that could arise from a (possibly hypothetical) intervention such as a reduction in exposure to m pesticides. The goal is to estimate the change in Y that would result from a shift from $\mathbf{X}$ to $\mathbf{X} + \boldsymbol{\delta}$.

When $m > 1$, there is a key distinction between the *agents* and what we term the *multivariate exposure*. This distinction is necessary because simple shifts in each of the m agents may produce complicated changes to the multivariate exposure comprised of the individual agents and any relationships among the agents (e.g. interactions). Let $\mathbf{Z} = z(\mathbf{X})$ denote the r-dimensional multivariate exposure comprised of the m individual agents and possibly complicated functions of the individual agents. The importance of the distinction between $\mathbf{X}$ and $\mathbf{Z}$ derives from the fact that health effects and confounding must be defined with regard to shifts in $\mathbf{Z}$ (agents and interacting functions of agents), even though scientific reasoning often focuses on changes in $\mathbf{X}$, as shifts the agents themselves can typically be targeted for manipulation via, for example, interventions. Even in a setting where an experimentalist can control $\mathbf{X}$ and $\boldsymbol{\delta}$, the nature of the relationships among each of the agents captured in $\mathbf{Z}$ may be entirely governed by natural phenomena and not under experimenter control.

In order to be explicit about the the effect of interest and the threat of confounding when estimating this effect, we turn to the potential outcomes notation that underlies a vast literature on causal inference (Rubin, 1978). With

$Y(\mathbf{Z})$ denoting the potential outcome that would be observed under multivariate exposure $\mathbf{Z}$, the average causal effect of a shift in $\mathbf{Z}$ from $\mathbf{z} = z(\mathbf{x})$ to $\mathbf{z}' = z(\mathbf{x}+\boldsymbol{\delta})$ is defined as:

$$\Delta_{\mathbf{z},\mathbf{z}'} = \Delta_{\mathbf{x},\boldsymbol{\delta}} = E[Y(\mathbf{z}') - Y(\mathbf{z})] = E[Y(z(\mathbf{x}+\boldsymbol{\delta})) - Y(z(\mathbf{x}))]. \quad (1)$$

The effect in (1) can be estimated from observed data under the assumption of *strongly ignorable multivariate exposure assignment* stating that, conditional on the *confounders* (denoted by $\mathbf{C}^*$), the potential outcomes $Y(\mathbf{Z})$ are independent of $\mathbf{Z}$:

$$Y(\mathbf{Z}) \perp\!\!\!\perp \mathbf{Z}|\mathbf{C}^* \quad (2)$$

for all possible values of $\mathbf{Z}$. Here, we assume that the set of confounders $\mathbf{C}^*$ is a subset of the measured covariates $\mathbf{C}$. This amounts to the standard "no unmeasured confounding" assumption, where in this case, care is taken to denote this conditional independence with respect to potential outcomes indexed by the multivariate exposure $\mathbf{Z}$, and not simply with respect to the agents $\mathbf{X}$. We illustrate the necessity of this distinction with an example in Web Appendix A.

Thus, the confounders of the multivariate exposure effect are those required to satisfy assumption (2), adjustment for which would allow the following quantity to serve as an estimator of $\Delta_{\mathbf{z},\mathbf{z}'}$:

$$\widehat{\Delta}_{\mathbf{z},\mathbf{z}'} = E[Y|z(\mathbf{x}+\boldsymbol{\delta}) = \mathbf{z}', \mathbf{C}^*] - E[Y|z(\mathbf{x}) = \mathbf{z}, \mathbf{C}^*], \quad (3)$$

representing an estimate of the average effect of a (hypothetical) intervention that acts on the agents by shifting them from $\mathbf{x}$ to $\mathbf{x}+\boldsymbol{\delta}$ and producing a change in multivariate exposure from $z(\mathbf{x})$ to $z(\mathbf{x}+\boldsymbol{\delta})$.

The methods developed in the subsequent section are designed to address the setting where the $\mathbf{C}^*$ required to satisfy (2) and, therefore, estimate the multivariate exposure effect with (3) are not known *a priori*. We introduce a data-driven method to identify $\mathbf{C}^*$ while offering a proper account of uncertainty. In practice, this will require 1) specification of the function $z(\mathbf{X})$ constituting the interrelationships among multiple agents and 2) specification of a statistical model to estimate $\Delta_{\mathbf{z},\mathbf{z}'}$, as defined in (3).

3 Model

3.1 Approach

The above discussion of confounding of the multivariate exposure effect is not tied to any specific statistical model. Here, we consider a setting where we are interested in several continuous agents (nutrients and pesticides) and a continuous outcome (lipid levels in the blood). We want to estimate the multivariate exposure effect in (3) with a linear regression model. Specifically, we assume the "augmented" linear model defined by Raftery et al. (1997):

$$Y_i = \boldsymbol{\beta}_0^{\boldsymbol{\alpha}} + \mathbf{Z}_i^T \boldsymbol{\beta}^{\boldsymbol{\alpha}} + \sum_{j=1}^{k} C_{ji}\alpha_j\eta_j + \epsilon_i. \quad (4)$$

In (4), $\alpha_j \in \{0,1\}$ indicates whether covariate C_j is included into the model. Hence, $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_k)^T$ identifies the set of covariate $\{C_j : \alpha_j = 1\}$ included into (4).

Conditionally on a pre-specified $\boldsymbol{\alpha}$, say $\boldsymbol{\alpha}_l = (1, 1, 0, \ldots, 1)$, we can estimate the posterior distribution of the exposure effect $\Pr(\Delta^{\boldsymbol{\alpha}_l}_{\mathbf{z},\mathbf{z}'} | \mathbf{Z}, \mathbf{C}, \mathbf{Y}, \boldsymbol{\alpha}_l)$ using standard Bayesian methods. To account for model uncertainty about the choice of confounders included in the model we estimate the model averaged posterior distribution of $\Delta_{\mathbf{z},\mathbf{z}'}$ as

$$\Pr\left(\Delta_{\mathbf{z},\mathbf{z}'} | \mathbf{Z}, \mathbf{C}, \mathbf{Y}\right) \approx \sum_{l \in \mathcal{A}} \Pr\left(\Delta^{\boldsymbol{\alpha}_l}_{\mathbf{z},\mathbf{z}'} | \mathbf{Z}, \mathbf{C}, \mathbf{Y}, \boldsymbol{\alpha}_l\right) \Pr\left(\boldsymbol{\alpha}_l | \mathbf{Z}, \mathbf{C}, \mathbf{Y}\right), \qquad (5)$$

where $\Pr\left(\boldsymbol{\alpha}_l | \mathbf{Z}, \mathbf{C}, \mathbf{Y}\right) \propto \Pr\left(\mathbf{Y} | \mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha}_l\right) \Pr\left(\boldsymbol{\alpha}_l\right)$ and $l \in \mathcal{A}$ indexes the k^2 possible confounder adjustment combinations (Raftery et al., 1997; Hoeting et al., 1999). The equality in (5) is approximate because the estimates of $\Delta^{\boldsymbol{\alpha}_l}_{\mathbf{z},\mathbf{z}'}$ for values of $\boldsymbol{\alpha}_l$ that exclude at least one element of $\mathbf{C}^*$ are not interpretable as estimates of the causal effect $\Delta_{\mathbf{z},\mathbf{z}'}$. The extent of approximation relates directly to the ability to assign high posterior weight only to values of $\boldsymbol{\alpha}$ that contain all elements of $\mathbf{C}^*$. This framework was motivated by Wang et al. (2012a) and further discussed in Zigler and Dominici (2014).

In practice the prior distribution $\Pr\left(\boldsymbol{\alpha}\right)$ is commonly assumed to be non-informative, $\Pr(\alpha_j = 1) = \Pr(\alpha_j = 0) = 0.5$ for $j = 1, \ldots, k$. Wang et al. (2012a) showed that applying BMA with this prior on $\boldsymbol{\alpha}$ will select covariates that are associated with Y only. Therefore, it will minimize prediction error for Y but will not minimize the confounding bias for $\widehat{\Delta}_{\mathbf{z},\mathbf{z}'}$.

Because the observed data cannot tell us which set of C_j satisfy (2), we learn from the data which covariates are associated with both $\mathbf{Z}$ and Y. We consider these covariates to be likely confounders and construct a prior that ensures that these likely confounders are included into the regression models with high probability.

3.2 Construction of prior to adjust for confounding

In this section we introduce an informative prior on $\boldsymbol{\alpha}$, so that we can estimate $\Delta_{\mathbf{z},\mathbf{z}'}$ by averaging only over models that include the confounders $\mathbf{C}^*$. We also compare the proposed prior with alternative choices and show full details of the calculations used in this section in Web Appendix B.

We start by calculating the confounding bias in $\widehat{\Delta}_{\mathbf{z},\mathbf{z}'}$ that would occur if we would fit model (4) without any adjustment, e.g. $\boldsymbol{\alpha} = \mathbf{0}$. This is the difference between the posterior mean of $\Delta_{\mathbf{z},\mathbf{z}'}$ under a model that includes all observed covariates ($\boldsymbol{\alpha} = \mathbf{1}$) and posterior mean of $\Delta_{\mathbf{z},\mathbf{z}'}$ under a model that does not include any observed covariates ($\boldsymbol{\alpha} = \mathbf{0}$). Under a flat prior on $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ and inverse-gamma prior on σ^2 (with fixed hyper-parameters), the confounding bias

of $\widehat{\Delta}_{\mathbf{z},\mathbf{z}'} = (\mathbf{z}' - \mathbf{z})^T \widehat{\boldsymbol{\beta}}$ is

$$E\left(\widehat{\Delta}_{\mathbf{z},\mathbf{z}'}|\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\alpha}=\mathbf{1}\right) - E\left(\widehat{\Delta}_{\mathbf{z},\mathbf{z}'}|\mathbf{Y},\mathbf{Z},\mathbf{C},\boldsymbol{\alpha}=\mathbf{0}\right)$$
$$= (\mathbf{z}' - \mathbf{z})^T \left(\mathbf{Z}^T\mathbf{Z}\right)^{-1} \mathbf{Z}^T\mathbf{C} \left(\mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{C}\right)^{-1} \mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{Y}, \tag{6}$$

where $\mathbf{P}_Z = \mathbf{Z}(\mathbf{Z}^T\mathbf{Z})^{-1}\mathbf{Z}^T$ is the perpendicular projection onto the column space of $\mathbf{Z}$ and $\mathbf{P}_Z^{\perp} = \mathbf{I} - \mathbf{P}_Z$. The confounding bias in (6) is zero when $\left(\mathbf{Z}^T\mathbf{Z}\right)^{-1}\mathbf{Z}^T\mathbf{C} = \mathbf{0}$ or $\mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{Y} = \mathbf{0}$ and, therefore, is zero when $\mathbf{C}$ is independent of either $\mathbf{Z}$ or Y (which would occur if no element of $\mathbf{C}$ were a confounder of the multivariate exposure effect).

Under the assumption that $(\mathbf{z}' - \mathbf{z}) = \mathbf{a}^T\mathbf{Z}$ for some vector $\mathbf{a}$, that is, the shift in exposure being studied is a linear combination of the observed multivariate exposures $\mathbf{Z}$, the confounding bias defined in (6) is equal to 0 if:

$$\sum_{l=1}^{r} \left[\sqrt{\zeta_l}\mathbf{q}_l^T \left(\mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{C}\right)^{-1} \mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{Y}\right]^2 = 0 \tag{7}$$

where $\mathbf{C}^T\mathbf{P}_Z\mathbf{C} = \sum_{l=1}^{k} \zeta_l \mathbf{q}_l\mathbf{q}_l^T$ is the spectral decomposition.

We construct our informative prior for $\boldsymbol{\alpha}$ by: 1) isolating the part of (7) that is a function only of $\mathbf{Z}$ and $\mathbf{C}$, so as not to have the prior depend on the outcome Y; and 2) assigning higher prior inclusion probabilities to the covariates that are linearly dependent with $\mathbf{Z}$ and therefore contribute most to the confounding bias expression in (7). Specifically, we assume that α_j has *a priori* a Bernouilli distribution with mean $\pi_j(\mathbf{Z},\mathbf{C},\lambda) = \text{Logit}^{-1}\{\lambda\omega_j(\mathbf{Z},\mathbf{C})\}$, where $\boldsymbol{\omega}(\mathbf{Z},\mathbf{C}) = \sum_{l=1}^{r}\left[\sqrt{\zeta_l}\mathbf{q}_l^T\left(\mathbf{C}^T\mathbf{P}_Z^{\perp}\mathbf{C}\right)^{-1}\right]^2$ is a k-vector taken from (7) and $\lambda \geq 0$ is a tuning parameter. Hence, $\lambda\omega_j(\mathbf{Z},\mathbf{C})$ is the prior log-odds of including C_j in regression model (4).

This prior specification has the following desirable properties. If C_j is linearly independent of $\mathbf{Z}$ then $\omega_j(\mathbf{Z},\mathbf{C}) = 0$ and the prior probability of including C_j is $\Pr(\alpha_j = 1|\mathbf{Z},\mathbf{C},\lambda) = 0.5$. When $\lambda > 0$, if C_j is linearly dependent with $\mathbf{Z}$ then $\omega_j(\mathbf{Z},\mathbf{C}) > 0$, and $\Pr(\alpha_j = 1|\mathbf{Z},\mathbf{C},\lambda) \in (0.5, 1)$ is proportional to the strength of this association. The stronger the association between C_j and $\mathbf{Z}$, the larger the potential confounding bias that could result from omitting C_j, and the larger $\omega_j(\mathbf{Z},\mathbf{C})$. Therefore, a covariate C_j that has high potential to cause confounding bias if omitted is assigned higher prior probability of inclusion than a covariate that has relatively low potential to cause confounding bias if omitted.

Another key requirement in building our informative prior for $\boldsymbol{\alpha}$ is to be able to exclude covariates that are associated only with $\mathbf{Z}$ and are independent from Y. We address this with the tuning parameter λ which controls the strength of the prior. When $\lambda \to \infty$ we assume that any covariate that is correlated with $\mathbf{Z}$ (even weakly) is forced into the model regardless of whether it is associated with Y (an undesirable property for a prior). In this situation (large λ) we prioritize confounding adjustment over model parsimony. In contrast, when

$\lambda = 0$ we assume that the prior on $\boldsymbol{\alpha}$ is flat, $\pi_j(\mathbf{Z}, \mathbf{C}, 0) = 0.5 \ \forall j$. In this situation ($\lambda = 0$) the prior provides no improvement to confounder adjustment and we prioritize model parsimony over confounding adjustment. Hence, the structure of the prior comes from $\omega_j(\mathbf{Z}, \mathbf{C})$ and the strength of the prior from λ. We discuss choices for λ in Section 3.3 and recommend a tuning method that balances confounder adjustment and model parsimony.

We complete the Bayesian specification for the normal linear model following that of Raftery et al. (1997) and provide details in Web Appendix B. The R package `regimes` provides software to use this method and reproduce the simulated data.

3.3 Relation to a penalized likelihood approach and selecting λ

The proposed approach can be formulated as a penalized likelihood problem where the prior $\Pr(\boldsymbol{\alpha}|\mathbf{Z}, \mathbf{C}, \lambda)$ translates to a penalty that reflects our goals of minimizing confounding bias and maximizing model parsimony. This will guide the selection of the tuning parameter λ.

We approximate the posterior model probability of $\Pr(\boldsymbol{\alpha}|\mathbf{Y}, \mathbf{X}, \mathbf{C}) \propto \Pr(\mathbf{Y}|\mathbf{Z}, \mathbf{C}, \boldsymbol{\alpha}) \times \Pr(\boldsymbol{\alpha}|\mathbf{Z}, \mathbf{C}, \lambda)$ using the Bayesian information criterion (BIC; Schwarz, 1978) as

$$\text{BIC}(\boldsymbol{\alpha}) = -2 \times ll(\mathbf{Y}, \mathbf{X}, \mathbf{C}; \boldsymbol{\alpha}, \widehat{\boldsymbol{\beta}}_{\boldsymbol{\alpha}}, \widehat{\boldsymbol{\eta}}_{\boldsymbol{\alpha}}, \widehat{\sigma}^2_{\boldsymbol{\alpha}}) - 2\lambda \sum_{j=1}^{k} \alpha_j \omega_j(\mathbf{Z}, \mathbf{C}) + \log(n) \sum_{j=1}^{k} \alpha_j. \tag{8}$$

The three terms on the the right hand side of (8) are, from left to right: 1) the negative profile log-likelihood of the normal linear regression model which is minimized when covariates predictive of Y are included into the regression model; 2) the negative prior log-odds of covariate inclusion defined in Section 3.2; and 3) the BIC sparsity penalty. Hence, the proposed method optimizes model fit while balancing confounder adjustment and model parsimony.

The second and third terms of (8) can be combined and viewed as a single penalty on the log-likelihood. The penalty for including covariate j is $\alpha_j [\log(n) - 2\lambda\omega_j(\mathbf{Z}, \mathbf{C})]$. The tuning parameter λ controls the balance between confounder adjustment (large λ) and model parsimony (small λ). A unique feature of this penalty is that it can be either positive, acting as a penalty, or negative, acting as an incentive for covariate inclusion.

This provides a theoretical basis to select λ such that we can achieve balance between parsimony and confounder adjustment. Specifically, we balance these competing interests by choosing λ so that the penalty $\alpha_j [\log(n) - 2\lambda\omega_j(\mathbf{Z}, \mathbf{C})]$ is on average zero. Setting $0 = k^{-1} \sum_{j=1}^{k} \alpha_j [\log(n) - 2\lambda\omega_j(\mathbf{Z}, \mathbf{C})]$ yields the balancing penalty

$$\lambda^* = \frac{k \log(n)}{2 \sum_{j=1}^{k} \omega_j(\mathbf{Z}, \mathbf{C})}. \tag{9}$$

In practice, $\lambda = \lambda^*$ can be seen as a benchmark choice. By taking $\lambda > \lambda^*$ we include more covariates and take a more conservative approach to confounder

adjustment at the expense of less parsimony. Alternatively $\lambda < \lambda^*$ prioritizes sparsity over confounder adjustment.

4 Simulation

4.1 Simulation scenario one

The first simulation scenario includes two agents and their interaction, $\mathbf{Z} = (X_1, X_2, X_1X_2)^T$, plus 100 covariates $\mathbf{C}$ with $n \in \{200, 500\}$ observations. We have simulated the data so that: $C_1, \ldots, C_{15}$, are associated with Y and X_1 and/or X_2; $C_{16}, \ldots, C_{20}$ are associated with Y and X_1X_2; $C_{21}, \ldots, C_{30}$ are associated with Y but not with $\mathbf{Z}$ (predictors of Y); $C_{31}, \ldots, C_{35}$ are associated with $\mathbf{Z}$ but not Y (instrumental variables) and should not be included in the model; and $C_{36}, \ldots, C_{100}$ are independent of both Y and $\mathbf{Z}$ (noise). Hence, the true model contains only covariates $C_1, \ldots, C_{30}$ and the set of confounders $\mathbf{C}^*$ is $\{C_1, \ldots, C_{20}\}$. Specifically, we simulate 1000 data sets as follows:

$$
\begin{aligned}
C_{ji} &\sim \mathrm{N}(0,1) \quad \text{for } j = 1, \ldots, 100 \\
X_{1i} &\sim \mathrm{N}\left(11^{-1/2} \sum_{j=1}^{10} C_{ji}, 11^{-1/2}\right) \\
X_{2i} &\sim \mathrm{N}\left(21^{-1/2}\left[\sum_{j=6}^{15} C_{ji} + X_{1i} \sum_{j=16}^{50} C_{ji} + \sum_{j=31}^{35} C_{ji}\right], 21^{-1/2}\right) \\
Y_i &\sim \mathrm{N}\left(\mathbf{Z}\boldsymbol{\beta} + \sum_{j=1}^{30} \eta_j C_{ji}, 1\right). \qquad (10)
\end{aligned}
$$

For X_{1i} and X_{2i} the regression coefficients $11^{-1/2}$ and $21^{-1/2}$ are chosen so that both agents have variance 1. Finally, $\{\beta_j\}_{j=1}^3$ and $\{\eta_j\}_{j=1}^{30}$, are simulated as independent Uniform$(0.2, 0.5)$.

We compare the proposed method, henceforth ACPME (adjustment for confounding in the presence of multivariate exposures), to five alternatives: 1) BMA (equivalent to ACPME with $\lambda = 0$, that is $\Pr(\alpha_j = 1|\mathbf{Z}, \mathbf{C}, \lambda) = 0.5\ \forall j$); 2) BayesPen using one exposure model for each agent but not for the interactions; 3) the full Bayesian regression model that includes all 100 measured covariates; 4) the true Bayesian regression model that controls for covariates 1 to 30 only; and 5) the unadjusted Bayesian regression model that regresses the outcome on $\mathbf{Z}$ with no covariate adjustment. For ACPME we use $\lambda = \lambda^*$ as described in Section 3.3. We estimate the effect of a simultaneous change in both agents from 0 to 1. The true value of the multivariate exposure effect is $\Delta_{\mathbf{0},\mathbf{1}} = \beta_1 + \beta_2 + \beta_3$.

Figure 1 shows the prior (panels 1a and 1b) and posterior (panels 1c and 1d) inclusion probabilities for each C_j with ACPME. The posterior inclusion probabilities for the C_j in $\mathbf{C}^*$ (covariates 1 to 20) under the ACPME approach are all near one. *A priori*, the covariates that are only associated with the

exposure and not the outcome (covariates 31 to 35) have high inclusion probabilities, but *a posteriori* these inclusion probabilities are less than 0.5. Hence, setting the tuning parameter to λ^* as defined in Section 3.3 balances the goals of parsimony and confounder adjustment and does not force covariates into the model that are only associated with $\mathbf{Z}$ and not Y.

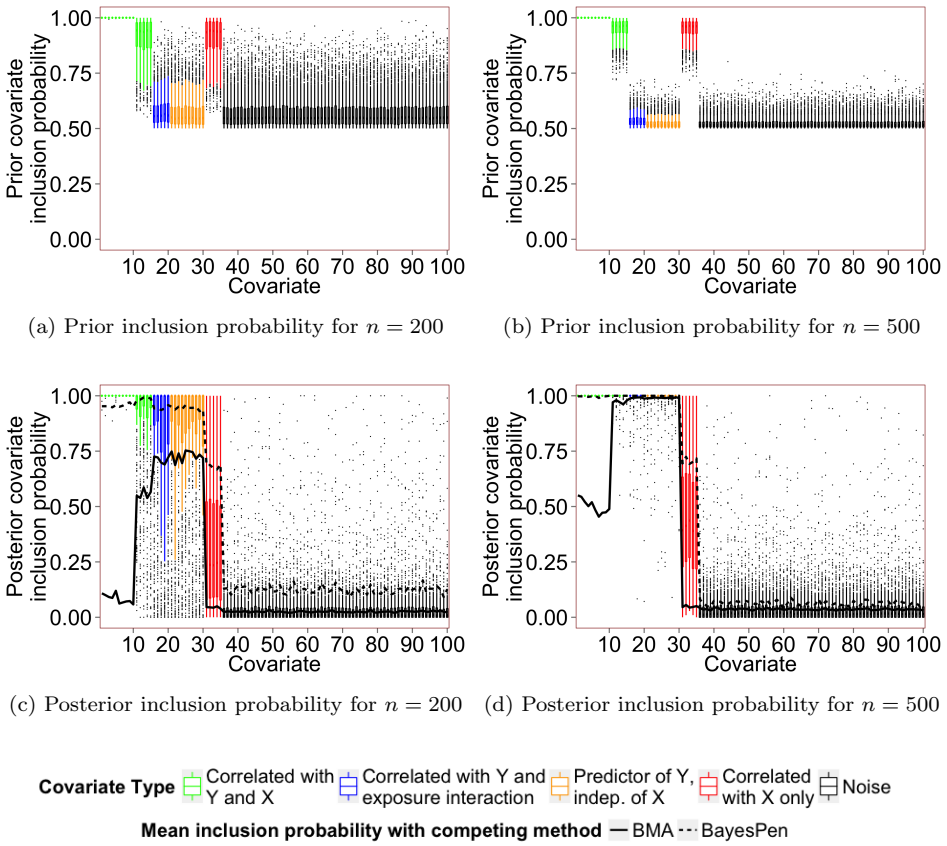
For comparison, Figures 1c and 1d show the mean posterior inclusion probability with BMA and the covariate selection rate with BayesPen. BMA has lower posterior inclusion probabilities than ACPME for the confounders. BayesPen selects true confounders at a high rate but includes covariates that are independent of the outcome and should not be included in the model (covariates 31 to 100) at a higher rate, relative to ACPME.

Table 1 show results for estimating the exposure effect $\boldsymbol{\Delta_{0,1}}$. At both sample sizes ($n = 200$ and 500), ACPME has lower root mean square error (RMSE) compared to all the alternatives except for the true model. In addition, ACPME has credible interval coverage near or at the nominal level. Estimates of $\boldsymbol{\Delta_{0,1}}$ from BMA are biased and have larger RMSE because the prediction-oriented approach under adjust for confounding. This highlights the effectiveness of the ACPME informative prior. BayesPen had lower RMSE than the full model but higher than ACPME indicating the importance of identifying confounders of the multivariate exposure rather than confounders of the individual agents.

4.2 Simulation with large number of agents

We conduct a second simulation to evaluate the performance of ACPME relative to alternative methods in the context of a high-dimensional multivariate exposure and the same $n = 200$ and 500. We generate a new data set for several specifications of $\mathbf{Z}$, where we allow the number of agents to vary from $m = 2, \ldots, 10$ and include all pairwise interactions so that $\mathbf{Z}$ has $m + m(m-1)/2$ columns (ranging from 3 to 55 including main effects and interactions).

For each sample size we simulate 100 covariates so that: $C_1, \ldots, C_{15}$ are associated with at least one of the agents $X_1, \ldots, X_m$ and Y; $C_{16}, \ldots, C_{25}$ are associated with Y and at least one of the interactions between agents; $C_{26}, \ldots, C_{30}$ are predictors of Y that are independent of $\mathbf{Z}$; and $C_{31}, \ldots, C_{100}$ are independent of both Y and $\mathbf{Z}$. Here, the true model includes covariates $C_1, \ldots, C_{30}$ and the set of confounders is $C_1, \ldots, C_{25}$. The specifics of the data generating



(a) Prior inclusion probability for $n = 200$

(b) Prior inclusion probability for $n = 500$

(c) Posterior inclusion probability for $n = 200$

(d) Posterior inclusion probability for $n = 500$

Figure 1: Prior (top) and posterior (bottom) inclusion probabilities for each covariate in simulation scenario one for our proposed method. The box plots show the distribution of prior and posterior probabilities across 1000 simulated data sets. For comparison, the lines in panels 1c and 1d show the average posterior inclusion probability for BMA (solid line) and the proportion of times each covariate was selected into the model with BayesPen (dashed line). This figure appears in color in the electronic version of this article.

Table 1: Simulation results for simulation scenario 1. The first four columns show the mean bias, mean RMSE, mean posterior SD or SE, and 95% interval coverage rate. The right most columns show statistics for covariate inclusion–the true inclusion rate defined as the mean probability that the true confounders and predictors of the outcome (covariates 1 to 30) are included into the regression model and the false selection rate defined as the mean probability that covariates independent of the outcome are included in the model (covariates 31 to 100). Covariates are consider included if they have posterior inclusion probability exceeding 0.5.

Method	Bias	RMSE	Mean SD / SE	95% Int. Coverage	True Inc. Rate	False Sel. Rate
$n = 200$						
ACPME	0.07	0.34	0.34	0.94	0.89	0.06
BayesPen	0.11	0.42	0.28	0.78	0.96	0.17
BMA	1.23	1.25	0.17	0.00	0.48	0.03
Unadjusted	1.65	1.66	0.21	0.00	0.00	0.00
Full	0.00	0.43	0.44	0.95	1.00	1.00
True	0.00	0.28	0.29	0.96	1.00	0.00
$n = 500$						
ACPME	0.02	0.18	0.19	0.96	1.00	0.07
BayesPen	0.03	0.21	0.18	0.91	1.00	0.11
BMA	0.66	0.78	0.15	0.24	0.84	0.04
Unadjusted	1.64	1.65	0.13	0.00	0.00	0.00
Full	0.02	0.21	0.21	0.96	1.00	1.00
True	0.01	0.17	0.17	0.95	1.00	0.00

method are as follows:

$$
\begin{aligned}
C_{ji} &\sim \mathrm{N}(0,1) \quad \text{for } j = 1,\ldots,100 \\
h_j &\sim \mathrm{Cat}(1,\ldots,m_1) \quad \text{for } j = 1,\ldots,10 \\
h_j &\sim \mathrm{Cat}(1,\ldots,m) \quad \text{for } j = 11,\ldots,25 \\
q_j &\sim \mathrm{Cat}(m_1+1,\ldots,m_2) \quad \text{for } j = 1,\ldots,15 \\
X^*_{ki} &\sim \mathrm{N}\left(\sum_{j=1}^{10} C_{ji}\mathbb{1}\{h_j = k\}X_{q_j i} + \sum_{j=11}^{25} C_{ji}\mathbb{1}\{h_j = k\}, 1\right) \\
Y_i &\sim \mathrm{N}\left(\mathbf{Z}\boldsymbol{\beta} + \sum_{j=1}^{30}\eta_j \mathbf{C}_{ji}, 1\right), \qquad (11)
\end{aligned}
$$

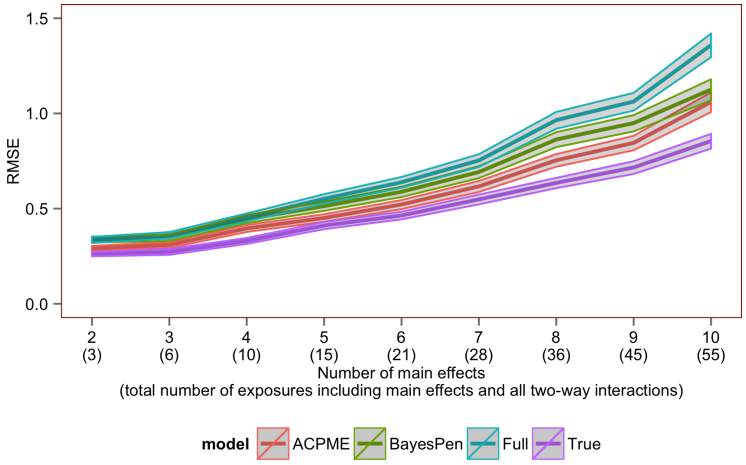
where $\mathrm{Cat}(1,\ldots,l)$ indicates a categorical random variable that takes integer values $1,\ldots,l$ with equal probability $1/l$, $m_1 = m/2$ rounded down to the nearest integer, $m_2 = m - m_1$, and, X_{ki} is X^*_{ki} scaled to have variance 1 and $\mathbf{Z}$ includes all $\mathbf{X}$ and all parities interactions. The regression coefficients $\{\beta_l\}_{l=1}^{r}$ and $\{\eta_j\}_{j=1}^{30}$, are independent Uniform$(0.2, 0.5)$. Web Figure 3 illustrates the correlation structure in the data.

Figure 2 shows the RMSE for estimating the exposure effect for one unit increase in each agent, $\Delta_{\mathbf{0},\mathbf{1}}$. At the smaller sample size ($n = 200$), ACPME has lower RMSE than BayesPen and both these two approaches have lower RMSE than the full model. The RMSE of ACPME improves relative to the alternative approaches as the dimension of the exposure increases. At $n = 500$, ACPME has near identical RMSE to the true model and there are no statistically significant differences between the methods. Hence, even when the multivariate exposure is large relative to sample size, the ACPME informative prior captures information about confounding covariates and the choice of tuning parameter λ^* is appropriate .

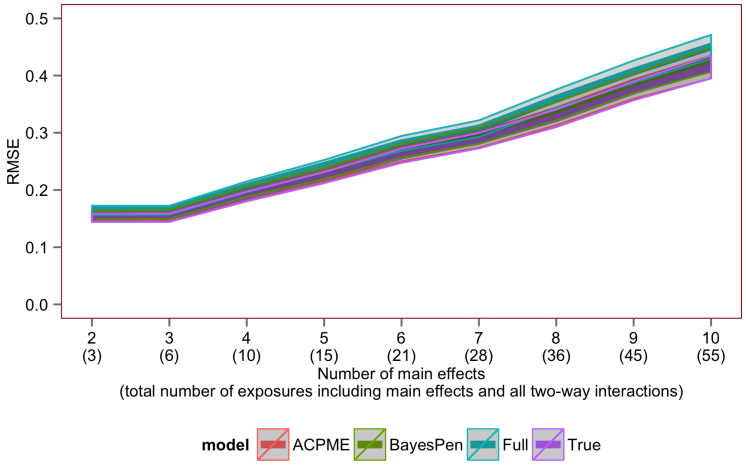
5 Analysis of the NHANES Data

5.1 Overview of the data and analysis

We apply ACPME to the NHANES data as previously described by Patel et al. (2012). We briefly recap here and note important differences for the multivariate analysis. The data combines laboratory and questioner data from the 1999-2000, 2001-2002, 2003-2004, and 2005-2006 surveys. Each survey is a non-overlapping sample representative of the general US population. The data include blood serum and urine biomarkers measurements of 132 nutrients and persistent pesticides (agents). We consider three lipid levels as outcomes: 1) low-density lipoprotein-cholesterol (LDL), 2) high-density lipoprotein-cholesterol (HDL), and 3) triglyceride. In an EWAS analysis, Patel et al. (2012) screened these agents independently for their marginal associations with each of the three outcomes. The EWAS controlled for a small, pre-specified set of individual co-



(a) $n = 200$



(b) $n = 500$

Figure 2: RMSE of the estimated exposure effect for simulation scenario two for $n = 200$ (panel 2a) and $n = 500$ (panel 2b). The x-axis shows the number of agents and in parentheses the dimension of the multivariate exposure including all main effects and two-way interactions. In all cases there are 100 additional covariates in the model of which 30 are true confounders or and predictors of the outcome that should be included in the model. The true model only includes the 30 important covariates, the multivariate exposure, and an intercept. This figure appears in color in the electronic version of this article.

variates (such as age, gender, and BMI) but not for co-exposures to the other nutrients and persistent pesticides. As noted by Patel and Ioannidis (2014) multiple agents are often highly correlated and confounding may underlie many of the strong correlations observed in EWAS. Hence, it is important to consider a multivariate exposure and further adjust for confounding.

We group the 132 agents into 24 mutually exclusive exposure groups of related agents, defined by Patel et al. (2012), that may effect the same biological pathways. We are interested in estimating the effect of a simultaneous change in exposure to all agents on each of the three outcomes. The multivariate exposure includes all agents within each of the 24 groups and their pairwise interactions. We consider agents in other groups and individual level covariates as potential confounders. The change of interest (earlier denoted by $\boldsymbol{\delta}$) is a one unit increase in each scaled agent within the group, equivalent to a one standard deviation increase of the agents on log scale. We conduct separate analyses to estimate the exposure effect of each of the 24 groups on the 3 outcomes (72 models total).

Different exposures were measured for different subsets of the sample frame. As such, the number of persons with available measurements for lipid levels, nutrient, and persistent pesticide exposures ranged by group. Because we are interested in a joint analysis of multiple agents we limit our analysis to the subset of individuals with complete data on multiple agents. This may result in selection bias. Estimates identified as potentially biased due to our selection approach are presented as faded in the analysis figures and are not discussed. Details for the selection process and screening for selection bias are in Web Appendix C.

Table 2 shows the 24 exposure groups, the sample size (n ranging from 158 to 1370), number of agents in each group (m ranging from 1 to 22), and the number of covariates included as potential confounders (k ranging from 22 to 92). The potential confounders include: agents in the other groups that are measured in that subsample; nine body measurements (weight; standing height; body mass index; upper leg length; maximal calf circumference; waist circumference; thigh circumference; triceps skinfold; and subscapular skinfold) and 13 demographic and socioeconomic status variables (age; age squared; poverty to income ratio; indicator for any heard disease; indicator of at least one chronic disease; indicators for race/ethnicity: black, Mexican-American, other hispanic, other race/ethnicity; indicator for female; indicators for SES tertile; indicators for education: less than high school, high school, or more than high school). In the last column of Table 2 we shows the ratio of p divided by n, which ranges from 0.77 to 0.03. Web Figure 4 shows which groups are included as covariates for the analysis of other exposures.

5.2 Comparison of estimates to the full and unadjusted models

We estimate the multivariate exposure effect with ACPME, the full model including all k potential confounders, and the unadjusted model that controls for

Table 2: Summary of the NHANES data by exposure group. The table shows the total number of subjects with complete observations (n); the number of agents (m); the dimension of the multivariate exposure including the main effect of each agent and each two-way interaction (r); the total number of potential confounders including the main effect of agents in other groups (k); the total number of independent variables including the multivariate exposure, potential confounders, and an intercept ($p = r + k + 1$); and the p/n ratio.

Exposure Group	Sample Size (n)	# Agents (m)	Dim($\mathbf{Z}$) (r)	# Potential Confounders (k)	# Indep. Variables (p)	p to n Ratio (p/n)
volatile compounds	179	10	55	82	138	0.77
pcbs	558	22	253	59	313	0.56
phenols	179	3	6	92	99	0.55
dioxins	201	5	15	83	99	0.49
furans - dibenzofuran	201	5	15	83	99	0.49
pest. - pyrethyroid	201	1	1	87	89	0.44
diakyl	225	6	21	76	98	0.44
pest. - phenols	225	4	9	78	88	0.39
pest. - chloroacetanilide	225	1	1	81	83	0.37
pest. - organophosphate	225	1	1	81	83	0.37
heavy metals	444	13	91	48	140	0.32
phthalates	387	11	66	51	118	0.30
pest. - organochlorine	288	7	28	53	82	0.28
nutrients - minerals	158	2	3	37	41	0.26
polyflourochemicals	444	10	55	51	107	0.24
hydrocarbons	292	9	45	22	68	0.23
phytoestrogens	432	6	21	49	71	0.16
nutrients - vitamin C	444	1	1	60	62	0.14
nutrients - carotenoid	1370	5	15	33	49	0.04
nutrients - vitamin A	1370	3	6	35	42	0.03
nutrients - vitamin B	1370	3	6	35	42	0.03
nutrients - vitamin E	1370	2	3	36	40	0.03
cotinine	1370	1	1	37	39	0.03
nutrients - vitamin D	1370	1	1	37	39	0.03

none of the potential confounders. Figure 3 presents the point estimates and 95% posterior intervals.

To highlight the advantage of ACPME we focus on exposure effect estimates for volatile compounds. In this case there are $n = 179$ individuals, $k = 82$ potential confounders, $m = 10$ agents and all 45 pairwise interactions ($\dim(\mathbf{Z}) = 55$). Hence, the p to n ratio is 0.77 (138/177) and a more parsimonious model is preferred. The effect of volatile compounds on HDL and triglyceride both change sign with confounder adjustment using ACPME and with the full model relative to the unadjusted model. In addition, using ACPME resulted about a 30% decrease in posterior standard deviation of $\widehat{\Delta}_{\mathbf{z},\mathbf{z}'}$ compared to the full model for all three outcomes as shown in Figure 4.

In general, the ACPME point estimates are similar to the full model. This suggests that all important confounders are included using ACPME. In contrast, the unadjusted model produces notably different posterior means in several cases suggesting that the unadjusted estimates are confounded. Figure 4 shows that estimates with ACPME had smaller posterior variance on average than estimates from the full model due to decreased model size. Hence, ACPME fully adjusts for observed confounding while decreasing posterior variance.

Figure 3 also shows significance at the 0.05 and 0.01 posterior probability level after Bonferroni adjustment (Web Figure 6 provides additional details on significance). We define significance level as the highest probability symmetric credible interval that does not contain zero. Using the unadjusted model, four groups (carotenoid, vitamin B, vitamin C, and cottoning) had a statistically significant effect on HDL levels and one group (carotenoid) has a statistically significant effect on LDL levels. However, only two–the effect of carotenoid on HDL and LDL levels–are significant after confounder adjustment with either ACPME or the fully adjusted model. Figure 3 shows the point estimates for the other three groups all shrink toward zero when adjusted for confounding. Hence, three of the five are likely false discoveries as a result of confounding.

There are two cases where the posterior probability of a significant exposure effect substantially increased in the more parsimonious ACPME model compared to the full model. Vitamin C was significant for LDL at the 0.01 level using ACPME but only at the 0.05 level with the full model and not at all with the unadjusted model. This highlights a new result not reported in Patel et al. (2012) where the analysis did not control for other exposures. Figure 3 shows that vitamin C is negatively associated with LDL levels and the posterior mean shifts away from 0 after confounder adjustment, while figure Figure 4 show a small decrease in posterior standard deviation using the more parsimonious ACPME compared the full model. ACPME identified several agents as important confounders: two E vitamins (α-tocopherol, γ-tocopherol), two carotenoids (combined lutein/zeaxanthin and thans-lycophene), and folate. In addition to these agents, body mass index, standing height, and weight were also included in the model with high posterior probability. Finally, the effect of vitamin D on HLD levels is significant at the 0.05 level only with ACPME. Combined lutein/zeaxanthin and γ-tocopherol were identified as important to adjust for in this case as well.

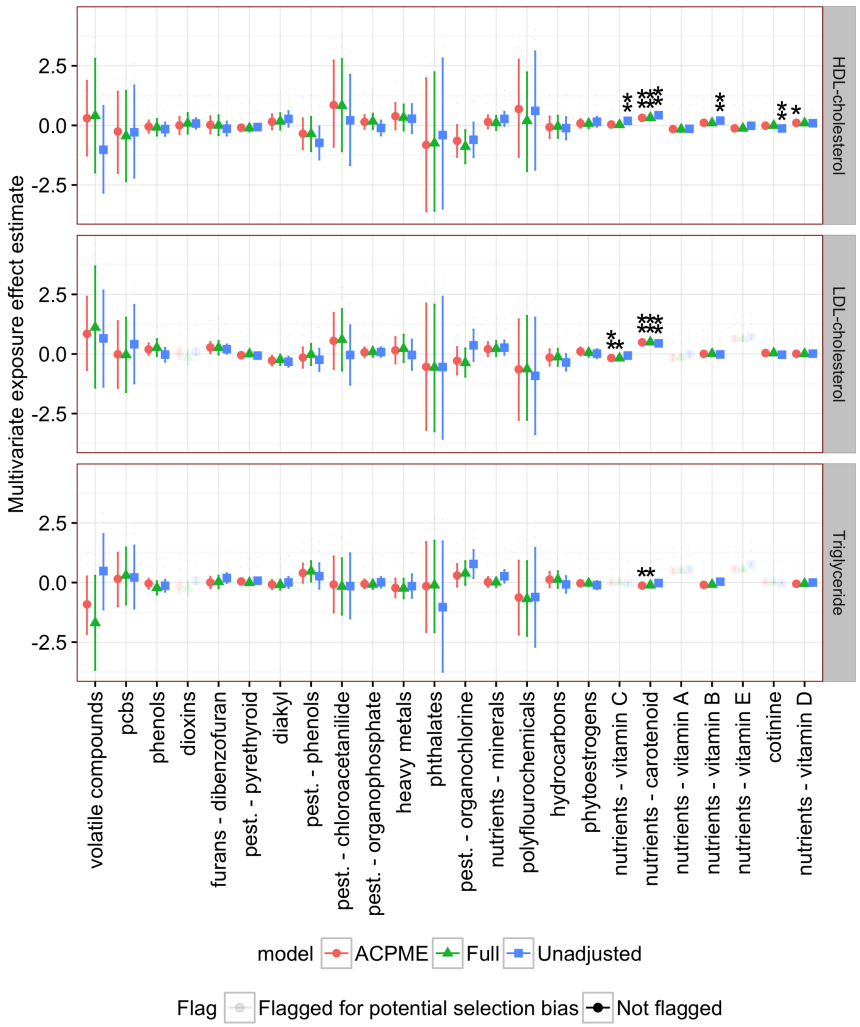


Figure 3: Point estimates and 95% credible intervals of the association between the multivariate exposure $\mathbf{Z}$ and each of the 3 outcome adjusted by the exposure to the other 23 groups and baseline covariates. Results are reported under the ACPME model, the full model ($\boldsymbol{\alpha} = \mathbf{1}$), and the unadjusted model ($\boldsymbol{\alpha} = \mathbf{1}$). Faded estimates were flagged for potentially selection bias. The black asterisks indicate 0.05 (*) and 0.01 (**) significance levels after Bonferroni adjustment for multiple comparisons. This figure appears in color in the electronic version of this article.

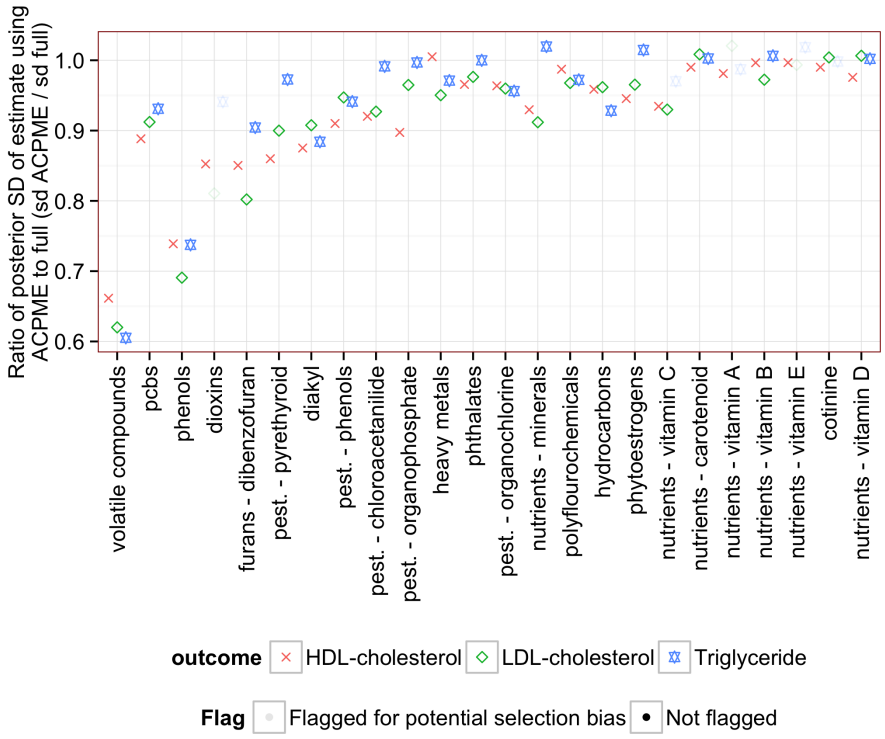


Figure 4: Ratio of the posterior standard deviation of exposure effect estimate under the ACPME model compared to the full model. Results are showed ranked by groups with the largest (left) to the smallest (right) p/n. In general, the estimates from ACPME have lower SD because of the reduced model size. Faded estimates were flagged for potentially selection bias. This figure appears in color in the electronic version of this article.

6 Discussion

In this paper we address the challenge of estimating the health effect of a multivariate exposure when there is uncertainty about which of the many measured covariates are important confounders. In this setting, it is imperative to identify a parsimonious model that is fully adjusted for confounding. We consider the special case of a multivariate exposure that includes several continuous agents and all their pairwise interactions. We estimate the multivariate exposure effect using a regression model and the BMA framework. To adjust for confounding, we develop an informative prior on covariate inclusion. As shown by our simulation study, the proposed method performs better than recently developed alternative methods.

We apply the proposed method to the NHANES data to estimate the effects of exposure to 132 nutrients and persistent pesticides grouped into 24 groups on lipid levels. When p is close to n, our analysis estimates the multivariate exposure effect that is fully adjusted for confounding and has smaller variance than the same estimate obtained under a model that includes all available covariates. As a result, we identified two groups of significant exposures that were not evident without the proposed method: the effects of vitamin C on LDL and vitamin D on HDL. We also identified multivariate exposures that are significant in their association with the outcome absent confounding adjustment, but lose statistical significance after adjustment for confounding.

While we used the formality of potential-outcomes notation to explicitly clarify notions of confounding of the multivariate exposure effect, the methodology proposed for estimation relies on the framework of linear regression. Making inference from a parametric regression model has several advantages in this context, such as permitting the use of model-averaging computations and providing a framework for constructing the proposed prior distribution. Nonetheless, the use of a regression model has inherent limitations for estimating effects that could be improved upon in future work (Vansteelandt, 2012).

The proposed method uses information from the relationship between the potential confounders ($\mathbf{C}$) and the multivariate exposure ($\mathbf{Z}$) to construct the prior. Therefore the ACPME prior does not depend on the outcome and indeed builds on other previous work that uses the covariate space to construct a prior. For example, in Zellner's g-prior the covariance structure of the prior is a function of the observed covariates but the strength of the prior is user-specified (Zellner, 1986). Similarly, the relative ordering of the prior inclusion probabilities on each covariate (C_j) is determined for ACPME by the covariate space ($\mathbf{Z}$ and $\mathbf{C}$) while the strength of the prior is defined by the user through the tuning parameter (λ).

The proposed method offers several advantages over previously developed confounder adjustment methods. First, our approach scales as the dimension of $\mathbf{Z}$ increases. Previously proposed confounder adjustment approaches have relied on exposure modeling (Wang et al., 2012a; Wilson and Reich, 2014; Wang et al., 2015), which require an additional exposure model for each agent. This is impractical when the number if individual agents is large. For example, the PCB

exposure group in our data analysis with 22 agents would require 22 separate exposure models using the approach of Wilson and Reich (2014).

A second advantage of the proposed method is that it explicitly addresses confounding of the multivariate exposure effect. As discussed in Section 2 and Web Appendix A, the confounders of the agents is different from the confounders of the multivariate exposure effect. To our knowledge, the proposed method is the first to explicitly address confounding of a multivariate exposure that includes interaction terms. In addition, the method immediately extends to other functional forms of the multivariate exposure $z(\mathbf{X})$. For example, higher order interactions or basis expansions.

Another advantage is the specific guidance provided for tuning the strength of the prior on covariate inclusion. We give a choice of tuning parameter (λ^*) that can be easily calculated *a priori*, balances model parsimony and confounder adjustment, and performed well in both the simulation and data analysis. In contrast, the method of Wang et al. (2012a) showed sensitivity to choice of tuning parameters, but did not give specific tuning guidance (Vansteelandt, 2012; Wang et al., 2012b) while Wilson and Reich (2014) provide an approach to tuning in the single agent case that relies on refitting several models and post-hoc selecting the best choice for tuning parameter.

A fourth advantage of the proposed method is that it allows covariates that are associated with the exposure but not the outcome to be dropped from the model. In contrast, Wang et al. (2012a) explicitly forces covariates associated with the exposure to be included in the outcome model when the tuning parameter approaches infinity; the choice highlighted in Wang et al. (2012a). These covariates, sometimes called instrumental variables, should not be included in the analysis as they can increase the variance of the estimated exposure effect.

With increased availability of high-dimensional exposure data (called exposome) there is growing interest in understanding the effect of the exposome on complex diseases. However, there is often uncertainty as to which covariates to include in the model to estimate the multivariate exposure effect. The proposed method fills a methodological gap to adjust for confounding when estimating multivariate exposure effects. This method is a valuable tool that can reliably be used in EWAS to estimate health effects of mixtures while simultaneously allowing a rigorous adjustment for confounding and guarantee model parsimony.

7 Supplemental Material

Web Appendices, Tables, and Figures referenced in Sections 2, 3, and 5 are available with this paper at the Biometrics website on Wiley Online Library.

Software for the ACPME method is available in the `R` package `regimes` available at `anderwilson.github.io/regimes/`.

Acknowledgements

Supported in part by NIH grants T32ES007142, P01CA134294, K99ES023504, R21ES025052, R21ES020152, U54HG007963, R21ES022585-01, R01ES019560, R21ES024012, R01GM111339, R01ES024332, R35CA197449, and 1P50MD010428-01. Also supported by HEI 4909, AHRQ HS021991, and PhRMA Foundation. This publication was made possible by USEPA grant RD-83479801. Its contents are solely the responsibility of the grantee and do not necessarily represent the official views of the USEPA. Further, USEPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

Bobb, J. F., Valeri, L., Claus Henn, B., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J., and Coull, B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*, 16(3):493–508.

Crainiceanu, C. M., Dominici, F., and Parmigiani, G. (2008). Adjustment uncertainty in effect estimation. *Biometrika*, 95(3):635–651.

Danaei, G., Pan, A., Hu, F. B., and Hernán, M. A. (2013). Hypothetical Midlife Interventions in Women and Risk of Type 2 Diabetes. *Epidemiology*, 24(1):122–128.

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(468):938–948.

Greenland, S. (2008). Invited commentary: variable selection versus shrinkage in the control of multiple confounders. *American Journal of Epidemiology*, 167(5):523–529.

Hernán, M. a., Brumback, B., and Robins, J. M. (2001). Marginal Structural Models to Estimate the Joint Causal Effect of Nonrandomized Treatments. *Journal of the American Statistical Association*, 96(454):440–448.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14(4):382–417.

Imai, K. and van Dyk, D. A. (2004). Causal Inference With General Treatment Regimes. *Journal of the American Statistical Association*, 99(467):854–866.

Louis, G. M. B. and Sundaram, R. (2012). Exposome: Time for transformative research. *Statistics in Medicine*, 31(22):2569–2575.

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(19):2937–2960.

Patel, C. J., Cullen, M. R., Ioannidis, J. P. A., and Butte, A. J. (2012). Systematic evaluation of environmental factors: Persistent pollutants and nutrients correlated with serum lipid levels. *International Journal of Epidemiology*, 41(3):828–843.

Patel, C. J. and Ioannidis, J. P. A. (2014). Studying the elusive environment in large scale. *JAMA*, 311(21):2173–4.

Patel, C. J., Rehkopf, D. H., Leppert, J. T., Bortz, W. M., Cullen, M. R., Chertow, G. M., and Ioannidis, J. P. A. (2013). Systematic evaluation of environmental and behavioural factors associated with all-cause mortality in the united states national health and nutrition examination survey. *International Journal of Epidemiology*, 42(6):1795–1810.

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*, 169(2):179–203.

Raftery, A. E., Madigan, D., and Hoeting, J. A. (1997). Bayesian model averaging for linear regression models. *Journal of the American Statistical Association*, 92(437):179.

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(427):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(1):41.

Rubin, D. B. (1978). Bayesian Inference for Causal Effects: The Role of Randomization. *The Annals of Statistics*, 6:34–58.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Taubman, S. L., Robins, J. M., Mittleman, M. A., and Hernán, M. A. (2009). Intervening on risk factors for coronary heart disease : an application of the parametric g-formula. *International Journal of Epidemiology*, 38(April):1599–1611.

Vansteelandt, S. (2012). Discussions. *Biometrics*, 68(3):175–678.

Vansteelandt, S., Bekaert, M., and Claeskens, G. (2012). On model selection and model misspecification in causal inference. *Statistical Methods in Medical Research*, 21(1):7–30.

Verschuren, L., Wielinga, P. Y., van Duyvenvoorde, W., Tijani, S., Toet, K., van Ommen, B., Kooistra, T., and Kleemann, R. (2011). A dietary mixture containing fish oil, resveratrol, lycopene, catechins, and vitamins E and C reduces atherosclerosis in transgenic mice. *The Journal of Nutrition*, 141(5):863–869.

Wang, C., Dominici, F., Parmigiani, G., and Zigler, C. M. (2015). Accounting for uncertainty in confounder and effect modifier selection when estimating average causal effects in generalized linear models. *Biometrics*, 71(3):654–665.

Wang, C., Parmigiani, G., and Dominici, F. (2012a). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–71.

Wang, C., Parmigiani, G., and Dominici, F. (2012b). Rejoinder: Bayesian Effect Estimation Accounting for Adjustment Uncertainty. *Biometrics*, 68(3):680–686.

Wild, C. P. (2005). Complementing the genome with an "exposome": The outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiology Biomarkers and Prevention*, 14(8):1847–1850.

Wilson, A., Rappold, A. G., Neas, L. M., and Reich, B. J. (2014a). Modeling the effect of temperature on ozone-related mortality. *The Annals of Applied Statistics*, 8(3):1728–1749.

Wilson, A. and Reich, B. J. (2014). Confounder selection via penalized credible regions. *Biometrics*, 70(4):852–861.

Wilson, A., Reif, D. M., and Reich, B. J. (2014b). Hierarchical dose-response modeling for high-throughput toxicity screening of environmental chemicals. *Biometrics*, 70(1):237–46.

Zanobetti, A., Austin, E., Coull, B. A., Schwartz, J., and Koutrakis, P. (2014). Health effects of multi-pollutant profiles. *Environment International*, 71:13–19.

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distributions. *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, 6:233–243.

Zigler, C. M. and Dominici, F. (2014). Uncertainty in Propensity Score Estimation: Bayesian Methods for Variable Selection and Model-Averaged Causal Effects. *Journal of the American Statistical Association*, 109(505):95–107.