Web-based Supplementary Materials for: Model-Averaged Confounder Adjustment for Estimating Multivariate Exposure Effects with Linear Regression by Ander Wilson, Corwin M. Zigler, Chirag J. Patel, and Francesca Dominici

Ander Wilson[1,*], Corwin M. Zigler[2], Chirag J. Patel[3], Francesca Dominici[2]

[1]Department of Statistics, Colorado State University

[2]Department of Biostatistics, Harvard T. H. Chan School of Public Health

[3] Department of Biomedical Informatics, Harvard Medical School

A. Exposure to a single agent versus multivariate exposure

In this section we illustrate that we cannot always identify the confounders of the effect of a multivariate exposure (the effect of a change in $\boldsymbol{Z}$ on outcome Y) by taking the union of the confounders of the effect of exposure to each single agent (e.g. X_1 and X_2) on Y.

A covariate C_1 can be a confounder of the effect of $\boldsymbol{Z}$ on Y but not a confounder of the effect of any single agent (e.g. X_1 or X_2) on Y, if C_1 is correlated with a function of agents (e.g. the interaction X_1X_2) but C_1 is not correlated with X_1 nor with X_2. More generally, this can occur when the covariate is balanced across levels of each individual agent but not balanced across levels of combinations of agents.

Confounder adjustment methods that rely on exposure modeling with a linear exposure model (e.g. Crainiceanu et al., 2008; Wang et al., 2012; Wilson and Reich, 2014; Wang et al., 2015) do not address this situation. These methods identify potential confounders by specifying an exposure model for each agent (one for X_1 and one for X_2), where the single agent is the dependent variable and all the potential confounders are the independent variables. Hence, potential confounders associated with the interaction terms only, but not with the main effect, will not be identified as confounders in these single agent exposure models.

To demonstrate this phenomenon, we construct a simple hypothetical example where we can calculate the closed-form confounding bias. We assume there are two agents $\boldsymbol{X} = (X_1, X_2)$, a multivariate exposure that includes an interaction $\boldsymbol{Z} = (X_1, X_2, X_1X_2)$, and a single covariate C_1. We assume that C_1 and X_1 are both distributed independent $\mathrm{N}(0, 1)$. As is common in environmental epidemiology X_1 and C_1 might affect the level of the other exposure X_2. We assume X_2 is $\mathrm{N}(X_1C_1, 1)$. Finally, the outcome Y is distributed $\mathrm{N}(X_1 +$

$X_2 + X_1X_2 + C_1, 1)$. In this case C_1 is linearly associated with X_1X_2 (covariance 1) and with Y, but not with X_1 or X_2 (because C_1 and X_1 are independent and centered on 0). We also assume that we are interested in estimating Δ, here defined as the effect of a change from no exposure to one unit of both agents (X_1 and X_2).

Because C_1 is not linearly associated with X_1 or X_2, any approach that relies on exposure modeling with a linear association between C_1 and X_1 and X_2 with two separate exposures models (e.g. Wang et al., 2012; Wilson and Reich, 2014; Wang et al., 2015) will not identify C_1 as a confounder and, therefore, will fail to identify the necessary confounders of the effect of $\boldsymbol{Z}$ on Y. Further, the covariate C_1 is balanced across levels of X_1 and X_2 but not across levels of the interaction X_1X_2 as shown in Web Appendix Figure 1. Under model (4), the truth is $\Delta = 3$. However, the model without C_1 has $E[\Delta|\boldsymbol{X}] = 3.25$ and is biased whereas the $E[\Delta|\boldsymbol{X}, C_1] = 3$ and is unbiased. Hence, the union of confounders identified by the collection of single agent exposure models does not include the full set of confounders for the association between the multivariate exposure ($\boldsymbol{Z}$) and the outcome (Y). In summary, new methods are needed for confounder adjustment in the multiple exposure setting that adapt to the multivariate exposure effect being estimated.

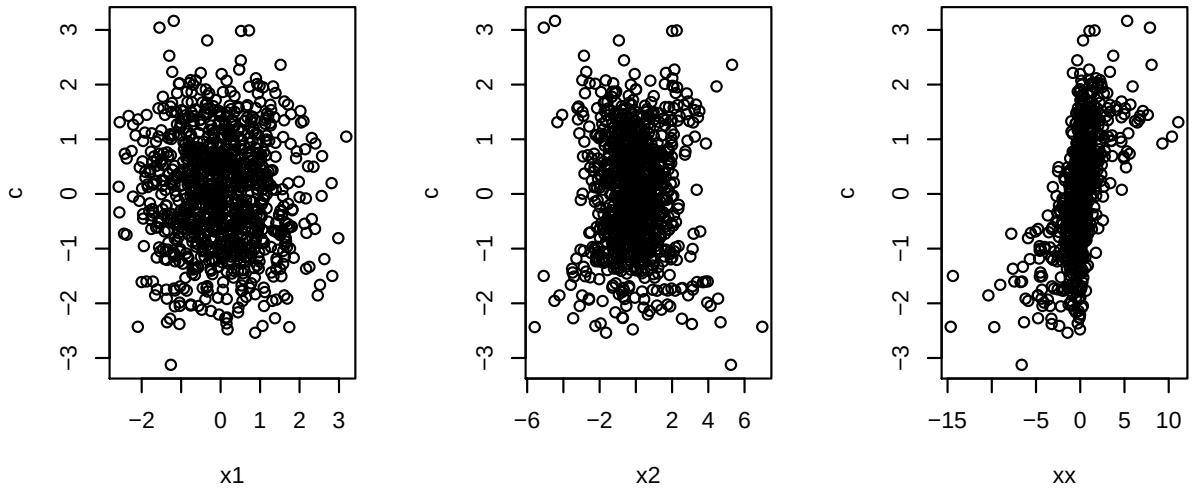


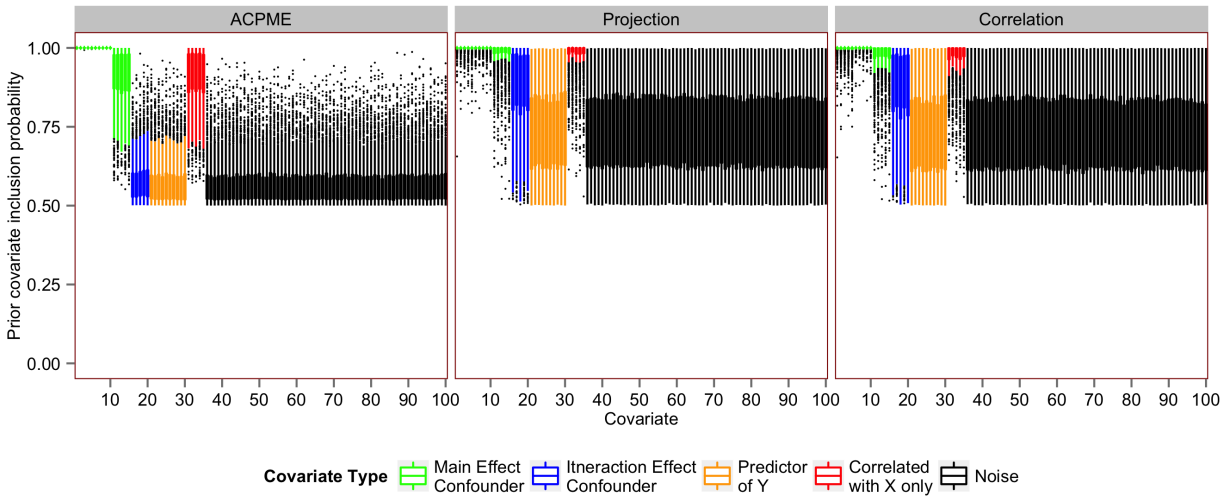
Figure 1: Plots showing balance of the covariate (y-axis) across levels X_1, X_2, and X_1X_2 (x-axis shown on panels from left to right). The data is for one simulated dataset of sample size 1000.

B. Additional details for model specification section

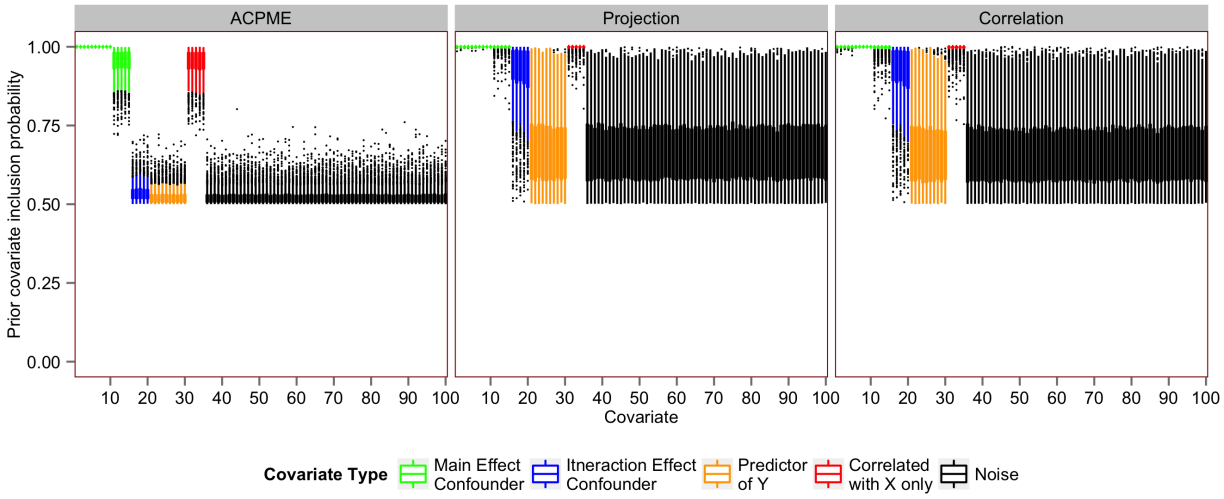
B.1 Alternative choices for prior odds

Any non-negative function that captures the linear association between $\boldsymbol{C}$ and $\boldsymbol{Z}$ can be used as $\omega_j(\boldsymbol{Z}, \boldsymbol{C})$. For example, $\omega_j(\boldsymbol{Z}, \boldsymbol{C}) = \sum_{k=1}^{r} \text{cor}(\boldsymbol{C}_j, \boldsymbol{Z}_k)^2$ or $\boldsymbol{Z}$, $\omega_j(\boldsymbol{Z}, \boldsymbol{C}) = \boldsymbol{C}_j^T \boldsymbol{P}_Z \boldsymbol{C}_j / \boldsymbol{C}_j^T \boldsymbol{C}_j$.

Both of these give $\omega_j(\boldsymbol{Z}, \boldsymbol{C}) = 0$ if C_j is linearly independent of $\boldsymbol{Z}$ (not a confounder) and positive if they are linearly dependent (potential confounders). However, these options tend to be identify potential confounders with less specificity at smaller sample sizes. As a result they assign increased prior inclusion probabilities to covariates that are not true confounders. Web Appendix Figure 2 compares these priors for the alternative formulations using the simulated data.



(a) $n = 200$



(b) $n = 500$

Figure 2: Comparison of ACPME prior to alternative priors discussed in Web Appendix B.1. The y-axis displayed the prior inclusion probability for the 1000 simulated datasets in scenario one. The top panel is for sample size $n = 200$ and the bottom for $n = 500$. The prior inclusion probability under ACPME is lower for the covariates that are not associated with the exposure which results in lower posterior false inclusion rates.

B.2 Details for bias calculation in (6)

This section contains additional details on the calculation in (6) of Section 3.2 of the main text. The estimated exposure effect of interest can be biased if important confounders are omitted from the model. With flat priors on $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ and inverse-gamma priors on σ^2 the posterior mean of $(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T$ is the ordinary least squares solution $E\{(\boldsymbol{\beta}^T, \boldsymbol{\eta}^T)^T | \boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{C}\} = \{(\boldsymbol{Z}, \boldsymbol{C})^T(\boldsymbol{Z}, \boldsymbol{C})\}^{-1}(\boldsymbol{Z}, \boldsymbol{C})^T\boldsymbol{Y}$, where $(\boldsymbol{Z}, \boldsymbol{C})$ is the full design matrix. The unadjusted model that include only the exposures but no covariates has multivariate-t posterior with mean $E(\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}) = (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Y}$. The difference between the posterior means for the unadjusted and the fully adjusted model is

$$
\begin{aligned}
&E\left(\boldsymbol{d}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}\right) - E\left(\boldsymbol{d}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{C}\right) \\
&\quad = \boldsymbol{d}^T(\boldsymbol{X}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Y} - (\boldsymbol{d}^T, \boldsymbol{0}^T)\{(\boldsymbol{Z}, \boldsymbol{C})^T(\boldsymbol{Z}, \boldsymbol{C})\}^{-1}(\boldsymbol{Z}, \boldsymbol{C})^T\boldsymbol{Y} \\
&\quad = \boldsymbol{d}^T(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\boldsymbol{Z}^T\boldsymbol{Y} - \boldsymbol{d}^T\left[\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{Y} - \left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}\right] \\
&\quad = \boldsymbol{d}^T\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}, \qquad (1)
\end{aligned}
$$

where $\boldsymbol{P}_Z$ is the perpendicular projection onto the column space of $\boldsymbol{Z}$ and $\boldsymbol{P}_Z^{\perp} = \boldsymbol{I} - \boldsymbol{P}_Z$. This difference is the confounding bias in the exposure effect estimate caused by excluding all covariates.

B.3 Details for calculation of (7)

This section contains additional details on the calculation in (7). We assume $\boldsymbol{d} = \boldsymbol{Z}\boldsymbol{a}$ for some vector $\boldsymbol{a}$. It follows that there is no confounding bias if the inner product of $\|E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}\right) - E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{C}\right)\|_2^2 = 0$. Using (1) in the Web Appendix this quantity can be rewritten as

$$
\begin{aligned}
&\left\|E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{X}\right) - E\left(\boldsymbol{Z}^T\boldsymbol{\beta}|\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{C}\right)\right\|_2^2 \\
&\quad = \left\|\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}\right\|_2^2 \\
&\quad = \boldsymbol{Y}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y} \\
&\quad = \sum_{l=1}^{k}\zeta_l\boldsymbol{Y}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{q}_l\boldsymbol{q}_l^T\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y} \\
&\quad = \sum_{l=1}^{r}\zeta_l\boldsymbol{Y}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{q}_l\boldsymbol{q}_l^T\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y} \\
&\quad = \sum_{l=1}^{r}\left[\zeta_l^{-1/2}\boldsymbol{q}_k l^T\left(\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{C}\right)^{-1}\boldsymbol{C}^T\boldsymbol{P}_Z^{\perp}\boldsymbol{Y}\right]^2 4 \qquad (2)
\end{aligned}
$$

where $\boldsymbol{C}^T\boldsymbol{P}_Z\boldsymbol{C} = \sum_{l=1}^{k}\zeta_l\boldsymbol{q}_l\boldsymbol{q}_l^T$ is the spectral decomposition. Assuming $\boldsymbol{Z}$ is rank $r \leq k$ then $\boldsymbol{C}^T\boldsymbol{P}_Z\boldsymbol{C}$ is rank r and the equality between the fourth and fifth lines of (2) is valid.

B.4 Selection of prior distributions for other parameters and computation

We complete the Bayesian specification for the normal linear model following that of Raftery et al. (1997). The prior for the regression coefficients is $(\boldsymbol{\beta}, \boldsymbol{\eta})^T \sim \mathrm{N}(\boldsymbol{\mu}_0, \sigma^2\phi^2\boldsymbol{\Sigma}_0)$ while the prior precision is $\sigma^{-2} \sim \mathrm{Gamma}(\nu/2, \kappa\nu/2)$. The hyper-parameters $\boldsymbol{\mu}_0$, ϕ, $\boldsymbol{\Sigma}_0$, ν, and κ are as described in Raftery et al. (1997). The posterior is approximated with the MC^3 method of Madigan et al. (1995). Assuming the outcome, exposures, and covariates are standardized $\boldsymbol{\mu}_0 = \mathbf{0}$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{I}$.

C. Additional Material for the Simulation Study

C.1 Additional results for simulation scenario one

Web Appendix Figure 3 shows the correlation structure in the data.

Web Appendix Table 1 shows additional results for simulation scenario one under the null and reduced correlation between the covariates and the outcome. In this case, $\boldsymbol{\beta} = \mathbf{0}$. The effect of the covariates $\{\eta_j\}_{j=1}^{30}$ are simulated from a uniform$(0.1, 0.2)$ distribution. The relationship between $\mathbf{C}$ and $\mathbf{Z}$ remains the same as in scenario 1 from the main text.

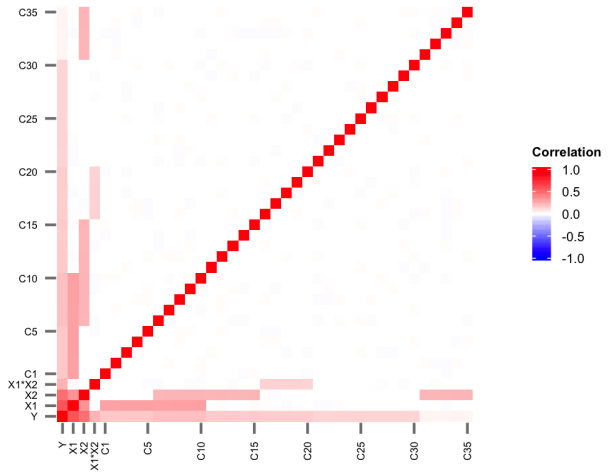
C.2 Additional results for simulation scenario two

For simulation scenario two we simulate the data as

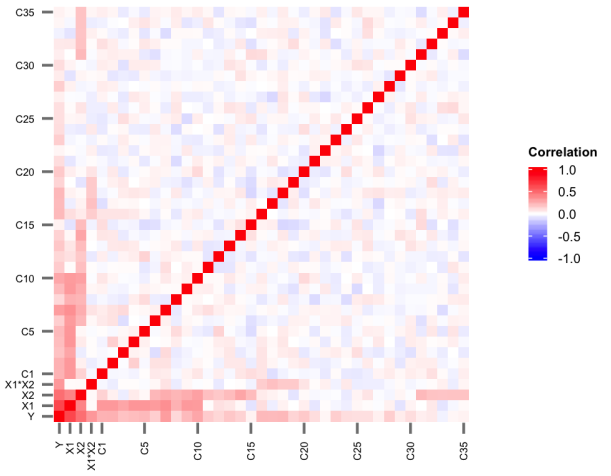
$$
\begin{aligned}
C_{ji} &\sim \mathrm{N}(0,1) \quad \text{for } j = 1, \ldots, 100 \\
h_j &\sim \mathrm{DiscreteUniform}(1, \ldots, m_1) \quad \text{for } j = 1, \ldots, 10 \\
h_j &\sim \mathrm{DiscreteUniform}(1, \ldots, m) \quad \text{for } j = 11, \ldots, 25 \\
q_j &\sim \mathrm{DiscreteUniform}(m_1 + 1, \ldots, m_2) \quad \text{for } j = 1, \ldots, 15 \\
X^*_{ki} &\sim \mathrm{N}\left(\sum_{j=1}^{10} C_{ji}\mathbb{1}\{h_j = k\}X_{q_j i} + \sum_{j=11}^{25} C_{ji}\mathbb{1}\{h_j = k\}, 1\right) \\
Y_i &\sim \mathrm{N}\left(\boldsymbol{Z}\boldsymbol{\beta} + \sum_{j=1}^{30} \eta_j \boldsymbol{C}_{ji}, 1\right),
\end{aligned} \tag{3}
$$

where $m_1 = m/2$ rounded down to the nearest integer, $m_2 = m - m_1$, and, X_{ki} is X^*_{ki} scaled to have variance 1. Finally, $\{\beta_l\}_{l=1}^{r}$ and $\{\eta_j\}_{j=1}^{30}$, are independent Uniform$(0.2, 0.5)$. Web Appendix Figure 4 shows the correlation structure in the data.

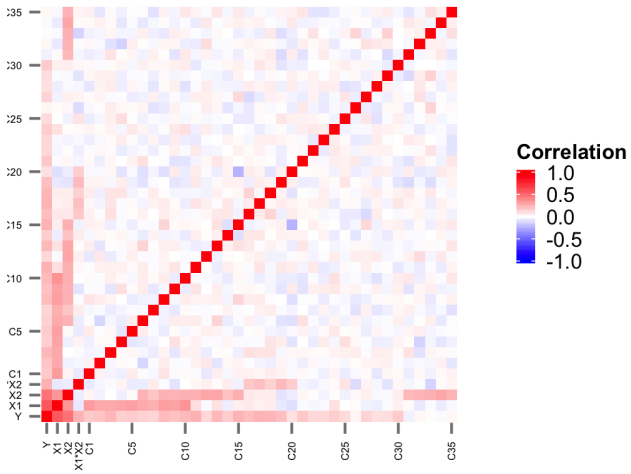
Web Appendix Table 2 and Web Appendix Table 3 shows results for simulation scenario 2 for $n = 200$ and $n = 500$ respectively. These results are also presented in Figure 2 of the main text.



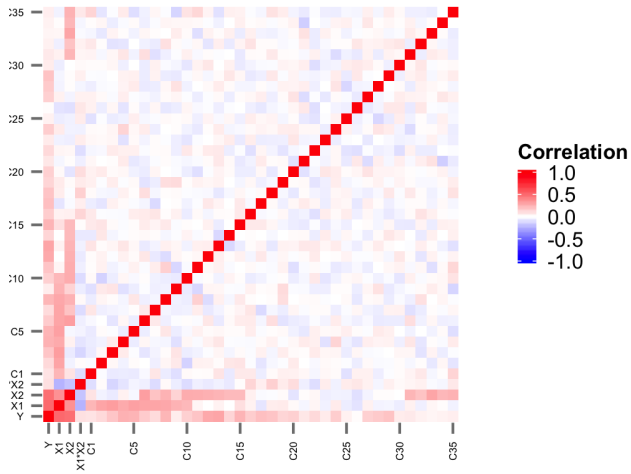
(a) Average correlation structure



(b) Correlation for simulated data set 1



(c) Correlation for simulated data set 2



(d) Correlation for simulated data set 3

Figure 3: Visualization of the structure for simulated data under scenario 1. Panel 3a shows the average correlation structure across 1000 datasets for the outcome Y, the multivariate exposures $\boldsymbol{Z} = (X_1, X_2, X_1X_2)$, and the first 35 covariates. The remaining 65 covariates are iid $\mathrm{N}(0, 1)$. Panels 3b – 3d show the covariance structure for the first three simulated data sets of sample size 500.

D. Additional Material for Data Analysis

D.1 Data preparation detail

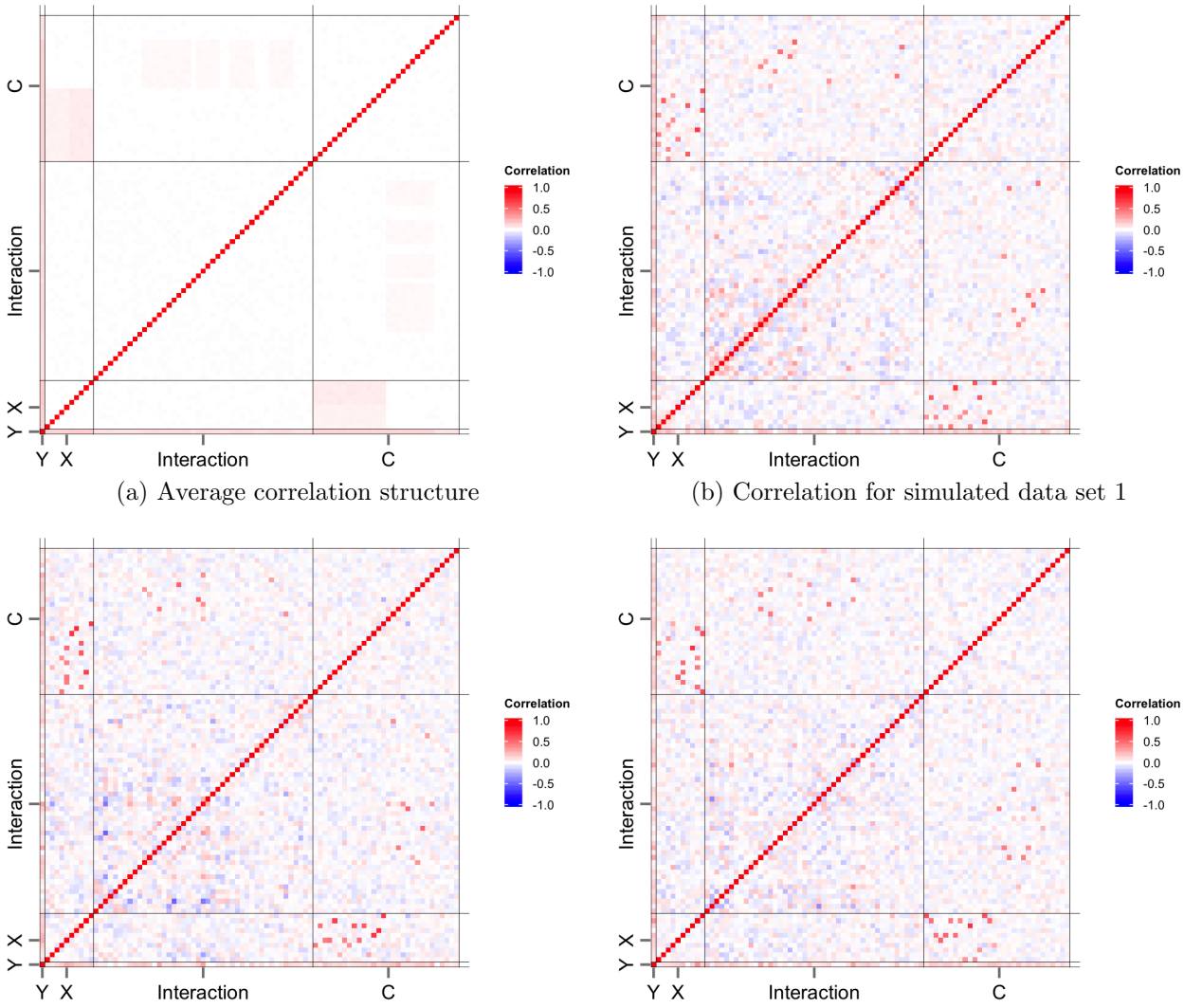
The process of limiting the data is as follows. First, for a given exposure group we limit the data to all subjects that have complete data for all agents in the exposure group and the covariates. Second, we identify other exposure groups that are complete for at least 1/3 of the sample identified in the first step. If this sample is at least 150 persons we use this sample. If the sample was less than 150 persons, we identify other exposure groups that are

Table 1: Simulation results for the alternative simulation scenario 1. This is the same as scenario one in the text but the true effect of $\boldsymbol{Z}$ on Y is null and there is reduced correlation between $\boldsymbol{C}$ and Y. 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 considered 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	0.16	0.39	0.33	0.90	0.62	0.12
BayesPen	0.28	0.45	0.19	0.49	0.62	0.14
BMA	0.63	0.65	0.14	0.01	0.21	0.07
Unadjusted	0.71	0.72	0.13	0.00	0.00	0.00
Full	0.03	0.45	0.43	0.94	1.00	1.00
True	0.01	0.30	0.29	0.94	1.00	0.00
$n = 500$						
ACPME	0.06	0.22	0.20	0.93	0.79	0.08
BayesPen	0.05	0.24	0.16	0.78	0.91	0.13
BMA	0.57	0.58	0.09	0.00	0.38	0.04
Unadjusted	0.70	0.71	0.08	0.00	0.00	0.00
Full	0.00	0.22	0.21	0.94	1.00	1.00
True	0.00	0.18	0.17	0.95	1.00	0.00

complete for at least 1/2 of the sample identified in 1 (a smaller group of covariates). We use this group if it has a sample of at least 150. Otherwise we do not include any exposure groups as covariates. Web Appendix Figure 5 illustrates covariate inclusion for the data analysis.

Our approach to limiting the data could result in selection bias. To address this we regressed the outcomes on each multivariate exposure without including any potential covariates in both the full dataset and the data used for analysis. Web Appendix Figure 6 shows the z-scores of the posterior means for the full sample using the posterior mean and standard deviation for the subsample used for analysis. Using z-scores to screen for selection bias we identified 8 exposure-outcome combinations that may be effected by selection bias. In these cases the z-scores exceeded the threshold of a 0.05 α-level two-sided test after Bonferroni corrections. These estimates are presented as faded or omitted in the analysis figures and are not discussed in the results. Web Appendix Figure 6 looks at potential selection bias caused by selecting a subsample with multiple exposures for analysis.



(a) Average correlation structure

(b) Correlation for simulated data set 1

(c) Correlation for simulated data set 2

(d) Correlation for simulated data set 3

Figure 4: Visualization of the structure for simulated data under scenario 2. Panel 4a shows the average correlation structure across 1000 datasets for the outcome Y, $m = 10$ agents, all pairwise interactions between the agents, and the first 30 covariates. The remaining 70 covariates are iid $\mathrm{N}(0, 1)$. Panels 4b – 4d show the covariance structure for the first three simulated data sets of sample size 500.

D.2 Potential confounders

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

Table 2: Simulation results for simulation scenario 2 and $n = 200$. The first two columns show the number of agents (m) and the dimension of the multivariate exposure r. The next 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 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.

Method	m	r	Bias	Mean RMSE	95% Int. SD / SE	True Inc. Coverage	False Sel. Rate	Rate
$n = 200$								
Full	2	3	-0.01	0.34	0.33	0.95	1.00	1.00
Full	5	15	-0.03	0.55	0.55	0.95	1.00	1.00
Full	10	55	0.12	1.36	1.38	0.95	1.00	1.00
Full	13	91	-0.22	4.12	4.52	0.95	1.00	1.00
True	2	3	-0.01	0.26	0.25	0.95	1.00	0.00
True	5	15	-0.01	0.41	0.40	0.94	1.00	0.00
True	10	55	0.05	0.85	0.85	0.94	1.00	0.00
True	13	91	0.00	1.29	1.35	0.96	1.00	0.00
BMA	2	3	0.87	0.92	0.18	0.06	0.58	0.03
BMA	5	15	1.45	1.58	0.45	0.18	0.53	0.03
BMA	10	55	1.34	1.81	0.94	0.66	0.58	0.07
BMA	13	91	0.42	2.64	1.10	0.62	0.68	0.34
Unadjusted	2	3	1.18	1.24	0.25	0.04	0.00	0.00
Unadjusted	5	15	2.83	2.89	0.54	0.00	0.00	0.00
Unadjusted	10	55	3.81	4.03	1.31	0.17	0.00	0.00
Unadjusted	13	91	4.12	4.59	2.06	0.48	0.00	0.00
ACPME	2	3	0.02	0.29	0.28	0.95	0.87	0.05
ACPME	5	15	0.08	0.45	0.43	0.94	0.91	0.09
ACPME	10	55	0.16	1.06	0.88	0.90	0.92	0.28
ACPME	13	91	-0.13	3.57	1.42	0.58	0.88	0.70
BayesPen	2	3	0.06	0.34	0.23	0.82	0.95	0.17
BayesPen	5	15	0.14	0.51	0.39	0.86	0.94	0.19
BayesPen	10	55	0.35	1.13	0.80	0.84	0.92	0.30
BayesPen	13	91	0.18	2.57	1.32	0.71	0.92	0.59

education: less than high school, high school, or more than high school).

D.3 Additional figures

Web Appendix Figure 7 presents the posterior probability that each exposure group results in a change (either positive or negative) in lipid level on negative log scale. We define the posterior probability of a change as the highest probability symmetric credible interval that does not contain zero. The horizontal lines indicate 0.05 and 0.01 significance levels after Bonferroni adjustment for multiple comparisons.

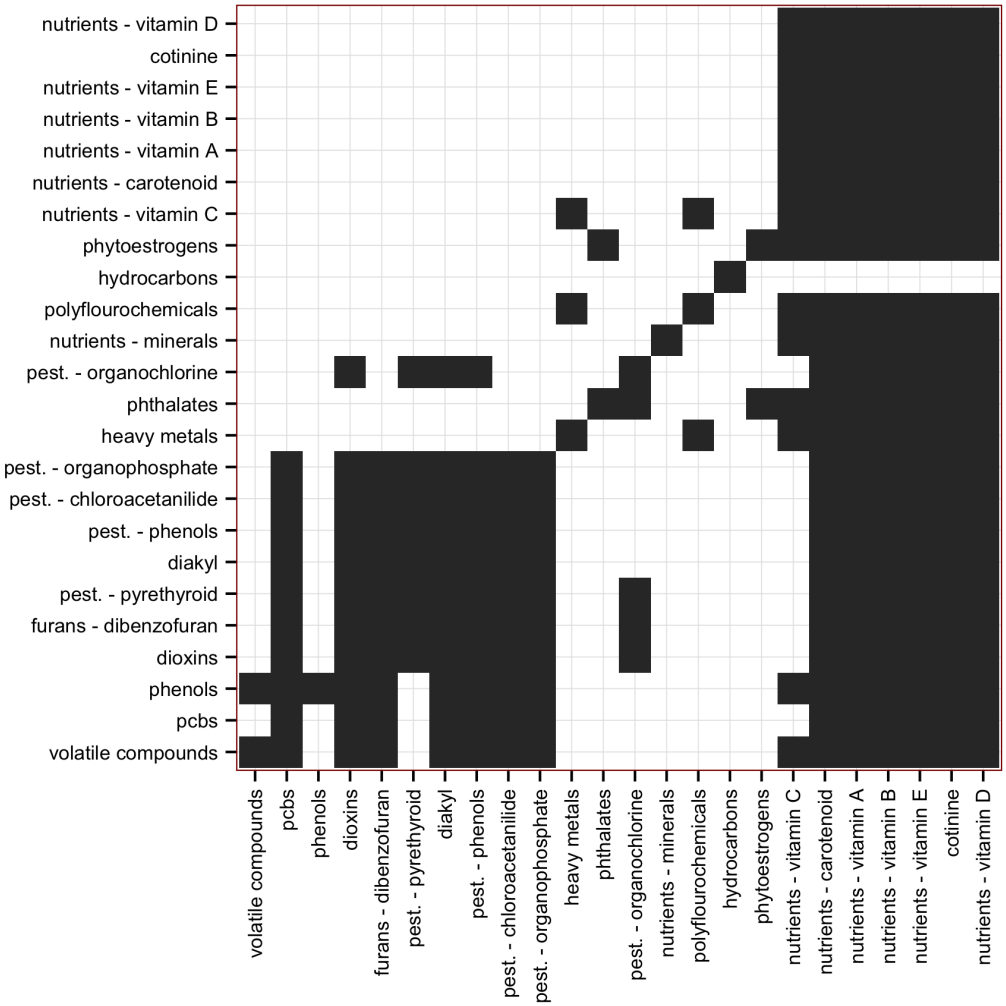


Figure 5: Covariate inclusion for each exposure group. The y-axis show the exposure group while the x-axis shows groups included as potential confounders in the analysis. Shaded cells indicate that a covariate group is included for the analysis of an exposure group.

D.4 Model diagnostics

When using a parametric model, unbalanced covariates across levels of the exposure can exacerbate sensitivity to model misspecification. We performed a series of diagnostics to assess as whether these model assumptions are reasonable. We found no evidence of model misspecification.

For each unique multivariate exposure group and outcome we refit the model using ordinary least squares (OLS) and included as covariates all C_j that has posterior model inclusion probability of at least 0.5.

Web Appendix Figures 8 and 9 show QQ-plots for the data analysis to assess the normality assumption on the residuals. There were no significant deviations from the theoretical normal quantiles.

Web Appendix Figures 10 and 11 show plots of the standardized residuals verse fitted values to assess misspecification of the regression model. While there are a small number of outliers there are no trends in the plots that indicate heteroskedasticity or misspecification of the linear predictors.

Finally, Web Appendix Figures 12 and 13 show smoothed plots of the standardized residuals verse the covariates C_j. This includes all continuous covariates regardless of the posterior model inclusions probability. This will detect if any covariates should have been included as a nonlienar term (e.g. quadratic). For simplicity of presentation and to better assess any trends we show only the smoothed trend of the residuals as a function of observed covariates. To ease comparisons, the y-axis scale is the same as in Web Appendix Figures 10 and 11. There are no notable trends in the figures to suggest misspecification.

References

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

Madigan, D., York, J., and Allard, D. (1995). Bayesian graphical models for discrete data. *International Statistical Review*, 63(2):215–232.

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.

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. (2012). Bayesian effect estimation accounting for adjustment uncertainty. *Biometrics*, 68(3):661–71.

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

Table 3: Simulation results for simulation scenario 2 and $n = 500$. The first two columns show the number of agents (m) and the dimension of the multivariate exposure r. The next 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 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.

Method	m	r	Bias	Mean RMSE	95% Int. SD / SE	True Inc. Coverage	False Sel. Rate	Rate
$n = 200$								
Full	2	3	-0.01	0.16	0.16	0.95	1.00	1.00
Full	5	15	-0.01	0.24	0.25	0.96	1.00	1.00
Full	10	55	-0.02	0.45	0.45	0.95	1.00	1.00
Full	15	120	0.01	0.75	0.73	0.94	1.00	1.00
Full	20	210	0.02	1.22	1.20	0.94	1.00	1.00
Full	25	325	0.03	2.36	2.44	0.96	1.00	1.00
True	2	3	-0.01	0.15	0.15	0.95	1.00	0.00
True	5	15	-0.01	0.22	0.23	0.97	1.00	0.00
True	6	21	-0.01	0.26	0.26	0.95	1.00	0.00
True	10	55	-0.02	0.42	0.41	0.95	1.00	0.00
True	15	120	0.01	0.66	0.65	0.94	1.00	0.00
True	20	210	0.01	1.05	1.02	0.94	1.00	0.00
True	25	325	0.07	1.69	1.74	0.96	1.00	0.00
BMA	2	3	0.31	0.49	0.15	0.55	0.91	0.04
BMA	5	15	0.32	0.52	0.25	0.69	0.91	0.02
BMA	10	55	0.21	0.54	0.43	0.89	0.93	0.02
BMA	15	120	0.24	0.77	0.67	0.91	0.93	0.03
BMA	20	210	0.25	1.16	1.03	0.92	0.92	0.05
BMA	25	325	0.27	2.07	1.67	0.89	0.88	0.20
Unadjusted	2	3	1.14	1.16	0.15	0.01	0.00	0.00
Unadjusted	5	15	2.81	2.83	0.32	0.00	0.00	0.00
Unadjusted	10	55	3.64	3.71	0.69	0.00	0.00	0.00
Unadjusted	15	120	4.10	4.26	1.14	0.06	0.00	0.00
Unadjusted	20	210	4.39	4.74	1.79	0.31	0.00	0.00
Unadjusted	25	325	4.71	5.55	2.96	0.63	0.00	0.00
ACPME	2	3	-0.01	0.15	0.15	0.95	0.99	0.05
ACPME	5	15	0.00	0.23	0.23	0.97	0.99	0.03
ACPME	10	55	-0.01	0.42	0.41	0.95	0.99	0.05
ACPME	15	120	0.02	0.68	0.66	0.94	1.00	0.09
ACPME	20	210	0.02	1.11	1.03	0.93	0.99	0.18
ACPME	25	325	0.03	2.15	1.79	0.90	0.98	0.46
BayesPen	2	3	-0.01	0.16	0.15	0.94	1.00	0.07
BayesPen	5	15	0.00	0.23	0.23	0.95	1.00	0.09
BayesPen	10	55	0.00	0.43	0.41	0.94	1.00	0.11
BayesPen	15	120	0.05	0.71	0.64	0.92	0.99	0.15
BayesPen	20	210	0.09	1.14	1.00	0.92	0.99	0.20
BayesPen	25	325	0.24	2.08	1.68	0.89	0.96	0.35

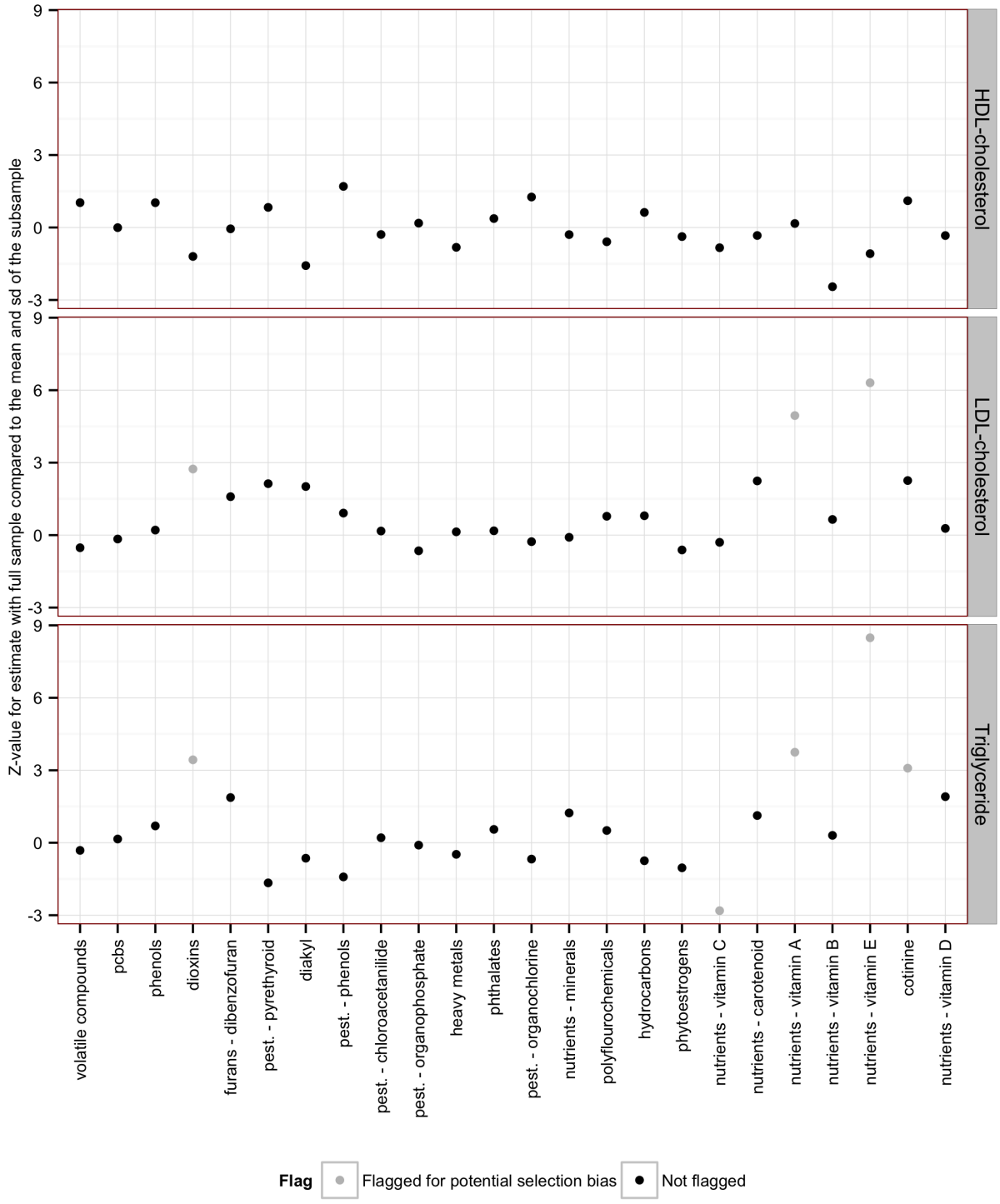


Figure 6: Comparison of the unadjusted estimate with the full sample and with the subsample used for analysis. Each point is the z-score for the posterior mean of the full estimate using the posterior mean and standard deviation of the subsample used. Grey points indicate that the z-score for the full estimate exceeded the threshold of a two-sided test at the 0.05 probability level after Bonferroni corrections.

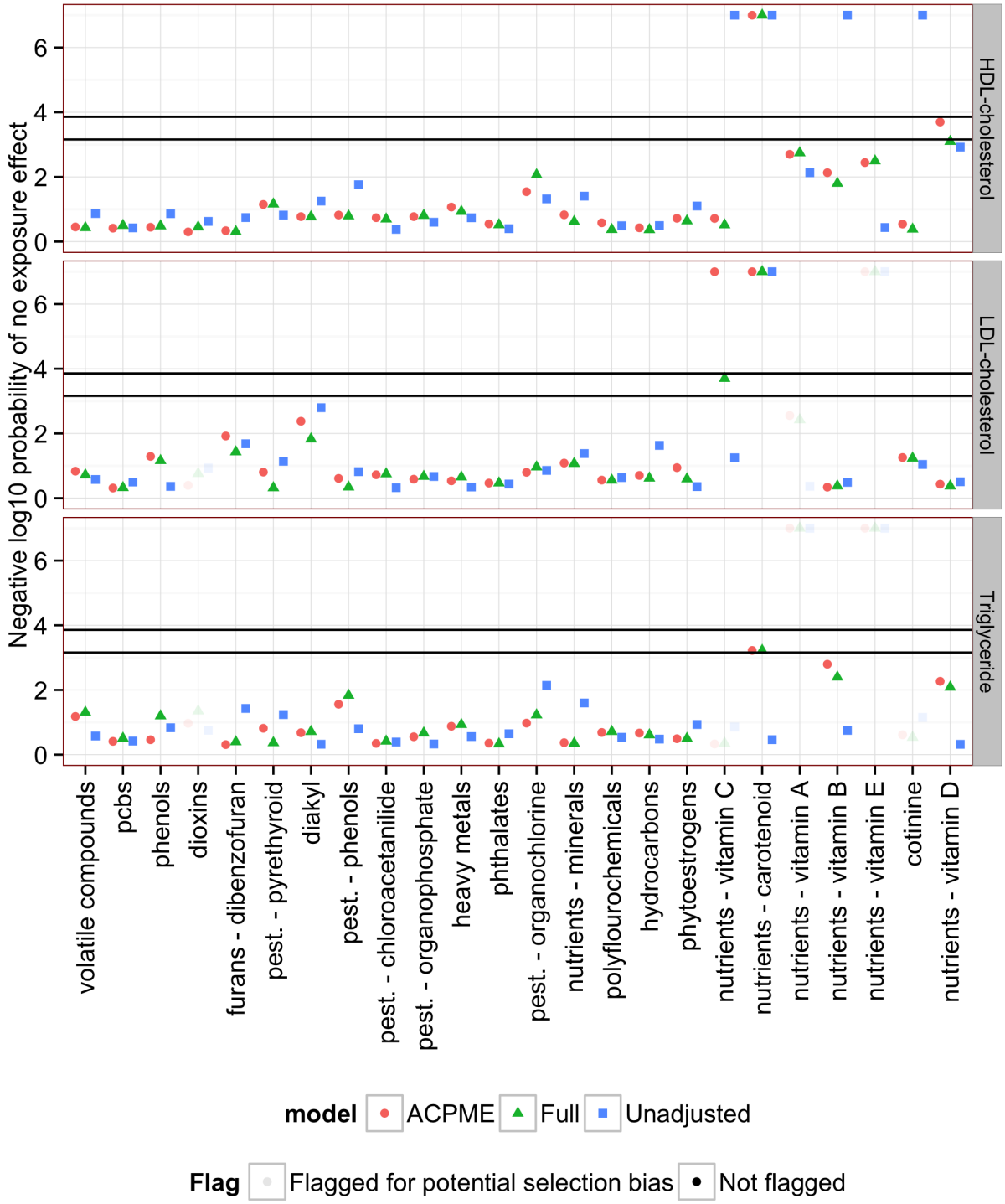


Figure 7: Posterior probability that the multivariate exposure effect is non-zero for each exposure group, outcome, and method on -$\log_{10}$ scale. We define significance level as the highest probability symmetric credible interval that does not contain zero. The black horizontal lines indicate the 0.05 and 0.01 significance level for the multivariate exposure effect. The probability levels are adjusted using Bonferroni corrections for 72 tests, the total number of tests calculated with each method. Faded estimates were flagged for potentially selection bias.

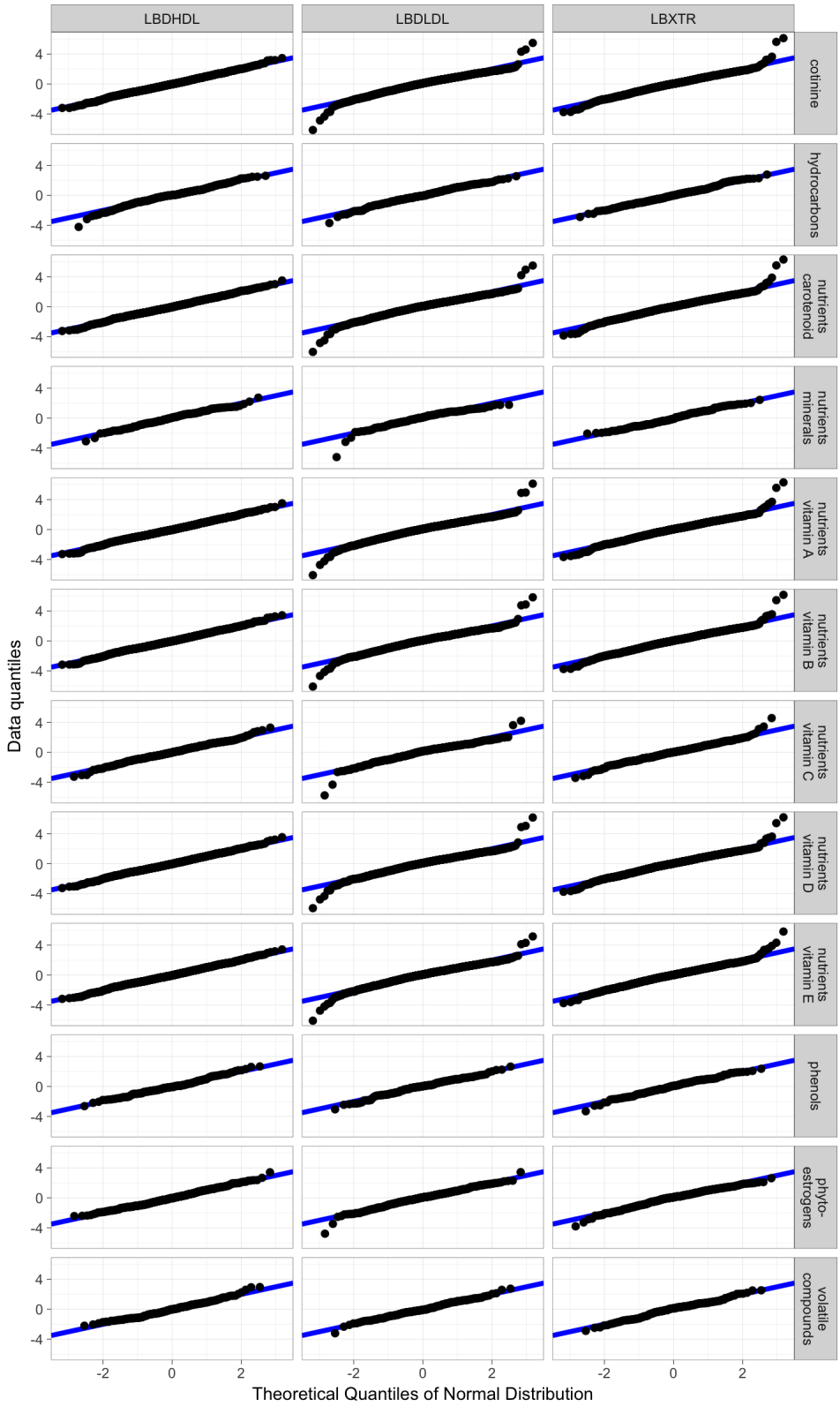


Figure 8: QQ-plots for the data analysis (part 1). The standardized residuals are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.

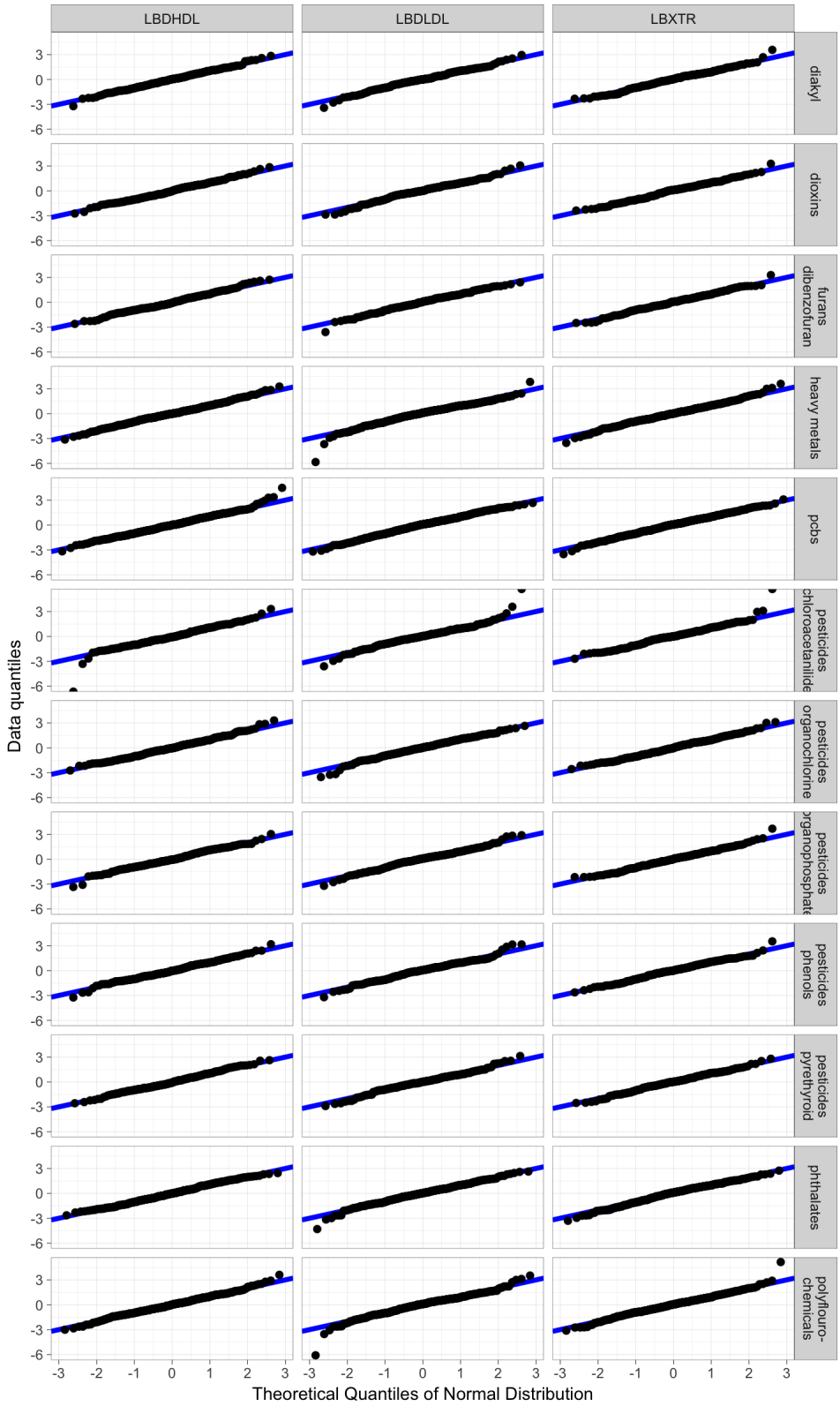


Figure 9: QQ-plots for the data analysis (part 2). The standardized residuals are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.

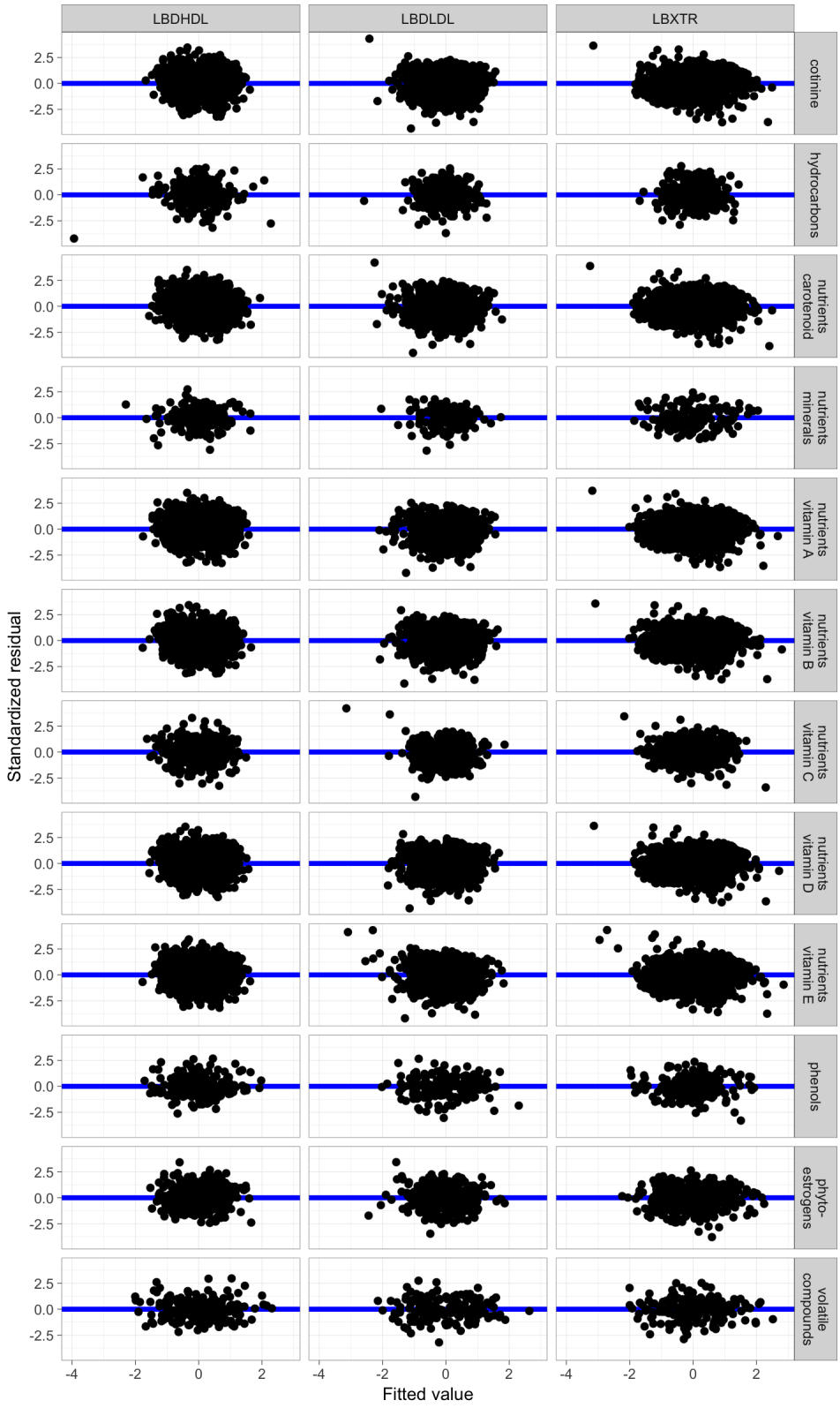


Figure 10: Residuals (y-axis) verse fitted values (x-axis) for the data analysis (part 1). The standardized residuals and fitted values are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.

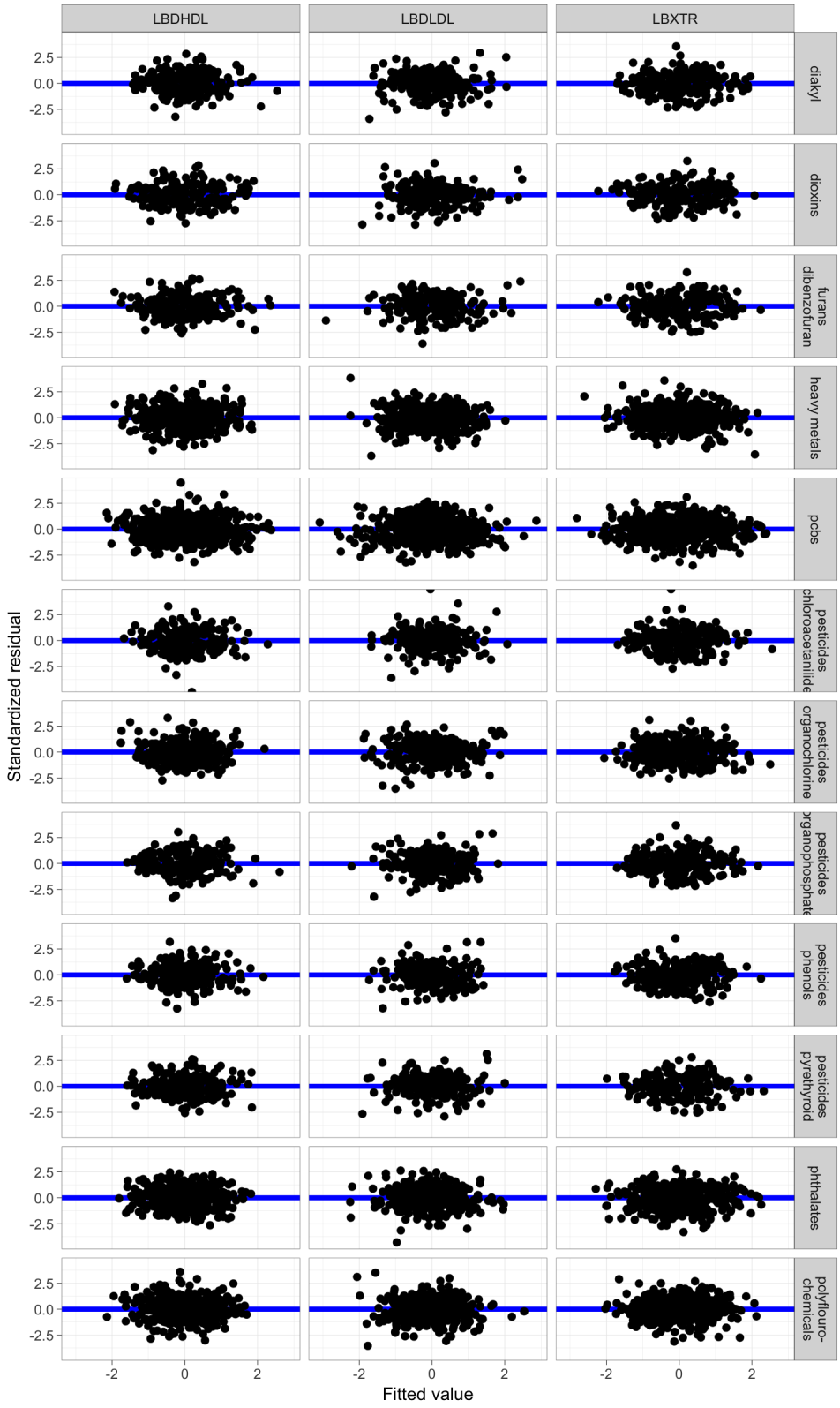


Figure 11: Residuals (y-axis) verse fitted values (x-axis) for the data analysis (part 2). The standardized residuals and fitted values are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.

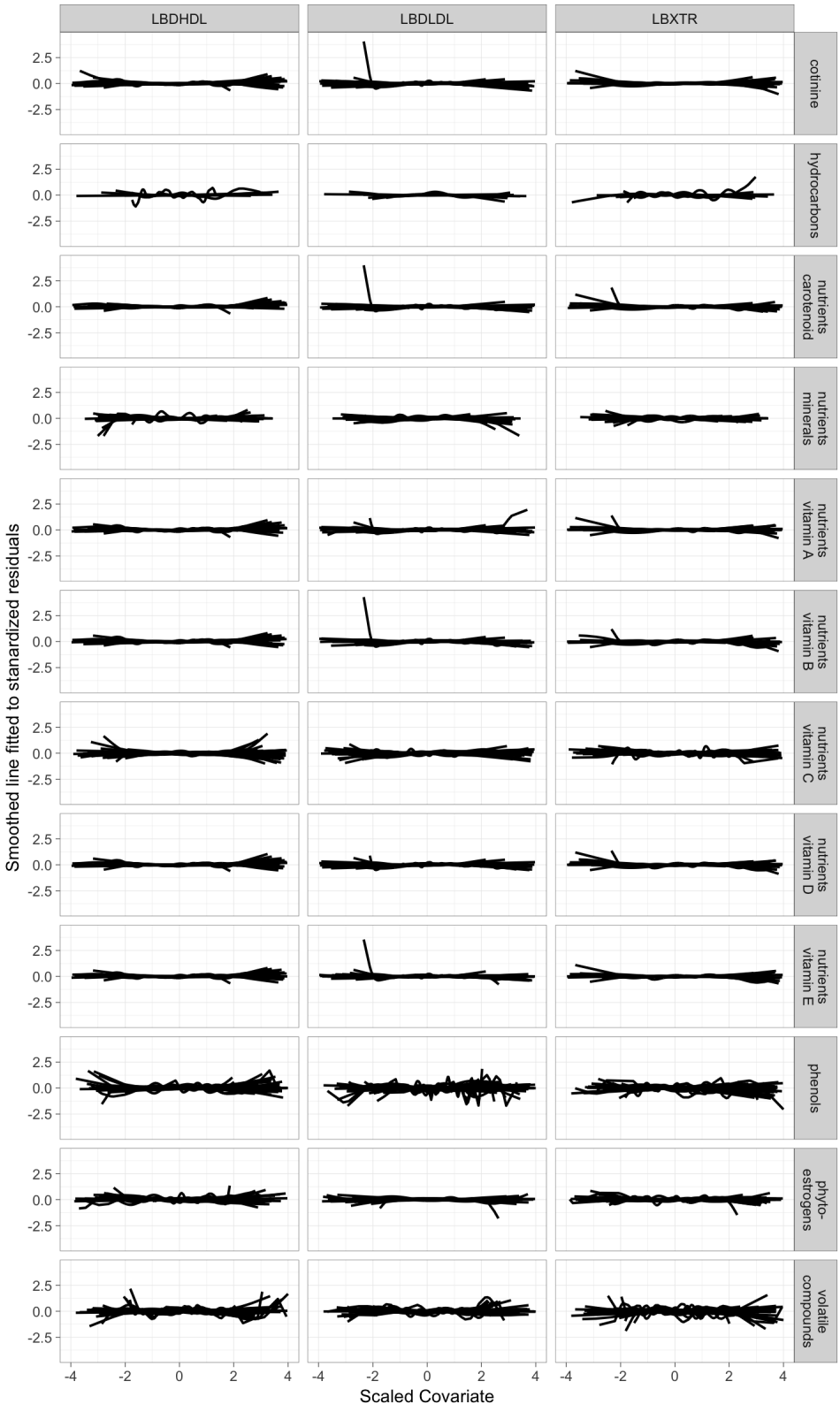


Figure 12: Smoothed residuals (y-axis) verse continuous covariates (x-axis) for the data analysis (part 1). To ease presentation the figure shows a smoothing spline fit thorough residuals instead of a point for each residuals. Each line represents a different covariate. The standardized residuals are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.

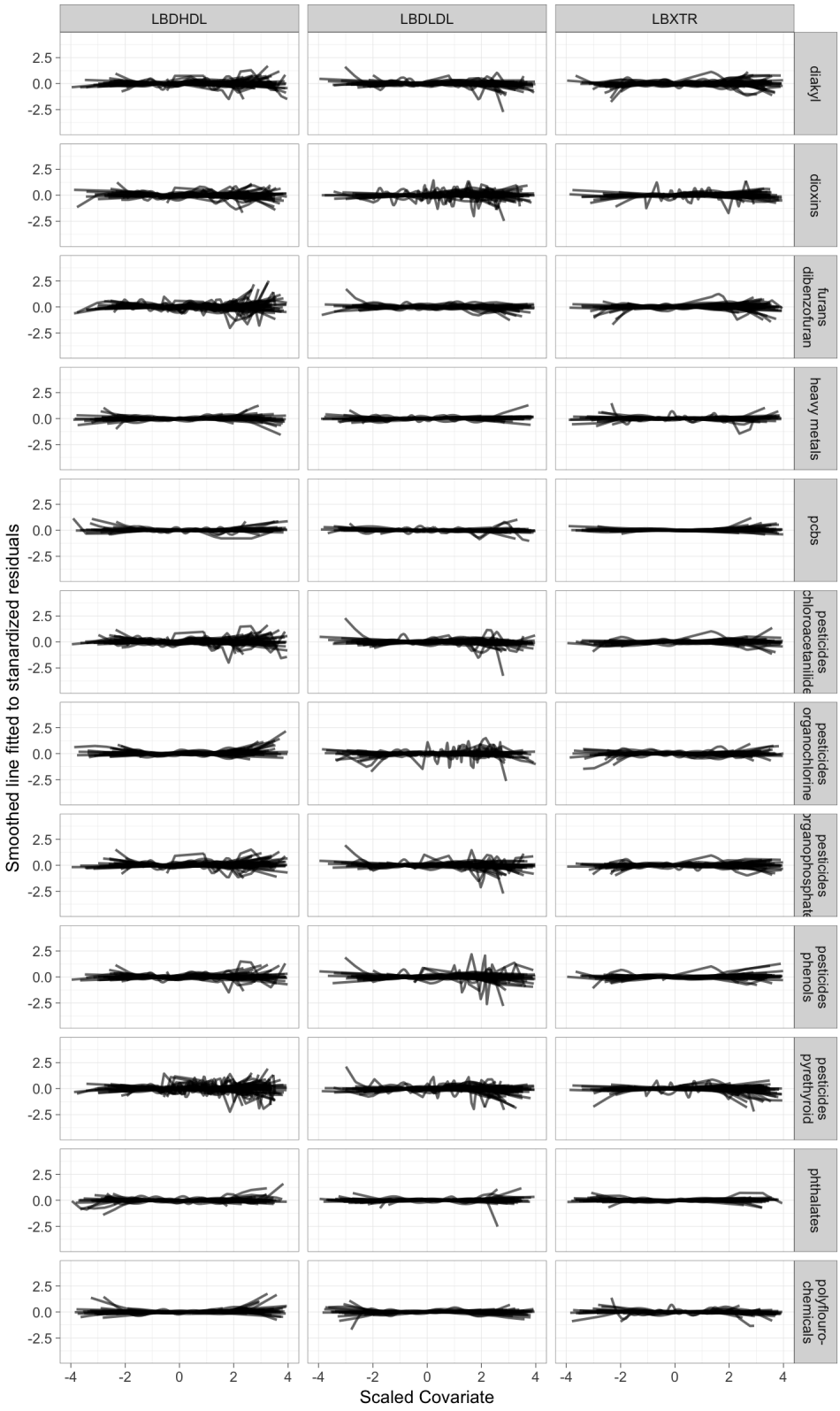


Figure 13: Smoothed residuals (y-axis) verse continuous covariates (x-axis) for the data analysis (part 2). To ease presentation the figure shows a smoothing spline fit thorough residuals instead of a point for each residuals. Each line represents a different covariate. The standardized residuals are obtained from an OLS model that includes as covariates all covariates that had a posterior probability of at least 0.5 of inclusion based on ACPME.