Simulations

Past simulation studies

Here, we preface our simulation study with an overview of examples in the literature which compare various methods for exposure mixture studies using simulations. @taylor_statistical_2016 conclude that, in general for exposure mixture studies, no single method consistently outperforms others across all situations and, importantly, that a method should be chosen based on the question of interest. Thus, for each study, we highlight not only the findings, but also the data-generating scenarios and the identified question of interest.

@lazarevic_performance_2020 compare the performance of a broad range of methods for accurate variable selection of important exposures. They simulated exposure data using a multivariate copula based on real-world data and the response by specifying a regression model with only a subset of truly significant exposures and a normal error term. Two correlation structures were considered — one with the original Spearman correlation matrix and one with the values halved — as well as two signal-to-noise ratios — one with an R^2 for the true model at 10% and one at 30%. They found that BKMR, along with three other flexible regression methods that allow for nonlinearity, provided more accurate variable selection results compared to two machine learning methods. Moreover, they observed that, in general, low signal-to-noise ratios had a stronger impact on performance than did increasing multicollinearity.

@hoskovec_model_2021 compare Bayesian methods, including BKMR, while considering 4 research questions: accurate estimation, selection of important exposures, exclusion of unimportant exposures, and identification of interactions. They use observed exposure and covariate data to simulate response data using regression relationships; they considered three exposure-response scenarios of varying complexity and included two-way multiplicative interaction terms. For each simulated dataset, they randomly assigned exposures to be active components of the mixture to incorporate variability in the data. Overall, they found that Bayesian methods outperformed traditional linear regressions, and that BKMR performed best when the exposure-response function takes on a complex form.

Most recently, @pesenti_comparative_2023 compare BKMR, BSR, and the Bayesian Least Absolute Shrinkage and Selection Operator (LASSO) for variable selection. Data were generated using a multivariate normal with moderate and strong correlation structures specified manually by the researchers. They found that, in situations with additive and linear exposure-response relationships, Bayesian LASSO was appropriate. Across the other scenarios, BKMR generally performed best, while BSR selected exposures with high heterogeneity when the sample size was smaller due to the influence of the degrees of freedom, d, tuning parameter. Notably, multicollinearity did not generally lead to spurious variable selection.

Finally, we briefly comment on studies by @sun_statistical_2013 and @barrera-gomez_systematic_2017, whose explicit goal is to compare methods for identifying interactions. Both studies generate exposure data using the correlation structure from an existing dataset; @sun_statistical_2013 uses a multivariate lognormal, while @barrera-gomez_systematic_2017 uses a multivariate normal. Both only consider two-way, multiplicative interactions. While neither of these studies consider the methods used in this thesis, they find that, in general, models that formally allow for interaction effects perform better than models that only allow for univariate additive effects.

Methods

The goal of our simulation study is to provide guidance on the choice between BSR and BKMR for characterizing a diverse range of complex interactions between predictors. In particular, we aim to extend findings from previous simulation studies by considering a more comprehensive range of interaction types, including different effect sizes, non-multiplicative interactions, and three-way interactions. We also explore interactions between exposures and categorical covariates, a previously understudied form of interaction in exposure mixture studies.

MADRES data

In order to make our simulations comparable to real-world exposure mixture studies, we based our simulation data on the Maternal And Developmental Risks from Environmental and Social Stressors (MADRES) pregnancy cohort. The MADRES cohort is an ongoing, prospective pregnancy cohort of predominantly lower-income, Hispanic women in Los Angeles, California, which began in 2015 [@bastain_study_2019]. Urine samples were collected by participants at their first visit, and questionnaires were administered during their first visit, with follow-ups at the first, second, and third trimesters. See @bastain_study_2019 for further details on study design.

@howe_prenatal_2020 previously examined the effect of prenatal metal mixtures of birth weight (BW) for gestational age (GA) in this cohort. They used BKMR to identify associations between metal mixtures and BW for GA, as well as BSR to conduct inference on interactions between metals. Using BKMR, they found that, of the metals in the mixture, mercury and nickel were most strongly associated with BW for GA. Moreover, BKMR results suggested that a potential interaction between mercury and nickel exists; however, when run through BSR, the PIP for this interaction was extremely small, despite being the highest of all two-way interactions.

Data from the study by @howe_prenatal_2020 were obtained from publicly available data in the Human Health Exposure Resource (HHEAR) Data Repository, which has been approved under Icahn School of Medicine at Mount Sinai IRB Protocol #16-00947. The Digital Object Identifiers associated with the urinary trace element data and epidemiological data are 10.36043/1945_159 and 10.36043/1945_177, respectively. All analyses were conducted in R v4.3.2 [@r_core_team_r_2013].

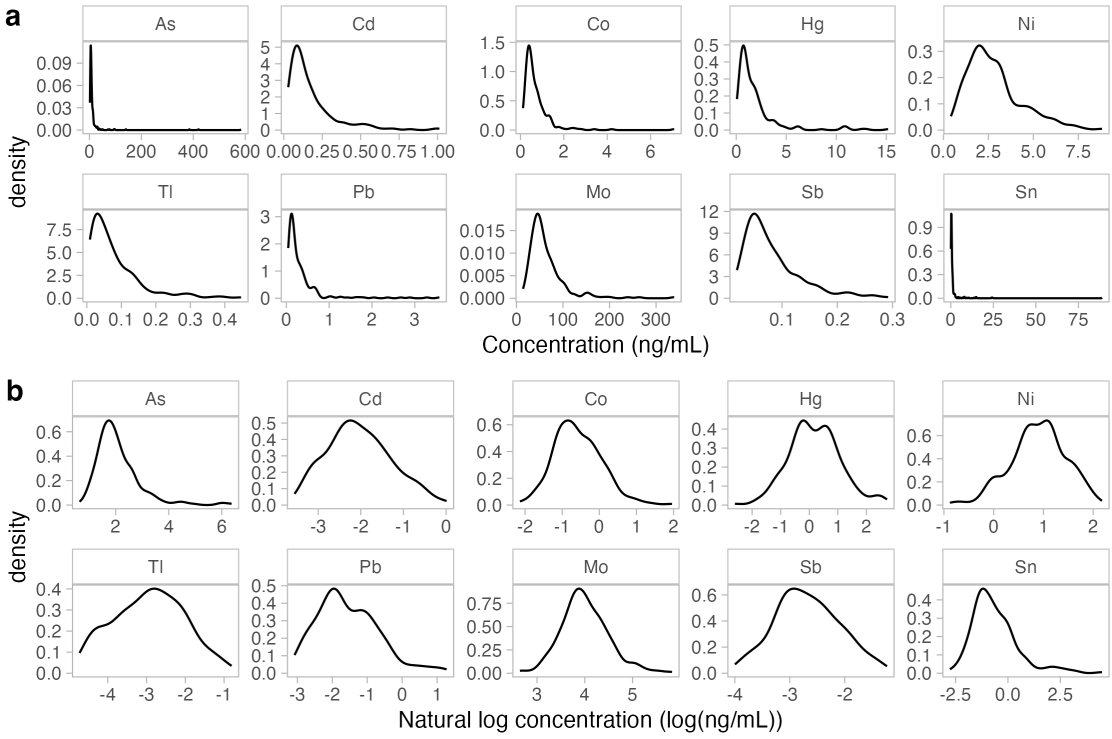

Figure 1: Distributions of original (a) and natural log transformed (b) concentrations of metals in MADRES cohort (n=252).

We followed the approach by @howe_prenatal_2020 for preparing the data for analysis. This resulted in retaining 10 metals in analysis: arsenic (As), cadmium (Cd), cobalt (Co), mercury (Hg), nickel (Ni), molybdenum (Mo), lead (Pb), antimony (Sb), tin (Sn), and thallium (Tl). @howe_prenatal_2020 used speciated As, but this was not available in HHEAR, so we used total As. Metals were expressed in nanograms

per milliliter (ng/mL) and natural log transformed to reduce right-skewness (Figure @ref(fig:logtransf)). Among the full range of covariates considered by @howe_prenatal_2020, we used the subset of 4 that were available in HHEAR: any smoke exposure during pregnancy, maternal prepregnancy body mass index (BMI), maternal age during firt trimester, and maternal race by ethnicity and birth place. We chose not to include study site, as there was a study site with only 1 participant. Race by ethnicity and birth place was collapsed into the following categories: non-Hispanic white, non-Hispanic black, non-Hispanic other, Hispanic born in the US, and Hispanic born outside the US. We observed 8 missing values for BMI in the data from HHEAR, which were not reported by @howe_prenatal_2020. We mean imputed these missing values. Distributions of covariates are shown in Figure @ref(fig:covdist). Our final analytic dataset included 252 participants, which was 10 fewer than in @howe_prenatal_2020, likely due to small discrepancies in their dataset and the one made available in HHEAR.

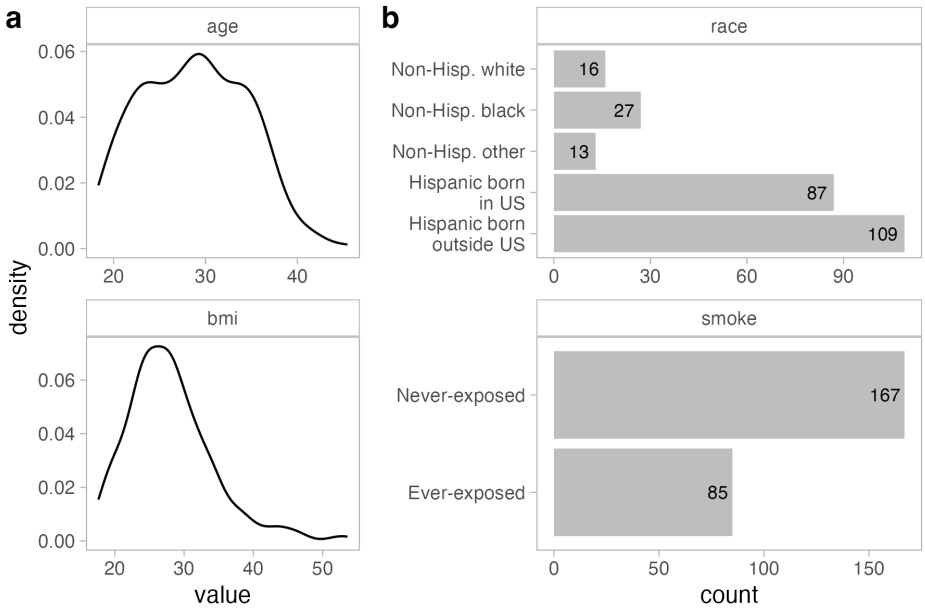

Figure 2: Distributions of continuous (a) and categorical (b) covariates in the MADRES cohort (n=252).

Using copulas to simulate predictor data

We simulated exposure and covariate data (hereafter referred to collectively as predictors) using a multivariate Gaussian copula fit on the 252 participants in the MADRES cohort. We used copulas as they can preserve both the correlation structure and marginal distributions from the observed data, allowing us to replicate conditions in a real-world scenario.

First, we briefly introduce copulas in the context of their use in this simulation, based on the presentation in @nelsen_introduction_2006. Copulas are joint cumulative distribution functions (CDFs) defined on the unit cube $[0,1]^n$ that capture the dependence between n uniformly distributed marginals. Sklar's theorem allows us to apply copulas to our observed data. Sklar's theorem states that, if $H(x_1, \dots x_n)$ is a joint CDF of the marginal CDFs $F_1(x_1), \dots, F_n(x_n)$, then there exists a copula C such that, for all $(x_1, \dots, x_n)$ in $(X_1, \dots, X_n)$,

$$H(x_1, \dots x_n) = C(F_1(x_1), \dots, F_n(x_n)).$$

Note that, by the probability integral transform, or the universality of the uniform, the CDFs $F_1(x_1), \dots, F_n(x_n)$ are distributed uniformly.

We used the `copula` package in R to fit copulas and generate random data [@hofert_copula_2023]. We transformed the observed continuous predictor values to uniform distributions based on their empirical marginal CDFs, a process called generating "pseudo-random" samples. We used the checkerboard copula approach for generating pseudo-random samples for smoke exposure, a binary variable [@genest_primer_2007]. We coded smoke exposure as 0's and 1's, generated a pseudo-random sample, and then "jittered" the values with uniform random noise. There is currently no widely accepted approach for generating pseudo-random samples from unordered categorical variables with more than two levels. Thus, we excluded race by ethnicity and birthplace from the copula model. While this means that our simulated datasets did not preserve any potential association between race and exposures, Figure @ref(fig:raceexp) suggests that there is little to no visible association between race and exposures in the observed dataset.

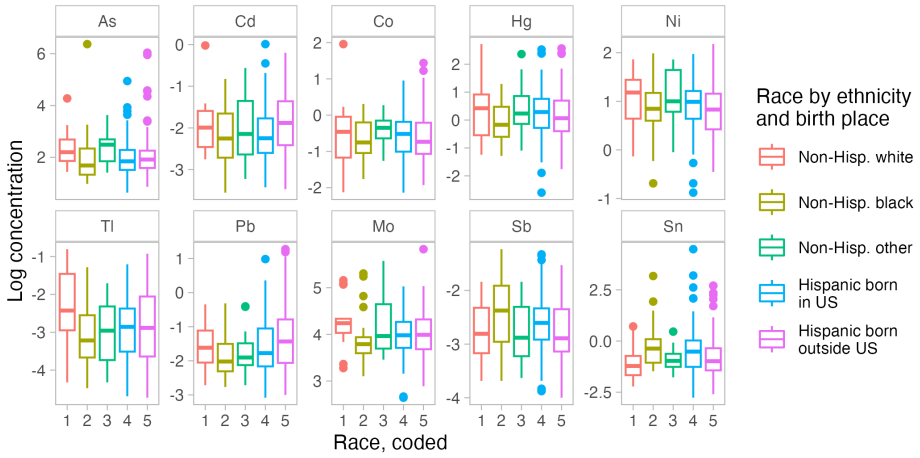

Figure 3: Association between race by ethnicity and birth place and metal exposures in the MADRES cohort (n=252).

Various families of copulas have been described, each of which specifies a different shape for the dependence structure. We performed model selection to identify the copula that best approximates the dependence structure of our data. We fit the set of multivariate copulas used by @lazarevic_performance_2020 in their simulation study, which included the Gaussian, t, Gumbel, Frank, Clayton, and Joe copulas. We fit two t copulas with 4 and 10 degrees of freedom, which controls dependence at the tails of the distributions, as well as a t copula where the degrees of freedom was determined during the fitting process. The Gumbel, Frank, Clayton, and Joe copulas require a θ parameter, which controls dependence between the distributions. We fit two versions of these copulas with $\theta = \{2, 4\}$. Among these, the Gaussian copula minimized the Akaike information criterion and maximized the likelihood, so we proceeded with this model. The Gaussian copula assumes a bivariate normal dependence structure between the marginal CDFs.

We simulated predictor data by randomly sampling from the fitted multivariate Gaussian copula distribution. All pseudo-random samples were then back-transformed to their original distributions using empirical marginal CDFs. We simulated the race by ethnicity and birthplace variable by randomly assigning observations to each of the five categories based on proportions in the observed dataset.

We generated one set of simulated datasets with the same sample size as the observed dataset (n=252), which is typical in many cohort studies. We also generated another set of simulated datasets with a larger sample size (n=1000), which has become increasingly common with the rise of larger-scale studies. The goal of this choice was to inform sample size considerations in study design. We verified that the original structure of the observed dataset were preserved by visually comparing univariate distributions of exposures (Figure @ref(fig:univexpsim)) and covariates (Figure @ref(fig:univcovsim)), as well as the correlation structure using Spearman's ρ (Figure @ref(fig:corsimssm)). Distributions of Spearman's correlation were approximately normal (Figure @ref(fig:cordistsm)). Plots for the larger size simulated datasets were similar (Figures @ref(fig:univexplg), @ref(fig:univcovlg), and @ref(fig:corsimslg)).

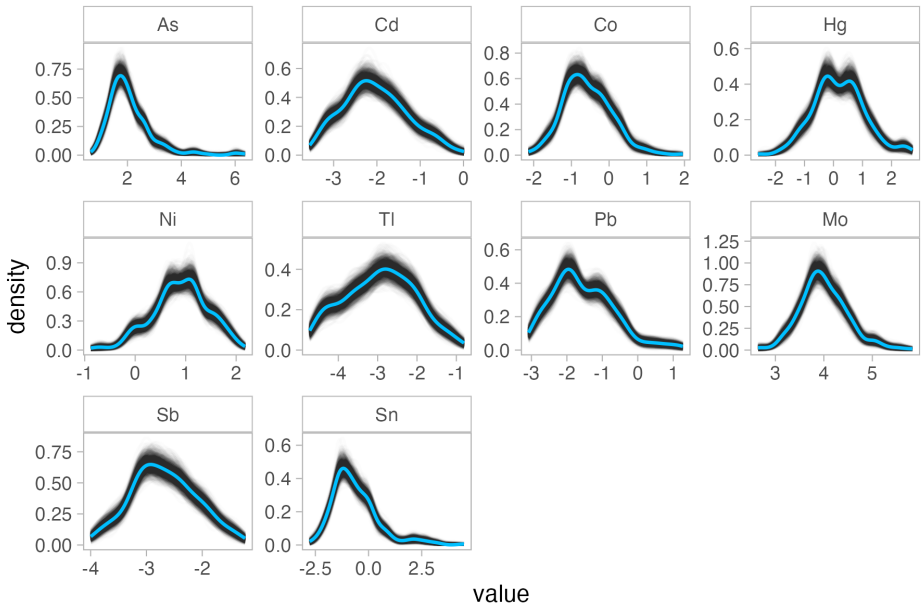

Figure 4: Distributions of log-transformed exposures from observed data (blue) and 2100 simulated smaller size (n=252) datasets (gray).

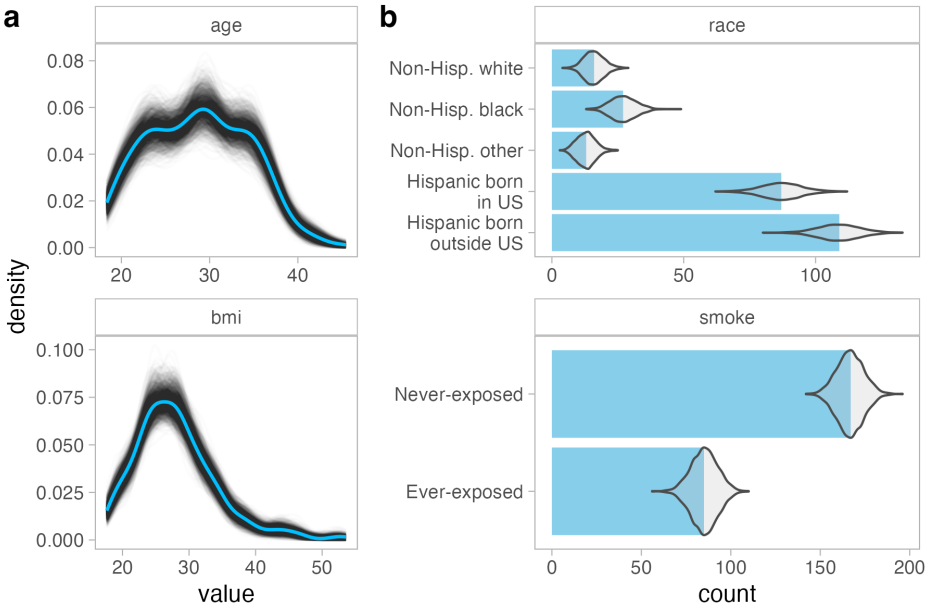

Figure 5: Distributions of continuous (a) and categorical (b) covariates from observed data (blue) and 2100 simulated smaller size (n=252) datasets (gray).

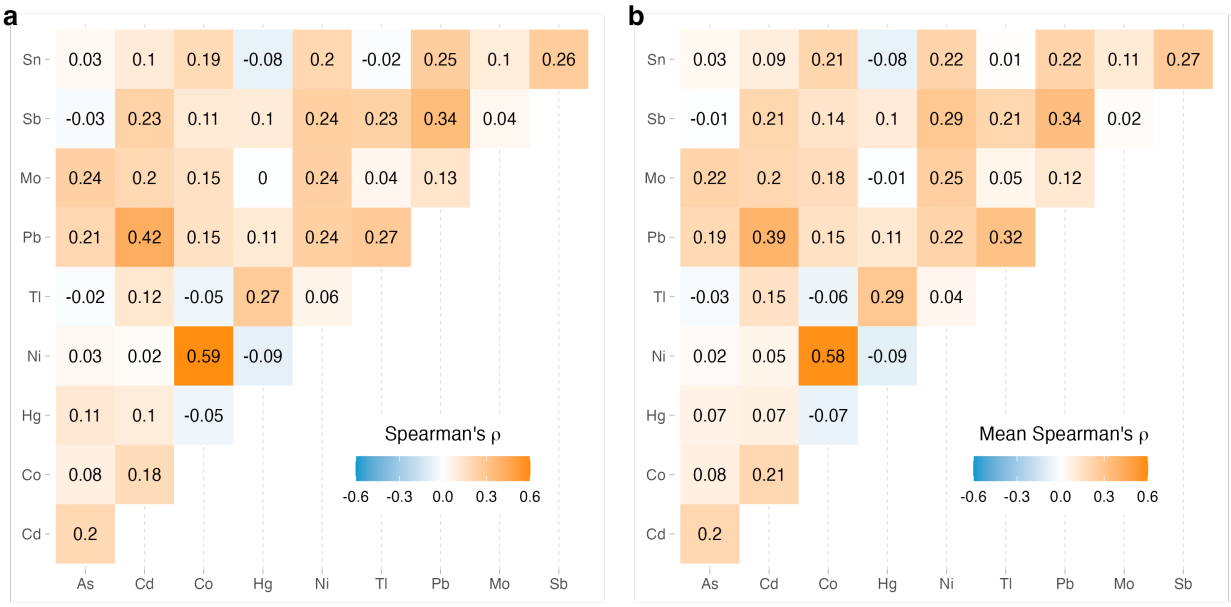

Figure 6: Spearman's correlation heat maps of exposures from observed data (a) and averaged across 2100 smaller size (n=252) simulated datasets (b).

Simulating predictor-response relationships

Health outcome responses were simulated under several different scenarios, each of which included different effect sizes and functional forms for the interactions. All scenarios were run for both the smaller (n=252) and larger (n=1000) sample sizes. In the first scenario, we specified a "base case" model:

$$Y = \text{Hg} + \frac{3}{1+\exp(-4\text{Ni})} + \frac{1.5}{1+\exp(-4\text{Sn})} - \text{Sb}^2 + 0.5\text{Sb} + \text{age} + 0.5\text{bmi} + 0.5\text{race}_{\text{black}} + 0.5\text{race}_{\text{hisp.non}} + 1.5\text{smoke} + \varepsilon,$$

where $\varepsilon \overset{\text{iid}}{\sim} N(0, 5)$. This model includes a linear term for Hg, two S-shaped logistic terms for Ni and Sn with varying effect sizes, and a symmetric inverse U-shaped quadratic term for Sb (Figure @ref(fig:univlines)). Moreover, we included covariate terms as linear effects in the model. We chose the standard deviation on the normal random error term in order to achieve an R^2 of around 0.1-0.3 in a multiple linear regression that included only the true functional form of the significant chemicals (Figure @ref(fig:rsqcheck)). This R^2 range approximates realistic signal-to-noise ratios in exposure mixture studies [@lazarevic_performance_2020].

In subsequent scenarios, we added an additional interaction term to the base case model. First, we considered interactions between two exposures. We defined four cases of interest: a two-way interaction between exposures that are univariately significant, a two-way interaction between exposures that are univariately insignificant, a two-way interaction between exposures that are moderately collinear, and a three-way interaction. For each case, we considered two functional forms — multiplicative and polynomial — and a lower and higher effect size, which we set by defining the weight on the interaction term in the model. The higher effect sizes were selected in order to achieve a power of approximately 0.5 at $\alpha = 0.05$ in the smaller sample size (n=252) case, using a multiple linear regression with the true functional form of the chemicals specified and the covariate terms included. The lower effect sizes were set equal to half of the higher effect size. Table @ref(tab:scenarios) shows the specification of interaction terms. See Appendix @ref(suppmethods), Figures @ref(fig:basesurf)-@ref(fig:cp2), for 3D surfaces of the two-way interaction terms.

Table 1: Specification of interaction terms in simulations.

	Effect size	
	Lower	Higher
Univariately significant		
Multiplicative	0.35Hg∗Ni	0.7Hg∗Ni
Polynomial	$0.13\text{Hg}*(\text{Ni}-1)^2$	$0.26\text{Hg}*(\text{Ni}-1)^2$
Univariately insignificant		
Multiplicative	0.35Cd∗As	0.7Cd∗As
Polynomial	$0.125\text{Cd}*(\text{As}-1)^2$	$0.25\text{Cd}*(\text{As}-1)^2$
Highly correlated		
Multiplicative	0.3Ni∗Co	0.6Ni∗Co
Polynomial	$0.1\text{Ni}*(\text{Co}-1)^2$	$0.2\text{Ni}*(\text{Co}-1)^2$
Three-way interaction		
Multiplicative	0.3Hg∗Ni∗Tl	0.6Hg∗Ni∗Tl
Polynomial	$0.09\text{Hg}*(\text{Ni}-1)^2*\text{Tl}$	$0.18\text{Hg}*(\text{Ni}-1)^2*\text{Tl}$

Next, we considered interactions between the race by ethnicity and birthplace covariate (hereafter referred to as race for concision) and an exposure. We are interested in cases where the health effects of an exposure are higher in one group compared to the rest. In a real-world scenario, such interactions can arise from excess amounts of social stress experienced by a group due to racism. To model this, we increased the coefficient Hg in Non-Hispanic Black individuals (n=27 in the original MADRES cohort) for the first scenario, and in Hispanic individuals born outside the US (n=109 in the original MADRES cohort) for the second scenario. The goal of this choice was to assess the impact of group size on detectability of an interaction, and to quantify the potential value of oversampling the minority group. For each scenario, we specified a lower effect size by increasing the coefficient on Hg by 1.5× (i.e. from 1∗Hg to 1.5∗Hg) in the target group, and a higher effect size by increasing the coefficient on Hg by 2× (i.e. from 1∗Hg to 2∗Hg).

This resulted in a total of 42 scenarios ([1 base case + 5 interaction cases × 2 effect sizes × 2 functional forms] × 2 sample sizes = 42). For each scenario, we generated 100 simulated datasets to fit our models on, resulting in a total of 4200 datasets.

Models

We ran four methods on our simulated datasets. All metal concentrations and continuous covariates were standardized in analysis to keep values scale-free.

To get a baseline, we ran a multiple linear regression, including all exposures and covariates as linear, additive terms in the model. We refer to this model as the naive MLR. Then, we ran a multiple linear regression with the true model explicitly specified by excluding non-significant exposures and specifying the known form of non-linear terms and non-additive interactions. We refer to this model as the oracle MLR. In scenarios with an interaction between race and Hg, we collapsed race into a binary variable indicating whether or not the original race category was interacting with Hg before running oracle MLR's, in order to simplify the detection of the interaction.

Next, we ran BKMR using the `bkmr` package in R [@bobb_statistical_2018; @bobb_bkmr_2022]. We chose to implement component-wise variable selection rather than hierarchical selection to make simulation results more interpretable, and because there was only moderate multicollinearity in the observed and simulated data. We specified the default priors [@bobb_bayesian_2015, and as listed in Chapter \@ref(bkmrprior)], which is common in the literature for BKMR [e.g., @lazarevic_statistical_2019, @howe_prenatal_2020, @pesenti_comparative_2023]. We ran the MCMC sampler for 50,000 iterations, as recommended by @bobb_statistical_2018, and discarded the first 25,000 iterations for burn-in. BKMR does not provide

the option to run multiple chains or to thin chains. For larger size datasets, we sped up computations by employing a Gaussian predictive process on 100 knots specified evenly across the predictor space.

We ran BSR using the `NLinteraction` package in R [@antonelli_nlinteraction_2018]. We specified the default priors [@antonelli_estimating_2020, and as listed in Chapter \@ref(bsrprior)], which is common in the literature for BSR [e.g., @howe_prenatal_2020; @pesenti_comparative_2023]. @antonelli_estimating_2020 suggests separately fitting models for degrees of freedom $d = \{1, 2, 3, 4\}$ and selecting the value for d which minimizes WAIC. Due to time constraints in this thesis, we first fit BSR on the grid of values for d using 5,000 MCMC iterations to obtain the empirical Bayes estimate for σ^2_β and then another 5,000 MCMC iterations to obtain the posterior distributions, discarding the first 2,500 iterations for burn-in each time. We selected d based on the WAIC criterion on these preliminary models. Then, we fit the full BSR model using 50,000 MCMC iterations to obtain the empirical Bayes estimate and then another 50,000 MCMC iterations to obtain the posterior distributions, discarding the first 25,000 iterations for burn-in each time. We ran two chains to verify convergence, thinning each chain by selecting every 8th iteration to reduce autocorrelation based on default settings. In a small test run on five smaller size datasets for each scenario containing interactions between exposures, as well as the base case, we found that using 5,000 iterations selected the same degrees of freedom as using 50,000 iterations 86% of the time (see Appendix @ref(suppmethods), Figure @ref(fig:comparedf)).

Finally, we ran stratified BKMR and BSR models in scenarios where we simulated an interaction between race and Hg. This involved running five separate models for each race category, each with the same settings specified above. For the smaller size datasets, we often observed convergence issues in BKMR within the smaller race categories. As such, for smaller size datasets, we also assess the impact of collapsing the three smaller race categories (Non-Hispanic white, black, and other) into one category before stratifying. This is a common practice in real-world studies where sample sizes for certain categories are low.

We checked convergence for a selection of BKMR and BSR models using trace plots (Appendix @ref(suppresults), Figure @ref(fig:traceplots)).

Model assessment

We assessed model performance based on detection of significant univariate chemicals as well as detection of interactions. For the naive and oracle MLRs, we considered a p-value less than 0.05 to indicate detection of a significant term. For BKMR and BSR, we used the median probability model, which considers a PIP greater than or equal to 0.5 to indicate detection of a significant term [@barbieri_optimal_2004].

While BSR provides PIP's to quantify detection of interactions, BKMR does not. As such, for BKMR, we considered formal detection of an interaction based on confidence intervals constructed around the estimated response. Specifically, we first calculated the difference in estimated response at a chemical's 0.25 and 0.75 quantiles. Then, we assessed whether this quantity differed at the 0.25 and 0.75 quantiles of one (or two, for three-way interactions) other chemicals in the interaction, while holding all other chemicals at their 0.5 quantiles, by constructing a 95% confidence interval of the difference in differences. We followed the code in the `SingVarIntSummaries()` function in the `bkmr` package for constructing confidence intervals [@bobb_bkmr_2022].

For both BKMR and BSR, we also visually assessed detection of interactions by plotting the estimated exposure-response surface for one chemical while fixing one (or two, for three-way interactions) other chemicals at their 0.1, 0.5, and 0.9 quantiles. In all scenarios, we calculated the sensitivity as the proportion of times a significant term was correctly detected. Due to time constraints in this thesis, we only calculated the false discovery rates, or, the proportion of times a significant term is incorrectly detected, for two-way interactions between chemicals. Future work should include the calculation of false discovery rates for all interactions.

For stratified models, we compared the estimated response across each separate model. Specifically, for BKMR, we computed a confidence interval for the difference in estimated response at the 0.25 and 0.75 quantiles of Hg on each subcategory of race, following the code in the `SingVarRiskSummaries()` function

in the `bkmr` package [@bobb_bkmr_2022]. We adjusted for multiple comparisons based on the Bonferroni procedure by constructing 5 simultaneous 99% confidence intervals, in order to achieve an overall 95% confidence level [dunn_multiple_1961; @vanderweele_desirable_2019]. We considered an interaction as correctly detected if (1) there was at least one overlap between the target group's confidence interval and all other confidence intervals, and (2) all other confidence intervals overlapped.

However, for BSR, we were not able to find a method in the literature for combining variances from estimated responses at two different sets of predictor values. Therefore, we visualized and compared the estimated exposure-response relationship in each of the stratified models as a qualitative way to assess for interactions. We also generated these diagnostic plots for BKMR. Moreover, due to time constraints in this thesis, we only considered sensitivity, as opposed to also considering false discovery rates. Future work should involve exploring methods for estimating contrasts along continuous predictors for a more formal method of inference on BSR output, as well as consideration of false discovery rates.

Results

Base case

We start by presenting results from models run on the base case scenario, in which the true relationship contained no interactions. Figure @ref(fig:basecasesig) displays the distribution of p-values and PIPs from this scenario, while Table @ref(tab:basecasetab) summarizes model sensitivity and false discovery rates based on these values. Note that insignificant chemicals were not included in the oracle MLR, which is why their distributions are omitted from this output.

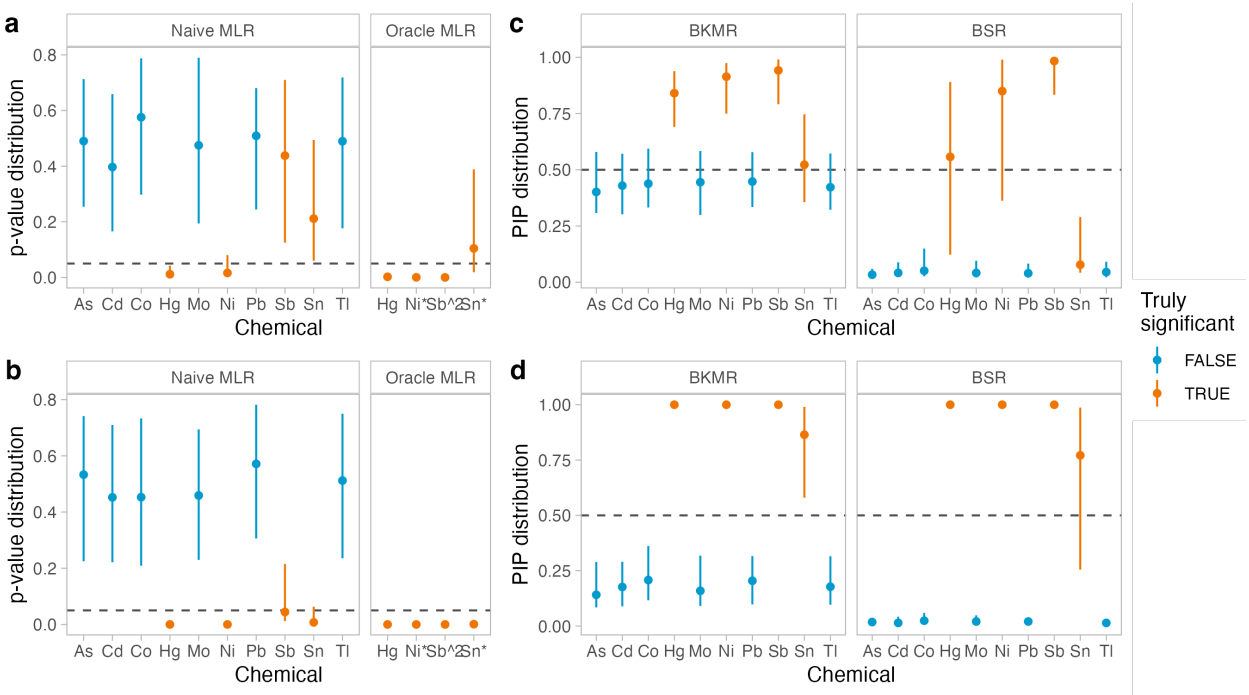

Figure 7: P-value distributions from smaller (a) and larger (b) size datasets and PIP distributions from smaller (c) and larger (d) size datasets.

The naive MLR does the best job at picking up the effects of Hg, with a sensitivity of 0.8 and 1 in the smaller and larger size datasets, respectively. This is likely because Hg is the only linear term in the model. Ni and Sn have an S-shaped curve with higher and lower effect sizes, respectively, so the naive MLR detects a slight linear signal from them. Sb, which has a U-shaped curve, is the hardest to pick up. On the other

Table 2: Sensitivity and false discovery rate (FDR) of chemicals in base case scenario.

	Sensitivity				FDR					
	Hg	Ni	Sb	Sn	As	Cd	Co	Mo	Pb	Tl
Naive MLR										
Small	0.80	0.70	0.12	0.24	0.05	0.07	0.02	0.07	0.03	0.08
Large	1.00	1.00	0.51	0.71	0.02	0.04	0.08	0.13	0.03	0.03
Oracle MLR										
Small	0.84	0.94	0.95	0.35	-	-	-	-	-	-
Large	1.00	1.00	1.00	0.89	-	-	-	-	-	-
BKMR										
Small	0.86	0.92	0.95	0.52	0.30	0.33	0.36	0.38	0.41	0.40
Large	1.00	1.00	1.00	0.77	0.12	0.12	0.14	0.13	0.08	0.13
BSR										
Small	0.51	0.67	0.88	0.17	0.06	0.04	0.04	0.04	0.04	0.05
Large	1.00	1.00	1.00	0.62	0.01	0.00	0.03	0.02	0.03	0.00

hand, the oracle MLR consistently detects Hg, Ni, and Sb. The smaller size oracle MLRs only occasionally pick up Sn, likely due to the lower effect size.

BKMR has similar sensitivity rates as the oracle MLR, ranging from 0.52 to 0.95 in the smaller size datasets and 0.77 to 1.00 in the larger size datasets. However, BKMR also has, by far, the highest false discovery rates, ranging from 0.30 to 0.41 in the smaller size datasets and 0.08 to 0.14 in the larger size datasets. This is likely due to the default choice of an inverse uniform distribution from 0 to 100 for the "slab" component on the "slab-and-spike" prior on r_m. Choosing a prior that assigns higher probability to smaller values of r_m should reduce the false discovery rate. In contrast, BSR tends to have slightly lower sensitivity rates than BKMR, raning from 0.17 to 0.88 in the smaller size datasets and 0.62 to 1.00 in the larger size datasets, but the false discovery rates are much lower, ranging from 0.04 to 0.05 in the smaller size datasets and 0.00 to 0.03 in the larger size datasets. Together, this suggests that the default prior choices provided by BSR are likely a better fit for this scenario.

Overall, this base case scenario confirms that the multiple regression and Bayesian models behave as expected. We also recognize that, for univariate significance metrics, BKMR tends to produce higher sensitivity and false discovery rates, while the opposite is true for BSR.

Univariate sensitivity

Now, we provide a brief overview of univariate sensitivity metrics from all scenarios with an interaction between chemicals. We are particularly interested in cases where it appears that the inclusion of an interaction term influences the detection rate of univariate chemicals.

Table @ref(tab:onewaytabsens) summarizes the sensitivity and false discovery rates of univariate chemicals in all scenarios with interactions between chemicals, comparing the form of interactions, effect sizes, size of datasets, and models.

In general, the detection rates of univariate chemicals in naive MLRs are not affected by the inclusion of interaction terms, in comparison to the base case. Though, in scenarios with a polynomial interaction between As and Cd, both of which are univariately insignificant, the larger size naive MLR detects a significant effect from Cd 67% of the time. Figures

See Figures @ref(fig:nsmunivp)-@ref(fig:slgunivp) in Appendix @ref(suppresults) for the full p-value and PIP distributions for all scenarios with interactions between chemicals.

Everything else is in appendix

Two-way interactions between chemicals

Probably only include false discovery rates for BKMR for, again, Hg and Ni to reduce complexity and run-time, note that future work would involve getting FDR's for all models

Hg Ni full output

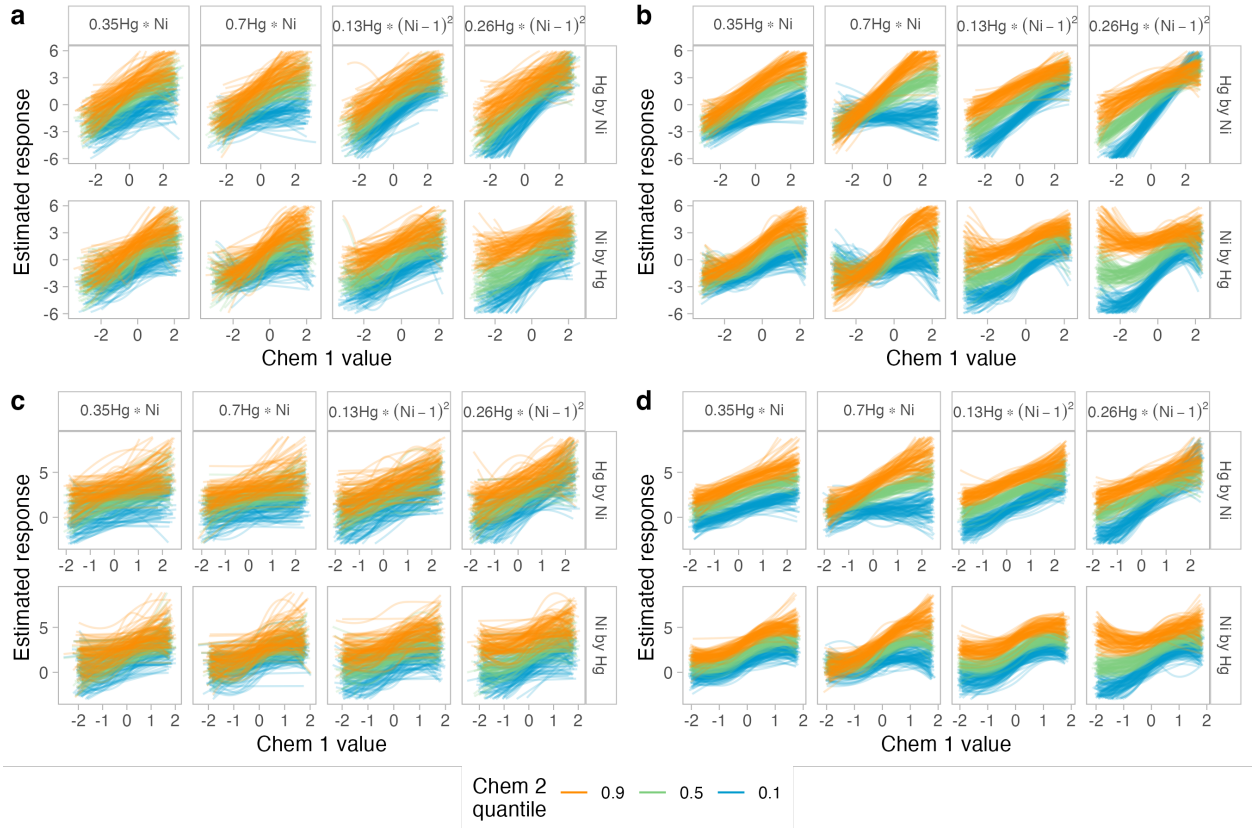

Table of all sensitivities

Table of FDRs

Everything else is in appendix

Three-way interactions between chemicals

Decide whether or not to include a figure???

Table of sensitivities (try alternative ways of thresholding BSR???)

Consider getting false discovery rates?

Interactions between race and an exposure

Figure of diagnostic plots

BKMR sensitivity

In general, interactions between a categorical variable and a continuous variable are harder to detect, both in the typical multiple linear regression framework as well as using BKMR and BSR. Even when the effect of Hg is doubled in the largest race category, the oracle MLR

Table 3: Sensitivity to interactions in all scenarios with two-way interactions between exposures.

		Small (n=252)			**Large (n=1000)**		
Interaction type	Effect size	Oracle	BKMR	BSR	Oracle	BKMR	BSR
Hg-Ni							
Multiplicative	Lower	0.24	0.01	0.07	0.48	0.21	0.28
	Higher	0.48	0.03	0.14	1.00	0.76	0.76
Polynomial	Lower	0.14	0.01	0.03	0.54	0.18	0.27
	Higher	0.50	0.11	0.10	1.00	0.88	0.77
Cd-As							
Multiplicative	Lower	0.15	0.00	0.00	0.59	0.04	0.01
	Higher	0.52	0.00	0.00	0.99	0.03	0.21
Polynomial	Lower	0.18	0.00	0.00	0.57	0.00	0.00
	Higher	0.52	0.01	0.02	0.99	0.57	0.39
Ni-Co							
Multiplicative	Lower	0.18	0.01	0.01	0.52	0.03	0.00
	Higher	0.54	0.00	0.02	0.97	0.05	0.05
Polynomial	Lower	0.16	0.00	0.00	0.53	0.00	0.00
	Higher	0.50	0.02	0.01	0.98	0.52	0.09

Table 4: False discovery rate of interactions in all scenarios with two-way interactions between exposures.

		Small (n=252)		**Large (n=1000)**	
Interaction type	Effect size	BKMR	BSR	BKMR	BSR
Hg-Ni					
Multiplicative	Lower	0.0011	0.0030	0.0034	0.0125
	Higher	0.0009	0.0032	0.0064	0.0180
Polynomial	Lower	0.0014	0.0014	0.0034	0.0148
	Higher	0.0020	0.0020	0.0048	0.0159
Cd-As					
Multiplicative	Lower	0.0011	0.0032	0.0039	0.0123
	Higher	0.0002	0.0018	0.0034	0.0168
Polynomial	Lower	0.0025	0.0016	0.0036	0.0136
	Higher	0.0018	0.0020	0.0059	0.0150
Ni-Co					
Multiplicative	Lower	0.0009	0.0023	0.0025	0.0120
	Higher	0.0011	0.0030	0.0016	0.0173
Polynomial	Lower	0.0020	0.0034	0.0036	0.0141
	Higher	0.0011	0.0009	0.0025	0.0150

Table 5: Sensitivity to trivariate interactions between Hg, Ni, and Tl.

		Small (n=252)			**Large (n=1000)**		
Interaction type	Effect size	Oracle	BKMR	BSR	Oracle	BKMR	BSR
Multiplicative	Lower	0.59	0.01	0	0.80	0.01	0.00
	Higher	0.69	0.02	0	0.98	0.01	0.01
Polynomial	Lower	0.55	0.01	0	0.81	0.01	0.00
	Higher	0.69	0.01	0	0.98	0.13	0.00

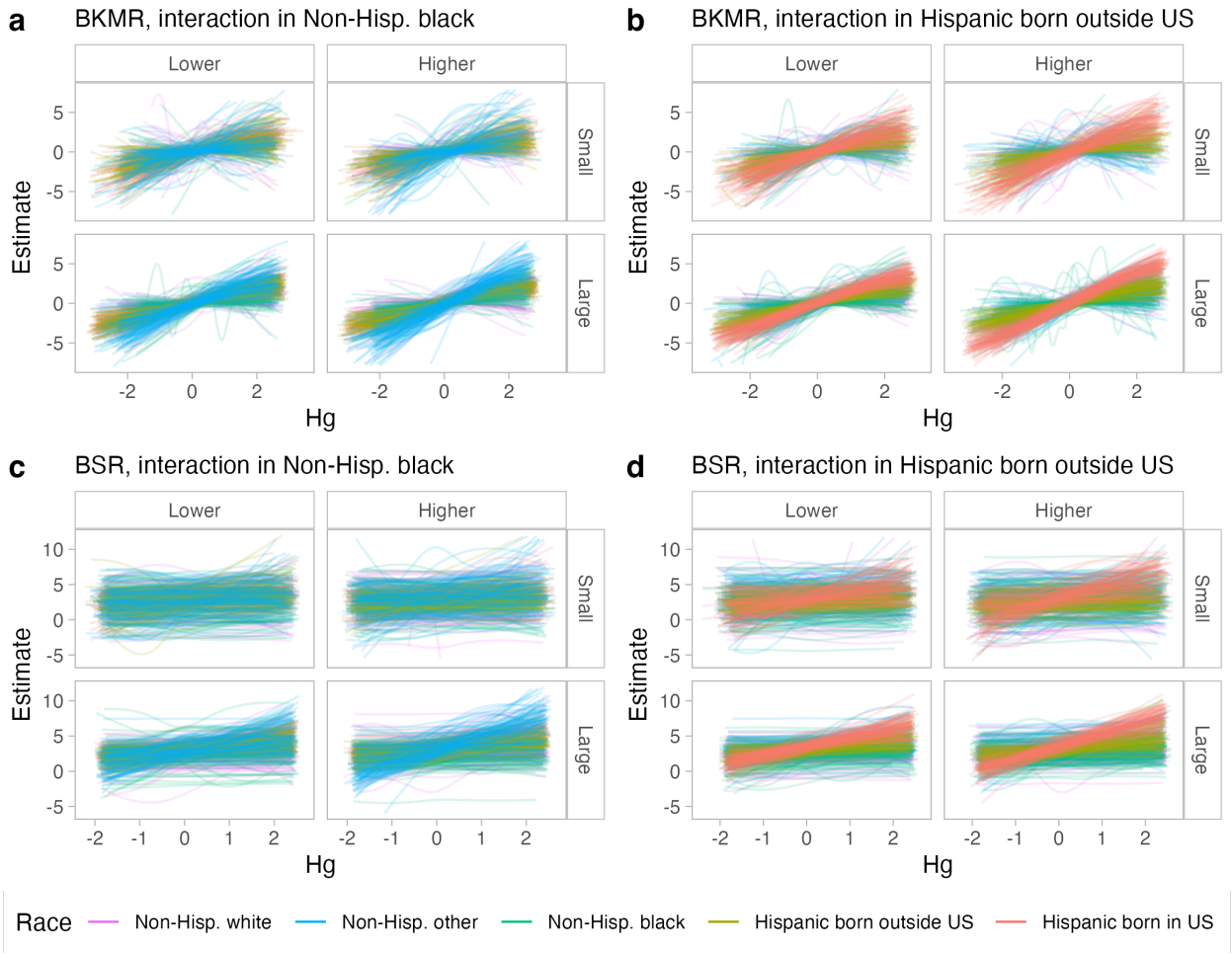

Figure 8: diagnostic plot of bkmr, bsr RE.

Table 6: Sensitivity to interactions between the categorical race variable and Hg.

		Small (n=250)			**Large (n=1000)**	
		Uncollapsed		Collapsed*		
Interaction in	Effect size	Oracle	BKMR	BKMR	Oracle	BKMR
non-Hispanic black†	Lower	0.07	0.00	0.00	0.21	0.01
	Higher	0.19	0.00	0.00	0.51	0.03
Hispanic born outside US‡	Lower	0.12	0.00	0.00	0.39	0.03
	Higher	0.24	0.02	0.03	0.83	0.21

* "Collapsed" refers to scenarios where the smallest three race categories are collapsed into one stratified model.
† Original n=27
‡ Original n=109

Table 7: Run-times.

			Multiplicative		**Polynomial**	
Model	Sample size	Base	Lower	Higher	Lower	Higher
Naive	Small	0.0032 s	0.0016 s	0.002 s	0.0016 s	0.0017 s
	Large	0.002 s	0.002 s	0.0035 s	0.0031 s	0.0025 s
Oracle	Small	0.0015 s	0.0015 s	0.0015 s	0.0019 s	0.0016 s
	Large	0.0044 s	0.0018 s	0.0018 s	0.0019 s	0.0024 s
BKMR	Small	13.56 m	15.21 m	15.51 m	15.20 m	15.17 m
	Large	1.64 h	1.68 h	1.62 h	1.64 h	1.59 h
BSR df	Small	30.11 m	33.15 m	32.92 m	33.81 m	35.11 m
	Large	57.22 m	58.02 m	1.06 h	57.75 m	1.05 h
BSR mod	Small	1.37 h	1.36 h	1.42 h	1.44 h	1.49 h
	Large	2.92 h	2.81 h	3.06 h	2.87 h	3.04 h

While BKMR and BSR are capable of detecting

In particular, the flexibility of BKMR and BSR mean that it is not possible to directly conduct inference on the effect sizes of specific chemicals

This section highlights the difficulty of using current methods for detecting

Run-time analysis

Discussion

Discuss results, link back to previous sim studies.

[*remember to label figures in appendix + add code*]

Table 8: Run-times race ethnicity. "NH Bl" and "H b/o US" represent scenarios in which the effect of Hg is increased in Non-Hisp. black and Hispanic born outside US categories, respectively.

		Small (n=252)		Large (n=1000)	
Model	Race	NH Bl	H b/o US	NH Bl	H b/o US
Naive	-	0.0027 s	0.0015 s	0.0034 s	0.0037 s
Oracle	-	0.0016 s	0.0023 s	0.0026 s	0.0018 s
BKMR	Non-Hispanic white	2.55 m	2.44 m	3.22 m	3.17 m
BKMR	Non-Hispanic black	4.01 m	4.00 m	4.25 m	4.23 m
BKMR	Non-Hispanic other	1.51 m	1.32 m	3.34 m	3.34 m
BKMR	Hispanic born in US	5.38 m	5.36 m	29.36 m	29.11 m
BKMR	Hispanic born outside US	5.64 m	5.65 m	53.43 m	50.97 m
BKMR	Collapsed non-Hispanic	4.15 m	4.19 m	-	-

[a] Original sample sizes of race category are as follows: Non-Hispanic white (n=16), Non-Hispanic black (n=27), Non-Hispanic other (n=13), Hispanic born in US (n=87), Hispanic born outside US (n=109), and Collapsed non-Hispanic (n=56=16+27+13)