Performance of variable and function selection methods for estimating the nonlinear health effects of correlated chemical mixtures: A simulation study

Nina Lazarevic1 | Luke D. Knibbs1 | Peter D. Sly2 | Adrian G. Barnett3

1School of Public Health, Faculty of Medicine, The University of Queensland, Herston, Queensland, Australia
2Children’s Health and Environment Program, Child Health Research Centre, The University of Queensland, South Brisbane, Queensland, Australia
3School of Public Health and Social Work, Faculty of Health, Queensland University of Technology, Kelvin Grove, Queensland, Australia

Correspondence
Nina Lazarevic, School of Public Health, Faculty of Medicine, The University of Queensland, Herston, QLD 4006, Australia.
Email: nina.lazarevic@uqconnect.edu.au

Funding information
Australian Government Research Training Program Scholarship; Australian National Health and Medical Research Council Senior Research Fellowship, Grant/Award Number: APP1117784; Australian National Health and Medical Research Council Project Grant/Award Number: APP1142222

Statistical methods for identifying harmful chemicals in a correlated mixture often assume linearity in exposure-response relationships. Nonmonotonic relationships are increasingly recognized (e.g., for endocrine-disrupting chemicals); however, the impact of nonmonotonicity on exposure selection has not been evaluated. In a simulation study, we assessed the performance of Bayesian kernel machine regression (BKMR), Bayesian additive regression trees (BART), Bayesian structured additive regression with spike-slab priors (BSTARSS), generalized additive models with double penalty (GAMDP) and thin plate shrinkage smoothers (GAMTS), multivariate adaptive regression splines (MARS), and lasso penalized regression. We simulated realistic exposure data based on pregnancy exposure to 17 phthalates and phenols in the US National Health and Nutrition Examination Survey using a multivariate copula. We simulated data sets of size N = 250 and compared methods across 32 scenarios, varying by model size and sparsity, signal-to-noise ratio, correlation structure, and exposure-response relationship shapes. We compared methods in terms of their sensitivity, specificity, and estimation accuracy. In most scenarios, BKMR, BSTARSS, GAMDP, and GAMTS achieved moderate to high sensitivity (0.52-0.98) and specificity (0.21-0.99). BART and MARS achieved high specificity (≥0.90), but low sensitivity in low signal-to-noise ratio scenarios (0.20-0.51). Lasso was highly sensitive (0.71-0.99), except for quadratic relationships (≤0.27). Penalized regression methods that assume linearity, such as lasso, may not be suitable for studies of environmental chemicals hypothesized to have nonmonotonic relationships with outcomes. Instead, BKMR, BSTARSS, GAMDP, and GAMTS are attractive methods for flexibly estimating the shapes of exposure-response relationships and selecting among correlated exposures.

KEYWORDS
chemical mixtures, endocrine-disrupting chemicals, multipollutant, nonmonotonicity, variable selection
1 | INTRODUCTION

Epidemiological studies of environmental exposures are increasingly focusing on the health effects of chemical mixtures, rather than individual chemicals. Mixtures analyses provide better representation of real-world exposure patterns and enable the adjustment of confounding by co-existing toxicants. However, chemical components are often highly correlated (eg, traffic-related pollutants), which may lead to instability in effect estimates and inflated standard errors in generalized linear models. This can be exacerbated by the inclusion of multiple nonlinear terms (eg, polynomials) when exposure-response relationships are nonlinear. These issues present challenges for the statistical identification of important mixture components, that is, harmful chemicals driving the association between a mixture and health outcome.

Methods for variable selection among multiple correlated exposures have been the subject of recent reviews, including penalized regression methods (eg, lasso and elastic net regression), dimension reduction methods (eg, sparse partial least squares and supervised principal components analysis), regression tree ensemble methods, and others (eg, weighted quantile sum regression). These methods produce a sparse solution, which increases interpretability; however, there is no assurance that the selected exposures are etiologic agents or that the excluded exposures are safe, particularly when exposures are highly correlated.

Recent simulation studies have characterized the sensitivity and rate of false discoveries of variable selection methods in the exposome context (ie, the totality of environmental exposures throughout the life-course) and in assessing interactions between chemical exposures. These studies have assumed that exposure-response relationships are linear. However, nonmonotonic relationships are biologically plausible and have been observed for metals with dose-dependent effects (eg, manganese is an essential nutrient at physiologic levels but also an environmental toxicant) and endocrine-disrupting chemicals (EDCs) that mimic the nonmonotonic effects of endogenous hormones on endocrine outcomes. The impact of nonmonotonicity on the performance of exposure selection methods has not been evaluated.

Standard implementations of many variable selection methods do not accommodate the simultaneous selection of groups of terms associated with an exposure, for example, spline basis functions. Therefore, methods for variable selection as well as function selection are required for the identification of important mixture components when exposure-response relationships are nonlinear. This can be performed, for example, through the use of both sparseness and smoothness penalties in maximum likelihood-based approaches (eg, the group lasso) and through selection indicators or spike-slab priors in Bayesian models. Alternatively, nonparametric methods may be used to estimate a multivariate exposure-response function.

In this study we compared the performance of six variable selection methods that can model nonlinear exposure-response relationships: Bayesian kernel machine regression (BKMR), Bayesian additive regression trees (BART), Bayesian structured additive regression with spike-slab priors (BSTARSS), generalized additive models with variable selection based on a double penalty (GAMDP) and thin plate regression spline shrinkage smoothers (GAMTS), and multivariate adaptive regression splines (MARS). We assessed the ability of each method to distinguish outcome-associated exposures and to reveal the shapes of both monotonic and nonmonotonic exposure-response relationships when varying the exposure correlation structure. We evaluated whether there was (1) an advantage to the use of these methods when exposure-response relationships are nonlinear, over lasso penalized regression assuming linearity, and (2) a cost associated with their use when exposure-response relationships are linear. Building on our previous work, we focused on exposure to EDC-mixtures during pregnancy, as this is a critical period of developmental sensitivity to environmental chemicals.

2 | METHODS

2.1 | NHANES data

We based our simulation study on environmental chemical data from 214 women with positive urinary pregnancy tests in the US National Health and Nutrition Examination Survey (NHANES). We merged data on urinary chemical concentrations measured in the biennial phthalates, phenols, and parabens surveys between 2005 and 2014 (Tables S1 and S2, Supplementary Material). Ten phthalate metabolites and seven phenols were retained that had fewer than 20% of
participants with measurements below the limit of detection (LOD) (Tables S1 and S2), and observations below this limit were assigned a value of the LOD/√2.18 Exposures were corrected for urinary creatinine levels (to account for urine dilution) and expressed in μg/g creatinine, then natural log-transformed and standardized (to keep parameters scale-free and preempt numerical accuracy issues).

2.2 Simulating exposure data

We simulated correlated exposure data using a multivariate t-copula with truncated kernel-smoothed empirical marginal distributions and the observed Spearman’s rank-order correlation structure from the NHANES data. We fitted a selection of multivariate copula types by maximum likelihood and chose the t-copula as it had the highest maximum likelihood compared with the Gaussian, Gumbel, Frank, Clayton, and Joe copulas. This was performed using the R packages copula (version 0.999-18) and copulaedas (version 1.4.2).19,20 The use of a multivariate copula enables the separate specification of the marginal exposure distributions and the dependence structure (ie, the correlation between exposures),20 which allowed us to also simulate a low correlation data set using half the observed Spearman correlation matrix and the same marginal distributions. This enabled us to assess whether the level of correlation affects method performance, using simulated data based on the empirical exposure distributions from the NHANES. The linearity of the dependence structure was first verified through maximal information-based nonparametric exploration statistics21 (Figure S1, Supplementary Material). We assessed fit graphically using scatterplots and nonparametric kernel density estimates of the original and simulated data (Figures S2-S5, Supplementary Material). Correlation heat maps are shown in Figure 1.

2.3 Exposure-response relationships and data-generating processes

Our data-generating processes differed by model size, exposure-response functions \( f_j \) \((j = 1, \ldots, J)\), the degree of correlation between exposures, and the signal-to-noise ratio. We specified two model sizes of \( J = 8 \) and \( J = 17 \) exposure variables \( x_{ij} \) \((i = 1, \ldots, N)\). A subset of exposures, \( J^* = 4 \), were assumed to be associated with the outcome; hence, we consider a “low-sparsity” setting in the \( J = 8 \) model and a “high-sparsity” setting in the \( J = 17 \) model. The response \( y_i \) for individual \( i \) was generated from the model

\[
y_i = \sum_{j=1}^{J^*} f_j(x_{ij}) + e_i,
\]

where \( e_i \sim N(0, \sigma^2) \).

For simplicity, we assumed no confounding by nonexposure variables, no interaction, and that only four phenols were associated with the outcome (ie, methylparaben [MPB], propylparaben [PPB], benzophenone-3 [BP3], and bisphenol A [BPA]). The remaining exposures were assumed not to be associated with the outcome in the \( J = 17 \) model, and the four exposures most correlated with the outcome-associated exposures were included in the \( J = 8 \) model (ie, mono-ethyl phthalate [MEP], mono-(2-ethyl-5-hydroxyhexyl) phthalate [MHH], mono(carboxynonyl) phthalate [CNP], and mono-(2-ethyl-5-oxohexyl) phthalate [MOH]).

We specified four exposure-response functions with two association strengths, shown in Figure 2: linear, nonlinear monotonic (S-shaped; using the log-logistic cumulative distribution function),22 nonmonotonic symmetric (inverse-U-shaped; using a quadratic function), and nonmonotonic asymmetric (skewed inverse-U-shaped; using a Dawson function).23 We specified two linear exposure-response association strengths, \( f_j(x_{ij}) = \beta_j x_{ij} \) with \( \beta_j = 2 \) for MBP and BP3 and \( \beta_j = 1 \) for PPB and BPA. To keep a constant association strength across functions, we scaled each function to have the same area under the curve as the linear function.

We assumed two signal-to-noise ratios, adjusting \( \sigma \) so that \( R^2 \) for the true model corresponded to 10% (“low” signal-to-noise ratio) and 30% (“high” signal-to-noise ratio).

The four exposure-response functions, two levels of exposure correlation, and two signal-to-noise ratios, gave 16 data-generating processes. We simulated data sets of sample size \( N = 250 \), replicating each data-generating process 100
times to give 1600 data sets. The sample size of 250 was based on our systematic review of studies assessing prenatal exposure to mixtures of EDCs where the interquartile range in sample sizes was 179 to 488.4 We estimated models based on the two model size and sparsity settings, producing 32 simulation scenarios.

2.4 Statistical methods

We compared six methods for identifying important mixture components and modeling nonlinear exposure-response relationships, to lasso penalized regression assuming linearity.
2.4.1 Bayesian kernel machine regression

BKMR is an approach for mixtures analyses that provides flexible estimation of a multivariate exposure-response function, represented by a Gaussian kernel machine.\textsuperscript{24} Expressed as a mixed model, BKMR with componentwise variable selection is specified as follows\textsuperscript{24,25}:

\[
y_i = h_i + z'_i \beta + e_i, \quad e_i \sim N(0, \sigma^2),
\]
\[
h = (h_1, \ldots, h_N)' \sim N(0, \tau K),
\]

where \(y_i\) is the response for individual \(i (i = 1, \ldots, N)\), \(h\) is a vector of subject-specific health effects \(h_i = h(x_i)\) with \(h(\cdot)\) representing a multivariate exposure-response function, \(x_i = (x_{i1}, \ldots, x_{ij})\) is the vector of \(J\) exposures for individual \(i\), and \(z_i\) is a vector of potential confounders. \(K\) is an \(N \times N\) kernel matrix, with \((i, k)\)-elements specified by the augmented Gaussian kernel function

\[
K(x_i, x_k; r) = \exp \left( -\sum_{j=1}^{J} r_j (x_{ij} - x_{kj})^2 \right),
\]

where \(r = (r_1, \ldots, r_J)'\) is a vector of parameters \(r_j\) that control the smoothness of \(h(\cdot)\), for which a spike-slab prior is assumed:

\[
r_j \mid \delta_j \sim (1 - \delta_j)P_0 + \delta_j \Gamma(a_r, b_r),
\]
\[
\delta_j \mid \pi \sim \text{Bernoulli}(\pi),
\]
\[
\pi \sim \text{Beta}(a_\pi, b_\pi).
\]
P₀ is the spike density with point mass at zero and a gamma distribution is used for the slab component. Here δⱼ are variable selection indicators with prior probability π. The posterior mean of δⱼ is the posterior inclusion probability of exposure j, that is, a measure of the importance of exposure j. The model is estimated by Markov chain Monte Carlo (MCMC), using the Metropolis-Hastings algorithm for r and λ = rσ⁻² (a convenient reparameterization), and a Gibbs sampler for the remaining parameters.²⁴

We used the bkmr package in R, version 0.2.0, and based our prior specifications on the default implementation.²⁴ We used a threshold of 0.5 on the posterior inclusion probabilities for variable selection.²⁶ We assigned π a Beta(1, 1) prior, so that the prior probability of variable inclusion is 0.5. Prior distributions for σ⁻² and λ were assumed to be Gamma with parameters (shape, rate) set to (aₕ,bₕ) = (0.001,0.001) and (aₗ,bₗ) = (1, 1), respectively. For the slab component of the prior on rⱼ, we specified a Gamma prior with mean and SD of 0.25, that is, (aₗ,bₗ) = (1, 4); these values were chosen by fitting frequentist kernel machine regression and observing which values of rⱼ produced appropriate levels of smoothing.²⁴ We chose tuning parameters that produced adequate acceptance rates in the Metropolis-Hastings steps (around 20% to 40%), that is, SDs of the gamma proposal distributions for λ of 0.5 and 1 for the low and high signal-to-noise ratio data sets, respectively, and for rⱼ of 0.1 for both the switching and refinement steps (except for the monotonic function and high signal-to-noise ratio data sets, which required a 0.2 SD in the refinement step). We ran the MCMC sampler for 10 000 iterations and discarded the first 8000 iterations. Convergence diagnostics are in Section 2 of the Supplementary Material. We assessed the sensitivity of our results to prior specification in Section 3 of the Supplementary Material.

2.4.2 Bayesian additive regression trees

BART is a nonparametric ensemble method, which models an outcome using a sum of regression trees.²⁷ BART flexibly captures nonlinearity and interactions, and imposes no assumptions on the functional forms of exposure-response relationships.²⁷ BART produces a measure of variable importance by tracking variable inclusion proportions, which enables variable selection with a user-defined threshold.²⁸ As BART is defined by a Bayesian statistical model, full posterior inference is possible, including exposure effect estimates and credible intervals.²⁷ The model is

\[ yᵢ = h(xᵢ) + eᵢ, \quad eᵢ \sim N(0, \sigma^2), \]

where yᵢ is the response and xᵢ = (xᵢ₁, …, xᵢⱼ)’ is a vector of J exposures for individual i (i = 1, …, N). The multivariate exposure-response function h(xᵢ) is approximated by a sum of K regression trees,

\[ h(xᵢ) \approx \sum_{k=1}^{K} g(xᵢ; T_k, M_k). \]

Here Tₖ is the kth regression tree with terminal node (ie, “leaf”) parameters Mₖ = {µ₁ₖ, …, µₙₖ}, for nₖ terminal nodes, and the function g(·) assigns µᵢₖ ∈ Mₖ to xᵢ.²⁷ A tree Tₖ consists of nonterminal decision rules (ie, binary splits of the form {xⱼ ≤ c} or {xⱼ > c} for given splitting variables xⱼ and splitting values c), and the set of terminal nodes.²⁷ Following a sequence of decision rules, each observation is assigned the leaf value µᵢₖ (l = 1, …, nₖ) of the terminal node. The fitted value \( \hat{y}_i = E(y_i|x_i) \) is then the sum of the K leaf parameters µᵢₖ assigned to observation i.²⁷

Individual trees are constrained via a regularization prior; each tree explains a different small portion of h(xᵢ) and the prior ensures that no individual tree is overly influential.²⁷ The regularization prior is composed of priors on the tree structure, leaf parameters, and error variance σ², which is assumed independent²⁷²⁹.\(^8\)

\[ P(T₁, M₁, …, Tₖ, Mₖ, σ²) = \left[ \sum_k \sum_l P(µₖ | Tₖ)P(Tₖ) \right] P(σ²) \]

P(Tₖ) takes a form which favors shallow tree structures with fewer splits,²⁹ with the probability that a depth d node is nonterminal of a (1 + d)⁻β, a ∈ (0, 1) and β ∈ [0, ∞).²⁷ We used the recommended default values for the hyperparameters α and β, of α = 0.95 and β = 2, which keeps individual trees small (ie, greatest probability on trees with 2 or 3 terminal nodes).²⁷ To complete the specification of P(Tₖ), a uniform prior is placed on the choice of splitting variable at each node, and a discrete uniform prior is specified for the splitting values.²⁷
For $P(\mu_k | T_k)$, a conjugate normal prior is used $\mu_k \sim \text{iid } N(\mu, \sigma_k^2)$, where $\mu$ is the center of the response range and $\sigma_k^2$ is selected so that the response range center $\pm n \times 2$ standard deviations corresponds approximately to 95% coverage of the observed response values. This shrinks the leaf parameters towards the response distribution center, weakening individual trees.

For $P(\sigma^2)$, a conjugate inverse Gamma prior is used, $\sigma^2 \sim \Gamma^{-1}(\nu, \lambda)$. The hyperparameter $\lambda$ was calibrated by first obtaining a data-based estimate $\hat{\sigma}^2$, then setting $\lambda$ such that $P(\sigma < \hat{\sigma}) = q$, that is, a larger quantile $q$ places more weight on values lower than $\hat{\sigma}$, and setting $\nu$ to obtain an appropriate shape. We chose $K = 50$ for the number of trees; while larger K have also been recommended, smaller K are preferred for variable selection.

We used the bartMachine package in R, version 1.2.4. The package allows hyperparameters to be chosen empirically using $k$-fold cross-validation; however, this involves a substantial computational burden not feasible in our simulation study. We assessed the sensitivity of our results to prior specification in Section 3 of the Supplementary Material.

BART is estimated by a Bayesian backfitting MCMC algorithm. We set the prior probabilities for proposing grow, prune, and change steps to $(0.2, 0.6, 0.2)$, to achieve adequate acceptance rates. We generated 2000 draws from the posterior after a burn-in of 4000 draws. Convergence diagnostics are in Section 2 of the Supplementary Material. For variable selection, we chose the local threshold as it gives the least sparse solutions.

### 2.4.3 Bayesian structured additive regression with spike-slab priors

Structured additive regression (STAR) is a flexible modeling framework that allows exposures to be modeled with arbitrary combinations of smooth interactions, random effects, spatial effects, and varying coefficient terms. Generalized additive models (GAMs) and generalized additive mixed models are special cases of STAR models. BSTARRS extends these models through specification of priors for penalized regression and variable selection, as well as function and model selection, that is, allowing both individual terms and groups of terms associated with an exposure to be selected or deselected. This enables the simultaneous identification of important exposures, their interactions, and flexible estimation of the shapes of exposure-response relationships. Importantly, BSTARRS differentiates between exposures with no effect, linear effects, and nonlinear effects. The model is:

\[ E(y_i | x_{i1}, \ldots, x_{iJ}) = h(\eta_i), \quad e_i \sim N(0, \sigma^2), \]

\[ \sigma^2 \sim \Gamma^{-1}(a, b), \]

where $y_i$ is an exponential-family distributed response for individual $i$ ($i = 1, \ldots, N$) with exposures $x_{ij}$ ($j = 1, \ldots, J$), $h(\cdot)$ is a known generalized linear model link function, and

\[ \eta_i = \eta_0 + z_i^T \delta + \sum_{j=1}^J f_j(x_i). \]

Here $\eta_0$ is an optional offset and $z_i$ includes terms not subject to selection (such as known linear confounders and a global intercept), with coefficients $\delta$ that are given a flat prior.

The terms $f_j(x_i)$ can be linear, factors, smooth functions of one or multiple exposures (eg, splines, tensor products, varying coefficients), random effects, Markov random fields, or interactions between terms. Smooth functions $f_j(x)$ can be represented by a linear combination of $d_j$ basis functions $B_j(\cdot)$, so that

\[ f_j(x) = \sum_{k=1}^{d_j} \beta_{j_k} B_{jk}(x) = B_j \beta_j, \]

where $B_{jk}(x) = (B_{jk}(x_1), \ldots, B_{jk}(x_N))'$. The prior specification assumes that model terms have been reparameterized to separate their penalized and unpenalized parts. The coefficient group $\beta_{j}$, of length $d_j$, is given a parameter-expanded
Normal mixture of inverse Gamma distributions prior, denoted by peNMIG(·)\(^{30}\):

\[
\beta_j = a_j \xi_j \sim \text{peNMIG}(\nu_0, w, a_\tau, b_\tau).
\]

The prior uses a multiplicative parameter expansion \(\beta_j = a_j \xi_j\) that enables simultaneous selection of groups of coefficients, with \(a_j\) representing the importance of a coefficient group \(\beta_j\) and \(\xi_j\) distributing the importance across the elements of \(\beta_j\). \(^{30}\) Each \(\xi_{jk}\) is given an iid Normal prior, with a mean of either 1 or \(-1\) with equal probability\(^{30}\):

\[
\xi_{jk} \mid m_{jk} \sim \text{iid } N(m_{jk}, 1)
\]

\[
m_{jk} \sim \frac{1}{2} I_1(m_{jk}) + \frac{1}{2} I_{-1}(m_{jk})
\]

The prior structure for \(a_j\) is\(^{30,32}\):

\[
a_j \mid \gamma_j, \tau_j^2 \sim N(0, \gamma_j \tau_j^2),
\]

\[
\gamma_j \mid w \sim w I_1(\gamma_j) + (1 - w) I_0(\gamma_j),
\]

\[
\tau_j^2 \sim \Gamma^{-1}(a_\tau, b_\tau),
\]

\[
w \sim \text{Beta}(a_w, b_w).
\]

Here \(\gamma_j\) is an indicator variable taking the value 1 with probability \(w\) and a small value \(v_0\) with probability \((1 - w)\). The hypervariance \(\tau_j^2\) follows an inverse Gamma distribution with shape and rate parameters \((a_\tau, b_\tau)\), \(a_\tau \ll b_\tau\). \(^{30}\) The prior variance \(\nu_j^2 = \gamma_j \tau_j^2\) is a bimodal mixture of inverse Gamma distributions with spike at \(\nu_j = v_0\) (scale \(v_0 b_\tau\)) and slab at \(\nu_j = 1\) (scale \(b_\tau\)). \(^{30}\) The spike part of the prior strongly shrinks coefficients towards zero if \(v_0\) is sufficiently small, and its posterior probability gives the probability of exclusion of \(\beta_j\) and \(f_j(x)\) from the model. \(^{30}\) The Beta prior on \(w\) incorporates prior knowledge on the sparsity of \(\beta\). \(^{30}\)

BSTARSS is implemented in the R package spikeSlabGAM, version 1.1-14.\(^{32}\) We set \(v_0 = 0.025\), \((a_\tau, b_\tau) = (5, 40)\), \((a_w, b_w) = (1, 1)\) (ie, a uniform prior on \(w\)), a vague prior on the error variance \((a_\sigma, b_\sigma) = (0.001, 0.001)\), and fit smooth effects using the default reduced-rank representation of 10 cubic B-spline basis functions with equidistant knots and second-order difference penalties. \(^{30}\) We ran five parallel chains, generating 2000 draws from each after a burn-in period of 8000 draws. For variable selection, we used a threshold of 0.5 on the posterior inclusion probability of any term associated with an exposure.\(^{26}\) Convergence diagnostics are in Section 2 of the Supplementary Material. We assessed the sensitivity of our results to prior specification in Section 3 of the Supplementary Material.

### 2.4.4 Generalized additive models that perform variable selection

GAMs are extensions of generalized linear models that allow smooth functions of covariates.\(^{33,34}\) The model is

\[
g\{E(y_i)\} = Z_i \theta + \sum_j f_j(x_{ij}),
\]

where \(y_i\) is an exponential-family distributed response with link function \(g\{\cdot\}\), \(f_j\) are smooth functions of the covariates \(x_j\), represented by spline bases, and \(Z_i\) contains linear terms with coefficient vector \(\theta\).\(^{33,34}\) GAMs are estimated by penalized maximum likelihood using the smoothing penalty

\[
\lambda_j \beta^T S_j \beta,
\]

where \(\beta^T S_j \beta\) measures the roughness of the \(f_j\), \(S_j\) is the penalty matrix, and \(\lambda_j\) are the smoothing parameters.\(^{33,34}\) An eigen decomposition of the penalty matrix

\[
S_j = U_j \Lambda_j U_j^T,
\]
where $U_j$ and $\Lambda_j$ are the eigenvector and eigenvalue matrices, respectively, will lack full rank if there are zero eigenvalues in $\Lambda_j$ associated with the penalty null space.$^{33}$ The smoothing penalty is therefore not usually able to remove exposures from a model, as it shrinks functions in the range space of a spline basis but not the null space (eg, it may penalize toward a straight line but not necessarily a flat line).$^{33}$ GAM with a double penalty (GAMDP) circumvents this by applying penalties on both the range space and null space

$$\lambda_j \beta^T S_j \beta + \lambda_j^* \beta^T U_j^* U_j^T \beta,$$

where $U_j^*$ are the eigenvectors corresponding to the zero eigenvalues of $\Lambda_j$, so that smooth terms may be removed from the model.$^{33}$

An alternative approach, GAM based on thin plate regression spline shrinkage smoothers (GAMTS), involves setting the zero eigenvalues of $\Lambda_j$ to a small value $\epsilon$, represented by $\tilde{\Lambda}_j$, so that the smoothing penalty becomes

$$\lambda_j \beta^T U_j \tilde{\Lambda}_j U_j^T \beta,$$

ensuring that the penalty is of full rank so that smooth terms may be removed from the model.$^{33}$ This is equivalent to setting $\lambda_j^* = \epsilon \lambda_j$ in the GAMDP approach.$^{33}$ The value $\epsilon$ is set to a small proportion of the positive eigenvalues of $\Lambda_j$, 0.1 by default.$^{33}$ Therefore, while GAMDP allows the degree of penalization to be determined from the data, GAMTS penalizes the null space to a lesser degree than the range space.$^{33}$

We used the R package `mgcv` (version 1.8-23),$^{34}$ specifying thin plate regression splines with a basis dimension of 10 in both approaches and restricted maximum likelihood smoothing parameter estimation. We assumed that an effective degrees of freedom of zero (at two decimal places) meant that an exposure was excluded from the model.$^{33}$

### 2.4.5 Multivariate adaptive regression splines

MARs is a nonparametric method that proceeds in a similar manner to stepwise regression, building a model from piecewise linear (or higher order) spline basis functions and their products.$^{35,36}$ The model is

$$\hat{y} = \beta_0 + \sum_m \beta_m h_m(X),$$

where $\hat{y}$ is the predicted response and $h_m$ is the $m$th basis function, or product of basis functions, with coefficient $\beta_m$. The basis functions take the form

$$(x_{ij} - t)_+ = \begin{cases} x_{ij} - t & \text{if } x_{ij} > t \\ 0 & \text{otherwise} \end{cases},$$

$$(t - x_{ij})_+ = \begin{cases} t - x_{ij} & \text{if } x_{ij} < t \\ 0 & \text{otherwise} \end{cases},$$

where $t$ is a knot location and $x_{ij}$ is the $j$th exposure for observation $i$.$^{35,36}$ Model building proceeds through a forward pass, entering basis functions in a hierarchical manner to build a large model that deliberately overfits the data, and subsequently a pruning pass in which a subset of basis functions are selected according to a generalized cross-validation or $k$-fold cross-validation criterion.$^{35,36}$ Following the pruning pass, an ordinary least squares (or generalized linear model) regression is used to obtain coefficient estimates.$^{37}$

We used the R package `earth` (version 4.6.0),$^{37}$ which we restricted to estimate an additive model, and performed 30 repetitions of 10-fold cross-validation. We estimated prediction intervals using a linear variance model and estimated relative exposure importance by comparing the number of model subsets in the pruning pass that include each exposure.$^{37}$

### 2.4.6 Lasso penalized regression

Lasso performs variable selection by shrinking the coefficients of exposure variables toward zero. Standard implementations of lasso assume linearity, and groups of terms associated with an exposure cannot be selected or dropped
simultaneously. When exposure-response relationships are linear, lasso has achieved comparable or superior performance in terms of sensitivity and specificity to competing variable selection methods. Lasso produces a family of solutions, parameterized by a tuning parameter. Selection of a solution is usually conducted via k-fold cross-validation; we used 10-fold cross-validation in the R package glmnet, version 2.0-13, and kept the fold assignment constant within each replication of the simulation.

### 2.5 Comparison of methods

We compared methods by the average sensitivity and specificity across the 100 replications, defined as the proportion of outcome-associated and outcome-unassociated exposures, respectively, that were correctly identified. We also measured the positive predictive value (defined as the proportion of selected exposures that were true positives) and the negative predictive value (defined as the proportion of nonselected exposures that were true negatives). We calculated the \( F_1 \)-statistic = \( 2 \times \) positive predictive value \( \times \) sensitivity/(positive predictive value + sensitivity), which reflects the ability of a method to detect outcome-associated exposures while avoiding the selection of unassociated exposures.

To further assess variable selection, we measured the proportion of replications in which all the outcome-associated exposures were ranked higher than the outcome-unassociated exposures. This was in terms of the posterior inclusion probabilities for BKMR and BSTARSS, variable inclusion proportions for BART, number of model subsets that include each exposure in MARS, and the effective degrees of freedom for GAMDP and GAMTS (which measures relationship complexity rather than variable importance). Additionally, we measured the mean proportion of outcome-associated exposures that were ranked higher than outcome-unassociated exposures across replications.

To assess estimation accuracy for outcome-associated exposures, we compared the estimated posterior mean (averaged over the post-burn-in MCMC samples) evaluated at the 25th, 50th, and 75th percentiles of each exposure (holding other exposures at their means), to the value of the simulated exposure-response curve using the mean-squared error (defined as the average of the squared differences across 100 replications), and the 90% credible interval coverage (defined as the proportion of times the true value was contained in the 90% credible interval). For MARS, we used predicted values and prediction intervals, rather than posterior means and credible intervals. We compared the ability of each method to estimate the shapes of the exposure-response curves graphically, by plotting posterior means (or predicted values for MARS) evaluated at every 10th percentile of one exposure distribution while holding other exposures at their means. As an oracle (ie, benchmark) method, we fitted standard GAMs to the true model of four phenols (specified as outlined in Section 2.4.4).

### 3 RESULTS

#### 3.1 Variable selection

##### 3.1.1 Sensitivity and specificity

Mean sensitivity and specificity for each method and simulation scenario are shown in Figure 3 (and Table S3, Supplementary Material). BKMR, BSTARSS, GAMDP, and GAMTS achieved moderate to high sensitivity and specificity across scenarios (BKMR 0.55 to 0.87 sensitivity and 0.57 to 0.99 specificity; BSTARSS 0.52 to 0.95 sensitivity and 0.21 to 0.99 specificity; GAMDP 0.75 to 0.98 sensitivity and 0.56 to 0.69 specificity; GAMTS 0.66 to 0.98 sensitivity and 0.61 to 0.75 specificity), with the exception of lower specificity for BSTARSS in low signal-to-noise ratio scenarios of S-shaped relationships. BART and MARS were highly specific (0.95 to 1.00 and 0.90 to 0.99, respectively), but markedly less sensitive than the other methods in the low signal-to-noise ratio (MARS 0.21 to 0.43) and low sparsity settings (BART 0.20 to 0.57).

When exposure-response relationships were linear, lasso achieved high sensitivity (0.83 to 0.99) but lower specificity (0.45 to 0.72) than most other methods. Lasso was competitive with BSTARSS, BKMR, GAMDP, and GAMTS in terms of sensitivity and specificity for S-shaped and asymmetric inverse-U-shaped exposure-response relationships (0.71 to 0.99 sensitivity and 0.43 to 0.76 specificity). However, lasso had very low sensitivity for quadratic relationships (0.12 to 0.27).
The sensitivity and specificity of BART, BSTARSS, and MARS were robust to changes in the correlation structure. BKMR was slightly more sensitive when halving the correlation between chemicals, whereas GAMDP and GAMTS were slightly less specific.

In almost all cases, increasing the signal-to-noise ratio improved sensitivity. Specificity was also improved for BKMR and BSTARSS, but was minimally affected by the signal-to-noise ratio for BART, GAMDP, GAMTS, MARS, and lasso.

### 3.1.2 Positive and negative predictive value

For BKMR and BSTARSS, we observed moderate to high positive predictive value (0.42 to 0.97) and high negative predictive value (0.70 to 0.97) in most scenarios (Figure 4 and Table S4, Supplementary Material). GAMDP and GAMTS had moderate positive predictive value in high sparsity scenarios (0.41 to 0.55) and higher positive predictive value in low sparsity scenarios (0.68 to 0.77), while negative predictive value was high overall (0.70 to 0.99). This was also the case for lasso when exposure-response relationships were monotonic or asymmetric inverse-U-shaped, but both positive and negative predictive value were lower when relationships were quadratic. In most scenarios, BART and MARS had higher positive predictive value (0.73 to 1.00) than the other methods, suggesting that the few exposures selected by BART and MARS were likely to be true positives (Table S4). BART and MARS had lower negative predictive value than the other methods in most low sparsity scenarios (0.55 to 0.85).

Overall, for the methods that can model nonlinearity, the positive and negative predictive values appeared minimally affected by the shape of the exposure-response relationships and the exposure correlation structure. For BKMR and BSTARSS, the signal-to-noise ratio had a strong positive impact on positive predictive value in the high sparsity scenarios (low signal-to-noise ratio scenario positive predictive value 0.42 to 0.83 and high signal-to-noise ratio scenario positive predictive value 0.73 to 0.97). In the low sparsity scenarios, higher signal-to-noise ratio tended to improve the negative predictive value to a greater degree than the positive predictive value for MARS, GAMDP, and GAMTS.
3.1.3 | F1-statistic

Considering the F1-statistic (Figure 5 and Table S3, Supplementary Material), a balanced measure of sensitivity and positive predictive value, BKMR, BSTARSS, GAMDP, and GAMTS were competitive in terms of mean and interquartile range of F1. BART and MARS had the least favorable performance in terms of F1 in the low sparsity scenarios, but had comparable performance to the other methods in the high sparsity scenarios.

Lasso tended to perform comparably with the methods that can model nonlinearity when exposure-response relationships were monotonic or asymmetric inverse-U-shaped. However, lasso consistently performed poorly with a low F1 when exposure-response relationships were quadratic.

3.1.4 | Ranking of exposures

BSTARSS had the highest proportion of replications in which the outcome-associated exposures were all ranked higher than the outcome-unassociated exposures in 22 of 32 scenarios (Figure 6 and Table S5, Supplementary Material), although BSTARSS was tied with or closely followed by BKMR, BART, and GAMDP, in most cases. This proportion appeared robust to changes in the exposure-response relationship and the exposure correlation structure. However, with the exception of MARS, the proportion was highly affected by the signal-to-noise ratio and sparsity setting: only 1% to 16% of replications yielded the correct ranking across any method in the high sparsity and low signal-to-noise ratio scenarios, whereas 36% to 79% did so in the low sparsity and high signal-to-noise ratio scenarios. MARS had the worst performance in almost all cases.

For the mean proportion of outcome-associated exposures that were ranked above outcome-unassociated exposures across replications (Figure 7 and Table S5), most outcome-associated exposures were ranked higher by all methods in the high signal-to-noise ratio and low sparsity scenarios (mean proportions 0.62 to 0.93). This was also the case in the high signal-to-noise ratio and high sparsity scenarios, except for GAMDP when relationships were linear and S-shaped (mean proportions 0.34 to 0.50). In the low signal-to-noise ratio scenarios, BSTARSS achieved the highest mean proportions (0.50
**FIGURE 5** Box plots of the F1-statistic by scenario and method. Box plots show the median value (dash) and interquartile range (IQR), with whiskers at ±1.5*IQR. Mean values denoted by diamonds [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 6** Proportion of replications in which all outcome-associated exposures were ranked above outcome-unassociated exposures (in terms of the posterior inclusion probabilities for BKMR and BSTARSS, variable inclusion proportions for BART, effective degrees of freedom for GAMDP and GAMTS, and number of model subsets in the pruning pass that include each exposure for MARS), by scenario [Colour figure can be viewed at wileyonlinelibrary.com]
3.2 Estimation of exposure-response curves

3.2.1 Estimation accuracy

BSTARSS, GAMDP, and GAMTS achieved similar mean-squared errors to the oracle method, with the exception of GAMTS in low signal-to-noise ratio scenarios of quadratic relationships (Figure 8 and Table S6, Supplementary Material). This was the case to a lesser degree for BKMR, which had poorer performance in some low signal-to-noise ratio scenarios. BART mean-squared errors were substantially higher than the oracle method, except when exposure-response relationships were asymmetric inverse-U-shaped. MARS also performed competitively when relationships were asymmetric inverse-U-shaped, but had poorer performance elsewhere when the correlation was halved. Similar conclusions were drawn when considering weak (PPB, BPA) and strong (MPB, BP3) exposure-response relationships separately (Supplementary Material Table S6).

3.2.2 Credible interval and prediction interval coverage

BSTARSS and BART achieved 90% credible interval coverage closely approaching 100% across all scenarios, meaning that the credible intervals were excessively wide (Figure 9 and Table S7, Supplementary Material). This was also the case for MARS prediction intervals. BKMR had a variable credible interval coverage between 75% and 100%. GAMDP and GAMTS had credible interval coverage between 55% and 96% when exposure-response relationships were
linear, S-shaped, or quadratic, and between 36% and 63% when relationships were asymmetric inverse-U-shaped. Similar conclusions were drawn when considering weak (PPB, BPA) and strong (MPB, BP3) exposure-response relationships separately (Supplementary Material Table S7).

### 3.2.3 | Graphical comparison of estimated and true curves

We present estimated curves for one replication, chosen at random, for MPB (Figure 10). Curves were well estimated overall, but all methods appeared to have some difficulty with the lower exposure portion of the asymmetric inverse-U-shaped curve. There was also some evidence of under-smoothing by BART, and a flat instead of quadratic relationship was estimated by MARS when the signal-to-noise ratio was low. We present all curves estimated by each method for MPB, in Supplementary Material Figures S6-S12.

### 4 | DISCUSSION

We assessed the performance of six methods for variable and function selection when exposure-response relationships are nonlinear, in a simulation study based on maternal exposure to 17 phthalates and phenols in the NHANES. Our results suggest that BKMR, BSTARSS, GAMDP, and GAMTS may be best suited for the analysis of mixtures of correlated chemicals when there is uncertainty regarding the shapes of exposure-response relationships. These four methods performed consistently well across scenarios, balancing moderate to high sensitivity, specificity, and positive and negative predictive values. Moreover, they estimated the shapes of exposure-response relationships with error comparable to an oracle method (GAM estimate of the true model). By contrast, BART and MARS had low sensitivity in low signal-to-noise ratio scenarios and the highest mean-squared errors.

Variable selection requires a choice of threshold for the posterior inclusion probabilities (in BSTARSS and BKMR) or variable inclusion proportions (in BART). Although the magnitude of the posterior inclusion probabilities may be sensitive to prior and tuning parameter selection, our choice of 0.5 (the median probability model) appeared reasonable for BSTARSS and BKMR (on average 4.4 exposures were selected by both methods). For BART, three threshold selection rules have been proposed, of which the local threshold that we used yields the least sparse solutions. However, BART selected too few exposures (2.2 on average) and its sensitivity increased with sparsity, suggesting that this threshold is overly stringent and may be more suitable for sparser problems. In practice, the adequacy of the selected threshold is not known and simplistic binary decision making should be avoided in favor of assessing the ranking of exposures according to their posterior inclusion probabilities or variable inclusion proportions. Selection thresholds are not required in GAMDP, GAMTS, and MARS; GAMDP and GAMTS use a shrinkage approach to remove exposures from a model, and MARS finds an optimal subset model with predictive performance comparable to a larger model.

Our additive (main effects only) data-generating processes favored BSTARSS, GAMDP, GAMTS, and MARS, which were specified to model univariate smooths for each exposure. BKMR is, by default, specified to model a multidimensional function for the exposures, and BART is a nonparametric method that imposes no structural assumptions. BKMR and BART do not therefore require a priori specification of interactions and are able to automatically identify pairwise and higher order interactions. However, if interactions are spurious artifacts in an additive model of univariate nonlinear exposure-response functions, BART and BKMR are at risk of interpreting nonlinearity as interaction. By contrast, MARS, BSTARSS, and the GAM-based methods allow the researcher to select the degree of interaction a priori. BSTARSS decomposes a multivariate exposure-response function into linear effects, smooth main effects, linear interactions, varying-coefficient terms (linear × smooth interactions), and smooth bivariate interactions. Similarly, GAMDP and GAMTS allow the specification of hierarchical interactions through tensor product smooths. This has the important advantage of allowing the researcher to test linearity (in BSTARSS) and interaction hypotheses; however, it can substantially increase the complexity of the model when there are a large number of exposures.

The use of informative priors in BKMR and BSTARSS (ie, priors for $r_j$ in BKMR and $\tau^2_j$ in BSTARSS that may influence posterior inference) may have given these methods an advantage. The variable selection performance of BART may be improved by specifying an informative prior, by giving subsets of exposures greater than equal weight of being selected as splitting variables. Although a correctly specified informative prior can increase power and decrease the chance of
**FIGURE 8**  Ratio of method to oracle mean-squared error for outcome-associated exposures, with each exposure evaluated at the 25th, 50th, 75th percentile and other exposures at their mean, by method and scenario. Dashed line at 1 is the targeted ratio [Colour figure can be viewed at wileyonlinelibrary.com]

**FIGURE 9**  90% credible interval coverage (prediction interval coverage for MARS) for outcome-associated exposures, with each exposure evaluated at the 25th, 50th, 75th percentile and other exposures at their mean, by method and scenario. Dashed line at 0.9 is the targeted coverage [Colour figure can be viewed at wileyonlinelibrary.com]
LAZAREVIC ET AL.

Statistics in Medicine

FIGURE 10  Estimated posterior means evaluated at every tenth percentile of MPB (methylparaben) with other exposures at their means, for each method and scenario in one replication, together with the true curves. SNR, signal-to-noise ratio [Colour figure can be viewed at wileyonlinelibrary.com]

a false positive finding.43 simulations have shown that even incorrectly specified informative priors may benefit variable selection by BART.28 A common concern is that informative priors are subjective and may excessively influence the posterior44; however, sensitivity analyses can assess the impact of varying strengths of prior information on posterior inference (Section 3, Supplementary Material). Moreover, incorporating external knowledge (eg, from experimental research45 or meta-analyses) may allow the researcher to explicitly model assumptions that may be a source of bias in conventional (ie, objective) modeling approaches.46

We would expect that linear penalized regression methods have a performance advantage over nonparametric methods when the data-generating process is linear, and this was the case for lasso regression. We also found that lasso variable selection is robust to some degree of nonlinearity, specifically, to settings in which the majority of the relationship is approximately linear and the turning point/s occur in the tails of the exposure distribution (ie, the S-shaped and asymmetric inverse-U-shaped relationships). However, the performance of lasso deteriorated substantially when exposure-response relationships were quadratic, presumably because methods that assume linearity may fit a horizontal line and fail to detect a symmetric U-shaped or inverse-U-shaped relationship. Moreover, we held the shapes of exposure-response relationships constant across outcome-associated exposures. In practice, shapes are likely to vary across chemicals (eg, a mix of symmetric and asymmetric, monotonic and nonmonotonic, relationships), which may affect the ability of lasso to identify and rank nonlinear exposures.

Although lasso is a sensitive method for identifying outcome-associated exposures when exposure-response relationships are linear or approximately linear, statistical inference is complicated by the highly nonnormal finite sample distributions and large sample properties that depend on the choice of tuning parameter.47 Data-driven approaches such as cross-validation for the selection of tuning parameters and model optimization may adversely impact variable selection stability, in both lasso and MARS.48 One major advantage of Bayesian penalized regression methods (and GAM-based methods if we take an empirical Bayes perspective on smoothing parameter estimation)33 is that inference is based on the marginal posterior of the exposure coefficients, meaning that these methods provide a measure of uncertainty in the coefficient estimates and inference does not depend on the tuning parameters or require asymptotic assumptions.49 Bayesian methods, therefore, achieve selection stability by allowing parameter and model uncertainty rather than requiring data
Moreover, uncertainty in the tuning parameter estimates can be assessed through their marginal posterior distributions.\textsuperscript{49} Although high correlation between chemicals may complicate the identification of outcome-associated exposures, we observed that the signal-to-noise ratio had a stronger impact on performance than changes to the exposure correlation structure. Penalized regression methods are robust to the effects of collinearity but no method may be able to discriminate between very highly correlated exposures. BKMR may be specified to perform hierarchical variable selection (ie, to estimate a joint posterior inclusion probability for a group of correlated exposures and conditional probabilities for each exposure),\textsuperscript{24} enabling selection of one exposure within the group. However, this can adversely impact stability as the selected exposure may vary with repeated sampling\textsuperscript{51} and may require consultation with subject-matter experts to choose a subset of exposures prior to statistical modeling. This is especially germane to MARS, which can select an exposure somewhat arbitrarily among a group of correlated exposures.\textsuperscript{41} Concurvity (ie, nonlinear dependence between exposures) may also adversely impact variable selection performance.\textsuperscript{30} This was not the case with the NHANES data but is an important consideration that should be assessed in studies of chemical mixtures (eg, using maximal information-based nonparametric exploration statistics\textsuperscript{21} or concurvity measures for GAMs,\textsuperscript{34} available in the R package mgcv).

All of the six methods for nonlinearity are implemented in accessible software and can estimate effects with a measure of uncertainty. However, confidence intervals in MARS are invalid as they do not take into account selection uncertainty, so only prediction intervals are available.\textsuperscript{37} All of the methods can adjust for linear confounders; however, confounders are subject to selection in MARS and in the bartMachine implementation of BART. Several other considerations are important when selecting a method. All six methods can accommodate both continuous (Gaussian) and binary outcomes (BKMR and BART use a Probit model,\textsuperscript{29,52} while BSTARSS uses a Binomial model).\textsuperscript{32} BSTARSS can additionally accommodate count outcomes (Poisson model),\textsuperscript{32} and the GAM-based methods and MARS can model any exponential-family outcome.\textsuperscript{34,37} BKMR and BSTARSS can incorporate random subject-specific intercepts,\textsuperscript{32,52} and random effects can be included in GAMDP and GAMTS.\textsuperscript{34} The R packages for BKMR, BSTARSS, GAMDP, GAMTS, and MARS require complete data sets, so an additional missing value method may be required such as multiple imputation by chained equations.\textsuperscript{53} By contrast, BART automatically handles missing values without imputation, through an extension of the partitioning mechanisms native to tree-based methods.\textsuperscript{29,54}

Nonparametric methods require careful specification and models should therefore be developed in consultation with subject-matter experts who understand the exposures and their potential health effects. Researchers should fully report their assumptions (including priors and tuning parameters) and use standard model checking procedures. We recommend that researchers select one method \textit{a priori}, as this increases robustness and reduces the chance of a false positive finding compared with using many methods and publishing the most appealing results.

### 4.1 Limitations

We limited our study to lower dimensional models of eight and 17 exposures and no confounding by nonexposure variables; however, all of the methods considered are applicable to higher dimensional settings, with the caveat that GAMTS and GAMDP require that the sample size is greater than the average basis size multiplied by the number of exposures.\textsuperscript{33} We did not include interactions between exposures in our data-generating processes. Although statistical interactions between chemicals in environmental epidemiology studies are possible, they have seldom been assessed,\textsuperscript{4} and are often conflated with biological interactions.\textsuperscript{55} Nevertheless, when analyzing correlated chemical mixtures, failure to consider nonlinearity may lead to biased and false positive interaction effects, and there may be ambiguity in the magnitude of interaction and nonlinear main effects when both are included in a model.\textsuperscript{42,56} Assessing the performance of methods for detecting nonlinear interactions is therefore an important area of future research. We considered exposure measurements at a single point in time; future research should address gestational windows of exposure susceptibility, by assessing the performance of methods that consider exposure trajectories, such as lagged kernel machine regression\textsuperscript{57} and lagged weighted quantile sum regression.\textsuperscript{58} Studies using data from surveys with complex designs, such as the NHANES, may need to incorporate survey weights; however, accounting for survey weights may not be straightforward in methods for variable and function selection. Finally, we focused on six variable selection methods that can model nonlinear exposure-response relationships, but other methods are available and have been reviewed elsewhere.\textsuperscript{4,13}
5 | CONCLUSIONS

We compared the performance of seven methods for identifying outcome-associated exposures in a correlated mixture, six that can model nonlinearity and one that assumes linearity, while varying the shapes of exposure-response relationships. We used a multivariate copula to simulate realistic exposure data based on prenatal exposure to 17 phthalates and phenols in the NHANES, which allowed us to assess whether the level of correlation was affecting method performance. Despite the recent popularity of machine learning methods in epidemiology, we found that MARS and BART were less competitive in low sparsity and low signal-to-noise ratio scenarios that are prevalent in epidemiology studies. For BKMR, BSTARSS, GAMDP, and GAMTS, in terms of variable selection performance, we found that there was little cost to their use over lasso penalized regression when exposure-response relationships were linear, and a distinct advantage to their use when exposure-response relationships were nonlinear. While variable and function selection methods may require thoughtful application, they are able to: estimate the shapes of exposure-response relationships, estimate effects at specified exposure levels with credible intervals, and, in a Bayesian setting, they can incorporate external information from experimental studies or meta-analyses. Although mechanisms for nonmonotonic EDC dose-response curves in cell-, tissue-, and animal-experimental studies are well understood, it is widely accepted that nonmonotonic relationships can occur in epidemiological studies, it is not yet known which shapes are likely to apply to specific EDCs and epidemiological endpoints. Our findings may inform method choice in studies seeking to identify harmful chemicals in a mixture and to determine the nature of exposure-response relationships.

ACKNOWLEDGEMENTS

N.L. is supported by an Australian Government Research Training Program Scholarship. L.D.K. acknowledges a National Health and Medical Research Council Project Grant (APP1142222). P.D.S. is a Senior Principal Research Fellow of the National Health and Medical Research Council, Australia. A.G.B. is supported by an Australian National Health and Medical Research Council Senior Research Fellowship (APP1117784). The authors would like to thank the participants involved in the NHANES. The authors are grateful to the associate editor and two anonymous reviewers whose comments contributed to improving the manuscript.

CONFLICT OF INTEREST

The authors declare that they have no competing interests.

ETHICS STATEMENT

The study was deemed to be exempt from ethics review under the Australian National Statement on Ethical Conduct in Human Research and The University of Queensland policy (Clearance Number: 2017001605).

DATA AVAILABILITY STATEMENT

The NHANES data are openly available at https://www.cdc.gov/nchs/nhanes/index.htm. The R code to reproduce this simulation study is available at https://github.com/n-lazarevic/.

ORCID

Nina Lazarevic https://orcid.org/0000-0001-6463-6719
Luke D. Knibbs https://orcid.org/0000-0002-0399-2370
Peter D. Sly https://orcid.org/0000-0001-6305-2201
Adrian G. Barnett https://orcid.org/0000-0001-6339-0374

REFERENCES

1. Billionnet C, Sherrill D, Annesi-Maesano I. Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol. 2012;22(2):126-141. https://doi.org/10.1016/j.annepidem.2011.11.004.
2. Davalos AD, Luben TJ, Herring AH, Sacks JD. Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. Ann Epidemiol. 2017;27(2):145-153.e1. https://doi.org/10.1016/j.annepidem.2016.11.016.
3. Hamra GB, Buckley JP. Environmental exposure mixtures: questions and methods to address them. Curr Epidemiol Rep. 2018;5(2):160-165. https://doi.org/10.1007/s40471-018-0145-0.
4. Lazarevic N, Barnett AG, Sly PD, Knibbs LD. Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives. Environ Health Perspect. 2019;127(2):026001. https://doi.org/10.1289/EHP2207.
5. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science. *Curr Environ Health Rep*. 2017;4(4):481-490. https://doi.org/10.1007/s40572-017-0162-z.

6. Sun Z, Tao Y, Li S, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health*. 2013;12(1):85. https://doi.org/10.1186/1476-069X-12-85.

7. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. 2005;14(8):1847-1850. https://doi.org/10.1158/1055-9965.EPI-05-0456.

8. Agier L, Portengen L, Chadeau-Hyam M, et al. A systematic comparison of linear regression-based statistical methods to assess exposome-health associations. *Environ Health Perspect*. 2016;124(12):1848-1856. https://doi.org/10.1289/EHP172.

9. Lentes V, Vermeulen R, Portengen L. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. *Occup Environ Med*. 2018;75(7):522-529. https://doi.org/10.1136/oemed-2016-104231.

10. Barrera-Gómez J, Agier L, Portengen L, et al. A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ Health*. 2017;16(1):74. https://doi.org/10.1186/s12940-017-0277-6.

11. Claus Henn B, Ettinger AS, Schwartz J, et al. Early postnatal blood manganese levels and children’s neurodevelopment. *Epidemiology*. 2010;21(4):433-439. https://doi.org/10.1097/EDE.0b013e3181d8e52.

12. Vandenberg LN, Colborn T, Hayes TB, et al. Hormones and endocrine-disrupting chemicals: low-dose effects and nonmonotonic dose responses. *Endocr Rev*. 2012;33(3):378-455. https://doi.org/10.1210/er.2011-1050.

13. Scheipl F, Kneib T, Fahrmeir L. Penalized likelihood and Bayesian function selection in regression models. *Adv Stat Anal*. 2013;97(4):349-385. https://doi.org/10.1007/s10182-013-0211-3.

14. Diamanti-Kandarakis E, Bourguignon J-P, Giudice LC, et al. Endocrine-disrupting chemicals: an Endocrine Society scientific statement. *Endocr Rev*. 2009;30:293-342. https://doi.org/10.1210/er.2009-0003.

15. Gillman MW. Developmental origins of health and disease. *N Engl J Med*. 2005;353(17):1848-1850. https://doi.org/10.1056/NEJMoa058187.

16. Meeker JD. Exposure to environmental endocrine disruptors and child development. *Arch Pediatr Adolesc Med*. 2012;166(6):E1-E7. https://doi.org/10.1001/archpediatrics.2012.241.

17. Soto AM, Vandenberg LN, Maffini MV, Sonnenschein C. Does breast cancer start in the womb? *Basic Clin Pharmacol Toxicol*. 2008;102(2):125-133. https://doi.org/10.1111/j.1126-5255.2008.00165.x.

18. Centers for Disease Control and Prevention (CDC). National Centre for Health Statistics (NCHS). National Health and Nutrition Examination Survey Data; 2019. https://www.cdc.gov/nchs/nhanes/index.htm. Accessed July 18, 2019.

19. Gonzalez-Fernandez Y, Soto M. Copulaedas: an R package for estimation of distribution algorithms based on copulas. *J Stat Softw*. 2014;58(1):1-34. https://doi.org/10.18637/jss.v058.i09.

20. Yan J. Enjoy the joy of copulas: with a package copula. *J Stat Softw*. 2007;21(4):1–21. https://doi.org/10.18637/jss.v021.i04.

21. Reshef DN, Reshef YA, Finucane HK, et al. Detecting novel associations in large data sets. *Science*. 2011;334(6062):1518-1524. https://doi.org/10.1126/science.1205438.

22. Ritz C, Baty F, Streibig JC, Gerhard D. Dose-response analysis using R. *PLoS ONE*. 2015;10(12):e0146021. doi:10.1371/journal.pone.0146021.

23. Hankin RKS. Special functions in R: introducing the gsl package. *R News*. 2006;6(4):24-26.

24. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*. 2015;16(3):493-508. https://doi.org/10.1093/biostatistics/kwu058.

25. Liu D, Lin X, Ghosh D. Semiparametric regression of multidimensional genetic pathway data: least-squares kernels machines and linear mixed models. *Biometrics*. 2007;63(4):1079-1088. https://doi.org/10.1111/j.1541-0420.2007.00799.x.

26. Barbieri MM, Berger JO. Optimal predictive model selection. *Ann Stat*. 2004;32(3):870-897.

27. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. *Ann Appl Stat*. 2010;4(1):266-298. https://doi.org/10.1214/09-AOAS285.

28. Bleich J, Kapelner A, George EI, Jensen ST. Variable selection for BART: an application to gene regulation. *Ann Appl Stat*. 2014;8(3):1750-1781. https://doi.org/10.1214/14-AOAS755.

29. Kapelner A, Bleich J. bartMachine: machine learning with Bayesian additive regression trees. *J Stat Softw*. 2016;70(4):1–40. https://doi.org/10.18637/jss.v070.i04.

30. Scheipl F, Fahrmeir L, Kneib T. Spike-and-slab priors for function selection in structured additive regression models. *J Am Stat Assoc*. 2012;107(500):1518-1532. https://doi.org/10.1080/01621459.2012.737742.

31. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin*. 2004;14(3):731-761.

32. Scheipl F. spikeSlabGAM: Bayesian variable selection, model choice and regularization for generalized additive mixed models in R. *J Stat Softw*. 2011;43(14):1-24. https://doi.org/10.18637/jss.v043.i14.

33. Marra G, Wood SN. Practical variable selection for generalized additive models. *Comput Stat Data Anal*. 2011;55(7):2372-2387. https://doi.org/10.1016/j.csda.2011.02.004.

34. Wood SN. *Generalized Additive Models: an Introduction with R*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC Press; 2017.

35. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*. New York: Springer Science & Business Media; 2009.

36. Friedman JH, Rosen CB. An introduction to multivariate adaptive regression splines. *Stat Methods Med Res*. 1995;4(3):197-217. https://doi.org/10.1177/096228029500400303.

37. Milborrow S. *Earth: Multivariate Adaptive Regression Splines*. R package version 4.6.0; 2011. https://CRAN.R-project.org/package=earth.
38. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw.* 2010;33(1):1-22. https://doi.org/10.1109/TPAMI.2005.127.

39. Van Rijsbergen CJ. *Information Retrieval.* London/Boston, MA: Butterworths; 1979.

40. Coull BA, Bobb JF, Wellenius GA, et al. Part I. Statistical learning methods for the effects of multiple air pollution constituents. *Res Rep Health Eff Inst.* 2015;183(Pt 1-2):5-50.

41. Morlini I. On multicollinearity and Concurvity in some nonlinear multivariate models. *JISS.* 2006;15(1):3-26. https://doi.org/10.1007/s10260-006-0005-9.

42. Matuschek H, Kliegl R. On the ambiguity of interaction and nonlinear main effects in a regime of dependent covariates. *Behav Res Methods.* 2018;50(5):1882-1894. https://doi.org/10.3758/s13428-017-0956-9.

43. Swartz MD, Yu RK, Shete S. Finding predictors influencing risk: comparing Bayesian stochastic search and standard variable selection methods applied to logistic regression models of cases and controls. *Stat Med.* 2008;27(29):6158-6174. https://doi.org/10.1002/sim.3434.

44. Morita S, Thall PF, Müller P. Evaluating the impact of prior assumptions in Bayesian biostatistics. *Stat Biosci.* 2010;2(1):1-17. https://doi.org/10.1007/s12561-010-9018-x.

45. Hamra G, Richardson D, MacLehose R, Wing S. Integrating informative priors from experimental research with Bayesian methods. *Epidemiology.* 2013;24(1):90-95. https://doi.org/10.1097/EDE.0b013e31827623ea.

46. Thomas DC, Witte JS, Greenland S. Dissecting effects of complex mixtures: who’s afraid of informative priors? *Epidemiology.* 2007;18(2):186-190. https://doi.org/10.1097/01.ede.0000254682.47697.70.

47. Pfeiffer RM, Redd A, Carroll RJ. On the impact of model selection on predictor identification and parameter inference. *Comput Stat.* 2017;32(2):667-690. https://doi.org/10.1007/s00180-016-0690-2.

48. Roberts S, Nowak G. Stabilizing the lasso against cross-validation variability. *Comput Stat Data Anal.* 2014;70:198-211. https://doi.org/10.1016/j.csda.2013.09.008.

49. Fahrmeir L, Kneib T, Konrath S. Bayesian regularisation in structured additive regression: a unifying perspective on shrinkage, smoothing and predictor selection. *Stat Comput.* 2009;20(2):203-219. https://doi.org/10.1007/s11222-009-9158-3.

50. Meinshausen N, Bühlmann P. Stability selection [discussion on the paper, comment by Richardson S]. *J R Stat Soc Ser B Stat Methodol.* 2010;72(4):417-473. https://doi.org/10.1111/j.1467-9868.2010.00740.x.

51. Kirk P, Witkover A, Bangham CRM, Richardson S, Lewin AM, Stumpf MPH. Balancing the robustness and predictive performance of biomarkers. *J Comput Biol.* 2013;20(12):979-989. https://doi.org/10.1089/cmb.2013.0018.

52. Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ Health.* 2013;12:1-6. https://doi.org/10.1289/ehp.1205179.

53. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. *Stat Methods Med Res.* 2007;16:219–242.

54. Howard GJ, Webster TF. Contrasting theories of interaction in epidemiology and toxicology. *Environ Health Perspect.* 2012;120(4):A143-A144. https://doi.org/10.1289/ehp.1205179.

**SUPPORTING INFORMATION**

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Lazarevic N, Knibbs LD, Sly PD, Barnett AG. Performance of variable and function selection methods for estimating the nonlinear health effects of correlated chemical mixtures: A simulation study. *Statistics in Medicine.* 2020;39:3947–3967. https://doi.org/10.1002/sim.8701