2015

Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk

Jenna Czarnota
Virginia Commonwealth University, czarnotajn@vcu.edu

Chris Gennings
Icahn School of Medicine

David C. Wheeler
Virginia Commonwealth University, dcwheeler@vcu.edu

Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk

Jenna Czarnota[1], Chris Gennings[2] and David C. Wheeler[1]

[1]Department of Biostatistics, School of Medicine, Virginia Commonwealth University, Richmond, VA, USA. [2]Department of Preventive Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA.

Supplementary Issue: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

ABSTRACT: In evaluation of cancer risk related to environmental chemical exposures, the effect of many chemicals on disease is ultimately of interest. However, because of potentially strong correlations among chemicals that occur together, traditional regression methods suffer from collinearity effects, including regression coefficient sign reversal and variance inflation. In addition, penalized regression methods designed to remediate collinearity may have limitations in selecting the truly bad actors among many correlated components. The recently proposed method of weighted quantile sum (WQS) regression attempts to overcome these problems by estimating a body burden index, which identifies important chemicals in a mixture of correlated environmental chemicals. Our focus was on assessing through simulation studies the accuracy of WQS regression in detecting subsets of chemicals associated with health outcomes (binary and continuous) in site-specific analyses and in non-site-specific analyses. We also evaluated the performance of the penalized regression methods of lasso, adaptive lasso, and elastic net in correctly classifying chemicals as bad actors or unrelated to the outcome. We based the simulation study on data from the National Cancer Institute Surveillance Epidemiology and End Results Program (NCI-SEER) case–control study of non-Hodgkin lymphoma (NHL) to achieve realistic exposure situations. Our results showed that WQS regression had good sensitivity and specificity across a variety of conditions considered in this study. The shrinkage methods had a tendency to incorrectly identify a large number of components, especially in the case of strong association with the outcome.

KEYWORDS: environment, WQS regression, penalized regression, lasso, elastic net, cancer risk

SUPPLEMENT: Computer Simulation, Bioinformatics, and Statistical Analysis of Cancer Data and Processes

CITATION: Czarnota et al. Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. *Cancer Informatics* 2015:14(S2) 159–171 doi: 10.4137/CIN.S17295.

RECEIVED: November 05, 2014. **RESUBMITTED:** February 10, 2015. **ACCEPTED FOR PUBLICATION:** February 12, 2015.

ACADEMIC EDITOR: J.T. Efird, Editor in Chief

TYPE: Original Research

FUNDING: The authors (JC) gratefully acknowledge support from the National Institute of Environmental Health Sciences grant (#T32 ES0007334). The authors confirm that the funder had no influence over the study design, content of the article, or selection of this journal.

COMPETING INTERESTS: JC has a conflict of interest due to being funded on a federal training grant. Other authors disclose no competing interests.

CORRESPONDENCE: dcwheeler@vcu.edu

Introduction

The connection between environmental chemical exposures and human health/diseases (eg, cancer) is a complex multidimensional problem that is of great interest to public health researchers. Further, exposure to environmental chemical mixtures varies across location and behavior patterns. For example, exposure patterns to polycyclic aromatic hydrocarbons (PAHs) from automobile exhaust in dense urban sites such as Detroit differ from those in rural, agricultural sites such as Iowa. However, there are many sources of PAHs besides automobile exhaust, and different sources may explain different levels across space. For example, PAH concentrations have been found to be associated with residence age.[1,2] Some environmentally persistent chemicals such as polychlorinated biphenyls (PCBs) may be relatively similar in urban and rural areas. Total PCB levels have been found to be positively associated with percentage of developed land or population density and housing age.[3] Regardless, exposure studies demonstrate complex correlation patterns among environmental chemicals that may vary across regions.[4]

Because individuals are exposed to multiple chemicals simultaneously, it is important to examine the relationship between chemical mixtures and disease risk. Our research focus is on developing analysis strategies that incorporate larger sets of chemicals, which are more representative to actual human exposure. In such a high-dimensional framework, some factors may increase risk, some may diminish risk, and others may have no effect on risk. The goal is to use an analysis strategy that most efficiently detects the signals and deemphasizes the noise in the data. Furthermore, we aim to develop and assess methods that accommodate site-specific exposure patterns and have the ability to accurately discern whether or not a chemical is associated with the outcome of interest.

Through weighted quantile sum (WQS) regression,[5] we are able to estimate a body burden index within a set of

correlated environmental chemicals, and further estimate the association between the index and an outcome of interest. Additionally, the estimated chemical weights allow us to make generalized inference concerning relative chemical importance. WQS regression is constrained to model associations between the outcome and chemicals that are in one direction (all non-negative or all non-positive), making it appropriate in a risk setting where the goal is to identify exposures that are positively associated with a health outcome. As such, WQS regression is designed for variable selection with less emphasis on risk prediction.

Few existing studies of chemical exposure and disease risk have made efforts to consider the impact of spatially varying exposure patterns on the effect of a chemical mixture. One exception is Czarnota et al.[4], where the authors take a site-specific approach in a preliminary effort to assess the effects of spatially varying exposure patterns among chemical mixtures on the risk of non-Hodgkin lymphoma (NHL). The objective in that work was to apply statistical methods to detect bad actors in environmental chemical mixtures, while considering different exposure patterns based on the geographic site. The focus here is on assessing through simulation studies the accuracy of WQS regression in detecting subsets of chemicals associated with health outcomes (binary and continuous) in site-specific analyses and in non-site-specific analyses. We based the simulation study on data from the National Cancer Institute Surveillance Epidemiology and End Results Program (NCI-SEER) case–control study of NHL to achieve realistic exposure situations, while setting which chemicals were truly associated with the health outcomes. For comparison, we also evaluated the performance of several penalized regression methods in correctly classifying chemicals as bad actors or unrelated to the outcome.

Methods

NCI-SEER study population. The NCI-SEER NHL population-based case–control study design has been previously described.[6,7] Briefly, the study was conducted in Iowa, Los Angeles County, and the metropolitan areas of Detroit (Macomb, Oakland, and Wayne counties) and Seattle (King and Snohomish counties). Eligible cases were aged 20–74 years, diagnosed with a first primary NHL between July 1998 and June 2000, and uninfected with HIV. In Seattle and Iowa, all consecutive cases were chosen. In Detroit and Los Angeles, all African American cases and a random sample of white cases were eligible for study, allowing for oversampling of African American cases. Of the 2,248 potentially eligible cases, 320 (14%) died before they could be interviewed, 127 (6%) were not located, 16 (1%) had moved away, and 57 (3%) had physician refusals. Of the 1,728 remaining cases, 1,321 (76%) participated. Controls were selected from Center for Medicare and Medicaid Services files (≥65 years) or the general population using random digit dialing (<65 years) and were frequency matched to cases by sex, age, race, and study site. Of the 2,409 potentially eligible controls, 2,046 were able to be located and contacted, and 1,057 (52%) of these subjects participated.

Computer-assisted personal interviews were conducted in the home of each participant. Interviewers asked about demographics, including race and education, age of the home, housing type, the presence of oriental rugs, pesticide use in the home and garden, residential and occupational histories, and other factors. As described in detail,[6,8] dust was collected between February 1999 and May 2001 from vacuum cleaners of participants who gave permission (93% of cases, 95% of controls) and who had used their vacuum cleaner within the past year and owned at least half of their carpets or rugs for five years or more (695 cases (57%), 521 controls (52%)). Dust samples from 682 cases (98%) and 513 controls (98%) were successfully analyzed between September 1999 and September 2001.

A total of 27 chemicals were measured in house dust (5 PCBs, 7 PAHs, and 15 pesticides). The PCBs were congeners 105, 138, 153, 170, and 180. The PAHs were benz(a)anthracene, benzo(a)pyrene, benzo(b)fluoranthene, benzo(k)fluoranthene, chrysene, dibenz(a,h)anthracene, and indeno(1,2,3-cd)pyrene. The pesticides were α-chlordane, γ-chlordane, carbaryl, chlorpyrifos, cis-permethrin, trans-permethrin, 2,4-dichlorophenoxyacetic acid (2,4-D), DDE, dichlorodiphenyltrichloroethane (DDT), diazinon, dicamba, methoxychlor, o-phenylphenol, pentachlorophenol, and propoxur. Extraction and analysis were performed on 2-g aliquots of dust samples using gas chromatography/mass spectrometry (GC/MS) in selected ion monitoring mode. Concentrations were quantified using the internal standard method. Usual detection limits were 20.8 ng/g of dust for α-chlordane, γ-chlordane, DDE, DDT, propoxur, o-phenylphenol, PAHs, and PCBs; 42–84 ng/g for chlorpyrifos, diazinon, cis-permethrin, dicamba, pentachlorophenol, and 2,4-D; and 121–123 ng/g for carbaryl and trans-permethrin. Changes in analytic procedures during the study resulted in increased detection limits for methoxychlor (from 20.7 to 62.5 ng/g). A small proportion of samples weighing less than 2 g had detection limits that were higher than the usual detection limits.

The laboratory measurements for the 27 analytes contained various types of missing data, primarily when the concentration was below the minimum detection level. To a lesser extent, missing data occurred when there was co-elution between the target chemical and interfering compounds. Chemical concentrations were assumed to follow a log-normal distribution, and data were imputed using a fill-in approach to create 10 complete data sets for each of the 27 analytes. Details about the imputation of analyte values have been published previously.[6,9] A total of 1,180 participants had available measurements of all 27 chemicals. Of these participants, 508 (43%) were controls and 672 (57%) were cases. With respect to study site, 202 (17%) were from Detroit, 340

(29%) from Iowa, 292 (25%) from Los Angeles, and 346 (29%) from Seattle.

Our primary interest in the NCI-SEER NHL study is the chemical exposure patterns and the correlation structure of the exposures. As illustrated in Figure 1, the concentrations among the PCBs were similar across the four sites, while the concentrations of PAHs and pesticides varied considerably by site. More specifically, concentrations for all seven PAHs were notably elevated in Detroit, while elevated concentrations of pesticides were seen in Iowa (eg, 2,4-D and methoxychlor) and Los Angeles (eg, α- and γ-chlordane and propoxur).

The site-specific distributions of the Pearson pairwise correlation coefficients among the log-transformed concentrations are shown in Figure 2. The observed pairwise correlation patterns are complex, with correlations ranging from slightly negative to nearly perfectly correlated within all four sites. When examining the correlations by chemical group, we see that for each site, the PAHs and PCBs demonstrated a high degree of intragroup correlation. The pesticides generally exhibited lesser intragroup correlation, with the exception of the pairwise correlations between metabolites and analogs. For each site, correlation within chemical group is further illustrated in Figure 3 and summarized in Table 1. We see that the PCBs were most highly correlated in Los Angeles, with 75% of the intragroup correlations greater than 0.81. Additionally, we see that the correlation among the PAHs was most pronounced in Detroit (pairwise correlations ranging from 0.95 to 0.99) where PAH exposure was the highest and least pronounced in Los Angeles (interquartile range (IQR) of 0.68–0.86) where PAH exposure concentration was the lowest. As demonstrated by the NHL data, chemical exposure patterns may vary in both concentration and correlation across space, illustrating the need to consider site-specific risk analyses in the context of environmental chemical exposure.

Simulation study design. Using each of the four site-specific correlation structures for the 27 chemicals in the NCI-SEER NHL study, we generated data on a site-specific basis using the observed mean concentrations (from the log scale) and standard deviations, for each of the following three correlation patterns: (1) 65% of the observed site-specific correlation structure (moderate correlation), (2) 30% of the observed site-specific correlation structure (mild correlation), and (3) 1% of the observed site-specific correlation structure (near independence). We did not include the observed correlation patterns to generate data as the resulting correlation matrices were generally singular and could not be inverted.

The following seven chemicals (one PCB, two PAHs, and four pesticides/insecticides) were set to be associated with the response variable: PCB180 (X5), benzo(k)fluoranthene (X8), benzo(a)pyrene (X9), 2,4-D (X19), DDE (X20), methoxychlor (X24), and propoxur (X27). These chemicals were chosen in an effort to represent a wide range of correlations between and among the bad actors and benign chemicals. Correlation within the seven selected chemicals ranged from −0.08 to 0.94, with a median correlation of 0.14. We also ensured that at least one chemical was selected from each chemical group in an attempt to capture the variation in exposure patterns across the study sites. PCB 180 was purposefully selected to represent the PCBs because of its known link to NHL.

Given that chemical exposure patterns differed across site, the association with the outcome variable was assumed true only under the condition that the observed site-specific concentrations were high enough to have a health effect. More specifically, we assumed the association was true within a site if and only if more than 25% of the site-specific concentrations were higher than the overall median concentration. This condition was satisfied for each of the seven specified chemicals in the Detroit, Iowa, and Seattle sites, and thus, all seven of the specified chemicals were set to be associated with the response when simulating data. For the Los Angeles site, over 75% of the observed concentrations for chemicals X8, X9, and X19 were below the overall median concentration. Therefore, only four chemicals (X5, X20, X24, and X27) were set to be correlated with the outcome in Los Angeles.

For each correlation pattern, we simulated data where the selected chemicals (ie, the pre-specified chemicals that satisfied the above condition) had a correlation of 0.1 with the response (weak association with outcome) as well as where the selected chemicals had a correlation of 0.3 with the response (strong association with outcome). To simulate the analyses of case–control study data, we calculated a disease indicator for subjects with the highest 30% of the simulated continuous response variable.

We simulated 100 data sets for each set of conditions (correlation pattern and outcome correlation combination), with a sample size of $N = 1{,}000$ for each site. The site-specific data sets were split into training and validation sets (50:50 split), and then concatenated to create an overall training and testing data set. Thus, the site-specific sample size was $N = 1{,}000$ (500 for estimation and 500 for validation), while the overall sample size was $N = 4{,}000$ (2,000 for estimation and 2,000 for validation).

Data were simulated for each site based upon the site-specific observed mean vector (on the log scale) and covariance matrix. The continuous response variable y was assumed to be normally distributed with a mean of 0 and variance of 1, $y \sim N(0,1)$, for each site. We used a dampening parameter $k \in (0.65, 0.30, 0.01)$ to diminish the site-specific correlations to 65%, 30%, and 1% of that observed using the equation $\tilde{\mathbf{R}}_j = k\left(\mathbf{R}_j - \mathbf{I}\right) + \mathbf{I}$, where $\mathbf{R}_j$ denotes the observed matrix of correlations for site j and $\tilde{\mathbf{R}}_j$ is the corresponding diminished correlation structure. In order to simulate a data set $\mathbf{Y} \sim N(\mathbf{M}, \Sigma)$ with a specified correlation structure for a continuous response variable y and predictors $x_1, x_2, \ldots, x_c$, the following method was used according to Carrico et al.[5]

Let $p = c + 1$, and define $\rho_{p\times p}$ as the correlation matrix between and among y and the chemicals in $\mathbf{X}$. Let $\Sigma_{p\times p}$ be the corresponding covariance matrix and $\mathbf{S}_{p\times 1}$ be the vector

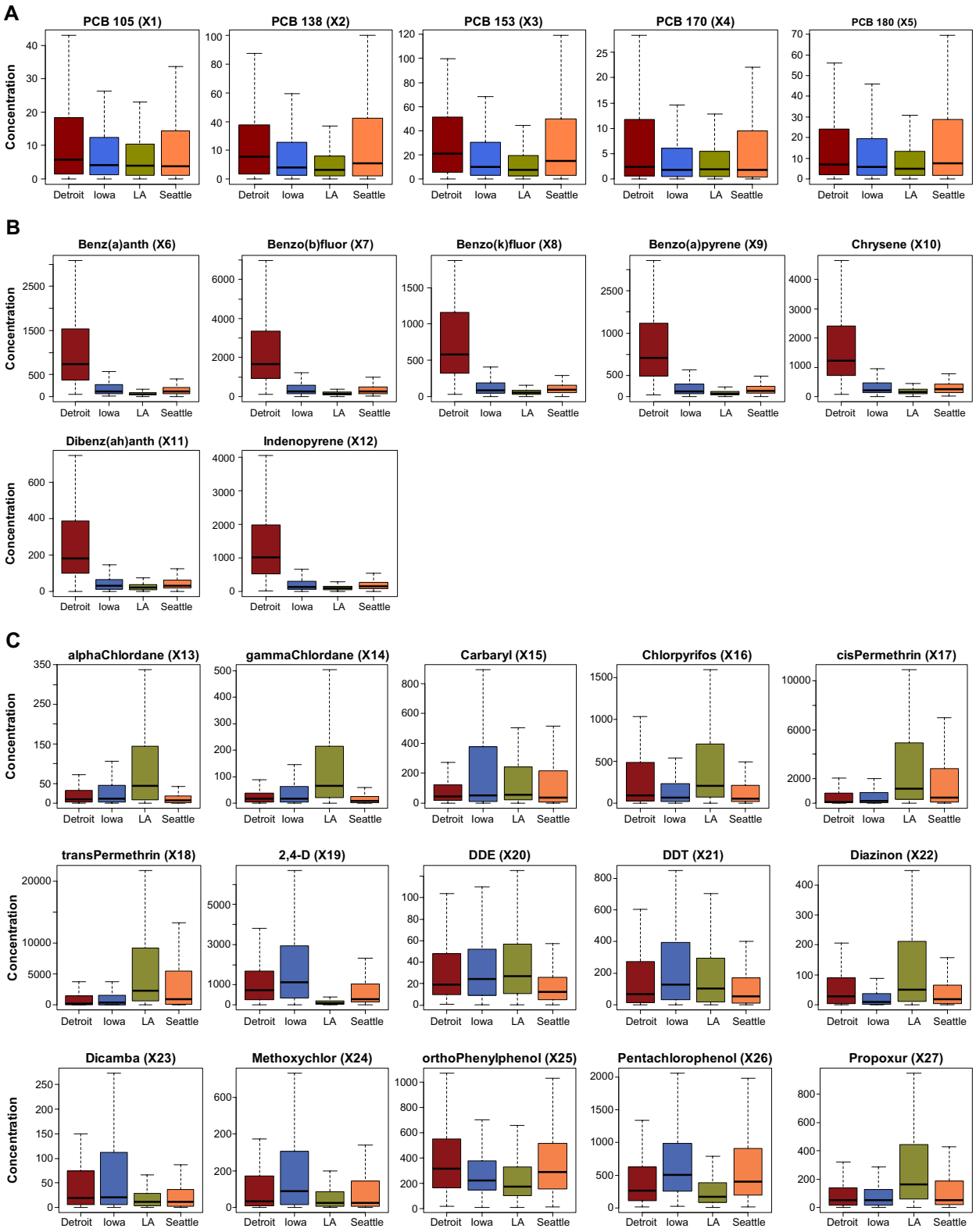

Figure 1. Distribution of chemical concentrations in the NCI-SEER NHL study by chemical group: (**A**) PCBs, (**B**) PAHs, and (**C**) pesticides/insecticides.

of standard deviations. Additionally, let $\mathbf{m}_{p\times 1}$ be the vector of observed sample means for the predictor variables and outcome y. We then define the matrix $\mathbf{D} = \text{diag}(\mathbf{S})$, a square matrix of dimension $p \times p$ with diagonal entries consisting of the standard deviations, and impose the desired correlation structure through the relationship between the correlation and variance given by $\Sigma = \mathbf{D}\rho\mathbf{D}$.

Next, calculate the Cholesky decomposition $\mathbf{U}$ of the covariance matrix Σ. That is, calculate $\mathbf{U}_{p\times p}$ such that $\Sigma = \mathbf{U}'_{p\times p}\mathbf{U}_{p\times p}$, noting that calculation of the Cholesky

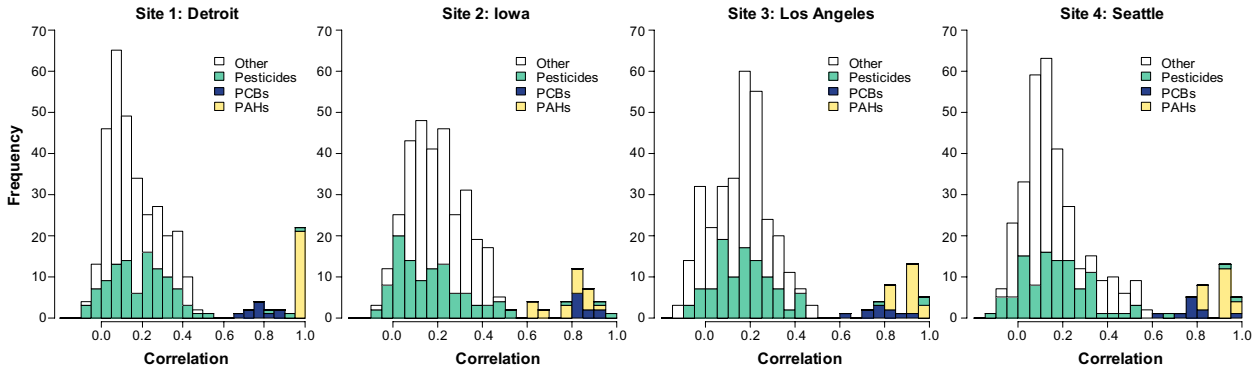

Figure 2. Distribution of Pearson pairwise correlation coefficients among chemical concentrations (on the log scale) by study site in the NCI-SEER NHL study.

decomposition requires that the covariance matrix be positive definite. Simulate $Z_i \sim N\left(\mathbf{0}_{p\times 1}, \mathbf{I}_p\right)$ for $i = 1,\ldots,n$ and define $\mathbf{Z}' = \left[\mathbf{Z}_1 \ldots \mathbf{Z}_n\right]$. In other words,

$$\mathbf{Z}_{n\times p} = \begin{bmatrix} \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_n \end{bmatrix},$$

where each row$_i$ is p-variate standard normal. Let $\mathbf{M}_{n\times p} = \left[\mathbf{m}_{p\times 1}\mathbf{1}_{1\times n}\right]'$ and define $\mathbf{Y}$ as $\mathbf{Y}_{n\times p} = \mathbf{M}_{n\times p} + \mathbf{Z}_{n\times p}\mathbf{U}_{p\times p}$. Here, $\mathbf{Y}$ is the newly generated data matrix with mean $E(\mathbf{Y}) = E(\mathbf{M} + \mathbf{ZU}) = \mathbf{M} + E(\mathbf{Z}) = \mathbf{M}$ and variance $\mathrm{Var}(\mathbf{Y}) = \mathrm{Var}(\mathbf{M} + \mathbf{ZU}) = \mathrm{Var}(\mathbf{M}) + \mathrm{Var}(\mathbf{ZU}) = \mathbf{U}'\mathrm{Var}(\mathbf{Z})\mathbf{U} = \mathbf{U}'\mathbf{U} = \Sigma$. Thus, $\mathbf{Y}_{n\times p}$ is distributed as $N_p(\mathbf{M}, \Sigma)$.

WQS regression. The primary method of risk analysis used in this study was WQS regression. The WQS approach estimates a weighted linear index in which the weights are empirically determined through the use of bootstrap sampling. The approach considers data with c components scored into quantiles that are reasonable to combine into an index and potentially have a common outcome. The weights are constrained to be between 0 and 1 and sum to 1, reducing dimensionality and addressing issues that arise with collinearity. For this analysis, the $c = 27$ chemical concentrations were scored into quartiles (to reduce the influence of outliers in skewed distributions), denoted by q_i, where $q_i = \{0,1,2,3\}$ for $i = 1$–c. For estimation of the weights, we split the simulated data into training and validation data sets of sizes N_t and N_v, respectively. A total of $B = 100$ bootstrap samples of size N_t were generated from the training data and used to estimate the unknown weights w_i that maximized the likelihood for $b = 1$–B for the following nonlinear model:

$$g(\mu) = \beta_0 + \beta_1\left(\sum_{i=1}^{c} w_i q_i\right) + z'\phi\Big|_b \qquad (1)$$

subject to the constraints

$$\sum_{i=1}^{c} w_i\Big|_b = 1 \quad \text{and} \quad 0 \le w_i \le 1 \quad \text{for } i = 1-c$$

In the above equation, w_i represents the weight for the ith chemical component q_i and the term $\sum_{i=1}^{c} w_i q_i$ represents a weighted index for the set of c chemicals of interest. Furthermore, $\mathbf{z}$ denotes a vector of covariates determined

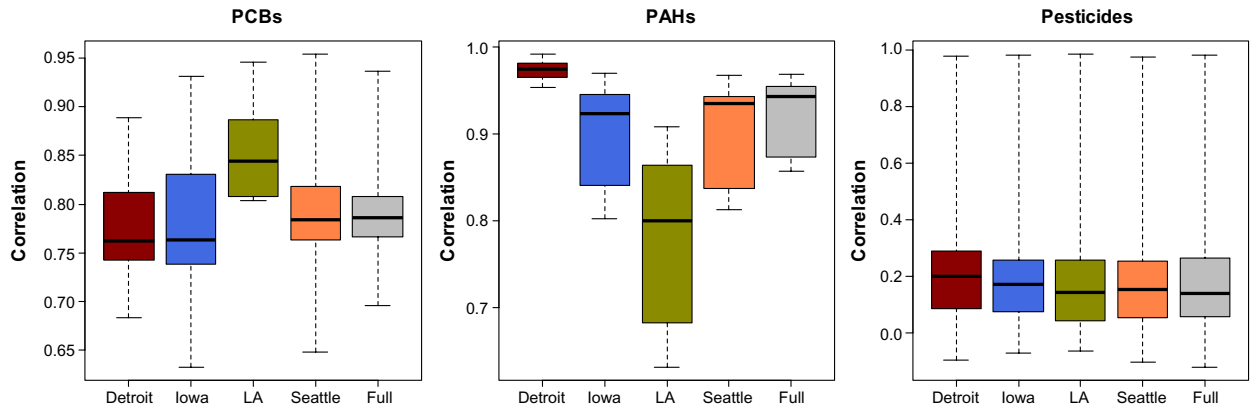

Figure 3. Distribution of the observed intergroup correlations for PCBs, PAHs, and pesticides by study site and across the full-study population.

Table 1. Correlations within chemical group by study site and across the full population in the NCI-SEER NHL study.

	PCBs		PAHs		PESTICIDES/INSECTICIDES	
	MEDIAN	RANGE	MEDIAN	RANGE	MEDIAN	RANGE
Detroit	0.76	[0.68, 0.89]	0.98	[0.95, 0.99]	0.20	[−0.09, 0.98]
Iowa	0.76	[0.63, 0.93]	0.92	[0.80, 0.97]	0.17	[−0.07, 0.98]
Los Angeles	0.84	[0.80, 0.95]	0.80	[0.63, 0.91]	0.14	[−0.06, 0.99]
Seattle	0.78	[0.65, 0.95]	0.93	[0.81, 0.97]	0.16	[−0.10, 0.97]
Full	0.79	[0.70, 0.94]	0.94	[0.86, 0.97]	0.14	[−0.12, 0.98]

prior to estimation of the weights, and g is a monotonic and differentiable link function that relates the mean, μ, to the predictor variables in the right-hand side of the equation. For this analysis, we considered a continuous outcome and a binary outcome with an identity link and a logit link for g for the respective outcome variables.

For each bootstrap sample, the significance of the estimated vector of weights was evaluated through the significance ($P \leq 0.05$) of $\hat{\beta}_1$, which corresponds to the parameter estimate for the weighted index. The weighted quantile score was then estimated as $\text{WQS} = \sum_{i=1}^{c} \overline{w}_i q_i$, where $\overline{w}_i = (1/n_B)\sum_{j=1}^{n_B} w_{ij}$, and n_B is defined as the number of bootstrap samples in which $\hat{\beta}_1$ was significant. Finally, the significance of WQS was assessed using the validation data set and the model

$$g(\mu) = \beta_0 + \beta_1 \text{WQS} + \mathbf{z}'\phi. \quad (2)$$

Simultaneous estimation of the unknown weights and parameters was achieved through the use of an optimization algorithm that maximized the nonlinear function in equation (1), subject to the linear constraint $\sum_{i=1}^{c} w_i = 1$ and the bounds $w_i \in [0,1]$. The nonlinear optimization was performed in R using the function solnp found in the package Rsolnp.[10] The algorithm employed belongs to the class of indirect solvers and implements the augmented Lagrange multiplier method with a sequential quadratic programing interior algorithm.

For each of the 100 simulated data sets for each set of simulation conditions (correlation pattern and outcome correlation), we performed WQS regression on the full data set (adjusted for study location) and separately at each site. The ranks used in WQS regression were calculated within each site for the site-specific analyses and overall for the site-adjusted analysis of the full data set. The process was performed twice, once using the continuous outcome variable and once using the binary outcome variable. Therefore, for each set of conditions, a total of five indices (four site-specific indices and one full-study index) were estimated for each outcome variable. The median number of correctly and incorrectly selected chemicals was calculated for each of the five indices across the 100 simulated studies. A chemical was identified as selected if it received a weight of at least 0.05. The significance of the five estimated indices in their respective validation data sets was also examined.

Comparison of WQS regression with lasso, adaptive lasso, and elastic net. Modern methods that address collinearity and high-dimensionality (eg, lasso, elastic net) have been demonstrated to be less accurate in the selection of potentially harmful chemicals compared with WQS regression.[5] To further assess the use of shrinkage regression models for evaluating effects of chemical exposures, we fitted lasso,[11] adaptive lasso,[12] and elastic-net[13] models to the 100 training data sets (of size $N_t = 500$) for each set of conditions (correlation pattern and outcome association) for both the continuous and binary response variables. In an effort to most closely parallel the site-adjusted model used in the estimation of WQS weights in these overall data sets, indicator variables for site were included in the lasso, adaptive lasso, and elastic-net models but were not subjected to the penalty (ie, these variables were forced to remain in the model). The penalized regressions were performed in R using the cv.glmnet and glmnet functions in the glmnet package.[14] For the lasso and adaptive lasso models, the tuning parameters were chosen using cross-validation and the one standard error rule.[15] For the elastic-net models, a grid search was performed using cross-validation, with the elastic-net mixing parameter allowed to vary from 0 to 1.

For the lasso, adaptive lasso, and elastic-net methods, chemicals related to the outcome variable were identified as correctly chosen if they were retained in the model with a positive coefficient, while chemicals not related to the outcome variable were identified as incorrectly chosen if they were retained in the model. The median and IQR for the number of correctly and incorrectly selected chemicals were reported, and the three methods were compared in terms of sensitivity and specificity.

Results

Sensitivity and specificity of WQS regression. The median number of correctly and incorrectly chosen chemicals across 100 samples for each setting is displayed in Tables 2 and 3, respectively. When association with the outcome was strong ($r = 0.3$), the estimated weights for sites Detroit, Iowa, and Seattle performed at least as well as the weights estimated using the full data set, in terms of both sensitivity and specificity. Based on the weights for these three sites, WQS regression correctly chose all seven chemicals at least half of the time (median value) for both the continuous and

Table 2. Median [IQR] number of correctly chosen chemicals for the five WQS indices across the 100 simulated data sets.

CONTINUOUS OUTCOME	TRUTH	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
		$k = 0.65$	$k = 0.30$	$k = 0.01$	$k = 0.65$	$k = 0.30$	$k = 0.01$
$WQS_{Detroit}$	7	5 [5,6]	6 [5,6]	6 [5,6]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{Iowa}	7	5 [5,6]	6 [5,6]	6 [5,6]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{LA}	4	4 [3,4]	4 [3,4]	4 [3,5]	4 [4,4]	4 [4,4]	4 [4,4]
$WQS_{Seattle}$	7	5 [5,6]	6 [5,6]	6 [5,6]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{Full}	7	6 [5,6]	6 [6,7]	6 [6,6]	6 [6,7]	7 [7,7]	7 [7,7]
BINARY OUTCOME							
$WQS_{Detroit}$	7	5 [4,5]	5 [4,6]	5 [4,6]	7 [6,7]	7 [7,7]	7 [7,7]
WQS_{Iowa}	7	5 [4,5]	5 [4,6]	5 [4,6]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{LA}	4	4 [3,4]	4 [3,4]	4 [3,4]	4 [4,4]	4 [4,4]	4 [4,4]
$WQS_{Seattle}$	7	5 [4,5]	5 [4,5]	5 [4,5]	7 [6,7]	7 [7,7]	7 [7,7]
WQS_{Full}	7	6 [6,6]	6 [6,7]	6 [6,7]	7 [7,7]	7 [7,7]	7 [7,7]

binary response variables, regardless of the correlation pattern among predictors. The weights estimated from the full-study population also correctly chose all seven chemicals at least half of the time (median value), with the exception of the setting in which the correlation among predictors was the strongest (65% of observed site-specific correlations) and the outcome was continuous (median of six correctly chosen chemicals). With respect to specificity, when outcome correlation was strong, the weights for sites Detroit, Iowa, and Seattle and the weights for the full-study population had a median value of 0 for incorrectly chosen chemicals at all settings. When association with the outcome was weak ($r = 0.1$), the weights estimated from the full data set may have slightly improved sensitivity, as the median number of correctly chosen chemicals for the overall analysis was often greater by one chemical (as compared to the median number correctly selected in Detroit, Iowa, and Seattle). This one chemical increase in sensitivity was seen across all correlation patterns for the binary outcome variable and in the case of moderate correlation (65% of that observed) among chemicals for the continuous variable. Similarly, the weights estimated from the full data set may have slightly increased specificity when outcome association was weak, as the site-specific weights tended to incorrectly choose one additional chemical.

From the results for the Los Angeles site, all four chemicals were correctly selected at least half of the time for each setting, but the number of incorrectly chosen chemicals ranged from two to four across the different settings. Because fewer chemicals (four) were set to be associated with the outcome within this site, it may have been advantageous to define a criterion for chemical selection unique to this site.

In summary, WQS regression had good sensitivity and specificity at all settings for the site-specific and overall analysis for both the continuous and binary outcome variables. Performance of the site-specific analyses was comparable to that of the overall analysis. We caution against overinterpretation of a one or two chemical difference in sensitivity and/or specificity, as any perceived improved performance for the overall analyses in comparison to the site-specific analyses may be a result of the four-fold increase in sample size in the estimation data set for the index derived from the full-study population. Furthermore, the results presented in this analysis are dependent upon chemical selection as defined by a minimum value of 0.05 for the estimated chemical weight. It was decided a priori that a chemical must receive at least 5% of the weights to be considered important. While this may be reasonably applied in practice, the method for best choosing a cutoff value is still an open area of research. The choice of cutoff value may be affected by number of chemicals, correlation structure, signal strength, etc.

With respect to chemical selection, we generally expect to see an increase in sensitivity and a decrease in specificity as the threshold weight for chemical selection is lowered. Figure 4 shows modified receiver operating curves (ROC) for the three different correlation structures among the chemicals in the setting of weak association with the (a) continuous outcome and (b) binary outcome, with the cutoff weight for chemical selection varied. The true positive rate (sensitivity) was calculated as the average percentage of correctly selected chemicals across the 100 simulations, and the false positive rate (1-sensitivity) was calculated as the average percentage of incorrectly selected chemicals across the 100 simulations. As the cutoff weight for chemical selection is lowered, we see an increase in both the average true and false positive rates as expected. The a priori chosen cutoff of 0.05 (ie, 5% of the total chemical weights) performed well, regardless of the level of correlation among chemicals or strength of association with outcome. When association with the continuous outcome was weak (Fig. 4A), the average false positive rate for the cutoff weight of 0.05 ranged from 2.3% to 4.5% across the correlation structures, while the average true positive rate ranged from 83.9% to 85.6%. Similarly, for the binary outcome (Fig. 4B),

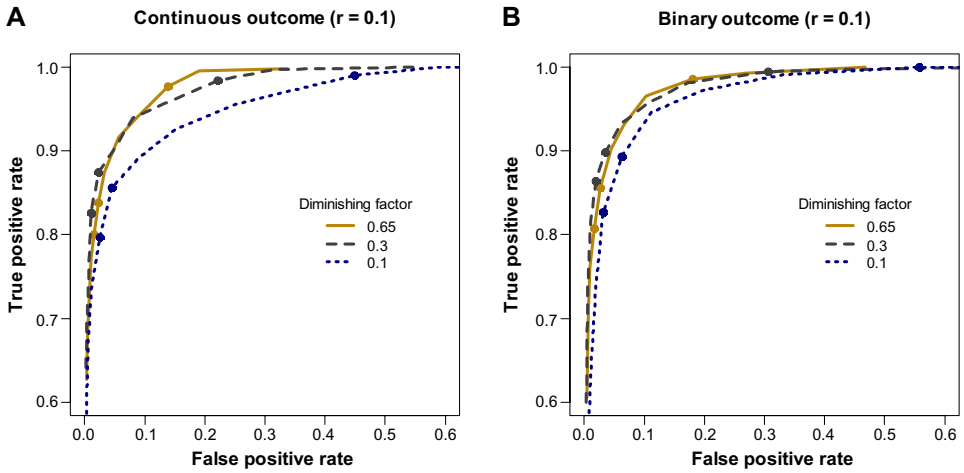

Figure 4. Modified ROC for the WQS index derived from the full-study population with varying weight thresholds for chemical selection for a continuous outcome (**A**) and a binary outcome (**B**). The true positive rate was calculated as the average percentage of correctly selected chemicals (ie, the average number of correctly selected chemicals divided by 7) over the 100 simulations, while the false positive rate was calculated as the average percentage of incorrectly selected chemicals (ie, the average number of incorrectly selected chemicals divided by 20). The points on each line represent (from left to right) the average true and false positive rates for weight thresholds of 0.01, 0.05, and 0.06. Diminishing factor refers to the level of correlation among the chemicals (ie, the amount by which the observed chemical correlations was diminished).

the average false positive rate for the cutoff weight of 0.05 ranged from 2.6% to 6.3% across the correlation structures, while the average true positive rate ranged from 85.6% to 89.9%. Finally, when association with the outcome was strong (results not shown), the average false positive rate for the cutoff weight of 0.05 was at most 0.05%, while the average true positive rate was at least 92.6% across the three correlation structures and both the continuous and binary outcomes.

Distribution of WQS regression weights. In practice, we also look at the distributions of the weights in deciding which chemicals are important. Figure 5 shows the distribution of the average weights across the 100 simulated samples for the seven chemicals assumed to be associated with outcome for each of the five indices. The plots focus on the setting in which there was weak correlation with the continuous outcome for (A) moderate correlation (65% of that observed) among chemicals and (B) correlation diminished to 1% of that observed. For both correlation structures, WQS appropriately placed considerable weight on the true bad actors and also placed negligible weight on chemicals uncorrelated with the outcome. The latter is demonstrated by the near zero weight placed on X8, X9, and X19 by the Los Angeles index. Also, as correlation among chemicals was diminished, reliability of the weights improved, as evident by the narrowed distributions in Figure 5(B).

When comparing the weights from the different indices, the index for Los Angeles tended to place greater emphasis on chemicals X5, X20, X24, and X27 compared with the other sites. This is likely because of the fact that these four chemicals were the only true bad actors in this site, and thus, the weights as a whole were divided over fewer components. Additionally, the weights for the full-study population analysis seem to demonstrate an averaging effect across the sites, as they appear to shift downward for the chemicals that were unassociated with outcome in Los Angeles (X8, X9, and X19). For the chemicals associated with outcome in all four sites (X5, X20, X24, and X27), the weights estimated by the overall analysis were slightly higher than those estimated in site-specific analysis. This may be attributable to the increased power (greater sample size) of the overall analysis.

When strongly correlated with the continuous outcome (data not shown), the findings were consistent with those discussed above, but the important chemicals (true bad actors) tended to receive higher average weights. With respect to the binary outcome (data not shown), the findings were again analogous.

Power of WQS regression. The estimated weights were applied to the validation data sets, and significance was assessed through the $\hat{\beta}_1$ parameter. The results of the significance tests across the simulated data sets are summarized in Tables 4 and 5. The average parameter estimates were all positively associated with outcome, suggesting that increased body burden (as estimated by the WQS index) was associated with increased risk. Overall power increased, as expected, as the association with the outcome variable strengthened and correlation among chemicals diminished. Additionally, the indices estimated using the continuous outcome variable had greater power in comparison to those using the binary outcome variable. Finally, the indices for the Los Angeles site exhibited lower power in comparison to the other indices. This is likely because the set of predictor variables as a whole was contributing less information, as fewer chemicals were set to be associated with the outcome. This became more pronounced, again as expected, when correlation among chemicals strengthened and association with the outcome was weakened.

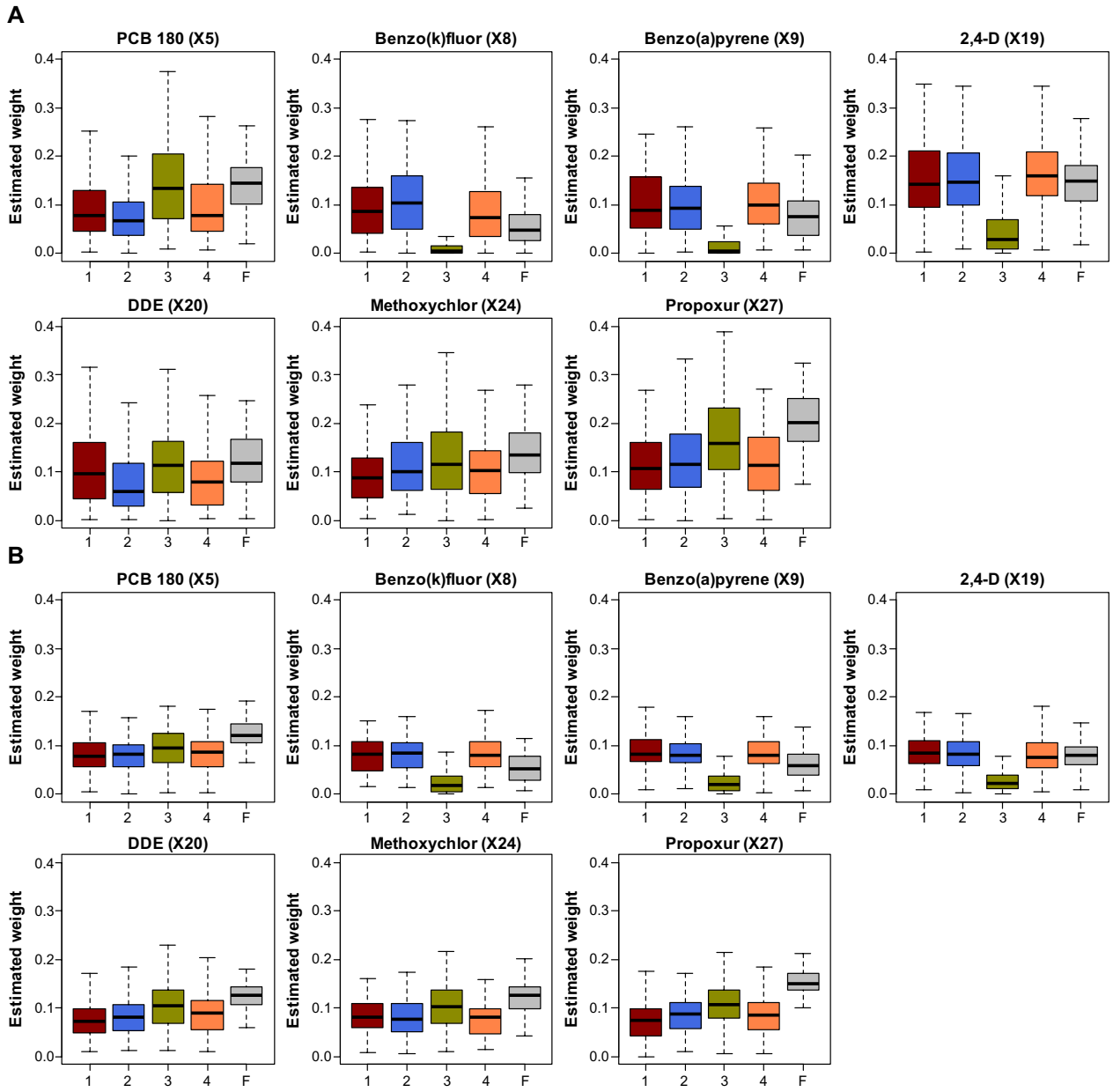

Figure 5. Distribution of WQS index weights for the five WQS indices across the 100 simulated data sets for the seven chemicals associated with the outcome. Site 1 = Detroit, 2 = Iowa, 3 = Los Angeles, 4 = Seattle, and F = full-study population. (**A**) Continuous outcome ($r = 0.1$). Correlation among chemicals diminished to 65% of the observed site-specific correlation structures. (**B**) Continuous outcome ($r = 0.1$). Correlation among chemicals diminished to 1% of the observed site-specific correlation structures.

Comparison of WQS regression with lasso, adaptive lasso, and elastic net regression. Lasso, adaptive lasso, and elastic-net regressions were performed on only the full-study population for both the binary and continuous outcomes for the six different simulation settings. The median number of correctly and incorrectly chosen chemicals for the lasso, adaptive lasso, elastic-net, and WQS regression models across 100 samples is given in Tables 6 and 7. When the predictors were strongly associated with the outcome, WQS and the traditional shrinkage methods demonstrated a high degree of sensitivity for both the continuous and binary response variables, regardless of the level of correlation among predictors. For the continuous outcome, each of the shrinkage methods correctly selected all seven chemicals at least half of the time, while WQS regression correctly selected a median of at least six of the seven chemicals. In the case of the binary outcome, each of the methods correctly chose all seven chemicals at least half of the time.

When considering the setting of weak association among the predictors and the outcome, WQS regression correctly selected six of the seven chemicals at least half of the time, for both the continuous and binary outcomes, regardless of the level of correlation among the predictors. Similarly, the median number of correctly chosen chemicals for elastic net ranged between six and seven for both the continuous and binary responses. In contrast, lasso and adaptive lasso demonstrated

Table 3. Median [IQR] number of incorrectly chosen chemicals for the five WQS indices across the 100 simulated data sets.

CONTINUOUS OUTCOME	MAX	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
		$k = 0.65$	$k = 0.30$	$k = 0.01$	$k = 0.65$	$k = 0.30$	$k = 0.01$
$WQS_{Detroit}$	20	1 [0,2]	1 [1,2]	2 [1,3]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{Iowa}	20	1 [.75,2]	2 [1,2]	2 [1,3]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{LA}	23	2 [1,3]	3 [2,3]	3 [2,4]	0 [0,0]	0 [0,0]	0 [0,0]
$WQS_{Seattle}$	20	1 [0,2]	1 [1,2]	2 [1,2]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{Full}	20	0 [0,1]	0 [0,1]	1 [0,1]	0 [0,0]	0 [0,0]	0 [0,0]
BINARY OUTCOME							
$WQS_{Detroit}$	20	2 [1,2]	2 [1,3]	3 [2,4]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{Iowa}	20	2 [1,3]	2 [1,3]	3 [2,3]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{LA}	23	3 [2,3]	3.5 [2,4]	4 [3,5]	0 [0,0]	0 [0,1]	1 [0,1]
$WQS_{Seattle}$	20	2 [1,3]	2 [2,3]	3 [2,4]	0 [0,0]	0 [0,0]	0 [0,0]
WQS_{Full}	20	0 [0,1]	1 [0,1]	1 [1,2]	0 [0,0]	0 [0,0]	0 [0,0]

diminished sensitivity, with the median number of correctly chosen chemicals ranging from four to seven for the continuous outcome and three to five for the binary outcome.

While the penalized regression models may have exhibited sensitivity that was comparable to that of WQS regression in several settings, the sensitivity of these traditional shrinkage methods was often overshadowed by their lack of specificity. WQS regression was highly specific, choosing at most a median of one incorrect chemical, regardless of the degree of correlation among predictors and regardless of the strength of association with the response. In contrast, as correlation among the chemicals increased, the penalized regression methods demonstrated a loss of specificity. In particular, when the predictors were strongly associated with the response (both continuous and binary), the penalized regression models chose a median of at least 14 incorrect chemicals in the presence of moderate or mild correlation among chemicals. Most notably, the lasso and elastic net had a tendency to select almost all

Table 4. Summary of testing results for the five WQS indices across the 100 simulated data sets for the continuous outcome variable.

	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
	$\hat{\beta}_1$	95% CONFIDENCE INTERVAL	% SIG.	$\hat{\beta}_1$	95% CONFIDENCE INTERVAL	% SIG.
$k = 0.65$						
$WQS_{Detroit}$	0.27	(0.11, 0.44)	91	1.00	(0.87, 1.13)	100
WQS_{Iowa}	0.29	(0.12, 0.46)	89	1.07	(0.94, 1.21)	100
WQS_{LA}	0.18	(0.00, 0.35)	58	0.69	(0.55, 0.83)	100
$WQS_{Seattle}$	0.31	(0.13, 0.50)	96	1.07	(0.93, 1.21)	100
WQS_{Full}	0.30	(0.20, 0.40)	100	0.94	(0.87, 1.02)	100
$k = 0.30$						
$WQS_{Detroit}$	0.39	(0.18, 0.60)	96	1.31	(1.16, 1.45)	100
WQS_{Iowa}	0.40	(0.18, 0.61)	99	1.35	(1.19, 1.51)	100
WQS_{LA}	0.26	(0.06, 0.45)	75	0.87	(0.74, 1.00)	100
$WQS_{Seattle}$	0.41	(0.15, 0.66)	94	1.35	(1.20, 1.50)	100
WQS_{Full}	0.40	(0.29, 0.51)	100	1.19	(1.11, 1.27)	100
$k = 0.01$						
$WQS_{Detroit}$	0.64	(0.32, 0.96)	99	1.91	(1.72, 2.11)	100
WQS_{Iowa}	0.62	(0.28, 0.97)	97	1.94	(1.74, 2.14)	100
WQS_{LA}	0.44	(0.17, 0.70)	87	1.28	(1.05, 1.51)	100
$WQS_{Seattle}$	0.66	(0.37, 0.96)	100	1.92	(1.74, 2.10)	100
WQS_{Full}	0.58	(0.45, 0.72)	100	1.64	(1.51, 1.76)	100

Notes: $\hat{\beta}_1$ is given as the average across 100 simulated data sets and % sig. denotes the % of data sets in which $\hat{\beta}_1$ was significant.

Table 5. Summary of testing results for the five WQS indices across the 100 simulated data sets for the binary outcome variable.

	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
	$\hat{\beta}_1$	95% CONFIDENCE INTERVAL	% SIG.	$\hat{\beta}_1$	95% CONFIDENCE INTERVAL	% SIG.
***k* = 0.65**						
$WQS_{Detroit}$	0.43	(0.07, 0.78)	61	2.00	(1.47, 2.53)	100
WQS_{Iowa}	0.45	(0.01, 0.89)	61	2.18	(1.70, 2.66)	100
WQS_{LA}	0.25	(−0.14, 0.65)	28	1.28	(0.91, 1.64)	100
$WQS_{Seattle}$	0.48	(0.06, 0.89)	67	2.16	(1.65, 2.66)	100
WQS_{Full}	0.51	(0.31, 0.72)	100	1.99	(1.73, 2.26)	100
***k* = 0.30**						
$WQS_{Detroit}$	0.65	(0.19, 1.10)	84	2.81	(2.11, 3.52)	100
WQS_{Iowa}	0.64	(0.10, 1.17)	74	2.97	(2.34, 3.60)	100
WQS_{LA}	0.35	(−0.08, 0.78)	31	1.70	(1.26, 2.14)	100
$WQS_{Seattle}$	0.62	(0.08, 1.17)	74	2.97	(2.32, 3.62)	100
WQS_{Full}	0.70	(0.45, 0.95)	100	2.67	(2.39, 2.96)	100
***k* = 0.01**						
$WQS_{Detroit}$	1.00	(0.29, 1.71)	84	4.82	(3.92, 5.72)	100
WQS_{Iowa}	0.97	(0.26, 1.67)	82	4.93	(3.95, 5.91)	100
WQS_{LA}	0.68	(0.09, 1.27)	61	2.67	(1.95, 3.38)	100
$WQS_{Seattle}$	1.03	(0.39, 1.67)	90	4.95	(3.95, 5.96)	100
WQS_{Full}	1.10	(0.77, 1.43)	100	4.21	(3.70, 4.72)	100

Notes: $\hat{\beta}_1$ is given as the average across 100 simulated data sets and % sig. denotes the % of simulated examples in which $\hat{\beta}_1$ was significant.

the chemicals when the chemicals were moderately correlated and strongly associated with the response. The relatively low specificity of these shrinkage methods appears to limit their role in risk evaluation of environmental chemical mixtures.

Discussion and Conclusion

WQS regression demonstrated good sensitivity and specificity for both site-specific models and the full-study population models across a variety of conditions considered in this study. WQS adequately detected important predictors, while simultaneously placing negligible weight on chemicals unassociated with outcome, for both continuous and binary response variables. Additionally, the WQS index was significantly and positively associated with the outcome when tested in the validation data sets, and generally demonstrated good power. Results improved as correlation among chemicals diminished and association with the outcome strengthened. In comparison to the shrinkage regression methods of lasso and elastic net, WQS performed well for sensitivity and specificity, while the lasso and elastic-net models exhibited good sensitivity but poor specificity. The shrinkage methods had a tendency to incorrectly identify a large number of components,

Table 6. Median [IQR] number of correctly selected chemicals for lasso, adaptive lasso, elastic-net, and WQs regressions across the 100 simulated data sets for the full-study population.

CONTINUOUS OUTCOME	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
	***k* = 0.65**	***k* = 0.30**	***k* = 0.01**	***k* = 0.65**	***k* = 0.30**	***k* = 0.01**
Lasso	7 [6,7]	4 [3,6]	5 [4,6]	7 [7,7]	7 [7,7]	7 [7,7]
Adaptive Lasso	6 [6,7]	4 [3,5]	5 [4,6]	7 [7,7]	7 [7,7]	7 [7,7]
Elastic Net	7 [6,7]	6 [5,7]	6 [5,7]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{Full}	6 [5,6]	6 [6,7]	6 [6,6]	6 [6,7]	7 [7,7]	7 [7,7]
BINARY OUTCOME						
Lasso	5 [3,6]	3 [0,5]	3 [0,5]	7 [7,7]	7 [7,7]	7 [7,7]
Adaptive Lasso	5 [4,6]	3 [1,5]	3 [2,5]	7 [7,7]	7 [7,7]	7 [7,7]
Elastic Net	7 [6,7]	7 [5.75,7]	6 [4,7]	7 [7,7]	7 [7,7]	7 [7,7]
WQS_{Full}	6 [6,6]	6 [6,7]	6 [6,7]	7 [7,7]	7 [7,7]	7 [7,7]

Table 7. Median [IQR] number of incorrectly selected chemicals for lasso, adaptive lasso, elastic net, and WQs regressions across the 100 simulated data sets for the full-study population.

CONTINUOUS OUTCOME	WEAK ASSOCIATION WITH OUTCOME			STRONG ASSOCIATION WITH OUTCOME		
	$k = 0.65$	$k = 0.30$	$k = 0.01$	$k = 0.65$	$k = 0.30$	$k = 0.01$
Lasso	7 [4,10]	0 [0,0.25]	0 [0,0]	20 [19,20]	18 [17,19]	0 [0,0]
Adaptive Lasso	4 [2.75,6]	0 [0,0]	0 [0,0]	17 [17,18]	14 [13,16]	0 [0,0]
Elastic Net	8 [5,11]	0.5 [0,5]	0 [0,1]	20 [19,20]	18.5 [18,19]	0 [0,0.25]
WQS_{Full}	0 [0,1]	0 [0,1]	1 [0,1]	0 [0,0]	0 [0,0]	0 [0,0]
BINARY OUTCOME						
Lasso	2 [0,7]	0 [0,0]	0 [0,0]	19 [19,20]	16 [15,17]	0 [0,1]
Adaptive Lasso	2 [1,4]	0 [0,0]	0 [0,0]	16 [15,16.25]	11 [9,13]	0 [0,0]
Elastic Net	10 [5,20]	5 [0,20]	0.5 [0,20]	19 [19,20]	18 [16,19]	1 [0,2]
WQS_{Full}	0 [0,1]	1 [0,1]	1 [1,2]	0 [0,0]	0 [0,0]	0 [0,0]

especially in the case of strong association with the outcome. This suggests that these methods may be limited for use in risk assessment, as they are unable to discern which chemicals are unassociated with health risk.

The WQS index weights for the full-study population demonstrated an averaging effect, suggesting that chemical weights estimated in an overall analysis may not be representative of the true bad actors within a site. Three chemicals were deemed unassociated with outcome in the Los Angeles site, as they were not present in high enough concentrations to satisfy the imposed definition of health risk. The overall analysis consistently identified these three chemicals unassociated with outcome in Los Angeles as bad actors. While this is representative of the data as a whole (these chemicals were set as truly bad actors in three of the four sites), it is not an accurate representation of the chemicals posing risk in Los Angeles. Additionally, the average weights assigned to these three chemicals by the index in the full-study population were lower in comparison to the weights assigned by the indices in Detroit, Iowa, and Seattle. The non-association in Los Angeles, therefore, seems to result in an underestimation (by the full index) of the importance of these chemicals in the sites in which they truly were bad actors.

With the goal of identifying chemicals that pose a significant health risk, it is of great importance to consider the toxicological principle that "the dose makes the poison," especially given that exposure patterns are spatially varying. Although a chemical may not be present in high enough concentrations to pose a health risk in one location, it may still pose a significant health risk at other locations. Though limited, these simulation studies suggest that use of an overall index may overstate the importance of a chemical in sites where the concentration is too low to constitute risk and may understate the importance of a chemical in locations where it is present in concentrations that are high enough to adversely affect health.

The simulation studies conducted in this analysis were largely reflective of the exposure patterns observed in the original NCI-SEER NHL study, incorporating the exposure concentrations and the complex correlation among chemicals on a site-specific basis. However, simulation of the data utilized Cholesky decomposition, which required that the covariance matrices be positive definite. As a result, the simulations only incorporated (at most) 65% of the observed correlation structures, as the covariance matrices became singular if they were any less diminished. Other studies have used methods such as ridging to allow the correlation to be persevered or even inflated,[16] which should be considered in future work. We expect that as correlation among chemicals increases, one will encounter cases in which WQS may not perform as well as was in this study. Finally, in the general context of risk assessment, it is a limitation that the observed chemical concentrations in the NCI-SEER NHL study were external measures of exposures, as what is found in house dust may not be truly reflective of an individual's absorption or ingestion of chemicals.

Acknowledgments

We greatly thank the following individuals involved with the study design and data collection of the NCI-SEER NHL study: Anneclaire J. De Roos (Drexel University School of Public Health), James R. Cerhan (Mayo Clinic), Richard K. Severson (Wayne State University), Wendy Cozen (University of Southern California), Patricia Hartge (NCI), Joanne Colt (NCI), and Mary Ward (NCI).

Author Contributions

Conceived and designed the experiments: JC, CG, DCW. Analyzed the data: JC. Wrote the first draft of the manuscript: JC. Contributed to the writing of the manuscript: JC, CG, DCW. Agree with manuscript results and conclusions: JC, CG, DCW. Jointly developed the structure and arguments for the paper: JC, CG, DCW. Made critical revisions and

approved final version: JC, CG, DCW. All authors reviewed and approved of the final manuscript.

REFERENCES

1. Whitehead T, Metayer C, Gunier RB, et al. Determinants of polycyclic aromatic hydrocarbon levels in house dust. *J Expo Sci Environ Epidemiol*. 2009;21:1–10.
2. Whitehead TP, Metayer C, Ward MH, et al. Persistent organic pollutants in dust from older homes: learning from lead. *Am J Public Health*. 2014;104(7):1320–6.
3. DellaValle CT, Wheeler DC, Deziel NC, et al. Environmental determinants of polychlorinated biphenyl concentrations in residential carpet dust. *Environ Sci Technol*. 2013;47(18):10405–14.
4. Czarnota J, Gennings C, Colt JS, et al. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study. Environ Health Perspect; http://dx.doi.org/10.1289/ehp.1408630.
5. Carrico C, Gennings C, Wheeler D, Factor-Litvak P. Characterization of a weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Biol Agricul Environ Stat*. 2014:1–21. ISSN: 1085-7117. DOI: 10.1007/s13253-014-0180-3. http://dx.doi.org/10.1007/s13253-014-0180-3.
6. Colt JS, Lubin J, Camann D, et al. Comparison of pesticide levels in carpet dust and self-reported pest treatment practices in four US sites. *J Expo Anal Environ Epidemiol*. 2004;14:74–83.
7. Wheeler DC, De Roos AJ, Cerhan JR, et al. Spatial-temporal analysis of non-Hodgkin lymphoma in the NCI-SEER NHL case-control study. *Environ Health*. 2011;10:63.
8. Colt JS, Severson RK, Lubin J, et al. Organochlorines in carpet dust and non-Hodgkin lymphoma. *Epidemiology*. 2005;16(4):516–25.
9. Lubin JH, Colt JS, Camann D, et al. Epidemiologic evaluation of measurement data in the presence of detection limits. *Environ Health Perspect*. 2004;112:1691–6.
10. Ghalanos A, Theussl S. 2012. Rsolnp: general non-linear optimization using augmented Lagrange multiplier method. R package version 1.14.
11. Tibshirani R. Regression shrinkage and selection via the lasso. *J Royal Stat Soc*. 1996;58(1):267–88.
12. Zou H. The adaptive lasso and its oracle properties. *J Am Stat Assoc*. 2006;101:1418–29.
13. Zou H, Hastie T. Regularization and variable selection via the elastic net. *J Royal Stat Soc B*. 2005;67(2):301–20.
14. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *J Stat Softw*. 2010;33(1):1–22.
15. Breiman L, Friedman J, Olshen R, Stone C. *Classification and Regression Trees*. Pacific Grove: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.
16. Carrico C. Characterization of a weighted quantile score approach for highly correlated data in risk analysis scenarios [VCU Theses and Dissertations]. Biostatistics: Virginia Commonweath University; 2013.