2023

Model-based Imputation of Below Detection Limit Missing Data and Group Selection in Bayesian Group Index Regression

Matthew Carli
Virginia Commonwealth University

Model-based Imputation of Below Detection Limit Missing Data and Group Selection in Bayesian Group Index Regression

A dissertation submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Biostatistics at Virginia Commonwealth University

Matthew Carli

Department of Biostatistics

Virginia Commonwealth University

Richmond, Virginia

July 26, 2023

Director:

David C. Wheeler, Ph.D., M.P.H., Department of Biostatistics

Committee Members:

Nitai D. Mukhopadhyay, Ph.D., Department of Biostatistics

Roy T. Sabo, Ph.D., Department of Biostatistics

Bernard F. Fuemmeler, Ph.D., M.P.H., Department of Health Behavior and Policy

Hua Zhao, Ph.D., Division of Epidemiology

Acknowledgement

I am indebted to many people without whose help this dissertation would not have been possible:

To my advisor, Dr. Wheeler, whose guidance was crucial to any success I can claim as a student. He has been a model of professional dedication and through his mentorship I have become a better statistician.

To the other members of my committee – Dr. Mukhopadhyay, Dr. Sabo, Dr. Fuemmeler, and Dr. Zhao – for their generous commitment of time and feedback during the course of my dissertation work.

To Dr. Thacker, for his interest in me as a student and person, and for the good times working under him as a Biostats consultant.

To Helen Wang, who has been a constant source of help and advice whenever I needed to use the departmental computer cluster.

To my graduating class – Alicia Richards, Martin Lavelle, and Xinxin Sun – for making that tough first year doable and for their continued support and friendship.

To Alex Karanevich, John Stansfield, Salem Rustom, and all the guys at St. Joseph's Catholic Church: friends unlooked for on the way.

To my family, especially Mom, who has always been there.

Finally, to my wife, my best friend and constant support.

Abstract

Model-based Imputation of Below Detection Limit Missing Data and Group Selection in Bayesian Group Index Regression

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Biostatistics at Virginia Commonwealth University

Matthew Carli, Virginia Commonwealth University, July 2023

Major Director: David C. Wheeler, Ph.D., M.P.H., Associate Professor, Department of Biostatistics

Investigations into the association between chemical exposure and health outcomes are increasingly focused on the role of chemical mixtures, as opposed to individual chemicals. The analysis of chemical mixture data required the development of novel statistical methods, one of these being Bayesian group index regression. A statistical challenge common to all chemical mixture analyses is the ubiquitous presence of below detection limit (BDL) data. We propose an extension of Bayesian group index regression that treats both regression effects and missing BDL observations as parameters in a model estimated through a Markov Chain Monte Carlo algorithm that we refer to as Pseudo-Gibbs imputation. The Pseudo-Gibbs approach enables the estimated parameters of the health effects model to inform the missing data imputations and vice versa, as well as accounting for the true variance of the BDL imputations. We conduct a simulation study showing greater power to detect chemical indices significantly associated with an outcome and sensitivity for identifying important chemicals within indices at high levels of BDL missing data. We apply our model to a case-control study on the effects of chemical exposure on childhood leukemia. We next address a problem specific to group index models: how to partition a given set of chemicals into groups to form the requisite indices. We first proposed a novel variable clustering algorithm using a variant on the traditional PCA algorithm called Robust PCA. We compared this clustering method with other variable clustering methods from the literature using a simulation study. Finally, we extended the variable clustering method identified previously to incorporate information from an outcome variable. This semi-supervised clustering extension incorporates the ability to constrain clusters based on the direction of association of individual chemicals with the outcome of interest. We apply both unsupervised variable clustering and semi-supervised clustering methods identified to a case-control study on the effects of chemical exposure on non-Hodkin's lymphoma.

Vita

Matthew Carli was born on September 25, 1990 in Glens Falls, New York. He graduated from Saratoga Springs High School in Saratoga Springs, New York in 2009. He received his Bachelor of Arts in English Literature and Chinese Language from Washington and Lee University in Lexington, Virginia in 2013.

List of Tables and Figures

Table of Contents

Introduction

A facet of modern life that is of increasing interest to both researchers and the general public is the unavoidable reality of exposure to chemicals. Whether introduced through agriculture [1], industry [2], or household items [3], the ubiquity of chemicals that are not fully understood has become a widespread cause for concern. Such concerns are not without merit, as commercial chemicals have been found in human tissues and in household air and dust samples in varying concentrations [4-5], motivating questions as to their impact on human health. Epidemiologic studies have identified environmental chemical exposure as a risk factor in a number of human diseases, including cancer, type 2 diabetes, cardiovascular disease, thyroid disease, and developmental disorders [6-10]. While the inquiry into potential environmental risk factors for disease has been valuable, historically studies have taken a simplifying approach in their investigations. Single-chemical regression studies that evaluate the association of individual chemicals with a health outcome have predominated. Some consider the total (i.e. summed) exposure for a chemical class, such as polychlorinated biphenyls (PCBs) [11]. Another such approach is the environment-wide association study (EWAS), where multiple chemical exposures from a single chemical family [12] or from many chemical families [13] are evaluated for association with the outcome independently, which are finally adjusted for multiple comparisons. A drawback of these approaches is that effects for simultaneous exposures cannot be estimated.

Increasingly, investigations into the health impact of chemical exposures highlight the fact that they exist as mixtures or combinations of many simultaneous exposures [14]. Therefore, epidemiologists and biostatisticians have sought to assess the joint impact of chemical mixtures on health outcomes, as opposed to estimating chemicals as independent risk factors [15-17]. A difficulty posed in the analysis of chemical mixtures is the lack of statistical independence among exposures. Correlations between chemicals of interest can range from close to zero to near perfect positive correlation, resulting in the poor performance of standard regression techniques [18]. Specifically, malperformance due to collinearity includes sign reversal and inflated standard errors of estimated regression coefficients. These problems can lead to erroneous conclusions about the health effects of particular chemicals and must be accounted for in any joint chemical analysis.

As interest in the problem has grown, various methods to jointly analyze chemical mixture data have been developed. Bayesian kernel machine regression (BKMR) utilizes a kernel function to handle highly correlated

chemical variables and relate them to a health outcome. An advantage of BKMR is the ability to investigate non-linear relationships and interaction effects of chemical predictors within mixtures [19]. Limitations of the BKMR method include heavy computational burden when analyzing large datasets [20-21] and the necessity of fixing most chemical exposure values when investigating the association of one or two chemicals within the mixture with the outcome [22-23]. Quantile g-computation is a method which draws from g-computation ideas found in causal effect estimation literature in order to jointly estimate the effect of chemicals and accommodate scenarios where chemicals have opposing effect directions or nonlinear effects [24]. A drawback of this quantile g-computation is that models are restricted to a single index, which in the presence of both positively and negatively correlated chemical variables could result in either an effect estimate dominated by one direction or attenuated to the null. Additionally, the method relies on multiple imputation procedures to deal with below detection limit missing data, which may be cumbersome to apply for many users [25].

Another method of interest, weighted quantile sum (WQS) regression, constructs a weighted index where important chemicals contribute relatively more to an overall score, which can then be used as an estimation of the mixture's effect [26-27]. Individual component chemicals are each given a weight to identify the most important components within the group. These weights are constrained to be between 0 and 1 and sum to 1. This method was later extended by the addition of a bootstrapping step for the estimation of the weights [18]. The advantage of WQS regression over traditional regression methods is that it allows for the highly correlated data commonly found in chemical mixtures to be analyzed while avoiding collinearity issues, and has been shown to have good sensitivity and specificity when identifying important exposures [18,28]. One limitation of WQS regression is that the index effect is constrained to one direction, positive or negative. This constraint does not reflect the reality of all chemical mixtures, which can have certain chemicals positively associated and others negatively associated with the same outcome. A second limitation is that estimates are made in a two-step, data-splitting process, where weights are empirically estimated and weighted indices formed from training data and index effect parameters subsequently estimated from validation data. This reliance on data-splitting reduces power in smaller studies.

To address the limitations of WQS regression, various extensions were made that we will refer to generally as group index regression methods. The first of these, grouped weighted quantile sum (GWQS) regression, extends the WQS model to allow for multiple chemical indices, each of which can have different magnitudes and direction of

association with the outcome [29-30]. This method is distinct from fitting a positive and negative WQS regression model separately to the same data, which requires two models in comparison to a single GWQS model. GWQS regression was shown in simulations to outperform WQS and other traditional regression methods when the chemicals analyzed contained more than one group of exposures with different health effects [31].

A further extension to index regression methods was the formulation of WQS regression in the Bayesian framework [32-33]. The Bayesian index model eliminates the need for data-splitting, as index effect estimates and their weights are estimated together as parameters in the Markov chain Monte Carlo (MCMC) algorithm. A second advantage of Bayesian index regression is that Bayesian models are generally more flexible than their frequentist counterparts. For example, spatial random effects and exchangeable random effects are readily incorporated into the Bayesian framework, and have been applied to investigate neighborhood deprivation and risk of elevated blood lead levels [33,34] and tobacco retail outlet rates [35]. GWQS regression was then also implemented in the Bayesian framework, which we will refer to as Bayesian group index regression. In simulations comparing Bayesian group index regression and GWQS regression, Bayesian group index regression was found to have slightly improved sensitivity, specificity, and power for finding significant effects [36].

A challenge common to nearly all chemical exposure analyses is the presence of below detection limit (BDL) missing observations. BDL missing data are an artifact of laboratory analysis, where any level of chemical analyte below a certain detection limit (DL) cannot be reliably measured [37]. Many solutions have been proposed to deal with this problem, and can be sorted into the following categories. The simplest fall under ad hoc substitution methods, where the BDL is replaced by 0, the DL, or some function of the DL (DL/2 being a common example). While such substitutions are easily implemented, they have been shown to result in biased parameter estimates and variances [38-40]. Another category of methods, single imputation (SI), encompasses a wide variety of imputation methods ranging from nonparametric Kaplan-Meier [41] to parametric maximum likelihood estimate methods [42], and have been shown to generate superior imputations than substitution methods. What these methods all share is that the imputations generated from them are afterwards treated as true observations, and as such fail to account for the variability of the imputation process [39]. An approach that was developed to remedy this limitation of SI methods is multiple imputation (MI). Where SI methods will impute all missing observations once to achieve a complete dataset

that can then be analyzed, MI calls for many of these complete datasets to be generated. Each dataset is then analyzed individually, generating a set of repeat parameter estimates. These repeat estimates are then pooled, giving a final set of estimates that reflect the between-imputation variability of the missing BDL observations [39].

Alternatively, imputation of BDLs can be done in a Bayesian framework. The MCMC algorithm allows for the treatment of missing observations as parameters to be estimated, allowing for the uncertainty of imputed values similar to that of MI [43]. Further, imputation models can be combined with analysis models, allowing for integration with Bayesian group index models. The most straightforward method of imputing missing covariate data is by drawing imputations jointly from a multivariate distribution [44], often a multivariate normal or *t* distribution. A joint distribution can be hard to define, however, when covariates containing missing data are diverse (a combination of continuous and binary variables, for example) or when non-normal models are required. The imputation of BDLs is an instance of the latter, as these bounded variables are best modelled by truncated distributions. One method developed to deal with these difficult covariate groupings is Fully Conditional Specification (FCS), which imputes missing observations one covariate at a time by a univariate conditional distribution, conditioned on all other variables in the model [45]. A common criticism of FCS is the potential for the various univariate conditional distributions to be incompatible, that is to fail to converge to any joint distribution [46-47]. Incompatibility can result in unsound imputations and biased estimates [48]. An alternative imputation method which addresses the issue of potential incompatibility is what we will refer to as Sequential Full Bayes (SFB) imputation. Similar to FCS, univariate conditional distributions for each covariate containing missing observations are used, but in this instance in order to factorize the joint distribution as a product of all the conditional distributions [49]. In this manner, the joint distribution of the imputation model is specified, avoiding any issues of incompatibility.

While BDL imputation is a problem common to all mixture analyses, there are also challenges inherent to group index models. One such challenge is determining the chemical groups that will form indices. When fitting a multi-group index regression model, either GWQS or Bayesian group index regression, the number and chemical composition of indices must be chosen. Past applications of such models have organized exposure variables into chemicals which share a structural similarity (e.g., PCBs, PAHs, metals) or usage (e.g., herbicides, insecticides) [31,36]. This grouping strategy could be viewed as one reliant on domain-specific knowledge, and has several advantages. Chemicals that are similar in either structure or use have a greater chance of being highly correlated with each other,

and if not grouped could give rise to multicollinearity effects. Indices grouped in this way also have ready interpretations as the joint effect of a recognizable class of chemicals on a health outcome. There are, however, some limitations to groups formed in this manner. It is not always the case the chemicals within the same structure or usage class belong in the same group. Pesticides, for example, are a relatively heterogenous group of chemicals that may ideally be split into more than one group or combined with other groups. Groups that mix chemicals with positive and negative associations with an outcome of interest are particularly problematic in the context of group index models, as the opposing direction of association will artificially bias such an index to the null. Additionally, there may be patterns in a chemical mixture beyond that of chemical structure or usage that an empirical measure of similarity would be able to ascertain. In these situations empirically derived groupings could not only provide better fitting models, but also identify and characterize previously unknown predictor relationships. It is therefore of interest to develop a method of grouping chemicals into indices based on some empirically-driven methodology.

The problem of selecting group composition in a Bayesian group index regression model can be viewed as a cluster analysis problem. Cluster analysis encompasses a wide variety of methods employed for different reasons, but all share the common aim of grouping similar items together [50]. The goal is to capture some underlying mechanism at work in the data that causes some observations to have greater resemblance to each other than to other observations [51]. Various categories of clustering methods have been developed, including partitional, hierarchical, density-based, grid-based, model-based, and discriminative algorithms [52-53]. A distinction between clustering methods relevant to preparation for group index regression is hard versus soft (or fuzzy) clustering algorithms. Hard clustering groups objects with strictly defined boundaries, forcing membership of each object into a single group. Soft clustering allows objects to potentially belong to many groups [54]. This flexibility is advantageous in situations where an object's cluster membership is unclear or where clusters overlap [55]. In application to a chemical mixture before group index regression, however, this potential for multiple group membership is undesirable, as Bayesian group index regression requires chemicals be assigned to only one group. Another important distinction between clustering methods we must consider is subject clustering algorithms as opposed to variable clustering algorithms. The majority of clustering algorithms are subject clustering algorithms, where samples are partitioned into similar groups. Variable clustering algorithms, on the other hand, partition a dataset into groups of features. Subject clustering algorithms can of course be used to cluster variables by transposing the data matrix in question, with

some examples including k-means [56], hierarchical clustering [57-58], self-organizing maps [59], and model-based approaches [60-61]. However, subject clustering methods generally use some sort of measure of distance (such as Euclidean distance) to quantify similarity, as opposed to a measure such as correlation that is more suitable for variables [62].

The task of grouping variables is closely related to dimension reduction, where a complex set of large numbers of variables are assumed to be governed by a much simpler system of a few hidden or latent variables [63]. Motivated by the analytical challenge of extremely high variable numbers seen in such fields as genomics [64], researchers have developed methods specifically designed for variable clustering. One example of this is Dirichlet Process Variable Clustering (DPVC), a model-based clustering method that partitions a set of variables according to the Chinese restaurant process (CRP). The CRP defines a distribution over a number of partitions that does not need to be specified by the user, but is estimated during the model-fitting process. The partitioning restricts each variable to membership in a single cluster, which are represented as a normally-distributed latent factor [62]. Principal component analysis (PCA) is a widely used dimension reduction technique that has inspired variable clustering methods. PCA is used to reduce the number of predictor variables and avoid modelling problems due to multicollinearity. PCA does this by creating new latent variables that are linear combinations of the original features of the data, which can then be subsequently used in regression, commonly referred to as PCA regression. While these latent variables are formed in such a way as to capture the maximum variation of the constituent data, interpreting the results of PCA regression can be difficult, as a single variable may account for some portion of the variance captured in multiple latent variables. Some variable clustering algorithms found in the literature seek to use PCA's ability to summarize multiple variables into latent variables while at the same time establishing hard partitions between variable clusters. Clustering of Variables around Latent Components (CLV) achieves this by maximizing the covariance between variables and the first principal component derived from iteratively shifting subsets of the dataset's variables [65]. A critique of CLV is that it only uses the first principal component as a latent variable, a potential underutilization of variance explained by the inclusion of further components that may better characterize the group's underlying structure. To address this limitation, the VARCLUST algorithm extended CLV to allow for cluster latent variable dimensions greater than one and the estimation of the dimensionality of each cluster [66]. Other researchers have used the ability to represent the PCA algorithm as a matrix decomposition to formulate novel methods. One such

method is robust PCA (RPCA), where the matrix decomposition seen in standard PCA is subject to further constraints, generating a denoised data matrix that can be used to cluster variables [67]. An extension to RPCA has extended the algorithm to specifically accommodate the characteristics of chemical mixture data, such as the common occurrence of BDL missingness [68]. These variable clustering methods and the potential for new method development offered by such algorithms as RPCA present opportunities for improved, empirically justified chemical groupings for use in Bayesian group index regression.

A final consideration in clustering methods of interest for Bayesian group index regression is the ability of an algorithm to incorporate information from the outcome variable into the generated clusters, as the end goal of analysis will always be to estimate the associations between the groups modelled and an outcome. Clustering algorithms have historically been unsupervised [69], meaning there is no label or information outside the objects themselves that would inform "true" clustering assignments [70-71]. Opposite unsupervised learning methods are supervised methods, defined as methods that generate a function to map input variables to a desired output variable [72]. Some examples include regression, random forest, and support vector machines [73]. An intermediate category, referred to as semi-supervised learning, aims at some combination of unsupervised and supervised methods [74], where information other than the labels normally seen in a supervised setting can be used to extend unsupervised methods [75]. This additional information can take many forms, such as previous partial classification of a subset of inputs or, of particular interest in our context, the relationship between inputs and an outcome variable [76]. An example is the "supervised clustering" algorithm of Bair and Tibshirani 2004, where clustering of variables is only performed on variables with a univariate association test statistic with the outcome that exceeds some cutoff value. This focus on the magnitude of association with the outcome is meant to prevent highly identifiable clusters that are nonetheless relatively unrelated to the outcome from interfering with the discovery of more relevant clusters [77]. Another semi-supervised method is known as constrained clustering, clustering where partial data in the form of user-provided labels or pairwise constraints are used to guide the algorithm towards a more appropriate data partitioning. These constraints are commonly in the form of must-link or cannot-link pairs [78]. A constraint that is crucial for the clustering of variables in Bayesian group index modelling is that variables with opposite direction of association with an outcome are not grouped together. Each of these semi-supervised clustering methods could easily be adapted to a previously unsupervised algorithm. Another way in which outcome information can be

included in the clustering process is through the combination of clustering and regression models, as seen implemented in Clusterwise Effect Regression (CLERE). The CLERE model redefines the beta parameter of traditional regression models as an unobserved random variable following a mixture of Gaussian distributions containing some number of input variables [79]. While the CLERE algorithm estimates its own regression coefficients for variables, we are primarily interested in the its cluster assignments that are informed by the regression's supervision by the outcome variable. Adapting unsupervised clustering methods and identifying those that perform best on chemical mixture data could allow for superior Bayesian group index regression models.

Our description above of the current state of mixture analysis and the challenges facing these analyses reveal opportunities to provide important contributions to statistical practice. Our overview of some relevant literature also points to potential solutions to these problems. We have identified three ways in which Bayesian group index regression can be extended or complemented to improve its performance and usability: the imputation of BDL missing data in combination with group index regression, the development and identification of variable clustering algorithms well-suited to chemical mixture data, and the extension of unsupervised clustering methods to incorporate information from outcome variables of interest. These aims are enumerated in more detail below.

Dissertation Specific Aims

This dissertation will address the following research aims.

Specific Aim 1: Extend Bayesian group index regression to the imputation of BDLs.

Specific Aim 2: Develop and identify the variable clustering method best suited for use in chemical mixture analysis with Bayesian group index regression.

Specific Aim 3: Develop semi-supervised extension to previously identified variable clustering method and identify method best suited for use in chemical mixture analysis with Bayesian group index regression.

Specific Aim 1: Extend Bayesian group index regression to the imputation of BDLs.

In the analysis of chemical mixture data, BDL missing observations are a commonly encountered problem. Although a variety of methods have been employed to impute these missing observations [38,40-42,80-81], MI methods are regarded as the best practice, as they incorporate the variance of the truly unknown status of BDL missing observations into subsequent parameter estimates [39]. MI methods accomplish this by generating multiple copies of a dataset, which are then imputed and analyzed in parallel, resulting in a set of parameter estimates for each imputed dataset. These sets of estimates are then pooled to arrive at the final parameter estimates. MI methods following this algorithm of parallel multiple imputation and pooling have been implemented in both the frequentist [39] and Bayesian frameworks [82-83]. Alternatively, the Bayesian framework allows for the specification of an imputation model in combination with the analysis model, so that missing BDL data is imputed and analysis model parameters are estimated simultaneously in the MCMC algorithm [84]. As BDL imputations are drawn from a specified prior distribution at each iteration of the MCMC algorithm, the uncertainty of these repeated imputations is reflected in the posterior distribution found at convergence [85]. The most straightforward way to implement these imputation models is by drawing imputations from a multivariate distribution such as a normal or *t* distribution [44]. Unfortunately, not all missing data are amenable to imputation through such relatively simple joint distributions. The specification of a joint distribution is difficult when missing data covariates are composed of diverse data types, such as a combination of continuous and binary variables, or when non-normal models are required. BDL missing data is an example of the latter, as the imputations bound between zero and the DL are best modelled by truncated lognormal distributions. Sampling from such multivariate truncated distributions is difficult and computationally expensive [86].

An alternative to joint multivariate imputation distributions that avoids these difficulties is the expression of a joint multivariate distribution through conditional univariate distributions. One such method, Fully Conditional Specification (FCS), imputes missing observations one covariate at a time by a univariate conditional distribution, conditioned on all other variables in the model. Each variable in the model is cycled through in this fashion until convergence to an assumed but unspecified joint posterior distribution is reached [45]. This assumption has been criticized, with the concern that failure to converge to a joint distribution may result in poor imputations and biased parameter estimates [47-48]. This theoretical concern is not considered serious by some authors [46,87], who point out that FCS has performed well in simulations and has shown to be robust to incompatibility in some scenarios [45]. Another method that seeks to specify a joint multivariate distribution using conditional univariate distributions is Sequential Full Bayes (SFB) imputation [49]. SFB differs from FCS in that the univariate conditional distributions are used in order to factorize the joint distribution as a product of all the conditional distributions [88]. In this manner, the joint distribution of the imputation model is specified, avoiding any issues of incompatibility.

In Aim 1, I propose to extend Bayesian group index regression to include FCS or SFB as imputation models to model BDL missing data. These imputation methods will be compared to the Multiple Imputation by Chained Equations (MICE) algorithm and to a SI imputation method in a simulated case-control study. We will evaluate their relative performance in terms of mean squared error (MSE), bias, and power in estimating group exposure effects, their sensitivity and specificity in identifying important chemicals within indices, and by comparing the deviance information criterion (DIC) and computation times of each model. I will also apply the imputation method identified by the simulation study to the California Childhood Leukemia Study (CCLS) in order to investigate the link between chemical mixtures found in house dust and childhood leukemia.

Specific Aim 2: Develop and identify the variable clustering method best suited for use in chemical mixture analysis with Bayesian group index regression.

The extension of single index weighted quantile sum regression to group index regression methods requires that a chemical mixture be partitioned into groups. Historically this has been done along the lines of shared chemical structure or usage [31,36,89]. In order to avoid undesirable chemical groupings and to justify the similarity of chemicals grouped with some objective measure, we consider the problem of group composition from the point of view of data clustering. Clustering is a broad field that encompasses many different approaches and methods. The

requirement of Bayesian group index regression that all chemicals be a member of only a single group excludes soft clustering methods from consideration in favor of hard clustering methods. Further, as we are grouping variables, not subjects, we focus on clustering methods developed for this purpose. The measure of similarity used in variable clustering methods, usually some measure of covariance or correlation, is more apt to the clustering of variables than the measures of distance normally employed in subject clustering algorithms [62]. Additionally, recent developments in data preprocessing for chemical mixture data offer the opportunity to develop variable clustering methods suited particularly for chemical mixture variables [68].

In Aim 2, I propose a novel variable clustering method that utilizes a variant of the RPCA algorithm optimized for use with chemical mixture data. I hypothesize that such a clustering algorithm will have superior performance over other variable clustering algorithms not designed with the characteristics of chemical mixture data in mind. To evaluate the proposed method, I will design a simulation study that will encompass a range of true group numbers and levels of background or noise correlations. The proposed method and a number of variable clustering algorithms taken from my literature review will be used to cluster the simulated data, after which Bayesian group index regression will be run with the previously derived groupings. The accuracy of generated clusters, as well as the quality of group index parameter estimates, will be the criteria of comparison. Additionally, I will apply the variable clustering method with the strongest performance along with Bayesian group index regression to the NCI-SEER NHL study in order to estimate the association between the chemical mixture found is study participants' house dust and NHL.

Specific Aim 3: Develop semi-supervised extension to previously identified variable clustering method and identify method best suited for use in chemical mixture analysis with Bayesian group index regression.

Clustering methods are generally known as unsupervised algorithms [69], due to their exploratory nature and the common lack of any set of labels for what would constitute "true" cluster assignments [71]. Nonetheless, there are in certain situations information available that, while it does not constitute the set of labels required for a supervised algorithm, when incorporated into the clustering process can improve the cluster assignments generated. This application of partial information to the clustering process is referred to as semi-supervised clustering [76]. In the case of clustering in preparation for Bayesian group index regression, where the association between chemical groups and a target outcome of interest will be estimated, it follows that information from the outcome variable will improve the clustering of chemical exposure variables. Of particular interest is the ability of a semi-supervised

clustering algorithm to prevent chemicals with opposite directions of association with the outcome from being grouped together. One method of semi-supervised clustering that used outcome variable information is referred to in the literature as "supervised clustering" [77]. This method works by defining a test statistic cutoff value, whereby any variable with an association strength less than the cutoff is not considered for clustering. If clustering in performed only on highly associated variables, it may improve the ability of clustering algorithm to separate those with opposite directions of association, while also generating clustering most relevant to the target outcome variable. A criticism of this method of semi-supervision is that it discards the variables that do not exceed the cutoff, however, in a group index application these usually discarded variables can be placed into a "null" index. Another semi-supervised clustering method that can incorporate information from an outcome variable is constrained clustering [78]. Constrained clustering methods allow for the user to define pairs of variables that should or should not be clustered. These pairs are sometimes referred to as must-link and cannot-link pairings. When these pre-defined restrictions are violated, a penalty is imposed, disincentivizing these pairings from occurring. In the case of clustering in preparation for group index regression, we can use cannot-link pairings to discourage the grouping of chemicals oppositely associated with an outcome variable.

In Aim 3, I propose to extend the unsupervised clustering algorithm identified in Aim 2 to incorporate information from the outcome variable during clustering. I propose two semi-supervised clustering methods: one using the "supervised clustering" method of semi-supervision to focus the clustering algorithm on highly associated chemicals, and the other using constrained clustering to discourage the clustering of oppositely associated chemicals. I will evaluate the proposed extensions using a simulation study, where they will be compared to the unsupervised algorithm they are based on and two other semi-supervised clustering methods: Clusterwise Effect Regression (CLERE) and Conclust. The simulated data will generally be structured so that little distinguishes distinct chemical groups besides their association with the outcome variable. The number of true groups and level of correlation noise between distinct groups will be varied between simulated scenarios. The groups generated by the clustering algorithms will subsequently be used in Bayes group index regression models. The performance of the clustering methods will be compared by their accuracy and by the quality of group index parameter estimates. The semi-supervised method with the best performance will then be applied to the NCI-SEER NHL study in order to estimate the association between the chemical mixture found is study participants' house dust and NHL.

References

1. Savci S. Investigation of Effect of Chemical Fertilizers on Environment. APCBEE procedia. 2012;1:287-292. doi:10.1016/j.apcbee.2012.03.047

2. Landrigan PJ, Fuller R, Acosta NJR, et al. The Lancet Commission on pollution and health. The Lancet (British edition). 2018;391(10119):462-512. doi:10.1016/S0140-6736(17)32345-0

3. Khalil, Memoona, et al. "Household chemicals and their impact." Environmental micropollutants. Elsevier, 2022. 201-232.

4. Centers for Disease Control and Prevention, 2009. Fourth National Report on Human Exposure to Environmental Chemicals. Available: http://www.cdc.gov/ExposureReport/pdf/FourthReport.pdf (accessed on 2/22/21).

5. Rudel, Ruthann A et al. “Semivolatile Endocrine-Disrupting Compounds in Paired Indoor and Outdoor Air in Two Northern California Communities.” *Environmental science & technology* 44.17 (2010): 6583–6590.

6. Airaksinen, Riikka et al. “Association Between Type 2 Diabetes and Exposure to Persistent Organic Pollutants.” *Diabetes care* 34.9 (2011): 1972–1979.

7. Anand, Preetha et al. “Cancer Is a Preventable Disease That Requires Major Lifestyle Changes.” *Pharmaceutical research* 25.9 (2008): 2200–2200.

8. Brent, Gregory A. “Environmental Exposures and Autoimmune Thyroid Disease.” *Thyroid (New York, N.Y.)* 20.7 (2010): 755–761.

9. Grandjean, Philippe, and Philip J Landrigan. “Neurobehavioural Effects of Developmental Toxicity.” *Lancet neurology* 13.3 (2014): 330–338.

10. Zeliger, Harold I. “Lipophilic Chemical Exposure as a Cause of Cardiovascular Disease.” *Interdisciplinary toxicology* 6.2 (2013): 55–62.

11. Bouchard, Maryse F et al. “Polychlorinated Biphenyl Exposures and Cognition in Older U.S. Adults: NHANES (1999-2002).” *Environmental health perspectives* 122.1 (2014): 73–78.

12. Chevrier, Jonathan et al. “Associations Between Prenatal Exposure to Polychlorinated Biphenyls and Neonatal Thyroid-Stimulating Hormone Levels in a Mexican-American Population, Salinas Valley, California.” *Environmental health perspectives* 115.10 (2007): 1490–1496.

13. Patel, Chirag J, Jayanta Bhattacharya, and Atul J Butte. “An Environment-Wide Association Study (EWAS) on Type 2 Diabetes Mellitus.” *PloS one* 5.5 (2010): e10746–e10746.

14. Backhaus, Thomas, and Michael Faust. “Predictive Environmental Risk Assessment of Chemical Mixtures: A Conceptual Framework.” *Environmental science & technology* 46.5 (2012): 2564–2573.

15. Lee, Yu-Mi, Jr Jacobs, and Duk-Hee Lee. “Persistent Organic Pollutants and Type 2 Diabetes: A Critical Review of Review Articles.” *Frontiers in endocrinology (Lausanne)* 9 (2018): 712–712.

16. Oulhote, Youssef et al. “Joint and Independent Neurotoxic Effects of Early Life Exposures to a Chemical Mixture: A Multi-Pollutant Approach Combining Ensemble Learning and g-Computation.” *Environmental epidemiology* 3.5 (2019): e063–.

17. Park, Sung Kyun et al. “Environmental Risk Score as a New Tool to Examine Multi-Pollutants in Epidemiologic Research: An Example from the NHANES Study Using Serum Lipid Levels.” *PloS one* 9.6 (2014): e98632–.

18. Carrico, Caroline et al. “Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting.” *Journal of agricultural, biological, and environmental statistics* 20.1 (2015): 100–120.

19. Bobb, Jennifer F et al. “Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures.” *Biostatistics (Oxford, England)* 16.3 (2015): 493–508.

20. Devick, K.L. et al. Bayesian Kernel Machine Regression-Causal Mediation Analysis. 2018, arXiv:1811.10453. arXiv.org e-Print archive. https://arxiv.org/abs/1811.10453.

21. Li, Haomin et al. “Health Effects of Air Pollutant Mixtures on Overall Mortality Among the Elderly Population Using Bayesian Kernel Machine Regression (BKMR).” *Chemosphere (Oxford)* 286 (2022): 131566–131566.

22. Bobb, Jennifer F et al. “Statistical Software for Analyzing the Health Effects of Multiple Concurrent Exposures via Bayesian Kernel Machine Regression.” *Environmental health* 17.1 (2018): 67–67.

23. Zhang, Yuqing et al. “Association Between Exposure to a Mixture of Phenols, Pesticides, and Phthalates and Obesity: Comparison of Three Statistical Models.” *Environment international* 123 (2019): 325–336.

24. Keil, Alexander P et al. “A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures.” *Environmental health perspectives* 128.4 (2020): 47004–.

25. Keil, Alexander (2021). qgcomp: Quantile G-Computation. R package version 2.8.0. https://cran.r-project.org/web/packages/qgcomp/vignettes/qgcomp-vignette.html

26. Christensen, YKL et al. “Multiple Classes of Environmental Chemicals Are Associated with Liver Disease: NHANES 2003–2004.” *International journal of hygiene and environmental health* 216.6 (2013): 703–709.

27. Gennings, Chris, Roy Sabo, and Ed Carney. “Identifying Subsets of Complex Mixtures Most Associated With Complex Diseases: Polychlorinated Biphenyls and Endometriosis as a Case Study.” *Epidemiology (Cambridge, Mass.)* 21.4 (2010): S77–S84.

28. Czarnota, Jenna et al. “Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk.” *Cancer informatics* Suppl.2 (2015b): 159–171.

29. Wheeler, David C, and Jenna Czarnota. Modeling Chemical Mixture Effects with Grouped Weighted Quantile Sum Regression, in: International Society for Environmental Epidemiology (ISEE). ISEE Conference Abstracts, Rome, Italy. (2016)

30. Wheeler, David C, and Matthew Carli. groupWQS: group weighted quantile sum regression. (2020b)

31. Wheeler, David C et al. “Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk.” *International journal of environmental research and public health* 18.2 (2021a): 504–.

32. Colicino, Elena et al. “Per- and Poly-Fluoroalkyl Substances and Bone Mineral Density: Results from the Bayesian Weighted Quantile Sum Regression.” *Environmental epidemiology* 4.3 (2020): e092–.

33. Wheeler, David C et al. “Bayesian Deprivation Index Models for Explaining Variation in Elevated Blood Lead Levels Among Children in Maryland.” *Spatial and spatio-temporal epidemiology* 30 (2019): 100286–.

34. Wheeler, David C et al. “Modeling Elevated Blood Lead Level Risk Across the United States.” *The Science of the total environment* 769 (2021c): 145237–145237.

35. Wheeler, David C et al. “Neighborhood Disadvantage and Tobacco Retail Outlet and Vape Shop Outlet Rates.” *International journal of environmental research and public health* 17.8 (2020): 2864–.

36. Wheeler, David C et al. “Bayesian Group Index Regression for Modeling Chemical Mixtures and Cancer Risk.” *International journal of environmental research and public health* 18.7 (2021b): 3486–.

37. Palarea-Albaladejo J, Martín-Fernández JA. Values below detection limit in compositional chemical data. *Analytica chimica acta*. 2013;764:32-43. doi:10.1016/j.aca.2012.12.029

38. Helsel, Dennis. “Less than obvious - statistical treatment of data below the detection limit.” *Environmental Science & Technology* 24 (1990)

39. Lubin, Jay H et al. “Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits.” *Environmental health perspectives* 112.17 (2004): 1691–1696.

40. Singh, Anita, and John Nocerino. “Robust Estimation of Mean and Variance Using Environmental Data Sets with Below Detection Limit Observations.” *Chemometrics and intelligent laboratory systems* 60.1-2 (2002): 69–86.

41. Gillespie BW, Chen Q, Reichert H, et al. Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. Epidemiology. 2010;21(4):S64-S70. doi:10.1097/EDE.0b013e3181ce9f08

42. Cohen AC. Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples. The Annals of Mathematical Statistics. 1950;21:557-569. doi:10.1214/aoms/1177729751

43. Lee KJ, Carlin JB. Multiple Imputation for Missing Data: Fully Conditional Specification Versus Multivariate Normal Imputation. *American journal of epidemiology*. 2010;171(5):624-632. doi:10.1093/aje/kwp425

44. Gelman, Andrew et al. *Bayesian Data Analysis, Third Edition, 3rd Edition*. 3rd edition. CRC Press, 2013.

45. Van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16:219-242. doi:10.1177/0962280206074463

46. Gelman, Andrew. “Parameterization and Bayesian Modeling.” *Journal of the American Statistical Association* 99.466 (2004): 537–545.

47. Li, Fan, Yaming Yu, and Donald B. Rubin. “Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines.” *Duke University Department of Statistical Science* (2012)

48. Chen, S.-H., and Edward H Ip. “Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible.” *Journal of Statistical Computation and Simulation* 85.16 (2015): 3266–3275.

49. Ibrahim, Joseph G, Ming-Hui Chen, and Stuart R Lipsitz. “Bayesian Methods for Generalized Linear Models with Covariates Missing at Random.” *Canadian journal of statistics* 30.1 (2002): 55–78.

50. Fung, Glenn. “A Comprehensive Overview of Basic Clustering Algorithms.” (2001) https://sites.cs.ucsb.edu/~veronika/MAE/clustering_overview_2001.pdf

51. Witten, I. H. (Ian H.), and Eibe. Frank. Data Mining : Practical Machine Learning Tools and Techniques . 2nd ed. Amsterdam ;: Morgan Kaufman, 2005. Print.

52. Nagpal, Arpita, Aman Jatain, and Deepti Gaur. “Review Based on Data Clustering Algorithms.” *2013 IEEE Conference on Information & Communication Technologies*. IEEE, 2013. 298–303.

53. Zhong Shi, and Joydeep Ghosh. “A unified framework for model-based clustering.” *Journal of Machine Learning Research* 4 (2003):1001–1037

54. Li, Jiamin, and Harold W. Lewis. "Fuzzy clustering algorithms—review of the applications." *2016 IEEE International Conference on Smart Cloud (SmartCloud)*. IEEE, 2016.

55. Oliveira JV de, Pedrycz W, Oliveira JV de (José V. *Advances in Fuzzy Clustering and Its Applications*. Wiley; 2007.

56. De Smet, Frank, Janick Mathys, Kathleen Marchal, Gert Thijs, Bart De Moor, and Yves Moreau. “Adaptive Quality-Based Clustering of Gene Expression Profiles.” Bioinformatics 18, no. 5 (2002): 735–46. https://doi.org/10.1093/bioinformatics/18.5.735.

57. Alon, U, N Barkai, D A Notterman, K Gish, S Ybarra, D Mack, and A J Levine. “Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays.” Proceedings of the National Academy of Sciences - PNAS 96, no. 12 (1999): 6745–50. https://doi.org/10.1073/pnas.96.12.6745.

58. Eisen, M.B. (Howard Hughes Medical Institute, P.T Spellman, P.O Brown, and D Botstein. “Cluster Analysis and Display of Genome-Wide Expression Patterns.” Proceedings of the National Academy of Sciences - PNAS 95, no. 25 (1998): 14863–68. https://doi.org/10.1073/pnas.95.25.14863.

59. Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. Proc. Natl. Acad. Sci. USA, Vol. 96(6):2907–2912, March 1999

60. Ghosh, D. and Chinnaiyan, A.M. Mixture modelling of gene expression data from microarray experiments. Bioinformatics, 18:275–286, 2002.

61. Yeung, K.Y., Fraley, C, Murua, A., Raftery, AE., Ruzz WL. Model-based clustering and data transformations for gene expression data. Bioinformatics, 17:977–987, 2001.

62. Palla K, Ghahramani Z, Knowles D. A nonparametric variable clustering model. *Advances in Neural Information Processing Systems*. 2012;25(4):2987-2995.

63. Carreira-Perpinán, Miguel A. "A review of dimension reduction techniques." Department of Computer Science. University of Sheffield. Tech. Rep. CS-96-09 9 (1997): 1-69.

64. Sherlock G. Analysis of large-scale gene expression data. *Current Opinion in Immunology*. 2000;12(2):201-205. doi:10.1016/S0952-7915(99)00074-6

65. Vigneau E, Qannari EM. Clustering of Variables Around Latent Components. *Communications in statistics Simulation and computation*. (2003).

66. Sobczyk P, Bogdan M, Josse J. Bayesian Dimensionality Reduction With PCA Using Penalized Semi-Integrated Likelihood. *Journal of computational and graphical statistics*. 2017;26(4):826-839. doi:10.1080/10618600.2017.1340302

67. Candès E, Li X, Ma Y, Wright J. Robust principal component analysis? *Journal of the ACM*. 2011;58(3):1-37. doi:10.1145/1970392.1970395

68. Gibson EA, Zhang J, Yan J, et al. Principal Component Pursuit for Pattern Identification in Environmental Mixtures. *Environmental health perspectives*. 2022;130(11):117008-. doi:10.1289/EHP10479

69. Ghahramani Z. (2004) Unsupervised Learning. In: Bousquet O., von Luxburg U., Rätsch G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_5

70. Morales, Eduardo F., and Hugo Jair Escalante. "A brief introduction to supervised, unsupervised, and reinforcement learning." Biosignal processing and classification using computational learning and intelligence. Academic Press, 2022. 111-129.

71. Yom-Tov, Elad. "An Introduction to Pattern Classification." *Advanced Lectures on Machine Learning*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2004. 1–20.

72. Nasteski, Vladimir. "An overview of the supervised machine learning methods." Horizons. b 4 (2017): 51-62.

73. Jiang, Tammy, Jaimie L. Gradus, and Anthony J. Rosellini. "Supervised machine learning: a brief primer." Behavior Therapy 51.5 (2020): 675-687.

74. Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

75. Zhu, Xiaojin, and Andrew B Goldberg. Introduction to Semi-Supervised Learning. Vol. 6. San Rafael, Calif. (1537 Fourth Street, San Rafael, CA 94901 USA): Morgan & Claypool Publishers, 2009.

76. Bair, Eric. "Semi-supervised clustering methods." Wiley Interdisciplinary Reviews: Computational Statistics 5.5 (2013): 349-361.

77. Bair, Eric, and Robert Tibshirani. "Semi-supervised methods to predict patient survival from gene expression data." PLoS biology 2.4 (2004): e108.

78. Bilenko M, Basu S, Mooney R. Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM; 2004:11-. doi:10.1145/1015330.1015360

79. Yengo, Loïc, Julien Jacques, and Christophe Biernacki. "Variable clustering in high dimensional linear regression models." Journal de la Société Française de Statistique 155.2 (2014): 38-56.

80. Helsel D. Much Ado About Next to Nothing: Incorporating Nondetects in Science. The Annals of Occupational Hygiene. 2009;54:257-262. doi:10.1093/annhyg/mep092

81. Persson T, Rootzen H. Simple and highly efficient estimators for a type I censored normal sample. Biometrika. 1977;64:123-128. doi:10.1093/biomet/64.1.123

82. Van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006;76:1049-1064. doi:10.1080/10629360600810434

83. Van Buuren S , Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45:1-67. doi:10.18637/jss.v045.i03

84. Carpenter JR, Kenward MG. *Multiple Imputation and Its Application*. 1st ed. Wiley; 2013.

85. Kaplan D, Yavuz S. An Approach to Addressing Multiple Imputation Model Uncertainty Using Bayesian Model Averaging. *Multivariate behavioral research*. 2020;55(4):553-567. doi:10.1080/00273171.2019.1657790

86. Lockwood, J. R., and Mark J. Schervish. "MCMC strategies for computing Bayesian predictive densities for censored multivariate data." Journal of Computational and Graphical Statistics 14.2 (2005): 395-414.

87. Gelman, Andrew, and Trivellore E. Raghunathan. "Using conditional distributions for missing-data imputation." *Statistical Science* 15 (2001): 268-69.

88. Erler NS, Rizopoulos D, Rosmalen J van, Jaddoe VW v., Franco OH, Lesaffre EMEH. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. Statistics in Medicine. 2016;35:2955-2974. doi:10.1002/sim.6944

89. Boyle J, Ward MH, Cerhan JR, Rothman N, Wheeler DC. Estimating mixture effects and cumulative spatial risk over time simultaneously using a Bayesian index low-rank kriging multiple membership model. Statistics in medicine. 2022;41(29):5679-5697. doi:10.1002/sim.9587

Specific Aim 1: Extend Bayesian group index regression to the imputation of BDLs.

Paper: "Imputation of Below Detection Limit Missing Data in Chemical Mixture Analysis with Bayesian Group Index Regression"

Authors: Matthew Carli, Mary H. Ward, Catherine Metayer, David C. Wheeler

Abstract

There is growing scientific interest in identifying the multitude of chemical exposures related to human diseases through mixture analysis. In this paper, we address the issue of below detection limit (BDL) missing data in mixture analysis using Bayesian group index regression by treating both regression effects and missing BDL observations as parameters in a model estimated through a Markov Chain Monte Carlo algorithm that we refer to as Pseudo-Gibbs imputation. We compare this with other Bayesian imputation methods found in the literature (Multiple Imputation by Chained Equations and Sequential Full Bayes imputation), as well as with a non-Bayesian single imputation method. To evaluate our proposed method, we conduct simulation studies with varying percentages of BDL missingness and strengths of association. We apply our method to the California Childhood Leukemia Study (CCLS) to estimate concentrations of chemicals in house dust in a mixture analysis of potential environmental risk factors for childhood leukemia. Our results indicate that Pseudo-Gibbs imputation has superior power for exposure effects and sensitivity for identifying individual chemicals at high percentages of BDL missing data. In the CCLS, we found a significant positive association between concentrations of PAHs in homes and childhood leukemia, as well as significant positive associations for PCBs and herbicides among children from the highest quartile of household income. In conclusion, Pseudo-Gibbs imputation addresses a commonly encountered problem in environmental epidemiology, providing practitioners the ability to jointly estimate the effects of multiple chemical exposures with high levels of BDL missingness.

Introduction

There are more than 350,000 chemicals and chemical mixtures registered for production and use globally [1]. Chemicals used for commercial purposes have been found in human tissues and in household air and dust samples in varying concentrations [2-4], motivating questions as to their impact on human health. Epidemiologic studies have identified environmental chemical exposure as a risk factor in a number of human diseases, including cancer, type 2 diabetes, cardiovascular disease, thyroid disease, and developmental disorders [5-10]. Increasingly, investigations into

the health impact of chemical exposures highlight the fact that they exist as mixtures of many simultaneous exposures [11-12]. Therefore, epidemiologists have sought to assess the joint impact of chemical mixtures on health outcomes as opposed to estimating chemicals as independent risk factors [13-15].

Several statistical methods have been developed for analyzing chemical mixtures that handle the highly correlated data commonly found in chemical mixtures [16], including weighted quantile sum (WQS) regression [17], quantile g-computation [18], and Bayesian kernel machine regression (BKMR) [19]. WQS regression is a two-step process that estimates a single exposure index from part of the data and then estimates the health effect for the exposure index from the remainder of the data. More recently, group index models were developed to allow for multiple chemical groups, where each of the groups can have different magnitudes and direction of association with the outcome [20-21]. There are both frequentist and Bayesian versions of group index models, with Bayesian models being able to estimate all model parameters simultaneously in one step [21-24].

One of the challenges of mixture analysis not fully accounted for in these methods is the commonly encountered problem of below detection limit (BDL) missing observations. A detection limit (DL) is defined as the lowest chemical concentration that can be distinguished from a concentration of zero with reasonable confidence [25]. These detection limits can vary between chemicals, assay methods, different laboratories, and with laboratory time [26-27]. Concentrations below this limit are not reported, leading to interval-censored distributions. Traditionally, analysts presented with this missing data problem have resorted to ad-hoc substitution methods for imputation, where the BDL is replaced by 0, the DL, or some function of the DL (DL/2 being a common example). Such simple substitution has subsequently been criticized for leading to biased parameter estimates and variances [28-30] and for introducing artificial patterns into the original data [31] and therefore is not recommended practice. Various alternative imputation methods that have been developed, such as maximum likelihood estimate (MLE), restricted MLE [32-33], reverse Kaplan–Meier [34], and empirical “robust fill-in” methods [28]. A criticism of these “fill-in” or single-imputation (SI) methods is that imputations are treated as truly observed data without accounting for their variance; however, there is also some evidence that suggests such methods are suitable at lower percentages of BDL missingness [29]. To address this criticism, multiple-imputation (MI) methods, which account for the variance of imputations, have also been developed [29].

Moving to the Bayesian framework, the most straightforward method of imputing missing covariate data is by drawing imputations jointly from a multivariate distribution [35], often a multivariate normal or *t* distribution. A joint distribution can be hard to define, however, when covariates containing missing data are diverse (a combination of continuous and binary variables, for example) or when non-normal models are required. The imputation of BDLs is an instance of the latter, as these bounded variables are best modelled by truncated distributions. One method developed to deal with these difficult covariate groupings is Fully Conditional Specification (FCS), which imputes missing observations one covariate at a time by a univariate conditional distribution conditioned on all other variables in the model. Each variable in the model is cycled through in this fashion until convergence to an assumed but unspecified joint posterior distribution is reached [36].

A common criticism of FCS is the potential for the various univariate conditional distributions to be incompatible, that is, to fail to converge to any joint distribution [37-38]. Incompatibility can result in unsound imputations and biased estimates [39]. Despite these theoretical concerns, FCS has performed well in simulations and has shown to be robust to incompatibility in some scenarios [40]. An alternative imputation method that addresses the issue of potential incompatibility is what we will refer to as Sequential Full Bayes (SFB) imputation [41]. Similar to FCS, univariate conditional distributions for each covariate containing missing observations are used but, in this instance, in order to factorize the joint distribution as a product of all the conditional distributions [42]. In this manner, the joint distribution of the imputation model is specified, avoiding any issues of incompatibility.

The non-Bayesian imputation methods described above have all been applied in the context of mixture analysis. While not recommended, naïve substitutions are still performed [43], likely due to the convenience of these methods. SI methods, which are more theoretically justified than substitution methods but are also relatively easy to implement, are also commonly employed [44-45]. MI procedures are increasingly used in chemical mixture analysis. Single imputation was performed for 10 datasets in a study of non-Hodgkin lymphoma that utilized WQS regression; however, the resulting estimates were not pooled [46]. A Bayesian MI method was later developed specifically for the imputation of BDLs encountered when performing WQS regression [47]. MI procedures have also been developed for BKMR [48] and quantile g-computation [49]. Bayesian imputation methods, by contrast, are not as commonly employed in mixture analysis. One example is found in a 2010 paper by Herring, where BDLs were imputed by a joint distribution specified as a product of marginal and conditional truncated normal distributions in the larger context of

regression analyses of chemical mixtures using a nonparametric Bayesian shrinkage prior [50]. Such simultaneous estimation of missing BDL observations along with the main parameters of interest (index effects and their component weights in the case of Bayesian group index regression) is an attractive solution to the BDL problem.

In this paper, the aim was to extend Bayesian group index regression to handle BDL missing data. To accomplish this aim, we implemented four imputation methods in combination with the Bayesian group index model. The first two are statistical methods that utilize FCS: the well-known Multiple Imputation by Chained Equations (MICE) [51], and what we will refer to as pseudo-Gibbs imputation. As its name implies, MICE involves multiple imputation, where many completely observed datasets are generated by FCS, estimates are calculated for each, and they are then finally pooled into a final result. Pseudo-Gibbs imputation, on the other hand, combines the imputation model (FCS) with the health effects model (Bayesian group index regression) in one Gibbs sampler algorithm from which parameter estimates of interest are derived. A third method utilizes SFB imputation. As with pseudo-Gibbs imputation, this imputation model is combined with the Bayesian group index health effects model in the same Gibbs sampler. Finally, in addition to these Bayesian methods, we consider a type of "fill-in" method where missing BDL observations are singly imputed from a truncated log-normal distribution, which we refer to as Prior imputation.

To evaluate the four imputation techniques mentioned above (MICE, pseudo-Gibbs, Prior, and SFB) in combination with Bayesian group index regression, we conducted a simulation study with varying percentages of BDL observations and compared the model performance. We then applied the best performing method to an investigation of the link between the household exposures and childhood leukemia in the California Childhood Leukemia Study (CCLS). The CCLS data are well-suited for such an analysis, as some of the chemical concentrations gathered in this study exhibit high degrees of BDL missingness. The results from this paper will provide practitioners with a method of analysis that can simultaneously impute BDL observations in a reasonable fashion while estimating the association of chemical mixtures to health outcomes.

Methods

Bayesian Grouped Index Regression

The Bayesian grouped index model in general form for a binary health outcome $y_i \sim Bernoulli(p_i)$ is specified through the log-odds of disease of the *i*th subject as

$$logit(p_i) = \beta_0 + \sum_{k=1}^{K} \beta_k \left(\sum_{j=1}^{C_k} w_{jk}\, q_{ijk} \right) + z_i^T \varphi. \quad (1)$$

On the left of the equation is the logit of the disease probability p_i, and on the right are the effects for the intercept β_0, chemical indices β_k, which estimate the health effects for exposure to the *k*th group of exposures, and a vector of covariates z_i^T with corresponding effects in vector φ. The number of exposures in each of the *K* indices can vary and is denoted by C_k. For each index, w_{jk} is the weight for the *j*th exposure in the *k*th index and denotes the relative importance of that exposure within the index. The value of each w_{jk} is constrained to be between 0 and 1, and when summed across an individual index must equal 1. For each index, q_{ijk} is the quantile score for the *j*th exposure in the *k*th index for the *i*th subject. Quantiles are used instead of raw chemical concentration data in order to limit the influence of outliers and to standardize the varying concentration scaling of different exposures. The definition of quantiles adopted (e.g. quartiles, deciles) is at the discretion of the user.

Finally, the model is completely specified by the assignment of prior distributions to the model parameters. For any given index, the weights $w_{1k}, \ldots, w_{C_k k}$ are assigned a Dirichlet prior with parameters $\alpha_{jk} = (\alpha_{1k}, \ldots, \alpha_{C_k k})$. This choice of prior ensures that the weights $w_{jk} \in (0,1)$ and $\sum_{j=1}^{C_k} w_{jk} = 1$. Each index effect is given a vague normal prior $\beta_k \sim Normal(0, \tau_k)$ with precision $\tau_k = 1/\sigma_k^2$ and $\sigma_k \sim Uniform(0,100)$. Any covariate effects also receive vague normal priors.

Inference on health effects and relative importance of chemical exposures is done through the joint posterior distribution. Markov chain Monte Carlo (MCMC) is used for model parameter estimation and convergence to the posterior is established using the Gelman-Rubin diagnostic statistic using two chains. Researchers who wish to use the Bayesian grouped index regression model as detailed in this paper may do so using the R package BayesGWQS [22], which implements Bayesian grouped index models using Just Another Gibbs Sampler (JAGS) [52].

Imputation Methods

As discussed above, missing data imputation is any method by which incomplete data are made complete by substitution with artificial or imputed data. The Bayesian methods implemented were chosen because they each take into account the additional variability of imputed observations. MICE does this through pooling multiple

imputations, while SFB and Pseudo-Gibbs imputation do so by drawing estimates from converged posterior distributions. The final imputation method, Prior imputation, is a single imputation method that was chosen to highlight circumstances where simpler imputation methods perform just as well as more complex ones, and circumstances where they are contraindicated.

Multiple Imputation by Chained Equations (MICE)

MICE imputes missing data through a series of what are referred to as "chained equations". Given a partially observed dataset, it is assumed the outcome and predictors have a multivariate distribution that is completely specified by some unknown vector of parameters. MICE seeks to obtain a posterior distribution for these unknown parameters without explicitly defining the joint distribution of the data. Imputation models are specified in a univariate fashion for each variable in the dataset, where missing values in any given variable are imputed by a conditional distribution conditioned upon all other variables. These are then linked by means of a Gibbs sampler, which iterates through imputations variable by variable until convergence is attained.

In our application to BDL imputation, our data is composed of a binary outcome *y* and all chemical exposures of interest x_j, where $j = 1, \dots, C$. We assume a multivariate distribution of these variables is completely specified by θ, a $p = C + 1$ length vector of unknown parameters. We obtain the posterior distribution of θ by iteratively sampling from the following conditional distributions:

$$P(y|x_1, \dots, x_C, \theta_1)$$

$$P(x_1|y, x_2, \dots, x_C, \theta_2)$$

$$\vdots$$

$$P(x_C|y, x_1, \dots, x_{C-1}, \theta_p). \quad (2)$$

The chained equations compose the following Gibbs sampler to impute BDLs which at the *t*th iteration draws

$${\theta_1}^{*(t)} \sim P(\theta_1|y^{obs}, x_1^{(t-1)}, \dots, x_C^{(t-1)})$$

$$y^{*(t)} \sim P\left(y\middle|y^{obs}, x_1^{(t-1)}, \dots, x_C^{(t-1)}, {\theta_1}^{*(t)}\right)$$

$$\vdots$$

$$\theta_p^{*(t)} \sim P\left(\theta_p \middle| x_C^{obs}, y^{(t)}, x_1^{(t)}, \ldots, x_{C-1}^{(t)}\right)$$

$$x_C^{*(t)} \sim P\left(x_C \middle| x_C^{obs}, y^{(t)}, x_1^{(t)}, \ldots, x_{C-1}^{(t)}, \theta_p^{*(t)}\right) \quad (3)$$

where $x_j^{(t)} = \left(x_j^{obs}, x_j^{*(t)}\right)$ [51]. One challenge specific to applying this method to the imputation of BDLs is that imputations from these conditional distributions could result in imputed values above the LOD of any particular chemical, contradicting knowledge we already have about that particular observation's value. For these cases, erroneous imputations are "post-processed", taking imputations above the LOD and re-imputing them by drawing from a uniform distribution $x_j^* \sim Uniform(0, LOD_j)$.

Prior Imputation

The Prior imputation method utilizes the so called "data block" in JAGS, where variables can be assigned distributions from which single imputations are drawn. These imputed values are subsequently treated as observed data in the MCMC estimation. This is a type of single imputation or "fill-in" method, which avoids the negative characteristics of ad hoc imputation methods but, because imputation happens only once, does not reflect the variability in the imputation process. There is some evidence, however, that this underestimation of variance is not reflected in parameter estimates when BDL percentage is below 30% [29]. Specific to our application of this method, BDLs were imputed to follow a truncated log-normal prior $BDL_{ij} \sim Lognorm(\mu_j, \tau_j)$ restricted to values within the range of [0, LOD_j]. Uniform and gamma distributions were assigned for the mean and precision hyperpriors, with mean $\mu_j \sim Uniform(0, LOD_j)$ and precision $\tau_j \sim Gamma(0.01, 0.01)$.

Pseudo-Gibbs Imputation

The Pseudo-Gibbs method imputes missing BDL observations by including them as model parameters in the MCMC along with the health-effects model parameters. This pseudo-Gibbs sampling process is similar to that of MICE, where variables are imputed one at a time and the variable being imputed at a particular moment is conditioned on all other variables in the model, current to their most recently updated value. However, the Pseudo-Gibbs method is a combination of imputation and health effects models and therefore the estimated parameters of the health effects model inform the missing data imputations and vice versa. While each BDL observation is estimated as an individual

parameter, BDLs from the same chemical share the same chemical-specific prior and hyperprior distributions. These distributions are the same as those detailed for the Prior imputation method, however, the values drawn from them are not single imputations, but estimations sampled repeatedly though MCMC. A distribution is estimated, giving full posterior inference. Convergence of the MCMC algorithm is evaluated using the Gelman-Rubin diagnostic statistic.

Sequential Full Bayes Imputation (SFB)

Similar to the FCS imputation model used in MICE and Pseudo-Gibbs imputation, the SFB imputation method relies on a sequence of multiplied univariate conditional distributions to express a joint distribution. Again we take chemical exposures of interest x_j, where $j = 1, \dots, C$. Their joint distribution can be written as follows,

$$P(x_1, \dots, x_C|\theta) = P(x_C|x_1, \dots, x_{C-1}, \theta_C)$$

$$\times P(x_{C-1}|x_1, \dots, x_{C-2}, \theta_{C-1}) \times \dots \times P(x_2|x_1, \theta_2) \times P(x_1|\theta_1) \qquad (4)$$

where θ_j is a distinct vector of parameters indexing the *j*th conditional distribution, with the set of $\theta_1, \dots, \theta_C$ vectors parameterizing the joint distribution [42]. In our application to BDL imputation, these conditional distributions follow a truncated log-normal prior restricted to values within the range of 0 and that chemical's LOD. Like the Pseudo-Gibbs method, the above imputation model is combined with the Bayesian group index regression model to give full posterior inference on all model parameters including the index effects and weights.

Simulation Study Design

To evaluate the performance of the four imputation methods, we generated chemical concentration data consisting of three groups (with five chemicals in the first group, four in the second, and five in the third) with a binary outcome. Each group contained a single important chemical which was set by assigning a true chemical weight of 1 to the important chemicals and 0 to nonimportant chemicals, thereby making the total weight for each group sum to 1. The chemical concentrations were given an across group correlation of 0.3 and a within group correlation of 0.7. The correlation structure was specified through a matrix and then converted into a covariance matrix. A mean vector and standard deviation vector were selected to generate the covariance matrix and hence allow construction of the data that was distributed as multivariate normal.

These predictor groupings and outcome were then used in two different signal-strength scenarios. These scenarios differed in the magnitude their index associations, measured in odds ratios (OR) . In Scenario 1, the first group had no association with the outcome (OR=1.0), while the second and third were associated with OR=0.80 and OR=1.25, respectively. Scenario 2 was generated in a similar fashion, except the second and third groups were associated with the outcome with OR=0.67 and OR=1.50, respectively. The sample size generated for both Scenarios 1 and 2 was 500 observations. BDLs were introduced to the data by eliminating the lowest observation values up to a certain DL, depending on the percentage of BDLs desired. For each scenario, BDLs were introduced at the 10, 30, 50, and 70 percentage levels.

After defining the true exposure effects, we created binary outcomes for case or control status to replicate a case-control study by having a relatively balanced number of cases and controls (50% ± 10% cases) in each iteration of data generation. The binary outcome *y* was distributed as $y \sim Binomial(n, p)$ where $p = \frac{1}{1+e^{\eta}}$ and $\eta = \beta_0^* + \sum_{k=1}^{3} \beta_k^* [\sum_{j=1}^{C_k} w_{jk}^* q_{ijk}]$, and the star notation indicates true parameter values. As no covariates were used in generation of the data, the term $z^T \phi = 0$. The number of quantiles used in all simulations was set at four when computing the weighted index for each group (i.e. $q_{ij} = 0,1,2,3$). Each simulation was done with 100 data sets.

To assess the relative performance of the three imputation methods, we calculated the mean squared error (MSE), bias, and power on each of the group exposure effects, as well as the sensitivity and specificity of identifying chemicals as important or not. We assessed model fit by comparing the deviance information criterion (DIC) of each method, and also compared the computation times. When calculating power, we examined the proportion of 95% credible intervals (CIs) of the odds ratios of chemical group associations that did not contain 1.00. We measured sensitivity by determining the proportion of important chemicals that were identified by the models as being important. This was done by determining if the estimated weight of the important chemicals produced by the models was greater than or equal to the threshold $\frac{1}{C_k}$. Likewise, we defined specificity as the proportion of the unimportant chemicals that were correctly deemed unimportant by the models. This was determined by checking if the estimated weights of the unimportant chemicals were less than the same threshold of $\frac{1}{C_k}$. DIC was defined as $DIC = \overline{D} + p_D$, where $\overline{D}$ is the posterior mean deviance [53] and p_D is the effective number of parameters [54], a measure of model complexity.

Data Analysis

We next applied our chosen imputation method along with Bayesian grouped index regression to an investigation of childhood leukemia in the California Childhood Leukemia Study (CCLS). The CCLS is a population-based case-control study carried out in 35 counties in California, 17 counties in the San Francisco Bay area and 18 in the Central Valley [55-56]. Between 1995 and 2012, cases ≤14 years old were ascertained within 72 hours of diagnosis from nine major pediatric clinical centers in the study area. Using California birth certificate information, controls were matched to cases on the basis of date of birth, sex, Hispanic ethnicity, and maternal race.

The parents of both case and control participants were initially interviewed to gather information about their child's exposure to suspected leukemia risk factors. Families who had not moved since the child's diagnosis date (reference date for controls) were interviewed a second time (Tier 2), during which carpet dust samples were collected. The second interview and dust sampling was limited to cases and controls <8 years old at diagnosis to ensure the samples reflected early-life chemical exposure of the child. Case-control matching was not maintained due to residential eligibility criteria and voluntary participation. There were 731 eligible participants (324 cases and 407 controls). Of these, 296 cases (91%) and 333 controls (82%) agreed to participate. Due to insufficient dust or interferences in the chemical analyses, some chemical concentrations were lost, leading to a final 277 cases and 306 controls (n=583) [57].

Dust samples were collected using either a high-volume small surface sampler (HVS3) or a household vacuum cleaner. As previously described in Colt et al. (2008), concentrations of 64 organic chemicals (ng/g dust) were measured using gas chromatography/mass spectrometry (GC/MS) in multiple ion monitoring mode after extraction with three different extraction methods. Nine metals were measured using microwave-assisted acid digestion combined with inductively coupled plasma/mass spectrometry (ICP/MS).

As discussed in Wheeler et al. (2021b), strong correlations ($r > 0.6$) between many chemicals in the CCLS data do not allow for the use of traditional regression methods. Bayesian group index regression, on the other hand, is well-suited for mixture analyses of such data. Our analysis investigated the association of 67 chemicals (Table S1 in the supplemental material) with risk of childhood leukemia. Out of the entire CCLS dataset, only chemical exposure

variables with at least 20% non-missing observations were included, as past experience has shown that higher levels of missingness contribute negligible information on potential relations with an outcome.

We organized exposures into seven chemical class indices: PCBs, PAHs, insecticides, herbicides, metals, the tobacco exposure markers of nicotine and cotinine, and PBDEs. The logic of these groupings was that the chemicals share a structural similarity (e.g., PCBs, PAHs, metals) or usage (e.g., herbicides, insecticides). In addition to these chemical exposure indices, we included child's age, sex, ethnicity, annual household income, mother's education level, mother's age at birth of child, and whether the child lived at the sampling residence since birth as controlling covariates in the model.

We first fit the 7-group exposure model and then evaluated high family income as a potential effect modifier because it was a consistently significant covariate in previous analyses [23-24]. To investigate potential effect moderation, we extended the 7-group model to include seven interaction terms between each index and the highest income level. We then conducted a stratified analysis, dichotomized into the highest income bracket ($75,000+) as one level and the lower five brackets ($0 - $74,999) as the second.

We chose the method of BDL imputation suggested by the results of the simulation study described above. There were additional, non-BDL missing data in the PBDE chemicals as they were measured a few years later than other chemicals on a subset of cases (n=181) and controls (n=214) due to insufficient amounts of dust; in total, PBDEs were not measured on 32.2% of Tier 2 participants [58]. These missing observations were imputed in a similar fashion as BDLs, but their log-normal distributions are not truncated. Continuous chemical concentrations (ng/g) were categorized into quartiles for regression. Convergence of all parameters of interest in models were checked via a Gelman-Rubin diagnostic statistic upper CI less than 1.10. We summarized the results using ORs for each chemical index along with 95% credible intervals and forest plots. Within each index significantly associated with the outcome, we assess the important chemical exposures using the estimated weights.

Results

Simulation Study

The estimated odds ratios and power for the Prior imputation, SFB, Pseudo-Gibbs, and MICE imputation methods for all scenarios are in Table 1. All imputation methods in each BDL scenario performed similarly for null effect parameters, with the exception of SFB and MICE imputation at 70% BDL, where Type I error rates were noticeably lower. For Scenario 1 (lower signal scenario), power was similar for all imputation methods, with Pseudo-Gibbs imputation resulting in slightly higher power in the 70% BDL case. This pattern was repeated in Scenario 2 (higher signal scenario), where the difference in power at 70% BDL in favor of the Pseudo-Gibbs method was much more apparent. Power was predictably higher in the more strongly associated Scenario 2, with values more than doubling for all imputation methods. In both scenarios power tended to decrease as BDL percentage increased, with the drop in power most apparent after the 30% BDL case. While the Pseudo-Gibbs method was best able to preserve power from decreasing as BDL percentage increased, absolute power in Scenario 1 at 70% BDL reached extremely low levels for all imputation methods.

Table 1: Estimated odds ratio (OR) and power values for Bayesian group index regression using four different imputation methods.

Parameter	**Prior Imputation**		**Sequential Full Bayes**		**Pseudo-Gibbs**		**MICE**	
10% BDL	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power
$\exp(\beta_1)$= 1.00	1.000	0.070	0.999	0.060	0.999	0.050	1.000	0.060
$\exp(\beta_2)$= 0.80	0.818	0.430	0.818	0.430	0.818	0.430	0.818	0.430
$\exp(\beta_3)$= 1.25	1.251	0.430	1.251	0.420	1.251	0.440	1.251	0.430
$\exp(\beta_1)$= 1.00	0.994	0.050	0.9934	0.040	0.993	0.040	0.994	0.050
$\exp(\beta_2)$= 0.67	0.658	0.900	0.658	0.900	0.658	0.900	0.658	0.900
$\exp(\beta_3)$= 1.50	1.553	0.910	1.553	0.920	1.553	0.920	1.554	0.920
30% BDL	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power
$\exp(\beta_1)$= 1.00	1.004	0.080	1.001	0.080	1.001	0.080	1.000	0.060
$\exp(\beta_2)$= 0.80	0.816	0.430	0.814	0.430	0.814	0.430	0.819	0.410
$\exp(\beta_3)$= 1.25	1.246	0.400	1.254	0.430	1.253	0.430	1.247	0.420
$\exp(\beta_1)$= 1.00	0.996	0.050	0.999	0.070	0.996	0.050	0.994	0.050
$\exp(\beta_2)$= 0.67	0.662	0.920	0.655	0.920	0.655	0.930	0.664	0.930
$\exp(\beta_3)$= 1.50	1.539	0.900	1.552	0.890	1.556	0.900	1.535	0.890
50% BDL	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power
$\exp(\beta_1)$= 1.00	1.002	0.050	1.004	0.070	1.003	0.070	1.002	0.070
$\exp(\beta_2)$= 0.80	0.824	0.370	0.828	0.340	0.812	0.400	0.823	0.380
$\exp(\beta_3)$= 1.25	1.241	0.390	1.236	0.350	1.253	0.370	1.234	0.340
$\exp(\beta_1)$= 1.00	0.995	0.040	0.995	0.030	0.994	0.050	0.991	0.060
$\exp(\beta_2)$= 0.67	0.667	0.880	0.664	0.880	0.651	0.890	0.681	0.880
$\exp(\beta_3)$= 1.50	1.521	0.870	1.551	0.880	1.557	0.870	1.498	0.860
70% BDL	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power	Estimated OR	Power
$\exp(\beta_1)$= 1.00	0.997	0.060	0.992	0.010	0.997	0.060	0.994	0.030
$\exp(\beta_2)$= 0.80	0.857	0.200	0.843	0.200	0.810	0.290	0.857	0.180
$\exp(\beta_3)$= 1.25	1.209	0.260	1.250	0.280	1.256	0.260	1.184	0.220
$\exp(\beta_1)$= 1.00	0.993	0.020	0.979	0.040	0.987	0.050	0.984	0.010
$\exp(\beta_2)$= 0.67	0.724	0.680	0.693	0.660	0.655	0.810	0.753	0.600
$\exp(\beta_3)$= 1.50	1.425	0.690	1.530	0.740	1.542	0.750	1.356	0.590

MSE and bias of the four imputation methods are compared in Table 2. Both MSE and bias remained relatively consistent as the percentage of BDLs grew. Other than a few exceptional instances, the MICE imputation method estimations had the lowest MSE. The differences in MSE were minimal for the 10% BDL case, and was one of the instances where another method (Pseudo-Gibbs) outperformed MICE. While differences in MSE were never extreme, they tended to be larger at higher levels of missingness. The Prior imputation method often had the next best MSE after MICE. The results for bias were less consistent. In Scenario 1, Pseudo-Gibbs imputation tended to have the lowest bias, and if not was a close second. In Scenario 2, however, Pseudo-Gibbs imputation was only the least biased for 10% BDL, and was at times the most biased imputation method. MICE and Prior imputation were least biased for 30% and 50% BDL, but had the highest bias of all simulations done at 70% BDL. SFB and Pseudo-Gibbs had the lowest and second-lowest bias for 70% BDL, respectively.

Table 2: MSE and bias of index effects from Bayesian group index regression using different imputation methods.

Parameter	**Prior Imputation**		**Sequential Full Bayes**		**Pseudo-Gibbs**		**MICE**	
10% BDL	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\exp(\beta_1)$= 1.00	0.012	-0.006	0.012	-0.007	0.011	-0.007	0.012	-0.006
$\exp(\beta_2)$= 0.80	0.017	0.014	0.017	0.014	0.017	0.014	0.017	0.014
$\exp(\beta_3)$= 1.25	0.014	-0.007	0.014	-0.007	0.014	-0.006	0.014	-0.006
$\exp(\beta_1)$= 1.00	0.012	-0.012	0.012	-0.012	0.012	-0.013	0.012	-0.012
$\exp(\beta_2)$= 0.67	0.015	-0.026	0.015	-0.025	0.015	-0.025	0.015	-0.026
$\exp(\beta_3)$= 1.50	0.017	0.027	0.017	0.027	0.016	0.027	0.017	0.028
30% BDL	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\exp(\beta_1)$= 1.00	0.012	-0.002	0.013	-0.005	0.013	-0.005	0.012	-0.006
$\exp(\beta_2)$= 0.80	0.017	0.012	0.018	0.009	0.017	0.008	0.016	0.015
$\exp(\beta_3)$= 1.25	0.014	-0.010	0.015	-0.004	0.014	-0.005	0.014	-0.009
$\exp(\beta_1)$= 1.00	0.012	-0.010	0.013	-0.008	0.012	-0.010	0.012	-0.012
$\exp(\beta_2)$= 0.67	0.014	-0.019	0.015	-0.030	0.015	-0.031	0.013	-0.015
$\exp(\beta_3)$= 1.50	0.017	0.018	0.018	0.025	0.018	0.028	0.016	0.015
50% BDL	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\exp(\beta_1)$= 1.00	0.014	-0.005	0.015	-0.003	0.015	-0.004	0.013	-0.004
$\exp(\beta_2)$= 0.80	0.018	0.021	0.021	0.024	0.020	0.006	0.017	0.021
$\exp(\beta_3)$= 1.25	0.014	-0.014	0.015	-0.019	0.015	-0.005	0.013	-0.020
$\exp(\beta_1)$= 1.00	0.013	-0.012	0.013	-0.012	0.014	-0.013	0.012	-0.015
$\exp(\beta_2)$= 0.67	0.015	-0.011	0.015	-0.017	0.017	-0.036	0.013	0.009
$\exp(\beta_3)$= 1.50	0.018	0.005	0.021	0.024	0.020	0.028	0.017	-0.010
70% BDL	MSE	Bias	MSE	Bias	MSE	Bias	MSE	Bias
$\exp(\beta_1)$= 1.00	0.020	-0.013	0.019	-0.018	0.022	-0.014	0.012	-0.012
$\exp(\beta_2)$= 0.80	0.024	0.058	0.024	0.041	0.026	0.0000	0.017	0.062
$\exp(\beta_3)$= 1.25	0.018	-0.042	0.021	-0.011	0.019	-0.005	0.016	-0.060
$\exp(\beta_1)$= 1.00	0.016	-0.015	0.025	-0.032	0.022	-0.024	0.014	-0.023
$\exp(\beta_2)$= 0.67	0.024	0.069	0.023	0.023	0.024	-0.034	0.023	0.112
$\exp(\beta_3)$= 1.50	0.025	-0.062	0.031	0.005	0.028	0.014	0.028	-0.109

The sensitivity and specificity of important chemical identification calculated for the four imputation methods is presented in Table 3. Sensitivity for both signal strength scenarios was very similar for all imputation methods until the 70% BDL case, where Pseudo-Gibbs imputation had consistently larger sensitivity values. Specificity values were very similar across all imputation methods for each combination of signal strength and level of missingness. SFB and Pseudo-Gibbs generally performed best by this statistic. Differences in specificity values increased as the percentage of BDLs increased, most notably in Scenario 2. The odds ratios further from OR=1.00 predictably resulted in higher values for both sensitivity and specificity. In both scenarios sensitivity and specificity tended to decrease as BDL percentage rose, with the largest decreases occurring between 50% and 70% BDL.

Table 3: Sensitivity and specificity for Bayesian group index regression using different imputation methods.

Parameter	**Prior Imputation**		**Sequential Full Bayes**		**Pseudo-Gibbs**		**MICE**	
10% BDL	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$\exp(\beta_1)$= 1.00	0.340	0.573	0.330	0.580	0.310	0.575	0.310	0.568
$\exp(\beta_2)$= 0.80	0.910	0.797	0.890	0.803	0.900	0.800	0.900	0.800
$\exp(\beta_3)$= 1.25	0.820	0.738	0.850	0.753	0.820	0.733	0.840	0.748
$\exp(\beta_1)$= 1.00	0.390	0.615	0.380	0.600	0.420	0.623	0.410	0.615
$\exp(\beta_2)$= 0.67	0.980	0.943	0.980	0.940	0.980	0.940	0.980	0.940
$\exp(\beta_3)$= 1.50	0.990	0.918	1.000	0.918	1.000	0.918	0.990	0.920
30% BDL	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$\exp(\beta_1)$= 1.00	0.280	0.573	0.320	0.560	0.320	0.550	0.290	0.568
$\exp(\beta_2)$= 0.80	0.870	0.797	0.890	0.800	0.900	0.800	0.890	0.793
$\exp(\beta_3)$= 1.25	0.820	0.705	0.860	0.723	0.840	0.713	0.850	0.703
$\exp(\beta_1)$= 1.00	0.380	0.580	0.360	0.593	0.360	0.600	0.400	0.613
$\exp(\beta_2)$= 0.67	0.980	0.920	0.970	0.920	0.980	0.927	0.980	0.920
$\exp(\beta_3)$= 1.50	0.990	0.893	0.990	0.903	0.990	0.900	0.990	0.903
50% BDL	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$\exp(\beta_1)$= 1.00	0.380	0.593	0.330	0.593	0.350	0.585	0.370	0.603
$\exp(\beta_2)$= 0.80	0.850	0.760	0.810	0.787	0.830	0.800	0.810	0.783
$\exp(\beta_3)$= 1.25	0.830	0.705	0.860	0.700	0.830	0.715	0.810	0.703
$\exp(\beta_1)$= 1.00	0.380	0.578	0.410	0.605	0.400	0.598	0.410	0.603
$\exp(\beta_2)$= 0.67	0.960	0.890	0.980	0.903	0.980	0.903	0.980	0.890
$\exp(\beta_3)$= 1.50	0.980	0.870	0.980	0.875	0.990	0.885	0.990	0.873
70% BDL	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity	Specificity
$\exp(\beta_1)$= 1.00	0.320	0.605	0.410	0.620	0.370	0.595	0.370	0.573
$\exp(\beta_2)$= 0.80	0.640	0.670	0.720	0.690	0.750	0.693	0.710	0.673
$\exp(\beta_3)$= 1.25	0.630	0.675	0.680	0.675	0.740	0.670	0.620	0.660
$\exp(\beta_1)$= 1.00	0.390	0.625	0.410	0.620	0.380	0.585	0.400	0.580
$\exp(\beta_2)$= 0.67	0.880	0.767	0.870	0.817	0.950	0.790	0.920	0.737
$\exp(\beta_3)$= 1.50	0.890	0.775	0.880	0.778	0.890	0.800	0.870	0.743

The model fit tended to decrease (lower DIC is better) for all imputation methods as the percentage of BDLs rose (Table 4). There were very slight differences in DIC at low levels of missingness. For Scenario 1, Prior imputation resulted in the best fit, whereas for Scenario 2 Pseudo-Gibbs and SFB performed best. For both signal levels SFB and

Pseudo-Gibbs had the lowest DIC as BDL percentage increased, and of the two SFB was slightly better in Scenario 1, while Pseudo-Gibbs was better in Scenario 2. These two methods also saw increases in pD as BDL percentages rose, indicating greater model complexity. Of the four methods, MICE saw the largest increase in DIC as BDL percentage rose. Considering runtime, the Prior imputation method was always the fastest running analysis at around 7 minutes (Table 4). MICE was the next best, with similar but slightly slower runtime (accomplished with parallel computing). The Pseudo-Gibbs and SFB methods were the slowest by far, taking nearly nine hours or more to complete at 10% BDL and nearly two days or more at 70% BDL, averaged over 100 datasets.

Table 4: Model fit statistics and computation time for Bayesian group index regression using different imputation methods.

Scenario 1	**Prior Imputation**	**Sequential Full Bayes**	**Pseudo-Gibbs**	**MICE**
10% BDL				
DIC	585.04	585.51	585.53	585.64
pD	5.04	5.03	5.25	5.21
Runtime (min)	7.32	679.71	538.20	7.78
30% BDL				
DIC	585.49	585.58	585.77	585.52
pD	5.39	5.68	5.46	5.10
Runtime (min)	7.31	1567.51	1333.01	7.93
50% BDL				
DIC	585.58	585.56	585.77	586.32
pD	5.15	5.83	6.21	5.52
Runtime (min)	7.03	2375.42	2108.65	8.31
70% BDL				
DIC	587.56	586.25	586.57	588.56
pD	5.05	8.69	9.30	5.59
Runtime (min)	6.33	3557.38	2686.91	9.67
Scenario 2	**Prior Imputation**	**Sequential Full Bayes**	**Pseudo-Gibbs**	**MICE**
10% BDL				
DIC	577.71	577.66	577.33	577.57
pD	5.98	6.05	5.70	5.79
Runtime (min)	7.19	683.38	565.97	7.89
30% BDL				
DIC	578.83	578.36	579.46	578.89
pD	6.07	7.08	7.26	5.86
Runtime (min)	7.22	1573.61	1304.99	7.97
50% BDL				
DIC	581.55	580.27	579.18	582.49
pD	6.53	8.01	8.06	6.35
Runtime (min)	6.90	2407.21	2067.70	8.16
70% BDL				
DIC	589.33	586.20	586.11	591.42
pD	5.53	13.42	15.91	6.40
Runtime (min)	6.29	3487.45	2711.79	8.51

Application of Pseudo-Gibbs imputation to house dust chemicals in the CCLS

The results of our simulation study indicate that for data with relatively high percentages of BDL observations the most suitable imputation method is Pseudo-Gibbs imputation. As the CCLS data have 23.9% of 67 chemical exposure variables with greater than 50% BDLs (n = 16), and 10.4% with 70% or more BDLs (n = 7), we applied this method of imputation when performing the following analysis. We first consider the non-stratified analysis. The odds ratios estimated for index effects and covariates are in Table 5. PAHs were the only index found to have a significant association with childhood leukemia (OR = 1.27, 95% CI: 1.01, 1.60). The PCB index was also positively associated with the outcome, although this effect was marginally significant (OR = 1.19, 95% CI: 0.96, 1.51). The two most heavily weighted chemicals in the PAHs index were benzo(k)fluoranthene and indeno(1,2,3 -c,d)pyrene, with posterior mean weights of 0.164 and 0.149, respectively. Looking at the forest plot of estimated index means and 95% CIs (Figure S1), we can see PBDEs was the most variable index estimate. Among the controlling covariates, the highest income category and residence since birth were significant and protective.

Table 5: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model (n = 583).

Variable	Odds Ratio	2.5% CI	97.5% CI
PCBs	1.19	0.96	1.51
Insecticides	0.64	0.39	1.00
Herbicides	1.17	0.82	1.69
Metals	0.79	0.59	1.06
PAHs	**1.27**	**1.01**	**1.60**
Tobacco	0.82	0.66	1.01
PBDEs	1.21	0.79	1.83
Child's age	1.01	0.92	1.12
Female	0.98	0.70	1.37
Child's Ethnicity:			
Hispanic	1.25	0.81	2.00
Non-Hispanic	1.42	0.91	2.27
Household Income:			
$15,000 - $29,999	1.02	0.47	2.15
$30,000 - $44,999	0.79	0.36	1.61
$45,000 - $59,999	0.78	0.34	1.66
$60,000 - $74,999	0.45	0.18	1.06
$75,000 or more	**0.38**	**0.17**	**0.79**
Income Missing	0.56	0.17	1.61
Mother's education:			
High school	1.25	0.63	2.81
Some college	1.22	0.60	2.84
Bachelor's or higher	1.21	0.57	2.89
Mother's age	1.01	0.98	1.05
Residence Since Birth	**0.66**	**0.44**	**0.96**

Our Bayesian group index regression of interaction effects between the chemical indices and the highest income bracket ($75,000 or more) resulted in a significant interaction between income and the metals index (OR = 0.45, 95% CI: 0.24, 0.82). In the subsequent analysis stratified on household income, three chemical indices were found to have significant associations with childhood leukemia risk in the highest income strata (>=$75,000, 107 cases, 159 controls) (Table 6). PCBs (OR = 1.55, 95% CI: 1.04, 2.36) and herbicides (OR = 2.02, 95% CI: 1.005, 3.99) had significant positive associations with childhood leukemia. The herbicide index had the strongest association, but was the most variable. The metals index (OR = 0.42, 95% CI: 0.25, 0.69) was inversely associated with childhood leukemia. Of the covariates, residence since birth was significantly inversely associated with risk. The forest plot of the index association estimates and their 95% CIs are presented in Figure S2. Of the four PCB chemicals, PCB 138 had the highest mean posterior weight of 0.31, followed by PCB 180 with a weight of 0.28. Among the herbicides, dacthal had the largest weight (0.51). In the metals index, arsenic was the most highly weighted chemical (inverse association), with a mean posterior weight of 0.37. The specific estimates for the lower income stratum and its forest plot are presented in Table S2 and Figure S3. There were no significant findings in the lower income stratum (<$75,000).

Table 6: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in highest income bracket (n = 266).

Variable	Odds Ratio	2.5% CI	97.5% CI
PCBs	**1.55**	**1.04**	**2.36**
Insecticides	0.51	0.19	1.12
Herbicides	**2.02**	**1.00**	**3.99**
Metals	**0.42**	**0.25**	**0.69**
PAHs	1.19	0.83	1.75
Tobacco	0.77	0.52	1.09
PBDEs	1.12	0.63	2.23
Child's age	0.98	0.83	1.15
Female	0.70	0.38	1.22
Child's Ethnicity:			
Hispanic	1.14	0.47	2.83
Non-Hispanic	1.62	0.87	3.18
Mother's education:			
High school	0.49	0.00	1930.56
Some college	0.20	0.00	730.17
Bachelor's or higher	0.36	0.00	1375.01
Mother's age	0.99	0.93	1.05
Residence since birth	**0.40**	**0.21**	**0.76**

Discussion

In this paper, we implemented four methods for the imputation of BDL missing data in the context of Bayesian group index regression and conducted a simulation study to evaluate the performance of these methods at two different association strengths (OR = 1.25 and 1.50) as well as at four different levels of BDL missingness (10%, 30%, 50%, and 70%). We found that the relative performance of the methods was similar across the two association strengths and across the 10% - 50% BDL levels, with some methods slightly outperforming others in certain scenarios judged by some metrics. Notably, the Prior imputation method performed consistently well across metrics in this BDL range. It was at times the best performing method, was rarely the worst, and when not the best performer was usually competitive.

Clear differences in performance were seen, however, in the 70% BDL range. At such high levels of missingness, Pseudo-Gibbs imputation was found to be the preferred method of imputation. A clear advantage of Pseudo-Gibbs imputation was that it consistently had more power to detect significant associations than other methods (with power differences of 10% or more in many instances). This superior performance was also apparent in sensitivity. Results were not so clear for specificity, bias, and DIC, where SFB imputation performed slightly better in some instances. While all imputation methods had approximately the same performance as judged by MSE, Pseudo-Gibbs imputation was often the weakest method by a slight margin. The greatest weakness of the Pseudo-Gibbs method is its runtime. While faster than SFB imputation, it proved to be much slower than either MICE or Prior imputation. Additionally, while Pseudo-Gibbs imputation had the highest power in Scenario 1 at 70% BDL, in absolute terms power was quite low. Detecting lower signal differences at such high levels of BDL missingness would likely require an increase in sample size even when using the Pseudo-Gibbs method.

Based on the findings described above, we recommend Pseudo-Gibbs imputation for data where the percentage of BDLs approaches 70% and the Prior imputation method for lower percentages. While 70% BDL missing data is an extreme level of missingness to simulate, such percentages are at times encountered in chemical exposure investigations (CCLS being an example), and previous statistical research has been done for BDL missingness at such levels [29,59]. It should be noted that while our simulated datasets had uniform levels of missingness across all chemical

exposure variables, this would be highly unlikely to occur in actual practice. While this represents a simplification from real conditions, we believe our results nonetheless offer useful guidelines for determining the most suitable method of BDL imputation. A further limitation of our results is that they are restricted to the particular scenarios simulated. At higher BDL levels the slow runtime of the Pseudo-Gibbs imputation can be justified most clearly by its improved performance in power and in sensitivity. While second to SFB in some metrics, the difference in their performance was negligible. Importantly, although Pseudo-Gibbs was relatively slow, the slowest method was SFB, an increase in runtime which is hard to justify by its performance. At lower percentages, Prior imputation offers a computationally efficient and convenient method that produces estimates competitive with the other methods presented.

Our decision to apply Pseudo-Gibbs imputation in our analysis of the CCLS data reflects the above observations. While BDL missingness is not uniform across all chemical predictors in the CCLS observational data, many exhibit BDL levels of 50% or more, with some of these extending to 70% or more (chemicals with 80% or more were excluded). In our application of Pseudo-Gibbs imputation to the CCLS observational data, we fit a seven-index model and found a positive and significant association between PAHs (OR = 1.27) and leukemia, with benzo(k)fluoranthene (weight = 0.164) and indeno(1,2,3 -c,d)pyrene (weight = 0.149) having the highest mean posterior weights. Previous research of this study population employing single-chemical models have found either significant or borderline significant associations between these two PAHs and childhood leukemia [60]. In stratified analysis, of the highest income category and all others., the chemical indices estimated for the high-income strata tended to be larger and have lower variance. Among children from high-income households, PCBs (OR = 1.55) and herbicides (OR = 2.02) were significantly and positively associated with childhood leukemia, while the metals index (OR = 0.42) was significantly inversely associated with risk.

The association of PCBs with leukemia reflects the findings of earlier work. In a previous study of the CCLS cohort, group index regression methods found a marginally significant association between PCBs and childhood leukemia, with PCB 138 contributing the most to the index effect [24]. Single-chemical logistic regression analyses have also found significant positive associations between leukemia and PCB138, as well as between leukemia and summed total PCB concentrations [56]. Similarly, the significant positive association found for herbicides (and the dominance of dacthal within the index) closely mirrors prior analyses of these data done using Bayesian group index regression

analysis with a different imputation approach [23] and GWQS regression [24]. Besides these mixture analyses, univariable logistic regression analyses have found similar associations between dacthal and childhood acute lymphocytic leukemia (ALL) risk [57]. The significant negative association observed for the metals index, and for arsenic in particular, have less support from previous research. While arsenic is a well-known risk factor in adult bladder cancer [61], there is little to no evidence of any link between arsenic and childhood cancer, including childhood leukemia [62]. While selection bias cannot be ruled out to explain the negative association in the current paper, further investigation is necessary to understand this association.

In summary, through our comparison of BDL imputation methods in the context of Bayesian group index regression, the Pseudo-Gibbs method of imputation performed best under conditions of high BDL missingness, whereas Prior imputation offers a suitable method of imputation at relatively low levels of BDL missingness. These methods and the guidance for their appropriate use allows researchers assessing environmental exposures to more rigorously handle the common problem of BDL missing data. While our application was to chemical exposure missing data, other fields (such as genomics) that frequently encounter such missing observations could also benefit from these methods.

References

1. Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environmental Science & Technology. 2020;54(5):2575-2584. doi:10.1021/acs.est.9b06379

2. Centers for Disease Control and Prevention. Fourth National Report on Human Exposure to Environmental Chemicals.; 2009. Accessed February 21, 2021. http://www.cdc.gov/ExposureReport/pdf/FourthReport.pdf

3. Rudel RA, Dodson RE, Perovich LJ, et al. Semivolatile Endocrine-Disrupting Compounds in Paired Indoor and Outdoor Air in Two Northern California Communities. Environmental Science & Technology. 2010;44:6583-6590. doi:10.1021/es100159c

4. Yilmaz B, Terekeci H, Sandal S, Kelestimur F. Endocrine disrupting chemicals: exposure, effects on human health, mechanism of action, models for testing and strategies for prevention. Reviews in Endocrine and Metabolic Disorders. 2020;21(1):127-147. doi:10.1007/s11154-019-09521-z

5. Zeliger HI. Lipophilic chemical exposure as a cause of cardiovascular disease. Interdisciplinary Toxicology. 2013;6:55-62. doi:10.2478/intox-2013-0010

6. Grandjean P, Landrigan PJ. Neurobehavioural effects of developmental toxicity. The Lancet Neurology. 2014;13:330-338. doi:10.1016/S1474-4422(13)70278-3

7. Terry MB, Michels KB, Brody JG, et al. Environmental exposures during windows of susceptibility for breast cancer: a framework for prevention research. Breast Cancer Research. 2019;21(1):96. doi:10.1186/s13058-019-1168-2

8. Ruiz D, Becerra M, Jagai JS, Ard K, Sargis RM. Disparities in Environmental Exposures to Endocrine-Disrupting Chemicals and Diabetes Risk in Vulnerable Populations. Diabetes Care. 2018;41(1):193-205. doi:10.2337/dc16-2765

9. Han J, Zhou L, Luo M, et al. Nonoccupational Exposure to Pyrethroids and Risk of Coronary Heart Disease in the Chinese Population. Environmental Science & Technology. 2017;51(1):664-670. doi:10.1021/acs.est.6b05639

10. Ghassabian A, Trasande L. Disruption in Thyroid Signaling Pathway: A Mechanism for the Effect of Endocrine-Disrupting Chemicals on Child Neurodevelopment. Frontiers in Endocrinology. 2018;9. doi:10.3389/fendo.2018.00204

11. Backhaus T, Faust M. Predictive Environmental Risk Assessment of Chemical Mixtures: A Conceptual Framework. Environmental Science & Technology. 2012;46:2564-2573. doi:10.1021/es2034125

12. Hernández AF, Tsatsakis AM. Human exposure to chemical mixtures: Challenges for the integration of toxicology with epidemiology data in risk assessment. Food and Chemical Toxicology. 2017;103:188-193. doi:10.1016/j.fct.2017.03.012

13. Oulhote Y, Coull B, Bind MA, et al. Joint and independent neurotoxic effects of early life exposures to a chemical mixture. Environmental Epidemiology. 2019;3:e063. doi:10.1097/EE9.0000000000000063

14. Lee YM, Jacobs Jr. DR, Lee DH. Persistent Organic Pollutants and Type 2 Diabetes: A Critical Review of Review Articles. Frontiers in Endocrinology. 2018;9:712. doi:10.3389/fendo.2018.00712

15. Park SK, Tao Y, Meeker JD, Harlow SD, Mukherjee B. Environmental Risk Score as a New Tool to Examine Multi-Pollutants in Epidemiologic Research: An Example from the NHANES Study Using Serum Lipid Levels. Meliker J, ed. PLoS ONE. 2014;9:e98632. doi:10.1371/journal.pone.0098632

16. Czarnota J, Gennings C, Wheeler DC. Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. Cancer Informatics. 2015;14:159-171. doi:10.4137/CIN.S17295

17. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of Weighted Quantile Sum Regression for Highly Correlated Data in a Risk Analysis Setting. Journal of Agricultural, Biological, and Environmental Statistics. 2015;20:100-120. doi:10.1007/s13253-014-0180-3

18. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. Environmental Health Perspectives. 2020;128(4):047004. doi:10.1289/EHP5838

19. 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. Environmental Health. 2018;17:67-77. doi:10.1186/s12940-018-0413-y

20. Wheeler D, Czarnota J. Modeling Chemical Mixture Effects with Grouped Weighted Quantile Sum Regression. In: International Society for Environmental Epidemiology (ISEE). ISEE Conference Abstracts; 2016.

21. Wheeler D, Carli M. groupWQS: group weighted quantile sum regression. Published online 2020.

22. Wheeler D, Carli M. BayesGWQS: Bayesian Grouped Weighted Quantile Sum Regression. Published online 2020.

23. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Bayesian Group Index Regression for Modeling Chemical Mixtures and Cancer Risk. International Journal of Environmental Research and Public Health. 2021;18:3486. doi:10.3390/ijerph18073486

24. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. International Journal of Environmental Research and Public Health. 2021;18:504. doi:10.3390/ijerph18020504

25. Analytical Methods Committee. Recommendations for the definition, estimation and use of the detection limit. The Analyst. 1987;112:199-204. doi:10.1039/an9871200199

26. Succop PA, Clark S, Chen M, Galke W. Imputation of Data Values That are Less Than a Detection Limit. Journal of Occupational and Environmental Hygiene. 2004;1:436-441. doi:10.1080/15459620490462797

27. He J. Mixture model based multivariate statistical analysis of multiply censored environmental data. Advances in Water Resources. 2013;59:15-24. doi:10.1016/j.advwatres.2013.05.001

28. Helsel DR. Less than obvious - statistical treatment of data below the detection limit. Environmental Science & Technology. 1990;24:1766-1774. doi:10.1021/es00082a001

29. Lubin JH, Colt JS, Camann D, et al. Epidemiologic Evaluation of Measurement Data in the Presence of Detection Limits. Environmental Health Perspectives. 2004;112:1691-1696. doi:10.1289/ehp.7199

30. Singh A, Nocerino J. Robust estimation of mean and variance using environmental data sets with below detection limit observations. Chemometrics and Intelligent Laboratory Systems. 2002;60:69-86. doi:10.1016/S0169-7439(01)00186-1

31. Helsel D. Much Ado About Next to Nothing: Incorporating Nondetects in Science. The Annals of Occupational Hygiene. 2009;54:257-262. doi:10.1093/annhyg/mep092

32. Cohen AC. Estimating the Mean and Variance of Normal Populations from Singly Truncated and Doubly Truncated Samples. The Annals of Mathematical Statistics. 1950;21:557-569. doi:10.1214/aoms/1177729751

33. Persson T, Rootzen H. Simple and highly efficient estimators for a type I censored normal sample. Biometrika. 1977;64:123-128. doi:10.1093/biomet/64.1.123

34. Gillespie BW, Chen Q, Reichert H, et al. Estimating Population Distributions When Some Data Are Below a Limit of Detection by Using a Reverse Kaplan-Meier Estimator. Epidemiology. 2010;21(4):S64-S70. doi:10.1097/EDE.0b013e3181ce9f08

35. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin D. Bayesian Data Analysis. 3rd ed. CRC Press; 2013.

36. van Buuren S, Brand JPL, Groothuis-Oudshoorn CGM, Rubin DB. Fully conditional specification in multivariate imputation. Journal of Statistical Computation and Simulation. 2006;76:1049-1064. doi:10.1080/10629360600810434

37. Li F, Yu Y, Rubin DB. Imputing Missing Data by Fully Conditional Models: Some Cautionary Examples and Guidelines.; 2012.

38. Gelman A. Parameterization and Bayesian Modeling. Journal of the American Statistical Association. 2004;99:537-545. doi:10.1198/016214504000000458

39. Chen SH, Ip EH. Behaviour of the Gibbs sampler when conditional distributions are potentially incompatible. Journal of Statistical Computation and Simulation. 2015;85:3266-3275. doi:10.1080/00949655.2014.968159

40. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Statistical Methods in Medical Research. 2007;16:219-242. doi:10.1177/0962280206074463

41. Erler NS, Rizopoulos D, Rosmalen J van, Jaddoe VW v., Franco OH, Lesaffre EMEH. Dealing with missing covariates in epidemiologic studies: a comparison between multiple imputation and a full Bayesian approach. Statistics in Medicine. 2016;35:2955-2974. doi:10.1002/sim.6944

42. Ibrahim JG, Chen MH, Lipsitz SR. Bayesian methods for generalized linear models with covariates missing at random. Canadian Journal of Statistics. 2002;30:55-78. doi:10.2307/3315865

43. Fruh V, Claus Henn B, Weuve J, et al. Incidence of uterine leiomyoma in relation to urinary concentrations of phthalate and phthalate alternative biomarkers: A prospective ultrasound study. Environment International. 2021;147:106218. doi:10.1016/j.envint.2020.106218

44. Hu JMY, Arbuckle TE, Janssen P, et al. Prenatal exposure to endocrine disrupting chemical mixtures and infant birth weight: A Bayesian analysis using kernel machine regression. Environmental Research. 2021;195:110749. doi:10.1016/j.envres.2021.110749

45. Mitro SD, Sagiv SK, Rifas-Shiman SL, et al. Per- and Polyfluoroalkyl Substance Exposure, Gestational Weight Gain, and Postpartum Weight Changes in Project Viva. Obesity. 2020;28(10):1984-1992. doi:10.1002/oby.22933

46. Czarnota J, Gennings C, Colt JS, et al. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study. Environmental Health Perspectives. 2015;123(10):965-970. doi:10.1289/ehp.1408630

47. Hargarten PM, Wheeler DC. Accounting for the uncertainty due to chemicals below the detection limit in mixture analysis. Environmental Research. 2020;186:109466. doi:10.1016/j.envres.2020.109466

48. A. Wang, K.L. Devick, J.F. Bobbs, A. Navas-Acien, B.A. Coull, L. Valeri. BKMR-CMA: A Novel R Command for Mediation Analysis in Environmental Mixture Studies. In: ISEE Conference Abstracts. ; 2020.

49. Alexander Keil. qgcomp: Quantile G-Computation. Published online 2021.

50. Herring AH. Nonparametric Bayes Shrinkage for Assessing Exposures to Mixtures Subject to Limits of Detection. Epidemiology. 2010;21(4):S71-S76. doi:10.1097/EDE.0b013e3181cf0058

51. Buuren S van, Groothuis-Oudshoorn K. mice : Multivariate Imputation by Chained Equations in R. Journal of Statistical Software. 2011;45:1-67. doi:10.18637/jss.v045.i03

52. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In: 3rd International Workshop on Distributed Statistical Computing. ; 2003:124-124.

53. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64:583-639. doi:10.1111/1467-9868.00353

54. Plummer M. Penalized loss functions for Bayesian model comparison. Biostatistics. 2008;9:523-539. doi:10.1093/biostatistics/kxm049

55. Colt JS, Gunier RB, Metayer C, et al. Household vacuum cleaners vs. the high-volume surface sampler for collection of carpet dust samples in epidemiologic studies of children. Environmental Health. 2008;7:6. doi:10.1186/1476-069X-7-6

56. Ward MH, Colt JS, Metayer C, et al. Residential Exposure to Polychlorinated Biphenyls and Organochlorine Pesticides and Risk of Childhood Leukemia. Environmental Health Perspectives. 2009;117:1007-1013. doi:10.1289/ehp.0900583

57. Metayer C, Colt JS, Buffler PA, et al. Exposure to herbicides in house dust and risk of childhood acute lymphoblastic leukemia. Journal of Exposure Science & Environmental Epidemiology. 2013;23:363-370. doi:10.1038/jes.2012.115

58. Ward MH, Colt JS, Deziel NC, et al. Residential Levels of Polybrominated Diphenyl Ethers and Risk of Childhood Acute Lymphoblastic Leukemia in California. Environmental Health Perspectives. 2014;122:1110-1116. doi:10.1289/ehp.1307602

59. Shoari N, Dubé JS. Toward improved analysis of concentration data: Embracing nondetects. Environmental Toxicology and Chemistry. 2018;37:643-656. doi:10.1002/etc.4046

60. Deziel NC, Rull RP, Colt JS, et al. Polycyclic aromatic hydrocarbons in residential dust and risk of childhood acute lymphoblastic leukemia. Environmental Research. 2014;133:388-395. doi:10.1016/j.envres.2014.04.033

61. Christoforidou EP, Riza E, Kales SN, et al. Bladder cancer and arsenic through drinking water: A systematic review of epidemiologic evidence. Journal of Environmental Science and Health, Part A. 2013;48:1764-1777. doi:10.1080/10934529.2013.823329

62. Engel A, Lamm SH. Arsenic exposure and childhood cancer--a systematic review of the literature. Journal of environmental health. 2008;71:12-16. http://www.ncbi.nlm.nih.gov/pubmed/18990928

Specific Aim 2: Develop and identify the variable clustering method best suited for use in chemical mixture analysis with Bayesian group index regression.

Paper: "Comparison of Variable Clustering Methods in the Context of Group Index Regression"

Authors: Matthew Carli, Mary H. Ward, James R. Cerhan, Nat Rothman, David C. Wheeler

Abstract

Recent scientific interest in the relationship between chemical exposures and human health has led to the development of new methods for mixture analysis. One of these new methods, Bayesian group index regression, allows for the modelling of chemical exposure variables in multiple groups that can vary in direction and magnitude of association with an outcome. A question this presents is how a set of variables should be partitioned into their respective groups. To address this aspect of Bayesian group index modeling, we compare five variable clustering methods to assess their ability to empirically cluster chemical exposure variables. To evaluate these clustering methods, we conduct simulation studies with varying numbers of true chemical clusters and between-cluster noise. We apply the best performing method to the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) non-Hodgkin Lymphoma (NHL) case-control study. Of the methods assessed, Clustering of Variables around Latent Variables (CLV) most accurately identified the correct clusters, resulting in better fitting Bayesian group index regression models. Our analysis of the NCI-SEER dataset found significant associations, one positive and one inverse, between two different indices of residential pesticide concentrations and NHL in subjects living in Iowa. In conclusion, CLV is a clustering method that is robust to noise and able to separate variables of opposite direction of association with the outcome, making it suitable for clustering in preparation for fitting a Bayesian group index regression model, providing practitioners a ready way to pre-process and analyze chemical mixture data.

Introduction

Synthetic chemicals are essential to modern economic life, with applications ranging from agriculture [1] to commercial products [2]. The number of chemicals on the market is believed to be around 75,000 to 140,000, however, data on important characteristics of these compounds such as environmental persistence, bioconcentration levels, and toxicity is minimal [3]. With historical examples of deleterious side effects of chemicals abounding [4-5], it is of increasing interest to understand the full impact of industrial chemicals, especially as it relates to human health. A particular area of growing scholarly interest is the study of chemical mixtures' impact on human

health, as researchers wish to evaluate the joint effect of the many chemicals that individuals are exposed to in their environments.

The study of chemical mixtures poses a set of statistical challenges, but the primary hindrance is that chemical mixtures tend to be highly correlated, which frustrates attempts at normal regression methods. To address this problem, a number of novel statistical methods have been developed. Bayesian kernel machine regression (BKMR) utilizes a kernel link function to test the association of a chemical mixture with an outcome, and has the advantage of effectively modelling nonlinear, nonadditive relationships and simplifying models through a hierarchical variable selection routine [6]. Quantile g-computation applies the framework of causal inference to modeling chemical mixtures, which allows for complex models including nonlinear and interaction effects [7]. Lastly, the index regression family of models groups the chemicals which compose a mixture into indices, allowing for both the total effect of the mixture on an outcome and the individual contribution of particular chemicals to the overall mixture effect to be estimated [8]. Originally developed for a single index, Bayesian group index regression is an extension that allows for multiple groups to be modeled in a Bayesian framework, eliminating the need for data splitting and preventing chemicals with associations in opposite direction from biasing index effect estimates towards the null [9].

While the ability to model multiple groups is a clear advantage over single index regression models, there still remains the question of how to partition a set of chemicals into their respective groups. In previous applications of group index regression models, chemicals were sorted into clusters with other compounds of similar chemical structure or usage [10]. For example, polychlorinated biphenyls (PCBs) would be grouped with other PCBs, or herbicides would be grouped with other herbicides. While this grouping strategy is logical, as highly correlated chemicals tend to be related to each other in one of these two ways, there are instances where this is not ideal. Pesticides, for example, are a relatively heterogenous group of chemicals that do not necessarily belong in the same group. Such groups can contain chemicals with both positive and negative associations to an outcome of interest, and might artificially bias such an index to the null. Additionally, there may be patterns in a chemical mixture beyond that of chemical structure or usage that an empirical measure of similarity would be able to ascertain.

The goal of empirically grouped chemical indices is related to the wider field of data clustering, where we seek to cluster features into groups most similar to one another. While the majority of work in the field of data clustering

has been devoted to the clustering of subjects, there has also been interest in the clustering of variables, particularly in the study of gene expression data [11-12]. Many well-known clustering algorithms have been employed to cluster variables, including K-means [13], hierarchical clustering [14-15], self-organizing maps [16], and model-based approaches [17-18]. Principal component analysis is also of interest for clustering, although interpretation of cluster membership and defining exact cluster boundaries is difficult with this technique [19]. Other methods have been developed specifically to suit the task of clustering variables, some of which will be discussed in detail below.

Our aim in this paper was to identify the variable clustering method most suited for identifying chemical groups in Bayesian group index regression models. To this end we chose five variable clustering methods for comparison. Three of these methods, Clustering of Variables around Latent Variables (CLV), Clustering Variables using Dimensionality Reduction (VARCLUST), and Robust principal component analysis (RPCA) clustering, involve the widely known PCA algorithm. CLV seeks to cluster variables around latent components, which are defined as the first principal component of a group of variables. VARCLUST expands on CLV by allowing these latent components to be defined by more than just the first principal component. RPCA clustering employs a variant of PCA robust to noise and corrupt observations to generate a denoised data matrix from which grouping labels for variables are derived. In addition to these methods, we investigated an agglomerative hierarchical clustering (AHC) algorithm that successively groups individual variables until a single group is formed out of smaller merged groups. Finally, we implemented a Dirichlet Process Variable Clustering (DPVC) model, where one of its primary attractive features is the ability to estimate group number as well as group composition.

To evaluate the performance of these five clustering techniques in tandem with Bayesian group index regression, we conducted a simulation study with varying levels of noise and true number of groups. Model performance was then compared along a number of metrics. We then applied the best performing clustering method and the Bayesian group index model to the National Cancer Institute (NCI) Surveillance, Epidemiology, and End results (SEER) non-Hodgkin Lymphoma (NHL) case-control study, an investigation of the link between environmental chemical exposures and NHL. The NHL study dataset includes highly correlated chemical exposure variables, which could be grouped based on chemical structure and usage, but may benefit from an empirical grouping rationale. This is especially true for the pesticides, which consist of heterogeneous chemical classes from which multiple groups could

be formed or combined with other groups. The results from this paper will provide a ready way to group chemicals when performing a Bayesian group index regression analysis.

Methods

Bayesian Group Index Regression

The Bayesian grouped index model in general form for a binary health outcome $y_i \sim Bernoulli(p_i)$ is specified through the log-odds of disease of the *i*th subject as

$$\text{logit}(p_i) = \beta_0 + \sum_{\text{k}=1}^{\text{K}} \beta_{\text{k}} \left(\sum_{j=1}^{C_k} w_{jk}\, q_{ijk} \right) + z_i^T \varphi.$$

On the left of the equation is the logit of the disease probability p_i, and on the right are the effects for the intercept β_0, chemical indices β_k, which estimate the health effects for exposure to the *k*th group of exposures, and a vector of covariates z_i^T with corresponding effects in vector φ. The number of exposures in each of the *K* indices can vary and is denoted by C_k. For each index, w_{jk} is the weight for the *j*th exposure in the *k*th index and denotes the relative importance of that exposure within the index. The value of each w_{jk} is constrained to be between 0 and 1, and when summed across an individual index must equal 1. For each index, q_{ijk} is the quantile score for the *j*th exposure in the *k*th index for the *i*th subject. Quantiles are used instead of raw chemical concentration data in order to limit the influence of outliers and to standardize the varying concentration scaling of different exposures. Further details on prior specification and inference have been discussed previously [9].

Grouping Methods

Clustering of Variables around Latent Variables

CLV is a variable clustering method that seeks to group variables by finding a cluster arrangement that maximizes the covariance of individual clusters with an associated latent variable. A fixed number of clusters and latent variables *K* are sought for a set of *p* variables $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}$ measured on *n* subjects. We denote the *K* clusters as $G_1, G_2, \dots, G_K$ and the *K* latent variables as $\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_K}$. The CLV algorithm seeks to maximize

$$\text{T} = \text{n} \sum_{k=1}^{K} \sum_{j=1}^{p} \delta_{kj} Cov^2(\boldsymbol{x_j}, \boldsymbol{c_k}),$$

under the constraint $\boldsymbol{c_k^T c_k} = 1$ where $\delta_{kj} = 1$ if the jth variable belongs to cluster G_K and $\delta_{kj} = 0$ otherwise. The quantity *T* can also be written as $\mathrm{T} = \frac{1}{n}\sum_{k=1}^{K} \boldsymbol{c_k^T X_k X_k^T c_k}$, where $\boldsymbol{X_k}$ is the matrix whose columns are formed with the variables belonging to G_k. $\boldsymbol{c_k}$ is defined as the first principal component of $\boldsymbol{X_k}$.

The optimal clustering is found through an iterative algorithm where variables move between clusters in order to achieve an increase in the criterion value *T*. The algorithm is as follows:

1. An initial *K* groups are formed through an agglomerative hierarchical clustering process that also utilizes the criterion *T*.
2. For each cluster G_k the latent component $\boldsymbol{c_k}$ is found by deriving the standardized principal component.
3. New clusters are formed by reassigning variables to a new group if its squared covariance with another latent variable is higher than its current group's.
4. Steps 2-3 repeat until stability is reached [20].

Clustering Variables Using Dimensionality Reduction (VARCLUST)

The VARCLUST algorithm is an extension of the CLV clustering method, which allows for more than one principal component to be used when modelling clusters. A dataset is partitioned into $\boldsymbol{X^i}$ clusters $i \in \{1, \dots, K\}$. Each individual cluster VARCLUST specifies is modelled as follows:

$$\boldsymbol{X^i} = \boldsymbol{M^i} + \boldsymbol{\mu^i} + \boldsymbol{E^i},$$

Where $\boldsymbol{M^i}$ is a matrix of rank *k*, the dimensionality or number of principal components used for this cluster, $\boldsymbol{\mu^i}$ is a vector of means, and $\boldsymbol{E^i}$ a matrix of centered normal noise distributed $N(0, \sigma_i^2 I)$. The denoised matrix $\boldsymbol{M^i}$ is decomposed as a product

$$\mathbf{M^i} = \boldsymbol{F^i_{n \times k_i} C^i_{k_i \times p_i}},$$

where the matrix $\boldsymbol{F^i}$ contains the *k* principal components and $\boldsymbol{C^i}$ the PCA coefficients. The VARCLUST algorithm seeks to find the optimal number of *K* clusters, the variable membership of the clusters, and the dimensionality *k* of each cluster. The algorithm is as follows:

1. One to a maximum value of $\vec{k}$ dimensions are considered for each cluster by maximizing the semi-integrated likelihood (PESEL) criterion, a metric developed to estimate the number of principal components in PCA that is similar to BIC [21].

2. PCA with principle components *k* is performed on columns of $\boldsymbol{X^i}$ partitioned to maximize the BIC distance from $\boldsymbol{F^i}$.

3. The number of clusters *K* is finally determined by choosing the model with the highest mBIC, a BIC metric modified to account for datasets with many variables [22].

Robust Principal Component Analysis

RPCA is an alternative to the widely used PCA that seeks to reduce dimensionality without being subject to PCA's brittleness to corrupted observations. It does this by decomposing any given data set into a low-rank matrix ***L*** containing the predominant pattern of the data and a sparse matrix ***S*** containing outliers and noise outside of the main pattern [23]. The original RPCA algorithm has since been extended to make it more suitable for environmental mixture analysis, including distinct penalties for BDL missing data, a nonnegativity constraint on the ***L*** matrix, and the replacement of the minimization of the nuclear norm of ***L*** for a rank-*r* projection of the ***L*** matrix. The following optimization problem is minimized:

$$\min_{\mathbf{L},\mathbf{S}} \mathbf{1}_{\mathrm{rank}(\mathbf{L})\leq \mathrm{r}} + \lambda\left|\left|\mathbf{S}\right|\right|_1 + \mu\left|\left|\mathbf{L}+\mathbf{S}-\mathbf{X}\right|\right|_{\mathrm{F}},$$

where ***X*** is the original data matrix, λ and μ are tuning parameters, $\mathbf{1}_{\mathrm{rank(L)}\leq \mathrm{r}}$ is an indicator function constraining ***L*** to be of rank $\leq r$, and $\left|\left|\boldsymbol{S}\right|\right|_1$ is the L1 norm of the sparse matrix ***S***. The final term is the error between predicted and observed values [24]. We apply the matrix decomposition of RPCA to cluster variables as follows:

1. We take the ***L*** matrix, as this contains the dominant patterns of the original data.

2. We perform traditional PCA on the denoised ***L*** matrix.

3. For a predetermined number of clusters *K*, we take only the first *K* principal components.

4. We assign each variable's group label as the principal component with the maximum absolute loading weight.

Agglomerative Hierarchical Clustering

AHC methods begin with each variable assigned to its own cluster, after which the two most similar clusters are joined. This process continues until only a single cluster remains [25]. There is no overlapping cluster membership, and clustering is done sequentially based on some measure of similarity of group pairs [26]. Graphically this process is represented as a tree diagram, and cluster assignments can be determined for any number of clusters by cutting the tree at varying heights. There are many similarity or distance measures that quantify how closely related clusters are to each other. We chose the Hoeffding D statistic as it is sensitive to many types of dependence [27]. For our assessment of AHC, we used the `varclus` function of the Hmisc R package, an agglomerative hierarchical clustering method for clustering variables [28].

Dirichlet Process Variable Clustering

DPVC is a Bayesian nonparametric model that partitions variables into highly correlated groups while simultaneously estimating the appropriate number of groups. The variables in a given dataset are partitioned using the Chinese Restaurant Process (CRP), which defines a distribution over clusters without necessitating the assignment of a maximum possible cluster number [29]. We once again take *X*, a data matrix of *N* observations and *P* variables. The CRP partitioning is expressed as

$$(c_1, \dots, c_P) \sim CRP(\alpha)$$

where $c_p = k$ denotes variable *p* belongs to cluster *k* and α is the concentration parameter. An attribute of the CRP is that each variable is restricted to belong to only one cluster. Each cluster is assigned a single latent factor $\mathrm{z_{kn}} \sim \mathrm{N}(0, \sigma_z^2)$ to model correlations between its variables. The observed data is modelled as $x_{pn} = g_p z_{c_p \mathrm{n}} + \epsilon_{pn}$ where g_p is a factor loading for variable *p* and $\epsilon_{pn} \sim N(0, \sigma_p^2)$ is Gaussian noise. A Gaussian prior $N\left(0, \sigma_g^2\right)$ is placed on every element g_d independently. Finally, the concentration parameter and variances (σ_g^2, σ_p^2) are respectively assigned gamma and inverse gamma priors [30].

Simulation Study Design

To evaluate the performance of these five variable grouping methods in the context of Bayesian group index regression, we simulated chemical concentration data in three scenarios with different numbers of true groups. All

scenarios were generated with a binary outcome and each group in every scenario contained a single important chemical that was set by assigning a true chemical weight of 1 to the important chemical and 0 to nonimportant chemicals. Chemical predictors were given between-group and within-group correlations that varied by scenario and are detailed below. For each set of simulated datasets, the correlation structures were specified through a matrix and subsequently converted into a covariance matrix. A mean vector and standard deviation vector were chosen to form the covariance matrix. This allowed the construction of the data distributed as multivariate normal.

Scenario 1 datasets were generated to have 14 chemical predictor variables clustered into 3 true groups. These groups were associated with the outcome with odds ratios (ORs) of 0.67, 1.00, and 1.50, with the positive and negatively associated groups containing 5 predictors and the null group containing 4 predictors. All three groups were given a within-group correlation of 0.5. Scenario 2 datasets had 22 predictors clustered into 5 groups. Three of these groups are the same as Scenario 1 with the two additional groups associated with the outcome with odds ratios of 0.50 and 2.00. These two highly associated groups each contained 4 predictors and were given a within-group correlation of 0.9. For both Scenarios 1 and 2 between-group correlation was set to be 0.0 (no noise), 0.1 (low noise), and 0.3 (moderate noise).

Scenario 3 datasets contained 31 predictors split into 7 true groups. These groups were associated with the outcome with odds ratios of 0.40, 0.50, 0.67, 1.00, 1.50, 2.00, and 2.50, each group containing 5, 3, 3, 5, 5, 5, and 5 predictors, respectively. The 0.50 and 2.00 groups were modelled to be highly correlated, each having a 0.7 within-group correlation and a slightly smaller 0.5 between-group correlation. Another such pair was formed with the 0.67 and 1.50 groups, each having a 0.7 within-group correlation and a 0.3 between-group correlation. These two pairs simulated the challenge of properly clustering a group of highly correlated chemicals that would ideally be split into separate groups due to their opposite association with the outcome. The remaining groups simulated were more distinct, with the 0.40 and 2.50 groups set to have a 0.9 between-group correlation and the null group having a 0.5 between-group correlation. Besides the variables in the two high-noise pairings, all other between-group correlation was set to 0.1.

Once true exposure effects and correlation structures had been defined for all scenarios, we created binary outcomes that replicated a case-control study by having a relatively balanced number of cases and controls (50% ±

10% cases) in each iteration of data generation. The binary outcome y was distributed as $y \sim Binomial(n, p)$ where $p = \frac{1}{1+e^{\eta}}$ and $\eta = \beta_0^* + \sum_{k=1}^{3} \beta_k^* [\sum_{j=1}^{C_k} w_{jk}^* q_{ijk}]$, and the star notation indicates true parameter values. No covariates were used in generation of the data, so the term $z^T \phi = 0$. The number of quantiles used in all simulations was set at four when computing the weighted index for each group (i.e. $q_{ij} = 0,1,2,3$). Each simulation group number and correlation structure combination used 100 data realizations.

We assessed the relative performance of our chosen grouping methods by determining the accuracy with which they assigned variables to their proper groups. Additionally, as some methods that we compared do not deterministically return a pre-assigned number of groupings, we reported the distribution of group numbers found across the 100 data realizations. For the groups formed, we calculated the bias, mean squared error (MSE), and power of the group exposure effects. We also calculated the sensitivity and specificity of identifying chemicals as important or not. We compared model fit with the deviance information criteria (DIC). To calculate power, we determined the proportion of 95% credible intervals (CIs) for ORs that did not include 1.00. We measured sensitivity by determining the proportion of important chemicals that were identified by the models as being important. This was done by determining if the estimated weight of the important chemicals produced by the models was greater than or equal to the threshold $\frac{1}{C_k}$. Important chemicals assigned to the wrong group were counted as errors. Likewise, we defined specificity as the proportion of the unimportant chemicals that were correctly deemed unimportant by the models. This was determined by checking if the estimated weights of the unimportant chemicals were less than the same threshold of $\frac{1}{C_k}$. DIC was defined as $DIC = \overline{D} + p_D$, where $\overline{D}$ is the posterior mean deviance [31] and p_D is the effective number of parameters [32], a quantification of model complexity.

Data Analysis

Applying the grouping method indicated by the results of the simulation study, we performed a Bayesian group index regression analysis of the NCI-SEER NHL case-control study to assess if there exists any association between our chemical exposure indices and NHL. The NCI-SEER NHL study is a population-based case control study of NHL with participants drawn from four study centers: the Detroit metropolitan region, Los Angeles County, the Seattle metropolitan region, and the state of Iowa. Patients diagnosed with NHL without a history of HIV at one of the above

four SEER registries between July 1, 1998 and June 30, 2000, and age 20 to 74 years old were included as cases. Controls were selected from the same four geographic regions using random-digit dialing for controls younger than 65 years old and Medicare eligibility files for controls 65 years and older. Controls were frequency matched to cases by age, sex, race, and SEER registry, and excluded if a history of either NHL or HIV was reported. In total, the study enrolled 2,378 eligible participants (1,321 cases and 1,057 controls). Further characterization of the study design and study population can be found in past publications [33-34].

To assess study participants' exposure to environmental chemicals, dust samples were taken from participants' homes from their vacuum cleaner bags or bagless vacuums if participants had their carpets and rugs for a minimum of five years. Further information on the dust collection eligibility, sampling, and laboratory methods can be found in previous publications [35-36]. Dust was analyzed for 27 chemicals and complete covariate data were available for 1,180 subjects (672 cases and 508 controls). Our analysis investigated the association of these 27 chemicals with the risk of NHL. Previous analyses of the NHL-SEER dataset that sought to group chemical exposure variables have largely done so on the basis of similar chemical structure or use, identifying three categories: polychlorinated biphenyls (PCBs) (congeners 105, 138, 153, 170, 180), polycyclic aromatic hydrocarbons (PAHs) (benz(a)anthracene, benzo(a)pyrene, benzo(b)fluoranthene, benzo(k)fluoranthene, chrysene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene), and pesticides (α-Chlordane, γ-Chlordane, carbaryl, dichlorodiphenyldichloroethylene (DDE), dichlorodiphenyltrichloroethane (DDT), *o*-phenylphenol, pentachlorophenol, propoxur, chlorpyrifos, *cis*-permethrin, *trans*-permethrin, 2,4-D, diazinon, dicamba, methoxychlor). In some analyses, pesticides were further split into two groups based on individual pesticide's direction of association with the outcome [37-39]. We determined the number and composition of groups in the Bayesian group index model using the grouping method indicated from our simulation studies. In addition to these chemical exposure indices, we adjusted for age, gender, race, and level of education. Age was treated as continuous, gender as binary (male vs. reference female), race as binary (white vs. reference black or other), and education as ordinal (grouped as <12 years, 12–15 years, and ≥16 years).

Due to the substantial differences between the chemical exposure profiles of the four study centers, we performed four separate analyses. Continuous chemical concentrations were categorized in quartiles for the regression. Convergence of all parameters of interest in models were checked via a Gelman-Rubin diagnostic statistic upper CI less than 1.10. We summarized the results using ORs for each chemical index along with 95% credible intervals.

Within each index that was significantly associated with the outcome, we assess the important chemical exposures using the estimated weights.

Results

Simulation Study

The results for the three between-group correlation variants of Scenario 1 are presented below in Tables 1-3. In the no noise variant, all grouping methods except DPVC formed three groups 100 percent of the time. Group numbers for the 100 data realizations are given in Tables 1-7 under each method name, with the group number designated by a “G” and the number of realizations following an equal sign (e.g. 3G = 100 meaning 3 groups found 100 times). This is not surprising, as DPVC is the only method that provides an estimated group number, while all other methods compared must have a group number designated. Despite under- or over-estimating the group number 38% of the time, DPVC still outperformed VARCLUST in terms of accuracy and DIC. While DPVC also outperformed VARCLUST in terms of power, sensitivity, and specificity, these figures only reflect the averages of instances where DPVC correctly specified the true group number. The best performing methods were CLV and AHC, which both returned the true group chemical composition with 100% accuracy. As a result, there is almost no difference in their performance metrics and these two methods performed best in terms of MSE, power, sensitivity, specificity, and DIC. Close second to these two methods was RPCA, which had only slightly lower accuracy, power, sensitivity, specificity, and DIC while matching their performance measured by bias and MSE.

Table 1: Scenario 1 (no noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV	β_1 = 0.62 (0.49, 0.79)	-0.07	0.02	0.95	1.00	0.94	1.00	568.57 (5.69)
(3G=100)	β_2 = 0.99 (0.79, 1.25)	-0.01	0.01	0.03	0.36	0.61		
	β_3 = 1.54 (1.22, 1.95)	0.03	0.02	0.93	0.98	0.95		
AHC	β_1 = 0.62 (0.49, 0.79)	-0.07	0.02	0.95	1.00	0.94	1.00	569.08 (5.94)
(3G=100)	β_2 = 1.00 (0.80, 1.25)	0.00	0.01	0.03	0.39	0.62		
	β_3 = 1.54 (1.22, 1.95)	0.03	0.02	0.93	0.98	0.95		
RPCA	β_1 = 0.63 (0.49, 0.80)	-0.06	0.02	0.94	0.97	0.93	0.97	570.09 (6.30)
(3G=100)	β_2 = 0.99 (0.79, 1.25)	-0.01	0.01	0.03	0.35	0.61		
	β_3 = 1.54 (1.20, 1.97)	0.03	0.02	0.87	0.96	0.95		
VARCLUST	β_1 = 0.64 (0.48, 0.86)	-0.04	0.04	0.79	0.87	0.90	0.70	574.22 (7.18)
(3G=100)	β_2 = 1.00 (0.79, 1.27)	0.00	0.01	0.03	0.16	0.52		
	β_3 = 1.50 (1.11, 1.98)	0.00	0.04	0.69	0.80	0.90		
DPVC*	β_1 = 0.65 (0.51, 0.85)	-0.02	0.04	0.83	0.90	0.93	0.86	572.84 (5.84)
(1G=5, 2G=31,	β_2 = 1.00 (0.80, 1.25)	0.00	0.01	0.03	0.37	0.64		
3G=62, 4G=2)	β_3 = 1.41 (1.09, 1.81)	-0.06	0.08	0.76	0.83	0.91		

* Performance metrics only averaged for instances of correct group number specification

These relative performances were largely replicated in the low noise variant, with the notable exception of DPVC. The ability of DPVC to correctly specify group number fell dramatically, and as a consequence average DIC rose substantially and accuracy fell. As three groups were only estimated for a single data realization, the ORs, CIs, bias, MSE, power, sensitivity, and specificity numbers are not averages but the results for that single instance of correct group number, which saw the worst performance among the clustering methods. While VARCLUST underperformed all clustering methods but DPVC, it proved more robust to noise than DPVC, with only a slight increase in bias and DIC, a slight decrease in accuracy, and small improvements in power and specificity. RPCA proved less robust to noise, as it failed to estimate three groups on one occasion and performed slightly worse across all performance metrics. As in the no noise variant, CLV and AHC performed best and retrieved the cluster membership of chemical variables with 100% accuracy. Compared to the no noise variant, there was a slight increase in bias and DIC for each method, as well as a small increase in specificity.

Table 2: Scenario 1 (low noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV	β_1 = 0.62 (0.49, 0.79)	-0.07	0.02	0.99	1.00	0.95	1.00	572.22 (6.87)
(3G=100)	β_2 = 1.00 (0.80, 1.26)	0.00	0.01	0.03	0.39	0.59		
	β_3 = 1.56 (1.23, 1.98)	0.04	0.02	0.90	0.98	0.95		
AHC	β_1 = 0.62 (0.49, 0.79)	-0.07	0.02	0.99	1.00	0.95	1.00	571.10 (5.89)
(3G=100)	β_2 = 1.00 (0.80, 1.26)	0.00	0.01	0.03	0.39	0.60		
	β_3 = 1.56 (1.23, 1.98)	0.04	0.02	0.90	0.98	0.95		
RPCA*	β_1 = 0.61 (0.47, 0.80)	-0.08	0.02	0.95	0.95	0.93	0.90	572.36 (6.33)
(2G=1, 3G=99)	β_2 = 1.00 (0.78, 1.29)	0.00	0.01	0.02	0.29	0.59		
	β_3 = 1.57 (1.20, 2.04)	0.04	0.03	0.80	0.94	0.94		
VARCLUST	β_1 = 0.64 (0.49, 0.85)	-0.04	0.04	0.79	0.85	0.92	0.67	574.59 (5.76)
(3G=100)	β_2 = 0.99 (0.79, 1.24)	-0.01	0.01	0.03	0.13	0.36		
	β_3 = 1.52 (1.15, 1.99)	0.01	0.04	0.72	0.84	0.91		
DPVC*	β_1 = 0.95 (0.63, 1.45)	0.35	0.30	0.48	0.72	0.92	0.38	586.99 (5.98)
(1G=93, 2G=6,	β_2 = 0.98 (0.77, 1.24)	-0.02	0.00	0.00	0.00	0.50		
3G=1)	β_3 = 0.99 (0.65, 1.50)	-0.42	0.34	0.46	0.64	0.91		

* Performance metrics only averaged for instances of correct group number specification

In the moderate noise variant, the previously noted trends continued. CLV and AHC continued to perfectly return the true chemical composition of the three groups. The increase in noise did, however, lead to an increase in bias, MSE, and DIC and to a decrease in power and specificity upon subsequent Bayesian group index regression. DPVC was unable to distinguish any partitions in the simulations under this level of noise, resulting in a lack of estimates for two of the groups. This led to an extreme bias towards the null for the single group estimated, as well as a large increase in DIC and poor performance across all metrics. The performance of VARCLUST, while still inferior to other methods, did not see a marked deterioration in performance, with only slight decreases in power and specificity and slight increases in bias, MSE, and DIC. RPCA, on the other hand, saw a significant drop in accuracy, power, and sensitivity, as well as a substantial increase in DIC. Bias, MSE and specificity were also slightly worse.

Table 3: Scenario 1 (moderate noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV (3G=100)	β_1 = 0.61 (0.46, 0.79) β_2 = 1.01 (0.78, 1.31) β_3 = 1.59 (1.21, 2.08)	-0.10 0.01 0.06	0.02 0.01 0.03	0.95 0.00 0.83	0.99 0.36 0.99	0.92 0.58 0.95	1.00	575.23 (6.44)
AHC (3G=100)	β_1 = 0.61 (0.46, 0.79) β_2 = 1.01 (0.78, 1.31) β_3 = 1.59 (1.21, 2.08)	-0.10 0.01 0.06	0.02 0.01 0.03	0.95 0.01 0.81	0.99 0.37 0.99	0.92 0.58 0.95	1.00	575.65 (6.85)
RPCA (3G=100)	β_1 = 0.61 (0.44, 0.87) β_2 = 1.01 (0.76, 1.36) β_3 = 1.57 (1.11, 2.18)	-0.09 0.01 0.05	0.04 0.02 0.04	0.75 0.02 0.69	0.86 0.16 0.80	0.90 0.60 0.88	0.69	578.02 (7.43)
VARCLUST (3G=100)	β_1 = 0.63 (0.47, 0.87) β_2 = 1.01 (0.80, 1.27) β_3 = 1.50 (1.12, 2.00)	-0.05 0.01 0.00	0.04 0.01 0.05	0.73 0.02 0.62	0.91 0.19 0.78	0.90 0.34 0.88	0.68	578.90 (6.37)
DPVC[1] (1G=100)	β_1 = 0.97 (0.67, 1.39) β_2 = NA β_3 = NA	0.37	0.23	0.28	0.73	0.90	0.36	590.42 (3.53)

[1] Failure to estimate correct number of groups resulted in NA values

The trends seen in Scenario 1 were largely followed in Scenario 2, with the three correlation variants presented below in Tables 4-6. In the no noise variant, the CLV and AHC methods once again were 100% accurate, and therefore had nearly identical results for all other performance metrics. RPCA was only slightly less accurate by comparison, due to the fact that in four instances it failed to assign chemicals to a fifth group. This resulted in a small increase in DIC and MSE compared to AHC and CLV, as well as a slightly worse performance in power, sensitivity, and specificity for the weaker signal strength β_2 and β_4. VARCLUST was the least accurate method, and in particular had poor performance measured by bias, MSE, power, sensitivity, and specificity for the less associated β_2 and β_4. It also had a markedly higher DIC than the three previously discussed methods. The increase in group number in this scenario impacted the performance of DPVC the most, as its rate of correctly specifying group number dropped to 39%. This resulted in the worst performance as measured by DIC. When the correct number of groups were modelled, however, DPVC's power, sensitivity, and specificity for β_2 and β_4 were second only to AHC and CLV. This was not the case for the more highly associated groups, where these metrics were worse in comparison to all other methods. Bias and MSE were also generally higher for DPVC.

Table 4: Scenario 2 (no noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV	β_1 = 0.49 (0.39, 0.61)	-0.03	0.01	1.00	0.94	0.94	1.00	521.68 (10.07)
(5G=100)	β_2 = 0.62 (0.48, 0.80)	-0.07	0.02	0.96	0.94	0.94		
	β_3 = 1.00 (0.78, 1.28)	0.00	0.01	0.05	0.48	0.59		
	β_4 = 1.56 (1.21, 2.01)	0.04	0.02	0.90	0.92	0.92		
	β_5 = 2.04 (1.64, 2.55)	0.02	0.01	1.00	0.92	0.91		
AHC	β_1 = 0.49 (0.39, 0.61)	-0.03	0.01	1.00	0.94	0.94	1.00	520.49 (9.88)
(5G=100)	β_2 = 0.62 (0.48, 0.80)	-0.07	0.02	0.96	0.94	0.94		
	β_3 = 1.00 (0.78, 1.28)	0.00	0.01	0.05	0.46	0.61		
	β_4 = 1.56 (1.21, 2.01)	0.04	0.02	0.90	0.92	0.92		
	β_5 = 2.04 (1.64, 2.55)	0.02	0.01	1.00	0.92	0.91		
RPCA*	β_1 = 0.48 (0.38, 0.61)	-0.03	0.02	1.00	0.93	0.93	0.95	522.24 (10.08)
(4G=4, 5G=96)	β_2 = 0.63 (0.48, 0.82)	-0.06	0.02	0.92	0.90	0.92		
	β_3 = 1.01 (0.78, 1.30)	0.01	0.02	0.03	0.43	0.58		
	β_4 = 1.55 (1.20, 2.00)	0.03	0.02	0.89	0.84	0.89		
	β_5 = 2.06 (1.64, 2.61)	0.03	0.02	1.00	0.91	0.91		
VARCLUST	β_1 = 0.48 (0.36, 0.63)	-0.04	0.06	0.96	0.74	0.87	0.72	539.66 (14.13)
(5G=100)	β_2 = 0.73 (0.55, 0.98)	0.09	0.06	0.53	0.47	0.75		
	β_3 = 1.00 (0.78, 1.27)	0.00	0.01	0.01	0.15	0.32		
	β_4 = 1.39 (1.01, 1.87)	-0.08	0.05	0.54	0.52	0.79		
	β_5 = 2.04 (1.55, 2.70)	0.02	0.05	0.96	0.73	0.84		
DPVC*	β_1 = 0.52 (0.41, 0.69)	0.05	0.07	0.91	0.72	0.88	0.83	546.38 (16.94)
(5G=39, 4G=37,	β_2 = 0.62 (0.47, 0.80)	-0.08	0.03	0.95	0.92	0.93		
3G=21, 2G=3)	β_3 = 0.99 (0.77, 1.26)	-0.01	0.01	0.00	0.46	0.58		
	β_4 = 1.54 (1.19, 1.98)	0.02	0.02	0.90	0.90	0.90		
	β_5 = 1.86 (1.44, 2.42)	-0.07	0.08	0.87	0.73	0.86		

* Performance metrics only averaged for instances of correct group number specification

The low noise variant continued the trend seen previously, with DPVC almost unable to partition chemicals into groups in the presence of noise, at maximum identifying two groups 12% of the time. DIC and accuracy were quite poor as a result, and other metrics suffered from the bias to the null that occurs when variables from oppositely associated groups were mixed. As in Scenario 1, the number of instances where RPCA under-specified the group number increased slightly in the presence of noise, and as a result its accuracy and DIC suffered. Bias and MSE were largely the same, while power, sensitivity, and specificity saw small decreases. VARCLUST's performance in terms of bias and MSE was stable compared to the no noise variant, yet was the weakest performing method besides DPVC. Accuracy saw a slight decrease along with power and sensitivity, while specificity was marginally better.

Table 5: Scenario 2 (low noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV	β_1 = 0.49 (0.39, 0.61)	-0.03	0.01	1.00	0.92	0.93	1.00	525.72 (10.20)
(5G=100)	β_2 = 0.62 (0.48, 0.80)	-0.08	0.02	0.93	0.94	0.93		
	β_3 = 1.00 (0.78, 1.29)	0.00	0.01	0.01	0.47	0.61		
	β_4 = 1.58 (1.22, 2.04)	0.05	0.02	0.88	0.84	0.90		
	β_5 = 2.07 (1.65, 2.61)	0.03	0.02	1.00	0.87	0.90		
AHC	β_1 = 0.49 (0.39, 0.61)	-0.03	0.01	1.00	0.92	0.92	1.00	525.66 (10.46)
(5G=100)	β_2 = 0.62 (0.48, 0.80)	-0.08	0.02	0.93	0.94	0.92		
	β_3 = 1.00 (0.78, 1.29)	0.00	0.01	0.01	0.48	0.61		
	β_4 = 1.58 (1.22, 2.04)	0.05	0.02	0.89	0.86	0.90		
	β_5 = 2.07 (1.65, 2.61)	0.03	0.02	1.00	0.88	0.91		
RPCA*	β_1 = 0.48 (0.37, 0.61)	-0.04	0.02	1.00	0.87	0.89	0.87	528.72 (9.15)
(4G=7, 5G=93)	β_2 = 0.64 (0.49, 0.87)	-0.04	0.02	0.81	0.86	0.88		
	β_3 = 1.01 (0.78, 1.33)	0.01	0.02	0.04	0.37	0.54		
	β_4 = 1.55 (1.17, 2.03)	0.03	0.02	0.88	0.80	0.87		
	β_5 = 2.08 (1.60, 2.73)	0.04	0.02	0.99	0.89	0.90		
VARCLUST	β_1 = 0.49 (0.37, 0.64)	-0.03	0.05	0.95	0.80	0.90	0.70	545.02 (13.43)
(5G=100)	β_2 = 0.73 (0.55, 0.97)	0.08	0.06	0.48	0.53	0.79		
	β_3 = 0.99 (0.77, 1.28)	-0.01	0.01	0.03	0.15	0.37		
	β_4 = 1.37 (1.02, 1.81)	-0.09	0.05	0.52	0.51	0.79		
	β_5 = 2.07 (1.59, 2.72)	0.03	0.05	0.96	0.76	0.88		
DPVC[1]	β_1 = 0.88 (0.49, 1.52)	0.57	0.75	0.61	0.71	0.88	0.25	600.13 (19.76)
(1G=88, 2G=12)	β_2 = NA							
	β_3 = NA							
	β_4 = NA							
	β_5 = 0.97 (0.55, 1.67)	-0.73	0.95	0.58	0.62	0.88		

* Performance metrics only averaged for instances of correct group number specification

[1] Failure to estimate correct number of groups resulted in NA values

In the moderate noise variant, DPVC once again failed to partition the variables 100% of the time, leading to high DIC, low accuracy, and estimates biased to the null. RPCA's rate of group number under-specification increased to 10%, with an attendant increase in DIC and decrease in accuracy. Bias increased for some groups and decreased for others, however, MSE was consistently higher than previous variants. Power, sensitivity, and specificity suffered, especially for β_2 and β_4. VARCLUST saw a significant increase in DIC, with only DPVC registering a worse model fit. While still underperforming other methods, the increase to moderate noise levels did not have any large impact on bias, MSE, or power, while accuracy slightly increased. Sensitivity and specificity both saw weaker performance. Once again CLV and AHC proved most robust to increased noise, each achieving 100% accuracy. The moderate noise, however, did lead to worse results in all performance metrics compared to low and no noise variants.

Table 6: Scenario 2 (moderate noise) performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV (5G=100)	β_1 = 0.48 (0.38, 0.62)	-0.03	0.01	1.00	0.85	0.93	1.00	538.65 (10.13)
	β_2 = 0.61 (0.45, 0.81)	-0.09	0.03	0.90	0.84	0.91		
	β_3 = 1.00 (0.75, 1.33)	0.00	0.02	0.03	0.50	0.61		
	β_4 = 1.60 (1.20, 2.14)	0.07	0.02	0.86	0.82	0.88		
	β_5 = 2.07 (1.62, 2.67)	0.03	0.02	1.00	0.85	0.90		
AHC (5G=100)	β_1 = 0.48 (0.38, 0.62)	-0.03	0.01	1.00	0.87	0.93	1.00	538.84 (9.94)
	β_2 = 0.61 (0.45, 0.81)	-0.09	0.03	0.90	0.86	0.92		
	β_3 = 1.00 (0.75, 1.33)	0.00	0.02	0.03	0.50	0.62		
	β_4 = 1.60 (1.20, 2.14)	0.07	0.02	0.86	0.82	0.88		
	β_5 = 2.07 (1.62, 2.67)	0.03	0.02	1.00	0.85	0.89		
RPCA* (4G=10, 5G=90)	β_1 = 0.46 (0.34, 0.62)	-0.08	0.04	0.99	0.80	0.90	0.77	547.40 (10.53)
	β_2 = 0.66 (0.48, 0.94)	-0.01	0.03	0.61	0.69	0.84		
	β_3 = 1.01 (0.74, 1.37)	0.01	0.02	0.01	0.27	0.57		
	β_4 = 1.49 (1.09, 2.04)	-0.01	0.04	0.61	0.64	0.80		
	β_5 = 2.15 (1.60, 2.92)	0.07	0.03	0.98	0.79	0.85		
VARCLUST (5G=100)	β_1 = 0.49 (0.37, 0.65)	-0.01	0.05	0.97	0.68	0.85	0.72	558.18 (13.16)
	β_2 = 0.72 (0.53, 1.00)	0.08	0.05	0.48	0.50	0.78		
	β_3 = 0.98 (0.75, 1.28)	-0.02	0.01	0.03	0.15	0.40		
	β_4 = 1.39 (1.02, 1.88)	-0.07	0.05	0.53	0.54	0.76		
	β_5 = 2.02 (1.49, 2.74)	0.01	0.04	0.93	0.66	0.86		
DPVC[1] (1G=100)	β_1 = 0.95 (0.64, 1.48)	0.65	0.65	0.53	0.69	0.90	0.23	602.52 (8.28)
	β_2 = NA							
	β_3 = NA							
	β_4 = NA							
	β_5 = NA							

* Performance metrics only averaged for instances of correct group number specification

[1] Failure to estimate correct number of groups resulted in NA values

Scenario 3 (Table 7) presents the first instance where AHC failed to maintain perfect accuracy. Compared to CLV, the AHC method had higher bias and MSE for the most weakly associated groups, and had markedly lower power, sensitivity, and specificity for these groups as well. AHC's DIC was also higher than both the CLV and RPCA methods. RPCA had the third highest accuracy, but suffered from a significant under-specification of group number, finding all seven groups only 79% of the time. For instances in which RPCA specified the correct number of groups, bias, MSE, sensitivity, and specificity were comparable to AHC, higher in some instances and lower in others. RPCA's power was generally better than AHC, but underperformed CLV. VARCLUST was once again the second least accurate method after DPVC, and had the second highest DIC. With seven groups VARCLUST's weakness in estimating groups besides the most highly associated are highlighted, with extreme bias and MSE for $\beta_2, \beta_3, \beta_5$, and β_6. Power, sensitivity and specificity was also poor for these groups. Unsurprising given the high levels of noise, DPVC generally failed to partition the data, with 84% rate of finding a single group, and never estimated the correct group number. The

effects of this on performance metrics were similar to previous instances. CLV once again had an accuracy of 100%, and the lowest DIC. Predictably, power suffered the most for the weakest associated groups. On the other hand, sensitivity was highest for these groups. CLV's bias was generally the lowest, with some exceptions for certain groups, while MSE was consistently the lowest.

Table 7: Scenario 3 performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV	β_1 = 0.37 (0.28, 0.48)	-0.07	0.02	1.00	0.73	0.91	1.00	482.10 (13.51)
(7G=100)	β_2 = 0.45 (0.34, 0.59)	-0.10	0.03	1.00	0.72	0.87		
	β_3 = 0.66 (0.51, 0.85)	-0.01	0.02	0.86	0.93	0.90		
	β_4 = 0.99 (0.75, 1.30)	-0.01	0.02	0.05	0.33	0.62		
	β_5 = 1.55 (1.18, 2.02)	0.03	0.01	0.89	0.90	0.84		
	β_6 = 2.21 (1.67, 2.95)	0.10	0.03	1.00	0.71	0.91		
	β_7 = 2.66 (2.05, 3.50)	0.06	0.02	1.00	0.74	0.91		
AHC	β_1 = 0.39 (0.30, 0.51)	-0.01	0.02	1.00	0.84	0.92	0.92	501.08 (13.64)
(7G=100)	β_2 = 0.55 (0.42, 0.72)	0.10	0.06	0.91	0.41	0.80		
	β_3 = 0.74 (0.57, 0.99)	0.11	0.04	0.54	0.57	0.79		
	β_4 = 0.99 (0.74, 1.31)	-0.01	0.01	0.02	0.29	0.63		
	β_5 = 1.35 (1.00, 1.79)	-0.11	0.04	0.51	0.48	0.77		
	β_6 = 1.82 (1.38, 2.40)	-0.09	0.05	0.93	0.43	0.83		
	β_7 = 2.52 (1.98, 3.25)	0.01	0.02	1.00	0.91	0.94		
RPCA*	β_1 = 0.38 (0.28, 0.51)	-0.04	0.03	0.99	0.68	0.90	0.86	499.03 (16.68)
(5G=1,	β_2 = 0.48 (0.35, 0.65)	-0.05	0.04	0.96	0.58	0.84		
6G=20,	β_3 = 0.70 (0.52, 0.97)	0.05	0.03	0.71	0.78	0.83		
7G=79)	β_4 = 0.97 (0.70, 1.35)	-0.03	0.02	0.03	0.24	0.62		
	β_5 = 1.41 (1.00, 1.94)	-0.06	0.04	0.58	0.58	0.76		
	β_6 = 2.08 (1.52, 2.87)	0.04	0.05	0.91	0.44	0.82		
	β_7 = 2.80 (2.05, 3.91)	0.11	0.09	0.99	0.68	0.88		
VARCLUST	β_1 = 0.40 (0.29, 0.53)	0.00	0.07	0.97	0.82	0.91	0.71	522.99 (16.46)
(7G=100)	β_2 = 0.62 (0.46, 0.86)	0.21	0.11	0.62	0.48	0.80		
	β_3 = 0.86 (0.62, 1.18)	0.25	0.09	0.15	0.22	0.64		
	β_4 = 1.02 (0.75, 1.38)	0.02	0.01	0.01	0.09	0.52		
	β_5 = 1.20 (0.87, 1.62)	-0.23	0.08	0.20	0.25	0.66		
	β_6 = 1.61 (1.13, 2.17)	-0.22	0.11	0.63	0.39	0.80		
	β_7 = 2.47 (1.81, 3.34)	-0.01	0.07	0.94	0.80	0.90		
DPVC[1]	β_1 = 0.84 (0.45, 1.58)	0.74	0.93	0.53	0.78	0.93	0.19	619.66 (29.13)
(1G=84, 2G=14,	β_2 = NA							
3G=1, 4G=1)	β_3 = NA							
	β_4 = NA							
	β_5 = NA							
	β_6 = NA							
	β_7 = 0.99 (0.53, 1.87)	-0.92	1.21	0.48	0.63	0.92		

* Performance metrics only averaged for instances of correct group number specification

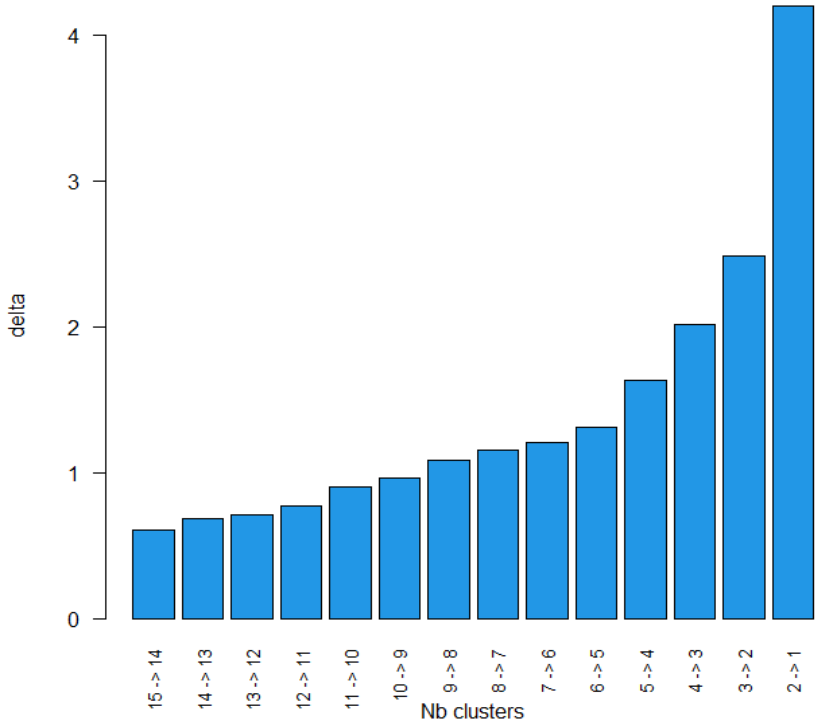
[1] Failure to estimate correct number of groups resulted in NA values

Application of CLV grouping and Bayesian group index regression to NCI-SEER NHL case-control study data

The results of our simulation study indicate that the CLV grouping method is best suited for grouping environmental chemical exposure variables before performing Bayesian group index regression. We applied the two-step process to

the NHL dataset. This was done for each study center separately; however, we will only present the Iowa results in the main tables as no significant associations were found in other study centers. The results for the other three study centers can be found in the supplemental materials. First, as suggested by the authors Vigneau and Qannari 2003, we assessed the proper group number for the data graphically (Figure 1).

Figure 1: Iowa Subset Group Number Plot



In the above plot we have the variation of the clustering criterion between a partition into *K* clusters and a partition into *K*-1 clusters. Generally, this variation tends to increase as cluster sizes decrease; however, we want to identify at what grouping number change the first instance of a large change in delta occurs. We identified this as the point between 5 and 4 groups, so we fixed the CLV clustering algorithm to 5 clusters. We characterized the five clusters and list their chemicals as follows: a group of pesticides named Group 1 (2,4-D, chlorpyrifos, *cis*-permethrin, and *trans*-permethrin), a second group of pesticides called Group 2 (dicamba, DDE, DDT, and propoxur), a group containing all PAHs named Group 3 (benz(a)anthracene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, chrysene, dibenz(ah)anthracene, and indeno(1,2,3-cd)pyrene), a group containing all PCBs and four pesticides named Group 4 (PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, carbaryl, methoxychlor, *o*-phenylphenol, pentachlorophenol), and a group containing the remaining pesticides called Group 5 (α-chlordane, γ-chlordane, diazinon). The odds ratios and 95% CIs estimated for our 5 index effects and covariates are in Table 8. Two indices

were significantly associated with NHL: Group 1 had an inverse association (OR = 0.58, 95% CI: 0.41, 0.78) and Group 2 had a positive association (OR = 1.50, 95% CI: 1.00, 2.11). None of the covariates were found to be significantly associated with NHL risk. The Group 1 index was dominated by 2,4-D, with a weight of 0.69. Propoxur and DDE were the most heavily weighted chemicals in the Group 2 index, with weights of 0.58 and 0.23 respectively.

Table 8: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Iowa

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	**0.58**	**0.41**	**0.78**
Group 2	**1.50**	**1.00**	**2.11**
Group 3	0.96	0.77	1.19
Group 4	0.99	0.70	1.41
Group 5	1.20	0.89	1.60
Male	1.07	0.70	1.66
White	1.39	0.16	9.47
Education	1.13	0.77	1.74
Age	0.99	0.97	1.01

Discussion

In this paper, we compared the performance of five variable clustering methods as a data pre-processing step before Bayesian group index regression. In our simulation study we evaluated these clustering methods at three levels of noise (none, low, and moderate) and varying numbers of true groups (3, 5, and 7). Additionally, we sought to determine if the clustering methods could successfully separate correlated but oppositely associated chemical variables that, if grouped together, lead to index estimates biased towards the null. We found that the performance of the DPVC and RPCA methods were particularly susceptible to higher levels of noise, and that the CLV and AHC methods were quite robust to even moderate levels of noise. The VARCLUST method uniformly underperformed the other methods. The CLV and AHC methods were nearly indistinguishable in terms of performance until the 7-group scenario, where CLV was slightly superior in retrieving true group status and separating oppositely associated variables. Notably, RPCA consistently failed to separate the correct number of groups at high group numbers.

Based on the findings described above, we recommend the CLV method for clustering of variables prior to Bayesian group index regression. Based on these results we used CLV to cluster the chemical exposures in the NHL study dataset and identified distinct chemical clusters for Detroit, Iowa, Los Angeles, and Seattle. We then fit four Bayesian group index regressions based on these empirical groupings. For Iowa, our empirical clusters partially reflected clustering by chemical structure, as there were two groups that contained all the PCBs and PAHs. The pesticides, however, were separated into three clusters and four were grouped with PCBs, characterizing the heterogeneity of these chemicals. This resulted in a five-index model, where we found a positive and significant association between what was labelled Group 2 (OR =1.50) and NHL, with propoxur (weight = 0.58) and DDE (weight = 0.23) having the highest mean posterior weights. A negative and significant association was also found between what was labelled Group 1 (OR = 0.58) and NHL, with the highest mean posterior weight attributed to 2,4-D (weight = 0.69). These significant associations between pesticides and NHL are supported by previous analyses of these data. Using group index regression methods, we previously found significant associations between two indices in the Iowa study center, one containing all pesticides with an univariate positive association with NHL status and the other containing all pesticides with an univariate inverse association. As in the analysis presented above, propoxur was the highest weighted chemical in the positively associated index, while 2,4-D was the highest in the inversely associated index [39]. Another analysis of the NCI-SEER NHL dataset, employing a single index regression, found a significant, positive association between all 27 chemicals and NHL in the Iowa subset, with propoxur and DDE among the highest weighted chemicals [37]. Interestingly, 2,4-D was assigned a marginal weight (0.005) in this positively associated index, effectively contributing no weight according to this method, but was found to have a significant inverse association with NHL in a single-chemical regression reported in the same analysis. This highlights the strength of the group index regression model to allow for multiple directions of associations among indices and the ability of the CLV clustering method to sort oppositely associated chemicals into their own groups. The chemicals of interest found in our positively associated pesticide index are supported by past single-chemical analyses, where similar associations as those found here were estimated for DDE and propoxur [36-37]. While 2,4-D is classed as a Group 2B, or possible carcinogenic to humans, by the International Agency for Research on Cancer working group [40], multiple investigations into its relationship with NHL have been conducted with inconclusive results [41-44].

In our analyses for the Los Angeles, Detroit, and Seattle centers we found no significant associations between any chemical index and NHL. This result is similar to our previous group index regression analysis that also looked at each study site individually [39]. These results differ from another previous analysis of the same data, where a two-step, frequentist grouped index regression model was used. In addition to a positively associated pesticide index, that analysis found a significant and positive association between dust concentrations of PCBs and NHL [38], which was consistent with analyses of individual and total PCBs [36] and with an analysis of some study participants' blood plasma [45]. The discrepancy of findings may be explained by the two-step procedure used, as failure to split the data into estimation and validation sets could lead to overfitting.

In conclusion, the two-step process of CLV clustering and Bayesian group index regression is a useful combination in the analysis of chemical mixture data. The ability to empirically determine group indices provides guidance when groupings are otherwise unclear and can be used to confirm groupings based on other rationales such as chemical structure. The CLV clustering algorithm showed a robustness to noisy, highly correlated data that is typical of chemical concentration data. Further, in application to both simulation and real data the method was able to separate oppositely associated variables into distinct clusters. These attributes allow practitioners to take full advantage of the Bayesian group index regression's ability to model indices with varying magnitudes and direction of association. While this combination has several strengths, there are also limitations that future research can address. While CLV was able to separate oppositely-associated chemicals, an approach that directly takes the outcome of interest into account could improve clustering performance. Similarly, covariates could also be accounted for in the clustering process.

References

1. Casida JE, Quistad GB. Golden Age of Insecticide Research: Past, Present, or Future? *Annual review of entomology*. 1998;43(1):1-16. doi:10.1146/annurev.ento.43.1.1
2. Calafat AM, Valentin-Blasini L, Ye X. Trends in Exposure to Chemicals in Personal Care and Consumer Products. *Current environmental health reports*. 2015;2(4):348-355. doi:10.1007/s40572-015-0065-9
3. Johnson AC, Jin X, Nakada N, Sumpter JP. Learning from the past and considering the future of chemicals in the environment. *Science (American Association for the Advancement of Science)*. 2020;367(6476):384-387. doi:10.1126/science.aay6637
4. Turusov V, Rakitsky V, Tomatis L. Dichlorodiphenyltrichloroethane (DDT): Ubiquity, Persistence, and Risks. *Environmental health perspectives*. 2002;110(2):125-128. doi:10.1289/ehp.02110125
5. Woodcock BA, Isaac NJB, Bullock JM, et al. Impacts of neonicotinoid use on long-term population changes in wild bees in England. *Nature communications*. 2016;7(1):12459-12459. doi:10.1038/ncomms12459
6. Bobb JF, Valeri L, Claus Henn B, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics (Oxford, England)*. 2015;16(3):493-508. doi:10.1093/biostatistics/kxu058
7. Keil AP, Buckley JP, O'Brien KM, Ferguson KK, Zhao S, White AJ. A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures. *Environmental health perspectives*. 2020;128(4):47004-. doi:10.1289/EHP5838
8. Czarnota J, Gennings C, Wheeler DC. Assessment of Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. *Cancer informatics*. 2015;2015(Suppl. 2):159-171. doi:10.4137/CIN.S17295
9. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Bayesian Group Index Regression for Modeling Chemical Mixtures and Cancer Risk. *International journal of environmental research and public health*. 2021;18(7):3486-. doi:10.3390/ijerph18073486
10. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. *International journal of environmental research and public health*. 2021;18(2):504-. doi:10.3390/ijerph18020504
11. Jiang D, Tang C, Zhang A. Cluster analysis for gene expression data: a survey. *IEEE transactions on knowledge and data engineering*. 2004;16(11):1370-1386. doi:10.1109/TKDE.2004.68
12. Sherlock G. Analysis of large-scale gene expression data. *Current Opinion in Immunology*. 2000;12(2):201-205. doi:10.1016/S0952-7915(99)00074-6
13. De Smet F, Mathys J, Marchal K, Thijs G, De Moor B, Moreau Y. Adaptive quality-based clustering of gene expression profiles. *Bioinformatics*. 2002;18(5):735-746. doi:10.1093/bioinformatics/18.5.735
14. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences - PNAS*. 1998;95(25):14863-14868. doi:10.1073/pnas.95.25.14863
15. Alon U, Barkai N, Notterman DA, et al. Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *Proceedings of the National Academy of Sciences - PNAS*. 1999;96(12):6745-6750. doi:10.1073/pnas.96.12.6745

16. Tamayo P, Slonim D, Mesirov J, et al. Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation. *Proceedings of the National Academy of Sciences - PNAS*. 1999;96(6):2907-2912. doi:10.1073/pnas.96.6.2907

17. Ghosh D, Chinnaiyan AM. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*. 2002;18(2):275-286. doi:10.1093/bioinformatics/18.2.275

18. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL. Model-based clustering and data transformations for gene expression data. *Bioinformatics*. 2001;17(10):977-987. doi:10.1093/bioinformatics/17.10.977

19. Quackenbush J. Computational analysis of microarray data. *Nature reviews Genetics*. 2001;2(6):418-427. doi:10.1038/35076576

20. Vigneau E, Qannari EM. Clustering of Variables Around Latent Components. *Communications in statistics Simulation and computation*. 2003;32(4):1131-1150. doi:10.1081/SAC-120023882

21. Sobczyk P, Bogdan M, Josse J. Bayesian Dimensionality Reduction With PCA Using Penalized Semi-Integrated Likelihood. *Journal of computational and graphical statistics*. 2017;26(4):826-839. doi:10.1080/10618600.2017.1340302

22. Bogdan M, Ghosh JK, Doerge RW. Modifying the Schwarz Bayesian information criterion to locate multiple interacting quantitative trait loci. *Genetics (Austin)*. 2004;167(2):989-999. doi:10.1534/genetics.103.021683

23. Candès E, Li X, Ma Y, Wright J. Robust principal component analysis? *Journal of the ACM*. 2011;58(3):1-37. doi:10.1145/1970392.1970395

24. Gibson EA, Zhang J, Yan J, et al. Principal Component Pursuit for Pattern Identification in Environmental Mixtures. *Environmental health perspectives*. 2022;130(11):117008-. doi:10.1289/EHP10479

25. Murtagh F, Contreras P. Algorithms for hierarchical clustering: an overview. *Wiley interdisciplinary reviews Data mining and knowledge discovery*. 2012;2(1):86-97. doi:10.1002/widm.53

26. Day WHE, Edelsbrunner H. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*. 1984;1(1):7-24. doi:10.1007/BF01890115

27. Hoeffding W. A Non-Parametric Test of Independence. *The Annals of mathematical statistics*. 1948;19(4):546-557. doi:10.1214/aoms/1177730150

28. Harrell FE. Hmisc: Harrell Miscellaneous. R package version 4.7-0. 2022. https://CRAN.R-project.org/package=Hmisc

29. Broderick T, Jordan MI, Pitman J. Cluster and Feature Modeling from Combinatorial Stochastic Processes. *Statistical science*. 2013;28(3):289-312. doi:10.1214/13-STS434

30. Palla K, Ghahramani Z, Knowles D. A nonparametric variable clustering model. *Advances in Neural Information Processing Systems*. 2012;25(4):2987-2995.

31. Spiegelhalter DJ, Best NG, Carlin BP, Van Der Linde A. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B, Statistical methodology*. 2002;64(4):583-639. doi:10.1111/1467-9868.00353

32. Plummer M. Penalized loss functions for Bayesian model comparison. *Biostatistics*. 2008;9:523-539. doi:10.1093/biostatistics/kxm049

33. Chatterjee N, Hartge P, Cerhan JR, et al. Risk of Non-Hodgkin's Lymphoma and Family History of Lymphatic, Hematologic, and Other Cancers. *Cancer epidemiology, biomarkers & prevention*. 2004;13(9):1415-1421. doi:10.1158/1055-9965.1415.13.9

34. Morton LM, Wang SS, Cozen W, et al. Etiologic heterogeneity among non-Hodgkin lymphoma subtypes. *Blood*. 2008;112(13):5150-5160. doi:10.1182/blood-2008-01-133587

35. 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. *Journal of exposure analysis and environmental epidemiology*. 2004;14(1):74-83. doi:10.1038/sj.jea.7500307

36. Colt JS, Severson RK, Lubin J, et al. Organochlorines in Carpet Dust and Non-Hodgkin Lymphoma. *Epidemiology (Cambridge, Mass)*. 2005;16(4):516-525. doi:10.1097/01.ede.0000164811.25760.f1

37. Czarnota J, Gennings C, Colt JS, et al. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study. *Environmental health perspectives*. 2015;123(10):965-965. doi:10.1289/ehp.1408630

38. Wheeler DC, Czarnota J. Modeling chemical mixture effects with grouped weighted quantile sum regression. *ISEE Conference Abstracts*. 2016.

39. Boyle J, Ward MH, Cerhan JR, Rothman N, Wheeler DC. Estimating mixture effects and cumulative spatial risk over time simultaneously using a Bayesian index low-rank kriging multiple membership model. *Statistics in medicine*. 2022;41(29):5679-5697. doi:10.1002/sim.9587

40. IARC. Agents Classified by the IARC Monographs, Volumes 1–129. Lyon, France: International Agency for Research on Cancer. 2021. Accessed March 20, 2023. http://monographs.iarc.fr/ENG/Classification/index.php

41. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. DDT, Lindane, and 2, 4-D. 2018.

42. Goodman JE, Loftus CT, Zu K. 2,4-Dichlorophenoxyacetic acid and non-Hodgkin's lymphoma, gastric cancer, and prostate cancer: meta-analyses of the published literature. *Annals of epidemiology*. 2015;25(8):626-636.e4. doi:10.1016/j.annepidem.2015.04.002

43. Ward MH, Lubin J, Giglierano J, et al. Proximity to Crops and Residential Exposure to Agricultural Herbicides in Iowa. *Environmental health perspectives*. 2006;114(6):893-897. doi:10.1289/ehp.8770

44. De Roos AJ, Fritschi L, Ward MH, et al. Herbicide use in farming and other jobs in relation to non-Hodgkin's lymphoma (NHL) risk. *Occupational and environmental medicine (London, England)*. 2022;79(12):795-806. doi:10.1136/oemed-2022-108371

45. De Roos AJ, Hartge P, Rothman N, et al. Persistent Organochlorine Chemicals in Plasma and Risk of Non-Hodgkin's Lymphoma. *Cancer research (Chicago, Ill).* 2005;65(23):11214-11226. doi:10.1158/0008-5472.CAN-05-1755

Specific Aim 3: Develop semi-supervised extension to previously identified variable clustering method and identify method best suited for use in chemical mixture analysis with Bayesian group index regression.

Paper: “Semi-supervised Clustering Methods Before Bayesian Group Index Regression”

Authors: Matthew Carli, Mary H. Ward, James R. Cerhan, Nat Rothman, David C. Wheeler

Abstract

Bayesian group index regression is a recently developed mixture analysis model that allows for the modelling of chemical exposure variables in multiple groups that can vary in direction and magnitude of association with an outcome. Before group index regression, the chemical mixture of interest must be partitioned into these groups. We propose two semi-supervised extensions of the clustering algorithm Clustering of Variables around Latent Variables (CLV). Our semi-supervised clustering methods seek to incorporate information from the target outcome variable to improve clusters while also preventing chemicals with opposite direction of association with the outcome from inclusion in the same group, which biases index effect estimates towards the null. We compare our proposed extensions with two other semi-supervised clustering algorithms and the unsupervised CLV algorithm. To evaluate these clustering methods, we conduct simulation studies characterized by true clusters with little to differentiate themselves from one another except their opposing association with an outcome variable. We apply the best performing method to the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) non-Hodgkin Lymphoma (NHL) case-control study to investigate associations between chemicals in house dust and risk of NHL. Our simulation study identified the constrained clustering algorithm Constrained Clustering by Tabu Search (Conclust) as the best performing clustering method. In our analysis of the NCI-SEER dataset we found three chemical indices with significant association with NHL, two in Iowa and one in Los Angeles. In conclusion, the semi-supervision of clustering as implemented by Conclust allows for the empirical partitioning of a chemical mixture while discouraging chemical groupings that are biased towards the null. These qualities are advantageous for the preparation of data for Bayesian group index regression and provides a rational two-step process for the analysis of chemical mixture data.

Introduction

Concern over the widespread proliferation of chemicals resulting from human activity is a decades-old phenomenon [1]. Nonetheless, due to the essential nature of these chemicals to modern life, the spread of such chemicals continues

[2]. This situation has led to increasing epidemiological interest in quantifying the impact of chemical pollution, with a particular focus on its effect on human health. As scientific investigation into this problem has grown more sophisticated, researchers have begun to prioritize the study of chemical mixtures' impact on human health, estimating the joint effect of the many chemicals one may encounter in day-to-day life as opposed to single chemicals in isolation.

The principal statistical challenge in the study of chemical mixtures is the tendency of mixture components to be highly correlated. This frustrates normal regression and has necessitated the development of novel statistical methods. Among these methods are Bayesian kernel machine regression [3], quantile g-computation [4], and various implementations of generalized additive models [5-6]. These methods have all been widely applied and described in detail. An additional family of methods, index regression models, have also been developed to deal with the problem of chemical mixture collinearity.

Initially, index regression models such as weighted quantile sum regression sought to model the impact of a mixture on an outcome by estimating the effect of a single index composed of all chemicals of interest [7]. Inside the index, chemical weights are estimated to model the contributions of individual chemicals to the overall mixture effect. A drawback of single-index modelling is that chemicals with opposite directions of association with the outcome could bias the index towards the null. To address this weakness, extensions such as group weighted quantile sum regression (GWQS) [8] and Bayesian group index regression [9] were developed, allowing for the modelling of multiple indices that could separately accommodate positively and negatively associated chemicals.

The extension to multiple groups introduced a new problem to index regression modelling. When fitting a multi-group index regression model, either GWQS or Bayesian group index regression, the number and chemical composition of indices must be chosen by the user. Past applications of such models have organized exposure variables into chemicals that share a structural similarity (e.g., PCBs, PAHs, metals) or usage (e.g., herbicides, insecticides) [8-10]. This grouping strategy could be viewed as one reliant on domain-specific knowledge, and has several advantages. Chemicals that are similar in either structure or use have a greater chance of being highly correlated with each other, and if not grouped could give rise to multicollinearity effects. Indices grouped in this way also have ready interpretations as the joint effect of a recognizable class of chemicals on a health outcome. There

are two weaknesses to this approach. One, there are some chemical groups, such as pesticides, that tend to be heterogeneous and contain chemicals with both positive and negative associations with an outcome. An index formed without accounting for this heterogeneity will have an effect estimate biased towards the null. Second, a rigid adherence to chemicals groups as defined above precludes the discovery of empirical patterns of similarity, which may see chemicals across structure or usage groups combined in a single index. Such empirically derived indices may identify and characterize previously unknown predictor relationships and result in better fitting models.

The task of selecting the number of groups and group composition in a Bayesian group index regression model can be viewed as a cluster analysis problem. Cluster analysis encompasses a wide variety of methods employed for different reasons, but all share the common aim of grouping similar items together [11]. The goal is to capture some underlying mechanism at work in the data that causes some observations to have greater resemblance to each other than to other observations [12]. Clustering algorithms have historically been categorized as "unsupervised learning" [13]. The goal of unsupervised learning is to describe associations and patterns among a set of input measures, as opposed to supervised learning, where the goal is to predict the value of an outcome measure based on a set of input measures [14]. An intermediate category, referred to as semi-supervised learning, aims at some combination of the two goals [15]. Semi-supervised clustering algorithms, for example, seek to supplement standard cluster analysis with additional information. This additional information can take many forms, such as previous partial classification of a subset of inputs or the relationship between inputs and an outcome variable [16]. In the case of clustering in preparation for group index regression, we hypothesize that clustering supplemented by information from the targeted outcome variable will result in improved models.

Our aim in this paper was to incorporate information from the relationship between chemical exposure variables and the outcome variable of interest to improve clustering for Bayesian group index regression. Specifically, we sought to supervise clustering so as to discourage the grouping of chemicals with opposite directions of association with an outcome. In a previous paper we identified an unsupervised clustering method, Clustering of Variables around Latent Variables (CLV), as best suited to clustering chemical exposure variables when compared to other unsupervised variable clustering algorithms [17]. While previous work has been done on incorporating outcome information with the CLV clustering algorithm [18-19], the focus of these methods has been on identifying superior predictive clusters of variables. These methods do not partition a collection of variables into mutually exclusive

clusters, and so are not applicable for use before group index regression. To improve the performance of this clustering method for Bayesian group index regression purposes, we developed two extensions that incorporate information from a target outcome variable. The first, Constrained Clustering of Variables around Latent Variables (cCLV), extends CLV to allow users to define pairwise constraints to potential cluster membership. In the context of Bayesian group index regression, these constraints are used to penalize proposed clusters that contain chemicals with opposite directions of association with an outcome. The second, Outcome-adjusted Clustering of Variables around Latent Variables (oCLV), determines a cutoff value for chemical variables' univariate association with the outcome below which they are not considered for CLV clustering. This focuses the CLV clustering on chemical exposure variables most relevant to the subsequent regression. We compared CLV and the two extensions above with two other clustering methods: Constrained Clustering by Tabu Search (Conclust), a pairwise constrained clustering algorithm similar to cCLV, and Clusterwise Effect Regression (CLERE), a combination of clustering and regression models.

We evaluated the performance of these five clustering algorithms with simulated data designed to model heterogeneous chemical mixtures containing variables with opposing directions of association with a target outcome. Three scenarios were generated to investigate the effect of varying levels of correlation and group number. We compared the performance of both the clustering itself and the subsequent performance of these group assignments in Bayesian group index regression estimates. After identifying the best performing clustering algorithm, we applied it and Bayesian group index regression to the National Cancer Institute (NCI) Surveillance, Epidemiology, and End results (SEER) non-Hodgkin Lymphoma (NHL) case-control study, an investigation of the link between environmental chemical exposures and NHL. This dataset contains many pesticides, a chemical category that consists of heterogeneous chemical classes that often have completely opposite directions of association with a given outcome. Our findings improve upon previous work done on clustering before group index regression, and will offer a rationale for forming indices in Bayesian group index regression that take the final aim of regressing upon an outcome into account from the beginning of the model building process.

Methods

Bayesian Group Index Regression

The Bayesian grouped index model in general form for a binary health outcome $y_i \sim Bernoulli(p_i)$ is specified through the log-odds of disease of the *i*th subject as

$$\text{logit}(p_i) = \beta_0 + \sum_{\text{k}=1}^{\text{K}} \beta_{\text{k}} \left(\sum_{j=1}^{C_k} w_{jk}\, q_{ijk} \right) + z_i^T \varphi.$$

On the left of the equation is the logit of the disease probability p_i, and on the right are the effects for the intercept β_0, chemical indices β_k, which estimate the health effects for exposure to the *k*th group of exposures, and a vector of covariates z_i^T with corresponding effects in vector φ. The number of exposures in each of the *K* indices can vary and is denoted by C_k. For each index, w_{jk} is the weight for the *j*th exposure in the *k*th index and denotes the relative importance of that exposure within the index. The value of each w_{jk} is constrained to be between 0 and 1, and when summed across an individual index must equal 1. For each index, q_{ijk} is the quantile score for the *j*th exposure in the *k*th index for the *i*th subject. Quantiles are used instead of raw chemical concentration data in order to limit the influence of outliers and to standardize the varying concentration scaling of different exposures. Further details on prior specification and inference have been discussed previously [9].

Proposed Semi-supervised Clustering Methods

Constrained Clustering of Variables around Latent Variables

cCLV is an extension to the CLV algorithm that incorporates known limitations on which chemicals can be grouped together into the clustering process. The original CLV algorithm seeks to group variables by finding a cluster arrangement that maximizes the covariance of individual clusters with an associated latent variable [20] (see CLV section below). cCLV works by imposing a penalty on this covariance score when pre-defined user constraints are violated. cCLV is part of a wider family of constrained clustering algorithms, a form of semi-supervised clustering where partial data in the form of user-provided labels or pairwise constraints are used to guide the algorithm towards a more appropriate data partitioning. These constraints are commonly in the form of must-link or cannot-link pairs [21]. In our application to Bayesian group index regression, our focus was to discourage the clustering of chemicals with opposing direction of association with the target outcome variable, which took the form of cannot-link constraint pairs.

Let $\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_p}$ be a set of *p* variables measured on *n* subjects. Our goal is to partition these variables in to a fixed number of clusters *K*. We denote the *K* clusters as $G_1, G_2, \ldots, G_K$ and the corresponding *K* latent variables as $\boldsymbol{m_1}, \boldsymbol{m_2}, \ldots, \boldsymbol{m_K}$. We seek to maximize the following cost function:

$$\mathrm{U} = \sum_{k=1}^{K} \sum_{j=1}^{p} \mathrm{r}_{jk} Cov^2(\boldsymbol{x_j}, \boldsymbol{m_k}) - \sum_{(\boldsymbol{x_j}, \boldsymbol{x_{j'}}) \in Q} w_{j,j'} I[G_k = G_{k'}]$$

where $\mathrm{r}_{jk} \in 0,1$ denotes whether or not the *j*th variable belongs to cluster G_k, *Q* is the set of $(\boldsymbol{x_j}, \boldsymbol{x_{j'}})$ pairwise constraints, $w_{j,j'}$ is the weight of the penalty imposed on constrained pairings, and $I[G_k = G_{k'}]$ is an indicator function that applies the penalty when said pairings occur. The cost function for the criterion quantity *U* is composed of two terms. The first is the cost from the original CLV directional clustering algorithm [20], the sum of squared covariances between chemical variables and their current cluster's latent variable. The second is the cost representing violations of cannot-link constraints. Imposed when two cannot-link variables are grouped together, the constraint penalty $w_{j,j'}$ is determined by the confidence in the constraint, with high confidence in constraints resulting in a relatively large weight. All constraints are given equal weight.

The optimal clustering is found through an iterative algorithm where first the latent variable of each cluster is defined, after which cluster membership of all variables is reassigned in accordance with the maximum covariance with the new latent variables. These two steps repeat to maximize *U.* In detail the algorithm is as follows:

1. An initial *K* groups are formed through either an agglomerative hierarchical clustering process or random assignment, as determined by the user.
2. For each cluster G_k the latent component $\boldsymbol{c_k}$ is found by deriving the first standardized principal component.
3. New clusters are formed by reassigning variables to a new group if its squared covariance with another latent variable is higher than its current group's. Expressed more formally, $\mathrm{r}_{jk} = 1$ if $\max_{\mathrm{k}'}\{\mathrm{Cov}^2(\boldsymbol{x_j}, \boldsymbol{m_{k'}})\} = \mathrm{Cov}^2(\boldsymbol{x_j}, \boldsymbol{m_k})$.
4. Steps 2-3 repeat until stability is reached, or until the maximum number of iterations have been performed [18].

Outcome-adjusted Clustering of Variables around Latent Variables

Similar to cCLV, oCLV extends the CLV algorithm to incorporate information from the outcome variable of interest during clustering. It is an adaption of a semi-supervised clustering method originally called "supervised clustering", where the univariate associations of dataset features with an outcome variable are ranked and then all but the most

highly associated features are discarded before clustering [22]. oCLV adapts this method, so that CLV clustering is performed solely on to the variables most highly associated with the outcome. This is accomplished as follows:

1. For each variable in the dataset, a test statistic A_j is calculated for the univariate association between the *j*th variable and the outcome.
2. A cut-off value *M* is chosen, so that only variables with $|A_j| > M$ will have the CLV clustering algorithm applied to them [16].

It has been previously noted that a downside of this method is that the variables that fall under the cutoff value are not analyzed and are discarded. This is not the case for use in a group index regression, where the usually discarded variables may be placed into a "null" index and still included in the analysis. The cutoff value *M* is determined by cross validation.

Comparison Grouping Methods

Clustering of Variables around Latent Variables

CLV is an unsupervised variable clustering method where the measure of similarity determining cluster membership is the covariance of variables with a latent variable. In previous work, we determined that the CLV algorithm was best suited to the clustering of chemical exposure variables in preparation for group index regression [17]. Therefore, we wish to compare CLV to semi-supervised methods to better characterize the utility of clustering with regard to information from an outcome variable.

As with cCLV, a fixed number of clusters and latent variables *K* are sought for a set of *p* variables $\boldsymbol{x_1}, \boldsymbol{x_2}, \dots, \boldsymbol{x_p}$ measured on *n* subjects. We denote the *K* clusters as $G_1, G_2, \dots, G_K$ and the *K* latent variables as $\boldsymbol{c_1}, \boldsymbol{c_2}, \dots, \boldsymbol{c_K}$. Vigneau et al. [20] proposed two variants of the CLV algorithm, a local mode informed by covariances and a directional mode informed by squared covariances. We opt for the directional mode, as the chemical mixture datasets we seek to cluster exhibit relatively few and weak negative correlations. The CLV algorithm seeks to maximize

$$\mathrm{T} = \sum_{k=1}^{K} \sum_{j=1}^{p} \mathrm{r}_{jk} Cov^2(\boldsymbol{x_j}, \boldsymbol{m_k}),$$

under the constraint $\boldsymbol{m_k^T m_k} = 1$ where $\mathrm{r}_{jk} = 1$ if the jth variable belongs to cluster G_K and $\mathrm{r}_{jk} = 0$ otherwise. The quantity *T,* or the total squared covariance of chemical variables $\boldsymbol{x_j}$ with current latent variables $\boldsymbol{m_k}$, is the same as the first term of the cCLV cost function without the penalty cost term. $\boldsymbol{m_k}$ is defined as the first principal component

of the variables comprising G_K. The optimization of T is accomplished with the same iterative algorithm as detailed in the cCLV section above.

Conclust

The Conclust method is another constrained clustering algorithm. As with cCLV, in our application of this method to grouping chemical variables for group index regression we sought to constrain clustering so as to discourage the grouping of chemicals with opposite associations with the outcome variable of interest. Conclust seeks to minimize the following cost function:

$$J = \sum_{h=1}^{k} \sum_{x_i \in X_h} \left|\left|\boldsymbol{x_i} - \mu_h\right|\right|^2 + \sum_{(x_i,x_j)\in M} w_{ij}\, I\left[l_i \neq l_j\right] + \sum_{(x_i,x_j)\in C} \overline{w_{ij}}\, I\left[l_i = l_j\right]$$

The first term is the squared distance between *i*th variable $\boldsymbol{x_i}$ and μ_h, the center of cluster *h*. The second term penalizes clusterings from the set of must-link constraints *M* that are not grouped together. In our application of Conclust must-link constraints were not specified. The third term penalizes violations of the cannot-link constraints found in set *C* with the weight $\overline{w_{ij}}$.

The cost function above is minimized through a local search algorithm called tabu search. This algorithm compares the cost of an initial solution to those in that solution's neighborhood (in our example the set of potential clusters that differ by a single chemical from the current solution). The neighbor found to have the lowest cost is then updated to be the best solution. To prevent the algorithm from getting stuck a local optima, previous solutions are added to the tabu list, a list of solutions that the algorithm is forbidden from choosing for some number of iterations [23]. Conclust's tabu search algorithm is as follows:

1. Data are scaled and clustering initialized using the weighted farthest-first scheme.
2. An empty tabu list is initialized.
3. Scan the neighborhood of the current clustering.
4. Select the best neighbor in the neighborhood.
5. Update the current best solution and tabu list.
6. Repeat steps 3 – 5 until iteration of time limit is reached, then return the best solution [24].

Clusterwise Effect Regression (CLERE)

CLERE is a model that simultaneously clusters covariates and performs regression on a target outcome variable. This is accomplished by taking the fixed β parameters of the standard regression model and instead considering them as unobserved random variables following a mixture of Gaussian distributions containing some number of components. These composite effect parameters can be expressed as $\beta_j \sim \sum_{k=1}^{g} \pi_k N(b_k, \gamma^2)$, where for each β_j a multinomial distributed random variable $\boldsymbol{z} = (z_{j1}, \ldots, z_{jg})$ of parameter $\boldsymbol{\pi} = (\pi_1, \ldots, \pi_g)^T$ is assumed. The full model can be written as

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \epsilon_i$$
$$\epsilon_i \sim N(0, \sigma^2)$$
$$\beta_j | \boldsymbol{z}_j \sim N\left(\sum_{k=1}^{g} b_k z_{jk}, \gamma^2\right)$$
$$\boldsymbol{z}_j = (z_{j1}, \ldots, z_{jg}) \sim M(\pi_1, \ldots, \pi_g).$$

Where z_{jk} indicates that variable *j* is a member of cluster *k* and $M(\pi_1, \ldots, \pi_g)$ is the multinomial distribution. Further details on the CLERE method have been published, both for continuous [25] and binary [26] outcomes. For use in combination with group index regression, variables were assigned to the cluster having the largest posterior probability. These cluster assignments were the used in the subsequent index regression.

Simulation Study Design

We compared the performance of these five clustering methods in tandem with Bayesian group index regression by simulating chemical concentration data where there was little to distinguish groups other than their relationship with the target outcome variable. Three simulation scenarios were generated that varied in strength of between group correlation and group number. Each scenario was generated with a binary outcome variable, as the NHL dataset we apply our chosen clustering method to has a binary outcome. True groups were each assigned a single important chemical that would dominate the group's overall effect, with important chemicals assigned a true chemical weight of 1 and unimportant chemicals assigned a weight of 0.

Scenario 1 was generated to have 15 chemical predictor variables clustered into 3 true groups. These groups were associated with the outcome variable with odds ratios (ORs) of 0.67, 1.00 and 1.50, with each group comprised of five predictors. The variables of the null group were correlated with each other at a strength of 0.5 and with the variables of all other groups at 0.1. The variables of the two remaining groups were given within-group and between-

group correlations of 0.3. Scenario 2 shared the same predictor number, true group number, and outcome associations as Scenario 1. It differed in that the within-group and between-group correlations of the groups with significant associations with the outcome were increased to 0.5.

Scenario 3 datasets were generated to have 25 chemical predictor variables clustered into 5 true groups with 5 chemical predictors each. These groups were associated with the outcome variable with ORs of 0.50, 0.67, 1.00, 1.50, and 2.00. The groups with lesser strength associations of 0.67 and 1.50 were modelled as highly correlated with between and within-group correlations of 0.5. The more highly associated groups and the null association group were modelled as being relatively distinct, with within-group correlations of 0.7 and between group correlations with all other groups of 0.1.

With true exposure effects and correlation structures determined, we generated binary outcomes that would model a case-control study, with cases and controls having a rough balance (50% ± 10% cases) in each iteration of data generation. The binary outcome y was distributed as $y \sim Binomial(n, p)$ where $p = \frac{1}{1+e^{\eta}}$ and $\eta = \beta_0^* + \sum_{k=1}^{3} \beta_k^* [\sum_{j=1}^{C_k} w_{jk}^* q_{ijk}]$, and the star notation indicates true parameter values. No covariates were used in generation of the data, making the term $z^T \phi = 0$. The number of quantiles used in all simulations was set at four when computing the weighted index for each group (i.e. $q_{ij} = 0,1,2,3$). 100 data realizations were generated for each scenario.

We assessed the relative performance of our clustering methods with a number of metrics. Most directly related to the clustering process, we measured the accuracy with which our methods assigned chemicals to their true groups, as well as the distribution of group numbers found across the 100 data realizations. These were recorded with the group number designated by a "G" and the number of realizations following an equal sign (e.g. 3G = 100 meaning 3 groups found 100 times). The other performance metrics reflect the impact of clusterings on the subsequent group index regression. Overall model fit was compared with the deviance information criteria (DIC). For estimated index effects, we calculated the bias, mean squared error (MSE), and power. For the chemical weights within indices, we calculated the sensitivity and specificity of properly identifying important and unimportant chemicals. We define power as the proportion of 95% credible intervals (CIs) for ORs that did not include 1.00. We measured sensitivity by determining the proportion of important chemicals that were identified by the models as being important. This was

done by determining if the estimated weight of the important chemicals produced by the models was greater than or equal to the threshold $\frac{1}{C_k}$. Important chemicals assigned to the wrong group were counted as errors. Likewise, we defined specificity as the proportion of the unimportant chemicals that were correctly deemed unimportant by the models. This was determined by checking if the estimated weights of the unimportant chemicals were less than the same threshold of $\frac{1}{C_k}$. DIC was defined as $DIC = \overline{D} + p_D$, where $\overline{D}$ is the posterior mean deviance [27] and p_D is the effective number of parameters [28], a quantification of model complexity. Bayesian grouped index regression was performed using the R package BayesGWQS [29], which implements Bayesian grouped index models using Just Another Gibbs Sampler (JAGS) [30].

Data Analysis

Using the grouping methods identified by our simulation study, we performed a Bayesian group index regression analysis of the NCI-SEER NHL case-control study. We investigated the potential association between the chemical exposure groups identified by our clustering method and NHL. The NCI-SEER NHL study is a population-based case control study of NHL with subjects taken from four study centers: the Detroit metropolitan region, Los Angeles County, the Seattle metropolitan region, and the state of Iowa. Patients diagnosed with NHL without a history of HIV at one of the above four SEER registries between July 1, 1998 and June 30, 2000, and age 20 to 74 years old were included as cases. Controls were selected from the same four study centers using random-digit dialing for controls younger than 65 years old and Medicare eligibility files for controls 65 years and older. Controls were frequency matched to cases by age, sex, race, and SEER registry, and excluded if a history of either NHL or HIV was reported. In total, the study enrolled 2,378 eligible participants (1,321 cases and 1,057 controls). Further details on study design and study population can be found in past publications [31-32].

To quantify exposure to environmental chemicals, dust samples were taken from study participants' homes. Details on dust collection eligibility, sampling, and laboratory methods can be found in previous publications [33-34]. Dust sampled during the collection process was analyzed for the presence of 27 chemicals. Covariates of interest were also collected from study participants, and complete covariate data were available for 1,180 subjects (672 cases and 508 controls). Our analysis explored the association between the 27 chemicals and NHL. These chemicals, considered from the standpoint of similar chemical structure or usage, fall into three categories: polychlorinated biphenyls

(PCBs) (congeners 105, 138, 153, 170, 180), polycyclic aromatic hydrocarbons (PAHs) (benz(a)anthracene, benzo(a)pyrene, benzo(b)fluoranthene, benzo(k)fluoranthene, chrysene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene), and pesticides (α-Chlordane, γ-Chlordane, carbaryl, dichlorodiphenyldichloroethylene (DDE), dichlorodiphenyltrichloroethane (DDT), *o*-phenylphenol, pentachlorophenol, propoxur, chlorpyrifos, *cis*-permethrin, *trans*-permethrin, 2,4-D, diazinon, dicamba, methoxychlor). The number of indices and the chemical variables that each contain were determined by the best performing grouping method identified in our simulation study. We included the controlling covariates of age, gender, race, and level of education in our Bayesian group index models. Age was treated as continuous, gender as binary (male vs. reference female), race as binary (white vs. reference black or other), and education as ordinal (grouped as <12 years, 12–15 years, and ≥16 years).

We conducted four separate analyses for each of the study centers, as the chemical exposure profiles of these different geographic regions varied significantly. We categorized the continuous chemical concentration data into quartiles in preparation for regression. Convergence of all parameters of interest in models were checked via a Gelman-Rubin diagnostic statistic upper CI less than 1.10. We summarized the results using ORs for each chemical index along with 95% credible intervals. When indices were found to be significantly associated with the outcome, we investigated the most important chemical contributors to the association using estimated weights.

Results

Simulation Study

The results for Scenario 1 are presented below in Table 1. Of our two proposed semi-supervised clustering methods, only cCLV improved on the results of CLV. While the accuracy of cCLV was lower than that of CLV, cCLV was markedly superior in terms of bias, power, and sensitivity. cCLV also exhibited a slight decrease in DIC, while MSE and specificity were roughly equivalent. oCLV performed worse than CLV on all performance metrics except MSE. Conclust performed best among all the clustering methods as measured by DIC, power, sensitivity, and specificity. On the other hand, it had the second-lowest accuracy, and had the highest MSE and bias, showing a tendency to overestimate index effect sizes. The clustering generated by CLERE performed similarly to Conclust in some respects, with the second lowest DIC, the lowest accuracy, and comparable power and sensitivity. CLERE differed in that it had relatively low bias, the lowest MSE, and the lowest specificity of any clustering method. Importantly, CLERE was the

only clustering method that failed to consistently find the true number of groups, underestimating true group number 10% of the time.

Table 1: Scenario 1 performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV (3=100)	β_1 = 0.72 (0.53, 1.03) β_2 = 0.99 (0.74, 1.33) β_3 = 1.38 (0.99, 1.86)	0.08 -0.01 -0.08	0.04 0.01 0.05	0.44 0.01 0.45	0.60 0.29 0.64	0.82 0.62 0.84	0.73	585.33 (7.16)
Conclust (3=100)	β_1 = 0.52 (0.36, 0.72) β_2 = 0.98 (0.76, 1.26) β_3 = 1.95 (1.40, 2.80)	-0.26 -0.02 0.26	0.08 0.03 0.09	1.00 0.11 1.00	0.99 0.24 0.99	0.93 0.53 0.93	0.59	571.86 (7.96)
CLERE* (1=1, 2=9, 3=90)	β_1 = 0.63 (0.49, 0.80) β_2 = 0.99 (0.68, 1.45) β_3 = 1.58 (1.24, 2.02)	-0.06 -0.01 0.05	0.03 0.04 0.02	0.96 0.07 0.96	1.00 0.28 0.97	0.60 0.65 0.51	0.46	572.35 (5.77)
oCLV (3=100)	β_1 = 0.75 (0.55, 1.09) β_2 = 0.99 (0.72, 1.36) β_3 = 1.33 (0.92, 1.83)	0.12 -0.01 -0.12	0.04 0.01 0.05	0.35 0.00 0.29	0.50 0.14 0.54	0.77 0.55 0.75	0.61	587.32 (7.59)
cCLV (3G=100)	β_1 = 0.68 (0.50, 0.96) β_2 = 0.96 (0.72, 1.29) β_3 = 1.52 (1.03, 2.15)	0.03 -0.04 0.01	0.05 0.02 0.05	0.54 0.05 0.59	0.58 0.23 0.77	0.80 0.60 0.87	0.65	582.94 (7.73)

* Performance metrics only averaged for instances of correct group number specification

The relative performance of the five clustering methods in Scenario 1 were quite similar in Scenario 2, although in absolute terms there was generally a decrease in performance due to Scenario 2's increased level of noise. The results for this scenario are presented below in Table 2. Once again, cCLV saw an increase in performance relative to CLV, while oCLV did not. cCLV's power, sensitivity, and bias were superior to CLV's by a significant margin. cCLV's DIC and accuracy were also slightly lower than CLV's. MSE and specificity were nearly the same for the two methods. oCLV's relative performance was closer to CLV's in Scenario 2 as opposed to Scenario 1, however, it still underperformed in terms of power, sensitivity, specificity, accuracy, and DIC. Bias and MSE were nearly the same for the two methods. Conclust once again performed best among all methods in power, sensitivity, and specificity. Its bias was still highest among the clustering methods, although it was less than seen in Scenario 1. In a reversal of the previous scenario, its DIC was slightly higher than that of CLERE, which had the lowest DIC of all methods compared. CLERE saw a significant drop in power in this scenario, as well as a slight drop in sensitivity. CLERE's specificity remained the lowest of all methods. Finally, the tendency of CLERE to underestimate true group number increased in this higher noise scenario, failing to find three groups 17% of the time.

Table 2: Scenario 2 performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Accuracy	DIC (pD)
CLV (3=100)	β_1 = 0.73 (0.53, 1.03)	0.09	0.05	0.39	0.62	0.80	0.73	587.29 (6.54)
	β_2 = 1.00 (0.77, 1.29)	0.00	0.01	0.00	0.24	0.62		
	β_3 = 1.38 (0.97, 1.89)	-0.09	0.05	0.38	0.68	0.85		
Conclust (3=100)	β_1 = 0.54 (0.38, 0.76)	-0.21	0.08	0.96	0.96	0.87	0.65	577.33 (7.96)
	β_2 = 0.99 (0.77, 1.28)	-0.01	0.03	0.12	0.24	0.55		
	β_3 = 1.87 (1.33, 2.70)	0.22	0.08	0.96	0.96	0.89		
CLERE* (2=17, 3=83)	β_1 = 0.62 (0.46, 0.83)	-0.07	0.04	0.88	0.95	0.63	0.45	575.95 (4.96)
	β_2 = 1.02 (0.71, 1.45)	0.02	0.03	0.05	0.20	0.62		
	β_3 = 1.59 (1.21, 2.10)	0.06	0.02	0.91	0.94	0.62		
oCLV (3=100)	β_1 = 0.74 (0.52, 1.09)	0.10	0.05	0.31	0.59	0.76	0.60	589.17 (6.99)
	β_2 = 0.98 (0.74, 1.31)	-0.02	0.02	0.08	0.15	0.51		
	β_3 = 1.37 (0.94, 1.92)	-0.09	0.05	0.35	0.60	0.78		
cCLV (3=100)	β_1 = 0.68 (0.48, 0.98)	0.02	0.05	0.49	0.61	0.81	0.67	585.39 (7.39)
	β_2 = 0.97 (0.73, 1.28)	-0.03	0.02	0.02	0.22	0.58		
	β_3 = 1.52 (1.03, 2.20)	0.02	0.05	0.50	0.78	0.85		

* Performance metrics only averaged for instances of correct group number specification

Scenario 3 featured five total true groups, with outer groups β_1 and β_5 corresponding to the two high signal, distinct groups and inner groups β_2 and β_4 corresponding to the lower signal, high noise groups. Neither of our proposed CLV extensions outperformed CLV in this scenario. cCLV had a slightly better model fit and slightly lower accuracy than CLV. cCLV registered a slight increase in bias and MSE compared to CLV, and while both methods had equally excellent power for the high signal groups, CLV slightly outperformed cCLV for the lower signal inner groups. Sensitivity and specificity had mixed results, with cCLV slightly outperforming CLV in sensitivity and specificity for β_2, but slightly underperforming for all other groups. Compared to CLV, oCLV has a slightly better model fit, comparable bias and MSE, and slightly better sensitivity for the outer groups. oCLV clearly unperformed CLV in terms of power for the inner groups and had worse scores across all groups for specificity. The larger number of true groups highlighted the tendency of CLERE to underestimate group number. CLERE was only able to correctly find five groups 40% of the time, with a correspondingly low accuracy score. In the 40 instances of correct group specification, CLERE had the highest bias and MSE, second best power, and sensitivity that was competitive in both the more weakly and more highly associated groups. Once again CLERE had the worst specificity. CLERE's model fit was the lowest of all compared methods, however the difference between its DIC and that of Conclust is mostly attributed to the savings in pD afforded by the high rate of mis-specifying group number. Conclust once again performed the strongest of all compared methods. This is most clearly seen in the power and DIC metrics. Conclust had the highest power for the two inner groups, where most other methods struggled. Its DIC was also markedly lower than all other methods

except CLERE. Conclust had the second highest bias and MSE after CLERE. The largest change in performance for Conclust between the other scenarios was its sensitivity, which was the lowest of all methods in Scenario 3. Conclust's specificity, on the other hand, was competitive with the best performing methods for the inner groups, but was slightly lower for the high signal outer groups.

Table 3: Scenario 3 performance metrics of Bayesian group index regression using five different grouping methods

Method	Beta OR (95% CI)	Bias	MSE	Power	Sensitivity	Specificity	Max Accuracy	DIC (pD)
CLV (5=100)	β_1 = 0.47 (0.37, 0.60)	-0.06	0.02	1.00	0.89	0.96	0.85	534.49 (9.92)
	β_2 = 0.67 (0.48, 0.95)	0.01	0.05	0.56	0.60	0.80		
	β_3 = 0.99 (0.77, 1.25)	-0.01	0.01	0.02	0.30	0.58		
	β_4 = 1.50 (1.06, 2.08)	0.00	0.05	0.59	0.63	0.83		
	β_5 = 2.12 (1.67, 2.73)	0.06	0.02	1.00	0.87	0.95		
Conclust (5=100)	β_1 = 0.42 (0.31, 0.58)	-0.16	0.07	1.00	0.62	0.87	0.80	526.96 (11.60)
	β_2 = 0.57 (0.42, 0.77)	-0.16	0.08	0.85	0.51	0.81		
	β_3 = 0.98 (0.74, 1.28)	-0.02	0.04	0.11	0.27	0.57		
	β_4 = 1.78 (1.31, 2.45)	0.17	0.07	0.84	0.55	0.84		
	β_5 = 2.33 (1.72, 3.23)	0.15	0.06	1.00	0.62	0.87		
CLERE* (2=1, 3=25, 4=34, 5=40)	β_1 = 0.39 (0.28, 0.54)	-0.25	0.14	1.00	0.84	0.65	0.33	520.01 (7.66)
	β_2 = 0.58 (0.43, 0.79)	-0.13	0.05	0.82	0.68	0.63		
	β_3 = 0.93 (0.62, 1.41)	-0.07	0.07	0.25	0.12	0.60		
	β_4 = 1.64 (1.16, 2.35)	0.09	0.06	0.72	0.50	0.63		
	β_5 = 2.70 (1.90, 3.90)	0.30	0.18	1.00	0.79	0.67		
oCLV (5=100)	β_1 = 0.48 (0.37, 0.63)	-0.03	0.02	1.00	0.93	0.82	0.48	533.85 (8.17)
	β_2 = 0.74 (0.52, 1.05)	0.11	0.06	0.37	0.55	0.74		
	β_3 = 1.00 (0.72, 1.38)	0.00	0.02	0.07	0.07	0.61		
	β_4 = 1.34 (0.96, 1.87)	-0.11	0.06	0.36	0.48	0.72		
	β_5 = 2.06 (1.57, 2.74)	0.03	0.02	0.99	0.91	0.82		
cCLV (5G=100)	β_1 = 0.46 (0.36, 0.59)	-0.07	0.02	1.00	0.87	0.96	0.81	532.98 (9.86)
	β_2 = 0.65 (0.45, 0.96)	-0.03	0.05	0.52	0.64	0.83		
	β_3 = 0.97 (0.73, 1.26)	-0.03	0.02	0.09	0.27	0.59		
	β_4 = 1.55 (1.07, 2.21)	0.04	0.06	0.56	0.53	0.79		
	β_5 = 2.17 (1.68, 2.86)	0.08	0.03	1.00	0.82	0.93		

* Performance metrics only averaged for instances of correct group number specification

Application to NCI-SEER NHL case-control study data

The results of our simulation study indicate that the Conclust grouping method is best suited for grouping environmental chemical exposure variables before Bayesian group index regression. We applied the two-step process of clustering followed by Bayesian group index regression to the NHL dataset, with a separate analysis performed for each study center subset.

We limit our results presented to the Iowa and LA subsets, as these were the only two study centers where significant group index associations were found. The results for the remaining two subset analyses can be found in

the supplemental materials. For the Iowa subset, among the varying group number models run, the models with the lowest DIC were the 2-group model (DIC = 454.5), the 3-group model (DIC = 452.4), and the 6-group model (DIC = 455.4). With no model clearly superior measured by DIC, we chose the 6-group model as this clustering arrangement found more results of interest. We characterize these six clusters and list their chemicals as follows: a singleton group composed solely of 2,4-D called Group 1, a group of pesticides named Group 2 (*cis*-permethrin and *trans*-permethrin), a singleton group composed solely of pentachlorophenol called Group 3, a group of pesticides named Group 4 (carbaryl, chlorpyrifos, DDT, methoxychlor, and *o*-phenylphenol), a group PAHs of called Group 5 (benz(a)anthracene, benzo(b)fluoranthene, benzo(a)pyrene, chrysene, and indeno(1,2,3-cd)pyrene), and a group consisting of pesticides, PAHs, and PCBs named Group 6 (dicamba, benzo(k)fluoranthene, dibenz(ah)anthracene, PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, α-chlordane, γ-chlordane, DDE, diazinon, and propoxur). The odds ratios and 95% CIs estimated for our 6 index effects and covariates are in Table 4. Two indices were significantly associated with NHL: Group 1 had an inverse association (OR = 0.67, 95% CI: 0.54, 0.84) and Group 6 had a positive association (OR = 1.82, 95% CI: 1.04, 3.29). No covariates were found to be significantly associated with NHL risk. The index effect estimate of Group 1, being a singleton group, can be attributed entirely to 2,4-D. Propoxur (w = 0.28), DDE (w = .10), γ-chlordane (w = 0.10), and α-chlordane (w = 0.08) were the most heavily weighted chemicals in Group 6.

Table 4: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Iowa

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	**0.67**	**0.54**	**0.84**
Group 2	0.95	0.76	1.16
Group 3	1.05	0.85	1.30
Group 4	0.96	0.64	1.40
Group 5	0.87	0.61	1.11
Group 6	**1.82**	**1.04**	**3.29**
Male	1.06	0.69	1.64
White	1.13	0.08	8.88
Education	1.08	0.71	1.66
Age	0.98	0.96	1.01

For the LA subset, 6-group model had a DIC of 405.3, more than 5 points lower than any competing model. We characterize these six clusters and list their chemicals as follows: a group of pesticides and PAHs called Group 1 (2,4-D, benzo(b)fluoranthene, chrysene, α-chlordane, γ-chlordane, DDT, *o*-phenylphenol, and pentachlorophenol), a singleton group composed solely of *trans*-permethrin called Group 2, a singleton group composed solely of *cis*-permethrin called Group 3, a singleton group composed solely of carbaryl called Group 4, a group of pesticides named Group 5 (chlorpyrifos, diazinon, and propoxur), and a group of PCBs, pesticides, and PAHs called Group 6 (dicamba, benz(a)anthracene, benzo(k)fluoranthene, benzo(a)pyrene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene, PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, DDE, and methoxychlor). The odds ratios and 95% CIs estimated for our 6 index effects and covariates are in Table 5. Group 5 was found to have a significant and negative association with NHL (OR = 0.69, 95% CI: 0.49, 0.97). None of the covariates were found to be significantly associated with NHL risk. The Group 5 index was dominated by diazinon, with a weight of 0.45.

Table 5: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in LA

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	1.02	0.65	1.62
Group 2	0.71	0.33	1.36
Group 3	1.46	0.77	3.15
Group 4	1.05	0.84	1.33
Group 5	**0.69**	**0.49**	**0.97**
Group 6	1.17	0.71	1.91
Male	0.90	0.56	1.41
White	1.17	0.70	2.03
Education	1.13	0.78	1.67
Age	1.00	0.98	1.02

Discussion

In this paper, we proposed two semi-supervised extensions to the unsupervised clustering algorithm CLV in order to improve groups defined for subsequent Bayesian group index regression and to discourage the grouping of chemicals with opposite directions of association with the target outcome variable. Our first extension, cCLV,

incorporates constraints that penalize chemical variable pairings deemed undesirable by the user. Our second extension, oCLV, determines the subset of chemical variables most associated with the outcome variable and focuses clustering on this group. We compared the performance of these two methods with two other semi-supervised clustering methods and one unsupervised method in partitioning a set of chemicals into the groups required for Bayesian group index regression. We designed a simulation study consisting of three scenarios: three true groups with moderate noise, three true groups with high noise, and five true groups with high noise. In each of these scenarios two groups were simulated such that the chemicals comprising one group had equal correlations with the chemicals of another group. The distinguishing factor between these groups was their opposite association with the simulated outcome variable. We hypothesized that by incorporating information from the outcome variable during clustering, semi-supervised clustering methods would outperform unsupervised methods, particularly in the task of preventing chemicals with opposite directions of association with an outcome from being grouped together.

Looking at the results of our simulation study, particularly the high noise pairs in each scenario, one notable difference between the clustering methods compared was the direction of the bias for index effects. Both the CLV and oCLV methods consistently estimated index effects that were biased towards the null, indicating that they regularly combined variables with opposing outcome associations into the same cluster. This is not surprising for the CLV method, as it is unsupervised. oCLV's behavior in this regard can be attributed to the fact that this method of supervision is primarily focused on the magnitude, not the direction of association with the outcome. Conversely, Conclust and CLERE consistently estimated index effects that were biased away from the null. Our proposed method cCLV also generally estimated index effects biased away from the null, with two exceptions in the β_1 group of Scenarios 1 and 2. Even in these two instances of bias towards the null, the bias was less than that of the CLV method.

The direction of bias for these high noise pairs helps explain the overall performance of the compared clustering methods. CLV and oCLV consistently underperformed in terms of power and DIC, and were often the lowest performers in sensitivity and specificity due to their inability to separate oppositely attracted variables. The three remaining semi-supervised methods saw stronger performance, albeit with some variation between them. The constraints implemented in cCLV saw an improvement over CLV in Scenarios 1 and 2, especially in power and sensitivity, while in Scenario 3 the two were roughly equivalent. CLERE consistently had good power and sensitivity

with relatively moderate bias and MSE, although it also consistently performed worst as measured by specificity. These averages only applied to the instances where CLERE returned the desired number of clusters, however, a weakness that appeared to compound in the face of higher noise and group number. Finally, Conclust was consistently best as measured by power, and was often the best clustering method in terms of sensitivity, specificity, and DIC. An apparent weakness of the method was the tendency to overestimate index effects, as evidenced by its relatively high bias and MSE. This overestimation stems from taking variables belonging to the null index that have a weak positive association with the outcome and adding them to significantly associated positive groups. This tendency might in fact be a strength when applied to real data, as our desire is to group empirically similar chemicals that also share the same direction of association with an outcome. The high bias and MSE are artifacts of our simulation's strict definition of "true" groups, whereas in real application such definitions do not exist. Based on these findings, we recommend the Conclust method for clustering variables prior to Bayesian group index regression. We applied Conclust to cluster chemical exposure variables in the NHL study dataset, individually for each of the study centers. In the Iowa subset of the Conclust analysis, we saw a significant departure from the traditional chemical structure and usage groupings. This largely entailed the breaking up of the heterogeneous pesticide category of chemicals, both into small one or two chemical groups or by folding pesticides into larger indices. The PAH category of chemical was also split, with most PAHs forming their own group (Group 5), and two PAHs added to the dominant positively associated index (Group 6). While a number of small or singleton groups were formed around chemicals with a negative association with the outcome, the majority of positively associated chemicals were clustered together in Group 6, which was found to have a positive and significant association with NHL (OR = 1.82), with propoxur (w = 0.28), DDE (w = .10), γ-chlordane (w = 0.10), and α-chlordane (w = 0.08) having mean posterior weights above an equal share in the index. A negative and significant association was also found between Group 1 (OR = 0.67) and NHL, with all the index weight attributed to the index's sole chemical 2,4-D.

In our previous work on unsupervised, empirical clustering of chemical variables before Bayesian group index regression, both positive and negative significant indices were found in the Iowa subset, dominated by propoxur and 2,4-D, respectively. The positive index had fewer chemicals, with fewer pesticides and no PAHs or PCBs, and had a smaller effect estimate (OR = 1.50), while the negative index was comprised of more chemicals and had a larger effect estimate (OR = 0.58) [17]. The differences in the two analyses' positive indices is an example of Conclust's

tendency to greedily aggregate variables of the same direction of association seen in the simulation study, resulting in larger effect estimates. In this instance we can see the benefit of this behavior, as Group 6 represents a sub-mixture consistent enough to be empirically grouped that encompasses most of the positive signal found in the overall mixture, all while avoiding bias towards the null from the accidental inclusion of negatively associated variables. Additionally, two chemicals of interest, γ-chlordane and α-chlordane, were identified, whereas in our previous analysis they were grouped separately in a non-significant index and therefore overlooked. The differences in the negative indices, on the other hand, show that the tendency to aggregate chemicals with the same direction of association is not absolute, and that the empirical similarity of chemicals is also important to the assignment of cluster labels. That 2,4-D was assigned as a singleton group is not surprising, as it heavily dominated the index formed by CLV in our previous analysis, with a mean posterior weight of 0.69.

The findings of this most recent analysis are also consistent with various other previous analyses of the NHL data. In a spatial Bayesian group index regression analysis of the NHL data, the indices were determined based on chemical structure and usage, with indices for PCBs, PAHs, and two pesticide indices to separate chemicals with opposing direction of association with the outcome. Both pesticide indices were found to be significant in the Iowa subset. The chemical found to be most important in the positively associated index was propoxur, follow by DDE, γ-chlordane, and α-chlordane. The most important chemical in the negatively associated index was 2,4-D [10]. In a single-index analysis of the NCI-SEER NHL dataset, a significant, positive association was found between all 27 chemicals and NHL in the Iowa subset, with propoxur, DDE, and γ-chlordane the highest weighted chemicals [35]. Single-chemical regression analyses between heavily-weighted chemicals in our significant indices and NHL also support our findings, where significant associations were found for propoxur, DDE, γ-chlordane, α-chlordane, and 2,4-D [33,35]. 2,4-D, a chemical classed as possibly carcinogenic to humans by the International Agency for Research on Cancer working group [36], has a strong negative signal in the Iowa subset of this dataset, which fits into a history of inconclusive investigations into the relationship between 2,4-D and NHL [37-40].

In the LA subset of our Conclust analysis we saw a similar breakdown of indices as with the Iowa subset, marked by a splitting of both pesticides and PAHs into separate indices, a number of singleton indices made of pesticides, and a large index composed of PCBs, PAHs, and pesticides. Group 5 was the only significant index, with a negative association with NHL (OR = 0.69). The index was dominated by diazinon, with a mean posterior weight of 0.45. This

result differs from our previous Bayesian group index analysis using CLV clustering, where no chemical exposure indices were found to be significant outside of the Iowa subset. While there was no significant index in the LA subset, the index that included diazinon (along with α-chlordane, γ-chlordane, and chlorpyrifos) had a nominally negative association with NHL and was dominated by diazinon with a weight of 0.43 [17]. In a previous spatial Bayesian group index analysis of the NHL data, diazinon was also the most highly weighted chemical in the LA subset negative association pesticide index, although the index effect itself was not found to be significant [10]. In a single-index regression analysis of the NHL data, the index for the LA subset was positive, so negative signal chemicals such as diazinon were given near-zero weight. Single-chemical regression of diazinon on NHL in this same study resulted in a negative association that was not statistically significant [35]. From these analyses we can see that there is a consistent signal of negative association between diazinon and NHL in the LA subset for this data. Our novel significant finding can be attributed to a previously untried combination of chemicals clustered in the same index. Diazinon has been classified as Group 2A, or probably carcinogenic to humans, by the International Agency for Research on Cancer working group [41].

We found no significant associations between any chemical index and NHL in the Detroit and Seattle subsets. This is consistent with two previous group index regression analyses performed on this dataset [10,17]. These results disagree with another group index analysis of the NHL data that used a two-step, frequentist approach and found a significant and positive association between PCBs and NHL [42]. Evidence for a positive association between PCBs and NHL is further supported by analyses of individual and total PCBs [34] and by an analysis of study participant's blood plasma [43]. This seeming discrepancy is attenuated by the fact that in our analysis of the Iowa subset the significant positive index included all PCB exposure variables in the study. While their inclusion in the index may have contributed to the larger index effect found compared to previous analyses, none of the PCBs were assigned weights indicating they were the largest contributors to the overall index effect.

As demonstrated in simulation and real data application, the two-step combination of Conclust and Bayesian group index regression has several strengths. There are limitations, however, that may motivate future work towards improving various aspects of our approach. Our simulation study assumed that the true group number was known, whereas the number of indices that should be modelled would be unknown in a research application. Our solution was to compare the model fit of various models with different group numbers. This process requires fitting many

models at the cost of time and computational resources. Future work on the estimation of group number would be valuable for the group index modelling approach. A second weakness of our current approach is the recourse to a two-step process. As our eventual group index regression involves the outcome variable of interest, it would be ideal to combine clustering informed by the outcome and the subsequent group index regression into a single model.

In conclusion, the incorporation of outcome variable information through semi-supervised clustering improves the chemical groups defined for Bayesian group index models. More specifically, constrained clustering as implemented in the Conclust algorithm allows for the partitioning of a chemical mixture that simultaneously maximizes the similarity of chemicals grouped while discouraging clusters containing chemicals with opposite direction of association with a target outcome, thus avoiding index effect estimates that are biased towards the null. The Conclust algorithm demonstrated the ability to separate oppositely associated variables into distinct clusters in both simulation and real data applications. As defining group composition of a chemical mixture is an essential step before performing Bayesian group index regression, this work informs practitioners on how best to empirically partition chemical mixtures without relying on assumptions inferred from chemical structure or usage.

References

1. Carson R, Darling L, Darling L. Silent Spring. Houghton Mifflin Company; 1962.

2. Wang Z, Walker GW, Muir DCG, Nagatani-Yoshida K. Toward a Global Understanding of Chemical Pollution: A First Comprehensive Analysis of National and Regional Chemical Inventories. Environmental science & technology. 2020;54(5):2575-2584. doi:10.1021/acs.est.9b06379

3. Bobb, Jennifer F, Linda Valeri, Birgit Claus Henn, David C Christiani, Robert O Wright, Maitreyi Mazumdar, John J Godleski, and Brent A Coull. "Bayesian Kernel Machine Regression for Estimating the Health Effects of Multi-Pollutant Mixtures." Biostatistics (Oxford, England) 16, no. 3 (2015): 493–508. https://doi.org/10.1093/biostatistics/kxu058.

4. Keil, Alexander P, Jessie P Buckley, Katie M O'Brien, Kelly K Ferguson, Shanshan Zhao, and Alexandra J White. "A Quantile-Based g-Computation Approach to Addressing the Effects of Exposure Mixtures." Environmental Health Perspectives 128, no. 4 (2020): 47004–. https://doi.org/10.1289/EHP5838.

5. 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(27):3947-3967. doi:10.1002/sim.8701

6. DeVille NV, Khalili R, Levy JI, Korrick SA, Vieira VM. Prenatal environmental exposures and associations with teen births. Journal of exposure science & environmental epidemiology. 2021;31(2):197-210. doi:10.1038/s41370-020-00262-9

7. Czarnota, Jenna, Chris Gennings, and David C. Wheeler. "Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk." Cancer informatics 14 (2015): CIN-S17295.

8. Wheeler, David C, Salem Rustom, Matthew Carli, Todd P Whitehead, Mary H Ward, and Catherine Metayer. "Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk." International Journal of Environmental Research and Public Health 18, no. 2 (2021a): 504–. https://doi.org/10.3390/ijerph18020504.

9. Wheeler, David C, Salem Rustom, Matthew Carli, Todd P Whitehead, Mary H Ward, and Catherine Metayer. "Bayesian Group Index Regression for Modeling Chemical Mixtures and Cancer Risk." International Journal of Environmental Research and Public Health 18, no. 7 (2021b): 3486–. https://doi.org/10.3390/ijerph18073486.

10. Boyle J, Ward MH, Cerhan JR, Rothman N, Wheeler DC. Estimating mixture effects and cumulative spatial risk over time simultaneously using a Bayesian index low-rank kriging multiple membership model. Statistics in medicine. 2022;41(29):5679-5697. doi:10.1002/sim.9587

11. Fung, Glenn. "A Comprehensive Overview of Basic Clustering Algorithms." (2001) https://sites.cs.ucsb.edu/~veronika/MAE/clustering_overview_2001.pdf

12. Witten, I. H. (Ian H.), and Eibe. Frank. Data Mining : Practical Machine Learning Tools and Techniques . 2nd ed. Amsterdam ;: Morgan Kaufman, 2005. Print.

13. Ghahramani Z. (2004) Unsupervised Learning. In: Bousquet O., von Luxburg U., Rätsch G. (eds) Advanced Lectures on Machine Learning. ML 2003. Lecture Notes in Computer Science, vol 3176. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-28650-9_5

14. Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). The elements of statistical learning: data mining, inference, and prediction (Vol. 2, pp. 1-758). New York: springer.

15. Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. Machine learning, 109(2), 373-440.

16. Bair, Eric. "Semi-supervised clustering methods." Wiley Interdisciplinary Reviews: Computational Statistics 5.5 (2013): 349-361.

17. Carli M, Ward HW, Cerhan JR, Rothman N, Wheeler CW. Comparison of Variable Clustering Methods in the Context of Group Index Regression. (2023). Manuscript submitted for publication.

18. Chen M, Vigneau E. Supervised clustering of variables. Advances in data analysis and classification. 2016;10(1):85-101. doi:10.1007/s11634-014-0191-5

19. Vigneau E. Clustering of variables for enhanced interpretability of predictive models. Informatica (Ljubljana). 2021;45(4). doi:10.31449/inf.v45i4.3283

20. Vigneau, Evelyne, and E. M. Qannari. "Clustering of variables around latent components." *Communications in Statistics-Simulation and Computation* 32.4 (2003): 1131-1150.

21. Bilenko M, Basu S, Mooney R. Integrating constraints and metric learning in semi-supervised clustering. In: Proceedings of the Twenty-First International Conference on Machine Learning. ACM; 2004:11-. doi:10.1145/1015330.1015360

22. Bair E, Tibshirani R. Semi-supervised methods to predict patient survival from gene expression data. PLoS Biol 2004, 2:e108. doi: 10.1371/journal.pbio.0020108.

23. Burke EK, Kendall G. Tabu Search. In: Search Methodologies. Springer; 2013:243-263. doi:10.1007/978-1-4614-6940-7_9

24. Hiep, T. K., Duc, N. M., & Trung, B. Q. (2016, December). Local search approach for the pairwise constrained clustering problem. In Proceedings of the 7th Symposium on Information and Communication Technology (pp. 115-122).

25. Yengo, Loïc, Julien Jacques, and Christophe Biernacki. "Variable clustering in high dimensional linear regression models." *Journal de la Société Française de Statistique* 155.2 (2014): 38-56.

26. Yengo, Loïc, Julien Jacques, and Christophe Biernacki. "VARIABLE CLUSTERING IN HIGH DIMENSIONAL PROBIT REGRESSION."

27. Spiegelhalter DJ, Best NG, Carlin BP, van der Linde A. Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2002;64:583-639. doi:10.1111/1467-9868.00353

28. Plummer M. Penalized loss functions for Bayesian model comparison. Biostatistics. 2008;9:523-539. doi:10.1093/biostatistics/kxm049

29. Wheeler D, Carli M. BayesGWQS: Bayesian Grouped Weighted Quantile Sum Regression. Published online 2020.

30. Plummer M. JAGS: A Program for Analysis of Bayesian Graphical Models using Gibbs Sampling. In: *3rd International Workshop on Distributed Statistical Computing*. ; 2003:124-124.

31. Chatterjee, Nilanjan, Patricia Hartge, James R Cerhan, Wendy Cozen, Scott Davis, Naoko Ishibe, Joanne Colt, Lynn Goldin, and Richard K Severson. "Risk of Non-Hodgkin's Lymphoma and Family History of Lymphatic, Hematologic, and Other Cancers." Cancer Epidemiology, Biomarkers & Prevention 13, no. 9 (2004): 1415–21. https://doi.org/10.1158/1055-9965.1415.13.9.

32. Morton, Lindsay M., Sophia S. Wang, Wendy Cozen, Martha S. Linet, Nilanjan Chatterjee, Scott Davis, Richard K. Severson, et al. "Etiologic Heterogeneity Among Non-Hodgkin Lymphoma Subtypes." Blood 112, no. 13 (2008): 5150–60. https://doi.org/10.1182/blood-2008-01-133587.

33. Colt, Joanne S, Jay Lubin, David Camann, Scott Davis, James Cerhan, Richard K Severson, Wendy Cozen, and Patricia Hartge. "Comparison of Pesticide Levels in Carpet Dust and Self-Reported Pest Treatment Practices in Four US Sites." Journal of Exposure Analysis and Environmental Epidemiology 14, no. 1 (2004): 74–83. https://doi.org/10.1038/sj.jea.7500307.

34. Colt, Joanne S, Richard K Severson, Jay Lubin, Nat Rothman, David Camann, Scott Davis, James R Cerhan, Wendy Cozen, and Patricia Hartge. "Organochlorines in Carpet Dust and Non-Hodgkin Lymphoma." Epidemiology (Cambridge, Mass.) 16, no. 4 (2005): 516–25. https://doi.org/10.1097/01.ede.0000164811.25760.f1.

35. Czarnota J, Gennings C, Colt JS, et al. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study. *Environmental health perspectives*. 2015;123(10):965-965. doi:10.1289/ehp.1408630

36. IARC. Agents Classified by the IARC Monographs, Volumes 1–129. Lyon, France: International Agency for Research on Cancer. 2021. Accessed May 25, 2023. http://monographs.iarc.fr/ENG/Classification/index.php

37. IARC Working Group on the Evaluation of Carcinogenic Risks to Humans. DDT, Lindane, and 2, 4-D. 2018.

38. Ward MH, Lubin J, Giglierano J, et al. Proximity to Crops and Residential Exposure to Agricultural Herbicides in Iowa. *Environmental health perspectives*. 2006;114(6):893-897. doi:10.1289/ehp.8770

39. De Roos AJ, Fritschi L, Ward MH, et al. Herbicide use in farming and other jobs in relation to non-Hodgkin's lymphoma (NHL) risk. *Occupational and environmental medicine (London, England)*. 2022;79(12):795-806. doi:10.1136/oemed-2022-108371

40. Goodman JE, Loftus CT, Zu K. 2,4-Dichlorophenoxyacetic acid and non-Hodgkin's lymphoma, gastric cancer, and prostate cancer: meta-analyses of the published literature. *Annals of epidemiology*. 2015;25(8):626-636.e4. doi:10.1016/j.annepidem.2015.04.002

41. International Agency for Research on Cancer Volume 112: Some organophosphate insecticides and herbicides: tetrachlorvinphos, parathion, malathion, diazinon and glyphosate. IARC Working Group. Lyon; 3–10 March 2015. IARC Monogr Eval Carcinog Risk Chem Hum

42. Wheeler DC, Czarnota J. Modeling chemical mixture effects with grouped weighted quantile sum regression. *ISEE Conference Abstracts*. 2016.

43. De Roos AJ, Hartge P, Rothman N, et al. Persistent Organochlorine Chemicals in Plasma and Risk of Non-Hodgkin's Lymphoma. *Cancer research (Chicago, Ill).* 2005;65(23):11214-11226. doi:10.1158/0008-5472.CAN-05-1755

Chapter 5: Conclusion

The work presented in the preceding chapters offers solutions to several problems encountered when working with Bayesian group index regression. We extended the model to simultaneously impute BDL missing data, provided novel methods for the grouping of chemicals before group index regression, and conducted simulation studies to identify the strongest performing candidates for these tasks. In this chapter, we summarize the findings of the previous chapters, discuss the implications of this research, and consider some remaining questions to motivate future research.

Research Summary

In Chapter 1, we began with a review of environmental chemical pollution, its potential for negatively impacting human health, and the desire of researchers to better quantify the risk these chemicals pose. We then discussed how human exposure to chemicals is better understood as the simultaneous exposure to a mixture of chemicals, as opposed to discrete exposures to individual chemicals, and the unique statistical challenges this model of exposure pose. The most immediate challenge is that traditional regression models are ill-suited to analyzing chemical mixtures due to the strong correlations between individual chemicals. We reviewed the work that has been done to overcome this aspect of chemical mixture analysis, focusing particularly on single and group index regression models. We then identified two problems commonly encountered when performing Bayesian group index regression: the presence of BDL missing data and how best to partition a chemical mixture into the groups required of the model. Reviewing the literature on BDL imputation, we noted the theoretical support for multiple imputation methods and their ability to be implemented in the Bayesian framework. We then discussed how partitioning chemical mixtures has traditionally been done on the basis of subject matter knowledge of chemicals, and hypothesized that an empirical basis for clustering chemicals could detect previously unknown patterns in the chemical mixture while avoiding the clustering of chemicals that would negatively impact model fit and index parameter estimates. In the wide field of clustering algorithms, we focused on hard variable clustering methods and semi-supervised clustering methods as being particularly well-suited to the chemical partitioning problem. In the following chapters we offered novel solutions to these analytical challenges.

In Chapter 2, we looked at the near ubiquitous presence of BDL missing data in chemical mixture data, and some of the various proposals for their imputation. Among these, Bayesian condition univariate imputation methods seemed promising, as they could be combined with Bayesian group index regression, would readily accommodate the truncated distributions needs to model BDL observations, and could account for the true variation of unknown imputed values. We hypothesized that the combination of an imputation model and the group index analysis model would result in superior parameter estimates. We incorporated two conditional univariate imputation methods with Bayesian group index regression: Pseudo-Gibbs and Sequential Full Bayes (SFB). We compared these extensions to Bayesian group index regression with a single imputation method and the well-known multiple imputation by chained equations (MICE) algorithm. We evaluated how these methods compared in terms of mean squared error (MSE), bias, power, sensitivity, specificity, DIC, and computation time in a simulation study. We found that the Pseudo-Gibbs imputation method outperformed the other methods at high levels of BDL missingness (70%), but that at lower percentages performance was similar for all methods compared. Based on this, we recommended the computationally efficient SI method for lower levels of BDL missing data, and the Pseudo-Gibbs method for higher levels. We applied Pseudo-Gibbs imputation to the California Childhood Leukemia Study (CCLS), as many of its chemical's exhibit BDL missingness of up to 50%, with a few reaching beyond 70%. We found a positive, significant association between the polycyclic aromatic hydrocarbon (PAH) index and childhood leukemia, with benzo(k)fluoranthene and indeno(1,2,3 -c,d)pyrene identified as important chemicals within the index. We then identified the income covariate as a likely effect modifier, and ran an analysis stratified by income. In the high income strata, we found three significant indices: the PCB and herbicide indices with positive associations and the metals index with a negative association. The herbicide dacthal, PCB 138, PCB 180, and the metal arsenic were identified as important chemicals.

In Chapter 3, we proposed a novel variable clustering algorithm that incorporated a variant of PCA, robust PCA (RPCA), that was modified for use with chemical mixture data. We compared this with three other variable clustering methods and a subject clustering method. We evaluated these five clustering methods in a simulation study, using the clustering algorithms to derive group assignments for chemicals that were then used in subsequent Bayesian group index regressions. The clustering assignments themselves were judged by their accuracy, while the quality of the following group index parameter estimates and model were measured by bias, MSE, power, sensitivity,

specificity, and DIC. We found that the Clustering of Variables around Latent Variables (CLV) and agglomerative hierarchical clustering (AHC) methods performed best, with CLV slightly outperforming AHC in the high group number scenario where there was moderate correlation overlap between two chemical groups with opposite direction of association with the outcome variable. We then applied CLV clustering in tandem with Bayesian group index regression to the National Cancer Institute (NCI) Surveillance, Epidemiology, and End Results (SEER) non-Hodgkin Lymphoma (NHL) case-control study, where we fit four separate models for the four study centers. In the Iowa subset, we fit a five-index model that included a PCB and pesticide index, a PAH index, and three pesticide indices. Two pesticide groups were found to be significantly associated with NHL: Group 1 (2,4-D, chlorpyrifos, *cis*-permethrin, and *trans*-permethrin) and Group 2 (dicamba, DDE, DDT, and propoxur). Group 1 was negatively associated with NHL, with 2, 4-D identified as an important chemical. Group 2 was positively associated with NHL, with propoxur and DDE identified as important chemicals. The grouping arrangement determined by CLV highlighted the general reasonableness of grouping based on chemical structure and usage, while also demonstrating that such grouping assignments could be replicated with empirical methods while also providing a rationale for the partitioning of heterogeneous chemical classes such as pesticides.

In Chapter 4, we proposed to extend the CLV clustering method identified in Chapter 3 to incorporate information from the outcome variable during clustering. We hypothesize that such a semi-supervised clustering algorithm would generate better chemical grouping assignments for subsequent Bayesian group index regression. Specifically, we sought to discourage clusters that combined chemicals with opposite directions of association with the outcome variable. Our first semi-supervised CLV method, constrained CLV (cCLV), extended CLV to allow for the definition of constraints by the user. In our application, the constraints took the form of cannot-link pairs that penalized any proposed groups that contained two chemicals with univariate associations of opposite direction with the outcome. Our second proposed extension, outcome-adjusted CLV (oCLV), adapted a “supervised clustering” algorithm that focuses clustering on only the most highly associated chemicals. We compared these two methods with the unsupervised CLV and two other semi-supervised methods: Constrained Clustering by Tabu Search (Conclust) and Clusterwise Effect Regression (CLERE). We evaluated how these methods compared in terms of mean squared error (MSE), bias, power, sensitivity, specificity, DIC, and computation time in a simulation study. We found that the Conclust method performed best, with consistently strong performance in terms of power, of sensitivity, specificity,

and DIC, and a tendency to greedily incorporate chemicals with the same direction of association with the outcome into the same group. We then applied the Conclust algorithm along with Bayesian group index regression to the NCI-SEER NHL case-control study, again fitting four separate models for the four study centers. In the LA subset we found one significant index, called Group 5 (chlorpyrifos, diazinon, and propoxur). It was negatively associated with NHL, with diazinon identified as an important chemical. In the Iowa subset, we fit a six-index model. In a departure from the analysis of this subset in Chapter 3, patterns similar to grouping based on chemical structure and use were not replicated, with both the pesticide and PAH classes of chemicals split into multiple and overlapping groups. Two significant indices were found: Group 1 (a singleton group of 2, 4-D alone), and Group 6 (dicamba, benzo(k)fluoranthene, dibenz(ah)anthracene, PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, α-chlordane, γ-chlordane, DDE, diazinon, and propoxur). Group 1 had a negative association, with index estimate attributed solely to 2, 4-D. Group 6 had a positive association, with propoxur, DDE, γ-chlordane, and α-chlordane identified as important chemicals. The chemicals of Group 6, a majority of the positively associated chemicals in the mixture, showcased Conclust's tendency to greedily combine chemicals of the same direction of association with the outcome. This was an improvement over our previous unsupervised clustering, as two additional chemicals of interest were identified. Overall, our findings demonstrate that supervision of chemical clustering with information from the outcome variable is an improvement over unsupervised methods in our application to group index models.

Implications

The results of our data applications to the CCLS and NCI-SEER NHL studies are supported by a number of previous analyses, and support further investigation of these associations. For instance, in Chapter 2 our non-stratified model found the PAH index positively associated with childhood leukemia, with benzo(k)fluoranthene and indeno(1,2,3 - c,d)pyrene identified as important chemicals. These two PAHs had previously been found to be either significantly or borderline significantly associated with childhood leukemia in single-chemical analyses of the CCLS data [1]. In the high income strata of our stratified model, the PCB index was found to be positively associated and significant, with PCB 138 as the highest contributor to the overall index effect. PCB 138 and summed PCBs were previously found to have a positive association with childhood leukemia in a logistic regression analysis [2]. Additionally, the significant and positively associated herbicide index, along with dacthal as the predominant chemical in the index, is very similar to previous group index analyses [3-4] that used different imputation methods, as well as to single-chemical logistic

regressions [5]. Our analysis of CCLS provides additional support for these previous findings, and could motivate additional investigations into the relationship between the chemicals identified and childhood leukemia.

In our data applications to the NCI-SEER NHL study in Chapters 3 and 4, we consistently found a significant positive index in the Iowa subset. The chemicals identified as important in the analyses, the pesticides propoxur, DDE, γ-chlordane, and α-chlordane, have also been found significantly associated with NHL in past analyses. In a single-index regression of the Iowa subset, the index was found to be significant and positively associated, with the chemicals propoxur, DDE, and γ-chlordane assigned the highest weights [6]. Single-chemical regression analyses also found significant positive associations for propoxur, DDE, γ-chlordane, and α-chlordane [6-7]. Once again our results support previous findings, and could motivate further investigations and discourage agricultural use of the pesticides identified.

Future Work

The research presented in the preceding chapters contributes and supports the use of several new tools for Bayesian group index regression. The methods introduced largely focused on the context of chemical mixture analysis, and in the case of our imputation extension attempted to make full use of the flexibility of Bayesian modelling. Improvements could be made, however, by widening the application of Bayesian group index regression to other types of data and by expanding the utilization of Bayesian methods. These improvements would require the development of new methods, which we detail below.

First, the application of Bayesian group index regression models could be expanded from our focus on chemical mixture data to the wider exposome. The exposome is defined as the total of exposures to which an individual is subjected to from birth to death, including internal processes regulated by gene expression and metabolism, external exposure such as pollution, radiation, and diet, and the influence of larger forces that shape the individual's place in the world such as social capital or economic status [8]. Some of the data sources in this list, such as gene expression [9], proteomic [10], and metabolomic data [11], are extreme cases of the "curse of dimensionality", where the number of variables greatly exceeds the number of subjects. Other data types, such as diet or measures of phycological well-being, are likely to be measured by some sort of ordinal scale. While this description of challenges encountered when trying to comprehensively model the variety of data types included in the exposome is not all-

encompassing, it is illustrative of the work that could be done both to characterize Bayesian group index regression's utility in such analyses and to contribute extensions to improve its performance.

Second, in Chapters 3 and 4 we identified clustering algorithms that were well-suited as data preparation algorithms before Bayesian group index regression. This two-step process is limited in that it requires complete data with no missing observations, BDL or otherwise, before clustering can occur. In order to make full use of the imputation extension proposed in Chapter 2, it would be ideal to incorporate the empirical clustering step into the larger Bayesian estimation algorithm. A related problem, that of determining the number of groups to fit in a group index model, is one that we did not address in our work. Knowledge of true group number was assumed in our simulations, and in Chapters 3 and 4 group number in data applications was decided by recommendations from the authors of the clustering methods applied. In Chapter 3 we evaluated a variable clustering method, Dirichlet Process Variable Clustering (DPVC), that did estimate group number. It did not perform well in the context of background correlation normally found in chemical mixture data, however. A Bayesian variable clustering method that does not exhibit this limitation would be a good candidate for combination with Bayesian group index regression.

Finally, an increase in the utilization of the flexibility of Bayesian modeling would likely require an improvement in the computational efficiency of our application of Bayesian group index regression. In Chapter 2, we found that our Pseudo-Gibbs imputation extension outperformed other imputation methods when BDL missingness was high, around 70%. Otherwise, it was much slower than single-imputation methods that performed similarly. This is not an ideal situation, as high levels of BDL missingness require the inclusion of a great number of parameters to the model, slowing time to convergence of the posterior distribution. Even in just this scenario, a faster algorithm would be of significant utility. The value of greater computational efficiency would only be compounded if further tasks, such as a great increase in the number of variables modeled, the estimation of clusters, and the estimation of group number, were considered.

Conclusion

We offer several solutions to problems commonly encountered when performing Bayesian group index regression. This was accomplished in the following ways. First, we extended Bayesian group index regression to simultaneously impute missing BDL observations in such a way that incorporates the variation of the unknown status of BDLs.

Second, we proposed a novel variable clustering algorithm and, from a selection of variable clustering algorithms, identified the one most suitable for use with Bayesian group index regression. This enables the definition of the chemical groups necessary for group index regression without relying on assumptions drawn from similarity of chemical's structure or use. Third, we extended the variable clustering algorithm previously identified to incorporate information from the outcome variable of interest. We compared these extensions with other semi-supervised clustering methods and identified the one most suitable for use with Bayesian group index regression. This offers superior chemical clusters to unsupervised methods, and discourages the grouping of chemical that may artificially bias indices to the null.

The methods we present share an emphasis on limiting analytical assumptions and focusing on the empirical realities of the data analyzed. Our proposed imputation method eschews simplistic and convenient replacements for the missing BDL observations that have been shown to lead to poor parameter estimates, and also does not assume the imputations are truly observed quantities. Our clustering methods avoid grouping assumptions based around chemical structure and usage. Public health practitioners can leverage these contributions to perform chemical mixture analyses while avoiding the uncertainty of questions related to improper imputations and chemical grouping.

As the investigation into the influence of chemical exposure on human health continues and expands into the larger context of the exposome, Bayesian group index modelling has the potential to be adapted to a wide variety of data types while leveraging the flexibility of Bayesian modelling for the determination of group composition and number. This will only be aided by more computationally efficient estimation algorithms. Improvements in mixture analysis such as those presented above offer a significant contribution to the future improvement of human health and wellbeing.

References

1. Deziel NC, Rull RP, Colt JS, et al. Polycyclic aromatic hydrocarbons in residential dust and risk of childhood acute lymphoblastic leukemia. Environmental Research. 2014;133:388-395. doi:10.1016/j.envres.2014.04.033
2. Ward MH, Colt JS, Metayer C, et al. Residential Exposure to Polychlorinated Biphenyls and Organochlorine Pesticides and Risk of Childhood Leukemia. Environmental Health Perspectives. 2009;117:1007-1013. doi:10.1289/ehp.0900583
3. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Bayesian Group Index Regression for Modeling Chemical Mixtures and Cancer Risk. International Journal of Environmental Research and Public Health. 2021b;18:3486. doi:10.3390/ijerph18073486
4. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Assessment of Grouped Weighted Quantile Sum Regression for Modeling Chemical Mixtures and Cancer Risk. International Journal of Environmental Research and Public Health. 2021a;18:504. doi:10.3390/ijerph18020504
5. Metayer C, Colt JS, Buffler PA, et al. Exposure to herbicides in house dust and risk of childhood acute lymphoblastic leukemia. Journal of Exposure Science & Environmental Epidemiology. 2013;23:363-370. doi:10.1038/jes.2012.115
6. Czarnota J, Gennings C, Colt JS, et al. Analysis of Environmental Chemical Mixtures and Non-Hodgkin Lymphoma Risk in the NCI-SEER NHL Study. *Environmental health perspectives*. 2015;123(10):965-965. doi:10.1289/ehp.1408630
7. Colt, Joanne S, Jay Lubin, David Camann, Scott Davis, James Cerhan, Richard K Severson, Wendy Cozen, and Patricia Hartge. “Comparison of Pesticide Levels in Carpet Dust and Self-Reported Pest Treatment Practices in Four US Sites.” Journal of Exposure Analysis and Environmental Epidemiology 14, no. 1 (2004): 74–83. https://doi.org/10.1038/sj.jea.7500307.
8. Wild CP. The exposome: from concept to utility. *International journal of epidemiology*. 2012;41(1):24-32. doi:10.1093/ije/dyr236
9. De Souza J, Carlos De Francisco A, Macedo DCD. Dimensionality Reduction in Gene Expression Data Sets. *IEEE access*. 2019;7:61136-61144. doi:10.1109/ACCESS.2019.2915519
10. Hilario M, Kalousis A. Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings in bioinformatics*. 2008;9(2):102-118. doi:10.1093/bib/bbn005
11. Boccard J, Rudaz S. Harnessing the complexity of metabolomic data with chemometrics. *Journal of chemometrics*. 2014;28(1):1-9. doi:10.1002/cem.2567

Supplemental Material

Chapter 2 Supplemental Materials

Table S1: List of chemicals and their group used in the CCLS analyses

Chemical	**Chemical Group**
PCB-118	PCB
PCB-138	PCB
PCB-153	PCB
PCB-180	PCB
DDE	Insecticide
DDT	Insecticide
Cyfluthrin(I)	Insecticide
Cyfluthrin(II)	Insecticide
Cyfluthrin(III)	Insecticide
Cyfluthrin(IV)	Insecticide
Carbaryl	Insecticide
Propoxur	Insecticide
Pentachlorophenol	Insecticide
gamma-Chlordane	Insecticide
alpha-Chlordane	Insecticide
Chlorpyrifos	Insecticide
Diazinon	Insecticide
Phosmet	Insecticide
cis-Permethrin	Insecticide
Methoxychlor	Insecticide
Cypermethrin(I)	Insecticide
Cypermethrin(II)	Insecticide
Cypermethrin(III)	Insecticide
Cypermethrin(IV)	Insecticide
trans-Permethrin	Insecticide
Piperonyl butoxide	Insecticide
o-Phenylphenol	Herbicide
Trifluralin	Herbicide
Simazine	Herbicide
mCPP	Herbicide
Dicamba	Herbicide
Dacthal	Herbicide
2,4-D	Herbicide
As	Metals
Cr	Metals
Cu	Metals
Pb	Metals
Sn	Metals
W	Metals
Zn	Metals
Indeno(1,2,3-c,d)pyrene	PAH
Dibenz(ah)anthracene	PAH
Dibenzo(ae)pyrene	PAH
Coronene	PAH

Benzo(a)anthracene	PAH
Benzo(a)pyrene	PAH
Benzo(b)fluoranthene	PAH
Nicotine	Tobacco
Cotinine	Tobacco
PBDE-28	PBDE
PBDE-47	PBDE
PBDE-99	PBDE
PBDE-100	PBDE
PBDE-153	PBDE
PBDE-154	PBDE
PBDE-183	PBDE
PBDE-196	PBDE
PBDE-197	PBDE
PBDE-203	PBDE
PBDE-206	PBDE
PBDE-207	PBDE
PBDE-208	PBDE
PBDE-209	PBDE

Figure S1: Forest plot of chemical group effects for childhood leukemia

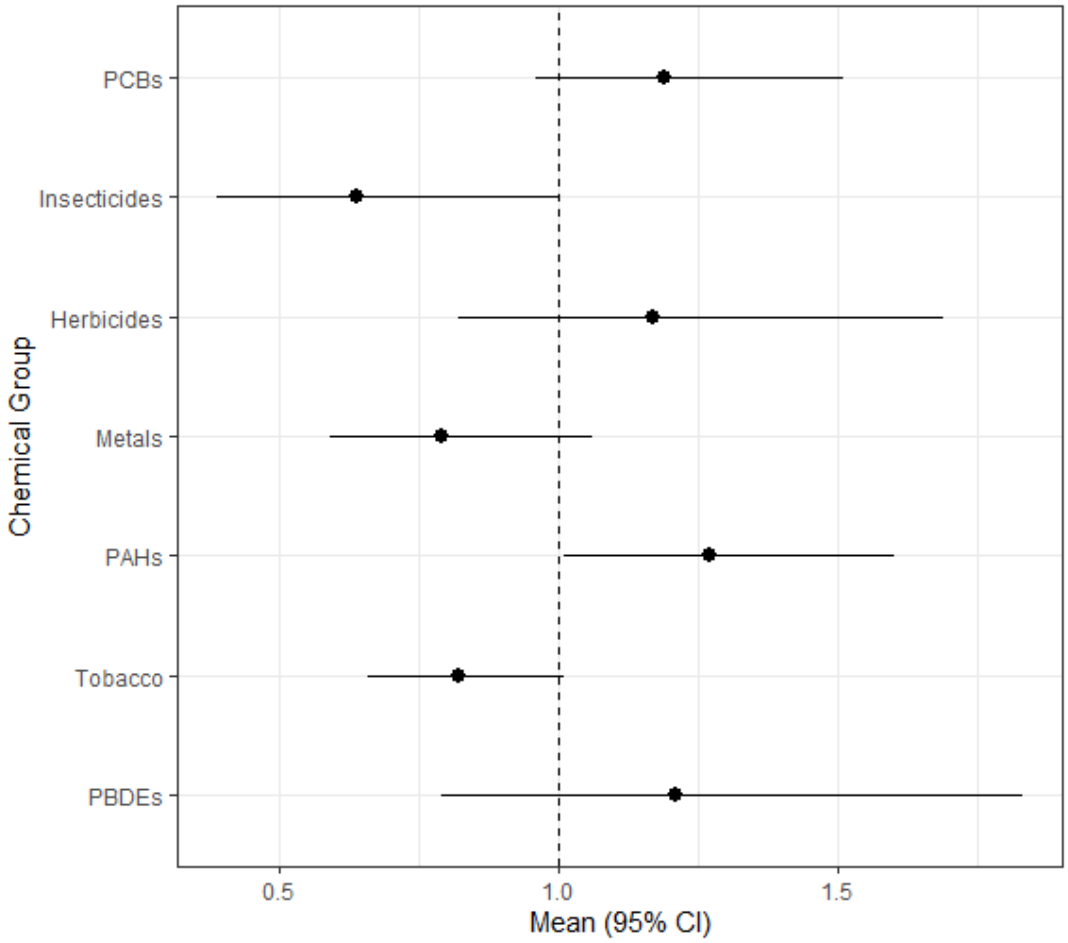


Figure S2: Forest plot of chemical group effects for childhood leukemia in children in the highest income bracket

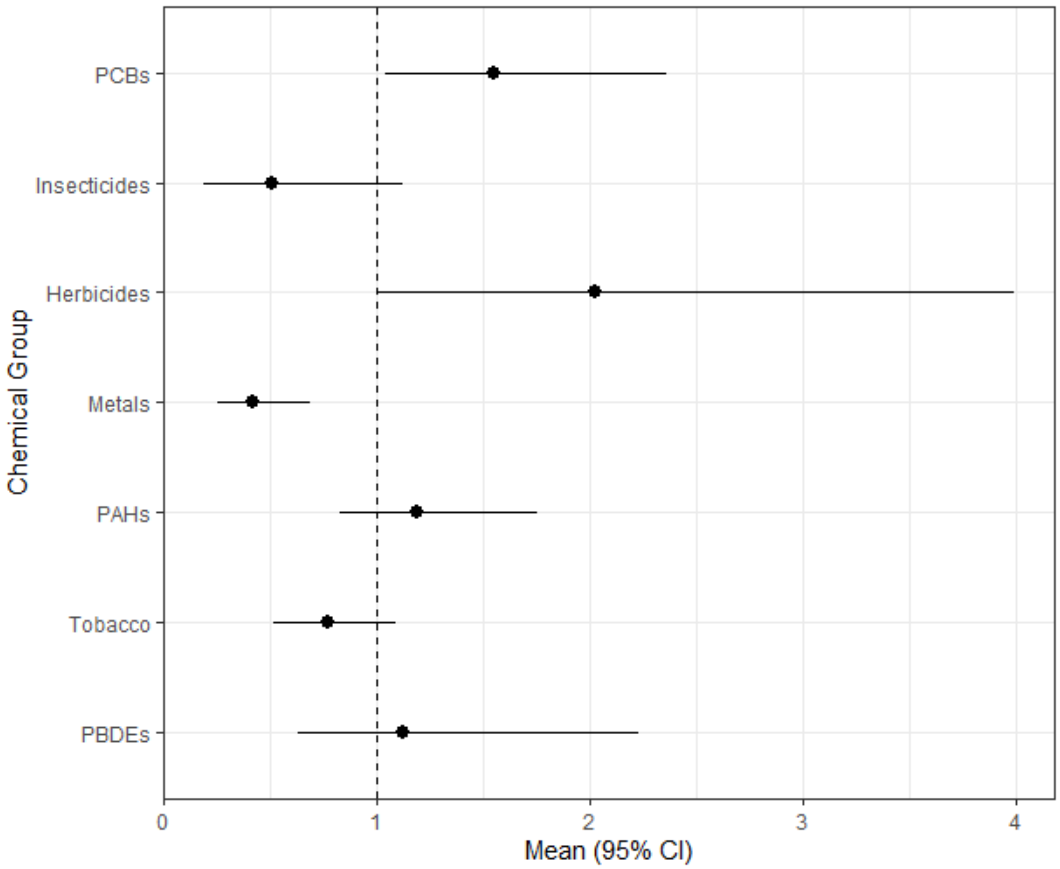
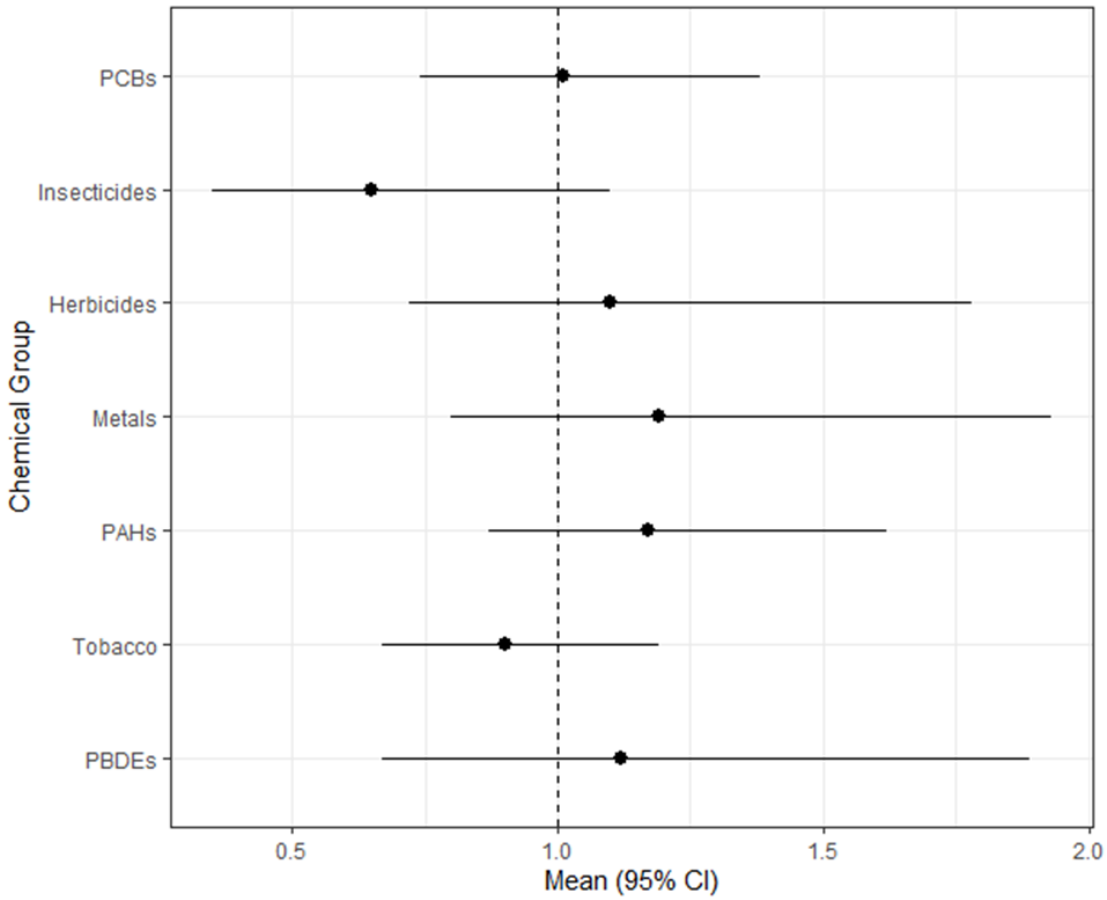


Table S2: Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in lower income brackets

Variable	**Odds Ratio**	**2.5% CI**	**97.5% CI**
PCBs	1.01	0.74	1.38
Insecticides	0.65	0.35	1.10
Herbicides	1.10	0.72	1.78
Metals	1.19	0.80	1.93
PAHs	1.17	0.87	1.62
Tobacco	0.90	0.67	1.19
PBDEs	1.12	0.67	1.89
Child's age	1.03	0.91	1.18
Female	1.40	0.86	2.36
Child's Ethnicity:			
Hispanic	1.44	0.83	2.70
Non-Hispanic	1.20	0.59	2.57
Mother's education:			
High school	1.32	0.61	3.10
Some college	1.39	0.61	3.35
Bachelor's or higher	0.75	0.29	1.89
Mother's age	1.02	0.98	1.07
Residence since birth	0.94	0.56	1.53

Figure S3: Forest plot of chemical group effects for childhood leukemia in children in the lower income brackets



Chapter 3 Supplemental Materials

For the Detroit subset, we fixed the CLV clustering algorithm to 5 clusters. We characterized the five clusters and list their chemicals as follows: a group of pesticides called Group 1 (2,4-D, dicamba, α-chlordane, γ-chlordane, diazinon, *o*-phenylphenol, pentachlorophenol), a group composed of all PAHs and one pesticide called Group 2 (benz(a)anthracene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, chrysene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene, methoxychlor), a group of all PCBs called Group 3 (PCB 105, PCB 138, PCB 153, PCB 170, PCB 180), a group of pesticides called Group 4 (DDE, DDT), and a group of the remaining pesticides called Group 5 (carbaryl, chlorpyrifos, *cis*-permethrin, *trans*-permethrin, propoxur). The odds ratios and 95% CIs estimated for our 5 index effects and covariates are in Table S1. No index effects were found to be significant. The race and age covariates were found to be significant.

Table S1. Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Detroit

Variable	Odds Ratio	2.5% CI	97.5% CI
Group I	0.73	0.40	1.24
Group 2	1.00	0.70	1.55
Group 3	1.40	0.94	2.10
Group 4	0.86	0.61	1.18
Group 5	1.20	0.81	1.88
Male	1.12	0.62	2.09
White	**2.52**	**1.01**	**6.44**
Education	0.95	0.57	1.51
Age	**0.95**	**0.91**	**0.98**

For the LA subset, we fixed the CLV clustering algorithm to 6 clusters. We characterized the six clusters and list their chemicals as follows: a group of pesticides called Group 1 (2,4-D, dicamba), a group of all PAHs called Group 2 (benz(a)anthracene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, chrysene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene), a group of all PCBs called Group 3 (PCB 105, PCB 138, PCB 153, PCB 170, PCB 180), a group of pesticides called Group 4 (carbaryl, *cis*-permethrin, *trans*-permethrin, propoxur), a group of pesticides called Group 5 (α-chlordane, γ-chlordane, chlorpyrifos, diazinon), and a group of the remaining pesticides called Group 6 (DDE, DDT, methoxychlor, *o*-phenylphenol, pentachlorophenol). The odds ratios and 95% CIs estimated for our 6 index effects and covariates are in Table S2. No index effects or covariates were found to be significant.

Table S2. Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in LA

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	1.03	0.81	1.32
Group 2	1.24	0.94	1.67
Group 3	1.20	0.91	1.61
Group 4	1.01	0.74	1.37
Group 5	0.81	0.54	1.25
Group 6	0.73	0.47	1.13
Male	0.90	0.55	1.41
White	1.16	0.70	2.01
Education	1.15	0.79	1.73
Age	1.00	0.98	1.03

For the Seattle subset, we fixed the CLV clustering algorithm to 6 clusters. We characterized the six clusters and list their chemicals as follows: a group of pesticides called Group 1 (2,4-D, dicamba, diazinon), a group of all the PAHs called Group 2 (benz(a)anthracene, benzo(b)fluoranthene, benzo(k)fluoranthene, benzo(a)pyrene, chrysene, dibenz(ah)anthracene, indeno(1,2,3-cd)pyrene), a group of all PCBS and one pesticide called Group 3 (PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, pentachlorophenol), a group of pesticides called Group 4 (carbaryl, chlorpyrifos, *cis*-permethrin, *trans*-permethrin, propoxur), a group of the two chlordane pesticides called Group 5 (α-chlordane, γ-chlordane), and a group of the remaining pesticides called Group 6 (DDE, DDT, methoxychlor, *o*-phenylphenol). The odds ratios and 95% CIs estimated for our 5 index effects and covariates are in Table S3. No index effects were found to be significant. The education covariate was found to be significant.

Table S3. Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Seattle

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	0.76	0.55	1.01
Group 2	1.02	0.83	1.25
Group 3	1.19	0.88	1.67
Group 4	1.05	0.78	1.45
Group 5	0.92	0.72	1.15
Group 6	0.99	0.67	1.45
Male	1.13	0.75	1.75
White	0.92	0.40	2.04
Education	**0.66**	**0.45**	**0.99**
Age	0.99	0.97	1.01

Chapter 4 Supplemental Materials

For the Detroit subset, among the group number models ran, the 5-group model had the lowest DIC (DIC = 258). We characterize these five clusters and list their chemicals as follows: a singleton group composed solely of 2,4-D called Group 1, a group of PAHs called Group 2 (benzo(b)fluoranthene, benzo(a)pyrene, chrysene, and indeno(1,2,3-cd)pyrene), a group of pesticides named Group 3 (*cis*-permethrin and *trans*-permethrin), a group of pesticides, PAHs, and PCBs named Group 4 (dicamba, dibenz(ah)anthracene, PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, carbaryl, α-chlordane, γ-chlordane, chlorpyrifos, DDE, DDT, diazinon, methoxychlor, and propoxur), and a group of PAHs and pesticides called Group 5 (benz(a)anthracene, benzo(k)fluoranthene, *o*-phenylphenol, and pentachlorophenol). The odds ratios and 95% CIs estimated for our 5 index effects and covariates are in Table S1. No indices were found to have a significant association with NHL, however, the gender covariate had a significant and positive association (OR = 2.80, 95% CI: 1.10, 7.36) and the age covariate had a significant and negative association (OR = 0.95, 95% CI: 0.91, 0.98).

Table S1. Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Detroit

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	1.00	0.76	1.31
Group 2	1.23	0.71	3.23
Group 3	1.14	0.87	1.52
Group 4	1.35	0.76	2.60
Group 5	0.55	0.18	1.12
Male	1.10	0.61	2.03
White	**2.80**	**1.10**	**7.36**
Education	0.99	0.60	1.65
Age	**0.95**	**0.91**	**0.98**

For the Seattle subset, among the group number models ran, the 6-group model had the lowest DIC (DIC = 481.4). We characterize these five clusters and list their chemicals as follows: : a singleton group composed solely of 2,4-D named Group 1, a group of pesticides named Group 2 (*cis*-permethrin and *trans*-permethrin), a singleton group composed solely of pentachlorophenol called Group 3, a group of pesticides, PAHs, and PCBs named Group 4 (dicamba, benzo(k)fluoranthene, dibenz(ah)anthracene, PCB 105, PCB 138, PCB 153, PCB 170, PCB 180, carbaryl, α-chlordane, γ-chlordane, chlorpyrifos, DDE, DDT, diazinon, methoxychlor, and propoxur), a group of PAHs called Group 5 (benz(a)anthracene, benzo(b)fluoranthene, benzo(a)pyrene, chrysene, and indeno(1,2,3-cd)pyrene), and a singleton group composed solely of *o*-phenylphenol called Group 6. The odds ratios and 95% CIs estimated for our 6 index effects and covariates are in Table S2. No indices were found to have a significant association with NHL, however, the education covariate had a significant and inverse association (OR = 0.65, 95% CI: 0.44, 0.99).

Table S2. Odds ratio estimates for chemical groups and demographic covariates from the Bayesian group index model for subjects in Seattle

Variable	Odds Ratio	2.5% CI	97.5% CI
Group 1	0.86	0.70	1.04
Group 2	1.07	0.89	1.30
Group 3	1.13	0.92	1.42
Group 4	0.89	0.52	1.34
Group 5	1.02	0.81	1.33
Group 6	1.06	0.87	1.31
Male	1.18	0.78	1.84
White	0.93	0.41	2.03
Education	**0.65**	**0.44**	**0.99**
Age	0.99	0.97	1.01