frontiers | Frontiers in Public Health

TYPE Review
PUBLISHED 09 May 2024
DOI 10.3389/fpubh.2024.1377685

OPEN ACCESS

EDITED BY
Dimirios Nikolopoulos,
University of West Attica, Greece

REVIEWED BY
Jianjun Xiang,
Fujian Medical University, China
Sean Mark Patrick,
University of Pretoria, South Africa

*CORRESPONDENCE
Tong Wang
✉ tongwang@sxmu.edu.cn

RECEIVED 28 January 2024
ACCEPTED 15 April 2024
PUBLISHED 09 May 2024

CITATION
Zhu G, Wen Y, Cao K, He S and Wang T (2024) A review of common statistical methods for dealing with multiple pollutant mixtures and multiple exposures.
Front. Public Health 12:1377685.
doi: 10.3389/fpubh.2024.1377685

A review of common statistical methods for dealing with multiple pollutant mixtures and multiple exposures

Guiming Zhu[1,2], Yanchao Wen[1,2], Kexin Cao[1,2], Simin He[1,2] and Tong Wang[1,2]*

[1]Department of Health Statistics, School of Public Health, Shanxi Medical University, Taiyuan, China, [2]Key Laboratory of Coal Environmental Pathogenicity and Prevention (Shanxi Medical University), Ministry of Education, Taiyuan, China

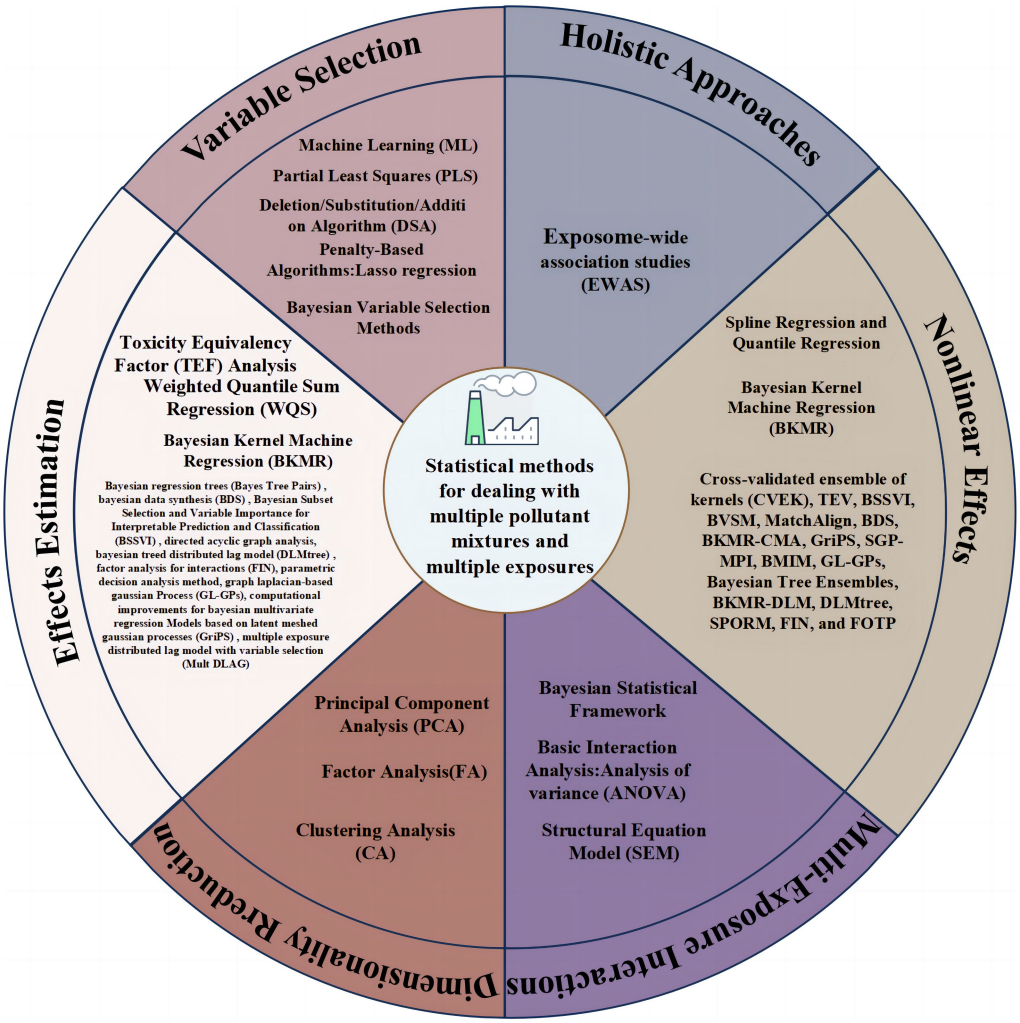
Traditional environmental epidemiology has consistently focused on studying the impact of single exposures on specific health outcomes, considering concurrent exposures as variables to be controlled. However, with the continuous changes in environment, humans are increasingly facing more complex exposures to multi-pollutant mixtures. In this context, accurately assessing the impact of multi-pollutant mixtures on health has become a central concern in current environmental research. Simultaneously, the continuous development and optimization of statistical methods offer robust support for handling large datasets, strengthening the capability to conduct in-depth research on the effects of multiple exposures on health. In order to examine complicated exposure mixtures, we introduce commonly used statistical methods and their developments, such as weighted quantile sum, bayesian kernel machine regression, toxic equivalency analysis, and others. Delineating their applications, advantages, weaknesses, and interpretability of results. It also provides guidance for researchers involved in studying multi-pollutant mixtures, aiding them in selecting appropriate statistical methods and utilizing R software for more accurate and comprehensive assessments of the impact of multi-pollutant mixtures on human health.

KEYWORDS

health effects, epidemiology, statistical methods, multi-pollutant mixtures, environment

1 Introduction

In the contemporary industrialized society, environmental concerns such as air pollution, water pollution, and soil contamination have gained significant attention (1–4). Some pollutants are metabolized because of their shorter half-lives, others, such as heavy metals, insecticides, flame retardants, persistent organic pollutants, and other endocrine-disrupting chemicals continue to accumulate in the human body and have significant and long-term effects on human health (5–8). For instance, the association between heavy metals toxicity and the development of neurodegenerative diseases and various ocular pathologies has been established, while concurrent exposure to heavy metals can elevate the risk of prostate cancer and thyroid enlargement (9–11). Polybrominated diphenyl ether is a persistent and pervasive environmental pollutant that disrupts the human endocrine system, leading to health implications such as developmental, thyroidal, and reproductive toxicity (12, 13). Particulate



GRAPHICAL ABSTRACT

matter along with nitrogen oxides in the atmospheric environment exhibit a close correlation with stroke incidence rate and mortality. The higher the concentration of particulate matter exposure, the greater risk of stroke (14). However, most studies are on single pollutants, they do not accurately reflect the real world since people are exposed to a combination of several dangerous substances at any given time, which might have antagonistic or synergistic effects. What's more, single-pollutant analysis methods often fall short in capturing the complexity and interactive effects of multi-pollutant mixtures (15). Lastly, the potential for spurious associations has increased in single-pollutant models, contributing to disagreements between studies. Consequently, accurately assessing the health effects of exposure to mixtures of environmental pollutants has become a focal point in current environmental epidemiology. In recent years, the focus of health effect assessments of environmental risk factors has shifted from the traditional single-pollutant approach to the study of mixtures of multiple pollutants (16–18). This shift aims to more accurately reflect the impact of environmental risk factors on human health.

In order to handle the multi-pollutant mixtures exposure, researchers have raised a number of problems that need to be addressed, as shown in Table 1.

The study of multi-pollutant mixtures is characterized by two primary focuses: (A) Estimating the effects of pollutants and (B) addressing the complexity associated with multi-pollutant mixtures.

Regarding effects estimation, the focus is divided into three aspects: (1) Overall effects of multi-pollutant mixtures; (2) Independent effects of components within multi-pollutant mixtures; and (3) Joint effects of mixture components.

Addressing the complexity of multi-pollutant mixtures involves three main aspects: (1) Addressing the challenge of high-dimensional data when multiple chemical substances are present in the model; (2) Resolving the issue of high correlations among pollutants to assess synergistic or antagonistic effects; and (3) Addressing interplay and non-linear effects among pollutants.

Thus, diverse statistical methods are introduced in this paper to address specific issues, as shown in Graphical abstract.

2 Methods for effects estimation

In estimating overall effects of multi-pollutant mixtures, two main approaches are commonly used: treating the mixtures as a single exposure or analyzing the weighted sum of exposures (17). For

TABLE 1 Key research questions on multi-pollutant mixtures exposure.

Authors or projects	Research questions
Kortenkamp (18, 19)	• Overall effects of mixtures, rather than single effect.
Ghassan (16) and Braun (17)	• Overall effects of mixtures; Weighting and effects of mixtures; Independent effects of each component of the mixture; Joint effects of each component of the mixture.
Gibson (20)	• Are specific exposure patterns present in the study population? What toxic substances are present in the mixture? Alternatively, what is the independent impact of each mixture member on the health outcomes of interest? Are there synergistic effects or interactions among mixture members? What is the overall impact of the mixture on the outcomes of interest?
Powering Research through Innovative methods for Mixtures in Epidemiology (PRIME) program (15)	• Overall effect estimation: What is the overall effect of the mixture, and what is the magnitude of the association? • Toxin identification: Which congeners/exposures are associated with the outcome? What exposure is most significant? • Pattern recognition: Are there specific exposure patterns in the data? • Predefined groups: What is the association between outcomes and pre-defined exposure groups? • Interactions and non-linearity: Are there interactions between exposures, and if so, which influences modify patterns? Is the exposure-response surface non-linear?
Stafoggia (21) and Yu (22)	• Dimensionality reduction; • Variable selection; • Observational grouping.

example, particle matter concentration serves as a comprehensive measure of particulate matter components in ambient air, assuming equal impact on health for all components (23). Alternatively, overall effect estimation can involve the weighted sum of individual component effects, with weights based on toxicological potency or contribution percentage (17). For instance, modeling phthalate metabolites' concentrations using molar sum or potency-weighted sum methods (24, 25). Independent effects estimation considers diverse pollution components' varied adverse effects, requiring the evaluation of overall mixture impact before assessing individual component effects (26). Joint effects estimation accounts for component interactions beyond additive impacts (18), considering mechanisms and pathways. For example, non-volatile carbonaceous particles like black carbon may require specific attention when studying joint health effects with volatile organic compounds. In this section we briefly introduce several methods of effects estimation. Figure 1 shows details of the effect estimation methods and R packages for their implementation.

2.1 Weighted quantile sum regression

Weighted quantile sum (WQS) regression is a convenient tool for addressing effect estimation problem, the high-dimensional and highly correlated issues among multiple pollutants, particularly among homogenous pollutants (27). It is widely used in studies of environmental exposure to multi-pollutant mixtures and enables the identification of high-risk factors. This model allows the construction of a weighted index in a supervised manner to assess the overall effects of environmental exposure and the contribution of each component in the mixture to the overall effect. WQS calculates individual exposure characteristics by weighting based on the correlation between exposure and outcome, resulting in a composite index value for various exposure components. This index characterizes the levels of mixed exposure to a range of exposure components and evaluates the impact of each component on health outcomes. Following Tanner's recommendation, introducing a bootstrap step in the WQS yields stable weights for exposure components and WQS index estimates (28). The core idea is to construct WQS to achieve dimensionality reduction, address multicollinearity issues, and filter high-risk factors through the weighting process. The most recent weight coefficient for each component in the exposure index represents its contribution to health outcomes.

WQS has advantages in analyzing multifactor exposure due to their simple model structure, small computational burden, and fast analysis speed. But the "directional consistency" precondition must be met, i.e., the effects of each component in the mixture are all in the same direction (all positive or all negative). Recent research has explored and developed new methods for WQS. For unidirectional hypotheses, methods such as quantile g-computation combined with the g-algorithm (29), grouped WQS (30), and Bayes group WQS model (31) have been developed. Focus has also been given to the lagged WQS to address time-varying exposure mixtures (32, 33).

2.2 Bayesian kernel machine regression

Bayesian kernel machine regression (BKMR) also provides a new approach to analyze multi-pollutant mixtures (34, 35). In contrast to WQS, BKMR provides probabilities included in the total effects of multi-pollutant mixtures, rather than estimate the percentage contribution of this effect but provides probabilities included in the total effects of multi-pollutant mixtures. It visualizes various exposure-response shapes. BKMR can also examine the independent impacts of mixture components by considering the effects of keeping other components constant at predetermined percentiles, such as the 50th percentile of the exposure distribution. BKMR does not require setting a parameter expression, allowing for the presence of nonlinear effects and interactions. It generates kernel functions based on the mixture variables included in the model, followed by Bayesian sampling and analysis methods to generate relationship curves between mixture components and disease variables included in the model.

In addition to analyzing the mixture's overall impacts and each component's effects separately, BKMR also estimates any possible interactions between the distinct components. Posterior inclusion

Objective	Overall effect estimate			Dimensionality reduction			Nonlinear effects	
Method	Weighted Quantile Sum	Bayesian Kernel Machine Regression	Toxicity Equivalency Analysis	Principal component analysis and Factor analysis	Cluster analysis	Partial least square method	Spline regression and Quantile regression	Bayesian Kernel Machine Regression
Advantages	Mitigate any issues related to overfitting and collinearity; in comparison to regularization methods, achieve a lower mean squared error and higher specificity. Assess the combined impact of multiple pollutants on health and prioritize the significant constituents within the mix.	Robustness addresses multicollinearity; resolves high-dimensional issues; estimates interactions between components in a mixture; achieves variable importance computation.	From a toxicological perspective, pollutants that share the same mechanism of action and endpoint exhibit additive toxicity. This is easily understandable and directly correlates with real exposure and toxicity data.	Non-parametric constraints; noise reduction; results interpretation facilitation; elimination of multicollinearity and achieving dimensionality reduction.	Clustering can be used for dimensionality reduction, reducing the dimensions of a dataset to make it more easily visualized and understood while retaining essential information. By identifying groups of objects with similar features, clustering can help identify features that share significant commonalities in the dataset. It reveals latent structures and patterns in the dataset, enabling researchers to better comprehend the data.	Reducing the dimensionality of multivariate data helps decrease the complexity of models. This is beneficial for handling high-dimensional data and mitigating the risk of overfitting. Partial least squares (PLS) often demonstrates superior modeling performance for small-sample data compared to other methods.	Spline regression models are highly flexible and can adapt to various nonlinear relationships of different shapes. They achieve smooth fits, helping to reduce the impact of noise and fluctuations, thus enhancing the stability of the model.Quantile regression is relatively robust to outliers and extreme values, as it estimates quantiles rather than means. This approach is better equipped to handle extreme cases in the data. Quantile regression provides estimates of regression coefficients at different quantiles, allowing for a more accurate capture of nonlinear relationships in various data distributions.	Enables nonlinear estimation and visualization
Limitations	The transformation of data into quantiles results in information loss; all chemical substances are constrained in the same direction as related to the outcomes.	The magnitude of the Performance Improvement Percentage (PIP) is highly sensitive to the selection of tuning parameters.	The requirement for available toxicity and exposure data for each chemical emphasizes that the assessment results depend on the selection of indicative chemicals and the quality of toxicological information. The uncertainty associated with chemicals will significantly impact the uncertainty of risk assessment results	Excluding quadratic terms, no consideration for interaction forms and non-linear relationships. Exposure variables are limited to continuous variables; components may lack biologically meaningful interpretations; the establishment of the number of principal components is not absolute.	Cluster selection, classification methods, and the determination of the number of clusters in clustering often require a strong degree of subjectivity based on prior knowledge.	PLS models are typically considered black-box models, making them challenging to interpret. The number of components may influence the results to a certain extent. PLS is sensitive to noise, so preprocessing of data is necessary before modeling to reduce the impact of noise.	In spline regression, the selection of parameters such as the number, placement, and smoothness of splines involves subjective decisions. Different choices may lead to different results. If too many splines are chosen carelessly, the model may overfit the data, resulting in poor generalization. Spline regression typically requires a relatively large number of data points to stabilize the estimation of nonlinear relationships, making it potentially less reliable in small-sample situations.In quantile regression, certain parameters, such as the choice of quantile levels, may also involve subjective decisions and require selection based on the researcher's questions and domain knowledge.	Bayesian kernel regression is often considered a black-box model, which is not easily interpretable and understandable, especially in applications that require transparency and interpretability. This may pose limitations in scenarios where clarity and interpretability are essential.
R package	gWQS	bkmr	NA	FactoMineR; factoextra; psych	factoextra	pls	rms；mgcv；gam; splines; quantreg	bkmr

Objective	Variable selection				Interaction effects estimate			
Method	Deletion/Substitution/Addtion	Penalized statistical methods (LASSO、Ridge Regression、Elastic-Net)	Machine learning	Bayesian variable selection method	Analysis of Variance (ANOVA)	Bayesian Statistical Framework	Structural Equation Model	Cross-Validated Kernel Ensemble
Advantages	Reduces false positive rates, enables mutual adjustment between variables, and allows exploration of interactions among chemical components; less sensitive to outliers.	LASSO (Least Absolute Shrinkage and Selection Operator): Robustly addresses multicollinearity; results in smaller coefficient variances compared to ordinary least squares regression. Ridge Regression: Effectively handles multicollinearity among predictor variables by constraining the size of parameters, reducing sensitivity to collinearity; lowers model variance, reducing the risk of overfitting. Performs well in high-dimensional datasets, aiding in controlling model complexity with numerous features. Elastic-Net: Robustly addresses multicollinearity; resolves high-dimensional issues; exhibits higher predictive accuracy than LASSO.	Machine learning methods can automatically sift through vast amounts of data to identify crucial features, thereby reducing the need for manual intervention, particularly in high-dimensional datasets. These methods excel in capturing nonlinear and intricate relationships within the data, enabling the discovery of significant features that traditional approaches might overlook. Furthermore, they have the capability to comprehensively consider interactions among multiple features, as opposed to merely individually screening each feature.	For small sample datasets, it exhibits superior performance and can provide uncertainty estimates regarding parameter and variable selection	Providing statistical significance information about different factors and interactions, one can compute effect sizes to aid in assessing the practical magnitude of the effects of factors and interactions. Regression analysis offers a clear explanation of the relationship between interaction terms and the dependent variable, enabling researchers to understand the interaction effects between exposures.	Enables estimation of interactions between multiple mixture components	Aids in estimating and understanding the network of relationships among variables (latent variables, observed variables, and error variables). This allows researchers to build complex models involving multiple independent variables, mediator variables, and interactions, facilitating a more comprehensive exploration of various relationships. Simultaneously estimating relationships between observed variables and latent variables helps capture the multi-component latent structure.	Simultaneously capturing nonlinearity and interaction effects among variables, it exhibits strong robustness, especially in situations with a small sample size
Limitations	When exploring interactions involving chemicals with low detection rates, the impact is limited.	Lasso: Lasso identifies only the crucial mixture components with a linear relationship to the outcome. Ridge Regression: Ridge regression is not an explicit variable selection method; it tends to assign non-zero coefficients to all predictor variables, even though some coefficients may be very close to zero. This implies that the model still contains unrelated or less relevant predictor variables, necessitating additional screening steps to identify important features. Ridge regression is typically unsuitable for handling categorical variables as it is based on a linear model for continuous variables. Choosing an appropriate regularization parameter (often used to adjust the strength of ridge regression) requires some subjective judgment or techniques like cross-validation, which may involve domain knowledge or trial and error. Elastic-Net: Post-selection statistical inference tools need to be applied, including generating confidence intervals; results may be susceptible to false positives.	Machine learning algorithms are often regarded as black-box models, posing challenges in terms of interpretability and understanding. This characteristic can make it difficult to interpret models after feature selection, particularly in applications that require transparency and interpretability. There may be limitations in explaining the selected features, potentially restricting the utility of the model in certain contexts. Additionally, there is a risk of overfitting to the training data during the feature selection process, leading to unstable feature choices.The selection of machine learning algorithms and feature selection methods tailored to specific problems necessitates domain knowledge and experimentation. Inappropriately chosen algorithms may result in inefficient or inaccurate feature selection processes.	Bayesian variable selection typically involves complex computations, especially in high-dimensional datasets. The computational cost of Bayesian methods can be substantial, and the choice of tuning parameters is highly sensitive.	Analysis of variance (ANOVA) typically assumes a linear relationship between the dependent variable and the independent variables. If the relationship is nonlinear, alternative methods may be necessary. ANOVA relies on crucial statistical assumptions such as normality, homoscedasticity, and independence. Considering multiple exposure interaction terms can increase the complexity of the model, making it challenging to interpret and understand. Regression analysis is often based on the assumption of a linear relationship. If the interaction effects between exposures are nonlinear, a regression model may struggle to accurately capture them. Utilizing regression analysis to explore interaction effects requires multiple tests for various hypotheses, potentially increasing the risk of multiple comparison issues that need correction.	PIP is highly sensitive to the selection of tuning parameters.	Necessary to meet some basic assumptions of conventional statistical analysis, such as linearity and normality. The model construction involves subjective decisions, including the selection of model fit indices, drawing path diagrams, and defining latent variables.	Good statistical performance in quantitative data
R package	DSA	Glmnet；lars；glinternet	Kernelshap；shapviz；caret; DALEX；mlr3；xgboost;e1071; randomForest; gbm; rpart; neuralnet; catboost	bkmr	glm	bkmr	Lavaan	CVEK

FIGURE 1
Overview of the methods and R packages for implementation.

probabilities (PIPs) generated by BKMR range from 0 (least important) to 1 (most important). Components with PIP ≥ 0.5 are identified as relatively important mixture components. BKMR can also be used to study possible three-way interactions. This is achieved by fixing one of the exposures at different quantile levels and visualizing the exposure-response functions for the remaining two exposures. Overall, BKMR has been widely used in environmental health research, including the analysis of continuous variables, binary variables, and repeated measurement data (36, 37). The advantages of this method include the ability to simultaneously assess the importance of each variable, analyze data with uncertainty, and easily extend the obtained results to longitudinal data.

Although BKMR can effectively assess the health effects of multi-pollutant mixtures, it has certain limitations. Firstly, when using the BKMR, the studied exposure variables must be continuous, and the size of PIPs is easily influenced by adjustment parameters. So, caution is required when interpreting results, as this method may obscure the underlying complex features of the data. If some components in the mixture are positively correlated while others are equally negatively correlated, the final overall result will appear as if there is no correlation, and other methods are needed to verify the estimation of their interactions. In addition, considering causality, time-varying exposure, or computational efficiency in massive datasets, the traditional implementation of BKMR may be limited. Several new methods have extended the BKMR strategy to address these limitations, such as bayesian kernel machine regression – causal mediation analysis (BKMR-CMA) (38), bayesian kernel machine regression distributed lag model (BKMR-DLM) (39).

2.3 Toxicity equivalency analysis

In addition to the two commonly used estimation methods mentioned above, pollutants with similar mechanisms of action and the same endpoint from a toxicological perspective exhibit additive toxicity. However, individual pollutants contribute differently to the overall health risk. Therefore, a normalization method, known as toxicity equivalency factor (TEF) analysis, a normalizing technique, is required. TEF is generally obtained by comparing the "starting point" of health risk assessments for standard reference compounds with the respective compounds. The exposure dose of a mixture, commonly represented as toxicity equivalent quantity (TEQ), is calculated by multiplying the TEF for each compound by its exposure metric and summing them. By combining TEQ with reference metrics such as the reference dose (RfD) or carcinogenic slope factor, the health risk of the mixture can be assessed (40–43). TEF represents the relative toxicity of an isomer of a compound and is set to 1 for the most toxic 2,3,7,8-TCDD. Other pollutants' toxicities are converted to their corresponding relative toxic intensities. Alternatively, TEF can be the toxicity equivalency factor for individual Polycyclic Aromatic Hydrocarbons, with a TEF of 0.001 for Pyrene. Daily total intake exposure metric from plasma polycyclic aromatic hydrocarbon levels based on

pharmacokinetic models (40–42). The results can be compared against specified standards to determine the presence of carcinogenic risk (43). To address non-linear problems, the acceptable concentration range model has been developed based on the RfD concept (44).

TEF has the advantage of being easy to understand and directly associated with real exposure and toxicity data. However, it requires available toxicity and exposure data for each chemical, making the assessment results dependent on the selection of indicative chemicals and the quality of toxicological information. Uncertainty in the chemicals significantly affects the uncertainty of risk assessment results.

2.4 Other methods for effects estimation

In addition to the three statistical methods for estimating effects mentioned above, there are also novel and unique methods for effect estimation, although they may have a narrower audience. These include bayesian regression trees (45), bayesian data synthesis (BDS) (46), bayesian subset selection and variable importance for interpretable prediction and classification (BSSVI) (47), directed acyclic graph analysis (48), bayesian treed distributed lag model (DLMtree) (49), factor analysis for interactions (FIN) (50), parametric decision analysis method (51), graph laplacian-based gaussian Process (GL-GPs) (52), computational improvements for bayesian multivariate regression models based on latent meshed gaussian processes (GriPS) (53), and multiple exposure distributed lag model with variable selection (54).

3 Methods for dimensionality reduction

When analyzing multi-pollutant mixtures with fewer components, the process is relatively simple, but the dimensionality of the data increases dramatically when multi-pollutant mixtures contain several components. Many statistical methods lack the capability to address this issue, and even methods designed to handle the complexity of high-dimensional data incur exponential time costs as the data dimensionality grows. Furthermore, the high correlation among components may lead to multicollinearity. For instance, analyzing correlated components with similar sources, exposure pathways, or metabolic processes, regardless of which one is individually studied, may yield biased conclusions. Faced with the challenges of high dimensionality and multicollinearity, a crucial aspect of studying the health impacts of multi-pollutant mixtures involves learning low-dimensional structures in the data to enhance interpretability and statistical efficiency, employing methods of dimensionality reduction proves to be a favorable approach. In this section we briefly introduce several dimensionality reduction methods. Figure 1 shows details of the methods for dimensionality reduction and R packages for their implementation.

3.1 Principal component analysis and factor analysis

Principal component analysis (PCA), introduced by Pearson for non-random variables and later extended to random vectors by Hotelling, transforms a set of potentially correlated variables into a set of linearly uncorrelated variables, referred to as principal components, through orthogonal transformation (55). The primary objective of PCA is to explain the majority of variance in the original data using fewer variables, converting highly correlated variables into ones that are independent or uncorrelated. When analyzing the relationship between multiple pollutant indicators and health, PCA can reduce the number of indicators for analysis, minimizing information loss from the original indicators and facilitating comprehensive data analysis. It simplifies high-dimensional exposure data into several orthogonal components usable for regression models, thus mitigating multicollinearity issues. For example, Smit applied PCA to estimate the relationship between the risk of asthma and eczema in school-age children and 16 pollutants in their mothers' serum. The study ultimately incorporated indicators from five principal components, explaining 70% of the variance in the outcome variable (56).

PCA's main limitations include difficulty in interpreting results, as the components are not in the same units as the original exposure variables, and the derived components may lack a direct relationship with study outcomes as they are derived in an unsupervised manner. Subsequently, PCA has evolved into methods like supervised PCA, which overcomes these issues by excluding "pollutants" that do not provide information directly related to the outcomes (57). Roberts applied this method to air pollution analysis, proposing a recursive algorithm that identifies the optimal predictor for study outcomes and combines it into several relevant principal components (58). Other developments include principal component pursuit (PCP), an analysis method based on matrix factorization, extended to multi-pollutant mixtures by Gibson. Through cross-validation in simulations, PCP identified the true number of patterns in all simulations, while PCA achieved this in only 32% of simulations, demonstrating PCP's superiority in most simulation scenarios (59). In addition to the above methods, Positive matrix factorization (PMF) is a variant of PCA applicable to multi-pollutant profiles, deriving air pollution sources from individual chemical components (60). Specifically, PMF decomposes the matrix of mixture data into two matrices—source contributions and source profiles. Source contributions represent the mass contribution of each source to the mixture measurements, while source profiles reflect the emission types from a given source. Source contributions are constrained to be non-negative, and the method can incorporate uncertainty measurements related to the data at each point (61–64).

Factor analysis (FA) is another commonly used dimensionality reduction method that groups variables based on the correlation matrix, creating common factors that represent the fundamental structure of the data. It decomposes multidimensional variables into a small number of common factors, where the fundamental idea is to break down original variables into two parts: one part is a linear combination of common factors that condenses a vast majority of information in the original variables, and the other part is special factors unrelated to common factors, reflecting the gap between the linear combination of common factors and the original variables. In other words, FA aggregates numerous variables into a few independent common factors with minimal loss or little loss of original data information. These common factors can reflect the essential information of numerous variables, reduce the number of variables, and reveal the inherent connections among variables. Perturbation FA is commonly used in multi-pollutant mixtures studies, focusing on

exploring the similarities and differences in exposure conditions among different groups. For example, Roy used this method to assessed the differences in exposure characteristics in biological or social structures based on race/ethnicity (65).

Both PCA and FA seek a small number of variables to comprehensively reflect the majority of information in all variables. While the number of variables is fewer than the original variables, the information contained is substantial, and the reliability of using these new variables for analysis remains high. Moreover, these new variables are uncorrelated, eliminating multicollinearity and achieving dimensionality reduction. In PCA, the newly determined variables are linear combinations of the original variables, obtained through coordinate transformation. In contrast, FA aims to explain the complex relationships present in many observed variables using a small number of common factors. It does not recombine original variables but decomposes them.

3.2 Clustering analysis

Clustering analysis (CA) organizes all data into clusters, groups of similar elements, where instances within the same cluster are similar to each other, while those in different clusters are dissimilar. Similarity among data is determined by defining a distance or similarity coefficient (66). Once several clusters are identified, the next step is to select a representative prototype for each cluster. CA can be matched with exposure data to define groups, and indicators of group members can then be used as predictor variables in health outcome regression models.

Clustering can be categorized into different groups based on techniques, with partition-based clustering being the most widely used, where k-means is a common approach (67). One advantage of the k-means method is its linear complexity, making its execution time proportional to the number of individuals, making it suitable for large datasets. However, the choice of initial centers and the number of clusters is arbitrary and can influence the results. Nevertheless, hierarchical classification can be applied to the cluster centers obtained from the k-means method. Clustering has been used in several studies to assess the impact of various pollutants. For instance, in time series analysis of air pollution, one study used k-means to divide days into five groups representing days with low pollution levels, high concentrations of crustal particles, high particle content from traffic and combustion of oil, days influenced by regional pollution sources, and days with high concentrations of particles from wood or oil burning (68). Some clusters were associated with pulse amplitude. Similarly, based on pollutant characteristics and community background, an evaluation was conducted on the correlation between NO_2, NO, and $PM_{2.5}$ concentrations and low birth weight (69).

The challenges of CA lie in the selection of clusters, classification methods, and the number of clusters. CA facilitates the distinct grouping of various entities, making it challenging to summarize them under a single label. The process typically necessitates the initial selection of appropriate distance metrics, clustering algorithms, and the number of clusters. These choices often based on users' subjective judgments by the user, and different selections may yield disparate clustering outcomes, thereby rendering the results subjective.

4 Methods for variable selection

When analyzing multi-pollutant mixtures with many components, it is not necessary to estimate the impact of each component of mixture, rather, the focus is on investigating the effects of a few crucial components that exhibit the maximum toxicity to human health and/or have the highest predictive power for the outcomes of interest. Therefore, it is imperative to employ appropriate methods for identifying or selecting important variables that represent the exposure-response relationship between individual exposures in the mixture and the outcomes. Uch methods are frequently known as "variable selection." In this section we briefly introduce several methods of variable selection. Figure 1 shows details of the methods for variable selection and R packages for their implementation.

4.1 Partial least squares

Partial least squares (PLS) regression combines principal component analysis and multivariate regression, taking into account the correlation between the outcomes and exposure variables (70). In essence, PLS regression searches for a linear decomposition of the exposure matrix that maximizes the covariance between exposure and outcomes. The exposed variable has a higher weight in the linear combination, the stronger the association between it and the outcome. PLS regression can also include multiple outcome variables. The optimal number of components can be selected based on cross-validated mean squared error (71). However, a drawback of PLS regression is that the interpretation of the linear combination can be challenging, especially in the presence of a large number of original exposure variables.

To address this limitation, Chun and Keles introduced a method called sparse PLS regression, which simultaneously combines variable selection and dimensionality reduction (72). This method results in a linear combination of exposure variables with reduced quantity. Sparsity is introduced into the loadings of exposure variables through penalty terms. The optimal number of components and sparsity parameters are selected based on cross-validated performance. This method has been applied in simulation studies related to exposure-health associations. In one simulation study involving 237 generated exposure covariates, 0 to 25 of which were related to the outcomes, sparse PLS regression demonstrated better sensitivity in distinguishing true predictive factors from correlated covariates (73).

4.2 Deletion/substitution/addition algorithm

The Deletion/Substitution/Addition (DSA) algorithm is a variable selection method (74, 75). The main steps include: (1) removal of selected variables; (2) substitution of selected variables with unselected ones; and (3) addition of new variables. By using five-fold cross-validation to minimize the root mean square error (L2 loss function) of the prediction equation, the number and particular kinds of variables included in the model are ascertained. To ensure selection stability, DSA is run with different seed numbers for 50 iterations. Subsequently, a binomial generalized linear model evaluation is conducted for multi-exposure variables. Variables included in the final model are those selected in at

least 6% ($n \geq 3$ times) or 10% ($n \geq 5$ times) of DSA iterations, and multicollinearity is validated in the final model.

Compared to traditional linear regression equations, this method reduces the false-positive rate, allows for mutual adjustments between variables, and explores interactions between chemicals. However, its effectiveness is limited when exploring interactions involving chemicals with low detection rates. The algorithm also provides the possibility of including interaction terms. In contrast to stepwise model selection procedures, DSA has the advantage of being less sensitive to outliers and permits movement between non-nested statistical models. In previous applications, the DSA algorithm has been utilized in multi-pollutant mixtures analysis, estimating the relationship between O_3, CO, NO_2, PM_{10} and lung function (76). However, DSA has faced criticism, particularly when the ratio of sample size to the number of candidate predictors is small, leading to inconsistent estimates. Moreover, its statistical properties for confidence intervals are compromised, when there is substantial correlation between predictors (77).

4.3 Penalty-based algorithms

Least absolute shrinkage and selection operator (LASSO) regression is highly similar to ordinary least squares, with the key difference lying in the estimation of coefficients through the minimization of a slightly different quantity, resulting in a shrinkage penalty on the coefficients' magnitudes (78). It penalizes the absolute size of regression coefficients based on the value of the tuning parameter λ. Consequently, LASSO can drive coefficients of irrelevant variables to zero, thereby performing automatic variable selection. When the tuning parameter λ is small, the results essentially converge to least squares estimation. Elastic net (ENET) combines the LASSO method with ridge regression (RR) (79), it includes first and second-order penalty terms on the regression coefficients. Thus, it not only selects the best subset of variables by precisely shrinking some effect estimates to zero through LASSO but also retains a set of highly correlated variables in a RR model with similar effect estimates. For instance, in the Veterans Affairs Normative Aging Study, the use of LASSO enables the selection of $PM_{2.5}$ components related to blood pressure (80). In a recent study, based on ENET penalized regression, two metabolites of phthalates were found to be consistently associated with impaired fetal growth (81). The group-lasso interaction-net method extends LASSO to select bidirectional interaction terms (82), allowing for the simultaneous use of LASSO while controlling the false discovery rate (83). A key characteristic of LASSO is the introduction of an L1 regularization term in estimation, leading to the precise compression of certain coefficients to zero, thereby achieving feature selection. Nevertheless, this excessive sparsity may render the model overly sensitive to noise, and the selected features may prove unstable across different datasets.

4.4 Machine learning approaches

Machine learning (ML) is a research methodology focused on discovering patterns within data and utilizing these patterns to make predictions. Variable selection is a crucial issue in the field of ML, as the predictive performance of models is influenced to some extent by the variables included in the model. The number of variables, variables' correlations, and the inclusion of important variables significantly impact the accuracy and efficiency of predictive models. Therefore, variable selection plays an indispensable role in constructing predictive models. Numerous ML algorithms are currently available for variable selection based on variable importance. Common methods include classification and regression trees (84), random forest (RF) models (85), support vector regression (SVM) (86), K-nearest neighbors (KNN) (87), naive bayes (88), neural networks (89), adaptive boosting (AdaBoost) (90), gradient boosting (GBM) (91), eXtreme gradient boosting (XGBoost) (92), light gradient boosting machine (LightGBM) (93), CatBoost (94), and others are examples of common techniques.

While ML demonstrates effective results, they often face challenges related to interpretability. For instance, models like XGBoost or LightGBM, comprised of N trees, make it difficult to understand how the features of a specific sample influence the final result. To address this issue, SHAP (shapley additive explanations) provides a method for explaining ML, offering detailed and interpretable information about model predictions (95). As the demand for incorporating complex high-dimensional data in environmental health research continues to grow, researchers are increasingly turning to ML. Recent studies have employed various ML such as AdaBoost, SVM, RF, decision tree classifier (DT), and KNN to identify the relationship between heavy metal exposure and coronary heart disease. Integrated with SHAP, these studies explained ML, determining the contributions of heavy metals such as cesium, thallium, antimony, dimethyl arsenic acid, barium, and arsenic acid in urine to the risk of coronary heart disease. This increases the likelihood that coronary heart disease can be detected and treated early (96). Some studies have also used multilayer perceptron, RR, gradient boosting decision tree, voting classifier, and KNN algorithms for generating optimal predictive models for multiple heavy metals causing hypertension. These studies integrated permutation feature importance analysis, Partial Dependence Plots, and SHAP methods into a single process, embedded within ML for model interpretation (97). However, most of the mentioned ML models are used for prediction and require comparisons based on accuracy, sensitivity/recall, specificity, negative predictive value, false positive rate, false negative rate, and F1 score.

4.5 Bayesian variable selection methods

ML algorithms such as RF can provide measures of variable importance for mixed components, but these measures do not succinctly capture the overall magnitude or direction of their associations. Variable selection techniques within the regression framework, such as LASSO, shrink individual regression coefficients to zero. However, these techniques are typically based on relatively simple models of mixed components parameters. To systematically address highly correlated exposures, the BKMR employs a hierarchical variable selection approach. This method can incorporate prior knowledge about the exposure variable/mixed component correlation structure to provide PIPs, as detailed in Section 2.2.

5 Methods for identifying multi-exposure interactions

Although various components in multi-pollutant mixtures may have completely independent effects on health outcomes, in many cases, there may be interactions among components in the mixtures.

Interactions represent the mutual dependence effects of two or more variables and can manifest as synergistic, additive, or antagonistic effects (98). A typical example of interaction is the additive synergistic effect of O_3 and particulate matter on the incidence of cardiovascular diseases (99). In the real world, interactions among various exposure pollutants may exist, and the analysis of these interactions aims to identify and explain their effects. Analyzing and interpreting interactions among multiple exposures can provide a more comprehensive understanding of exposure patterns and identify cooperative effects between specific exposures under certain conditions. In this section we briefly introduce several methods for identifying interaction effects. Figure 1 shows details of the methods for identifying multi-exposure interactions and R packages for their implementation.

5.1 Basic interaction analysis

In interaction factor analysis, analysis of variance (ANOVA) is commonly used to test whether interaction effects among multiple exposures are significant. By comparing the *F*-values or *p*-values of individual factors and interactions, it is possible to determine which factors exhibit interaction effects. Additionally, regression analysis can also be used to explore the interaction effects of exposures, perform significance tests, and create a relationship model between exposure interaction terms and outcomes.

5.2 Bayesian statistical framework

Apart from BKMR, Antonelli utilized a semi-parametric Bayesian sparse prior regression framework to generate variable importance scores for each exposure and each pairwise interaction in the mixture (100).

5.3 Structural equation model

A technique called the structural equation model (SEM) combines particular covariance and regression sets between certain variables into a single coherent model (101). It is used to test and estimate relationships between observed data and latent variables, as well as to assess the fit of theoretical models. SEM combines various techniques such as FA and path analysis, allowing researchers to simultaneously explore complex relationships between multiple variables. In SEM, a measurement model can be constructed to capture measurement errors and covariances among different exposure factors and health outcomes. This aids in accurately measuring these factors and accounting for measurement errors. It is also possible to determine the causal links between various exposure factors and health outcomes using structural models.

SEM is useful for estimating and understanding the network of relationships between variables (latent, observed, and error variables) and it is also employed to estimate the degree of model fit and allows for the presence of measurement errors in independent and dependent variables (102). As researchers turn to modeling multi-pollutant mixtures, SEM is increasingly used to estimate the impact of multi-pollutant mixtures on health (103, 104). For instance, SEM assessed the relationship between respiratory function, tobacco smoke exposure, and volatile organic compound exposure in a nationally representative sample of adolescents, revealing associations between respiratory function and certain types of volatile organic compounds (104). It is noteworthy that a critical feature of SEM analysis is its requirement to meet some basic assumptions of traditional statistical analyses, such as linearity and normality; otherwise, the obtained statistical data may be unreliable.

6 Methods for nonlinear effects

Numerous epidemiological studies have identified nonlinear associations (U-shaped, inverted U-shaped, J-shaped, etc.) between mixed pollutant exposures and health outcomes. For example, the relationship between plasma heavy metals concentrations and type 2 diabetes (105), as well as the association between volatile organic compounds and heart rate variability index (106). Ignoring the potential nonlinearity may result in biased conclusions. Therefore, a better approach is to fit the nonlinear relationship between exposure and outcome. In this section we briefly introduce several methods for estimating nonlinear effects. Figure 1 shows details of the methods for nonlinear effects and R packages for their implementation.

6.1 Spline regression and quantile regression

To overcome the limitations of polynomials, spline methods are often used for curve fitting, employing a piecewise function strategy instead of complex polynomials. One commonly used method in pollution studies is restricted cubic spline (RCS), which fits the curve relationship between a variable and an outcome using restricted cubic spline terms. For instance, Zhou combined RCS with logistic regression to estimate the relationship between typical heavy metal contents (lead, cadmium, mercury, and manganese) in the blood of adults and the metabolic syndrome (107). Similarly, generalized additive models can fit spline line models without specifying nodes automatically, allowing the fitting of spline terms like B-splines, natural splines, thin plates, etc., to control the impact of nonlinear confounding factors. This is achieved by fitting curves of corresponding nonlinear terms of pollutants.

Quantile regression (QR) is a regression analysis method that allows modeling different quantiles of the dependent variable, it can handle issues like non-normal error distribution, heteroscedasticity, and outliers. QR can also be used to fit the nonlinear relationship between pollutants and outcomes. For example, a study used linear regression and QR to investigate the relationship between the increase in concentrations of pollutants (PM_{10}, $PM_{2.5}$, NO_2, and O_3) and changes in birth weight by using linear regression and QR (108).

6.2 Bayesian kernel machine regression

BKMR can also handle nonlinear relationships between exposures. Exposure and result interactions are frequently nonlinear, and BKMR is an efficient way to capture these kinds of nonlinear relationships between contaminants. BKMR has been applied to a dataset on metal exposure and neurodevelopment in Bangladeshi children, indicating the presence of non-additive and nonlinear

exposure-response functions between metals and a summary measure of psychomotor development (109).

6.3 Other methods for nonlinear effects

Methods highlighted in other categories can also be employed to address nonlinear problems, including cross-validated ensemble of kernels (110), TEV, BSSVI, BVSM, MatchAlign, BDS, BKMR-CMA, GriPS, SGP-MPI, BMIM, GL-GPs, Bayesian Tree Ensembles, BKMR-DLM, DLMtree, SPORM, FIN, and FOTP.

7 Holistic approaches to mixture studies

As understanding of environmental pollution deepens and technology advances, the study of a single pollutant can no longer match the analysis of the total health impact of pollutants on the human body. So, researchers are increasingly recognizing the need to analyze complex interactions leading to or exacerbating diseases in mixtures. It is imperative to evaluate the connections between various risk factors and modifying factors from multiple biological dimensions. Similar to exploring the impact of genetic factors on chronic diseases through genome-wide association studies, exposome-wide association studies (EWAS) facilitate the investigation of non-genetic risk factors.

Initially proposed by Wild (111), EWAS can be represented as $P = G + E$, where an individual's phenotype, encompassing health and physical characteristics, is the sum of genetic factors (G) and environmental factors (E) (112). Rappaport also argue that exposure should not be limited to directly encountered chemicals but should consider a broader range of exposures, such as microbial exposure and life stress (113). EWAS provides a conceptual framework to understand the complex network of interactions between genes and the environment, as well as their causal relationships with diseases. It facilitates a holistic analysis of the impact of genetics and the environment on human diseases, including DNA sequences, epigenetic DNA modifications, gene expression, metabolite analysis, and the intricate and dynamic interactions among environmental factors, all of which can influence disease phenotypes.

EWAS research does not solely focus on a single exposure but systematically addresses multiple exposures and their mutual influences, thereby increasing the complexity of the study (114, 115). For instance, a recent study utilized data from the National Health and Nutrition Examination Survey (NHANES) and retained exposure factors, including 75 laboratory variables (clinical and biological biomarkers of environmental chemical exposure) and 64 lifestyle variables (63 dietary variables and 1 physical exercise variable). This study described the associations between body mass index, nutrition, clinical factors, and environmental factors among adolescents (116).

8 Conclusion

In conclusion, the statistical analysis of health effects resulting from the multi-pollutant mixtures is a key challenge in current environmental epidemiological research. In this paper, we review multi-pollutant mixtures statistical methods. It is essential to note that while examining scientific ideas, complementary approaches should be taken into account and statistical methods should be selected with the particular scientific problems in mind. By selecting appropriate statistical methods, considering the combined effects of various pollutants, incorporating interdisciplinary collaboration and emerging technological tools, a more accurate and comprehensive assessment of the impact of mixed environmental pollutant exposure on human health can be achieved. This will contribute to the scientific basis for environmental protection and the formulation of public health policies, promoting sustainable development for human health.

To facilitate the application of the discussed statistical methods, we summarize the advantages and limitations of commonly used statistical methods, corresponding R packages, and the above basic statistical analyses were conducted using the NHANES dataset within gWQS package (Please refer to Figure 1; Supplementary material for statistical analysis code). This serves as a convenient resource for researchers to directly apply these methods.

Author contributions

GZ: Writing – review & editing, Visualization, Writing – original draft. YW: Visualization, Writing – original draft, Writing – review & editing. KC: Visualization, Writing – review & editing. SH: Visualization, Writing – review & editing. TW: Writing – review & editing, Funding acquisition, Supervision.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. This study was supported by the National Natural Science Foundation of China (Grant numbers: 82073674 and 82373692).

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpubh.2024.1377685/full#supplementary-material

References

1. Holgate S. Air pollution is a public health emergency. *BMJ*. (2022) 378:o1664. doi: 10.1136/bmj.o1664

2. Münzel T, Hahad O, Daiber A, Landrigan PJ. Soil and water pollution and human health: what should cardiologists worry about? *Cardiovasc Res*. (2023) 119:440–9. doi: 10.1093/cvr/cvac082

3. Boelee E, Geerling G, van der Zaan B, Blauw A, Vethaak AD. Water and health: from environmental pressures to integrated responses. *Acta Trop*. (2019) 193:217–26. doi: 10.1016/j.actatropica.2019.03.011

4. Tariq M, Iqbal B, Khan I, Khan AR, Jho EH, Salam A, et al. Microplastic contamination in the agricultural soil-mitigation strategies, heavy metals contamination, and impact on human health: a review. *Plant Cell Rep*. (2024) 43:65. doi: 10.1007/s00299-024-03162-6

5. Fu Z, Xi S. The effects of heavy metals on human metabolism. *Toxicol Mech Methods*. (2020) 30:167–76. doi: 10.1080/15376516.2019.1701594

6. Zhang D, Lu S. Human exposure to neonicotinoids and the associated health risks: a review. *Environ Int*. (2022) 163:107201. doi: 10.1016/j.envint.2022.107201

7. Feiteiro J, Mariana M, Cairrão E. Health toxicity effects of brominated flame retardants: from environmental to human exposure. *Environ Pollut*. (2021) 285:117475. doi: 10.1016/j.envpol.2021.117475

8. Yu Y, Quan X, Wang H, Zhang B, Hou Y, Su C. Assessing the health risk of hyperuricemia in participants with persistent organic pollutants exposure – a systematic review and meta-analysis. *Ecotoxicol Environ Saf*. (2023) 251:114525. doi: 10.1016/j.ecoenv.2023.114525

9. He JL, Li GA, Zhu ZY, Hu MJ, Wu HB, Zhu JL, et al. Associations of exposure to multiple trace elements with the risk of goiter: a case-control study. *Environ Pollut*. (2021) 288:117739. doi: 10.1016/j.envpol.2021.117739

10. Vennam S, Georgoulas S, Khawaja A, Chua S, Strouthidis NG, Foster PJ. Heavy metal toxicity and the aetiology of glaucoma. *Eye (Lond)*. (2020) 34:129–37. doi: 10.1038/s41433-019-0672-z

11. Lim JT, Tan YQ, Valeri L, Lee J, Geok PP, Chia SE, et al. Association between serum heavy metals and prostate cancer risk – a multiple metal analysis. *Environ Int*. (2019) 132:105109. doi: 10.1016/j.envint.2019.105109

12. Gomes J, Begum M, Kumarathasan P. Polybrominated diphenyl ether (PBDE) exposure and adverse maternal and infant health outcomes: systematic review. *Chemosphere*. (2024) 347:140367. doi: 10.1016/j.chemosphere.2023.140367

13. Linares V, Bellés M, Domingo JL. Human exposure to PBDE and critical evaluation of health hazards. *Arch Toxicol*. (2015) 89:335–56. doi: 10.1007/s00204-015-1457-1

14. Tian F, Cai M, Li H, Qian Z(M), Chen L, Zou H, et al. Air pollution associated with incident stroke, Poststroke cardiovascular events, and death: a trajectory analysis of a prospective cohort. *Neurology*. (2022) 99:e2474–84. doi: 10.1212/WNL.0000000000201316

15. Joubert BR, Kioumourtzoglou MA, Chamberlain T, Chen HY, Gennings C, Turyk ME, et al. Powering research through innovative methods for mixtures in epidemiology (PRIME) program: novel and expanded statistical methods. *Int J Environ Res Public Health*. (2022) 19:1378. doi: 10.3390/ijerph19031378

16. Hamra GB, Buckley JP. Environmental exposure mixtures: questions and methods to address them. *Curr Epidemiol Rep*. (2018) 5:160–5. doi: 10.1007/s40471-018-0145-0

17. Braun JM, Gennings C, Hauser R, Webster TF. What can epidemiological studies tell us about the impact of chemical mixtures on human health? *Environ Health Perspect*. (2016) 124:A6–9. doi: 10.1289/ehp.1510569

18. Kortenkamp A. Ten years of mixing cocktails: a review of combination effects of endocrine-disrupting chemicals. *Environ Health Perspect*. (2007) 115:98–105. doi: 10.1289/ehp.9357

19. Kortenkamp A. Low dose mixture effects of endocrine disrupters: implications for risk assessment and epidemiology. *Int J Androl*. (2008) 31:233–40. doi: 10.1111/j.1365-2605.2007.00862.x

20. Gibson EA, Goldsmith J, Kioumourtzoglou MA. Complex mixtures, complex analyses: an emphasis on interpretable results. *Curr Environ Health Rep*. (2019) 6:53–61. doi: 10.1007/s40572-019-00229-5

21. Stafoggia M, Breitner S, Hampel R, Basagaña X. Statistical approaches to address multi-pollutant mixtures and multiple exposures: the state of the science. *Curr Environ Health Rep*. (2017) 4:481–90. doi: 10.1007/s40572-017-0162-z

22. Yu L, Liu W, Wang X, Ye Z, Tan Q, Qiu W, et al. A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture. *Environ Pollut*. (2022) 306:119356. doi: 10.1016/j.envpol.2022.119356

23. Hamra GB, Guha N, Cohen A, Laden F, Raaschou-Nielsen O, Samet JM, et al. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. *Environ Health Perspect*. (2014) 122:906–11. doi: 10.1289/ehp/1408092

24. Wolff MS, Engel SM, Berkowitz GS, Ye X, Silva MJ, Zhu C, et al. Prenatal phenol and phthalate exposures and birth outcomes. *Environ Health Perspect*. (2008) 116:1092–7. doi: 10.1289/ehp.11007

25. Varshavsky JR, Zota AR, Woodruff TJ. A novel method for calculating potency-weighted cumulative phthalates exposure with implications for identifying racial/ethnic disparities among U.S. reproductive-aged women in NHANES 2001–2012. *Environ Sci Technol*. (2016) 50:10616–24. doi: 10.1021/acs.est.6b00522

26. Zhang B, Weuve J, Langa KM, D'Souza J, Szpiro A, Faul J, et al. Comparison of particulate air pollution from different emission sources and incident dementia in the US. *JAMA Intern Med*. (2023) 183:1080–9. doi: 10.1001/jamainternmed.2023.3300

27. Carrico C, Gennings C, Wheeler DC, Factor-Litvak P. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. *J Agric Biol Environ Stat*. (2015) 20:100–20. doi: 10.1007/s13253-014-0180-3

28. Tanner EM, Bornehag CG, Gennings C. Repeated holdout validation for weighted quantile sum regression. *MethodsX*. (2019) 6:2855–60. doi: 10.1016/j.mex.2019.11.008

29. Zhang Y, Dong T, Hu W, Wang X, Xu B, Lin Z, et al. Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: comparison of three statistical models. *Environ Int*. (2019) 123:325–36. doi: 10.1016/j.envint.2018.11.076

30. 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. *Int J Environ Res Public Health*. (2021) 18:504. doi: 10.3390/ijerph18020504

31. Wheeler DC, Rustom S, Carli M, Whitehead TP, Ward MH, Metayer C. Bayesian group index regression for modeling chemical mixtures and Cancer risk. *Int J Environ Res Public Health*. (2021) 18:3486. doi: 10.3390/ijerph18073486

32. Gennings C, Curtin P, Bello G, Wright R, Arora M, Austin C. Lagged WQS regression for mixtures with many components. *Environ Res*. (2020) 186:109529. doi: 10.1016/j.envres.2020.109529

33. Bello GA, Arora M, Austin C, Horton MK, Wright RO, Gennings C. Extending the distributed lag model framework to handle chemical mixtures. *Environ Res*. (2017) 156:253–64. doi: 10.1016/j.envres.2017.03.031

34. Bobb JF, Valeri L, Claus Henn B, Christiani DC, Wright RO, Mazumdar M, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics*. (2015) 16:493–508. doi: 10.1093/biostatistics/kxu058

35. Bobb JF, Claus Henn B, Valeri L, Coull BA. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. *Environ Health*. (2018) 17:67. doi: 10.1186/s12940-018-0413-y

36. Chen L, Sun Q, Peng S, Tan T, Mei G, Chen H, et al. Associations of blood and urinary heavy metals with rheumatoid arthritis risk among adults in NHANES, 1999–2018. *Chemosphere*. (2022) 289:133147. doi: 10.1016/j.chemosphere.2021.133147

37. Tan Y, Fu Y, Yao H, Wu X, Yang Z, Zeng H, et al. Relationship between phthalates exposures and hyperuricemia in U.S. general population, a multi-cycle study of NHANES 2007–2016. *Sci Total Environ*. (2023) 859:160208. doi: 10.1016/j.scitotenv.2022.160208

38. Devick KL, Bobb JF, Mazumdar M, Claus Henn B, Bellinger DC, Christiani DC, et al. Bayesian kernel machine regression-causal mediation analysis. *Stat Med*. (2022) 41:860–76. doi: 10.1002/sim.9255

39. Wilson A, Hsu HL, Chiu YM, Wright RO, Wright RJ, Coull BA. Kernel machine and distributed lag models for assessing windows of susceptibility to environmental mixtures in children's health studies. *Ann Appl Stat*. (2022) 16:1090–110. doi: 10.1214/21-aoas1533

40. Yang Z, Guo C, Li Q, Zhong Y, Ma S, Zhou J, et al. Human health risks estimations from polycyclic aromatic hydrocarbons in serum and their hydroxylated metabolites in paired urine samples. *Environ Pollut*. (2021) 290:117975. doi: 10.1016/j.envpol.2021.117975

41. Haddad S, Withey J, Laparé S, Law F, Krishnan K. Physiologically-based pharmacokinetic modeling of pyrene in the rat. *Environ Toxicol Pharmacol*. (1998) 5:245–55. doi: 10.1016/S1382-6689(98)00008-8

42. Viau C, Diakité AS, Ruzgyté A, Tuchweber B, Blais C, Bouchard M, et al. Is 1-hydroxypyrene a reliable bioindicator of measured dietary polycyclic aromatic hydrocarbon under normal conditions? *J Chromatogr B*. (2002) 778:165–77. doi: 10.1016/S0378-4347(01)00465-0

43. Lei B, Zhang K, An J, Zhang X, Yu Y. Human health risk assessment of multiple contaminants due to consumption of animal-based foods available in the markets of Shanghai, China. *Environ Sci Pollut Res*. (2015) 22:4434–46. doi: 10.1007/s11356-014-3683-0

44. Gennings C, Shu H, Rudén C, Öberg M, Lindh C, Kiviranta H, et al. Incorporating regulatory guideline values in analysis of epidemiology data. *Environ Int*. (2018) 120:535–43. doi: 10.1016/j.envint.2018.08.039

45. Mork D, Wilson A. Estimating perinatal critical windows of susceptibility to environmental mixtures via structured Bayesian regression tree pairs. *Biometrics*. (2023) 79:449–61. doi: 10.1111/biom.13568

46. Feldman J, Kowal DR. A Bayesian framework for generation of fully synthetic mixed datasets. *arXiv: Methodology*. (2021). doi: 10.48550/arXiv.2102.08255

47. Kowal DR. Bayesian subset selection and variable importance for interpretable prediction and classification. *J Mach Learn Res*. (2021) 23:108. doi: 10.48550/arXiv.2104.10150

48. Jin B, Peruzzi M, Dunson D B. *Bag of DAGs: flexible & scalable modeling of Spatiotem-poral dependence*, (2021).

49. Mork D, Wilson A. Treed distributed lag nonlinear models. *Biostatistics*. (2020) 23:754–71. doi: 10.1093/biostatistics/kxaa051

50. Ferrari F, Dunson DB. Bayesian Factor analysis for inference on interactions. *J Am Stat Assoc*. (2021) 116:1521–32. doi: 10.1080/01621459.2020.1745813

51. Kowal DR. Fast, optimal, and targeted predictions using parameterized decision analysis. *J Am Stat Assoc*. (2020) 117:1875–86. doi: 10.1080/01621459.2021.1891926

52. Dunson DB, Wu HT, Wu N. Diffusion based Gaussian processes on restricted domains. *arXiv: Methodology*. (2020). doi: 10.48550/arXiv.2010.07242

53. Peruzzi M, Banerjee S, Dunson D B, Finley AO. *Grid-parametrize-Split (GriPS) for improved scalable inference in spatial big data analysis*, (2021).

54. Antonelli J, Wilson A, Coull B. Multiple exposure distributed lag models with variable selection. *Biostatistics*. (2021) 2021:1. doi: 10.1289/isee.2021.O-SY-069

55. Ben Salem K, Ben AA. Principal component analysis (PCA). *Tunis Med*. (2021) 99:383–9. doi: 10.1201/b10345-2

56. Smit LA, Lenters V, Høyer BB, Lindh CH, Pedersen HS, Liermontova I, et al. Prenatal exposure to environmental chemical contaminants and asthma and eczema in school-age children. *Allergy*. (2015) 70:653–60. doi: 10.1111/all.12605

57. Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. *J Am Stat Assoc*. (2006) 101:119–37. doi: 10.1198/016214505000000628

58. Roberts S, Martin MA. Using supervised principal components analysis to assess multiple pollutant effects. *Environ Health Perspect*. (2006) 114:1877–82. doi: 10.1289/ehp.9226

59. Gibson EA, Zhang J, Yan J, Chillrud L, Benavides J, Nunez Y, et al. Principal component pursuit for pattern identification in environmental mixtures. *Environ Health Perspect*. (2022) 130:117008. doi: 10.1289/EHP10479

60. Paatero P, Tapper U. Positive matrix factorization: a non-negative factor model with optimal utilization of error estimates of data values†. *Environmetrics*. (1994) 5:111–26. doi: 10.1002/env.3170050203

61. Krall JR, Strickland MJ. Recent approaches to estimate associations between source-specific air pollution and health. *Curr Environ Health Rep*. (2017) 4:68–78. doi: 10.1007/s40572-017-0124-5

62. Krall JR, Mulholland JA, Russell AG, Balachandran S, Winquist A, Tolbert PE, et al. Associations between source-specific fine particulate matter and emergency department visits for respiratory disease in four U.S. cities. *Environ Health Perspect*. (2016) 125:97–103. doi: 10.1289/EHP271

63. Dai L, Bind MA, Koutrakis P, Coull BA, Sparrow D, Vokonas PS, et al. Fine particles, genetic pathways, and markers of inflammation and endothelial dysfunction: analysis on particulate species and sources. *J Expo Sci Environ Epidemiol*. (2016) 26:415–21. doi: 10.1038/jes.2015.83

64. Siponen T, Yli-Tuomi T, Aurela M, Dufva H, Hillamo R, Hirvonen MR, et al. Source-specific fine particulate air pollution and systemic inflammation in ischaemic heart disease patients. *Occup Environ Med*. (2014) 72:277–83. doi: 10.1136/oemed-2014-102240

65. Roy A, Lavine I, Herring AH, Dunson DB. Perturbed factor analysis: accounting for group differences in exposure profiles. *Ann Appl Stat*. (2021) 15:1386. doi: 10.1214/20-AOAS1435

66. Reid S, Tibshirani R. Sparse regression and marginal testing using cluster prototypes. *Biostatistics*. (2016) 17:364–76. doi: 10.1093/biostatistics/kxv049

67. Steinley D. K-means clustering: a half-century synthesis. *Br J Math Stat Psychol*. (2006) 59:1–34. doi: 10.1348/000711005X48266

68. Ljungman PL, Wilker EH, Rice MB, Austin E, Schwartz J, Gold DR, et al. The impact of multipollutant clusters on the association between fine particulate air pollution and microvascular function. *Epidemiology*. (2016) 27:194–201. doi: 10.1097/EDE.0000000000000415

69. Coker E, Liverani S, Ghosh JK, Jerrett M, Beckerman B, Li A, et al. Multi-pollutant exposure profiles associated with term low birth weight in Los Angeles County. *Environ Int*. (2016) 91:1–13. doi: 10.1016/j.envint.2016.02.011

70. Wold H. Estimation of principal components and related models by iterative least squares. *Multivar Anal*. (1966):1.

71. Mevik B-H, Wehrens R. The pls package: principal component and partial least squares regression in R. *J Stat Softw*. (2007) 18:1–23. doi: 10.18637/jss.v018.i02

72. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *J R Stat Soc Series B Stat Methodol*. (2010) 72:3–25. doi: 10.1111/j.1467-9868.2009.00723.x

73. Agier L, Portengen L, Chadeau-Hyam M, Basagaña X, Giorgis-Allemand L, Siroux V, et al. A systematic comparison of linear regression–based statistical methods to assess Exposome-health associations. *Environ Health Perspect*. (2016) 124:1848–56. doi: 10.1289/EHP172

74. Sinisi SE, Van Der Laan MJ. Deletion/substitution/addition algorithm in learning with applications in genomics. *Stat Appl Genet Mol Biol*. (2004) 3:1–38. doi: 10.2202/1544-6115.1069

75. Sun Z, Tao Y, Li S, Ferguson KK, Meeker JD, Park SK, et al. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health*. (2013) 12:85. doi: 10.1186/1476-069X-12-85

76. Beckerman BS, Jerrett M, Martin RV, van Donkelaar A, Ross Z, Burnett RT. Application of the deletion/substitution/addition algorithm to selecting land use regression models for interpolating air pollution measurements in California. *Atmos Environ*. (2013) 77:172–7. doi: 10.1016/j.atmosenv.2013.04.024

77. Dominici F, Wang C, Crainiceanu C, Parmigiani G. Model selection and health effect estimation in environmental epidemiology. *Epidemiology*. (2008) 19:558–60. doi: 10.1097/EDE.0b013e31817307dc

78. Tibshirani R. Regression shrinkage and selection via the lasso. *J R Stat Soc Series B Stat Methodol*. (1996) 58:267–88. doi: 10.1111/j.2517-6161.1996.tb02080.x

79. Zou H, Hastie TJ. Regularization and variable selection via the elastic net. *J R Stat Soc Series B Stat Methodol*. (2005) 67:301–20. doi: 10.1111/j.1467-9868.2005.00503.x

80. Dai L, Koutrakis P, Coull BA, Sparrow D, Vokonas PS, Schwartz JD. Use of the adaptive LASSO method to identify PM2.5 components associated with blood pressure in elderly men: the veterans affairs normative aging study. *Environ Health Perspect*. (2016) 124:120–5. doi: 10.1289/ehp.1409021

81. Lenters V, Portengen L, Rignell-Hydbom A, Jönsson BAG, Lindh CH, Piersma AH, et al. Prenatal phthalate, Perfluoroalkyl acid, and organochlorine exposures and term birth weight in three birth cohorts: multi-pollutant models based on elastic net regression. *Environ Health Perspect*. (2016) 124:365–72. doi: 10.1289/ehp.1408933

82. Lim M, Hastie T. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat*. (2015) 24:627–54. doi: 10.1080/10618600.2014.938812

83. Huang H. Controlling the false discoveries in LASSO. *Biometrics*. (2017) 73:1102–10. doi: 10.1111/biom.12665

84. Loh WY. Classification and regression trees. *Wiley Interdiscip Rev Data Min Knowl Discov*. (2011) 1:14–23. doi: 10.1002/widm.8

85. Biau G. Analysis of a random forests model. *J Mach Learn Res*. (2012) 13:1063–95.

86. Smola AJ, Schölkopf B. A tutorial on support vector regression. *Stat Comput*. (2004) 14:199–222. doi: 10.1023/B:STCO.0000035301.49549.88

87. Peterson LE. K-nearest neighbor. *Scholarpedia*. (2009) 4:1883. doi: 10.4249/scholarpedia.1883

88. Webb GI, Keogh E, Miikkulainen R. Naïve Bayes. *Encycl Mach Learn*. (2010) 15:713–4. doi: 10.1007/978-0-387-30164-8_576

89. Bishop CM. Neural networks and their applications. *Rev Sci Instrum*. (1994) 65:1803–32. doi: 10.1063/1.1144830

90. Margineantu D D, Dietterich T G. *Pruning adaptive boosting. ICML*, (1997): 211–218.

91. Friedman JH. Stochastic gradient boosting. *Comput Stat Data Anal*. (2002) 38:367–78. doi: 10.1016/S0167-9473(01)00065-2

92. Chen T, Guestrin C. Xgboost: a scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining, (2016): 785–794.

93. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. Lightgbm: a highly efficient gradient boosting decision tree. *Adv Neural Inf Proces Syst*. (2017) 30:3146–3154. doi: 10.5555/3294996.3295074

94. Prokhorenkova L, Gusev G, Vorobev A, Dorogush AV, Gulin A. CatBoost: unbiased boosting with categorical features. *Adv Neural Inf Proces Syst*. (2018) 31. doi: 10.48550/arXiv.1706.09516

95. Lundberg SM, Lee S-I. A unified approach to interpreting model predictions. *Adv Neural Inf Proces Syst*. (2017) 30:4768–4777. doi: 10.48550/arXiv.1705.07874

96. Li X, Zhao Y, Zhang D, Kuang L, Huang H, Chen W, et al. Development of an interpretable machine learning model associated with heavy metals' exposure to identify coronary heart disease among US adults via SHAP: findings of the US NHANES from 2003 to 2018. *Chemosphere*. (2023) 311:137039. doi: 10.1016/j.chemosphere.2022.137039

97. Li W, Huang G, Tang N, Lu P, Jiang L, Lv J, et al. Effects of heavy metal exposure on hypertension: a machine learning modeling approach. *Chemosphere*. (2023) 337:139435. doi: 10.1016/j.chemosphere.2023.139435

98. Mauderly JL, Samet JM. Is there evidence for synergy among air pollutants in causing health effects? *Environ Health Perspect*. (2009) 117:1–6. doi: 10.1289/ehp.11654

99. Liu C, Chen R, Sera F, Vicedo-Cabrera AM, Guo Y, Tong S, et al. Interactive effects of ambient fine particulate matter and ozone on daily mortality in 372 cities: two stage time series analysis. *BMJ*. (2023) 383:e075203. doi: 10.1136/bmj-2023-075203

100. Antonelli J, Mazumdar M, Bellinger DC, Christiani D, Wright R, Coull B. Estimating the health effects of environmental mixtures using Bayesian semiparametric regression and sparsity inducing priors. *Ann Appl Stat*. (2017) 14:275–75. doi: 10.48550/arXiv.1711.11239

101. Davalos AD, Luben TJ, Herring AH, Sacks JD. Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. *Ann Epidemiol*. (2017) 27:145–153.e1. doi: 10.1016/j.annepidem.2016.11.016

102. Tomarken AJ, Waller NG. Structural equation modeling: strengths, limitations, and misconceptions. *Annu Rev Clin Psychol*. (2005) 1:31–65. doi: 10.1146/annurev.clinpsy.1.102803.144239

103. Stein CM, Morris NJ, Nock NL. Structural equation modeling. *Methods Mol Biol*. (2012) 850:495–512. doi: 10.1007/978-1-61779-555-8_27

104. Shook-Sa BE, Chen DG, Zhou H. Using structural equation modeling to assess the links between tobacco smoke exposure, volatile organic compounds, and respiratory function for adolescents aged 6 to 18 in the United States. *Int J Environ Res Public Health*. (2017) 14:1112. doi: 10.3390/ijerph14101112

105. Shan Z, Chen S, Sun T, Luo C, Guo Y, Yu X, et al. U-shaped association between plasma manganese levels and type 2 diabetes. *Environ Health Perspect*. (2016) 124:1876–81. doi: 10.1289/EHP176

106. Wang B, Cheng M, Yang S, Qiu W, Li W, Zhou Y, et al. Exposure to acrylamide and reduced heart rate variability: the mediating role of transforming growth factor-β. *J Hazard Mater*. (2020) 395:122677. doi: 10.1016/j.jhazmat.2020.122677

107. Zhou J, Meng X, Deng L, Liu N. Non-linear associations between metabolic syndrome and four typical heavy metals: data from NHANES 2011–2018. *Chemosphere*. (2022) 291:132953. doi: 10.1016/j.chemosphere.2021.132953

108. Lamichhane DK, Lee S-Y, Ahn K, Kim KW, Shin YH, Suh DI, et al. Quantile regression analysis of the socioeconomic inequalities in air pollution and birth weight. *Environ Int*. (2020) 142:105875. doi: 10.1016/j.envint.2020.105875

109. Valeri L, Mazumdar MM, Bobb JF, Claus Henn B, Rodrigues E, Sharif OIA, et al. The joint effect of prenatal exposure to metal mixtures on neurodevelopmental outcomes at 20–40 months of age: evidence from rural Bangladesh. *Environ Health Perspect*. (2017) 125:067015. doi: 10.1289/EHP614

110. Liu JZ, Deng W, Lee J, Lin PD, Valeri L, Christiani DC, et al. A cross-validated ensemble approach to robust hypothesis testing of continuous nonlinear interactions: application to nutrition-environment studies. *J Am Stat Assoc*. (2022) 117:561–73. doi: 10.1080/01621459.2021.1962889

111. Wild CP. Complementing the genome with an "exposome": the outstanding challenge of environmental exposure measurement in molecular epidemiology. *Cancer Epidemiol Biomarkers Prev*. (2005) 14:1847–50. doi: 10.1158/1055-9965.EPI-05-0456

112. Wild CP. The exposome: from concept to utility. *Int J Epidemiol*. (2012) 41:24–32. doi: 10.1093/ije/dyr236

113. Rappaport SM, Smith MT. Epidemiology. Environment and disease risks. *Science*. (2010) 330:460–1. doi: 10.1126/science.1192603

114. Khoury MJ, Wacholder S. Invited commentary: from genome-wide association studies to gene-environment-wide interaction studies--challenges and opportunities. *Am J Epidemiol*. (2009) 169:227–30. doi: 10.1093/aje/kwn351

115. Thomas D. Gene--environment-wide association studies: emerging approaches. *Nat Rev Genet*. (2010) 11:259–72. doi: 10.1038/nrg2764

116. Haddad N, Andrianou X, Parrish C, Oikonomou S, Makris KC. An exposome-wide association study on body mass index in adolescents using the National Health and nutrition examination survey (NHANES) 2003–2004 and 2013–2014 data. *Sci Rep*. (2022) 12:8856. doi: 10.1038/s41598-022-12459-z