Environmental Pollution 306 (2022) 119356

Contents lists available at ScienceDirect

Environmental Pollution

journal homepage: www.elsevier.com/locate/envpol

A review of practical statistical methods used in epidemiological studies to estimate the health effects of multi-pollutant mixture☆

Linling Yu [a,b,1], Wei Liu [a,b,1], Xing Wang [a,b,1], Zi Ye [a,b], Qiyou Tan [a,b], Weihong Qiu [a,b], Xiuquan Nie [a,b], Minjing Li [a,b], Bin Wang [a,b,2], Weihong Chen [a,b,*,2]

[a] Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

[b] Key Laboratory of Environment and Health, Ministry of Education & Ministry of Environmental Protection, and State Key Laboratory of Environmental Health (Incubating), School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China

ARTICLE INFO

Keywords:
Multi-pollutant mixture
Health effect
Environmental epidemiology
Statistical method
Bayesian kernel machine regression

ABSTRACT

Environmental risk factors have been implicated in adverse health effects. Previous epidemiological studies on environmental risk factors mainly analyzed the impact of single pollutant exposure on health, while in fact, humans are constantly exposed to a complex mixture consisted of multiple pollutants/chemicals. In recent years, environmental epidemiologists have sought to assess adverse health effects of exposure to multi-pollutant mixtures based on the diversity of real-world environmental pollutants. However, the statistical challenges are considerable, for instance, multicollinearity and interaction among components of the mixture complicate the statistical analysis. There is currently no consensus on appropriate statistical methods. Here we summarized the practical statistical methods used in environmental epidemiology to estimate health effects of exposure to multi-pollutant mixture, such as Bayesian kernel machine regression (BKMR), weighted quantile sum (WQS) regressions, shrinkage methods (least absolute shrinkage and selection operator, elastic network model, adaptive elastic-net model, and principal component analysis), environment-wide association study (EWAS), etc. We sought to review these statistical methods and determine the application conditions, strengths, weaknesses, and result interpretability of each method, providing crucial insight and assistance for addressing epidemiological statistical issues regarding health effects from multi-pollutant mixture.

1. Introduction

The environment is the basis for human survival. Some unfavorable factors in the environment such as chemical pollutants may have potential adverse effects on human health. The impact of these environmental chemicals on health has always been a major challenge for the public health or medical community. Most epidemiological studies on the health effects from environmental pollutant exposure have focused on one single factor, such as nitrogen dioxide (Meng et al., 2021), acrylamide (Wang et al., 2020b), or single metal (Xiao et al., 2021; Zhou et al., 2020). However, humans are exposed to a complex multi-pollutant mixture rather than a single pollutant in the real world. The adverse health effects of multi-pollutant mixture have attracted great attention from the academic community. In 2020, the US National Institute of Environmental Health Science had called the Powering Research through Innovative Methods for Mixture in Epidemiology (PRIME) program to explore statistical methods that could be used in studies of multi-pollutant mixture (NIH, 2020). There is an urgent need for multi-pollutant mixture studies to inform the formulation of more sustainable environmental regulations (Kortenkamp and Faust, 2018).

Traditionally, in order to examine the health effects of multi-pollutant mixture, two or more pollutants are commonly included in a regression model to estimate the association attributable to each single pollutant after adjusting for potential confounding factors (Ranganathan

☆ This paper has been recommended for acceptance by Dr. Da Chen.

* Corresponding author. Department of Occupational and Environmental Health, School of Public Health, Tongji Medical College, Huazhong University of Science and Technology, Wuhan, Hubei 430030, China.
E-mail address: wchen@mails.tjmu.edu.cn (W. Chen).

[1] These authors contributed equally to this work and should be considered co-first authors.

[2] These authors contributed equally to this article.

https://doi.org/10.1016/j.envpol.2022.119356
Received 7 January 2022; Received in revised form 11 April 2022; Accepted 21 April 2022
Available online 27 April 2022

et al., 2017). This method may lead to highly unstable results when incorporating two or more highly correlated (collinear) pollutants (Ranganathan et al., 2017). To produce a reduced model, effects from multi-pollutant mixture exposure are also often assessed by using a step-by-step algorithm (such as forward stepwise regression or backward stepwise regression) (Billionnet et al., 2012). In such statistical methods, multiple pollutants were incorporated into a single model, but their effects were considered independently, which may well lead to errors in estimation and interpretation (Billionnet et al., 2012). Therefore, the development of statistical methods for assessing the health effects of multi-pollutant mixture is a valuable topic to be explored.

The interdisciplinary approaches have been deemed to be a well solution for this problem as the academic discussion deepens. Overlaps of multidiscipline (environmental science, epidemiology, biostatistics, etc.) in solving the problem of multi-pollutant mixture exposure have sparked numerous new ideas. In recent years, various interdisciplinary methods, such as the Bayesian method and biometric method, have been developed to solve the analytic problem of assessing health effects from exposure to multi-pollutant mixture (Bobb et al., 2015; Zheng et al., 2020a). In this review, we focused on practical statistical methods that can be used to address the potential health effects from multi-pollutant mixture exposure, and proper examples were therewith presented. The characteristics, application conditions, advantages, limitations, and result interpretability of each statistical method were summarized.

2. Characteristics of multi-pollutant mixture data

The characteristics of multi-pollutant mixture data and the possible relationship of various pollutants were described below and graphically shown in Fig. 1.

2.1. High-dimensionality

Epidemiological researchers are increasingly evaluating exposure variables in the context of numerous environmental pollutants and multi-biomarkers, which opens up new opportunities to identify multiple health determinants; but high-dimensional data also bring tremendous analytical challenges that are intractable to evaluate with traditional statistical methods. The curse of dimensionality is the biggest problem encountered in analyzing high-dimensional data (Patel et al., 2018). The complexity and time cost of analyzing high-dimensional data increase exponentially as the number of dimensions increases (Chattopadhyay and Lu, 2019).

2.2. Multicollinearity

Multicollinearity is the most concerned statistical challenge in the study of multi-pollutant mixture. Multicollinearity could be induced when simultaneously analyzing correlated pollutants with similar sources, exposure pathways, and/or metabolic processes (Vrijheid et al., 2016). For example, ambient air fine particulate matter (particulate matter with diameter $\leq$2.5 μm, PM2.5) and ozone are strongly correlated, and their formation is coupled because of similar sources, secondary nature, and interactions of their precursors (Kampa and Castanas, 2008). The study of each of them independently may lead to biased conclusions. In fact, the association between ozone and respiratory disease may be caused by $PM_{2.5}$ or other pollutants correlated with ozone. Multicollinearity can result in an inflated standard error in a generalized linear model (GLM), leading to unstable coefficient estimates and difficulty in assessing the relative importance of individual pollutant (Lazarevic et al., 2019). Adding interaction and nonlinear terms into the model may worsen the problem, and in extreme cases, the maximum likelihood technique may not converge to a solution (Schisterman et al., 2017).

2.3. Interaction

Although substances may have completely independent actions, there are many cases in which two substances may interact at the same time. Interaction represents the interdependence effect of two or more variables (Mauderly and Samet, 2009). The interaction among pollutants can be synergistic, additive, or antagonistic. The synergistic effect can be detected when the overall effect of combined multi-pollutants is significantly greater than a simple addition of their separate predicted effects (Mauderly and Samet, 2009). The synergistic effect between asbestos and cigarette smoking on lung cancer is a well-documented example (Ngamwong et al., 2015). The antagonistic effect refers to the effect of the combination less than the sum of individual effects. When two or more pollutants combine to produce an effect equal to the sum of the effect of individual components, an additive effect occurs. The assessment of interactions increases the methodological complexity of mixture analysis. The traditional way to test for the existence of interaction is by including the product terms of the potentially interacting variables in the model (Lazarevic et al., 2019). However, the statistical power of such a statistical test is low, especially when examining interactions among three or more pollutants (Bellavia et al., 2019). The interaction effects of several pollutants may be imprecisely characterized unless there is a strong interaction or abundant data (Bellavia et al., 2019).

2.4. Nonlinear effects

Nonlinear relationship between multi-pollutant mixture and outcome was found in a large number of epidemiological studies. The

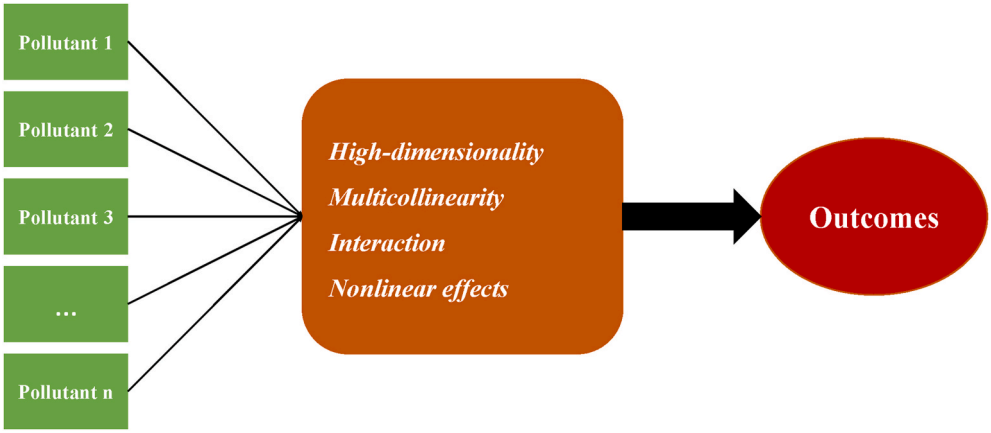

Fig. 1. Characteristics of multi-pollutant mixture data.

association between plasma manganese concentrations and type 2 diabetes was found to be nonlinear (Shan et al., 2016). Similar findings have been shown for the associations between some volatile organic compounds (VOCs) and specific heart rate variability indices (Wang et al., 2020a), blood glucose (Xu et al., 2021), and asthma (Billionnet et al., 2011). In such conditions, commonly used standard regression models are hardly applicable because there are nonlinear (U-shaped, inverted U-shaped, J-shaped, etc.) rather than linear associations between multi-pollutant mixture and specific outcomes. Since multi-pollutants have common complex and nonlinear effects on outcomes, exploring the interaction among multi-pollutant mixture without considering the potential nonlinear effects of each component can lead to biased conclusions. For example, given two exposure variables that interact and have U-shaped relationships with the outcomes, assuming linearity may simply yield an antagonistic effect of two exposure variables on the outcome, although in fact such antagonistic effect may only occur when the exposure level is in a specific range, while no interaction or even synergistic effect may be exist when the exposure variable level is within another specific range. In other words, in the case of nonlinear relationship between exposure variable and the outcome, the specific interaction of two exposure variables on the outcome are likely to be varied according to the different levels of exposure variables, while linearity assumption overlooks the variation of interaction and yields biased or spurious interaction.

3. Statistical methods

Over the years, in order to assess the health effects of exposure to the multi-pollutant mixture as accurately as possible, a wide variety of statistical techniques have been developed with consideration of the abovementioned characteristics of environmental mixture data. A review of frequently used practical statistical methods was elaborated below and summarized in Table 1 and Table 2.

3.1. Bayesian kernel machine regression (BKMR)

BKMR model, first proposed by Bobb and colleagues in 2015, combines Bayesian and statistical learning methods to regress an exposure-response function iteratively by a Gaussian kernel function (Bobb et al., 2015). The BKMR model is given by:

$$Y_i = h(z_{i1}, \ldots, z_{iM}) + \beta z_i + e_i$$

where Y_i denotes the response for individual i ($i = 1, \ldots, n$), z_{iM} is the mth exposure variable, h() is the unknown exposure-response function to be estimated, β represents coefficient, and the residuals $e_i \sim N(0, \sigma^2)$ are assumed to be independent and identically normally distributed with a common variance.

BKMR was developed specifically to study the effect of multi-pollutant mixture on outcome. This method not only is capable of assessing the overall mixture effect and the effects of individual components within the mixture but also permits assessments of potential interactions among the components (Bobb et al., 2018). Posterior inclusion probabilities (PIPs), an index in a range of 0 (least important) to 1 (most important) yielded by BKMR (Bobb et al., 2018), can identify relative important mixture components through variable selection with PIP $\geq$0.5 (Coker et al., 2018). BKMR can be used to investigate possible three-way interactions by visualizing the bivariate exposure-response function for a third exposure fixed at different quantiles (Bobb, 2017). Valeri L et al. used BKMR to assess the joint effects of in utero exposure to arsenic, manganese, and lead on children's neurodevelopment, which is among the first designed to address the health effects of multi-pollutant mixture by using BKMR (Valeri et al., 2017). Recently, we have used the BKMR model to assess the effect of multiple phthalate co-exposure on liver function (Yu et al., 2021b). Collectively, BKMR has been widely used in data application scenarios common in environmental health, including continuous and binary outcomes and repeated measures (Hou et al., 2020; Yu et al., 2021b; Zhang et al., 2019).

Although BKMR can largely address the scientific questions on features of the health effects of multi-pollutant mixture, there are still some limitations. First, exposure variables are limited to continuous variables when analyzed by using the BKMR model (Bobb et al., 2018). Second, the magnitude of the PIPs is sensitive to the choice of tuning parameters. Third, caution is warranted in interpreting results since they may obscure potentially complex features of the data. An apparent null overall association may be observed if parts of the components within the mixture are positively associated while the other parts are negatively associated with an outcome in a similar magnitude (Bobb et al., 2018). Fourth, formal procedures for assessing the statistical significance of pollutants mixture interactions are not readily available yet, and interaction estimates are largely based on subjective interpretation of two-way dose-responses figures (Bobb et al., 2018). Nonetheless, it is feasible to evaluate interactions by combining other statistical methods, such as generalized additive models (GAMs) (Shah-Kulkarni et al., 2020; Zheng et al., 2020b).

3.2. Weighted quantile sum (WQS), quantile g-computation, and grouped weighted quantile sum (GWQS)

WQS regression model can be used to combine highly correlated exposure variables into a weighted additive index. The index is used to estimate an overall mixture effect through a bootstrap resampling procedure based on a training data set (a random subset of the data) (Carrico et al., 2015). Briefly, for each pollutant grouping, WQS regression outputs an empirically weighted sum of the quartiles of pollutant group components. The weighted index is included in a regression model, with adjustment for covariates. When the corresponding regression coefficient is statistically significant, the weights (constrained to sum to one) estimated by maximum likelihood are used to identify important components. (Carrico et al., 2015). The WQS model takes the form as follow:

$$g(\mu) = \beta_0 + \beta_1 WQS + z'\varphi$$

$$WQS = \sum w_i q_{ij}$$

where, g(μ) is a nonlinear link function that can be generalized to continuous, binary, and other distributions; β_0 is the intercept; β_1 is the parameter estimate for the co-exposure index; *WQS* represents the weighted index for the pollutants of interest. z' and φ is vector of covariates and regression coefficients, respectively; w_i is the unknown weight for the component; and q_{ij} is the quantile rank assigned to each subject per variable.

WQS regression model possesses the advantage of incorporating multi-pollutants into a single score, thus avoiding any issue of overfitting and collinearity (Bellavia et al., 2019). It estimates the effect of the multi-pollutant mixture on the outcome as well as the weights of mixture components to determine the relative contributions (Carrico et al., 2015). The weights are constrained to be in a range of 0–1 and summed to 1. The weight is evaluated according to the impact of each component on outcome, the greater the impact, the higher the weight. WQS regression has been increasingly used in environmental epidemiological studies to elaborate the effects of multi-pollutant mixture on metabolic diseases, respiratory diseases, and neurological disorders (Araki et al., 2020; Tanner et al., 2020; Zhang et al., 2019). However, some substantial limitations of the WQS regression method should be noted. One is that WQS regression assumes that the components of multi-pollutant mixture have linear additive effects (Keil et al., 2020). Besides, all components are constrained to have the same (despite the truth may be entirely different) direction of associations (positive or inverse) with the outcome (Keil et al., 2020; Wheeler et al., 2021).

Table 1
Summary of statistical methods.

Method	Brief characterization	Goal	Strength	Limitation	The applicable types of exposure variables	Implementation method in R software	Research example
Bayesian kernel machine regression (BKMR)	A nonparametric method that combines Bayesian and statistical learning methods to regress an exposure-response function iteratively by a Gaussian kernel function.	Simultaneously estimate multi-pollutant mixture health effects while identifying the importance of individual components within the mixture.	Robust to address multicollinearity; address the problem of high-dimensionality; estimate interactions among components of a mixture; investigate possible three-way interactions.	The magnitude of the PIP is sensitive to the choice of tuning parameters.	continuous	*bkmr* package	Yu et al. (2021b), Valeri et al. (2017)
Weighted quantile sum (WQS)	Combine pollutants into a weighted additive index, which is used to estimate an overall mixture effect through a bootstrap resampling procedure based on a training data set (a random subset of the data)	Estimate multi-pollutant mixture health effect and rank important mixture components.	Avoid any issue of overfitting and collinearity; with lower mean-squared error and higher specificity than shrinkage methods.	Data information loss due to data transformation to quantiles; all chemicals are constrained to have the same direction associated with the outcome.	continuous	*wqs* package *gWQS* package	Zhang et al. (2019), Daniel et al. (2020)
Quantile g-computation	A new approach combining WQS and g-computation.	Estimate multi-pollutant mixture health effect and rank important constituents.	Allow inference on multi-pollutant mixture effects that is unbiased with proper confidence interval coverage at sample sizes.	Data information loss due to data transformation to quantiles; the marginal structural model may not adequately capture the dose-response function when the underlying model is not smooth.	continuous	*qgcomp* package	Keil et al. (2020), Yu et al. (2021a)
Grouped weighted quantile sum (GWQS)	A WQS-based method, which allows for multiple groups of pollutants to be considered in the model and the components of the multi-pollutant mixture are allowed to have different magnitudes and directions.	Estimate multi-pollutant mixture health effect and rank important chemicals.	With better power to detect true exposure effects compared to WQS; with better sensitivity and specificity than WQS for identifying important chemicals.	Data information loss due to data transformation to quantiles.	continuous	*groupWQS* package	Wheeler et al. (2021), Navas-Acien et al. (2021)
Least absolute shrinkage and selection operator (LASSO)	A shrinkage (penalized regression) method that pushes minimums of coefficients to exactly zero via directly shrinking the sum of the absolute values of coefficients.	Identify important mixture components.	Robust to address multicollinearity; with lower coefficient variance than ordinary least squares regression.	Only one chemical may be chosen and the others were dropped among a set of highly correlated chemicals, resulting in an impression that the dropped chemicals are not associated with the outcome; only identify important mixture components that have a linear relationship with the outcome.	continuous and categorical	*glinternet* package *glmnet* package *lars* package	McEligot et al. (2020) , Zhou et al. (2021)
Elastic network model (ENM)	A method with penalty parameters of both ridge regression and LASSO.	Identify important mixture components.	Robust to address multicollinearity; address the problem of high-dimensionality; greater prediction accuracy than LASSO.	Post-selection statistical inference tools are needed for inference, including yielding confidence intervals; the results may be troubled by false positives.	continuous and categorical	*glmnet* package	Yitshak-Sade et al. (2020), Navas-Acien et al. (2021)
Adaptive elastic-net model (AENM)	An adaptive version of ENM.	Identify important mixture components.	Robust to address multicollinearity; address the problem of high-dimensionality; greater prediction	The contribution of a single pollutant cannot be calculated; extrapolation is limited; higher-order	continuous and categorical	*gcdnet* package	Navas-Acien et al. (2021) , Wang et al. (2020)

(*continued on next page*)

Table 1 (*continued*)

Method	Brief characterization	Goal	Strength	Limitation	The applicable types of exposure variables	Implementation method in R software	Research example
			accuracy than LASSO; lower false positive than ENM by limiting the retention of variables with different weights.	interaction forms and nonlinear relationships other than square terms are not considered.			
Principal component analysxis (PCA)	An unsupervised machine learning approach for data dimensionality reduction.	Replace a large number of related variables with a smaller set of unrelated variables while preserving as much information as possible about the original variables; identify important mixture components; multi-pollutant mixture health effects estimation.	No parameter limit; noise removal; make the results easy to understand.	Exposure variables are limited to continuous variables; components may not have biologically-relevant interpretations; the establishment of the number of principal components is not absolutely fixed.	continuous	*princomp* function *prcomp* function *psych* package	Roberts and Martin, (2006), Zhou et al. (2016)
Environment-wide association study (EWAS)	A comprehensive analysis that screens numerous environmental factors to determine which factors are associated with the interested outcomes.	Identify outcomes-associated environmental risk factors, which are consequently considered relatively important.	Hundreds of chemicals (or even any environmental variable) can be tested simultaneously, high dimensional data are encouraged.	No accepted criteria for quality control and inclusion of environmental factors	continuous and categorical	*qqman* package; *ggplot2* package	Uche et al. (2020), Lee et al. (2020)
Classification and regression tree (CART)	A non-parametric statistical method that repeatedly partitions data into subsets with similar internal features.	Classification; identification of important mixture components.	Deal with missing values; immune to outliers; without additivity or linearity assumption in the assessment of the exposure-response relationship; no assumption about distribution for variables; allow multicollinearity.	Not suitable for small data sets; somewhat unstable.	continuous and categorical	*rpart* package	Gass et al. (2014), Yang et al. (2020)
Random forest regression (RFR)	An ensemble machine learning method developed from CART.	Classification; identification of important mixture components.	No data distribution assumptions are required; the interaction and high-order relationships among predictors is able to be detected; deal with the nonlinear problem and certain data missing.	Generate highly predictive models that are difficult to interpret.	continuous and categorical	*randomForest* package	Lin et al. (2021), Smit et al. (2015)
Structural equation modeling (SEM)	A methodology to integrate specific sets of covariance and regressions among specified variables into a single cohesive model.	Identify the important mixture components; estimate multi-pollutant mixture health effects.	With the ability to simultaneously model multiple outcomes and multiple exposure variables; estimate the degree of model fitting and allow measurement errors in independent and dependent variables.	Require greater subject matter knowledge for correct specification than single exposure-outcome models.	continuous and categorical	*Lavaan* package	Huang et al. (2017), Wang et al. (2019)

To compensate for such limitations, researchers have developed a series of improved methods to estimate the effect of multi-pollutant mixture exposure. Recently, Keil et al. demonstrated quantile g-computation, a new approach combining WQS and g-computation (Keil et al., 2020). It has the advantages of the simplicity of WQS inference and the flexibility of g-computation and allows for different directions of associations as well as nonlinearity and non-additivity of the effects of the individual pollutant and the mixture as a whole. Quantile g-computation is able to estimate the parameters of a marginal structural model, however, which may not adequately capture the dose-response function when the underlying model is not smooth (a limitation) (Keil et al., 2020). Xu et al. studied the relationship between metal mixtures and hypertension by using quantile g-computation (Van den Dries et al., 2021). Besides, grouped weighted quantile sum (GWQS) regression was proposed by Wheeler et al., in 2021. Multiple groups of pollutants are allowed to be included in the GWQS regression model, and the components of the multi-pollutant mixture are allowed to have different magnitudes and directions (Wheeler et al., 2021). Wheeler et al. used GWQS to analyze the associations between childhood leukemia and 49 chemicals, which were put into 6 groups (Wheeler et al.,

Table 2
Summary of statistical methods to solve the challenges of multi-pollutant mixture analysis.

Methods	Reduce data dimensionality	Address multicollinearity	Evaluate interaction	Evaluate nonlinearity	Identify important mixture component
Bayesian kernel machine regression (BKMR)	✓	✓	✓	✓	✓
Weighted quantile sum (WQS)	✓	✓	×	×	✓
Quantile g-computation	✓	✓	×	✓	✓
Grouped weighted quantile sum (GWQS)	✓	✓	×	×	✓
Least absolute shrinkage and selection operator (LASSO)	✓	✓	×	×	✓
Elastic network model (ENM)	✓	✓	✓	✓	✓
Adaptive elastic-net model (AENM)	✓	✓	✓	✓	✓
Principal component analysis (PCA)	✓	✓	×		✓
Environment-wide association study (EWAS)	✓	NA	×	×	✓
Classification and regression tree (CART)	✓	✓	✓	NA	✓
Random forest regression (RFR)	✓	✓	✓	✓	✓
Structural equation modeling (SEM)	×	✓	✓	×	✓

NA: not applicable.

2021). Nevertheless, it is worth noting that partial data information can be lost when performing WQS, quantile g-computation, and GWQS as they transform continuous exposure variables into quantiles, which is also a common limitation of these WQS-based methods.

3.3. Shrinkage methods

The mainly reviewed shrinkage methods in this section were least absolute shrinkage and selection operator (LASSO), elastic network model (ENM), adaptive elastic-net model (AENM), and principal component analysis (PCA). Indeed, in the usual analysis, the important components in the mixture are determined, instead of including all the components of the analysis. Some frequentist shrinkage methods display good abilities to identify important components in a highly correlated setting by imposing multiple penalties on the size of the coefficients, and this process is called regularization. Ridge regression shrinks the sum of the squares of the coefficients, and the coefficients of "unimportant" components (weakly associated with the outcome) decrease toward zero (Hoerl and Kennard, 1970). The ridge regression objective function is given by:

$$J_\omega = \min_{\omega} \left\{ ||X_\omega - y||^2 + \alpha ||\omega||^2 \right\}$$

The higher the value of the coefficient α, the more significant the effect of the penalty term. Ridge regression is where the penalty term is the l_2 norm of the parameter. A more radical method is the LASSO regression, where the penalty term is the l_1 norm of the parameter. And LASSO regression pushes the minimums of coefficients to exactly zero via directly shrinking the sum of the absolute values of coefficients (Tibshirani, 1996). A study assessed associations between 29 variables and breast cancer using LASSO regression (McEligot et al., 2020). However, LASSO model is a fitted linear model, and pollutants that have a nonlinear relationship with the outcome should be excluded. Only one pollutant could be chosen and the others may be dropped among a set of highly correlated chemicals, giving the impression that the dropped pollutants are not associated with the outcome. The ENM includes penalty parameters of both ridge regression and LASSO and maintains the regularization of ridge regression based on LASSO, resulting in a good performance in variable selection (Zou and Hastie, 2005).

In the comparative analysis with single-pollutant ordinary least squares regression models, ENM has been shown to well overcome the multicollinearity of three chemicals (Lenters et al., 2016). ENM has distinct superiority in components selection from large amounts of chemicals and in performing grouped selection by assigning coefficients of similar magnitude. The beauty of ENM is that it inherits both conservatism of ridge regression and sensitivity of LASSO. For ENM, cross-validation is a common-sense method for determining penalty parameters, but the results may be troubled by false positive. However, the combination of cross-validation and covariance test could avoid suffering from inflated false discovery rate. Similar to traditional penalty regression, the component selection was solely based on statistical association. A lot of studies chose to estimate the coefficients using the traditional regression model after component selection to overcome such drawbacks, but the coefficients obtained by ENM and the traditional regression model may be opposite (Aitken et al., 2021; Yitshak-Sade et al., 2020). The approach combining ENM and traditional regression that ignores the uncertainty of component selection.

The adaptive version of ENM is AENM, where two penalty parameters were ascertained typically based on 5-fold cross-validation for minimal prediction errors to define beta coefficients. Compared with ENM, the component selection is more stringent in AENM, which takes collinearity into account and ensures larger coefficients are penalized less, whereas smaller coefficients are shrunk faster to zero, resulting in a much lower false positive rate (Zou and Zhang, 2009). Wang et al. identified significant components associated with waist circumference among 18 blood and urinary heavy metals of U.S. adults by using AENM (Wang et al., 2018). A total of 18 main effects, 18 squared effects (i.e., nonlinear effects), and 153 pairwise interactions of 18 heavy metals served as candidates for contributing factors of waist circumference were screened by AENM. These screened components, which indicated excellent associations with outcomes, were retained in the model to calculate an environmental risk score (ERS) (Wang et al., 2018). Then, a traditional regression model including ERS (explanatory variable), health outcome, and confounders was fitted to assess the health effects of the mixture. Limitations of AENM include failure to estimate the quantitative contribution of a single chemical to the outcome or ERS. Besides, the calculated ERS is more suitable for the current research population with poor extrapolation. Moreover, other higher-order forms of interaction and nonlinear relationships in addition to square terms are not taken into consideration in AENM.

PCA reduces the high-dimensional exposure data into several orthogonal components, which can be used in regression models, and therefore eliminates multicollinearity (Bair et al., 2006; Zhou et al., 2016). PCA procedure is mainly divided into the following steps: 1) centralize all the features. 2) Find the covariance matrix C. 3) Find the eigenvalues and the corresponding eigenvectors of covariance matrix C. 4) Project the original features onto the selected feature vector to achieve dimensionality reduction. PCA is the most commonly used linear dimensionality reduction method. In the past 20 years, PCA has been commonly used to analyze the effects of multi-pollutant mixture on human health (Fig. 2). Details on the literature search are provided in supplementary materials. When analyzing the health effects of multi-pollutant indicators, PCA can be used to reduce the numbers of indicators that need to be analyzed with minimum loss of information contained in the original indicators, so as to achieve the purpose of the comprehensive analysis of the collected data. Smit et al. applied PCA to

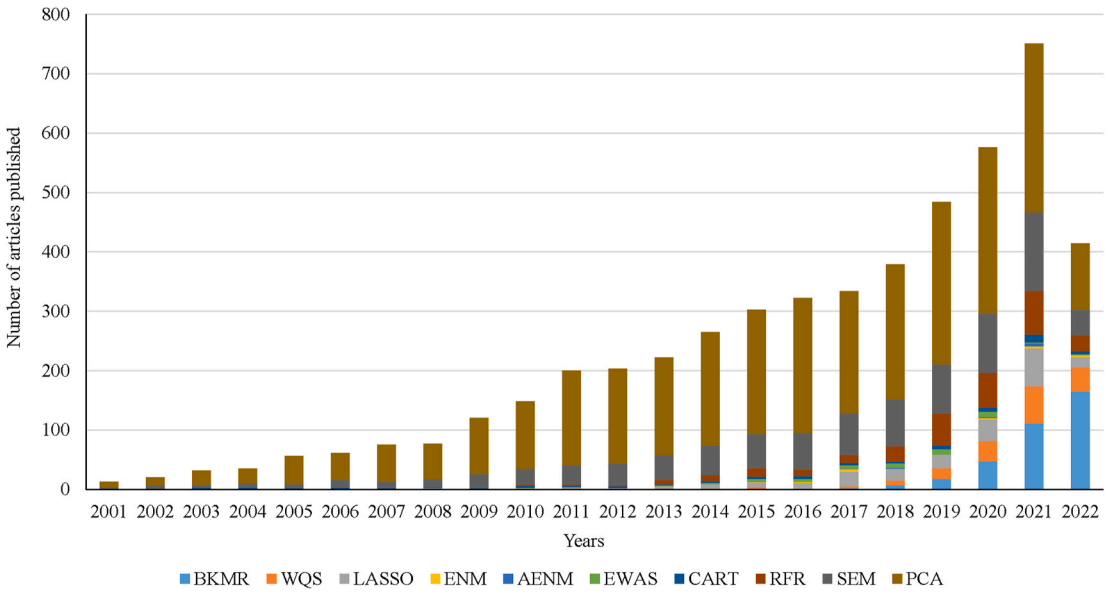

Fig. 2. Trends in the number of articles with multi-pollutant mixture analyzed by each statistical method.
Data were searched on PubMed from January 1, 2001 to April 1, 2022.

estimate the associations between 16 pollutants in maternal serum and the risks of asthma and eczema in school-aged children (Smit et al., 2015). This study eventually included the scores of five principal components that explained 70% of the variance in the multiple logistic regression models (Smit et al., 2015).

3.4. Environment-wide association study (EWAS)

EWAS is a relatively new concept first introduced by Patel et al., in 2010 (Patel et al., 2010). It was inspired by genome-wide association studies (GWAS) and similar to GWAS in methodology, but differs from GWAS in interested predictors: environmental factors in EWAS while genetic phenotypes in GWAS. Hundreds of chemicals and factors could be tested simultaneously for their associations with a health outcome by utilizing EWAS which is a strategy for analyzing the health effects of multi-pollutant mixtures. The general flow of the analysis is as follows: 1) the total sample is split into a discovery set and a test set. 2) All regression models are tested in the discovery set, where multiple tests are corrected by conducting false discovery rate (FDR) or Bonferroni correction. 3) Chemicals/factors that passed the FDR or Bonferroni correction threshold in the discovery set are further analyzed in the test set, where multiple tests are also corrected by conducting FDR or Bonferroni correction. 4) Chemicals/factors that survive FDR or Bonferroni correction from both the discovery set and the test set are statistically significant and finally reported (Uche et al., 2020). The reason for using discovery set and test set is to improve the external validity of the discovery. EWAS results can be visualized by constructing an "exposure globe" (Patel and Manrai, 2015).

Environmental factors differ from relatively stable genetic factors in that they have great spatiotemporal variabilities and can be controlled by measures (Zheng et al., 2020a). The putative pollutants identified by EWAS can be used not only for disease risk prediction, but also for disease prevention and intervention. To date, the EWAS approach has been applied to assess the effects of multi-pollutants on several diseases, such as cardiovascular diseases (Zhuang et al., 2018), type 2 diabetes (Patel et al., 2010), hypertension (New and Bennett, 2020), metabolic syndrome (Lind et al., 2013), and chronic kidney diseases (Lee et al., 2020).

EWAS encourages high-dimensional data for analysis. The results could identify a list of potential risk factors associated with the interested outcomes in a high-dimensional and agnostic manner, then generate new hypotheses. When the type and quantity of exposure variables are very large, the efficiency and effectiveness of EWAS in variable screening are superior to that of ENM. As to EWAS, the major challenges are that there are no accepted criteria for quality control and inclusion of environmental factors (Zheng et al., 2020a). In addition, research on multi-pollutant mixture by using EWAS is constrained by data availability since EWAS is generally used in the analysis of very large numbers of variables.

3.5. Classification and regression tree (CART) and random forest regression (RFR)

As a type of decision tree methodology first developed by Breiman in 1984, CART is a simple nonparametric regression approach that can be used to identify at-risk populations in public health (Breiman et al., 1984). The dependent variable can be either categorical (i.e., classification tree) or continuous (i.e., regression tree). The independent variable can be any combination of categorical and continuous variables. CART is an objective segmentation technique that takes different combinations of the independent variables and iteratively splits them into the "left branch", or "right branch" of the tree and finally finds the best combination in the tree generated by different combinations (Fig. 3). There are many types of splitting criteria. For classification trees, the most commonly used is the GINI index. For regression trees, the least square residuals and least absolute residuals are used to measure. (Shimokawa et al., 2015). Such recursive partitioning continues until there is no noticeable improvement in the partitioning or the variation in child nodes is small enough (Strobl et al., 2009). CART can produce a visual output, a multilevel structure that is similar to the branches of a tree (Fig. 3). Based on the multilevel structure, we can find out which environmental pollutant exposure may have a potential impact on the interested outcomes. Examples of previous studies using CART for public health analyses include epidemiological studies which assessed risk factors for mortality and morbidity from specific diseases (Carmelli et al., 1997). CART methodology has been increasingly applied in clinical research and in assessing the effects of multi-pollutant mixture on health (Gass et al., 2014; Marshall, 2001; Mazenq et al., 2017; Yang et al., 2020). CART can deal with missing values automatically through surrogate segmentation and is immune to outliers. In addition, additivity and linearity in the exposure-response relationship does not need to be assumed in CART. It is not suitable for small data sets, whereas it is

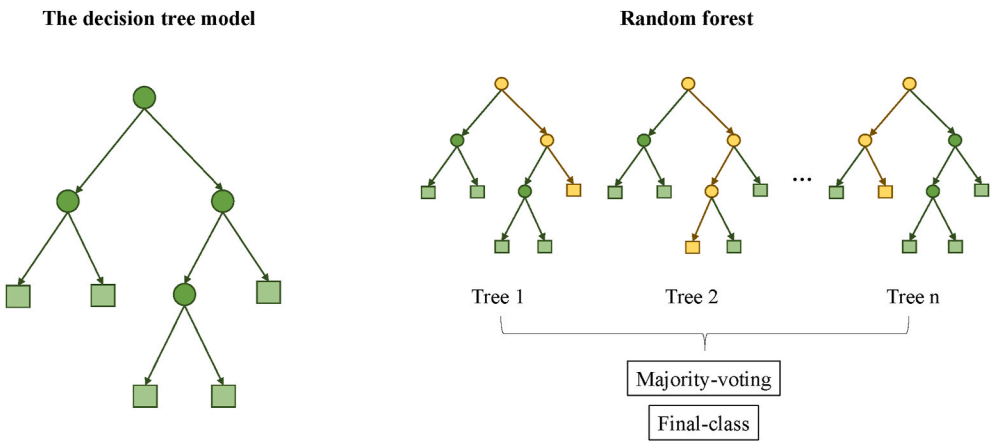

Fig. 3. Classification and regression tree (CART) and random forest regression (RFR). CART iteratively splits the node into two or more child nodes until certain specified conditions are met. RFR is an ensemble learning approach based on CART. Each tree in the ensemble is built based on the principle of recursive partitioning. In RFR, each node is split into two or more child nodes by using the best node in a randomly selected subset of predictors. Data in each child node are used to predict the values of the dependent variable. Results from all child nodes are then combined to produce final predictions.

suitable for categorical exposure variables because CART requires many splits to create a linear or nonlinear association.

RFR is an ensemble learning method developed from CART (Fig. 3) (Breiman, 2001). Each tree in the ensemble is constructed in a certain "random" way and is built based on the principle of recursive partitioning (Strobl et al., 2009). In RFR, each tree is created from a different row sample, and each node is split into two or more child nodes by using samples with different features. Data in each child node of each tree are used for its own individual prediction. Results from these predictions are then combined and averaged to produce final prediction. RFR is different from multiple linear regression because no data distribution assumption is needed (Davalos et al., 2017). Without specifying terms, RFR is able to detect the interactions and higher-order relationships among predictors during the modeling process (Strobl et al., 2009). However, RFR has been criticized for generating highly predictive models that are difficult to interpret. In addition, both CART and RFR are prone to overfitting. Yang et al. conducted a CVD prediction model study in China, and evaluated multiple methods by using a multivariate regression model as a benchmark for performance evaluation (Yang et al., 2020) The results show that RFR is superior to other methods such as the multivariate regression model and CART (Yang et al., 2020).

3.6. *Structural equation model (SEM)*

SEM has become a very popular approach in data analysis since the 1970s. It is a methodology to integrate specific sets of covariance and regressions among specified variables into a single cohesive model (Tomarken and Waller, 2005). The SEM process includes the following steps: 1) Model formulation. The initial model is generally formed based on theoretical research or practical experience before model estimation. 2) Model identification, which determines whether the parameter estimation of the set model has a unique solution. 3) Model estimation, which is to minimize the relationship between the sample variance/-covariance and the variance/covariance estimated by the model. The most commonly used estimation method is maximum likelihood. 4) Model evaluation, which is a necessary process to evaluate whether the model fits the data. There are a variety of methods used for model evaluation and various indices used for model fitting assessment (e.g., absolute fitting indices and incremental fitting indices). 5) Model correction. If the fitted model is not good, the model needs to be reset or modified. Modification indices are used as diagnostic indices to help modify model settings.

SEM facilitates understanding and estimating a network of relationships among variables (latent variables, visible variables, and error variables) (Fig. 4) (Stein et al., 2017). SEM is also used to estimate the degree of model fitting and allows measurement errors in independent and dependent variables (Stein et al., 2017). As researchers move towards multi-pollutant mixture modeling, SEM has been increasingly used to estimate the effects of multi-pollutant mixture exposure on health (Shook-Sa et al., 2017; Wang et al., 2019). Noteworthy, one of the

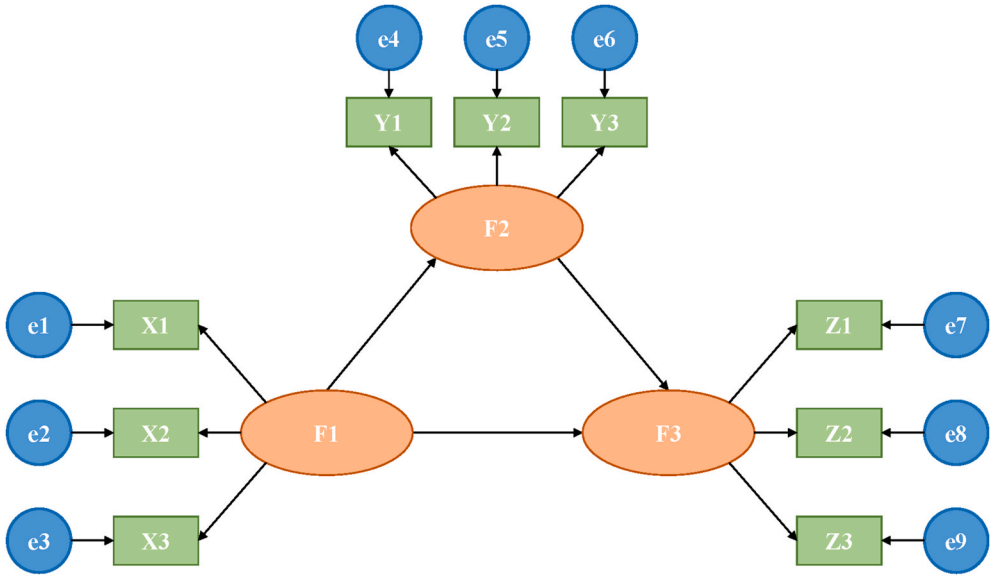

Fig. 4. Structural equation modeling (SEM). F1, F2, and F3 are latent variables. X1∼X3, Y1∼Y3, and Z1∼Z3 are visible variables. e1∼e9 are error variables. The measurement model is composed of e1∼e3, X1∼X3, and F1. The structural model is composed of F1, F2, and F3. The relationships between latent, visible and error variables in the model can be tested to estimate the direct, indirect, and total effects of independent variables on dependent variables.

most important characteristics of SEM analysis is that it should be based on a specific theoretical basis, which is also the main reason why SEM is regarded as a confirmatory rather than exploratory statistical method. The analysis process of SEM should be based on a clear theoretical concept or logical reasoning. From the point of view of statistical principles, SEM must comply with a number of basic assumptions of traditional statistical analysis, such as linearity and normality; otherwise, the statistical data obtained will not be convincing.

4. Discussion

In this review, we focused on the practical statistical methods used for estimating the effects of exposure to multi-pollution mixture on health, which currently are available in kinds of literature. A detailed summary of reviewed statistical methods with R package/function suggestions is presented in Tables 1 and 2. The independent variables of several statistical methods such as BKMR, WQS, quantile g-computation, GWQS and PCA are limited to continuous. The dependent variables in our review methods can be continuous or categorical.

It is highly recommended that the characteristics, application conditions, advantages, limitations, result interpretability, complexity, and computational cost of the proposed model, the data characteristics, and the purposes of study should be taken into account carefully when choosing an appropriate statistical method. Nonparametric methods such as CART and RFR may be suitable tools when little is known about the internal relationships among the components of a multi-pollutant mixture. Because one function of these nonparametric methods is to explore potentially complex interactions among the components of mixture and their impact on the outcome by relaxing parametric assumptions (Strobl et al., 2009). SEM is appropriate when the goal is to identify the relationships between multi-pollutant mixture and multiple outcomes and then evaluate the structure and relationship among pollutants (Tomarken and Waller, 2005). Besides, it is difficult to use SEM to assess the contribution of individual pollutants to the health outcome, or to separate potential hazardous factors from confounders and nuisance variables because of multicollinearity. When the study purpose is to conduct variable screening (identify relatively important mixture component), all statistical methods reviewed in this paper are practicable (Table 2). LASSO, ENM, and AENM are immune to multicollinearity to a certain extent and ensure convergence by forcing a penalty on the magnitude of the coefficients. WQS-based methods (WQS, quantile g-computation, and GWQS) are substitutions of LASSO and ENR for multi-pollutant mixture analysis with higher accuracy, improved specificity, and comparable sensitivity in the selection of correlated exposure variables (Carrico et al., 2015). BKMR can also be used as a variable selection tool, concurrently estimating the importance (i.e., PIP) of groups of highly correlated exposure variables and individual pollutants within the group (Bobb et al., 2015). It is worth mentioning and noting that the WQS-based methods and BKMR were proposed specifically for investigating the health effects of multi-pollutant mixture. Both WQS-based methods and BKMR provide a mixture analysis that considering multicollinearity. Linear, nonlinear, and non-additive mixture exposure-effect relationships can be estimated by quantile g-computation and BKMR. When the health effect of pollutants were nonlinear and non-additive, the quantile g-computation outputs an unbiased estimate of health effect and relevant variance, while the results outputted by WQS may be biased (Keil et al., 2020). Since each statistical approach used in assessing health effects from multi-pollutant mixture exposure possesses distinctive strengths and limitations, future researches may need to perform multiple comparative strategies with the combined use of multiple applicable statistical approaches (Zhang et al., 2019; Zheng et al., 2020b). An example of choosing statistical methods is given in supplementary materials.

Although the advances in statistics will benefit for addressing the analysis of health effects from multi-pollutant mixture exposure, some challenges require attention. The study of multi-pollutant mixture effect may be susceptible to measurement error (Billionnet et al., 2012). In large-scale epidemiological studies, errors in assessing pollutant exposure of individuals are somewhat inevitable. Measurement error can bias exposure-outcome effect estimates in an unpredictable way. In addition, a specific interested chemical is usually chosen based on the known concerns and measured by existing methods, which means that the research may be affected by the observational bias (Braun et al., 2016). In order to correct bias from measurement error, we need expert knowledge or prior knowledge on the measurement error mechanism (e.g., classical or Berkson) (Bateson et al., 2007). In the absence of prior information, sensitivity analysis may be able to examine the extent of inference affected by measurement error revision. Several values of chemicals below the limit of detection (LOD) are frequently encountered when addressing the health effects of multi-pollutant mixture. Various single-imputation methods have been developed, such as maximum likelihood estimate (MLE), restricted MLE, reverse Kaplan–Meier, and empirical "robust fill-in" methods (Carli et al., 2022; Gillespie et al., 2010). In addition, multiple-imputation procedures have also been developed for BKMR, WQS regression and quantile g-computation (Lubin et al., 2004; Carli et al., 2022; Hargarten and Wheeler, 2020). These abovementioned methods allow us to estimate the health effect of exposure to multi-pollutant mixture with high percentage of values below LOD. More targeted imputation methods need to be developed in future studies.

Epidemiological studies are uniquely positioned to provide insight into the potential health risk from exposure to multi-pollutant mixture that are more representative of exposure patterns in the real-word. Notably, although researches on the health effects of multi-pollutant mixture are important to potentially formulate environmental pollutant control regulations, assessment for the independent effect of exposure to a single pollutant is still crucial, especially when assessing the health effect of an emerging unknown pollutant. We hope that this review will aid researchers in selecting appropriate statistical methods to estimate the health effects of multi-pollutant mixture. At present, it is still difficult to assert which model is most suitable for evaluating the health effects of multi-pollutant mixture given the unique data and purposes of different studies as well as the unique characteristics (application conditions, strengths, weaknesses, etc.) of different models. More advanced approaches are encouraged to be developed to uncover multiple environmental determinants of disease.

5. Conclusions

This review elaborated on the practical statistical methods suitable for epidemiological studies to estimate the effects of multi-pollutant mixture on human health. The characteristics, applications, goals, advantages, disadvantages, and specific implemented R package/function of each method were summarized. These statistical methods are of great help in disentangling the authentic effects. Researchers are highly recommended to take summarized information (characteristics, application conditions, advantages, limitations, and result interpretability of each statistical method) as well as research objectives and the characteristics of data they have into consideration when selecting an appropriate statistical method. The combined application of multiple suitable methods with conclusions drawn based on the intersection of the corresponding multiple results is also suggestive. More advanced and improved statistical methods are encouraged to be developed to more accurately and easily assess the health risks from exposure to multi-pollutant mixture.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT authorship contribution statement

Linling Yu: Conceptualization, Methodology, Visualization, Writing – original draft. **Wei Liu:** Conceptualization, Methodology, Writing – review & editing, and. **Xing Wang:** Conceptualization, Methodology, Writing – review & editing. **Zi Ye:** Methodology, Writing – review & editing. **Qiyou Tan:** Methodology, Writing – review & editing. **Weihong Qiu:** Methodology, Writing – review & editing. **Xiuquan Nie:** Methodology, Writing – review & editing, and. **Minjing Li:** Methodology, Writing – review & editing. **Bin Wang:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, and. **Weihong Chen:** Conceptualization, Methodology, Writing – review & editing, Supervision, Project administration, All the authors have read, edited and approved the manuscript.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We sincerely appreciated all participants recruited in the study and the support from the study team.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envpol.2022.119356.

References

Aitken, E.H., Damelang, T., Ortega-Pajares, A., Alemu, A., Hasang, W., Dini, S., Unger, H. W., Ome-Kaius, M., Nielsen, M.A., Salanti, A., Smith, J., Kent, S., Hogarth, P.M., Wines, B.D., Simpson, J.A., Chung, A., Rogerson, S.J., 2021. Developing a multivariate prediction model of antibody features associated with protection of malaria-infected pregnant women from placental malaria. Elife 10. https://doi.org/10.7554/eLife.65776.

Araki, A., Ait Bamai, Y., Bastiaensen, M., Van den Eede, N., Kawai, T., Tsuboi, T., Miyashita, C., Itoh, S., Goudarzi, H., Konno, S., Covaci, A., Kishi, R., 2020. Combined exposure to phthalate esters and phosphate flame retardants and plasticizers and their associations with wheeze and allergy symptoms among school children. Environ. Res. 183, 109212. https://doi.org/10.1016/j.envres.2020.109212.

Bair, E., Hastie, T., Paul, D., Tibshirani, R., 2006. Prediction by supervised principal components. J. Am. Stat. Assoc. 101, 119–137. https://doi.org/10.1198/016214505000000628.

Bateson, T.F., Coull, B.A., Hubbell, B., Ito, K., Jerrett, M., Lumley, T., Thomas, D., Vedal, S., Ross, M., 2007. Panel discussion review: session three–issues involved in interpretation of epidemiologic analyses–statistical modeling. J. Expo. Sci. Environ. Epidemiol. 17 (Suppl. 2), S90–S96. https://doi.org/10.1038/sj.jes.7500631.

Bellavia, A., James-Todd, T., Williams, P.L., 2019. Approaches for incorporating environmental mixtures as mediators in mediation analysis. Environ. Int. 123, 368–374. https://doi.org/10.1016/j.envint.2018.12.024.

Billionnet, C., Gay, E., Kirchner, S., Leynaert, B., Annesi-Maesano, I., 2011. Quantitative assessments of indoor air pollution and respiratory health in a population-based sample of French dwellings. Environ. Res. 111, 425–434. https://doi.org/10.1016/j.envres.2011.02.008.

Billionnet, C., Sherrill, D., Annesi-Maesano, I., 2012. Estimating the health effects of exposure to multi-pollutant mixture. Ann. Epidemiol. 22, 126–141. https://doi.org/10.1016/j.annepidem.2011.11.004.

Bobb, J.F., 2017. Example Using the Bkmr R Package with Simulated Data from the NIEHS Mixtures Workshop.

Bobb, J.F., Valeri, L., Claus Henn, B., Christiani, D.C., Wright, R.O., Mazumdar, M., Godleski, J.J., Coull, B.A., 2015. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16, 493–508. https://doi.org/10.1093/biostatistics/kxu058.

Bobb, J.F., Claus Henn, B., Valeri, L., Coull, B.A., 2018. Statistical software for analyzing the health effects of multiple concurrent exposures via Bayesian kernel machine regression. Environmental health : a global access science source 17, 67. https://doi.org/10.1186/s12940-018-0413-y, 67.

Braun, J.M., Gennings, C., Hauser, R., Webster, T.F., 2016. What can epidemiological studies tell us about the impact of chemical mixtures on human health? Environ. Health Perspect. 124, A6–A9. https://doi.org/10.1289/ehp.1510569.

Breiman, L., 2001. Random forest. Mach. Learn. 45, 5–32. https://doi.org/10.1023/A:1010933404324.

Breiman, L.B., Friedman, J.H., Olshen, R.A., Stone, C.J., 1984. Classification and Regression Trees, second ed. Brooks/Cole (Pacific Grove).

Carli, M., Ward, M.H., Metayer, C., Wheeler, D.C., 2022. Imputation of below detection limit missing data in chemical mixture analysis with bayesian group index regression. Int. J. Environ. Res. Publ. Health 19. https://doi.org/10.3390/ijerph19031369.

Carmelli, D., Zhang, H., Swan, G.E., 1997. Obesity and 33-year follow-up for coronary heart disease and cancer mortality. Epidemiology 8, 378–383. https://doi.org/10.1097/00001648-199707000-00005.

Carrico, C., Gennings, C., Wheeler, D.C., Factor-Litvak, P., 2015. Characterization of weighted quantile sum regression for highly correlated data in a risk analysis setting. J. Agric. Biol. Environ. Stat. 20, 100–120. https://doi.org/10.1007/s13253-014-0180-3.

Chattopadhyay, A., Lu, T.P., 2019. Gene-gene interaction: the curse of dimensionality. Ann. Transl. Med. 7, 813. https://doi.org/10.21037/atm.2019.12.87.

Coker, E., Chevrier, J., Rauch, S., Bradman, A., Obida, M., Crause, M., Bornman, R., Eskenazi, B., 2018. Association between prenatal exposure to multiple insecticides and child body weight and body composition in the VHEMBE South African birth cohort. Environ. Int. 113, 122–132. https://doi.org/10.1016/j.envint.2018.01.016.

Daniel, S., Balalian, A.A., Insel, B.J., Liu, X., Whyatt, R.M., Calafat, A.M., Rauh, V.A., Perera, F.P., Hoepner, L.A., Herbstman, J., Factor-Litvak, P., 2020. Prenatal and early childhood exposure to phthalates and childhood behavior at age 7 years. Environ. Int. 143, 105894. https://doi.org/10.1016/j.envint.2020.105894.

Davalos, A.D., Luben, T.J., Herring, A.H., Sacks, J.D., 2017. Current approaches used in epidemiologic studies to examine short-term multipollutant air pollution exposures. Ann. Epidemiol. 27, 145–153. https://doi.org/10.1016/j.annepidem.2016.11.016 e141.

Gass, K., Klein, M., Chang, H.H., Flanders, W.D., Strickland, M.J., 2014. Classification and regression trees for epidemiologic research: an air pollution example. Environ. Health 13, 17. https://doi.org/10.1186/1476-069x-13-17.

Gillespie, B.W., Chen, Q., Reichert, H., Franzblau, A., Hedgeman, E., Lepkowski, J., Adriaens, P., Demond, A., Luksemburg, W., Garabrant, D.H., 2010. Estimating population distributions when some data are below a limit of detection by using a reverse Kaplan-Meier estimator. Epidemiology 21 (Suppl. 4), S64–S70. https://doi.org/10.1097/EDE.0b013e3181ce9f08.

Hargarten, P.M., Wheeler, D.C., 2020. Accounting for the uncertainty due to chemicals below the detection limit in mixture analysis. Environ. Res. 186, 109466. https://doi.org/10.1016/j.envres.2020.109466.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67. https://doi.org/10.2307/1267351.

Hou, J., Yin, W., Li, P., Hu, C., Xu, T., Cheng, J., Li, T., Wang, L., Yu, Z., Yuan, J., 2020. Joint effect of polycyclic aromatic hydrocarbons and phthalates exposure on telomere length and lung function. J. Hazard Mater. 386, 121663. https://doi.org/10.1016/j.jhazmat.2019.121663.

Huang, Hui, Wan Mohamed Radzi, C.W., Salarzadeh Jenatabadi, H., 2017. Family environment and childhood obesity: a new framework with structural equation modeling. Int. J. Environ. Res. Public Health 181. https://doi.org/10.3390/ijerph14020181.

Kampa, M., Castanas, E., 2008. Human health effects of air pollution. Environ. Pollut. 151, 362–367. https://doi.org/10.1016/j.envpol.2007.06.012.

Keil, A.P., Buckley, J.P., O'Brien, K.M., Ferguson, K.K., Zhao, S., White, A.J., 2020. A quantile-based g-computation approach to addressing the effects of exposure mixtures. Environ. Health Perspect. 128, 47004. https://doi.org/10.1289/ehp5838.

Kortenkamp, A., Faust, M., 2018. Regulate to reduce chemical mixture risk. Science 361, 224–226. https://doi.org/10.1126/science.aat9219.

Lazarevic, N., Barnett, A.G., Sly, P.D., Knibbs, L.D., 2019. Statistical methodology in studies of prenatal exposure to mixtures of endocrine-disrupting chemicals: a review of existing approaches and new alternatives. Environ. Health Perspect. 127, 26001. https://doi.org/10.1289/EHP2207, 26001.

Lee, J., Oh, S., Kang, H., Kim, S., Lee, G., Li, L., Kim, C.T., An, J.N., Oh, Y.K., Lim, C.S., Kim, D.K., Kim, Y.S., Choi, K., Lee, J.P., 2020. Environment-wide association study of CKD. Clin. J. Am. Soc. Nephrol. 15, 766–775. https://doi.org/10.2215/cjn.06780619.

Lenters, V., Portengen, L., Rignell-Hydbom, A., Jönsson, B.A., Lindh, C.H., Piersma, A.H., Toft, G., Bonde, J.P., Heederik, D., Rylander, L., Vermeulen, R., 2016. 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. 124, 365–372. https://doi.org/10.1289/ehp.1408933.

Lin, Z., Lin, S., Neamtiu, I.A., Ye, B., Csobod, E., Fazakas, E., Gurzau, E., 2021. Predicting environmental risk factors in relation to health outcomes among school children from Romania using random forest model - an analysis of data from the SINPHONIE project. Sci. Total Environ. 784, 147145. https://doi.org/10.1016/j.scitotenv.2021.147145.

Lind, P.M., Risérus, U., Salihovic, S., Bavel, B., Lind, L., 2013. An environmental wide association study (EWAS) approach to the metabolic syndrome. Environ. Int. 55, 1–8. https://doi.org/10.1016/j.envint.2013.01.017.

Lubin, J.H., Colt, J.S., Camann, D., Davis, S., Cerhan, J.R., Severson, R.K., Bernstein, L., Hartge, P., 2004. Epidemiologic evaluation of measurement data in the presence of detection limits. Environ. Health Perspect. 112, 1691–1696. https://doi.org/10.1289/ehp.7199.

Marshall, R.J., 2001. The use of classification and regression trees in clinical epidemiology. J. Clin. Epidemiol. 54, 603–609. https://doi.org/10.1016/s0895-4356(00)00344-9.

Mauderly, J.L., Samet, J.M., 2009. Is there evidence for synergy among air pollutants in causing health effects? Environ. Health Perspect. 117, 1–6. https://doi.org/10.1289/ehp.11654.

Mazenq, J., Dubus, J.C., Gaudart, J., Charpin, D., Viudes, G., Noel, G., 2017. City housing atmospheric pollutant impact on emergency visit for asthma: a classification and regression tree approach. Respir. Med. 132, 1–8. https://doi.org/10.1016/j.rmed.2017.09.004.

McEligot, A.J., Poynor, V., Sharma, R., Panangadan, A., 2020. Logistic LASSO regression for dietary intakes and breast cancer. Nutrients 12. https://doi.org/10.3390/nu12092652.

Meng, X., Liu, C., Chen, R., Sera, F., Vicedo-Cabrera, A.M., Milojevic, A., Guo, Y., Tong, S., Coelho, M., Saldiva, P.H.N., Lavigne, E., Correa, P.M., Ortega, N.V., Osorio, S., Garcia Kyselý, J., Urban, A., Orru, H., Maasikmets, M., Jaakkola, J.J.K., Ryti, N., Huber, V., Schneider, A., Katsouyanni, K., Analitis, A., Hashizume, M., Honda, Y., Ng, C.F.S., Nunes, B., Teixeira, J.P., Holobaca, I.H., Fratianni, S., Kim, H., Tobias, A., Íñiguez, C., Forsberg, B., Åström, C., Ragettli, M.S., Guo, Y.L., Pan, S.C., Li, S., Bell, M.L., Zanobetti, A., Schwartz, J., Wu, T., Gasparrini, A., Kan, H., 2021. Short term associations of ambient nitrogen dioxide with daily total, cardiovascular, and respiratory mortality: multilocation analysis in 398 cities. Bmj 372, n534. https://doi.org/10.1136/bmj.n534.

Navas-Acien, A., Domingo-Relloso, A., Subedi, P., Riffo-Campos, A.L., Xia, R., Gomez, L., Haack, K., Goldsmith, J., Howard, B.V., Best, L.G., Devereux, R., Tauqeer, A., Zhang, Y., Fretts, A.M., Pichler, G., Levy, D., Vasan, R.S., Baccarelli, A.A., Herreros-Martinez, M., Tang, W.Y., Bressler, J., Fornage, M., Umans, J.G., Tellez-Plaza, M., Fallin, M.D., Zhao, J., Cole, S.A., 2021. Blood DNA methylation and incident coronary heart disease: evidence from the strong heart study. JAMA Cardiol 6, 1237–1246. https://doi.org/10.1001/jamacardio.2021.2704.

New, A., Bennett, K.P., 2020. A precision environment-wide association study of hypertension via supervised cadre models. IEEE J Biomed Health Inform 24, 916–925. https://doi.org/10.1109/jbhi.2019.2918070.

Ngamwong, Y., Tangamornsuksan, W., Lohitnavy, O., Chaiyakunapruk, N., Scholfield, C. N., Reisfeld, B., Lohitnavy, M., 2015. Additive synergism between asbestos and smoking in lung cancer risk: a systematic review and meta-analysis. PLoS One 10, e0135798. https://doi.org/10.1371/journal.pone.0135798.

NIH, 2020. Powering Research through Innovative Methods for Mixtures in Epidemiology (PRIME) Program Meeting.

Patel, C.J., Manrai, A.K., 2015. Development of exposome correlation globes to map out environment-wide associations. Pac Symp Biocomput 20, 231–242.

Patel, C.J., Bhattacharya, J., Butte, A.J., 2010. An Environment-Wide Association Study (EWAS) on type 2 diabetes mellitus. PLoS One 5, e10746. https://doi.org/10.1371/journal.pone.0010746.

Patel, N.P., Sarraf, E., Tsai, M.H., 2018. The curse of dimensionality. Anesthesiology 129, 614–615. https://doi.org/10.1097/aln.0000000000002350.

Ranganathan, P., Pramesh, C.S., Aggarwal, R., 2017. Common pitfalls in statistical analysis: logistic regression. Perspect Clin Res 8, 148–151. https://doi.org/10.4103/picr.PICR_87_17.

Roberts, S., Martin, M.A., 2006. Using supervised principal components analysis to assess multiple pollutant effects. Environ. Health Perspect. 114, 1877–1882. https://doi.org/10.1289/ehp.9226.

Schisterman, E.F., Perkins, N.J., Mumford, S.L., Ahrens, K.A., Mitchell, E.M., 2017. Collinearity and causal diagrams: a lesson on the importance of model specification. Epidemiology 28, 47–53. https://doi.org/10.1097/ede.0000000000000554.

Shah-Kulkarni, S., Lee, S., Jeong, K.S., Hong, Y.-C., Park, H., Ha, M., Kim, Y., Ha, E.-H., 2020. Prenatal exposure to mixtures of heavy metals and neurodevelopment in infants at 6 months. Environ. Res. 182, 109122. https://doi.org/10.1016/j.envres.2020.109122.

Shan, Z., Chen, S., Sun, T., Luo, C., Guo, Y., Yu, X., Yang, W., Hu, F.B., Liu, L., 2016. U-shaped association between plasma manganese levels and type 2 diabetes. Environ. Health Perspect. 124, 1876–1881. https://doi.org/10.1289/ehp176.

Shimokawa, A., Kawasaki, Y., Miyaoka, E., 2015. Comparison of splitting methods on survival tree. Int. J. Biostat. 11, 175–188. https://doi.org/10.1515/ijb-2014-0029.

Shook-Sa, B.E., Chen, D.G., Zhou, H., 2017. 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. Publ. Health 14. https://doi.org/10.3390/ijerph14101112.

Smit, L.A., Lenters, V., Høyer, B.B., Lindh, C.H., Pedersen, H.S., Liermontova, I., Jönsson, B.A., Piersma, A.H., Bonde, J.P., Toft, G., Vermeulen, R., Heederik, D., 2015. Prenatal exposure to environmental chemical contaminants and asthma and eczema in school-age children. Allergy 70, 653–660. https://doi.org/10.1111/all.12605.

Stein, C.M., Morris, N.J., Hall, N.B., Nock, N.L., 2017. Structural equation modeling. Methods Mol. Biol. 1666, 557–580. https://doi.org/10.1007/978-1-4939-7274-6_28.

Strobl, C., Malley, J., Tutz, G., 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol. Methods 14, 323–348. https://doi.org/10.1037/a0016973.

Tanner, E.M., Hallerbäck, M.U., Wikström, S., Lindh, C., Kiviranta, H., Gennings, C., Bornehag, C.G., 2020. Early prenatal exposure to suspected endocrine disruptor mixtures is associated with lower IQ at age seven. Environ. Int. 134, 105185. https://doi.org/10.1016/j.envint.2019.105185.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. Roy. Stat. Soc. B 58, 267–288.

Tomarken, A.J., Waller, N.G., 2005. Structural equation modeling: strengths, limitations, and misconceptions. Annu. Rev. Clin. Psychol. 1, 31–65. https://doi.org/10.1146/annurev.clinpsy.1.102803.144239.

Uche, U.I., Suzuki, S., Fulda, K.G., Zhou, Z., 2020. Environment-wide association study on childhood obesity in the U.S. Environ. Res. 191, 110109. https://doi.org/10.1016/j.envres.2020.110109.

Valeri, L., Mazumdar, M.M., Bobb, J.F., Claus Henn, B., Rodrigues, E., Sharif, O.I.A., Kile, M.L., Quamruzzaman, Q., Afroz, S., Golam, M., Amarasiriwardena, C., Bellinger, D.C., Christiani, D.C., Coull, B.A., Wright, R.O., 2017. 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. 125, 067015 https://doi.org/10.1289/ehp614.

Van den Dries, M.A., Keil, A.P., Tiemeier, H., Pronk, A., Spaan, S., Santos, S., Asimakopoulos, A.G., Kannan, K., Gaillard, R., Guxens, M., Trasande, L., Jaddoe, V. W.V., Ferguson, K.K., 2021. Prenatal exposure to nonpersistent chemical mixtures and fetal growth: a population-based study. Environ. Health Perspect. 129, 117008. https://doi.org/10.1289/ehp9178.

Vrijheid, M., Casas, M., Gascon, M., Valvi, D., Nieuwenhuijsen, M., 2016. Environmental pollutants and child health-A review of recent concerns. Int. J. Hyg Environ. Health 219, 331–342. https://doi.org/10.1016/j.ijheh.2016.05.001.

Wang, X., Mukherjee, B., Park, S.K., 2018. Associations of cumulative exposure to heavy metal mixtures with obesity and its comorbidities among U.S. adults in NHANES 2003-2014. Environ. Int. 121, 683–694. https://doi.org/10.1016/j.envint.2018.09.035.

Wang, L., Hou, J., Hu, C., Zhou, Y., Sun, H., Zhang, J., Li, T., Gao, E., Wang, G., Chen, W., Yuan, J., 2019. Mediating factors explaining the associations between polycyclic aromatic hydrocarbons exposure, low socioeconomic status and diabetes: a structural equation modeling approach. Sci. Total Environ. 648, 1476–1483. https://doi.org/10.1016/j.scitotenv.2018.08.255.

Wang, X., Bhramar, M., Carrie, A.K., William, H.H., Stuart, B., Siobán, D.H., Sung, K.P., 2020. Urinary metal mixtures and longitudinal changes in glucose homeostasis: the study of Women's Health Across the Nation (SWAN). Environ. Int. 145 https://doi.org/10.1016/j.envint.2020.106109.

Wang, B., Cheng, M., Yang, S., Qiu, W., Li, W., Zhou, Y., Wang, X., Yang, M., He, H., Zhu, C., Cen, X., Chen, A., Xiao, L., Zhou, M., Ma, J., Mu, G., Wang, D., Guo, Y., Zhang, X., Chen, W., 2020a. Exposure to acrylamide and reduced heart rate variability: the mediating role of transforming growth factor-β. J. Hazard Mater. 395, 122677. https://doi.org/10.1016/j.jhazmat.2020.122677.

Wang, B., Qiu, W., Yang, S., Cao, L., Zhu, C., Ma, J., Li, W., Zhang, Z., Xu, T., Wang, X., Cheng, M., Mu, G., Wang, D., Zhou, Y., Yuan, J., Chen, W., 2020b. Acrylamide exposure and oxidative DNA damage, lipid peroxidation, and fasting plasma glucose alteration: association and mediation analyses in Chinese urban adults. Diabetes Care 43, 1479–1486. https://doi.org/10.2337/dc19-2603.

Wheeler, D.C., Rustom, S., Carli, M., Whitehead, T.P., Ward, M.H., Metayer, C., 2021. Assessment of grouped weighted quantile sum regression for modeling chemical mixtures and cancer risk. Int. J. Environ. Res. Publ. Health 18. https://doi.org/10.3390/ijerph18020504.

Xiao, L., Li, W., Zhu, C., Yang, S., Zhou, M., Wang, B., Wang, X., Wang, D., Ma, J., Zhou, Y., Chen, W., 2021. Cadmium exposure, fasting blood glucose changes, and type 2 diabetes mellitus: a longitudinal prospective study in China. Environ. Res. 192, 110259. https://doi.org/10.1016/j.envres.2020.110259.

Xu, T., Wang, B., Wang, X., Yang, S., Cao, L., Qiu, W., Cheng, M., Liu, W., Yu, L., Zhou, M., Wang, D., Ma, J., Chen, W., 2021. Associations of urinary carbon disulfide metabolite with oxidative stress, plasma glucose and risk of diabetes among urban adults in China. Environ. Pollut. 272, 115959. https://doi.org/10.1016/j.envpol.2020.115959.

Yang, L., Wu, H., Jin, X., Zheng, P., Hu, S., Xu, X., Yu, W., Yan, J., 2020. Study of cardiovascular disease prediction model based on random forest in eastern China. Sci. Rep. 10, 5245. https://doi.org/10.1038/s41598-020-62133-5.

Yitshak-Sade, M., Fabian, M.P., Lane, K.J., Hart, J.E., Schwartz, J.D., Laden, F., James, P., Fong, K.C., Kloog, I., Zanobetti, A., 2020. Estimating the combined effects of natural and built environmental exposures on birthweight among urban residents in Massachusetts. Int. J. Environ. Res. Publ. Health 17, 8805. https://doi.org/10.3390/ijerph17238805.

Yu, G., Jin, M., Huang, Y., Aimuzi, R., Zheng, T., Nian, M., Tian, Y., Wang, W., Luo, Z., Shen, L., Wang, X., Du, Q., Xu, W., Zhang, J., 2021a. Environmental exposure to perfluoroalkyl substances in early pregnancy, maternal glucose homeostasis and the risk of gestational diabetes: a prospective cohort study. Environ. Int. 156, 106621. https://doi.org/10.1016/j.envint.2021.106621.

Yu, L., Yang, M., Cheng, M., Fan, L., Wang, X., Xu, T., Wang, B., Chen, W., 2021b. Associations between urinary phthalate metabolite concentrations and markers of liver injury in the US adult population. Environ. Int. 155, 106608. https://doi.org/10.1016/j.envint.2021.106608.

Zhang, Y., Dong, T., Hu, W., Wang, X., Xu, B., Lin, Z., Hofer, T., Stefanoff, P., Chen, Y., Wang, X., Xia, Y., 2019. Association between exposure to a mixture of phenols, pesticides, and phthalates and obesity: comparison of three statistical models. Environ. Int. 123, 325–336. https://doi.org/10.1016/j.envint.2018.11.076.

Zheng, Y., Chen, Z., Pearson, T., Zhao, J., Hu, H., Prosperi, M., 2020a. Design and methodology challenges of environment-wide association studies: a systematic review. Environ. Res. 183, 109275. https://doi.org/10.1016/j.envres.2020.109275.

Zheng, Y., Zhang, C., Weisskopf, M.G., Williams, P.L., Claus Henn, B., Parsons, P.J., Palmer, C.D., Buck Louis, G.M., James-Todd, T., 2020b. Evaluating associations between early pregnancy trace elements mixture and 2nd trimester gestational glucose levels: a comparison of three statistical approaches. Int. J. Hyg Environ. Health 224, 113446. https://doi.org/10.1016/j.ijheh.2019.113446, 113446.

Zhou, Yuhan, Ma, W., Zeng, Y., Yan, C., Zhao, Y., Wang, P., Shi, H., Lu, W., Zhang, Y., 2021. Intrauterine antibiotic exposure affected neonatal gut bacteria and infant growth speed. Environ. Pollut. 289 https://doi.org/10.1016/j.envpol.2021.117901.

Zhou, Y., Sun, H., Xie, J., Song, Y., Liu, Y., Huang, X., Zhou, T., Rong, Y., Wu, T., Yuan, J., Chen, W., 2016. Urinary polycyclic aromatic hydrocarbon metabolites and altered lung function in Wuhan, China. Am. J. Respir. Crit. Care Med. 193, 835–846. https://doi.org/10.1164/rccm.201412-2279OC.

Zhou, M., Xiao, L., Yang, S., Wang, B., Shi, T., Tan, A., Wang, X., Mu, G., Chen, W., 2020. Cross-sectional and longitudinal associations between urinary zinc and lung function among urban adults in China. Thorax 75, 771–779. https://doi.org/10.1136/thoraxjnl-2019-213909.

Zhuang, X., Guo, Y., Ni, A., Yang, D., Liao, L., Zhang, S., Zhou, H., Sun, X., Wang, L., Wang, X., Liao, X., 2018. Toward a panoramic perspective of the association between environmental factors and cardiovascular disease: an environment-wide association study from National Health and Nutrition Examination Survey 1999-2014. Environ. Int. 118, 146–153. https://doi.org/10.1016/j.envint.2018.05.046.

Zou, H., Hastie, T., 2005. Regularization and variable selection via the elastic net. J. Roy. Stat. Soc. B 67, 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, H., Zhang, H.H., 2009. On the adaptive elastic-net with a diverging number of parameters. Ann. Stat. 37, 1733–1751. https://doi.org/10.1214/08-aos625.