HHS Public Access

Author manuscript

Curr Epidemiol Rep. Author manuscript; available in PMC 2019 June 01.

Published in final edited form as:

Curr Epidemiol Rep. 2018 June ; 5(2): 160–165. doi:10.1007/s40471-018-0145-0.

Environmental exposure mixtures: questions and methods to address them

Ghassan B Hamra[*,1] and **Jessie P Buckley**[1,2]

[1]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, MD, USA

[2]Department of Environmental Health and Engineering, Johns Hopkins Bloomberg School of Public Health, MD, USA

Abstract

Purpose of this review—This review provides a summary of statistical approaches that researchers can use to study environmental exposure mixtures. Two primary considerations are the form of the research question and the statistical tools best suited to address that question. Because the choice of statistical tools is not rigid, we make recommendations about when each tool may be most useful.

Recent Findings—When dimensionality is relatively low, some statistical tools yield easily interpretable estimates of effect (e.g., risk ratio, odds ratio) or intervention impacts. When dimensionality increases, it is often necessary to compromise this interpretablity in favor of identifying interesting statistical signals from noise; this requires applying statistical tools that are oriented more heavily towards dimension reduction via shrinkage and/or variable selection.

Summary—The study of complex exposure mixtures has prompted development of novel statistical methods. We suggest that further validation work would aid practicing researchers in choosing among existing and emerging statistical tools for studying exposure mixtures.

Keywords

complex mixtures; environmental epidemiology; Bayesian methods; machine learning

Introduction

Environmental epidemiology, like many other specialties of epidemiology, historically concerned itself with identifying the health effects of single exposures, treating the effects of co-occurring exposures as something to adjust away. This focus could be rationally motivated by a desire to inform meaningful public health policy or reluctantly justified by limitations in statistical and laboratory tools to measure and analyze exposure data. As we

[*]Corresponding Author: Ghassan B Hamra, Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore, MD 21205, Tel: +1 (443) 287 0122, ghassanhamra@jhu.edu.

Compliance with Ethical Standards

Conflict of Interest

Ghassan B Hamra and Jessie P Buckley declare no conflicts of interest.

Human and Animal Rights and Informed Consent

This article does not contain any studies with human or animal subjects performed by any of the authors.

have learned more about the health effects of aggregate exposures that are often the subject of policy (ex: $PM_{2.5}$), interest has shifted to consider their constituents. Advances in technology have allowed researchers to better measure constituents of aggregate exposures, as well as multiple exposures in the environment and in biological samples. In tandem, statistical tools have been developed that allow us to better study the health impacts of multiple exposures. Both of these advances have supported a shift in interest to exposure mixtures and their potential health impacts. The term *mixture* may narrowly focus on a subset of compounds in a chemical class or broadly describe all the exposures a human may experience in his/her lifetime (i.e, the human exposome[1, 2]). The National Institute for Environmental Health Sciences has highlighted research into the effects of complex exposure mixtures as a priority area in its 2012–2017 strategic plan, and advances continue in this area both in terms of laboratory and statistical tools.

In this commentary, we will discuss the categories of primary research questions that may be posed in the presence of exposure mixtures, discuss the statistical tools that can be used for each category, and explain how the complexity of available data contribute to the choice of statistical tools.

Framing research questions about mixtures

The first step in any analysis is identifying a research question; when considering exposure mixtures the question may differ while the exposures of interest remain the same. Braun et al.[3] described three types of questions that researchers might consider when studying exposure mixtures. We expand on their classification and suggest four: (a) what is the effect of an aggregate mixture, (b) what is the effect of a sum of mixture components, (c) what are the independent effects of mixture components, and (d) what are the joint effects of mixture components. These research questions tie directly to measurement of the exposure variables and how they are parameterized. For simplicity, we compare possible formulations of these four structures using an aggregate exposure, Z, two exposures that represent components of Z, X_1 and X_2, and a binary outcome D:

a. $P(D = 1) = \alpha + \beta * Z$

b. $P(D = 1) = \alpha + \beta(\omega_1 X_1 + \omega_2 X_2)$

c. $P(D = 1) = \alpha + \beta_1 * X_1 + \beta_2 * X_2$

d. $P(D = 1) = \alpha + \beta_1 * X_1 + \beta_2 * X_2 + \beta_3 * X_1 * X_2$

Notably, these four structures can be modeled using the same exposure set but each structure addresses a different research question. Here, we will elaborate on these four structures and connect them to the goals of a study.

Aggregate exposure (a) studies represent the simplest and most common model structure. This structure treats a mixture as a single exposure. Often, exposures are measured in such a way that they cannot be disaggregated. For example, $PM_{2.5}$ is an aggregate measure of all compounds sampled in ambient air that have an aerodynamic diameter less than 2.5 microns. [4] To compare the estimated effects of $PM_{2.5}$ across research studies, we must assume that the health effects of all component exposures within this particle size range are equal, or that

the composition of $PM_{2.5}$ is uniform across different geographic areas where it is measured, two assumptions we know to be false.[5] Nonetheless, studying aggregate exposures may lend itself more directly to evaluations of public health interventions. For example, it is more straightforward to evaluate an intervention to reduce/remove $PM_{2.5}$ from ambient air than one designed to intervene on specific components of $PM_{2.5}$.

We define summed effects (b) studies as those that combine individual mixture components, often using weights. Exposure weights may be based on their expected potency relative to a reference exposure, a concept drawn from toxicology,[6] or based on their percent contribution to the total mixture effect.[7] We use the description of *summed* over the term *cumulative* used by previous authors,[3] because the latter suggests an exposure quantity that is calculated based on repeated exposures, over time, to the same chemical. There are many examples of summed approaches, which often employ toxic equivalency factors and related metrics. For instance, concentrations of multiple phthalate biomarkers have been modeled using molar sums[8] or a potency-weighted sum.[9] Like aggregate effects, summed effects estimate a single value for the effect of the mixture; however, in this case, the contribution of each component is typically scaled based on a reference exposure or percent contribution to the total mixture effect. When exposures are given a weight of 1, the summed effects formulation reduces down to the aggregate formulation; in our example, this simply means X_1 and X_2 sum to Z.

Independent effect (c) studies are concerned with isolating the effect of each mixture component; they are often motivated by results of studies examining health effects of aggregate exposures and seek to identify which mixture components have the most pronounced effect or account for co-pollutant confounding. As mentioned above, the abundance of particles in the $PM_{2.5}$ range will differ based on the source of the exposure and the individual components within this size range are well known to differ in terms of their expected health impacts. Research suggesting an effect of $PM_{2.5}$ on an outcome fits into the framework of a type (a) study and is the first step into a deeper investigation of etiologically relevant exposures, such as those represented by a type (c) study, as $PM_{2.5}$ itself contains both harmful and innocuous components.

Joint effect (d) studies are motivated by both prior research and the belief that exposure effects may be more complex than the simple, additive structures in expression (c). Many approaches have been developed to assess joint effects; however, special consideration must be given to articulating the effects being estimated.[10] Suppose X_1 and X_2 are two volatile organic compounds that fall in the $PM_{2.5}$ range; in this case, we would be interested in estimating an *interaction* between the exposures. This is justified because their effects are likely acting together in a biological sense, such as via a shared mechanistic pathway. However, if X_2 were an individual's sex at birth, then we are likely considering *effect measure modification*; this is because sex at birth is not modifiable. In this case, our goal is not to quantify a biologically relevant joint effect; rather, we are quantifying an effect of the main exposure, X_1, that is expected to differ by sex at birth. Our subsequent discussion focuses primarily on methods designed to assess interaction.

Different tools for different questions

Use of traditional statistical tools, such as univariate linear regression, can be used to address any of the exposure structures we describe; however they are typically most useful for (a) and (b). When there are few exposures to consider, exposures are not strongly correlated, or when sample size is very large, traditional tools may also be suitable for (c) and (d). More often, the computational complications of studying independent and joint effects of multiple, correlated exposures within a single regression model become apparent. As the number of exposures increases, potentially exceeding the number of observations, alternative statistical tools are needed that apply parameter shrinkage and/or dimension reduction techniques. The choice of the most appropriate tool depends on the hypothesis being tested, the relative complexity of the problem, and the researcher's desire for easily interpretable results.

When studying aggregate or summed exposures, traditional statistical tools with which researchers are largely familiar may be appropriate. For example, these tools are adequate to assess departures from linearity for single exposures using quadratic, cubic, or spline terms because the models are not overburdened with high dimensional exposure data. More recently, procedures have been developed to quantify the summed mixture effect as well as the contribution of exposures within the mixture. Two such approaches include weighted quantile sums and Bayesian Kernel Machine Regression (BKMR); these approaches can be viewed as a version of exposure structure (b). Weighted quantile sums regression provides a summed mixture effect and an estimate of the percent contribution to that effect from each exposure within the mixture; exposures with a percent contribution of close or equal to zero are viewed as null effects. Notably, this procedure requires the assumption that any effects of exposures on the outcome are all in the same direction (positive or negative).[7] BKMR provides a summed mixture effect as well; however, rather than an estimate of the percent contribution to the effect, BKMR provides the probability of inclusion in the summed mixture effect, and allows for examining multiple exposure-response shapes.[11] BKMR is also capable of examining independent effects of mixture components, and considers the impact of individual components holding the others constant at pre-specified percentile values (such as the 50^{th} percentile of the exposure distribution).[11]

Bayesian hierarchical models (BHM) are another appealing method, because they handle correlated exposure data and can yield highly interpretable results. BHMs incorporate prior information on exposure effects to stabilize model estimates of correlated exposure variables. A common BHM approach for estimating independent effects assumes, before the analysis, that the effects of a group of exposures on the outcome of interest are exchangeable, and applies a shared prior distribution to those exposure variables.[12, 13] This allows borrowing of information between exposure effects, thereby leveraging the available data and the model form via the specified prior.[14] If the data do not support th assumption of exchangeability, then the effects of exposures that do not conform will depart from the shared mean. As long as exposures are on the same scale, such as $\mu g/m^3$ or parts per billion (ppb) as is common in air pollution research, the estimated effects may be a risk difference, risk ratio, or odds ratio, all of which are familiar to most researchers. Importantly, the prior used need not be a shared mean. Researchers may apply any number of priors,

including a mixture prior that allows an effect estimate to arise from multiple distributions, without losing interpretability of the results.[15]

We also note that Bayesian methods can be useful for studying aggregate and summed exposure effects when an informative prior exists and it is desirable to integrate that prior, quantitatively, into an analysis. A quantitative prior may take the form of estimated effects from prior studies, including meta-analyses, concerning specific compounds. However, priors may also be drawn directly from toxicology and experimental research. For example, if laboratory and animal experiments show that certain components in a mixture have a higher or lower effect on the outcome relative to a reference compound in the mixture, this knowledge can be integrated using order constraints.[16, 17] Such an approach allows researchers to quantitatively connect concepts such as toxic equivalence and relative potency from toxicology into observational epidemiology research.

As the number of exposures in the mixture increases, into the hundreds or even thousands, the approaches described above may be less appealing because, in some cases, they cannot handle the computational burden of higher dimensional exposure data. When there are many exposures, researchers can consider a number of machine learning algorithms to study complex exposure mixtures. These methods do not return easily interpretable effect estimates with which researchers are familiar; however, they can indicate which exposures more strongly predict the outcome, an insight that can help direct further research. These tools have been adopted for informatics and data science problems because of their ability to handle high dimensional data problems.

Machine learning tools may be useful for considering the independent and joint effects of higher dimensional exposure data. These tools fall into two general classes: penalized estimators or decision trees. The former applies a penalty to the likelihood to achieve the goals of shrinkage and/or variable selection. A useful aspect of penalized estimators is that we retain the ability to explore the dose-response shape and potential joint effects of exposures and obtain interpretable coefficients. A necessary compromise is that exposures are standardized so that the estimated effect applies to a 1 unit change in the standard deviation of each exposure's distribution. Commonly applied penalized estimators are partial least squares,[18, 19] elastic net,[20] least absolute shrinkage and selection operator (or LASSO), and ridge regression; elastic net and LASSO can both be derived from Bayesian inference, as well.[21, 22] Chadeau-Hyam et al. and Stafoggia et al. provide useful summaries of these estimation procedures in the context of OMICs research.[23, 24] Decision trees operate by splitting the exposures into categories based on cutpoints (ex: exposed vs unexposed, below vs above 50^{th} percentile) and identifies the subgroups (or *branches*) that best predict the outcome of interest. Because results can be highly sensitive to this classification, methods for fitting multiple trees and aggregating their results have been developed. Examples of these approaches are random forest[25] and Bayesian additive regression trees (BART). [26] Other tools, such as bagging [27] and boosting[28] are also applied to regression trees to improve their performance.

Another suite of methods for examining joint effects includes dimension reduction techniques that focus on and leverage information contained in the exposure matrix itself.

Dimension reduction methods reduce a high dimensional exposure matrix based on the amount of variability explained by each exposure, producing lower dimension exposure summaries. The most common of these approaches is principal components analysis (PCA), which uses orthogonal transformations to reduce a set of correlated variables into a smaller set of uncorrelated variables, or principal components. When PCA is unsupervised, selection of exposures for each component is agnostic to the health outcome of interest. Components identified with PCA can be regressed on an outcome using principal components regression. However, because components are constructed based on the amount of variance explained by each exposure distribution and do not account for their relationship to the outcome, they are not often generalizable across studies. Supervised PCA is less common but has the added benefit of incorporating outcome associations in the construction of summary variables. Additionally, cluster analysis methods can be used to examine joint effects of exposure mixtures. Cluster analysis classify *observations* into groups with similar mixture profiles; this is in contrast to dimension reduction techniques that group *variables* that are alike. For example, hierarchical and *k*-means clustering approaches were used to identify days with similar air pollution mixture profiles,[29] and these clusters were subsequently related to mortality.[30]

Finally, we highlight g-methods as an alternative suite of tools that can address questions about aggregate, summed, individual, or joint effects from the perspective of public health interventions.[31] In the setting of exposure mixtures, g-methods may be particularly useful for estimating effects of environmental health policies targeted at altering the *source* of an exposure mixture. For example, studies have used g-methods to estimate effects of interventions to reduce NO_2, a marker for traffic exposures, or ozone concentrations below a specified level or exposure standard.[32, 33] Recent Bayesian extensions to g-methods can incorporate any Bayesian estimation method, such as BKMR, BHM or Bayesian LASSO, to model complex underlying exposure-outcome associations while still producing a simple summary estimate that quantifies the effect of a specified intervention on a mixture or its components.[34]

Future considerations

While we have described several methods to handle distinct exposures that occur as a mixture, additional statistical challenges warrant attention to improve the study of exposure mixtures. The methods described above have predominantly been applied to exposure mixtures ascertained at a single point in time. However, investigators are increasingly measuring exposures at multiple time points to estimate effects of cumulative exposures as well as examine critical windows of susceptibility. The problem of exposure mixtures becomes increasingly complex when exploring time windows, given both within and between correlation of mixture components. Methods that expand the consideration of exposure mixtures as time varying include BHM, WQS, BKMR, and the Bayesian g-formula.[34–37]

Methods to correct for inaccurate or missing exposure information are another area in need of development within the complex mixture setting. Dependent misclassification is an underrecognzied but important potential source of bias in studies estimating effects of

multiple exposure biomarkers on health.[38] Similarly, estimates of multiple exposures obtained via extrapolation with tools such as land use regression[39] may be prone to correlated exposure measurement error. Several approaches have been developed to formally address exposure measurement error.[40–43] However, with rare exception, [44, 45] these methods have not been applied to studies of exposure mixtures.

We believe Bayesian methods are particularly appealing for studying exposure mixtures because they are flexible and can incorporate prior information. Notably, effects of mixtures can be examined in a single framework along with corrections for measurement error [40] and other issues, such as values outside the limits of detection.[46] Bayesian approaches are also capable of handling time varying exposures, for example, by applying priors that shrink estimates across time windows towards a common mean when researchers believe that effects of time varying exposures are similar. These methods can be computationally intensive, taking hours or days to converge. Notably, researchers are making headway in resolving this, such as development of STAN software for more efficient sampling.[47] Researchers can also apply a laplace approximation[48] of the Bayesian posterior for more efficient computation.

Conclusion

Studying exposure mixtures is an active and evolving area in environmental epidemiology. We have summarized many of the most compelling methods to study mixtures. Notably, the choice of method will be partly motivated by the complexity of the exposure mixture. To this end, researchers exploring potential methods should clearly articulate their research question and understand the exposure structure driving the study. These considerations are important in choosing among statistical approaches to balance their benefits against their computational costs and interpretability.

Acknowledgments

JPB was supported by funding from the National Institutes of Health (U24 OD023382).

References

1. Wild CP. Complementing the genome with an “exposome”: the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol Biomarkers Prev. 2005; 14(8):1847–50. [PubMed: 16103423]
2. Wild CP. The exposome: from concept to utility. Int J Epidemiol. 2012; 41(1):24–32. [PubMed: 22296988]
3**. Braun JM, et al. What Can Epidemiological Studies Tell Us about the Impact of Chemical Mixtures on Human Health? Environmental Health Perspectives. 2016; 124(1):A6–A9. [PubMed: 26720830]
4. Hamra GB, et al. Outdoor particulate matter exposure and lung cancer: a systematic review and meta-analysis. Environ Health Perspect. 2014; 122(9):906–11. [PubMed: 24911630]
5. Chung Y, et al. Associations between long-term exposure to chemical constituents of fine particulate matter (PM2.5) and mortality in Medicare enrollees in the eastern United States. Environ Health Perspect. 2015; 123(5):467–74. [PubMed: 25565179]
6. Howard GJ, Webster TF. Contrasting theories of interaction in epidemiology and toxicology. Environ Health Perspect. 2013; 121(1):1–6. [PubMed: 23014866]

7. Czarnota J, Gennings C, Wheeler DC. Assessment of weighted quantile sum regression for modeling chemical mixtures and cancer risk. Cancer Inform. 2015; 14(Suppl 2):159–71. [PubMed: 26005323]

8. Wolff MS, et al. Prenatal phenol and phthalate exposures and birth outcomes. Environ Health Perspect. 2008; 116(8):1092–7. [PubMed: 18709157]

9. 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(19): 10616–10624. [PubMed: 27579903]

10. VanderWeele TJ. On the distinction between interaction and effect modification. Epidemiology. 2009; 20(6):863–71. [PubMed: 19806059]

11. Bobb JF, et al. Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics. 2015; 16(3):493–508. [PubMed: 25532525]

12. Hamra G, MacLehose R, Richardson D. Markov chain Monte Carlo: an introduction for epidemiologists. Int J Epidemiol. 2013; 42(2):627–34. [PubMed: 23569196]

13. MacLehose RF, Hamra GB. Applications of Bayesian Methods to Epidemiologic Research. Current Epidemiology Reports. 2014:1–7.

14. Gelman A, Hill J, Yajima M. Why We (Usually) Don't Have to Worry About Multiple Comparisons. Journal of Research on Educational Effectiveness. 2012; 5(2):189–211.

15. MacLehose RF, et al. Bayesian methods for highly correlated exposure data. Epidemiology. 2007; 18(2):199–207. [PubMed: 17272963]

16. Hamra G, et al. Integrating informative priors from experimental research with Bayesian methods: an example from radiation epidemiology. Epidemiology. 2013; 24(1):90–5. [PubMed: 23222512]

17. Hamra GB, et al. Lung cancer risk associated with regulated and unregulated chrysotile asbestos fibers. Epidemiology. 2016

18. Wold, H. Partial Least Squares, in Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc; 2004.

19. Wold S, Sjostrom M, Eriksson L. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems. 2001; 58(2):109–130.

20. Zou H, Hastie T. Regularization and variable selection via the elastic net. JR Statist Soc B. 2005; 67(Part 2):301–320.

21. Li Q, Lin N. The Bayesian Elastic Net. Bayesian Analysis. 2010; 5(1):151–170.

22. Park T, Casella G. The Bayesian Lasso. Journal of the American Statistical Association. 2008; 103(482):681–686.

23. Chadeau-Hyam M, et al. Deciphering the complex: methodological overview of statistical models to derive OMICS-based biomarkers. Environ Mol Mutagen. 2013; 54(7):542–57. [PubMed: 23918146]

24*. Stafoggia M, et al. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science. Current Environmental Health Reports. 2017; 4(4):481–490. [PubMed: 28988291]

25. Ho, TK. Random decision forests. Proceedings of the Third International Conference on Document Analysis and Recognition (Volume 1); IEEE Computer Society; 1995. 278

26. Chipman HA, George EI, McCulloch RE. BART: Bayesian additive regression trees. Ann Appl Stat. 2010; 4(1):266–298.

27. Breiman L. Bagging Predictors. Machine Learning. 1996; 24(2):123–140.

28. Freund, Y; Schapire, RE. Experiments with a new boosting algorithm. Proceedings of the Thirteenth International Conference on International Conference on Machine Learning; Bari, Italy: Morgan Kaufmann Publishers Inc; 1996. 148–156.

29. Austin E, et al. A framework for identifying distinct multipollutant profiles in air pollution data. Environ Int. 2012; 45:112–21. [PubMed: 22584082]

30. Zanobetti A, et al. Health effects of multi-pollutant profiles. Environ Int. 2014; 71:13–9. [PubMed: 24950160]

31. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. Mathematical Modelling. 1986; 7(9–12):1393–1512.
32. Snowden JM, et al. Population intervention models to estimate ambient NO2 health effects in children with asthma. J Expo Sci Environ Epidemiol. 2015; 25(6):567–73. [PubMed: 25182844]
33. Moore K, et al. Ambient ozone concentrations and cardiac mortality in Southern California 1983–2000: application of a new marginal structural model approach. Am J Epidemiol. 2010; 171(11): 1233–43. [PubMed: 20439309]
34. Keil AP, et al. A Bayesian approach to the g-formula. Stat Methods Med Res. 2017
35. Bello GA, et al. Extending the Distributed Lag Model framework to handle chemical mixtures. Environ Res. 2017; 156:253–264. [PubMed: 28371754]
36. Liu SH, et al. Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. Biostatistics. 2017
37. Richardson DB, et al. Hierarchical latency models for dose-time-response associations. Am J Epidemiol. 2011; 173(6):695–702. [PubMed: 21303803]
38. Pollack AZ, et al. Correlated biomarker measurement error: an important threat to inference in environmental epidemiology. Am J Epidemiol. 2013; 177(1):84–92. [PubMed: 23221725]
39. Basagana X, et al. Measurement error in epidemiologic studies of air pollution based on land-use regression models. Am J Epidemiol. 2013; 178(8):1342–6. [PubMed: 24105967]
40. MacLehose RF, et al. Bayesian methods for correcting misclassification: an example from birth defects epidemiology. Epidemiology. 2009; 20(1):27–35. [PubMed: 19234399]
41. Cole SR, Chu H, Greenland S. Multiple-imputation for measurement-error correction. Int J Epidemiol. 2006; 35(4):1074–81. [PubMed: 16709616]
42. Kuchenhoff H, Mwalili SM, Lesaffre E. A general method for dealing with misclassification in regression: the misclassification SIMEX. Biometrics. 2006; 62(1):85–96. [PubMed: 16542233]
43. Rosner B, Willett WC, Spiegelman D. Correction of logistic regression relative risk estimates and confidence intervals for systematic within-person measurement error. Stat Med. 1989; 8(9):1051–69. [PubMed: 2799131]
44. Keller JP, et al. Covariate-Adaptive Clustering of Exposures for Air Pollution Epidemiology Cohorts. Ann Appl Stat. 2017; 11(1):93–113. [PubMed: 28572869]
45. Dionisio KL, Chang HH, Baxter LK. A simulation study to quantify the impacts of exposure measurement error on air pollution health risk estimates in copollutant time-series models. Environ Health. 2016; 15(1):114. [PubMed: 27884187]
46. Herring AH. Nonparametric bayes shrinkage for assessing exposures to mixtures subject to limits of detection. Epidemiology. 2010; 21(Suppl 4):S71–6. [PubMed: 20526202]
47. Carpenter B, et al. Stan: A Probabilistic Programming Language. 2017; 76(1):32.
48. Rue H, Martino S, Chopin N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. Journal of the Royal Statistical Society Series B-Statistical Methodology. 2009; 71:319–392.
49. Valeri L, 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(6):067015. [PubMed: 28669934]
50. Lenters V, 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(3):365–72. [PubMed: 26115335]
51. Liaw A, Wiener M. Classification and Regression by randomForest. R News. 2002; 2(3):18–22.
52. Agay-Shay K, et al. Exposure to Endocrine-Disrupting Chemicals during Pregnancy and Weight at 7 Years of Age: A Multi-pollutant Approach. Environ Health Perspect. 2015; 123(10):1030–7. [PubMed: 25956007]
53. Milligan, GW. Encyclopedia of Statistical Sciences. John Wiley & Sons, Inc; 2004. Cluster Analysis.

Table 1

Summary of statistical tools used to study environmental exposure mixtures and research questions they may address.

Statistical tools	Research question				High dimensional (>10 exposures)	Key references
	Aggregate	Summed	Independent	Joint		
Traditional regression methods	x	x	x	x		n/a*
Bayesian hierarchical models	x	x	x	x		[13, 15]
Weighted quantile sums (WQS)		x			x	[7]
Bayesian Kernel Machine Regression (BKMR)		x	x	x		[11, 49]
Least Absolute Shrinkage and Selection Operator (LASSO)			x	x	x	[46]
Ridge regression			x	x	x	
Elastic net			x	x	x	[20, 50]
Partial Least Squares (PLS)			x	x	x	[23]
Random forest			x		x	[51]
Bagging			x		x	
Boosting			x		x	
Bayesian Additive Regression Trees (BART)			x		x	[26]
Principal Components Analysis (PCA)				x	x	[23, 52]
Cluster analysis				x	x	[29, 53]
Bayesian g-formula	x	x	x	x	x	[34]

*No citation is included for traditional regression methods, as these are familiar to public health researchers.