H2020 program	The Exposome Project for Health and Occupational Research
Grant agreement number	874703
Title	Report on tutorial for the application of a suite of 'multiple exposure methods'
Date	2022
Responsible author	Dr Susan Peters
E-mail	s.peters@uu.nl
Co-authors	Wenxin Wan (UU)
	Dr Lützen Portengen (UU)
	Dr Tanja Krone (TNO)
	Dr Calvin Ge (UU)
	Prof Roel Vermeulen (UU)

Contents

Background

It is well recognised that workers are usually exposed to multiple exposures at workplaces (Sejbaek et al., 2020). An example was given by a nationwide survey that over 81% Australian workers were exposed to more than one carcinogen, and 26% to more than five carcinogens (McKenzie et al., 2020). Understanding the health effect from multiple occupational exposures has increasingly been a focus both for aetiology research and regulatory agencies (Fourneau et al., 2021, Bosson-Rieutort et al., 2020). This aim has been further extended to the concept of "working-life exposome" that represents all occupational and related non-occupational exposure risk factors (e.g., environment, lifestyle, behavioural and socio-economic) (Pronk et al., 2021). The multiple exposure-response relationships that can be captured with the working-life exposome would lay the groundwork for evidence-based and cost-effective preventions, and ultimately improve working life health by reducing the burden of non-communicable diseases (Pronk et al., 2021).

Epidemiological studies with the focus on the health effect associated with multiple exposures have been mainly explored for environmental exposures such as chemical mixtures. A number of novel methods have been developed and applied in this field, which have been comprehensively reviewed in (Braun et al., 2016, Gibson et al., 2019a, Huang et al., 2018, Lazarevic et al., 2019, Taylor et al., 2016, Vuong et al., 2020). Additionally, several papers have reviewed the statistical approaches by incorporating the "exposome" concept and state-of-the-art machine learning approaches (Agier et al., 2016, Bellinger et al., 2017, Bi et al., 2019, Billionnet et al., 2012, Braun et al., 2016, Guillien et al., 2021, Kino et al., 2021, Oskar and Stingone, 2020, Santos et al., 2020).

By summarising these reviews and adapting the research questions for which the methods were developed to the EPHOR objectives, we here put forward two generalised research questions:

(1) Identification of important exposure(s) and independent effects; and

(2) Estimation of joint health effects.

1. *Identification of important exposure(s) and independent effects*

Because of the large number of occupational and non-occupational exposures that workers are exposed to, it is necessary to identify exposures that are most strongly associated with adverse health outcomes, including individual exposures or groups of highly correlated and related exposures with a common source (Braun et al., 2016). This leads to an important question epidemiology can address: from a range of possible exposures, what are the important exposures that contribute to the health

outcomes? Several studies and ample research have been focused on this question. An example in occupational epidemiology was given where researchers were interested in identifying important metals (from 16 measured metals) that are associated with the alteration of cardiovascular function (Zhang et al., 2017). Another example involves the studies of exposome-wide association analyses (EWAS), where Patel and colleagues investigated the association between 188 environmental and life-style factors and serum lipid (Patel et al., 2012).

Whilst the increasing number of measured exposures can potentially unravel previously unidentified exposure-response relationships, the large number of exposure variables brings challenges to statistical modelling. Specifically, multiple comparison problem occurs when one consider a set of statistical inferences simultaneously, resulting in many false discoveries (Type I error) of exposure-response relations (Greenland, 2008). This is particularly the case for EWAS-related studies with a very large number of exposure variables and no strong a-priori hypotheses. Another challenge lays in the nature of occupational exposures where exposure levels of co-occurring exposures often show strong correlations. For conventional multiple regression models, such high degree of correlation between exposures would lead to unstable effect estimates and inflated standard errors (multicollinearity), and effects for individual exposures would be hard to disentangle.

2. *Estimation of joint health effects*

The other research question is the estimation of a joint health effect (also cumulative/overall effect). This refers to a potential summarised effect resulting from combined exposures to multiple occupational, environmental, and social stressors (Gibson et al., 2019a, Huang et al., 2018, Taylor et al., 2016). The motivation behind it is the overall effect from a large number of small-effect exposures (below regulatory limits) could still exist and pose adverse effect to public health (Gibson et al., 2019a). One main challenge is to identify and account for interactions (non-additive effects) between exposure variables, and this becomes more challenging when more exposures are involved (Barrera-Gomez et al., 2017). The other challenge is to capture more complex, nonlinear exposure-response relationships. Many conventional statistical approaches like multiple linear regression assume linear effects, which may not sufficiently reflect the actual complex relationship. Therefore, more advanced statistical approaches are needed to estimate the joint health effect from multiple exposures, while accounting for the non-additive and nonlinear exposures-response relationships.

In this tutorial we present an overview of a suite of existing statistical approaches that can be applied to multiple exposure-response modelling in the context of working-life exposome. The basic principles, implementation, pro/cons, and their extensions are discussed and may serve as a useful guide for statistical analyses with multiple exposures.

Method selection for multiple exposure analyses within EPHOR

The selection of methods that are included in this tutorial followed the following steps:

1) **Method collection:** We produced a list of modern methods by a comprehensive review of the existing review papers and model comparison papers that are related to multiple exposure analyses.
2) **Evaluation criteria:** We developed detailed criteria that evaluate the performance, availability, and interpretability of the methods; and obtained feedback by reaching out to the EPHOR consortium and by holding a working session at the consortium meeting on 12th November 2021.
3) **Method evaluation:** Based on the developed criteria and the received feedback, two experts (LP and TK, both with >10 years' experience in biostatistics) were involved in several rounds of discussions to evaluate the list of methods and come up with a smaller set of methods upon agreement.

The details of this selection process are provided in Appendices A and B.

Selected methods for tutorial

The feedback from the working session at consortium meeting was incorporated into the expert-based evaluation of methods. Details of the feedback are shown in Appendix A. Based on all the acquired information, we have highlighted three methods, which are considered to be most appropriate for the multiple exposure analyses for EPHOR project:

- BKMR (Bayesian kernel machine regression);
- Random Forest; and
- LASSO (Least absolute shrinkage and selection operator).

These three methods complement each other in terms of strengths and weaknesses, and are all relatively easy to apply in epidemiological analyses (as shown in Figure 1). These features will inform method selection, which depends on the specific research question (Appendix B).

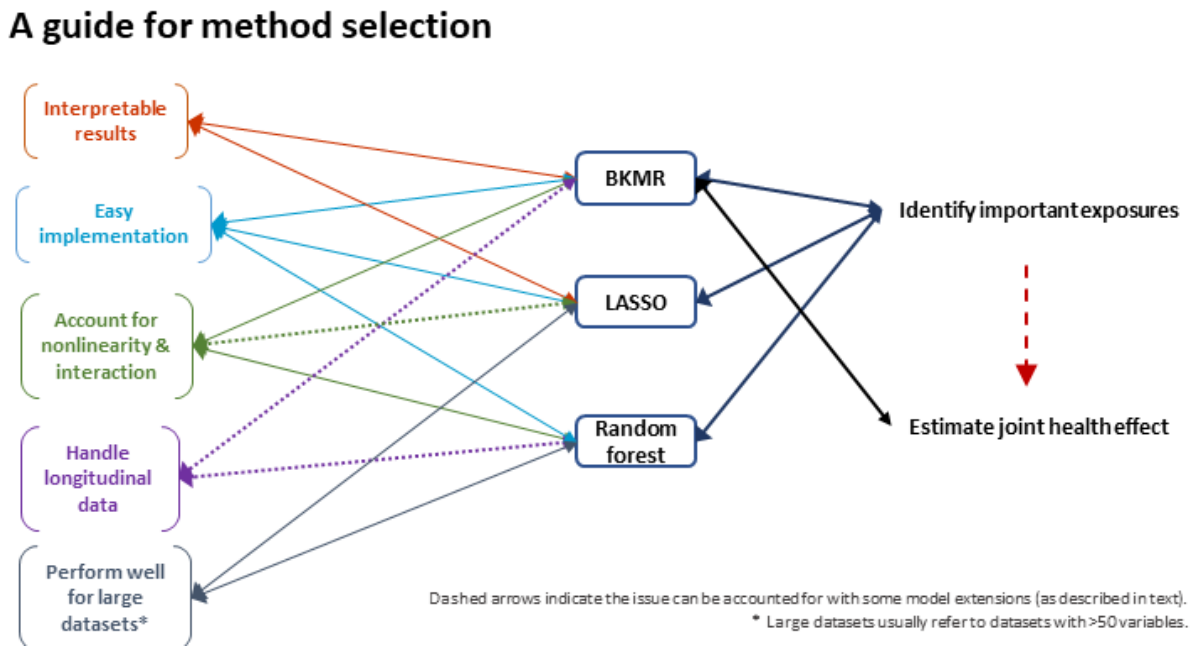


Figure 1. Overview of the three selected methods

General recommendations for method selection are as follows:

- For researchers who seek simple models for variable selection, LASSO (combined with stability selection, details are in later section) and random forest are more preferred to apply for large datasets. LASSO could provide more interpretable results than random forest, and so-called "relaxed" LASSO models could be used to produce coefficients with less bias.
- For researchers who have experience with Bayesian analyses (e.g., MCMC modelling), BKMR could be a more appropriate approach that provides more comprehensive outputs (e.g., exposure-response relationships and interaction). However, BKMR in its present form cannot be used for case-control studies and struggles when used on datasets with large number of observations.

More description about the strengths and weakness of the three methods can be found in Table 1. Specifically for analyses with many exposure variables of interest (e.g., more than 20), random forest could be used as a variable pre-selection method, followed by BKMR to examine the exposure-response relationships for selected variables in more detail. Note, however, the potential for bias and overfitting whenever model structure or parameter settings are based on results from earlier analyses using the same data. To assess the potential for bias, several different methods could be used in parallel to check the consistency of results. Also, empirical knowledge of the co-occurring exposures at workplaces still plays an important role in the variable selection process.

We have also considered the research question of estimating joint health effect to closely interlink with the question of identifying important exposures. Once the important exposures relating to the health outcome are identified, subsequent models (e.g., multiple linear regression) can be developed to investigate the interaction effect (risk of co-exposed to multiple exposures). It should be acknowledged that current development in environmental mixture analyses has been limited to produce reliable estimates of overall health effect from multiple environmental exposures. Some methods such as weighted quantile sum regression (WQSR) (Carrico et al., 2015) have been considered, but were not highlighted here because of its strong assumption on the direction of effects and inability to account for interaction and nonlinearity (at least in its present form). In the process of method selection, we have also identified Bayesian profile regression (BPR) and multivariate adaptive regression spline (MARS) as promising approaches for multiple exposure analyses, but more validation and discussion are needed to confirm their performance. Interested readers could refer to relevant sources for BPR (Molitor et al., 2010, Papathomas et al., 2011, Pirani et al., 2015) and MARS (Lu et al., 2021, Nacar et al., 2020, Nieto et al., 2015)).

This does not suggest the unselected methods are inferior compared to the selected. Researchers should select appropriate methods based primarily on research goals and the features of the datasets. The analytic objectives and correlation structure should also be considered in part with methodology strength and weakness for the method selection. Below we provide the full list of methods we have evaluated based on the criteria. It is expected to serve as a reference for method selection in multiple exposure analyses.

Please note that exploration of the dataset (see Appendix C) is recommended to precede any main analyses.

Description of the selected methods

BKMR (Bayesian kernel machine regression)

Introduction

The BKMR (Bayesian kernel machine regression) model (Bobb et al., 2015) is a semi-parametric technique to estimate individual and joint health effect from multiple exposures. For continuous outcome, the model is given by:

$$\boldsymbol{Y_i = h(z_{i1}, \cdots, z_{iM}) + x_i'\beta + \varepsilon_i,}$$

where Y_i denotes the response for individual *i* (*i* = 1, …, n), z_{iM} is the m^{th} exposure variable, *h* denotes flexible exposure-response function to be estimated, $\boldsymbol{\beta}$ represents the effect of the covariates and x_i the vector of covariate. The residuals ε_i are assumed to be independent and identically (iid) normally distributed with a common variance.

BKMR is one of a few approaches that were developed with the explicit goal of environmental mixture modelling (Bobb et al., 2015). Since it was first developed in 2015, BKMR has been widely applied in both toxicology and epidemiological studies, leading to some exposure-health relations that were previously unidentified (Valeri et al., 2017). This method became more popular with the release of the *bkmr* package in R (Bobb et al., 2018). A search of "BKMR" in abstract section in PubMed results in 125 papers before September 2021.

Implementation

The main implementation follows the application of the R package *bkmr* from its developers (Bobb, 2017, Bobb et al., 2018). The main steps and the outputs are shown in Figure 2. Briefly, three "modes" are available for the application: no variable selection, the component-wise variable selection (BKMR-VS) and hierarchical variable selection (BKMR-HVS). BKMR-VS is suited for situations where: (1) there is a relatively small number of exposure variables, or (2) the correlation between exposure variables is low, or (3) the researcher intends to include all the exposure variable into the model. BKMR-HVS, on the other hand, is suited for highly correlated exposures variables which can be grouped based on correlation or empirical knowledge. The outputs from BKMR-HVS will inform the selection of variables from the correlated exposures within one group.

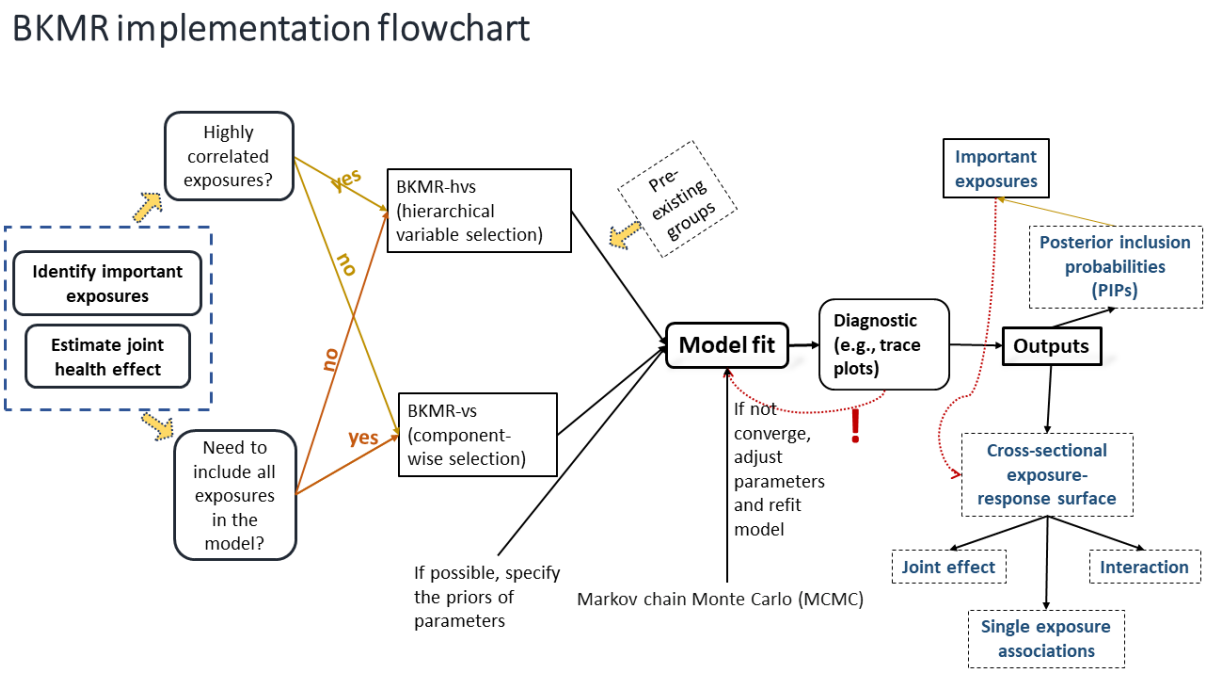


Figure 2. The implementation flowchart for *bkmr* package

MCMC (Markov Chain Monte Carlo) is used to fit the BKMR model. Inputs to the models include outcome as a vector Y, exposure variables matrix Z, and optional confounders/covariates matrix X. A kernel function is used to estimate the flexible exposure-response function *h(·)*. Among the choices of several kernel functions, Gaussian kernel is often used. The model fit can be very computational expensive for large sample size, but can be made more efficient using a Gaussian predictive process could speed up the process (Bobb et al., 2018).

Since it is a Bayesian model, pre-specification of priors could be helpful, but the default ones work well with appropriately scaled data. Convergence check with diagnostic data is an essential step after the model fit. The most common approach is to visually check the trace plot (by using the embedded code – *TracePlot()* in the *bkmr* package).

The output contains the posterior inclusion probabilities (PIPs) and the estimated exposure-response function. The PIPs indicate the importance of exposure variables, while the exposure-response function can be analysed cross-sectionally to produce the single-exposure associations, interaction assessment, and joint health effect estimate. The *bkmr* package include visualisation functions embedded in *ggplot*, providing a range of options for result presentation. The presentation and interpretation of results can be flexible. Several recent publications could provide reference for the interpretation of results (Valeri et al., 2017, Bauer et al., 2020). An online tutorial for the package application is also available (https://jenfb.github.io/bkmr/overview.html).

Strengths and limitations

The main benefit of BKMR is its comprehensive outputs that serves most research objectives in environmental mixture modelling (i.e., identification of important exposures and individual effect, joint health effect estimate, and informal interaction assessment). BKMR also produces an estimate of model uncertainty (in the form of credible intervals). It would be convenient for researchers to answer their research questions by applying this model alone, without the need to apply multiple models at one time. The second advantage is the ability to adequately accommodate interaction and nonlinearity in a single model. This is achieved with by using a kernel function (e.g., Gaussian kernel) (Bobb et al., 2015). Finally, BKMR provides a way to account for highly correlated exposures with the implementation of hierarchical variable selection. This provides relative measurements of exposure variables importance within a pre-specified group and among groups, supporting the variable selection.

Limitations include the requirement of a not too large sample size and the model instability when the number of exposure variables gets very large (>50). BKMR is also computationally expensive (i.e., slow).

Some useful extensions

One important variation of BKMR is Lagged kernel machine regression (LKMR) (Liu et al., 2018b). Similar to BKMR, LKMR can account for complex non-linear and non-additive effect of the mixture, additionally, identify critical exposure windows of mixtures. For working-life exposome research, LKMR could be useful to estimate the health effect of time-varying exposures to multi-pollutant mixtures from cohort studies. It has demonstrated its efficiency in high serial correlation among the time-varying exposures by considering each time window separately (Liu et al., 2018b). The computational efficiency of LKMR has been later improved by an extension of MFVB-LKMR with a procedure called “mean field variational approximation” which reduce the time of running the algorithm to just minutes (Liu et al., 2018a). Its R codes can be found on https://github.com/shelleyhliu/VB-LKMR-Simulations. However, both LKMR and MFVB-LKMR can only be used for studies with a small number of measured time windows (e.g., blood and other biomarkers), while some semi-continuous exposure measurement such as weekly air pollution measurement may not be suitable.

Random Forest

Introduction

Random forest analysis is a non-parametric approach for classification and regression problems (Breiman, 2001). It focuses on prediction of the outcome rather than understanding of the underlying process. While it is a complex model, the basic idea is elegant and follows the simple "divide and conquer" principle (i.e., ensemble): sample fractions of the data, grow a randomised tree predictor on each small piece, then combine the trees together to make prediction based on all individual trees (Biau and Scornet, 2016). This simple but effective strategy leads to its successful applications in addressing various practical problems.

However, epidemiological questions are causal inference by nature. Random Forest can therefore be used as a variable selection method to identify important exposures, without giving the parametric estimate of effect size. Such importance is often evaluated by extent of increase in prediction error when the variable is rearranged, where little error in prediction accuracy implies low importance (Breiman, 2001). Gini index is commonly used to indicate the variable importance.

Application of random forest has been limited in occupational epidemiological studies, where among only a handful of examples, Faramawi and colleagues (Faramawi et al., 2021) investigated which occupational risk factors were associated with increased pancreatic cancer risk in a case-control study for poultry plant workers.

Implementation

There are many R package options for random forest to make the implementation friendly. Except the "classic" *randomForest* package (RColorBrewer and Liaw, 2018), there are *VSURF* (Genuer et al., 2015), ranger (Wright and Ziegler, 2015), and many others. Some studies have compared the performance of those R packages with different types of datasets (Speiser et al., 2019), which suggested that the *VSURF* R package performs better than other packages in general, compared with other packages that were compared with. The packages *varSelRF* (Diaz-Uriarte and Diaz-Uriarte, 2017), *Boruta (Kursa et al., 2020)*, and *ranger* are also suited for datasets with many variables. Details of the implementation process can be found in respective R package description, and for popular R packages, there are tutorials available for readers to follow (*ranger* R package for example: https://uc-r.github.io/random_forests).

Strengths and limitations

One important advantage of random forest is the high performance in dealing with high-dimensional data (Biau and Scornet, 2016). Random forest also demonstrate capacity to capture complex nonlinearities and interaction (Breiman, 2001). Moreover, the implementation of random forest is very user-friendly, and the parameters are easy to tune.

The high prediction accuracy by random forest is at the expense of a lower interpretability compared with conventional classification and regression tree (CART) (James et al., 2013). The results from random forest can only indicate the variable importance related to outcome, without any further details regarding the extent of effect, the interaction and dose-response relations. Random forests are found to be biased while dealing with categorical variables, and are usually computationally demanding (Strobl et al., 2007).

Some useful extensions

Random survival forests is a useful extension from random forests method to analyse right-censored survival data (Ishwaran et al., 2008). Several packages are available such as *randomForestSRC* (tutorial available https://luminwin.github.io/randomForestSRC/articles/getstarted.html). An example of application of random survival forest is where Dietrich and colleagues (Dietrich et al., 2016) investigated the disease-associated variables in complex data (with time-to-event outcome, high dimensional metabolomics, and prospective cohort design), and the results demonstrate random survival forests method as a promising approach for such type of data.

LASSO (Least absolute shrinkage and selection operator)

Introduction

Based on the common approach from ordinary least square (OLS) that minimise the residual sum of squares, LASSO performs variable selection by imposing a shrinkage penalty on the size of coefficients towards zero – the l_2-norm (Tibshirani, 1996). Such constrain leads to some unimportant coefficients that are exactly 0 and hence gives more interpretable models. LASSO finds coefficients by minimising the following quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

where i stands for observation index i = 1, 2, …, n_i; j (j = 1, 2, …, p) stands for the number of variable indexes; the first part is the RSS (residual sum of squares) that is the same with what ordinary least square tries to minimise, and the λ represents the penalty term that forces some of the coefficients to be exactly zero when λ is sufficient large ($\lambda \geq 0$). The tuning parameter λ can be determined with cross-validation.

Since it was developed in 1996, it has been widely applied in many fields of research and many extensions have been further developed. Similar to random forest, LASSO has been mainly used to identify exposure variables that are associated with certain health outcomes in epidemiological studies. For example, Zhang and colleagues applied LASSO to identify metal components associated with cardiac autonomic responses among welders (Zhang et al., 2017). An extensive simulation study also demonstrated the penalised approaches (i.e., LASSO and Elastic Net) performs particularly well for correlated exposures in case-control studies (Lenters et al., 2018).

Implementation

The R package *glmnet* offers a simple and fast implementation of LASSO algorithm. It has also been widely used to fit generalised linear and similar models (e.g., Elastic net) via penalised maximum likelihood (Friedman et al., 2021). A comprehensive guide of the package can be followed via https://glmnet.stanford.edu/articles/glmnet.html. Cross-validation can also be implemented with the *cvfit()* command to determine an appropriate λ for our model.

A proper selection of the penalisation parameter λ is an important part of LASSO adjustment. It control the number of variables to be selected (degree of shrinkage) and model bias (Lahiri, 2021). For high-dimensional data, the quality of the variable selection can be improved with stability selection

(Meinshausen and Bühlmann, 2010). Based on the subsampling in combination with high-dimensional selection algorithms, stability selection can provide finite sample control for the per-family error rate, thus offering a transparent principle to choose a proper amount of regularisation or penalisation (i.e., λ.) (Meinshausen and Bühlmann, 2010). The stability selection can be easily implemented with R package *stabs* (Hofner et al., 2017) (https://cran.r-project.org/web/packages/stabs/stabs.pdf).

Strengths and limitations

Compared with other two methods, LASSO is advantageous in terms of producing interpretable results and simple implementation. The algorithm produces the effect estimates for the selected exposure variables, which fits well for epidemiological interpretation. LASSO also demonstrate robustness to multicollinearity, and the ability to accommodate different outcome types. It was originally suggested that the LASSO would be best for dataset with a small to moderate number of exposure variables with moderate-sized effects (Tibshirani, 1996).

The pure LASSO is also limited by its high false discovery rate, although this can be partly addressed with stability selection as mentioned previously. For a group of exposure variables with high pair-wise correlation, LASSO tends to select one variable completely at random. Additionally, pure LASSO algorithm assumes linear exposure-response relationships and no interaction, which could be unrealistic for real-world epidemiological data. Several extensions were developed to partly account for those limitations.

Some useful extensions

Elastic net is one important extension that uses a weighted sum of lasso and ridge regression penalties to improve the algorithm performance (Zou and Hastie, 2005). One added value of the elastic net is the algorithm accounts for a grouping effect where strongly correlated variables tend to be in or out of the model together, thus reduce the false discovery rate (Zou and Hastie, 2005). Elastic net is also reported to perform better for high-dimensional data (number of variables (*p*) much higher than number of observation (*n*)) compared with LASSO (Zou and Hastie, 2005). One example of its application is that Zhang and colleagues studied the association between some highly correlated environmental chemical contaminants and birth weight using elastic net regression (Zhang et al., 2017). Implementation of elastic net can also be performed with the *glmnet* package.

Hierarchical group-Lasso regularization is an extension for learning linear interaction models that satisfy strong hierarchy (interaction can be present only if both of its main effects are present) (Lim

and Hastie, 2015). It can model pairwise interactions for both categorical and continuous variables and shows good performance for high-dimensional data (e.g., genome-wide association study). Implementation of this approach can be realised with the R package *glinternet* (Michael, 2021). Another extension is generalized additive model selection (GAMSEL) which can be used to model low-complexity curves – an approach to accommodate nonlinear effect (Chouldechova and Hastie, 2015). This derivative can be implemented with *gamsel* package in R (Alexandra, 2018).

Table 1. Overview of the three selected methods for multiple exposure modelling (BKMR, Random Forest and LASSO)

Method	Analytical objective	Outcome type	Strengths	Weaknesses	R packages	Extension
BKMR	exposure-response surface estimate	continuous; categorical	- address multiple research questions - accommodate interaction and nonlinearity - provide a way to account for correlated exposures	- require only moderate sample size - limited number of exposure variables - computationally expensive	*bkmr*	LKMR and MFV-LKMR extension for longitudinal data
Random Forest	variable selection	continuous; categorical	- account for interaction and nonlinearity - perform well for high-dimensional data - easy to tune parameters	- results are hard to interpret - computational expensive	*randomForest; ranger*	Random survival forests; XGBoost
LASSO	variable selection	continuous; categorical	- able to control for confounders - flexible data input type - extensions are available to account for grouped and correlated exposures, interaction, and nonlinearity	- additional tools are needed to make statistical inference - assume linear exposure-response relationships	*glmnet*	Elastic net; GAMSEL; Hierarchical group-Lasso regularization

References

AGIER, L., PORTENGEN, L., CHADEAU-HYAM, M., BASAGANA, X., GIORGIS-ALLEMAND, L., SIROUX, V., ROBINSON, O., VLAANDEREN, J., GONZALEZ, J. R., NIEUWENHUIJSEN, M. J., VINEIS, P., VRIJHEID, M., SLAMA, R. & VERMEULEN, R. 2016. A Systematic Comparison of Linear Regression-Based Statistical Methods to Assess Exposome-Health Associations. *Environ Health Perspect,* 124**,** 1848-1856.

ALEXANDRA, C. T., HASTIE; VITALIE, SPINU 2018. gamsel: Fit Regularization Path for Generalized Additive Models}.

BARRERA-GOMEZ, J., AGIER, L., PORTENGEN, L., CHADEAU-HYAM, M., GIORGIS-ALLEMAND, L., SIROUX, V., ROBINSON, O., VLAANDEREN, J., GONZALEZ, J. R., NIEUWENHUIJSEN, M., VINEIS, P., VRIJHEID, M., VERMEULEN, R., SLAMA, R. & BASAGANA, X. 2017. A systematic comparison of statistical methods to detect interactions in exposome-health associations. *Environ Health,* 16**,** 74.

BAUER, J. A., DEVICK, K. L., BOBB, J. F., COULL, B. A., BELLINGER, D., BENEDETTI, C., CAGNA, G., FEDRIGHI, C., GUAZZETTI, S., OPPINI, M., PLACIDI, D., WEBSTER, T. F., WHITE, R. F., YANG, Q., ZONI, S., WRIGHT, R. O., SMITH, D. R., LUCCHINI, R. G. & CLAUS HENN, B. 2020. Associations of a Metal Mixture Measured in Multiple Biomarkers with IQ: Evidence from Italian Adolescents Living near Ferroalloy Industry. *Environ Health Perspect,* 128**,** 97002.

BEHRENS, J. T. 1997. Principles and procedures of exploratory data analysis. *Psychological Methods,* 2**,** 131-160.

BELLINGER, C., MOHOMED JABBAR, M. S., ZAIANE, O. & OSORNIO-VARGAS, A. 2017. A systematic review of data mining and machine learning for air pollution epidemiology. *BMC Public Health,* 17**,** 907.

BI, Q., GOODMAN, K. E., KAMINSKY, J. & LESSLER, J. 2019. What is Machine Learning? A Primer for the Epidemiologist. *Am J Epidemiol,* 188**,** 2222-2239.

BIAU, G. & SCORNET, E. 2016. A random forest guided tour. *Test,* 25**,** 197-227.

BILLIONNET, C., SHERRILL, D., ANNESI-MAESANO, I. & STUDY, G. 2012. Estimating the health effects of exposure to multi-pollutant mixture. *Ann Epidemiol,* 22**,** 126-41.

BOBB, J. F. 2017. bkmr: Bayesian Kernel Machine Regression.

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. *Environ Health,* 17**,** 67.

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.

BOSSON-RIEUTORT, D., SARAZIN, P., BICOUT, D. J., HO, V. & LAVOUE, J. 2020. Occupational Co-exposures to Multiple Chemical Agents from Workplace Measurements by the US Occupational Safety and Health Administration. *Ann Work Expo Health,* 64**,** 402-415.

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**,** A6-9.

BREIMAN, L. 2001. Random forests. *Machine Learning,* 45**,** 5-32.

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.

CHIU, Y. H., BELLAVIA, A., JAMES-TODD, T., CORREIA, K. F., VALERI, L., MESSERLIAN, C., FORD, J. B., MINGUEZ-ALARCON, L., CALAFAT, A. M., HAUSER, R., WILLIAMS, P. L. & TEAM, E. S. 2018. Evaluating effects of prenatal exposure to phthalate mixtures on birth weight: A comparison of three statistical approaches. *Environ Int,* 113**,** 231-239.

CHOULDECHOVA, A. & HASTIE, T. 2015. Generalized additive model selection. *arXiv preprint arXiv:1506.03850*.

DIAZ-URIARTE, R. & DIAZ-URIARTE, M. R. 2017. Package 'varSelRF'.

DIETRICH, S., FLOEGEL, A., TROLL, M., KUHN, T., RATHMANN, W., PETERS, A., SOOKTHAI, D., VON BERGEN, M., KAAKS, R., ADAMSKI, J., PREHN, C., BOEING, H., SCHULZE, M. B., ILLIG, T., PISCHON, T., KNUPPEL, S., WANG-SATTLER, R. & DROGAN, D. 2016. Random Survival Forest in practice: a method for modelling complex metabolomics data in time to event analysis. *Int J Epidemiol,* 45**,** 1406-1420.

FARAMAWI, M. F., ABOUELENEIN, S. & JOHNSON, E. 2021. A case-control study of occupational risk factors for pancreatic cancer in poultry plant workers: a random forest approach. *J Public Health (Oxf)*.

FOURNEAU, C., SANCHEZ, M., PEROUEL, G., FRERY, N., COUTROT, T., BOULANGER, G., COURRIER, B., PERNELET-JOLY, V. & BASTOS, H. 2021. The French 2016-2020 National Occupational Health Plan: a better understanding of multiple exposures. *Environnement Risques Santé,* 20**,** 377-382.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. & NARASIMHAN, B. 2021. Package 'glmnet'. *CRAN R Repositary*.

GENUER, R., POGGI, J.-M. & TULEAU-MALOT, C. 2015. VSURF: an R package for variable selection using random forests. *The R Journal,* 7**,** 19-33.

GIBSON, E. A., GOLDSMITH, J. & KIOUMOURTZOGLOU, M. A. 2019a. Complex Mixtures, Complex Analyses: an Emphasis on Interpretable Results. *Curr Environ Health Rep,* 6**,** 53-61.

GIBSON, E. A., NUNEZ, Y., ABUAWAD, A., ZOTA, A. R., RENZETTI, S., DEVICK, K. L., GENNINGS, C., GOLDSMITH, J., COULL, B. A. & KIOUMOURTZOGLOU, M. A. 2019b. An overview of methods to address distinct research questions on environmental mixtures: an application to persistent organic pollutants and leukocyte telomere length. *Environ Health,* 18**,** 76.

GREENLAND, S. 2008. Multiple comparisons and association selection in general epidemiology. *Int J Epidemiol,* 37**,** 430-4.

GUILLIEN, A., CADIOU, S., SLAMA, R. & SIROUX, V. 2021. The Exposome Approach to Decipher the Role of Multiple Environmental and Lifestyle Determinants in Asthma. *Int J Environ Res Public Health,* 18.

HOFNER, B., HOTHORN, T. & HOFNER, M. B. 2017. Package 'stabs'.

HUANG, H., WANG, A., MORELLO-FROSCH, R., LAM, J., SIROTA, M., PADULA, A. & WOODRUFF, T. J. 2018. Cumulative Risk and Impact Modeling on Environmental Chemical and Social Stressors. *Curr Environ Health Rep,* 5**,** 88-99.

ISHWARAN, H., KOGALUR, U. B., BLACKSTONE, E. H. & LAUER, M. S. 2008. Random survival forests. *The annals of applied statistics,* 2**,** 841-860.

JAMES, G., WITTEN, D., HASTIE, T. & TIBSHIRANI, R. 2013. *An introduction to statistical learning*, Springer.

KINO, S., HSU, Y. T., SHIBA, K., CHIEN, Y. S., MITA, C., KAWACHI, I. & DAOUD, A. 2021. A scoping review on the use of machine learning in research on social determinants of health: Trends and research prospects. *SSM Popul Health,* 15**,** 100836.

KURSA, M. B., RUDNICKI, W. R. & KURSA, M. M. B. 2020. Package 'Boruta'.

LAHIRI, S. N. 2021. Necessary and sufficient conditions for variable selection consistency of the LASSO in high dimensions. *The Annals of Statistics,* 49**,** 820-844.

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.

LE BORGNE, F., CHATTON, A., LEGER, M., LENAIN, R. & FOUCHER, Y. 2021. G-computation and machine learning for estimating the causal effects of binary exposure statuses on binary outcomes. *Sci Rep,* 11**,** 1435.

LENTERS, V., VERMEULEN, R. & PORTENGEN, L. 2018. Performance of variable selection methods for assessing the health effects of correlated exposures in case-control studies. *Occup Environ Med,* 75**,** 522-529.

LIM, M. & HASTIE, T. 2015. Learning interactions via hierarchical group-lasso regularization. *J Comput Graph Stat,* 24**,** 627-654.

LIU, S. H., BOBB, J. F., CLAUS HENN, B., SCHNAAS, L., TELLEZ-ROJO, M. M., GENNINGS, C., ARORA, M., WRIGHT, R. O., COULL, B. A. & WAND, M. P. 2018a. Modeling the health effects of time-varying complex environmental mixtures: Mean field variational Bayes for lagged kernel machine regression. *Environmetrics,* 29.

LIU, S. H., BOBB, J. F., LEE, K. H., GENNINGS, C., CLAUS HENN, B., BELLINGER, D., AUSTIN, C., SCHNAAS, L., TELLEZ-ROJO, M. M., HU, H., WRIGHT, R. O., ARORA, M. & COULL, B. A. 2018b. Lagged kernel machine regression for identifying time windows of susceptibility to exposures of complex mixtures. *Biostatistics,* 19**,** 325-341.

LU, R., DUAN, T., WANG, M., LIU, H., FENG, S., GONG, X., WANG, H., WANG, J., CUI, Z., LIU, Y., LI, C. & MA, J. 2021. The application of multivariate adaptive regression splines in exploring the influencing factors and predicting the prevalence of HbA1c improvement. *Ann Palliat Med,* 10**,** 1296-1303.

MALOVINI, A., BELLAZZI, R., NAPOLITANO, C. & GUFFANTI, G. 2016. Multivariate Methods for Genetic Variants Selection and Risk Prediction in Cardiovascular Diseases. *Front Cardiovasc Med,* 3**,** 17.

MCKENZIE, J. F., EL-ZAEMEY, S. & CAREY, R. N. 2020. Prevalence of exposure to multiple occupational carcinogens among exposed workers in Australia. *Occup Environ Med,* 78**,** 211-217.

MEINSHAUSEN, N. & BÜHLMANN, P. 2010. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* 72**,** 417-473.

MICHAEL, L. T., HASTIE 2021. glinternet: Learning Interactions via Hierarchical Group-Lasso Regularization.

MOLITOR, J., PAPATHOMAS, M., JERRETT, M. & RICHARDSON, S. 2010. Bayesian profile regression with an application to the National Survey of Children's Health. *Biostatistics,* 11**,** 484-98.

N, L., L, K., P, S. & A, B. 2019. Performance of variable selection methods for estimating the non-linear health effects of correlated chemical mixtures. *Environmental Epidemiology,* 3**,** 224-225.

NACAR, S., METE, B. & BAYRAM, A. 2020. Estimation of daily dissolved oxygen concentration for river water quality using conventional regression analysis, multivariate adaptive regression splines, and TreeNet techniques. *Environ Monit Assess,* 192**,** 752.

NIETO, P. J., ANTON, J. C., VILAN, J. A. & GARCIA-GONZALO, E. 2015. Air quality modeling in the Oviedo urban area (NW Spain) by using multivariate adaptive regression splines. *Environ Sci Pollut Res Int,* 22**,** 6642-59.

OSKAR, S. & STINGONE, J. A. 2020. Machine Learning Within Studies of Early-Life Environmental Exposures and Child Health: Review of the Current Literature and Discussion of Next Steps. *Curr Environ Health Rep,* 7**,** 170-184.

PAPATHOMAS, M., MOLITOR, J., RICHARDSON, S., RIBOLI, E. & VINEIS, P. 2011. Examining the joint effect of multiple risk factors using exposure risk profiles: lung cancer in nonsmokers. *Environ Health Perspect,* 119**,** 84-91.

PATEL, C. J., CULLEN, M. R., IOANNIDIS, J. P. & BUTTE, A. J. 2012. Systematic evaluation of environmental factors: persistent pollutants and nutrients correlated with serum lipid levels. *Int J Epidemiol,* 41**,** 828-43.

PEARSON, R. K. 2018. *Exploratory data analysis using R*, CRC Press.

PIRANI, M., BEST, N., BLANGIARDO, M., LIVERANI, S., ATKINSON, R. W. & FULLER, G. W. 2015. Analysing the health effects of simultaneous exposure to physical and chemical properties of airborne particles. *Environ Int,* 79**,** 56-64.

PRONK, A., LOH, M., KUIJPERS, E., ALBIN, M., SELANDER, J., GODDERIS, L., GHOSH, M., VERMEULEN, R., PETERS, S. & MEHLUM, I. S. 2021. S-135 Applying the exposome concept to working-life health: The EU EPHOR project. BMJ Publishing Group Ltd.

RCOLORBREWER, S. & LIAW, M. A. 2018. Package 'randomForest'. *University of California, Berkeley: Berkeley, CA, USA*.

SANTOS, S., MAITRE, L., WAREMBOURG, C., AGIER, L., RICHIARDI, L., BASAGANA, X. & VRIJHEID, M. 2020. Applying the exposome concept in birth cohort research: a review of statistical approaches. *Eur J Epidemiol,* 35**,** 193-204.

SEJBAEK, C. S., PEDERSEN, J., SCHLUNSSEN, V., BEGTRUP, L. M., JUHL, M., BONDE, J. P., KRISTENSEN, P., BAY, H., RAMLAU-HANSEN, C. H. & HOUGAARD, K. S. 2020. The influence of multiple occupational exposures on absence from work in pregnancy: a prospective cohort study. *Scand J Work Environ Health,* 46**,** 60-68.

SONG, Q., LI, R. Z., ZHAO, Y., ZHU, Q. Y., XIA, B., CHEN, S. Q. & ZHANG, Y. H. 2018. Evaluating effects of prenatal exposure to phthalates on neonatal birth weight: Structural equation model approaches. *Chemosphere,* 205**,** 674-681.

SPEISER, J. L., MILLER, M. E., TOOZE, J. & IP, E. 2019. A Comparison of Random Forest Variable Selection Methods for Classification Prediction Modeling. *Expert Syst Appl,* 134**,** 93-101.

STAFOGGIA, M., BREITNER, S., HAMPEL, R. & BASAGANA, X. 2017. Statistical Approaches to Address Multi-Pollutant Mixtures and Multiple Exposures: the State of the Science. *Curr Environ Health Rep,* 4**,** 481-490.

STANIAK, M. & BIECEK, P. 2019. The landscape of r packages for automated exploratory data analysis. *arXiv preprint arXiv:1904.02101*.

STROBL, C., BOULESTEIX, A. L., ZEILEIS, A. & HOTHORN, T. 2007. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics,* 8**,** 25.

SUN, Z., TAO, Y., LI, S., FERGUSON, K. K., MEEKER, J. D., PARK, S. K., BATTERMAN, S. A. & MUKHERJEE, B. 2013. Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environ Health,* 12**,** 85.

TANNER, E., LEE, A. & COLICINO, E. 2020. Environmental mixtures and children's health: identifying appropriate statistical approaches. *Curr Opin Pediatr,* 32**,** 315-320.

TAYLOR, K. W., JOUBERT, B. R., BRAUN, J. M., DILWORTH, C., GENNINGS, C., HAUSER, R., HEINDEL, J. J., RIDER, C. V., WEBSTER, T. F. & CARLIN, D. J. 2016. Statistical Approaches for Assessing Health Effects of Environmental Chemical Mixtures in Epidemiology: Lessons from an Innovative Workshop. *Environ Health Perspect,* 124**,** A227-A229.

TIBSHIRANI, R. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological),* 58**,** 267-288.

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.

VUONG, A. M., YOLTON, K., BRAUN, J. M., LANPHEAR, B. P. & CHEN, A. 2020. Chemical mixtures and neurobehavior: a review of epidemiologic findings and future directions. *Rev Environ Health,* 35**,** 245-256.

WICKHAM, H. & GROLEMUND, G. 2016. *R for data science: import, tidy, transform, visualize, and model data*, " O'Reilly Media, Inc.".

WIRTH, R. & HIPP, J. CRISP-DM: Towards a standard process model for data mining. Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining, 2000. Springer-Verlag London, UK.

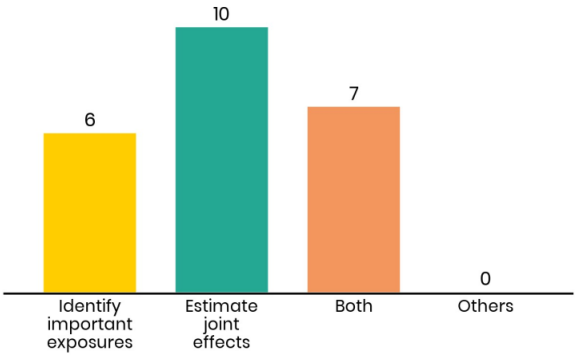
WRIGHT, M. N. & ZIEGLER, A. 2015. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *arXiv preprint arXiv:1508.04409*.

ZHANG, J., CAVALLARI, J. M., FANG, S. C., WEISSKOPF, M. G., LIN, X., MITTLEMAN, M. A. & CHRISTIANI, D. C. 2017. Application of linear mixed-effects model with LASSO to identify metal components associated with cardiac autonomic responses among welders: a repeated measures study. *Occup Environ Med,* 74**,** 810-815.

ZOU, H. & HASTIE, T. 2005. Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology),* 67**,** 301-320.

Appendix A – Feedback collected at EPHOR consortium meeting 12 Nov 2021

Which research question(s) interests you most?



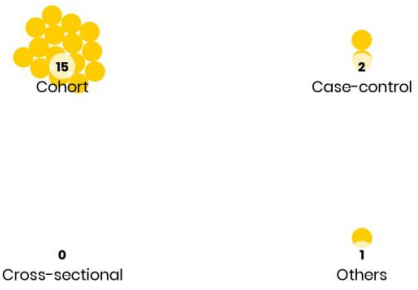
23

What are the other research questions?

Joint effect over time (protracting exposures)

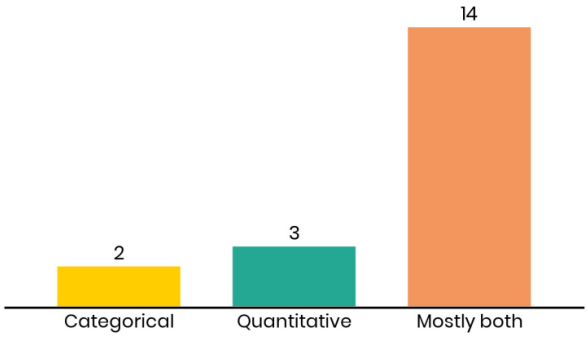
1

Your data are from what study design?



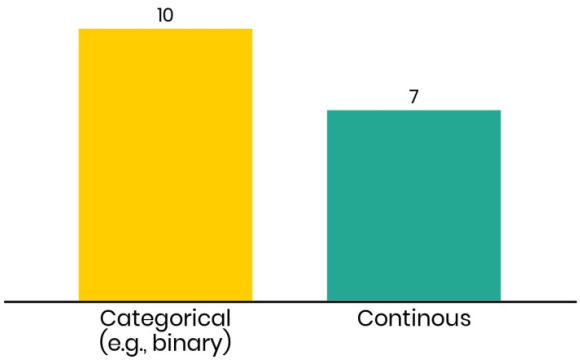
16

What is usually the type of your EXPOSURE data?



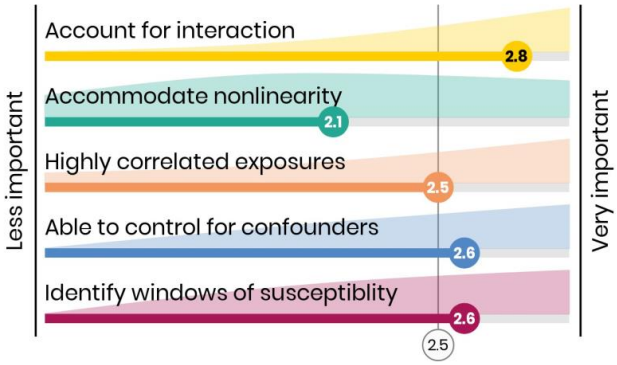
19

What is usually the type of your OUTCOME?



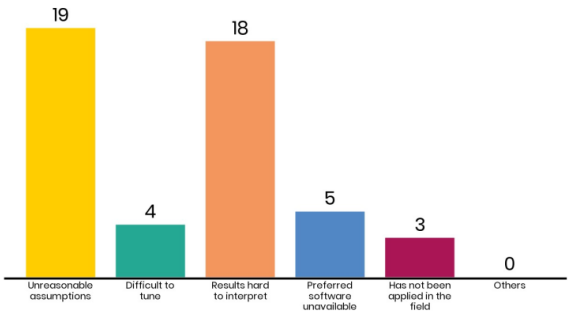
17

Which are important for a model to address?



15

What make you decide NOT to use one method?



21

What are the OTHER important issues for you to consider when applying a method?

Timing of exposure, temporal patterns

Important to be able to stratify analysis

2

Appendix B – Selection of methods for multiple exposure analysis

Method collection

A systematic literature search has been performed to provide the experts with a complete list of methods that are suited for the purpose of the tutorial and the proposed research questions. Keywords including "multiple exposure", "co-exposure", "multipollutant", "multi-pollutant", and "mixture" were searched with Google Scholar, PubMed, Scopus, and Web of Science. The list of methods for expert-rating was mainly synthesized by cross-comparing recent methodological reviews on multiple exposures/exposome/multipollutant mixtures, mostly in the fields of environmental epidemiology and toxicology (Agier et al., 2016, Lazarevic et al., 2019, Stafoggia et al., 2017, Taylor et al., 2016, Kino et al., 2021, Malovini et al., 2016, Guillien et al., 2021, Billionnet et al., 2012, Tanner et al., 2020, Vuong et al., 2020, Santos et al., 2020, Bellinger et al., 2017, Bi et al., 2019, Braun et al., 2016, Gibson et al., 2019a, Huang et al., 2018, Oskar and Stingone, 2020).

In addition to publications gathered through the literature search, several high-impact papers with focus on the application of machine learning have also been reviewed and integrated into the list (Bi et al., 2019, Bellinger et al., 2017, Oskar and Stingone, 2020). Several model comparison papers were also reviewed (Agier et al., 2016, Sun et al., 2013, N et al., 2019, Malovini et al., 2016, Le Borgne et al., 2021, Barrera-Gomez et al., 2017, Chiu et al., 2018, Lenters et al., 2018, Song et al., 2018, Gibson et al., 2019b). To ensure a complete list of methods, we also used the network of papers to capture previously unidentified studies and methods (via https://www.connectedpapers.com/). Flexibility was also given to experts that allows them to add previously unlisted methods while giving scores.

This process produced a total of 49 statistical methods (including their derivatives), see complete list in Appendix D.

Evaluation criteria

The criteria were developed as a guide for experts to formally evaluate the methods in terms of performance, availability, and interpretability. Efforts in comprehensive literature review, rounds of discussion with experts, and external feedback were made to devise the criteria as follows:

1) Performance

The performance of a method indicates the inherent capacity to handle complex exposure-outcome relationships by accounting for challenges such as multicollinearity. This criterion would also reflect the method's ability to address questions that are important to environmental/occupational epidemiologist such as interaction and nonlinearity. In addition, the method performance would also be assessed under different research questions and datasets with model comparison studies. This domain would essentially answer: "What can the model do?" and "On the relative scale, how good the model is?".

	-/0/+	Note
1.1 Does the model rely on assumptions that are widely applicable?		
1.2 Can indicate model uncertainty (e.g., FDR)?		
1.3 Can distinguish confounders from other covariates?		
1.4 Reasonable handling of multicollinearity? [e.g., -1 for not helpful at all; 0 for average (e.g., randomly drop/select correlated ones); 1 for methods with more justified approaches.]		
1.5 Can account for nonlinearity in a principled way?		
1.6 Can identify time windows of susceptibility to exposures?		
1.7 Can assess two-way interaction in a principled way?		
1.8 Can assess non-linear and/or high-dimensional interaction effect?		

2) Availability

The availability refers to how accessible the method can be applied by non-statisticians (i.e., whether the readers can easily find other useful resources beyond the tutorial). A widely applied method would imply its popularity, accessibility, and most likely, its usefulness for the research questions applied. Moreover, more studies that applied the method would mean more examples are available for readers to support the interpretation of the produced results.

	-/0/+	Note
2.1 Is there an existing package/program/codes in R/STATA/SAS/WINBUGS/GitHub for this method?		
2.2 Is there an existing tutorial/textbook chapter for this method?		
2.3 Has there been application(s) in environmental epidemiology-related studies?		

3) Interpretability

Interpretability indicates the extent to which the results can be interpreted and applied in policymaking. Policymakers would first want to know the risk of bias associated with the methods (whether the produced results can be trusted, which can be reflected by the model's ability to adjust for confounders, or the justification of parameter selection). This criterion essentially indicates the feasibility of the method being integrated in risk assessment process.

	-/0/+	Note
3.1 Can indicate importance of individual exposures? [0 for list of important variables; 1 for providing a quantified "importance" of the exposures]		
3.2 Can provide summarised effect estimates and associated confidence intervals?		

3.3 Are results insensitive to model parameter choices? [-1 for very difficult tuning process, 0 for sensible tuning, 1 for no need of tuning]		

Appendix C – Exploratory Data Analysis

For most statistical analyses, it is recommended to perform exploratory data analysis (or pre-analysis) before main analyses. A well-conducted pre-analysis would inform researchers about data structure and support method selection.

In general, descriptive analyses include the following parts:

a) to identify missing data and choose the right replacement strategy

b) to examine distributions of variables and select transformation method

c) to identify outliers and decide strategies (e.g., IQR and PCA)

In addition to those tasks, for "multiple exposures" research questions, one could pay more attention to:

d) to summary cohort demographic characteristics

e) to explore patterns of correlation between exposures

f) to assess the unadjusted and adjusted associations between each exposure and outcome

More comprehensive discussions of exploratory data analyses have been presented elsewhere (Pearson, 2018, Behrens, 1997, Wickham and Grolemund, 2016, Wirth and Hipp, 2000). Several R packages are available for automatic exploratory analyses such as *DataExplorer* and *visdat,* and a complete list and comparison of different R packages for automatic exploratory analyses have been described (Staniak and Biecek, 2019).We also recommend researchers to employ visualisations to explore directions of associations and possible nonlinear relationships.

Appendix D – Full list of methods for multiple exposure analysis

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
Standard methods													
Single-exposure models (a.k.a. exposome/environment-wide association study ExWAS)	+	+	+	+	-	+	+	-	-	--	1	many	Single exposure analyses could be implemented in any chosen model (incl. models using splines). Interpretation of results is extremely difficult (impossible) if there is the potential for co-exposure confounding. Interactions could be included in e.g., bi-pollutant models.
Full multiple regression models (including models using regression splines)	+	+	-	+	0	-	+	-	+	+	2	many	Only useful when there is a limited number of exposures (p<<n) and there is no strong multi-collinearity. If not, precision will be low and interpretation becomes very hard.
Including univariate or bivariate smoothing splines	+	+	-	+	0	-	+	-	+	+	3	mgcv	See above. This is actually a form of penalized regression.
Methods estimating "mixture effects"													BKMR is included as a machine-learning method
Weighted Quantile Sum (WQS) regression													The estimated "mixture" effect is a weighted average of exposure-specific effects and does not account for interaction.

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
Generalized WQS	0	0	0	-	-	-	+	+	-	0/-	5	gWQS	No estimate of precision of weights is provided
Bayesian Generalized WQS	0	0	0	-	-	-	+	+	+	0/-	5	BayesGWQS	Uses JAGS. Bayesian framework allows uncertainty estimation of weights
Lagged WQSR	0	0	0	-	-	+	+	+	-	0/-	5	lwqs	Adaptation of gWQS for longitudinal data
BWS (Bayesian weighted sums)	0	0	0	-	-	-	+	+	-	0/-	5	rjags	Implemented with JAGS script
Non-parametric (semi-parametric) methods													
MARS (Multivariate adaptive regression spline)	0	0	0	+	+	-	+	-	0	+	7	earth	Relatively good for models with not too many predictors. Includes a variable importance measure.
NPB	-	0	-	-	+	-	+	-	+	+	5	mmpack	Bayesian model with DPP, only for continuous (gaussian) outcomes
DSA	+	0	0	0	0	-	+	-	0/-	+	6	dsa	package no longer available (but partDSA is)
Methods relying on dimension reduction													Usefulness depends on whether the latent variables can be interpreted. For the unsupervised methods these are the factors that explain the correlations.
Based on PCA													

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
PCA regression	+	+	+	-	-	-	+	+	-	-	4	prcomp, kernlab	PCA and extensions like non-linear (kernel) PCA.
NSPCA (non-negative sparse PCA)	+	+	+	-	-	-	+	+	-	-	4	nsprcomp	A bit easier to interpret, but more restricted in the type of PCA models that can be fitted.
SPCA (supervised PCA)	+	-	+	-	-	-	-	+	-	-	4	superpc	Not so familiar with this
ECM (Exposure continuum mapping)	+	+	+	0	0	-	-	-	-	-	4	ECM	Self-organising maps followed by spatial analysis.
Partial Least Squares (PLS) regression													
PLS regression	0	0	+	-	-	-	-	+	-	-	5	mixOmics	
sparse PLS regression	0	0	+	-	-	-	-	+	-	-	6	mixOmics	
Structural Equation Modelling (SEM)													
Maximum Likelihood SEM	0	0	0	-	-	-	+	+	-	0	5	lavaan	Single factors assumed for each group of variables.
Bayesian Factor Analysis (BFA)	0	0	-	-	-	-	+	+	+	0	5	blavaan	Bayesian methods tend to be computationally demanding
Bayesian Profile Regression (BPR)	0	0	-	-	+	-	-	-	+	-	5	PreMiuM	

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
Penalized regression methods													
Lasso/Elastic Net	+	0	+	-	-	-	+	-	0	+	7	glmnet	Pure lasso methods tend to select too many variables (geared towards prediction), but stability selection can be used to reduce false positive rates.
Grouped Lasso	+	0	+	-	-	-	+	+	0	+	7	grplasso	
Interaction Lasso	0	0	+	-	+	-	+	-	0	+	7	glinternet	strong hierarchy only
Hierarchical Interaction Lasso	0	0	+	-	+	-	+	-	0	+	7	hierNet	both strong and weak hierarchy
Non-linear Lasso	0	0	+	+	-	-	+	-	0	+	7	gamsel	based on group-lasso
Both non-linear and interaction effects	+	0	-	+	+	-	+	-	0	0	6	plsmselect	Combination of mgcv-based selection and glmnet, only suitable with a small number of non-linear effects.
PCA Lasso	+	0	+	-	-	-	+	+	0	0	6	pcLasso	Lasso for PCA regression
Generalized Linear Mixed Model (GLMM) lasso	0	+	+	+	-	-	+	-	0	+	7	glmmLasso	
Alternative frequentist penalties (MCP)	0	0	+	-	-	-	+	+	0	+	7	ncvreg, grpreg, grpregOverlap	Non-convex penalties have lower error rates, but are more difficult to computationally and have not been

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
													implemented for models with interactions or non-linear models.
Bayesian horseshoe regression	0	+	0	-	-	-	+	-	+	+	7	brms	Computationally demanding, but avoids the problem of using one penalty to achieve both selection and estimation.
Variable selection methods													
Best subsets	0	0	-	-	-	-	+	-	-	0	5	bestsubset	on Github
Stepwise selection	+	+	-	-	0	-	+	-	-	0	5	base	
Bayesian model averaging (BMA)	+	0	0/+	-	0	-	+	-	+	+	6	BAS	Limited to linear models only
R2GUESS	-	0	+	-	0	-	0	-	+	+	7	R2GUESS	Only available for gaussian outcomes
Structured additive regression models using spike-and-slab variable selection	+	+	0/-	+	+	-	+	-	+	+	7	spikeSlabGAM	Sampling method struggles with strongly correlated variables.
Non-parametric Varying-coefficient models using spike-and-slab variable selection	-	+	0/-	-	-	+	0	-	+	+	6	NVCSSL	Only available for continuous outcomes
Machine learning algorithms													

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
CART	0	0	+	+	+	-	0	-	0	0	6	rpart	Needs more tuning than random forest
CART for survival analysis	+	0	+	+	+	-	0	-	0	0	6	LongCART	
Random forests	0	0	+	+	+	-	0	-	0	0	8	ranger, many others	Easy to tune
RF for survival analysis	+	0	+	+	+	-	0	-	0	0	7	randomForestSRC	
Boosted trees (incl. stochastic gradient boosting)	+	0	+	+	+	-	0	-	0	0	8	xgboost	May be difficult to tune
Boosted regression models (incl. GAMBOOST)	+	0	+	+	+	-	0	-	0	0	7	mboost	Not so useful for interactions
Support vector machines (SVM)													Geared towards classification, but can be used for regression also. Works well with large p, but not with large n (although there are implementations that try to deal with that, e.g., github liquidSVM).
SVM for classification and regression	+	0	0	+	+	-	0	-	0	0	7	kernlab	
SVM for survival outcomes	+	0	0	+	+	-	0	-	0	0	7	survivalsvm	
Kernel semi-parametric models for continuous outcomes	-	0	0	+	+	-	0	-	0	0	7	KPSM	

Method	outcome	design	high-dim	non-linear	interaction	temporal	confounder	grouping	uncertainty	interpretation	Overall	R packages	Comments
Bayesian Kernel Machine Regression (BKMR)	0	?	-	+	+	-	+	+	+	+	8	bkmr	Method of choice for smaller datasets with <50 or so exposures
Lagged Kernel Machine Regression (LKMR)	0	?	-	+	+	+	+	+	+	+	7	bkmrdlm	BKMR for distributed lag models
Bayesian Multiple Index Models (BMIM)	0	?	-	-	-	-	+	+	+	0	6	bsmim2 (github)	BKMR with index functions (WQS)
Artificial neural networks (ANN)	+	0	+	+	+	+	0	-	0	0	7	keras, neuralnet	Can be very difficult to tune
Super Learner algorithm	NA	NA	NA	NA	NA	NA	NA	NA	NA	-	-	SuperLearner	Heterogeneous ensemble method (i.e., using different types of models) and therefore as good as the methods that make up the ensemble. Computationally demanding.

Note: generally, "+" means the model performs well for that criterion; "0" means moderate performance, and "-1" indicates its limited ability for the corresponding criterion. Particularly for "outcome" criterion, "+" means model includes GLM+survival; "0" means model includes at least binomial+gaussian; and "-" means either one of these. For "design", "+": allows for clustering; "0": no mixed models; and "-1": only prospective designs. "?" indicate uncertainty. "High-dim" stands for the model performance for large datasets. Other columns correspond to the previously described criteria.