Beyond the single-outcome approach: A comparison of outcome-wide analysis methods for exposome research

Augusto Anguita-Ruiz, Ines Amine, Nikos Stratakis, Lea Maitre, Jordi Julvez, Jose Urquiza, Chongliang Luo, Mark Nieuwenhuijsen, Cathrine Thomsen, Regina Grazuleviciene, et al.

▶ **To cite this version:**

Augusto Anguita-Ruiz, Ines Amine, Nikos Stratakis, Lea Maitre, Jordi Julvez, et al.. Beyond the single-outcome approach: A comparison of outcome-wide analysis methods for exposome research. Environment International, 2023, 182 (3), pp.108344. 10.1016/j.envint.2023.108344 . hal-04326209

HAL Id: hal-04326209
https://hal.science/hal-04326209

Submitted on 6 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Environment International 182 (2023) 108344

Contents lists available at ScienceDirect

Environment International

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

Full length article

Beyond the single-outcome approach: A comparison of outcome-wide analysis methods for exposome research

Augusto Anguita-Ruiz [a,b], Ines Amine [c], Nikos Stratakis [a], Lea Maitre [a,d,e], Jordi Julvez [a,b,f,m], Jose Urquiza [a], Chongliang Luo [g], Mark Nieuwenhuijsen [a,d,e], Cathrine Thomsen [h], Regina Grazuleviciene [i], Barbara Heude [j], Rosemary McEachan [k], Marina Vafeiadi [l], Leda Chatzi [l], John Wright [k], Tiffany C. Yang [k], Rémy Slama [c], Valérie Siroux [c], Martine Vrijheid [a,d,e], Xavier Basagaña [a,d,e,*]

[a] ISGlobal, 08003 Barcelona, Spain
[b] CIBEROBN (CIBER Physiopathology of Obesity and Nutrition), Instituto de Salud Carlos III, 28029 Madrid, Spain
[c] University Grenoble Alpes, Inserm U 1209, CNRS UMR 5309, Team of Environmental Epidemiology Applied to the Development and Respiratory Health, Institute for Advanced Biosciences, 38000 Grenoble, France
[d] Universitat Pompeu Fabra (UPF), 08003 Barcelona, Spain
[e] CIBER Epidemiología y Salud Pública (CIBERESP), 28029 Madrid, Spain
[f] Epidemiology and Environmental Health Joint Research Unit, Foundation for the Promotion of Health and Biomedical Research in the Valencian Region, FISABIO-Public Health, FISABIO-Universitat Jaume I-Universitat de València, Av. Catalunya 21, 46020 Valencia, Spain
[g] Division of Public Health Sciences, Washington University School of Medicine in St. Louis, 600 S Taylor Ave, St. Louis, MO 63110, USA
[h] Department of Food Safety, Norwegian Institute of Public Health (NIPH), Oslo, Norway
[i] Department of Environmental Science, Vytautas Magnus University, 44248 Kaunas, Lithuania
[j] Université Paris Cité and Université Sorbonne Paris Nord, Inserm, INRAE, Center for Research in Epidemiology and StatisticS (CRESS), F-75004 Paris, France
[k] Bradford Institute for Health Research, Bradford Teaching Hospitals NHS Foundation Trust, Bradford, UK
[l] Department of Social Medicine, School of Medicine, University of Crete, Heraklion, Crete, Greece
[m] Institut d'Investigació Sanitària Pere Virgili (IISPV), Clinical and Epidemiological Neuroscience Group (NeuroÈpia), 43204 Reus (Tarragona), Catalonia, Spain

ARTICLE INFO

Handling Editor: Adrian Covaci

Keywords:
Outcome-wide analysis
Multi-outcome analysis
Exposome analysis
Environmental epidemiology

ABSTRACT

Outcome-wide analysis can offer several benefits, including increased power to detect weak signals and the ability to identify exposures with multiple effects on health, which may be good targets for preventive measures. Recently, advanced statistical multivariate techniques for outcome-wide analysis have been developed, but they have been rarely applied to exposome analysis. In this work, we provide an overview of a selection of methods that are well-suited for outcome-wide exposome analysis and are implemented in the R statistical software. Our work brings together six different methods presenting innovative solutions for typical problems arising from outcome-wide approaches in the context of the exposome, including dependencies among outcomes, high dimensionality, mixed-type outcomes, missing data records, and confounding effects. The identified methods can be grouped into four main categories: regularized multivariate regression techniques, multi-task learning approaches, dimensionality reduction approaches, and bayesian extensions of the multivariate regression framework. Here, we compare each technique presenting its main rationale, strengths, and limitations, and provide codes and guidelines for their application to exposome data. Additionally, we apply all selected methods to a real exposome dataset from the Human Early-Life Exposome (HELIX) project, demonstrating their suitability for exposome research. Although the choice of the best method will always depend on the challenges to be faced in each application, for an exposome-like analysis we find dimensionality reduction and bayesian methods such as reduced rank regression (RRR) or multivariate bayesian shrinkage priors (MBSP) particularly useful, given their ability to deal with critical issues such as collinearity, high-dimensionality, missing data or quantification of uncertainty.

* Corresponding author at: C/ del Dr. Aiguader, 88, 08003 Barcelona, Spain.
E-mail address: xavier.basagana@isglobal.org (X. Basagaña).

https://doi.org/10.1016/j.envint.2023.108344
Received 28 July 2023; Received in revised form 16 October 2023; Accepted 20 November 2023
Available online 22 November 2023
0160-4120/© 2023 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

1. Introduction

The concept of the "exposome" encompasses all environmental exposures to which an individual is exposed from conception onward (Santos et al. 2020; Wild 2005). An increased number of studies have evidenced the driving role of the early-life exposome in the development of most common non-communicable diseases, pointing to prenatal and early childhood as windows of opportunity for disease prevention (Agier et al. 2019; Maitre et al. 2021; Vrijheid et al. 2020). Most of the exposome research has traditionally focused on the study of different health outcomes separately, under what is known as the single-outcome approach (Barrera-Gómez et al. 2017; Vrijheid et al. 2020). Although single-outcome analysis is a valid approach for assessing exposome-health associations, in some cases, it may not provide a comprehensive enough vision, especially when exposures are presumed to have multiple effects on health. In the last years, advanced methods for outcome-wide analysis have been developed, which can provide a more holistic perspective. In outcome-wide analysis, the effect of an environmental exposure on multiple health outcomes is simultaneously investigated (Vanderweele 2017; VanderWeele et al. 2020). Although traditionally applied for the evaluation of single exposures in environmental epidemiology, outcome-wide analysis has barely been extended to exposome research, where multiple exposures are investigated jointly.

Outcome-wide studies can offer several benefits, particularly in exposome studies where exposures can influence many health outcomes that are usually manifested in populations under a pattern of comorbidity (e.g., cardiometabolic health, respiratory problems, and neurodevelopment and cognitive dysfunction). From the methodological point of view, outcome-wide analysis allows consideration of the correlations among the multiple exposures and outcomes. This can be used to borrow information from other exposure-outcome pairs when estimating the effect of a particular exposure on a given outcome, bringing in an increased ability to detect signals that may not be strong enough to be detected in a traditional hypothesis-driven analysis of a single outcome (Kundu et al. 2021). This is especially true in the scenario of responses that share similar predictors or exposures exerting similar influences across different outcomes, or in the case of complex disorders arising from multiple correlated phenotypes. Moreover, by allowing some flexibility in the functions estimated for each health outcome, some of these models can also lead to the discovery of ambiguous risk factors previously unnoticed (e.g., exposures that may be beneficial for some outcomes but harmful for others) (VanderWeele 2017; VanderWeele et al. 2020). An outcome-wide exposome analysis also helps avoiding selective reporting of results and may better control multiple testing. From the perspective of public health recommendations, the outcome-wide approach might also be of utility, since the identification of key exposure agents with multiple and simultaneous effects on health can lead to the development of preventive measures with higher impacts on population health.

Outcome-wide research also poses some challenges, such as the increased dimensionality requiring larger sample sizes, the increased risk of false positives, the need for complex confounder-adjusting strategies, or the presence of incomplete data records. Fortunately, advanced multivariate statistical methods have become available in recent years offering solutions to some of these problems (Bai and Ghosh 2018; Cao et al. 2022; Chen and Huang 2012; Chun and Keleş 2010; Kim et al. 2009; Kundu et al. 2021; Luo et al. 2018; Obozinski et al. 2011; Peng et al. 2010; Turlach et al. 2005; Wang et al. 2015a). Most of these methods stem from the field of omics, in which the identification of "master regulators" (i.e., genes affecting the expression or molecular profiles of many targets at the same time) is of great interest. Despite the parallelisms existing between omics and exposome fields (e.g., high-dimensional settings with low sample sizes, complex noise and correlation structures present in data, etc.), these techniques have been rarely applied for the study of environmental exposure-health associations. On the contrary, most of the outcome-wide approaches conducted to date in the context of environmental hazards have opted for the use of multivariate regressions or seemingly unrelated regressions, which at best only modestly improve efficiency compared to that achieved when fitting a separate linear regression model for each outcome (Descarpentrie et al. 2023; Kim et al. 2022; Okuzono et al. 2022; Steptoe and Fancourt 2020). One of the plausible reasons for the low spread of the new available multivariate techniques for outcome-wide analysis in exposome research could be the lack of guidelines and recommendations on how to apply them to the exposome context, or the requirements in some cases of adaptation for dealing with certain exposome particularities.

In this work, we provide an overview of a selection of methods that are well-suited for outcome-wide exposome analyses and are implemented in the R statistical software. Specifically, we compare each of the techniques, presenting the main rationale behind them, their strengths, and limitations, and provide guidelines for their application to exposome data. In addition, we apply all selected methods to a real exposome dataset from the HELIX project, in which multiple health outcomes and a rich exposome characterization are available for 6 longitudinal European birth cohorts (Maitre et al. 2018). By synthesizing the most promising methods in this field, illustrating their application to a real dataset, and providing analysis codes, this work aims to provide a valuable resource for researchers in the field of exposome analysis who seek to identify and investigate in a more comprehensive way the complex relationships between environmental exposures and multiple health outcomes.

The layout of this manuscript is as follows. In Section 2, we describe the main analytical challenges behind outcome-wide analysis with exposome data and present each of selected methods, comparing their suitability, strengths and limitations. In Section 3, we introduce additional steps that should be considered after the main outcome-wide analysis to increase the robustness of findings. Section 4 includes details on the codes and software resources generated in this paper. In Section 5, we present the application of all selected methods to a real exposome dataset and detail the analysis findings. Finally, we conclude with a general discussion in Section 6.

2. Methods for outcome-wide analysis in exposome research

In exposome research, there are many examples of situations in which an analysis of multiple outcomes simultaneously can be useful. A clear example is the identification of key risk exposures with small but relevant effects on different health domains, such as the case of passive tobacco exposure during childhood and pregnancy, which has been revealed as an adverse factor for cardiometabolic or respiratory health, as well as for behavioral and neurodevelopmental problems in children (Agier et al. 2019; Maitre et al. 2021; Vrijheid et al. 2020). Alternatively, one could be interested in the evaluation of pollutants and chemicals and their effect on a group of health parameters which are strongly correlated to each other; for example, the group of parameters framed under the umbrella of the metabolic syndrome definition (e.g., blood pressure, lipids levels, glucose, and body mass index). For these and other types of outcome-wide approaches in exposome research, a number of different advanced techniques have recently become available (Bai and Ghosh 2018; Cao et al. 2022; Chen and Huang 2012; Chun and Keleş 2010; Kim et al. 2009; Kundu et al. 2021; Luo et al. 2018; Obozinski et al. 2011; Peng et al. 2010; Turlach et al. 2005; Wang et al. 2015a). In this work, we selected a representative group of the most promising and useful outcome-wide analysis techniques, considering their suitability for exposome data and their ability to handle the common hurdles encountered when working with multiple health domains in the context of the exposome. These challenges involve:

1) The curse of dimensionality. The high dimensionality of the exposome is a well-known complication in the case of low-to-moderate

sample sizes; it is exacerbated when multiple outcomes are considered as it requires the exploration of how exposome domains could be affecting each of the different studied phenotypes. This requires study populations with a large enough sample size (i.e., "one in ten rule" for each outcome) to robustly test multiple research hypotheses as well as to incorporate effective strategies for feature selection during the model training. Another issue related to the high dimensionality problem in outcome-wide analysis is the increased probability of making one or more false discoveries when testing multiple hypotheses (family-wise error rate), which should also be addressed through the incorporation of adequate inference techniques.

2) The need for adjustment for confounding effects is a constant in all epidemiological studies in order to avoid bias and reduce both false negative and positive results. The identification of the different sources of bias, and their adequate incorporation into models is not a straightforward task in exposome-wide association studies (ExWAS). Usually, a simplification approach is adopted by selecting the same group of confounders for all exposures. In outcome-wide analysis, again, the situation gets more complicated, with different ways to approach it (Vanderweele 2017). For this reason, it is highly advisable that outcome-wide analysis techniques incorporate specific mechanisms or strategies to control for the effect of confounders.
3) Heterogeneity in the type of outcomes and correlation structures among outcomes. It is common to find phenotypes of interest with different variable domains (continuous, binary, counts), and the functions for modeling their relation with the exposome are usually different. For this reason, it is desirable to count on outcome-wide methods especially designed for the joint modeling of different types of outcomes. Additionally, within the context of health alterations influenced by the exposome, it is typical to find correlation structures among outcomes, such as those mentioned in the example of metabolic syndrome. It is plausible, therefore, to hypothesize that an exposure affecting two correlated outcomes should present a similar association with both of them in terms of effect size. The consideration of these situations by outcome-wide methods during the modelling would result in an increased estimation efficiency.
4) The presence of missing values in outcomes. Missing data represent an important source of uncertainty and loss of accuracy in epidemiological studies. When they are present in exposures, multiple imputation techniques are usually adopted in order to preserve the sample size of studies and to correct potential biases. Nevertheless, some authors point out that there are no clear benefits in imputing missing values when these are present in the outcome variables (von Hippel 2007). As an alternative, researchers tend to exclude study subjects with any missing observations in the phenotype, under the so-called complete case analysis, but this is at best of low statistical power and at worst provides biased estimates. In an outcome-wide analysis, the situation escalates given the increased number of assessed phenotypes, which could drastically reduce the sample size if one opts for the complete cases analysis. Besides, there could be a different number of participants available for each outcome. In this context, since there are no clear guidelines on how to proceed, it is advisable that selected outcome-wide methods are designed to deal with missing data in the outcome variables.
5) The need for a readily-available implementation in open-source software. Given the novelty of the outcome-wide approach, many of the methods encountered in the literature are not accompanied by the release of an analysis package for their implementation. For this reason, we restricted the search to those methods that were properly implemented in the open-source R software, commonly employed by the epidemiology community, so their use can be widely spread in the exposome research.

Our search identified six different methods that can be grouped into four main categories; 1) Regularized multivariate linear regression techniques (Graph-Guided Fused Lasso (GFLasso) and GroupRemMap) (Kim et al. 2009; Wang et al. 2015a); 2) Multi-task learning (Multi-task L2,1-norm regularized regression model (MTL_L21)) (Cao et al. 2022); 3) Dimensionality reduction approaches (Sparse Reduced-Rank Regression (sRRR) and Mixed-response reduced-rank generalized linear regression model (mRRR)) (Chen and Huang 2012; Luo et al. 2018); and 4) Bayesian extensions of the multivariate regression framework (Multivariate Bayesian Model with Shrinkage Priors (MBSP)) (Bai and Ghosh 2018). In all of them, exposome-health associations for each outcome and exposure are quantified in the form of regression coefficients (beta estimates, β). Output beta estimates are presented as a $P \times Q$ matrix where P refers to the number of assessed exposures and Q to the number of outcomes. In this matrix, each row represents the coefficient estimates for one predictor on all outcomes, and each column represents the coefficient estimates for all predictors on one outcome.

2.1. Preliminary concepts: Regularization in a nutshell

Before starting with the description of each method, we briefly introduce the general concept of regularization, which appears as a common theme across selected techniques. The term regularization refers to a set of mathematical strategies especially designed to select the relevant predictors for an outcome variable while mitigating the risk of overfitting in a high-dimensionality-low sample size setting, ultimately improving the predictive performance and interpretability of the model. This is achieved by imposing constraints during the estimation of regression coefficients (e.g., driving the coefficients of uninformative predictors to zero). For that, regularization techniques add penalties to the objective function during the data modelling. The objective function, in essence, is a mathematical expression that characterizes the probability of the observed data as a function of some parameters (e.g. regression coefficients linking environmental exposures and the health outcomes). The estimation process consists of finding values of the regression coefficients that optimize the objective function. By adding regularization penalties, one forces estimated coefficients to be less faithful to the training data and more generalizable in other situations; pursuing a balance between fitting the data closely and keeping the model's complexity in check.

Two common regularization strategies involve using L1 and L2 penalties. These penalties influence the shape of the objective function differently. In L1 regularization, the objective function has an extra term (known as (λ)) penalizing models in which the sum of the absolute values of the regression coefficients is large. As a result, estimated coefficients are shrunk and some are driven to zero. This property is particularly valuable in variable selection, allowing researchers to identify the most important exposures for the research question. On the other hand, In L2 regularization, there is an extra term penalizing models in which the squared sum of regression coefficients is large. This penalizes large estimated coefficients and results in distributing the weights more evenly across variables. This can be advantageous when dealing with correlated exposures, helping to prevent multicollinearity issues and stabilizing coefficient estimates.

In the context of the selected methods, different regularization strategies are adopted pursuing the generation of simpler models, and the reduction of overfitting in outcome-wide analysis.

2.2. Regularized multivariate regression framework

The techniques GFLasso and GroupRemMap are frequentist multivariate techniques based on the regularized linear regression framework (Kim et al. 2009; Wang et al. 2015a). In the single-outcome scenario, the most popular regularization techniques are the Lasso, Ridge regression, or Elastic-Net, which use L1 and L2 penalties (or their combination) to prevent overfitting in high-dimensional data. As previously mentioned, these techniques employ penalties to shrink to/near zero the coefficient estimates of non-relevant predictors, which contribute little to the

minimization of the error (Tibshirani 1996). The natural extension of these regularized techniques into the multivariate framework is equivalent to fitting *Q* single-outcome regularized models separately. In other words, for a collection of outcomes, each outcome would be treated as independent of all the others, and every outcome would be regressed on a common set of exposures via its own Lasso, ignoring the possible relations existing among models. For the multi-outcome scenario, the proposed methods take the basic idea of the multivariate Lasso and go a step further. Besides the L1 penalty, they also incorporate other types of penalties into the objective function (e.g., bridge penalty, fusion penalty, etc.), so the coefficients of similar exposures/outcomes are forced to adopt similar values. Through these strategies, GFLasso, and GroupRemMap, borrow information from different coefficients and outperform the basic multivariate linear regression and its regularized extensions. Below, a brief description and the main concept behind each technique are presented.

- GFLasso: The innovation in GFLasso is that, in addition to using L1 penalty, which induces overall sparsity, it employs another penalty called 'fusion penalty' (γ) that fuses regression coefficients across correlated phenotypes. This penalty is added to the objective function penalizing the difference among the estimated coefficients for the effect of the *i*th exposure on two correlated outcomes. A larger value for the γ□penalty will lead to a greater fusion effect, and therefore, a greater similarity among the coefficients of the effect of the *i*th exposure on correlated outcomes. GFLasso, therefore, lays on the drawing of the graph dependency structure underlying the outcome variables in the dataset, which can be quantified by an outcome correlation matrix. Starting from this correlation matrix, two types of fusion penalties are proposed in GFLasso, depending on whether they use an unweighted or weighted connectivity phenotype graph as a guide. In the case of the unweighted approach, the correlation matrix for outcomes is first transformed into a matrix of 0 s and 1 s according to a correlation coefficient threshold defined by the user, so the γ penalty is equally applied only to the pairs of correlated outcomes. Instead, the weighted approach does not need any threshold defined by the user and will apply the γ penalty to all pairs of outcomes. In this case, a higher correlation coefficient for a pair of outcomes will increase the fusing effect among these two phenotypes. More details on the mathematical formulations for the solution can be found in the original publication (Kim et al. 2009). According to simulation studies, the weighted approach is expected to generally outperform the unweighted approach, especially in cases with low correlations within outcomes. When using this method, the user will need to define the penalty values λ and γ, which are usually selected by cross-validation. As a result, GFLasso favors the selection of predictors with effects on multiple outcomes, flattening the coefficients of each exposure on correlated outcomes. An open-source package called "gflasso" is available in R through GitHub[1] for the implementation of this approach. We implemented a patch modification to the main functions of the package in order to allow the method to correct for the effect of confounders. The corrected codes are available at the supplementary GitHub repository.[2] Although the method was developed to work with continuous outcomes only, it could be extended to the case of binary phenotypes and logistic regression in the future.
- The GroupRemMap Method: The main idea behind GroupRemMap is quite similar to the solution proposed in GFLasso. However, in this case, the types of penalties employed are slightly different (Wang et al. 2015a). Here, the main goal of the model is to promote the selection of exposures affecting the majority of the outcomes while at the same time, considering the relations among predictors. I.e., the method favors the selection of groups of related exposures rather than single variables. This is achieved through the use of three different penalties: an L1 penalty that controls the overall sparsity of the output coefficient matrix; a γ-type penalty term that encourages a group selection effect ('bridge' penalty); and an L2 penalty, which combined with the bridge penalty, induces row sparsity in the matrix of coefficients, such that groups of exposures that have effects on many outcomes are more likely to enter the final model. The combination of the L1 penalty and the bridge penalty on L2 norm of grouped predictors is presented as the GroupRemMap penalty. As a result, GroupRemmap favors selection of predictors with effect in majority of outcomes, flattening the coefficients of exposures belonging to groups. For that, GroupRemMap relies on the existence of expert knowledge for grouping exposures (e.g., according to exposome domains, families, biological pathways involved in their health effects, etc.). The selection of tuning parameters is again done through K-fold cross-validation. An R package called "groupRemMap" is available through CRAN. As it happens with GFLasso, GroupRemMap is designed to work only on continuous outcomes.

2.3. Multi-task learning

Next, we describe the more sophisticated group of techniques called multi-task learning (MTL) (Cao et al. 2018, 2022). MTL with joint predictor selection is based on the concept of cross-task regularization; a type of regularization that penalizes the complexity of the whole coefficient matrix rather than individual estimates, aiming to identify a row-wise sparse structure for it that maximizes prediction accuracy (for all tasks/outcomes). That is to say, each selected exposure needs to affect all outcomes under study (which is an assumption not always held. This model is known as L2,1-norm regularized regression model (MTL_L21) and represents a complex optimization model (Liu et al. 2012). In MTL_L21, the value of an L1 penalty needs to be specified, and according to such value the variable selection (overall sparsity of the matrix of coefficients) will be more or less strict. The optimal value for the penalty can be determined by cross-validation according to the error of derived models. Interestingly, superior prediction performance and biological plausibility have been demonstrated for MTL when compared to single-task regularized approaches (e.g., standard Lasso or elastic-net recursively applied to each outcome) (Cao et al. 2018). Additionally, MTL_L21 includes an L2—like penalty for group selection (which is optional). This technique was derived in the omics field but has never been applied to exposome research. By applying these models to the outcome-wide analysis in the exposome context, one may identify exposures simultaneously associated with multiple illness phenotypes. This and other MTL variants have been implemented in the R library "RMTL", where authors propose to solve the optimization problem by means of a solver based on the accelerated gradient descent method (Cao et al. 2019). By changing their loss function, these methods are able to work on continuous or binary outcomes (logistic loss for classification or least square loss for linear regression). Nevertheless, they cannot simultaneously model a mix of continuous and binary responses.

2.4. Dimensionality reduction techniques

Other suitable models identified are the well-known dimensionality reduction techniques, that combine the predictor variables into fewer features that can be explained as latent factors that drive the variation in the multiple response variables (Chen and Huang 2012; Luo et al. 2018). These techniques, in comparison to regularized multivariate linear regressions, increase computational efficiency and predictive accuracy, especially in the case of high-dimensional settings. Among the different techniques, we highlight the group of reduced-rank regression methods (RRR). These methods try to identify latent factors that maximize the quantity $Corr^2(X,Y)*Var(Y)$, where X and Y are centered exposures (number of individuals $\times$ *P*) and outcomes (number of individuals $\times$ *Q*)

[1] https://github.com/krisrs1128/gflasso.
[2] https://github.com/AugustoAnguita/exposome_outcomewide.

matrices respectively. Therefore, they take advantage of interrelations among the outcome variables to improve predictive accuracy. RRR makes a restriction on the rank of the output regression coefficient matrix, which reduces the number of parameters to be estimated and improves the efficiency of estimation. Thanks to that, RRR is also efficient dealing with multicollinearity among predictors. Often, these techniques also incorporate the use of penalties, such as L1 and L2, during parameter estimation so that feature selection is achieved. Within the group of RRR, we identified two techniques that could be useful for outcome-wide analysis in the exposome context; the sparse RRR (sRRR) and the mixed-outcomes RRR (mRRR). sRRR is a variant of the reduced-rank regression (RRR) method that incorporates sparsity constraints to achieve feature selection. In addition to the standard RRR method that imposes a rank constraint on the regression coefficient matrix, the sRRR method also incorporates a penalty term in the objective function to encourage the sparsity of the coefficient matrix. This penalty term is a group-lasso type penalty, which shrinks some rows of the coefficient's matrix towards zero. The use of sparsity constraints in sRRR allows for the identification of a small set of important exposure variables that are most strongly associated with the outcome variables while ignoring irrelevant or weakly associated exposures. On the other hand, mRRR integrates multivariate outcomes of mixed types belonging to an exponential dispersion family. The types of outcomes, for example, cover continuous, binary, and count outcomes, which are commonly seen in exposome research. Although mRRR does not incorporate feature selection as sRRR, it is able to deal with missing data records in the outcomes. Both techniques are implemented in the R package "rrpack".

2.5. Bayesian approaches

Bayesian models allow the incorporation of prior knowledge about the parameters being estimated, which is particularly useful in low-sample size settings such as the ones typically faced in exposome research. Bayesian approaches also allow a better quantification of uncertainty about parameters through the posterior probability distribution. Finally, they tend to be more flexible than frequentist techniques since they can handle a wide range of models, including complex hierarchical models with multiple levels of uncertainty. All these facts make them well-suited for the outcome-wide analysis in exposome research. Recently, a method known as sparse multivariate Bayesian estimation with shrinkage priors (MBSP) has become available (Bai and Ghosh 2018). This method is a multivariate Bayesian technique that incorporates a Bayesian version of lasso penalties (known as global–local shrinkage (GL) priors) for the generation of a sparse model of predictors affecting multiple outcomes. Particularly, it uses GL priors belonging to the "three parameter beta normal (TPBN) family", which includes the horseshoe, the Strawderman-Berger, and the normal-exponential-gamma (NEG) priors. Interestingly, this method may be used for sparse multivariate estimation for P, and Q of any size, showing good performance even in the scenario of ultra-high-dimensional settings where P is allowed to grow nearly exponentially with the number of individuals. By examining the 95 % posterior credible intervals for every element in each row of the posterior conditional distribution of estimated coefficients, the MBSP model can also be used for variable selection. The level of sparsity of the final model in MBSP can be controlled locally through the choice of prior and at the global level through an hyperparameter, here called tau (τ). Finally, MBSP is able to account for the effect of confounders, which can be forced into the model. MBSP is implemented in a comprehensive R package called MBSP. For applying the model, the user needs to set which exact type of GL shrinkage priors to use, and the additional parameter τ (which controls the amount of global shrinkage). The selection of which GL shrinkage priors to use will shape the prior beliefs about the importance and distribution of the different coefficients for exposure-outcome associations, thereby influencing the variable selection and the estimation procedure. Nevertheless, the choice among the three priors is not a trivial question, especially if the researcher has no strong preconceived beliefs on the underlying associations. In general, according to previous simulation studies (Kundu et al. 2021), the horseshoe prior has a reasonably good performance across a variety of empirical experiments and situations and is more computationally efficient than other approaches. The NEG prior could be of preference if the researcher is specifically interested in capturing both common effects shared across multiple outcomes and individual effects specific to each outcome. The MBSP R package incorporates a strategy to help researchers identifying the most appropriate prior for each situation. For that, the user can try different priors and finally determine the optimal model by the inspection of goodness of fit criteria such as the Deviance Information Criterion (DIC) or the Watanabe–Akaike information criterion (WAIC). For that task, WAIC may be preferred since it averages over the posterior probabilities and it does not require the them to be approximately normal.

2.6. Overview of the properties of the different methods

A general overview of all selected methods and their solutions for outcome-wide analysis in exposome research can be found in Table 1. These techniques differ a bit conceptually, but they also differ in the situations they are able to accommodate. Most of the techniques are specially designed to deal with high-dimensional datasets. Among all methods, MBSP was specially designed to work in high or ultra-high dimensional settings, being able to retrieve a small number of exposures affecting the multiple outcomes even when input data contain hundreds of thousands of variables, favoring interpretability and summarization of the problem. Although most of the techniques are designed to work only on continuous outcomes, an option for mixed-type outcomes has been proposed (e.g., mRRR). Regarding missing data in the outcomes, again only the mRRR is able to deal with incomplete records, which suggests this functionality is a key aspect to be considered by future methodological developments. Since the methods GFLasso, MTL_L21, sRRR and MBSP did not incorporate a solution for accounting for the effect of confounders, whenever possible, we implemented this functionality. The new versions of these functions are available either in their latest CRAN releases (for sRRR and MBSP) or in our GitHub repository[3](for GFLasso). The only method in which it was not possible to add this implementation was the MTL approach. In this case, therefore, it is advisable to conduct alternative ways of confounding treatment (such as residualization of predictors/outcomes on confounders (partialling-out approach)) (Demissie and Cupples 2011).

One of the key disadvantages of the proposed methods, and of most machine learning methods in general, is their inability to quantify uncertainty, focusing on accurately producing a point estimate (e.g., via solving an optimization problem) but neglecting the reproducibility/replicability of the results. This, given the small sample sizes that are usually faced with in exposome research, can result in a high rate of false positives and negatives. Among the proposed techniques, only MBSP addresses the lack of uncertainty quantification in high-dimensional inferences, attempting to approximate the full posterior distribution by quantifying uncertainty instead of simply producing a point estimate.

Although the level of the sparsity of the output models will depend on the chosen size of penalties, the ability to obtain a parsimonious model with optimal hyperparameters for each technique is variable. The methods deriving a less sparse model are GFlasso and mRRR; the ones with medium level are MTL and sRRR; while the rest are able to retrieve a highly sparse model (GroupRemMap, MBSP), with no need of post-curation of findings.

Differences can also be found among methods in the type of relationships they look at. While GFLasso, GroupRemMap and MBSP allow intra-exposure sparsity (an exposure not necessarily need to affect all

[3] https://github.com/AugustoAnguita/exposome_outcomewide.

Table 1
Comparison of selected methods for outcome-wide analysis in exposome research.

	GFLasso	GroupRemMap	MTL_L21	mRRR	sRRR	MBSP
Group	Regularized multivariate regression framework	Regularized multivariate regression framework	Multitask Learning based on regularized regression	Dimensionality reduction techniques	Dimensionality reduction techniques	Sparse multivariate Bayesian estimation with shrinkage priors
Goal	Uses penalties to promote the selection of **exposures affecting multiple outcomes**.	Uses penalties to promote the selection of **exposures affecting the majority of the outcomes**.	Uses cross-task regularization to promote the selection of **exposures affecting all outcomes**.	Promotes the selection of **exposures affecting all the outcomes**.	Promotes the selection of **exposures affecting all the outcomes**.	Promotes the selection of only **few exposures affecting the majority of the outcomes**.
Strategy	It considers the **correlation structure existing among outcomes** encouraging similar (or dissimilar) responses to be explained by a similar (or dissimilar) predictors.	It considers the **correlation structure existing among predictors** favoring the selection of groups of related exposures affecting multiple outcomes rather than single variables.	It also allows considering the **correlation structure existing among predictors** favoring the selection of groups of correlated exposures affecting all outcomes.	Assume that all the **outcomes and exposures are associated through a shared low dimensional subspace**. It allows considering the **correlation structure existing among outcomes**.	Assume that all the **outcomes and exposures are associated through a shared low dimensional subspace**. It allows considering the **correlation structure existing among outcomes**.	It considers the **correlation structure existing among outcomes** encouraging similar (or dissimilar) responses to exhibit similar (or dissimilar) coefficients for the same predictor.
Type of outcomes	Only continuous	Only continuous	All continuous or All binary	Mixed outcomes	Only continuous	Only continuous
Missing data in Outcomes	No	No	No	Yes	No	No
Variable Selection	Yes	Yes	Yes	No	Yes	Yes
Allow adjusting for the effect of confounders	Yes (functionality added)	Yes	No (partialling-out as an alternative)	Yes	Yes (functionality added)	Yes (functionality added)
Quantification of uncertainty	No	No	No	No	No	Yes (95 % credible intervals)
Level of sparsity	Low	High	Medium	Low	Medium	High
Reference	Kim et al. (2009) Bioinformatics	Wang *et al.* (2015) Stat Biosciences	Han *et al.* (2018) Bioinformatics	C. Luo et al. (2018) Journal of Multivariate Analysis	Chen *et al.* (2012) Journal of the American Statistical Association	R Bai *et al.* (2018) Journal of Multivariate Analysis

outcomes under study), the rest of the methods rely on the assumption all outcomes need to share the same set of associated exposures.

Finally, it is important to state that, previous to data analysis, for all these methods it is critical to scale all predictor and outcomes to the same scale. This way, we avoid the bias of methods toward the selection variables with highest variance and wide domains, further allowing the inter-outcome comparison of coefficients.

3. Approaches for the curation of outcome-wide findings: post-selection inference

Some of the proposed methods are far from retrieving a parsimonious final model as shown in Table 1. In some cases, it is because they do not include a proper feature selection step (mRRR), while in others, despite including it, the optimal penalty does not always result in the desired level of sparsity (GFLasso, GroupRemMap, and sRRR). In these cases, therefore, we need a post-selection inference strategy to further filter out the output variables, resulting in a reduced list of associated exposures. This strategy is known as valid inference after data exploration (VIDE) (Kuchibhotla et al. 2022). For this, although no specific strategy has been defined in the literature for outcome-wide analysis, we propose several alternatives. It is important to note that for these approaches, it is mandatory to have standardized beta coefficients (where the scales of the exposures and outcomes have been standardized previous to data modelling):

1) Selection of exposures based on the magnitude of their individual estimated effects: In this procedure, we propose a basic filtering strategy in which we keep for interpretation only those exposures presenting a β estimate above a certain threshold for a certain percentage of the outcomes. The choice of the threshold for estimated effect, and the percentage of affected outcomes will rely on the characteristics of the problem and the preferences of the researcher. A reasonable but arbitrary option could be to select only those exposures presenting a β higher than the 50th percentile of estimates in at least 60 % of the outcomes under study.
2) Selection by row-wise norm: This approach tends to filter out exposures according to their estimated overall effect on all outcomes under the study, thereby promoting the selection of "master-regulators". This approach consists of calculating the Euclidean norm for each p^{th} row of the coefficient matrix. The Euclidean norm (also called the vector magnitude, Euclidean length, or 2-norm) of a vector v with Q elements is defined by $\sqrt{\sum_{k=1}^{Q}|\beta_k|^2}$. A higher value will indicate a higher overall effect of the p^{th} exposure on the outcomes under study. Once this value is calculated for all exposures, one could select a number of exposures presenting the top values for the norm. The choice of the final set of selected exposures will depend on the initial dimensionality of the dataset and the desired level of interpretability.
3) Resampling strategies: Bootstrapping and other resampling techniques have been proposed in the literature as useful VIDE techniques. In this case, we adapt the idea of *Bolasso* (bootstrap-enhanced least absolute shrinkage operator) (Francis R. Bach 2008), in which Lasso is run for several bootstrapped replications of a given sample, and then the results of the Lasso bootstrap estimates are intersected to provide consistent model selection. In the context of our problem, we recommend running this bootstrapping procedure at a minimum of 1000 times. Then, a *P*-value-like inference could be estimated with

"*H_0: exposure is not selected*". For a specific exposure, if it is selected N times out of 1000 runs, the *P*-value is calculated as (1000-N)/1000. The derived *P*-values can be used to reduce the number of selected features. Of note, for this approach, it is necessary to identify the optimal hyperparameters in each of the runs. Likewise, these approaches can only be applied if the output models already incorporate variable selection, i.e. they provide a list of selected or non-selected variables.

The selection of one or another approach should be made carefully by the researcher depending on the specific characteristics of the problem and the desired interpretability of the final output model.

4. Code availability

As an online companion to this paper, we have created a GitHub repository[4]and Rpubs website,[5] where we provide scripts and a simulated multivariate dataset for running each of these methods. In this repository, we cover all required steps for applying these techniques; including data standardization, model parameters calibration, output curation strategies and visualization.

5. Application of selected methods on a real exposome dataset: the helix study

In this section, we apply all selected methods to a real exposome dataset demonstrating their suitability and giving recommendations for implementation. The research dataset employed here derives from the HELIX project (Human Early-Life Exposome). The HELIX project gathers data from 6 longitudinal European birth cohorts with the aim of evaluating the effect of environmental risk factors on mothers' and children's health. HELIX cohorts include the *BIB* (Born in Bradford) (United Kingdom), *EDEN* (Étude des Déterminants pré et postnatals du développement et de la santé de l'ENfant) (France), *INMA* (INfancia y Medio Ambiente) (Spain), *KANC* (Kaunus Cohort) (Lithuania), *MoBa* (Norwegian Mother and Child Cohort Study) (Norway), and *Rhea* (Mother-Child Cohort in Crete) (Greece). General details of the study design can be found elsewhere (Maitre et al. 2018; Warembourg et al. 2019). Here, we focus on a HELIX subcohort of 881 children according to the following criteria of eligibility: 1) age 6 to 11 years at the moment of outcome evaluation; 2) complete address history; 3) no serious health problems that may affect the clinical testing or the child safety; and 4) having complete data on all health outcomes. In the 881 children, many environmental exposures were evaluated to define childhood exposome (age 6 to 11 years). Collected exposures comprise three main parts of the exposome: outdoor exposures, chemical exposures, and lifestyle and social factors. All variables incorporated in the dataset have been appropriately pre-processed previous to analysis (normalized and scaled, outliers removed, and missing values imputed) as described elsewhere (Maitre et al. 2018; Warembourg et al. 2019). In total, early-life exposome data was composed of 133 variables. A detailed description of all included exposures can be found in Supplementary Table 1. Regarding outcome data, 9 continuous health outcomes were investigated (Supplementary Table 2). Outcomes were assessed at the same time as the exposome and included parameters related to (1) obesity and cardiometabolic health, (2) respiratory health, and (3) cognition and mental health.

In total, our dataset included 133 exposures (Supplementary Table 1) and 9 outcomes (Supplementary Table 2) in 881 individuals. The main research question was to identify exposures simultaneously affecting multiple health outcomes. The way exposures affect each outcome does not need to be necessarily the same, so we identified both overall risk exposures negatively affecting most of the outcomes and also ambiguous factors (being protective for some outcomes but risky for others). For a fair comparison of all presented methods, we selected only continuous outcomes. Within the exposures, we restricted only to continuous or ordinal variables. Previous to the analysis, we inverted the domain of those outcomes in which higher values represent a healthy status so that the interpretation of β estimates always goes in the same direction; a positive/negative β value will always indicate a risk/protective relationship among the *p*-th exposure and the *q-th* outcome. In all the approaches, both outcomes and predictors were centered and scaled in order to obtain comparable estimates.

Additional details on the implementation of each method (input sequence of hyperparameters, details on cross-validation procedures, post-inference selection technique, or computing efficiency) can be found in Supplementary Table 4.

The strategy adopted for confounding control involved considering as potential confounders all variables temporally prior to the measured exposures that might affect the exposures, and at least one of the outcomes. Selected potential confounders included paternal age, maternal age, maternal body mass index, gestational age, trimester of conception, educational level of both parents, parity, maternal marital status, history of asthma for both parents, child height, child sex, child age, sibling position, child age of first nursery attendance, cohort of recruitment and child ethnicity. In GFLasso, GroupRemMap, sRRR, mRRR and MBSP, confounder adjustment was done during the modelling by forcing their inclusion into final models. In MTL_L21 model, this strategy was not possible due to method limitations. Thus, we opted for the alternative of regressing out the effect of confounders on both outcomes and exposures before the data analysis (partialling-out approach).

The whole matrix of estimated coefficients for all exposures and outcomes by each of the methods can be found in Supplementary Table 5. Output models were quite different in terms of sparsity degree (Fig. 1 **and** Supplementary Fig. 1-6). The sparsest model was generated by the Bayesian approach MBSP (with only 4 selected exposures affecting studied outcomes: copper, lead, DDE and HCB). On the contrary, the least sparse model was obtained by the GFLasso method (45 selected exposures). In general, all methods were quite consistent in the identification of exposures with effects on the multiple outcomes. Especially, the most overlapping findings were evidenced for the pairs (sRRR-MTL_L21, and mRRR-sRRR). The methods showing more consistent results (selected exposures more repeated in the rest of the methods) were the MBSP, GroupRemMap, and sRRR, while the ones deriving the most singular results were the mRRR and GFLasso. This may be because these models were the ones selecting higher number of features. The specific results obtained in each approach can be explored in Supplementary Fig. 1-6. Besides selected methods, a single-outcome approach (Exposome-Wide Association Study) was also used as a comparison against the multi-outcome approaches. Results can be found at **Supplementary figure 7**. After focusing exclusively on the overlapping findings of outcome-wide approaches, we can derive some interesting conclusions. On the one hand, the methods were able to identify "master-regulators" negatively or positively affecting most of the outcomes under study. This was the case for exposure to copper, indoor pollutants like benzene or $PM_{2.5}$, tobacco smoking, and sedentariness, which were evidenced as global risk factors for all outcomes under study (obesity and cardiometabolic health, respiratory health, and cognition and mental health). On the opposite, the size of green spaces near children's schools, or exposure to certain chemicals such as PBDEs and PCBs were shown to be protective factors, especially for cognition and behavioral outcomes. As expected, ambiguous risk factors were also identified, highlighted in orange in Fig. 1. Examples are sleep duration, which was reported as a protective factor for most of the outcomes but negatively associated with respiratory health, and the Family Affluence score, which was a negative factor for obesity and cardiometabolic traits, but protective for the rest. Other interesting findings involved the global protective role of family richness or the practice of moderate-to-vigorous

[4] https://github.com/AugustoAnguita/exposome_outcomewide.
[5] https://rpubs.com/aanguita/outcome_wide_analysis.

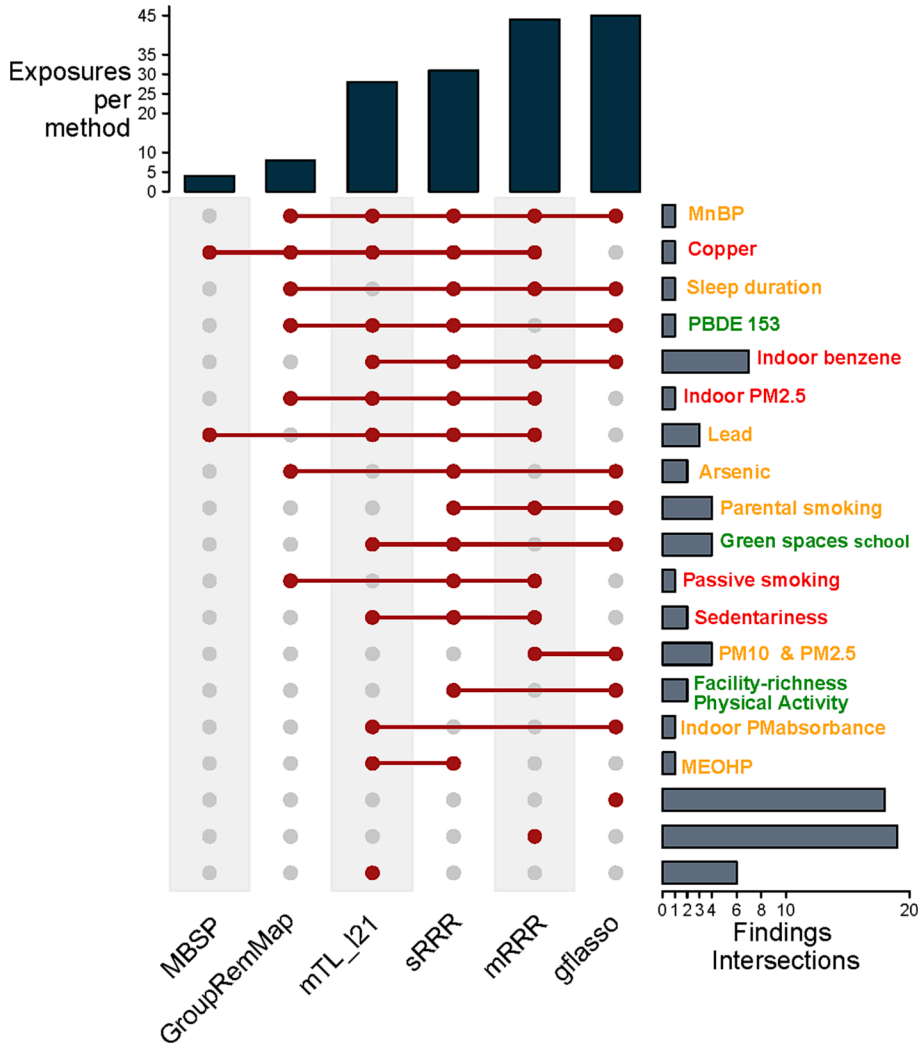

Fig. 1. Selected exposures in each of the methods. The upset plot provides an efficient way to visualize intersections of more than 5 sets compared to the traditional approaches (i.e. the Venn Diagram). Here, the top part of the upset plot shows the number of selected exposures by each of the methods (according to criteria defined in Supplementary table 4). The horizontal red lines and the histogram on the right part (Y-axis) of the figure refer to the overlapping findings between models (e.g., the first horizontal line connecting GroupRemMap, MTL_L21, sRRR, mRRR, and gflasso indicates that all these methods identified the same exposure (in this case, MnBP) affecting multiple outcomes). Thereby, the bar plot (findings intersections) refers to the number of exposures selected for each group or combination of methods. For all groups of exposures identified by at least two models, we indicate the name of one of them on top of each histogram bar. The red/green colour in exposure names indicates that the exposure has been evidenced as a global risk/protective factor for all outcomes in at least one method. The orange colour refers to exposures reported as ambiguous risk factors (i.e. risk factor for some outcomes and protective factor for others). The last rows showing red dots without lines connecting them, represent the exposures identified exclusively by each method, Abbreviations; GFLasso: Graph-Guided Fused Lasso, MBSP: sparse multivariate Bayesian estimation with shrinkage priors, mRRR: mixed-outcomes reduced-rank regression, MTL: Multi-task Learning, sRRR: Sparse reduced-rank regression. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

physical activity for all outcomes under study. Regarding the single-outcome approach, it only revealed associations for three group of exposures (metals and chemicals) and barely affected two groups of outcomes, thereby possibly reflecting an under-selection problem.

6. Discussion

In this work, we present outcome-wide analysis as an interesting approximation for the study of the exposome and its multiple effects on health. Particularly, we take advantage of the recent emergence of advanced statistical multivariate techniques from the field of omics to propose a group of methods that could help in the discovery of environmental exposures with simultaneous effects on multiple chronic non-communicable diseases. For selected methods, we present a general overview highlighting their suitability for dealing with the typical challenges of exposome data, pros and cons, and also give recommendations for their application. Identified methods are grouped into four categories: regularized multivariate regression techniques, multi-task learning approaches, dimensionality reduction approaches, and Bayesian extensions of the multivariate regression framework. According to simulations conducted in previous studies, selected methods have shown better performance than the standard single-outcome approach, in which one fits separate models on each response, ignoring the possible interrelations among response variables (Bai and Ghosh 2018; Cao et al. 2022; Chen and Huang 2012; Kim et al. 2009; Luo et al. 2018; Wang et al. 2015a). Thanks to that, they offer increased power to detect weak signals which could be not strong enough to be detected by the standard approach. On the other hand, single-outcome analyses allow a more tailored analysis, particular sensitivity analyses, a better discussion of results, and a more focused interpretation.

In Table 1, we present a general overview of each selected method. Initially developed for dealing with omics data, most of these techniques

can accommodate high dimensional datasets without suffering a loss of estimation efficiency. Likewise, some of them also allow correcting for the effect of confounders, which is a crucial step in environmental epidemiology. According to the specificities of each problem, and depending on the goal of each research, one could navigate through Table 1 to choose the most suitable approach. For example, we may not always expect that evaluated exposures affect all outcomes under study. In this situation, MTL_L21, mRRR and sRRR (which do not allow intra-row sparsity) might not be preferred options.

Despite all the benefits highlighted in the manuscript for these approaches, some drawbacks have also been identified. For example, the need for priori knowledge about the grouping structure of predictors and the long computing times for the identification of hyperparameters in GroupRemMap pose a burden for applying the method. The lack of proper confounding adjustment strategies in MTL_L21 may be a concern as well, especially in the context of multi-cohort exposome data, where cohort-specific effects are usually expected. In that case, alternatives such as partialling-out the effect of confounders on both predictors and outcomes should be considered. In some cases, we also found that the level of sparsity of output models is sometimes not the one we could desire for an adequate interpretation of findings. In that case, we have proposed three different strategies that could help researchers further curate selected exposures (e.g., restricting to those having a higher impact on most of the studied outcomes). On the other hand, we find the important issue of dealing with mixed-type outcomes. In exposome research, we often are interested in assessing the effect of exposures on both continuous measurements (scores, biochemical measurements) and clinical diagnoses (presence or not of a disease). This functionality is at the moment only possible with the method mRRR. In the rest of the methods, the only alternative is to restrict the analysis to a single type of outcome and further incorporate proper standardization strategies so exactly the same importance is given to each outcome during the modeling. Regarding their ability to deal with false positives, even though penalized approaches provide better control of false discovery rates than unpenalized ones, the rate of false discoveries is not formally controlled. Therefore, future work could consider recent developments connecting penalization and false discovery rate control to be applied to the techniques presented here (Miller and Breheny 2023). Finally, the issue of dealing with missing data records in outcomes is another functionality that is only implemented in mRRR. These two critical aspects, along with the inclusion of a proper strategy for adjusting for the effects of confounders are key points that should be addressed in future developments. We especially encourage the development of more Bayesian-type methods which, given their flexibility, can properly deal with all these aspects. For future work, we also encourage the release of statistical analysis packages accompanying each technical development since this is key point for their spreading among the exposome analysts' community.

Our paper was not exhaustive and other tools are available for multi-outcome exposome analysis that were not selected because of: 1) their similarity with these approaches, or 2) they did not include an implementation in the R software. These include some lasso-type techniques (Guo et al. 2010; Obozinski et al. 2011; Turlach et al. 2005), dimensionality reduction techniques such as sparse partial least squares (sPLS) (Chun and Keleş 2010), sparse canonical correlation analysis (sCCA) (Chen et al., 2017) or parallel independent component analysis (pICA) (Hardoon and Shawe-Taylor, 2011), as well as some other Bayesian approaches (Ando 2011; Kundu et al. 2021).

Despite their potential usefulness, most outcome-wide methods have not been applied to exposome research yet. On the contrary, the few outcome-wide analyses that can be found in the literature with environmental exposure data are restricted to the application of standard multivariate regressions (Amadou et al. 2023; Kim et al. 2022; Okuzono et al. 2023). To date, only one example of using one of the proposed methods with exposome data has been found, in which the relation of 138 exposures and 32 reproductive biomarkers are screened in 796 Chinese men using the GFLasso (Wang et al. 2021). On the contrary, successful applications of some of the proposed methods can be found in the literature in the field of molecular and biomedical data (Curtis et al. 2013; de Abreu e Lima et al. 2018; Li et al. 2017). For example, MTL has been successfully applied in the past for the identification of shared imaging features that simultaneously predict two subtypes of bipolar disorders (Wang et al. 2015b), or for the study of the shared behavioral rhythms that simultaneously predict ten symptoms of Schizophrenia (Tseng et al. 2020).

In order to discuss the applicability of each method to exposome data and give more specific details on their implementation, we decided to apply them all to a real dataset from the HELIX project. From the results, it seems that the GFLasso tends to estimate more homogeneous effects (the same exposure tends to affect most of the outcomes similarly). On the other hand, MBSP, GroupRemMap, and GFlasso are the only ones allowing intra-exposure sparsity (the same exposure does not necessarily affect all outcomes under study). This is important since in the exposome context we do not expect all exposures to systematically affect all assessed outcomes. Another big difference encountered among methods is the level of the sparsity of output models, which in some cases required the application of additional curating strategies. In the current application, MSBP derived the sparsest model with only 4 selected exposures. Normally, the variables that are included in an exposome analysis are pre-selected because there is some plausibility that they may have an effect (usually small) on studied outcomes, so a situation with many causal hits is in principle plausible. Nevertheless, this will always depend on the context, facing sometimes scenarios with just a few causal exposures or others with many of them. Future work should therefore include simulation studies for the identification of the most suitable outcome-wide method for the identification of causal exposures in the different scenarios (e.g., few causal exposures, or numerous causal exposures).

As mentioned above, a key disadvantage of most of the proposed methods, and of machine learning methods in general, is their inability to quantify uncertainty, focusing on accurately producing a point estimate (e.g., via solving an optimization problem) but neglecting the reproducibility/replicability of the results in other cohorts. This, given the small sample sizes we usually are faced with in exposome research, often drives us to deal with false positives and negatives in our results. Among the included techniques, only MBSP addresses the lack of uncertainty quantification in high-dimensional inferences. Therefore, this should be another key issue to be considered in future developments, as others have discussed (Dunson 2018). Fortunately, there is growing literature seeking to address the lack of uncertainty quantification in high-dimensional inferences; for example, focused on penalized optimization methods, such as Lasso (Basu et al. 2021; Miller and Breheny 2023). On this matter, and considering the growing popularity of machine learning approaches, exposome researchers should clearly state that highly predictive black box algorithms, along with an estimate of the important variables, are not enough. Instead, we crucially need tools to tell us how reliable our variable selection decisions are given the sample size, dimensionality, and correlation structure of the data at hand.

In general, there is not a unique method with solutions for all exposome data challenges. Therefore, the selection of the most suitable technique will depend on the characteristics of each specific problem (e.g., some methods will deal better with collinearity among predictors and outcomes than other).

Returning to the HELIX showcase, we found high consistency among the exposures affecting multiple health parameters identified by each of the methods, with some key exposures identified by almost all the approaches (Fig. 1 **and** Supplementary Fig. 1-7). According to most of the methods, exposure to copper, indoor pollutants (like benzene or PM2.5), tobacco smoking, and sedentary habits were found to be global risk factors for obesity and cardiometabolic health, respiratory health, and cognition and mental health. On the other hand, the size of green spaces

close to schools and exposure to chemicals such as PBDEs and PCBs were protective factors, especially for cognition and behavioral outcomes. The findings related to PBDEs and PCBs were unexpected, and highlight a more complex exposure system than expected. Specifically, previous literature have been inconsistent for these associations in the childhood stage, with some findings supporting our results and some others in the opposite direction (Julvez et al. 2021; Maitre et al. 2021). Among the plausible explanations, it highlights the strong lipophilic nature of these components (they are stored mainly in fat tissue and not in blood, where they were measured), and a derived residual confounding due to obesity trajectories or other unmeasured factors. In any case, this association remains inconclusive and cautious interpretation should be made. We also identified ambiguous risk factors, such as sleep duration, which was reported as protective for most outcomes but not for respiratory health. Family Affluence score was found to be harmful for obesity and cardiometabolic traits, but protective for the rest of outcomes. Additionally, interesting findings included the protective role of family richness, blue spaces in cities and the practice of moderate-to-vigorous physical activity for all outcomes under study (Supplementary Fig. 1-6). Individually, these exposures have been previously evidenced by HELIX papers but as factors affecting isolated outcomes (Agier et al. 2019, 2021; Granum et al. 2020; Julvez et al. 2021; Maitre et al. 2021; Nieuwenhuijsen et al. 2019; Vrijheid et al. 2020; Warembourg et al. 2019, 2021), reinforcing the importance of conducting more holistic approaches. In comparison to outcome-wide multivariate approaches, the single-outcome analysis identified only four exposure-health associations out of the nine assessed outcomes **Supplementary figure 7**. These outcomes were those related to cardiometabolic health, which are probably the ones showing stronger effects from environmental exposures. Therefore, this might indicate the higher risk of false negatives and the decreased ability to detect weak associations of the single-outcome approach, which otherwise is overcome in an outcome-wide analysis.

Another approach to deal with multiple outcomes in exposome research is to combine them all into a composite score in the form of a general health score (also known as multimorbidity index). This approach was recently applied to the HELIX project data (Amine et al. 2023). In that study, results highlighted the same exposures that were highlighted here as having multiple effects on health. Interestingly, some of the identified relations were unexpected associations (e.g., exposure to HCB was evidenced as a protective factor for the general health score). In our showcase, as a result of the modeling of each exposome-outcome association separately and not as a composite score, we actually see how many of these factors are not actually protective but ambiguous risk factors. The same was observed for the exposures "having a pet in the family", or "contact with friends and family", which in our use case are evidenced as good for some health parameters but harmful for others.

With the present methodological work, we provide a valuable resource for researchers in the field of exposome analysis seeking to investigate the complex relationships among environmental exposures and comorbidity patterns. In our comparison, we noticed some problems that should be addressed in future developments and encourage the exposome community embrace the use of these multivariate techniques. Future lines of research might involve the systematic comparison of these techniques through simulation studies emulating an exposome-like scenario.

Declaration of competing interest

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

Data availability

Data will be made available on request.

Acknowledgment

ATHLETE project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874583. This publication reflects only the authors' view and the European Commission is not responsible for any use that may be made of the information it contains. We acknowledge support from the grant CEX2018-000806-S funded by MCIN/AEI/ 10.13039/501100011033, and support from the Generalitat de Catalunya through the CERCA Program. We also thank the support from the grant FJC2021-046952-I funded by MCIN/AEI/ 10.13039/501100011033 and, by "European Union NextGenerationEU/PRTR" and acknowledge funding from the Ministry of Research and Universities of the Government of Catalonia (2021-SGR-01563).

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.envint.2023.108344.

References

Agier, L., Basagaña, X., Maitre, L., Granum, B., Bird, P.K., Casas, M., et al., 2019. Early-life exposome and lung function in children in Europe: an analysis of data from the longitudinal, population-based HELIX cohort. Lancet Planet Heal 3, e81–e92. https://doi.org/10.1016/S2542-5196(19)30010-5.

Agier, L., Basagaña, X., Hernandez-Ferrer, C., Maitre, L., Uria, I.T., Urquiza, J., et al., 2021. Association between the pregnancy exposome and fetal growth. Int. J. Epidemiol. 49, 572–586. https://doi.org/10.1093/IJE/DYAA017.

Amadou, C., Heude, B., de Lauzon-Guillain, B., Lioret, S., Descarpentrie, A., Ribet, C., et al., 2023. Early origins of metabolic and overall health in young adults: An outcome-wide analysis in a general cohort population. Diabetes Metab. 49:101414. https://doi.org/10.1016/J.DIABET.2022.101414.

Amine, I, Guillien, A, Anguita-Ruiz, A, Casas, M, Garcia-Aymerich, J, Grazuleviciene, R, et al. 2023. Environmental exposures in early-life and general health in childhood.; doi:10.21203/RS.3.RS-2640215/V1.

Ando, T., 2011. Bayesian variable selection for the seemingly unrelated regression models with a large number of predictors. J. Japan Stat. Soc. 41.

Francis R. Bach. 2008. Bolasso: model consistent Lasso estimation through the bootstrap. Proc 25th Int Conf Mach Learn 33–40; doi:10.1145/1390156.1390161.

Bai, R., Ghosh, M., 2018. High-dimensional multivariate posterior consistency under global–local shrinkage priors. J. Multivar. Anal. 167, 157–170. https://doi.org/10.1016/J.JMVA.2018.04.010.

Barrera-Gómez, J., Agier, L., Portengen, L., Chadeau-Hyam, M., Giorgis-Allemand, L., Siroux, V., et al., 2017. A systematic comparison of statistical methods to detect interactions in exposome-health associations. Environ Heal 16:74. https://doi.org/10.1186/s12940-017-0277-6.

Basu, T., Einbeck, J., Troffaes, M.C.M., 2021. Uncertainty quantification in lasso-type regularization problems. Optim Under Uncertain with Appl. Aerosp. Eng 81–109. https://doi.org/10.1007/978-3-030-60166-9_3/COVER.

Cao H, Meyer-Lindenberg A, Schwarz E. 2018. Comparative Evaluation of Machine Learning Strategies for Analyzing Big Data in Psychiatry. Int J Mol Sci 2018, Vol 19, Page 3387 19:3387; doi:10.3390/IJMS19113387.

Cao, H., Zhou, J., Schwarz, E., 2019. RMTL: an R library for multi-task learning. Bioinformatics 35, 1797–1798. https://doi.org/10.1093/BIOINFORMATICS/BTY831.

Cao, H., Hong, X., Tost, H., Meyer-Lindenberg, A., Schwarz, E., 2022. Advancing translational research in neuroscience through multi-task learning. Front. Psychiatry 13:2557. https://doi.org/10.3389/FPSYT.2022.993289/BIBTEX.

Chen, L., Huang, J.Z., 2012. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J. Am. Stat. Assoc. 107, 1533–1545. https://doi.org/10.1080/01621459.2012.734178.

Chen, S., Huang, L., Qiu, H., Nebel, M.B., Mostofsky, S.H., Pekar, J.J., Lindquist, M.A., Eloyan, A., Caffo, B.S., 2017. Parallel group independent component analysis for massive fMRI data sets. PLoS One 12, e0173496.

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

Curtis, R.E., Kim, S., Woolford, J.L., Xu, W., Xing, E.P., 2013. Structured association analysis leads to insight into Saccharomyces cerevisiae gene regulation by finding multiple contributing eQTL hotspots associated with functional gene modules. BMC Genomics 14, 1–17. https://doi.org/10.1186/1471-2164-14-196/TABLES/5.

de Abreu e Lima F, Li K, Wen W, Yan J, Nikoloski Z, Willmitzer L, et al. 2018. Unraveling lipid metabolism in maize with time-resolved multi-omics data. Plant J 93: 1102–1115; doi:10.1111/TPJ.13833.

Demissie, S., Cupples, L.A., 2011. Bias due to two-stage residual-outcome regression analysis in genetic association studies. Genet. Epidemiol. 35, 592–596. https://doi.org/10.1002/GEPI.20607.

Descarpentrie, A., Bernard, J.Y., Vandentorren, S., Melchior, M., Galéra, C., Chia, A., et al., 2023. Prospective associations of lifestyle patterns in early childhood with socio-emotional and behavioural development and BMI: An outcome-wide analysis of the EDEN mother–child cohort. Paediatr. Perinat. Epidemiol. 37, 69–80. https://doi.org/10.1111/PPE.12926.

Dunson, D.B., 2018. Statistics in the big data era: Failures of the machine. Stat Probab Lett 136, 4–9. https://doi.org/10.1016/J.SPL.2018.02.028.

Granum, B., Oftedal, B., Agier, L., Siroux, V., Bird, P., Casas, M., et al., 2020. Multiple environmental exposures in early-life and allergy-related outcomes in childhood. Environ. Int. 144 https://doi.org/10.1016/j.envint.2020.106038.

Guo, M., Mao, X., Ji, Q., Lang, M., Li, S., Peng, Y., et al., 2010. miR-146a in PBMCs modulates Th1 function in patients with acute coronary syndrome. Immunol. Cell Biol. 88, 555–564. https://doi.org/10.1038/icb.2010.16.

Hardoon, D.R., Shawe-Taylor, J., 2011. Sparse canonical correlation analysis. Mach. Learn. 83, 331–353. https://doi.org/10.1007/s10994-010-5222-7.

Julvez, J., López-Vicente, M., Warembourg, C., Maitre, L., Philippat, C., Gützkow, K.B., et al., 2021. Early life multiple exposures and child cognitive function: A multicentric birth cohort study in six European countries. Environ. Pollut. 284 https://doi.org/10.1016/j.envpol.2021.117404.

Kim, E.S., Chen, Y., Nakamura, J.S., Ryff, C.D., VanderWeele, T.J., 2022. Sense of Purpose in Life and Subsequent Physical, Behavioral, and Psychosocial Health: An Outcome-Wide Approach. Am. J. Health Promot. 36, 137–147. https://doi.org/10.1177/08901171211038545.

Kim, S., Sohn, K.A., Xing, E.P., 2009. A multivariate regression approach to association analysis of a quantitative trait network. i204–i212 Bioinformatics 25. https://doi.org/10.1093/BIOINFORMATICS/BTP218.

Kuchibhotla AK, Kolassa JE, Kuffner TA. 2022. Post-Selection Inference. 505–527.

Kundu, D., Mitra, R., Gaskins, J.T., 2021. Bayesian variable selection for multioutcome models through shared shrinkage. Scand. J. Stat. 48, 295–320. https://doi.org/10.1111/SJOS.12455.

Li, Q., Zhu, D., Zhang, J., Hibar, D.P., Jahanshad, N., Wang, Y., et al. 2017. Large-scale Feature Selection of Risk Genetic Factors for Alzheimer's Disease via Distributed Group Lasso Regression.

Liu, J., Ji, S., Ye, J., 2012. Multi-Task Feature Learning Via Efficient l2,1-Norm Minimization. Proc 25th Conf Uncertain Artif Intell UAI 2009 339–348.

Luo, C., Liang, J., Li, G., Wang, F., Zhang, C., Dey, D.K., et al., 2018. Leveraging mixed and incomplete outcomes via reduced-rank modeling. J. Multivar. Anal. 167, 378–394. https://doi.org/10.1016/J.JMVA.2018.04.011.

Maitre, L., De Bont, J., Casas, M., Robinson, O., Aasvang, G.M., Agier, L., et al., 2018. Human Early Life Exposome (HELIX) study: a European population-based exposome cohort. BMJ Open 8, e021311.

Maitre, L., Julvez, J., López-Vicente, M., Warembourg, C., Tamayo-Uria, I., Philippat, C., et al., 2021. Early-life environmental exposure determinants of child behavior in Europe: A longitudinal, population-based study. Environ. Int. 153 https://doi.org/10.1016/j.envint.2021.106523.

Miller, R., Breheny, P., 2023. Feature-specific inference for penalized regression using local false discovery rates. Stat. Med. 42, 1412–1429. https://doi.org/10.1002/SIM.9678.

Nieuwenhuijsen, M.J., Agier, L., Basagaña, X., Urquiza, J., Tamayo-Uria, I., Giorgis-Allemand, L., et al., 2019. Influence of the urban exposome on birth weight. Environ. Health Perspect. 127 https://doi.org/10.1289/EHP3971.

Obozinski, G., Wainwright, M.J., Jordan, M.I., 2011. Support union recovery in high-dimensional multivariate regression. Ann. Stat. 39, 1–47. https://doi.org/10.1214/09-AOS776.

Okuzono, S.S., Shiba, K., Kim, E.S., Shirai, K., Kondo, N., Fujiwara, T., et al., 2022. Ikigai and subsequent health and wellbeing among Japanese older adults: Longitudinal outcome-wide analysis. Lancet Reg. Heal - West Pacific 21. https://doi.org/10.1016/j.lanwpc.2022.100391.

Okuzono, S.S., Wilkinson, R., Shiba, K., Yazawa, A., VanderWeele, T., Slopen, N., 2023. Residential instability during adolescence and health and wellbeing in adulthood: A longitudinal outcome-wide study. Health Place 80. https://doi.org/10.1016/J.HEALTHPLACE.2023.102991.

Peng, J., Zhu, J., Bergamaschi, A., Han, W., Noh, D.-Y., Pollack, J.R., et al., 2010. Regularized Multivariate Regression for Identifying Master Predictors with Application to Integrative Genomics Study of Breast Cancer. Ann. Appl. Stat. 4, 53–77. https://doi.org/10.1214/09-AOAS271SUPP.

Santos, S., Maitre, L., Warembourg, C., Agier, L., Richiardi, L., Basagaña, X., et al., 2020. Applying the exposome concept in birth cohort research: a review of statistical approaches. Eur. J. Epidemiol. 35, 193–204. https://doi.org/10.1007/S10654-020-00625-4/FIGURES/1.

Steptoe, A., Fancourt, D,. 2020. An outcome-wide analysis of bidirectional associations between changes in meaningfulness of life and health, emotional, behavioural, and social factors. Sci Reports 2020 101 10:1–12; doi:10.1038/s41598-020-63600-9.

Tibshirani, R., 1996. Regression Shrinkage and Selection Via the Lasso. J. R. Stat. Soc. Ser. B 58, 267–288. https://doi.org/10.1111/J.2517-6161.1996.TB02080.X.

Tseng, V.W.S., Sano, A., Ben-Zeev, D., Brian, R., Campbell, A.T., Hauser, M., et al. 2020. Using behavioral rhythms and multi-task learning to predict fine-grained symptoms of schizophrenia. Sci Reports 2020 101 10:1–17; doi:10.1038/s41598-020-71689-1.

Turlach, B.A., Venables, W.N., Wright, S.J., 2005. Simultaneous Variable Selection.; doi: 10.1198/004017005000000139.

Vanderweele, T.J., 2017. Outcome-wide Epidemiology. Epidemiology 28, 399–402. https://doi.org/10.1097/EDE.0000000000000641.

VanderWeele, T.J., Mathur, M.B., Chen, Y., 2020. Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies. Stat. Sci. 35, 437–466. https://doi.org/10.1214/19-STS728.

Von Hippel, P.T., 2007. Regression with Missing Ys: An Improved Strategy for Analyzing Multiply Imputed Data. Sociol. Methodol. 37, 83–117. https://doi.org/10.1111/j.1467-9531.2007.00180.x.

Vrijheid, M., Fossati, S., Maitre, L., Márquez, S., Roumeliotaki, T., Agier, L., et al., 2020. Early-Life Environmental Exposures and Childhood Obesity: An Exposome-Wide Approach. Environ. Health Perspect. 128, 1–14. https://doi.org/10.1289/EHP5975.

Wang, Y., Liu, K., Han, Q., Yang, H., Zhou, N., Sun, L., et al., 2021. An exposomic approach with 138 chemical and non-chemical exposures to predict 32 biomarkers of male reproductive damages: A case study of college students in Chongqing, China. Sci. Total Environ. 767, 144380 https://doi.org/10.1016/J.SCITOTENV.2020.144380.

Wang, X., Qin, L., Zhang, H., Zhang, Y., Hsu, L., Wang, P., 2015a. A regularized multivariate regression approach for eQTL analysis. Stat. Biosci. 7, 129–146. https://doi.org/10.1007/S12561-013-9106-9.

Wang, X., Zhang, T., Chaim, T.M., Zanetti, M.V., Davatzikos, C., 2015b. Classification of MRI under the presence of disease heterogeneity using multi-task learning: Application to bipolar disorder. Lect Notes Comput Sci (including Subser Lect Notes Artif Intell Lect Notes Bioinformatics) 9349, 125–132. https://doi.org/10.1007/978-3-319-24553-9_16/COVER.

Warembourg, C., Maitre, L., Tamayo-Uria, I., Fossati, S., Roumeliotaki, T., Aasvang, G. M., et al., 2019. Early-Life Environmental Exposures and Blood Pressure in Children. J. Am. Coll. Cardiol. 74, 1317–1328. https://doi.org/10.1016/J.JACC.2019.06.069.

Warembourg, C., Nieuwenhuijsen, M., Ballester, F., de Castro, M., Chatzi, L., Esplugues, A., et al., 2021. Urban environment during early-life and blood pressure in young children. Environ. Int. 146 https://doi.org/10.1016/j.envint.2020.106174.

Wild, C.P., 2005. Complementing the genome with an 'exposome': the outstanding challenge of environmental exposure measurement in molecular epidemiology. Cancer Epidemiol.Biomarkers Prev 14, 1847–1850. https://doi.org/10.1158/1055-9965.EPI-05-0456.