5. Data science and the exposome

The main advantage of the holistic exposome framework over traditional "one-exposure-one-disease" approaches is that it provides an unprecedented conceptual structure for the study of multiple environmental hazards (including urban, chemical, lifestyle, and social risks) and their combined effects. Indeed, classical single pollutant models are unclear as to whether the analysed association can be attributed to the pollutant effect or to another correlated exposure not considered directly in the analysis. Such models are also unable to capture the interactions and cumulative effects derived from the exposure mixture. Furthermore, given the increasing availability of complex environmental health data thanks to the emergence of new technologies (including electronic health records, high throughput omics platforms, wearable sensors, etc.), there is a growing need for more advanced statistical approaches that focus on complex mixtures of exposures.

However, the analysis of such complex data comes with numerous challenges, including, for instance, the typically high correlations between exposures of the same family (e.g. air pollutants and lifestyle), and the ability to capture cumulative low dose effects, assess interactions, and identify important components of the mixture. Recently, a series of different methods have been developed to take into account multiple exposures and their interactions, including the use of mixture analysis methods; the integration of the selection, shrinkage and grouping of correlated variables (e.g. LASSO, elastic-net, adaptive elastic-net); the application of dimension reduction techniques (e.g. principal component and partial least square analyses); Bayesian model averaging (BMA), and Bayesian kernel machine regression (BKMR). Among the limitations of these approaches, however, are the lack of model selection stability (the case of shrinkage methods), the lack of interpretability of the latent variables (the case of dimension reduction techniques), and an overall computational inefficiency (the case of Bayesian models). Moreover,

they are rarely applied in the context of large (>100 variables), heterogenous exposome data (omics, categorical/continuous variables).

Dimensionality reduction

One way of handling multivariate exposomic data (even without resorting to omics) is to employ methods of dimensionality reduction, especially that of feature selection. Feature extraction, by contrast, is less frequently employed since it can complicate the interpretation of the results, given our interest in the effect of a particular exposure on health. However, a number of methods have been developed that seek to analyse groups of correlated exposures. In this way, the dimensionality of the input can be reduced while ensuring interpretability of results.

Combined effect of exposures

Index methods serve to measure the combined effect of exposures. As well as being easy to interpret, they provide both a single parameter estimate for the mixture of exposures and weights to show the contribution of each exposure. However, all index methods suffer from an inability to consider the interactions between the exposures that contribute to the same index. This can, in part, be addressed by using response surface methods, albeit that it potentially hinders interpretation. This tension between interpretability and complexity when choosing between the two types of model has been eased somewhat by recently developed methods (i.e. multiple index models) that combine some of the advantages of both methodologies. These have the advantage of providing readily interpretable indices, while accommodating non-linear and non-additive relationships between exposure indices and the health outcome (McGee et al., 2023).

Bayesian techniques[1] are also useful since they can be used in a manner that naturally penalises complex models yet they are sufficiently flexible to incorporate a variable selection mechanism. They also help obtain the distributions of any quantity that can be derived from the model output.

1. These are statistical methods that involve updating beliefs or probabilities about hypotheses based on prior knowledge and observed data, allowing for the incorporation of uncertainty and the estimation of parameters through probability distributions.

Machine learning and prediction

Machine learning methods – including ensemble methods (such as random decision forests and XGBoost), neural networks and support vector machines – have the potential to increase the predictability of the outcome by capturing more complex information (e.g. complex interactions, non-linear relationships, etc.) from the exposome data. Models that combine multiple statistical techniques into an ensemble can provide even better results, since the different methods employed may be able to capture different data patterns.

Causal models

Finally, causal models – including mediation analyses using omics data, g-computation methods, and the causal random forest – have gained popularity in environmental epidemiology (Bind, 2019), including when making estimates for mixtures of exposures. Indeed, causal questions are what ultimately drive interventions and policy change. Causal mediation analysis with exposome data can help prioritise the environmental factors that have the greatest impact on health.