Tutorial: Exposome analysis techniques

We have developed an **interactive online tutorial** that demonstrates the application of various exposome analysis techniques. Crucially, a sample dataset is included to facilitate hands-on practice.

The tutorial is presented in a *Jupyter Notebook* format, enabling users to read detailed explanations, view R code excerpts for each analytical step, and execute the code directly within the same interface (**Figure 1**). This integrated environment allows users to immediately observe the outputs and follow the reasoning behind each method. This format is particularly user-friendly for those with limited programming experience, as it is fully web-based and does not require local installation of R or any packages—thus avoiding compatibility issues and technical barriers (**Figure 2**). Additionally, by leveraging Google's cloud computing infrastructure, the tutorial significantly reduces computational demands on the user's device, making advanced exposome analysis methods accessible to a broader audience, regardless of their technical resources.

The whole tutorial repository can be found at the following GitHub LINK, and the specific URL to the notebook is LINK

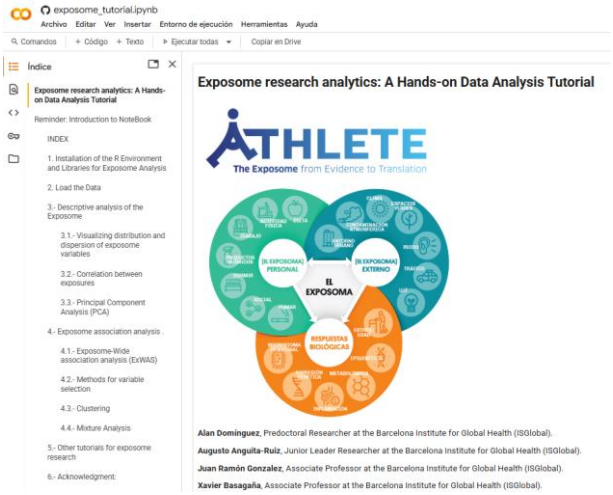


Figure 1. Overview of the Jupyter Notebook tutorial.

Figure 2. Screenshot of the tutorial showing how explanations and code chunks are alternated in a user-friendly and interactive environment.

1. Objective:

The **objective** of this tutorial was to offer a **general introduction to the different statistical approaches** necessary to address the main questions of **exposome research** (covering main aspects of descriptive analysis and association analysis). It is therefore adapted to a **general scientific audience**. In the following pages, we describe the overall contents of the tutorial and provide several screenshots to show how it is implemented.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 874583

2. Tutorial content.

2.1. Tutorial Dataset:

For the tutorial, we used a semi-synthetic dataset that mimics typical exposome data. It consisted of a matrix of environmental exposures, a set of health outcomes, and several covariates that act as confounders in the association between exposures and outcomes. The dataset was semi-simulated using real data from the HELIX sub-cohort, part of the **ATHLETE** project. For more details on the HELIX project and the origin of the collected data, we recommend the following publication: https://bmjopen.bmj.com/content/8/9/e021311 and the project website: https://athleteproject.eu/helix-cohort/ . In particular, with this tutorial, we demonstrate how to preprocess an exposome dataset and how it should be prepared for analysis. The dataset was embedded in a specific R object called an *ExposomeSet*, designed exclusively for exposome analyses. This object is the required input for various functions in the *rexposome* R package. The *ExposomeSet* was constructed by integrating the three main sources of information: environmental exposures, health outcomes, and confounders. The code to create this object is shown in **Figure 3**.

```
[ ]  exp <- rexposome::loadExposome(
       exposures = expo2[expo.list],
       description = codebook[expo.list,],
       phenotype = dat,
       description.famCol = "family"
       )
```

```
[ ]  dplyr::glimpse(exp)
```

```
Formal class 'ExposomeSet' [package "rexposome"] with 7 slots
  ..@ assayData        :<environment: 0x570df8e51e10>
  ..@ phenoData        :Formal class 'AnnotatedDataFrame' [package "Biobase"] with 4 slots
  ..@ featureData      :Formal class 'AnnotatedDataFrame' [package "Biobase"] with 4 slots
  ..@ experimentData   :Formal class 'MIAME' [package "Biobase"] with 13 slots
  ..@ annotation       : chr(0)
  ..@ protocolData     :Formal class 'AnnotatedDataFrame' [package "Biobase"] with 4 slots
  ..@ .__classVersion__:Formal class 'Versions' [package "Biobase"] with 1 slot
```

Figure 3. Screenshot of the tutorial: preparation of exposome dataset.

2.2. Descriptive Analysis:

In the first part of the tutorial, the concept of descriptive analysis in exposomics was addressed, through which the first conclusions about the data are drawn. Among other objectives, descriptive analysis aims to identify possible outliers, confounding factors, or variables that might require transformations before analysis. At the same time, descriptive analysis allows a preliminary comparison of the experimental groups under study, the examination of the existing patterns of correlation among exposure factors, and the identification of grouping phenomena in the data (both at the level of individuals and features). **All of these are essential steps to choose the most appropriate subsequent statistical approach.**

Some of the contents covered in this section included:

- **Visualization of the distribution and concentration of exposome variables.**

  We employed different visualisations (such as histograms of boxplots) to check the distribution of categorical and continuous variables, further stratifying by population groups of interest (e.g., sex).

- **Correlation between exposures.**

  Correlation between variables is an important phenomenon to take into consideration when we want to do exposome analysis. To look at the intra- and inter-family correlation of different exposures, we showed how to use the function *rexposome::correlation*.

- **Principal Component Analysis (PCA) applied to exposome variables.**

  We also employed PCA as an unsupervised machine learning algorithm for exploratory analysis and to reduce the dimensionality of data (**Figure 4**).

With these three sub-sections, the tutorial aimed to demonstrate how we can extract important conclusions on the dataset and take critical pre-processing decisions before secondary analyses. For example, in our example, we demonstrated how the exposures under study (exposome features) cluster in specific regions in the exposure space, indicating that some features are similar to each other. Furthermore, we showed how the individuals or participants also cluster, which could indicate possible patterns within the data (e.g., cohort structure).

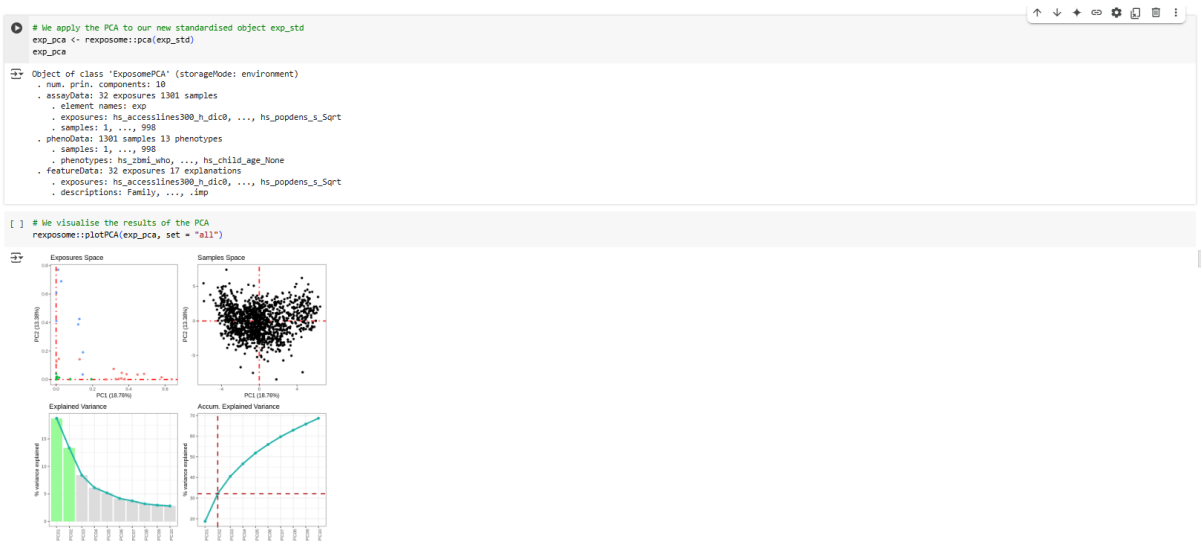


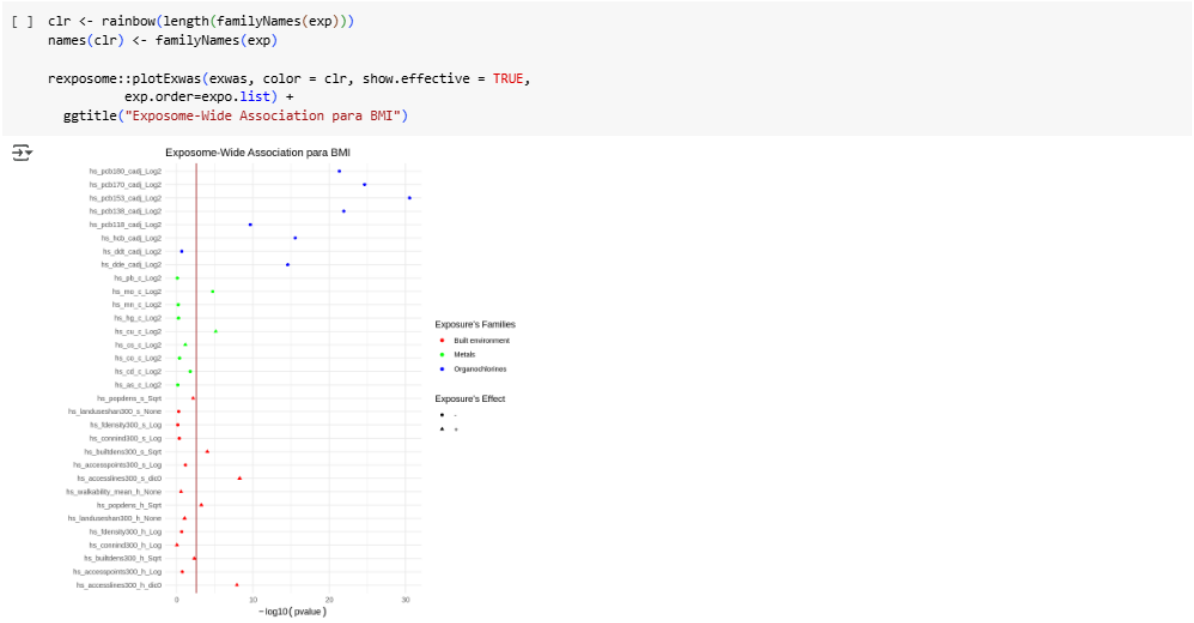
Figure 4. Screenshot of the tutorial: Dimensionality Reduction Analysis.

2.3. Association Analysis:

Association analysis aims to identify possible environmental exposure factors associated with different health parameters. In this section of the tutorial, different holistic analytical approaches were presented for the study of the effects of multiple exposure factors (and their mixtures) on health. This included simple models such as ExWAS (Exposome-Wide Association Analysis), or others for the study of interactions or non-linear phenomena (e.g., Bayesian Kernel Machine Regression). An introduction was also presented to clustering methods or exposure mixture methods (e.g., Weighted Quantile Sum Regression). In each sub-section, concepts of great importance in exposome analysis were introduced, such as feature selection or multiple testing correction.

- **Exposure Wide Association Analysis (ExWas)**

  In this sub-section, we presented the ExWAS (Exposome-Wide Association Study) method as an approach designed to handle high-dimensional data and discussed the importance of controlling for multiple testing. In addition to providing code and guidance on how to perform this analysis, we also demonstrate how to visualise the results using gold-standard plots, such as forest plots and volcano plots (**Figure 5**). The tutorial also included "example questions" and interpretations based on the data, helping readers to better understand how to extract meaningful conclusions from these visualisations.



- **Question 1: Is ExWas analysis controlled by multiple testing? Yes, we can define it explicitly in the function and all derived p-values will be corrected for multiple testing. This correction can be done according to the total number of analyses or, considering just the effective number of tests (considering correlation)**

- **Question 2: If a participant is exposed to PCB153, can we say that if he/she is also exposed to PCB118, will his/her BMI be reduced? ExWAS does not tell us anything about co-exposure effects and co-ocurrence of events. It just looks at the association between each individual exposure and the health outcome (adjusted for potential confounders) independently**

Figure 5. Screenshot of the tutorial: ExWAS analysis interpretation.

- **Methods for variable selection (Stepwise, Elastic net, DSA)**

  In this section, we presented more advanced modelling approaches beyond ExWAS, which allow for improved identification of variables or exposures that are not associated with a health outcome, while also accounting for co-exposure effects. The methods introduced are among the most widely used in exposome research and included stepwise selection, elastic net regression, and the Deletion-Substitution-Addition (DSA) algorithm.

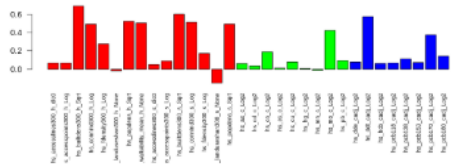
- **Clustering**

  Identifying exposure patterns or specific subgroups within the population is of particular interest in exposome research. For this reason, we included a dedicated sub-section on the

most commonly used clustering technique in this field. In particular, we showcase how the **k-means** algorithm can be applied to identify groups of individuals with similar exposure profiles—distinct from those of other groups in the population. These exposure-based clusters can then be used in subsequent analyses to explore associations between exposures and health outcomes (**Figure 6** and **7**).

- **Cluster 1 (N = 239):** They live in populated, dense, trafficable areas; exposed to Mo and DDT.

```
[ ] options(repr.plot.width=10, repr.plot.height=4) # definimos el tamaño del gráfico en colab
    par(mar = c(8, 4, 4, 2) + 0.1)  # c(bottom, left, top, right) ajustamos los margenes
    barplot(as.numeric(clus.means[1,2:ncol(clus.means)]),
            col=c(rep("red",15),rep("green",9),rep("blue",8)),
            names.arg=names(clus.means)[-1],
            cex.names=.7,
            las=2,
            srt=90)
```



- **Cluster 2 (N = 425):** High exposure to DDT but low exposure to other organochlorines; Low population density (possibly rural).

```
[ ] options(repr.plot.width=10, repr.plot.height=4) # we define the size of the graph for colab
    par(mar = c(8, 4, 4, 2) + 0.1)  # c(bottom, left, top, right) we adjust the margings
    barplot(as.numeric(clus.means[2,2:ncol(clus.means)]), col=c(rep("red",15),rep("green",9),rep("blue",8)),
            names.arg=names(clus.means)[-1],
            cex.names=.6,
            las=2,
            srt=90)
```

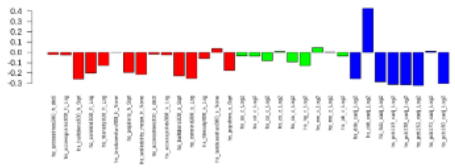


Figure 6. Screenshot of the tutorial: K-means analysis.

Finally, we can explore the association of each of the clusters with the health outcome of interest by adjusting for some covariates of interest.

```
mod_cluster <- lm(hs_zbmi_who ~ as.factor(km.res$cluster) + h_cohort + e3_sex_None + e3_yearbir_None, data = data) # we adjust the analysis by cohort, sex and year of birth
summary(mod_cluster)
```

```
Call:
lm(formula = hs_zbmi_who ~ as.factor(km.res$cluster) + h_cohort +
    e3_sex_None + e3_yearbir_None, data = data)

Residuals:
     Min       1Q   Median       3Q      Max
-3.7247 -0.7393 -0.0751  0.7175  3.8131

Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              -0.31055    0.34124  -0.910  0.36296
as.factor(km.res$cluster)2  0.44636    0.09062   4.926 9.51e-07 ***
as.factor(km.res$cluster)3 -0.19275    0.09686  -1.990  0.04681 *
as.factor(km.res$cluster)4  0.90059    0.13091   6.879 9.37e-12 ***
as.factor(km.res$cluster)5  0.26706    0.10806   2.471  0.01359 *
h_cohort2                  0.38921    0.29804   1.306  0.19182
h_cohort3                  0.95897    0.23162   4.140 3.69e-05 ***
h_cohort4                  0.35031    0.11296   3.101  0.00197 **
h_cohort5                  0.24656    0.21464   1.149  0.25090
h_cohort6                  0.50678    0.11340   4.469 8.56e-06 ***
e3_sex_Nonemale            0.15818    0.06253   2.530  0.01154 *
e3_yearbir_None2004       -0.21431    0.18550  -1.155  0.24816
e3_yearbir_None2005       -0.16931    0.23934  -0.707  0.47944
e3_yearbir_None2006       -0.14731    0.26219  -0.562  0.57432
e3_yearbir_None2007        0.10043    0.31818   0.316  0.75232
e3_yearbir_None2008        0.18256    0.33244   0.549  0.58299
e3_yearbir_None2009        0.49407    0.45248   1.092  0.27508
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.116 on 1284 degrees of freedom
Multiple R-squared:  0.1314,    Adjusted R-squared:  0.1205
F-statistic: 12.14 on 16 and 1284 DF,  p-value: < 2.2e-16
```

Figure 7. Screenshot of the tutorial: Exposome-health association analysis based on k-means cluster groups.

- **Mixture analysis:**

  Finally, we introduced mixtures analysis, a crucial step in modeling the health effects of environmental exposures. Unlike traditional approaches that assess exposures individually, mixtures analysis reflects real-life scenarios, where individuals are simultaneously exposed to multiple risk factors. For example, low levels of a single pollutant may have negligible or undetectable health effects, but combined exposure to several pollutants can result in significant impacts. Studying these complex mixtures is a key focus in environmental epidemiology and a central objective of the **ATHLETE project**.

  In this sub-section, we presented two advanced methods that enable the assessment of the joint effects of multiple exposures, as well as potential interactions between them:

  - Weighted Quantile Sum Regression (WQS)
  - Bayesian Kernel Machine Regression (BKMR)

  The tutorial guides how to implement these models in R, covering key aspects such as data preprocessing, interpretation of results, and recommended visualisation techniques.

2.4. Other tutorials:

Since this tutorial was thought to compile main analyses that can be done in exposome research, we decided to include links to other specific tutorials that have been developed during the past years within the context of ATHLETE project, serving as a concentrated resource for exposome data analysis.

3. Conclusion:

This tutorial effectively walks users through a modern exposome analysis pipeline using real-world tools. Its use of a synthetic yet realistic dataset allows users to focus on mastering methods rather than data cleaning. With its balance of theory and practice, it serves as a valuable resource for epidemiologists, biostatisticians, and environmental health researchers aiming to work with high-dimensional exposure data.