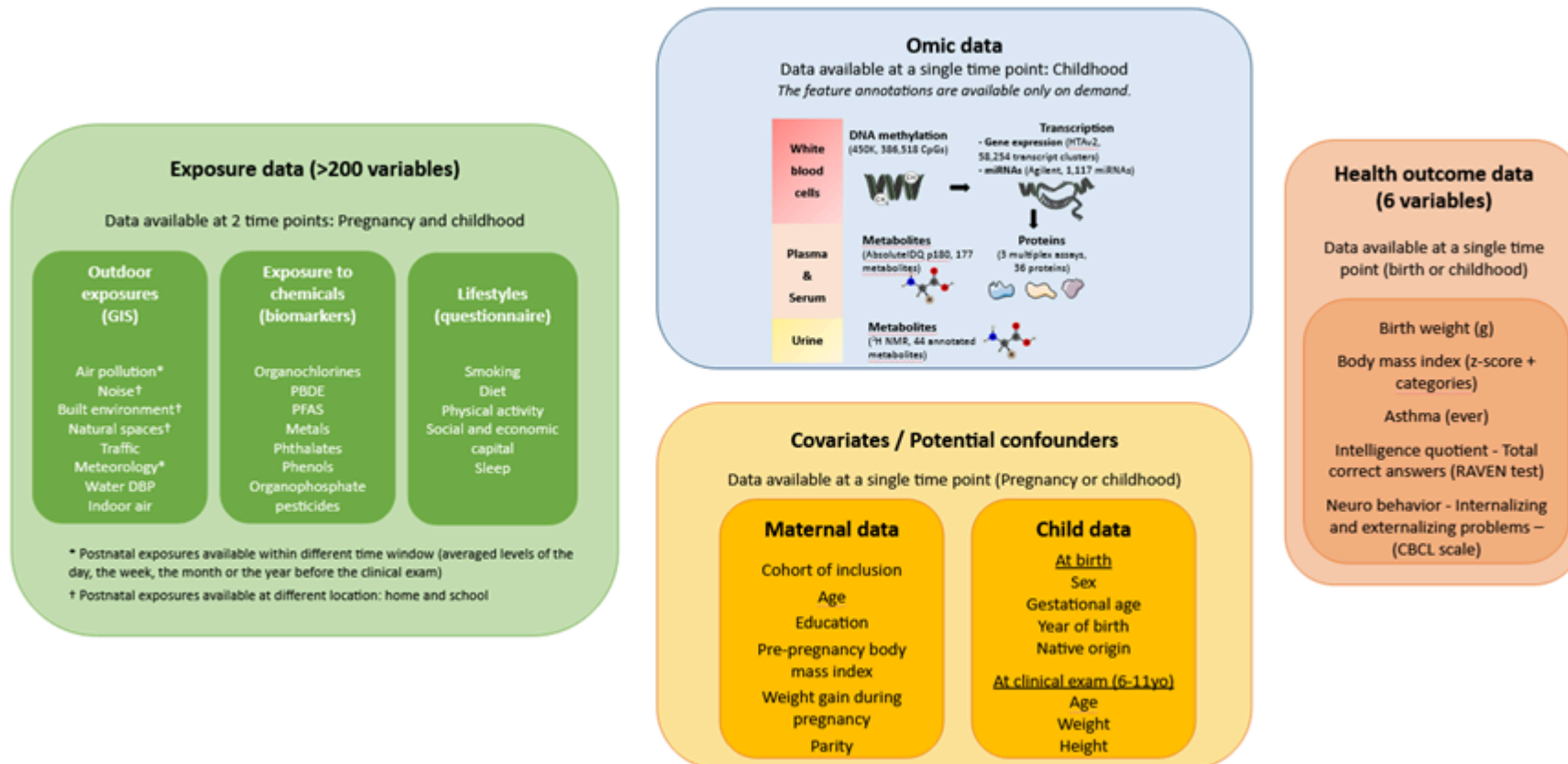


## Data description:



The data provided were simulated following the structure of the HELIX subcohort database. Further details on the HELIX project and the origin of this dataset can be found in this publication: <https://pubmed.ncbi.nlm.nih.gov/30206078/> and in this website: <https://www.projecthelix.eu/index.php/es/data-inventory>.

An overview of available data is shown in the figure. They can be downloaded from [here](#)

Two sets of datasets are available: 1 with complete case data and 1 including missing data. Each set is composed by:

- 1 dataset for exposures: exposome
- 1 dataset for covariates: covariates
- 1 dataset for outcomes: phenotype

For omic data we have also provided data following HELIX subcohort database

- [Serum Metabolome](#): ExpressionSet called `metabol_serum` of 1198 individuals and 177 metabolites (see [here](#) for a description). -
- [Urine Metabolome](#): ExpressionSet called `metabol_urine` of 1192 individuals and 44 metabolites (see [here](#) for a description).
- [Gene expression](#): ExpressionSet called `genexpr` (see [here](#) what an ExpressionSet is) of 1007 individuals and 28,738 transcripts with annotated gene symbols.
- [Methylation](#): GenomicRatioSet called `methy` (see [here](#) what a GenomicRatioSet is) of 918 individuals and 386,518 CpGs

To facilitate the analysis, exposure variables were transformed to approach normal distribution.

You can refer to the [codebook](#) for variable description (variable name, domain, type of variable, transformation, ...)

Health outcome data:

- 2 continuous variables: Birthweight and child BMI z-score
- 2 count variables: Intelligence quotient and Neuro behavior
- 2 categorical variables: Asthma (two categories) and child BMI (4 categories, that can be combined if needed)

## **Challenge examples:**

### **Challenge 1: Combined effects of exposures**

- Determine if there are particular combinations of exposures (“cocktail effects”), high-order interactions or exposure patterns that are particularly harmful or beneficial for one or several health outcomes.
- Handle the multicenter design of the study (i.e. center may be a strong determinant of exposure patterns).
- Control for potential confounders.

**Challenge 2: Using omics data to improve inference on the link between exposome and health.**

- Incorporate the different omics layers into the analysis linking the exposome and one or more health outcomes.
- Show how the extra information available in the omics data can improve the inference of an analysis using only exposome and health (e.g. improvement in statistical power, ...).
- Control for potential confounders and multicenter design.

**Challenge 3: Multi-omics analysis**

- Incorporate different layers of omics data (including exposome as one of the layers) to find patterns that can explain variations in one or more health outcomes.
- Control for potential confounders and multicenter design.
- Maximize power in a context of moderate sample size.

**Challenge 4: Causal structure in the exposome**

- Define hypothesised causal relationships between the different exposures and one health outcome, and incorporate this information into the analysis.
- Compare this approach with *agnostic* analyses that perform variable selection treating all exposures in the same way.
- Illustrate how one can answer a large number of causal questions referring to different exposures using causal inference techniques for high-dimensional data
- Incorporation of mediation analysis and high-dimensional mediation analysis is a welcome addition.

**Challenge 5: Visualization techniques**

- Tools to visualize the complex relationships between the different components of the analysis, with the main aim of illustrating determinants of health effects.
- Strong interest in visualizing the magnitude and direction of the associations.
- Incorporate multiple data types.
- Find the right balance between complexity and clarity.