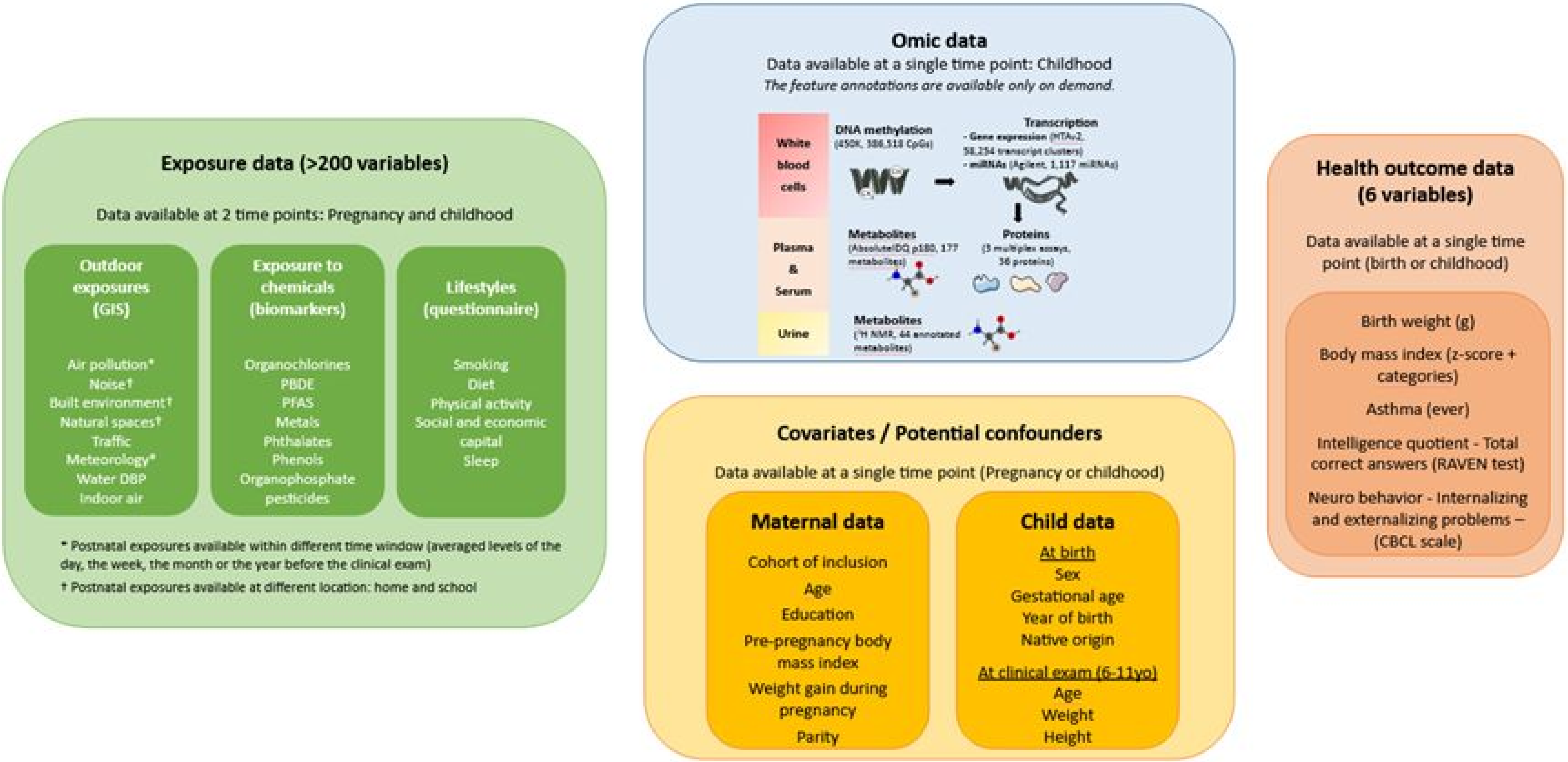
Data description:


Exposure data (>200 variables)
Data available at 2 time points: Pregnancy and childhood
Outdoor exposures (GIS)
Air pollution*
Noise†
Built environment†
Natural spaces†
Traffic
Meteorology*
Water DBP
Indoor air
Exposure to chemicals (biomarkers)
Organochlorines
PBDE
PFAS
Metals
Phthalates
Phenols
Organophosphate pesticides
Lifestyles (questionnaire)
Smoking
Diet
Physical activity
Social and economic capital
Sleep
* Postnatal exposures available within different time window (averaged levels of the day, the week, the month or the year before the clinical exam)
† Postnatal exposures available at different location: home and school
Omic data
Data available at a single time point: Childhood
The feature annotations are available only on demand.
White blood cells
DNA methylation (450K, 386,518 CpGs)
Transcription
- Gene expression (HTAv2, 58,254 transcript clusters)
- miRNAs (Agilent, 1,117 miRNAs)
Plasma & Serum
Metabolites (AbsoluteIDQ p180, 177 metabolites)
Proteins (3 multiplex assays, 36 proteins)
Urine
Metabolites (^{1}H NMR, 44 annotated metabolites)
Covariates / Potential confounders
Data available at a single time point (Pregnancy or childhood)
Maternal data
Cohort of inclusion
Age
Education
Pre-pregnancy body mass index
Weight gain during pregnancy
Parity
Child data
At birth
Sex
Gestational age
Year of birth
Native origin
At clinical exam (6-11yo)
Age
Weight
Height
Health outcome data (6 variables)
Data available at a single time point (birth or childhood)
Birth weight (g)
Body mass index (z-score + categories)
Asthma (ever)
Intelligence quotient - Total correct answers (RAVEN test)
Neuro behavior - Internalizing and externalizing problems – (CBCL scale)

The data provided were simulated following the structure of the HELIX subcohort database.

An overview of available data is shown in the figure.

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
- OMIC (coming soon)

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.