**Lab 1: High-dimensional exposure data in the Exposome dataset**

The aim of this lab is to introduce you to the key parts of the workshop that were discussed during the first day. Specifically:

* Explore correlations and exposure networks
* Run univariate and multivariable regression (linear and logistic)
* Check for non-linearity in exposure–outcome associations
* Apply and interpret penalized regression methods (Ridge, LASSO, Elastic Net)

The data is derived from the Exposome research project and freely available online. In the labs we work with a subset of the data. The dataset is available as an excel file. See the codebook file for the variable names and explanation.

The variables in the dataset are as follows:

**Outcomes**

* **bw**: continuous birthweight outcome
* **lbw**: binary low birthweight (<2500g) derived variable

**Exposures groups**

* Total: 56 exposure variables
* 16 pollutants (variables starting with a “p”)
* 22 phthalates (variables starting with a “ph”)
* 18 metals (variables starting with a “m”)

1. **Prepare the dataset**

* Load and inspect the dataset
* Identify missing values (visualize with gg\_miss\_var)
* Explore outcome distributions (summary statistics, histograms, binary categorization for LBW)
* Explore selected exposure distributions (histograms)
* Scale exposures (standardization, log-transformations)

1. **Investigate the data**

* Compute Spearman correlations between exposures
* Visualize correlation matrix (corrplot)
* Build exposure networks with qgraph to explore clustering and collinearity
* Identify highly correlated exposures (e.g., >0.8, >0.9)

1. **Check univariate associations**

* Logistic regression for binary outcome (LBW)
* Linear regression for continuous outcome (BW)
* Adjust p-values for multiple testing (Bonferroni)
* Visualize results using for example volcano plots
* Compare strongest associations across exposures

1. **Non-Linearity**

* Fit restricted cubic splines for continuous exposures
* Visualize spline fits with prediction bands
* Identify exposures with evidence of non-linear relationships

1. **Multivariable regression**

* Fit multivariable regression model including all exposures
* Assess multicollinearity (Variance Inflation Factor)
* Discuss interpretation and limitations in highly correlated data

1. **Penalized regression**

* Apply Ridge, LASSO, and Elastic Net regression using glmnet
* Compare variable selection and stability of coefficients
* Use cross-validation to identify optimal penalty parameters
* Extend models to adjust for potential confounders (covariates with penalty = 0)
* Discuss interpretation vs prediction