Aim one analysis notes

Table of Contents

[01: Pre-analysis scripts 1](#_Toc123727915)

[`01\_clean\_raw\_nhanes\_data.R` 1](#_Toc123727916)

[`01\_mgi\_partial\_correlations.R` 1](#_Toc123727917)

[`01\_prepare\_phenomes.R` 1](#_Toc123727918)

[02: Analysis scripts 1](#_Toc123727919)

[`02\_cooccurrence\_analysis.R` 2](#_Toc123727920)

[`02\_regression\_based\_weights\_using\_nhanes.R` 2](#_Toc123727921)

# 01: Pre-analysis scripts

Pre-analysis scripts either perform cleaning necessary for further analysis or do not require data to be cleaned prior to their running. There are currently three pre-analysis scripts:

- `01\_clean\_raw\_nhanes\_data.R`

- `01\_mgi\_partial\_correlations.R`

- `01\_prepare\_phenomes.R`

## `01\_clean\_raw\_nhanes\_data.R`

This may no longer be necessary since NHANES data preparation and cleaning is being integrated into the regression based weighting script.

## `01\_mgi\_partial\_correlations.R`

This simply calculates partial correlations in MGI. It takes quite a long time to run (~7 hours). By default, Pearson correlation coefficients are adjusted for age, sex, and first four PCs of the genotype data. The script can be run without the PCs. Correlations are outcome agnostic and rely on the full phecode indicator matrix.

- [ ] Create a script for calculating analogous partial correlations in UKB

## `01\_prepare\_phenomes.R`

This script takes several inputs including the outcome of interest, the version of the MGI data, the version of the UKB data, time thresholds, and information on matching (variables, caliper, ratio) and prepares matched, time-restricted covariate and phecode datasets for use in cooccurrence analysis.

This script has been `optparse`-ified, meaning it can be run directly from the command line. In other words, the script can be looped over any number of the inputs and run quickly and simultaneously (e.g., many different outcomes).

It is considered operational.

# 02: Analysis scripts

Analysis scripts perform analyses like PCA or co-occurrence analyses. There currently are three analysis scripts:

- `02\_cooccurrence\_analysis.R`

- `02\_pca\_analysis.R`

- `02\_regression\_based\_weights\_using\_nhanes.R`

## `02\_cooccurrence\_analysis.R`

This script relies on outputs from `01\_prepare\_phenomes.R`. Assuming `01\_prepare\_phenomes.R` has been run for a given combination of outcome, data versions, and time thresholds.

It runs \*unweighted\* co-occurrence analyses for a given outcome, set of data versions (MGI and UKB), and set of time thresholds. Co-occurrence models are Firth bias-corrected logistic regression and are adjusted for age at time threshold (continuous), sex, and length of follow-up. There is an option to run using adjusted saddle point approximation via SPAtest as well.

This script outputs a table with the following variables:

- Phecode (corresponding to the independent variable)

- Beta

- SE beta

- P-value

This script has been `optparse`-ified, meaning it can be run directly from the command line. In other words, the script can be looped over any number of the inputs and run quickly and simultaneously (e.g., many different outcomes).

It is considered operational.

An alternative script is being drafted to calculate Beta regression-based weights using NHANES data to run \*weighted\* logistic regression models. This script is provisionally names `02\_regression\_based\_weights\_using\_nhanes.R`.

## `02\_regression\_based\_weights\_using\_nhanes.R`

This script is currently under construction.

The goal is to generate weights using NHANES data.

It first reads and processes NHANES data.

It then estimates , where is estimated using NHANES data to fit

And is estimated using ALL (MGI and NHANES) data to fit

Predicted values using only MGI data are generated on both models and generate weights

Notably, these models omit cancer (of any kind).