Sex differences in PheEWAS analysis plan

The purpose of this analysis is to discover phenome-environment-wide associations that have differential effects between sexes. Most of the pipeline, phenotype and variable selection will follow Nikki's plan

1. Initial QC process
    - Drop any variables that are indeterminate according to the NHANES data dictionary
    - Remove variables that don't have 4 year weights
    - Drop any non-environmental exposures (examples physical fitness)
    - Determine covariates and phenotypes
    - From the remainder of the variables split them based on their type
    - Manually inspect the definition of ambiguous variables and determine their type
    - Merge binary into categorical variables
2. Make an adjustment to the lipid variable (LBDLDL) based on statin medication
3. Split by adults (greater than 18 yo) and subadults, and use only adults thereafter
4. Split dataset into discovery (series 1 and 2) and replication (series 3 and 4) datasets
    - Make sure the phenotypes from the dataset categorized under biochemistry and blood lab test/measures are included in at least one series from Discovery and Replication
    - Nikki's list of 58 phenotypes already has taken care of the previous step
5. Drop any variable that has a missing value in the covariate list
6. Separate datasets into phenotypes and exposures and continue with QC
7. Remove phenotypes with more than 90% of samples with 0 value
    - Nikki's list of phenotypes already has those removed
8. Transform variables due to skewness and non-normality
    - The list of transformations is again replicated from Nikki's analyses
    - Those with negative skew are mirrored by subtracting the maximum value plus 1
    - Non normal distributions are log transformed

Given the complexity of the ewas models, it is easier and more convenient to run a stratified ewas and test for sex differences after. Winkler et al (2017) recommend following two approaches in parallel in genome-wide association studies if there is no prior hypothesis on sex differences: to run a genome-wide difference test between sexes, and another approach that first filters for an overall association and then test for the difference between sexes with a Bonferroni corrected alpha.

Phenome-environment-wide sex difference test

Considering two sexes, $i = 1, 2$, let Y_p be a vector of phenotypes, where $p = 1..., P$ considering P phenotypes, and X_q is a vector of environmental exposures, $q = 1..., Q$, considering Q environmental exposures. We write the linear regression as:

$$Y_{ip} = X_{iq}\beta + \epsilon$$

Our interest will be focused on β_{ipq} which is the beta coefficient of the effect of the environmental exposure q on phenotype p, in sex i, with its corresponding standard error se_{ipq}. For the phenome-environment-wide sex difference test we will estimate the *difference test* as:

$$Z_{diff} = \frac{\beta_{1pq} - \beta_{2pq}}{\sqrt{se^2_{1pq} - se^2_{2pq}}}$$

We will use the Bonferroni correction for multiple testing.

Filtering by overall association

The second pipeline we will use to estimate sex differences incorporates a filtering based on overall association before the difference test. On a stratified approach, the *overall test* is given by:

$$Z_{overall} = \frac{\frac{\beta_{1pq}}{se^2_{1pq}} + \frac{\beta_{2pq}}{se^2_{2pq}}}{\sqrt{\frac{1}{se^2_{1pq}} + \frac{1}{se^2_{2pq}}}}$$

Both the filtering by overall test and the subsequent difference test will use the Bonferroni correction for multiple testing.

References

Winkler, Thomas W., Anne E. Justice, L. Adrienne Cupples, Florian Kronenberg, Zoltán Kutalik, Iris M. Heid, and GIANT consortium. 2017. "Approaches to Detect Genetic Effects That Differ between Two Strata in Genome-Wide Meta-Analyses: Recommendations Based on a Systematic Evaluation." PloS One 12 (7): e0181038. https://doi.org/10/gbqbnm.