Sex differences in PheEWAS analysis plan

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

Selection of variables

To assure consistency in the pre-selection of variables, we selected them based on the category used in NHANES. Therefore, for phenotypes, we selected `biochemistry`, `blood`, and `hormones` variables. The `biochemistry` category include diverse blood and urine biomarkers. The `blood` category includes complete blood count measurements. Finally, the `hormones` category includes estimation of different hormones in blood samples.

In the case of exposures, the list is more or less self explanatory: 'alcohol use', 'bacterial infection', 'cotinine','diakyl', 'dioxins', 'food component recall', 'furans', 'heavy metals', 'housing', 'hydrocarbons', 'nutrients', 'occupation', 'pcbs', 'perchlorate', 'pesticides', 'phenols', 'phthalates', 'phytoestrogens', 'polybrominated ethers', 'polyflourochemicals', 'sexual behavior', 'smoking behavior', 'smoking family', 'social support', 'street drug', 'sun exposure', 'supplement use', 'viral infection', and 'volatile compounds'.

In terms of the covariates, those are: 'black', 'mexican', 'other_hispanic', 'other_eth', 'SES_LEVEL', 'RIDAGEYR', 'SDDSRVYR', 'BMXBMI'.

Finally, some categories were left out, such as: - acrylamide: too few variables in a single survey cycle - aging: too few variables in 2 cycles (telomeres) (might add it) - allergen test: only in one survey cycle - blood pressure: not sure where else to classify them - body measures: not sure where else to classify them - cognitive functioning: not sure where else to classify them - disease: not sure where else to classify them - immunization: not sure where else to classify them - pharmaceuticals: not sure where else to classify them - physical fitness (cardiovascular fitness): not sure where else to classify them

We start with 8 predefined covariates and 55 phenotypes.

1. Selecting participants:
   - Keep only participants older than 18 years (`'RIDAGEYR' >= 18`)
   - Drop any participant that has a missing value in the covariate list
2. Variable selection and cleanup:
   - Keep only variables with weights in the survey weight file
   - Some weight variables from replication dataset are not in the weights_replication file, so we'll remove them
   - Recode `SMQ077` and `DDB100` values of 7 and 9 ("Refused/Don't Know") to NA
   - Drop physical fitness measurements, indeterminate variables and age groups

    - Drop variables that are transformations of phenotype variables (`LBXSCRINV`,`URXUMASI`,`LBXSIR`)
    - Drop body measure variables
    - Make an adjustment to the lipid variable (`LBDLDL`) based on statin medication ($LDL = LDL/0.7$) (Question: shouldn't we control for all medications?)
3. Split the dataset into four cohorts: discovery females and males, replication females and males
4. QC process (in each cohort independently):
    - Categorize variables
    - Remove constant variables
    - Manually categorize unknown variables
    - Remove categorical variables (because it won't estimate β coefficients)
    - Remove variables that have less than 200 non-NA values
    - Remove binary variables that have less than 200 values in a category
    - Remove continuous variables with 90% of non-NA observations equal to zero
5. Make sure there are no discrepancies in variable categories across the four cohorts, and keeping variables that were kept across cohorts
6. Variable transformation:
    - Merge cohorts
    - Log-transform continuos variables that are highly skewed (greater than -0.5 or 0.5)
    - Normalize variables (min-max approach)
    - Remove outliers
7. Split back into the four cohorts
8. Remove unnecessary variables (male, female, and white)

Given the complexity of the EWAS models, it is easier and more convenient to run a stratified EWAS and test for sex differences after. Therefore, we will run four separate EWAS models for each cohort independently and estimate the corresponding parameters (β, se). Winkler et al (2017) recommends 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 to search for opposite effects, and another approach that first filters for an overall association and then test for the difference between sexes to search for those with differences in the size of the effect, or for those that there is no effect in one sex. The following categories will be used to refer to those differences in effects:

- Qualitative: exposures that have opposite effects between sexes
- Quantitative: exposures have the same direction of effect, but the effect is larger in one sex
- Pure: exposures have an effect in only one sex and not in the other

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:

$$Z_{diff} = \frac{\beta_{1pq} - \beta_{2pq}}{\sqrt{se_{1pq}^2 + se_{2pq}^2}}$$

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:

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

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

Step-by-step analysis

Finally, the pipeline will follow these steps:

1. To detect qualitative effect differences we will run the difference test and select those with a Bonferroni corrected α. Then, we will keep only those with opposite effects (different directions of β coefficients, and nominally significant in both sexes)
2. To detect quantitative and pure effect differences we will first filter by an overall effect $p - value < 0.05$ and then select those with a difference test with a Bonferroni corrected α. We will classify those with β coefficients in the same direction and with a nominal p-value in both sexes as a quantitative difference, while those with a nominal p-value in only one sex will be classified as a pure difference

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.