CARDIA-ADPQS-replication

LekkiWood (LekkiWood@Gmail.com)

2026-02-04

Table of contents

Important Notes - read me first

Version control

- Always check that you have the most recent version of this document, which - unless I am sending you unfinalized work - is available here.
- An easy check for version control is to make sure this date: 2026-02-04. is the same as on the GitHub file here.
- The code for this is analysis available in the same repository (targets master file here and individual functions here)

Step 1: Cleaning and Formatting Proteins

Input file names

- A table of protein abundances: SMP_IntensityNormalized_20251005.csv
- Sample information to link TOPMed IDs to unique MESA SHARe ID and exam combinations: Mapping_SMP_Plate_20251005.csv
- Keys to link Olink IDs to names compounds: MESAOlink3k_proteinKeys_03292023.csv
- A file to bridge SHARe ids (sidno) with MESA IDs (idno) MESA-SHARE_IDList_Labeled.csv

Raw file info

- The raw protein abundance file contained information on N=3040 protein assays, including those used for QC.
- When removing assays for QC, the raw protein abundance file contained information on N=2941 proteins.
- The protein abundance file contained information on N=14051 sample IDs (i.e., unique participant/exam combinations), including bridging samples.
- After removing QC samples (including bridging, controls, and one duplicate) the protein abundance file contained information on N=12739 sample IDs (i.e., unique participant/exam combinations).

Table 1: Final N by exam

Exam	N_Pps
1	5949
5	3917
6	2873

Formatting

- Bridging (and other QC) samples were removed.
- Protein assays used for QC were removed.
- Proteins that should be excluded due to QC warnings (variable "QC_warning" set to "EXCLUDED") were removed, even though these do not have NPX values.
- Data were put into wide format, with "SampleID" as the unique ID, "OlinkID" forming the variable names (protein identifiers), and values taken from the "NPX" column.
- In wide format, the file contained information on N=12739 unique sample IDs.
- In wide format, the file contained information on N=0 duplicated sample IDs. [1]
- SHARe IDs, and subsequently MESA IDs, were merged into the file with exam information.
- At this point, the range of unique SHARe ID by exam combinations was N=0 - 1. This indicates no sample ID were duplicated in the assays.
- The formatted protein file was used to calculate the coefficient of variation (CV) using the formula: CV = sqrt(2^(sigma^2)-1).
- A variable called "Retain" was created to indicate whether each protein was (1) unique (i.e., included on only one panel); (2) duplicated, and across all panels had the lowest CV; or (3) duplicated, and across all panels did not have the lowest CV.
- A final table of protein abundances, with additional variables for SHARe ID, MESA ID, Exam, TOPMed ID and Batch, was created after the steps above, with proteins duplicated across more than one panel cleaned such that only the one with the lowest CV is retained. This file was used in the analysis
- The number of participants, stratified by exam, in the final file is available in Table 1: