CARDIA-ADPQS-replication

LekkiWood (LekkiWood@Gmail.com)

2026-02-05

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: 2026-02-05. 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:

Step 2: Build traits

Input files

- Covariates MESAe1FinalLabel02092016.dta
- Diet data (for exclusions) E1_nutrients_new.csv
- Incident events MESAEvThru2020AllCohort_20241120.dta

Coding

Outcomes

- CVD event data was time to any hard CVD event, as defined by MESA.
- The total number of person years included in the analysis is 7.7767079×10^4.

Covariates

- Race was coded as a factor variable; coded 1= European-American, 2 = Chinese-American, 3 = African-American, 4 = Hispanic-American
- Sex was coded as a factor variable; coded 1 = female, 2 = male
- Age was baseline age
- BMI was baseline BMI in kg/m2 with height and weight taken by trained study staff
- Diabetes was binary; 1 = no diabetes, 2 = diabetes according to ADA 2003 criteria (fasting glucose, use of medication, self-reported diagnosis) and so includes both treated and untreated diabetes
- Smoking is a continuous variable of pack years smoked over the lifetime at baseline
- Physical activity (PA) is total moderate of vigorous physical activity in Met-Min / week
- Caloric intake (energy) is calories / day
- eGFR is eGFR at baseline
- Systolic blood pressure was taken seated at baseline
- HDL is HDL-C in mg/dL at baseline
- Total cholesterol is is mg/dL at baseline
- Use of hypertension medication is a factor variable; coded 0 = No, 1 = yes
- Use of cholesterol lowering medication is a factor variable for use of any lipid-lowerting medication; coded 0 = No, 1 = yes

Note: Visceral fat was not available

Sample description

- There are N=5947 individuals with Olink protein data at exam 1.
- Of these, N=28 individuals did not have CVD data, leaving a sample of N=5919
- Of these, N=254 individuals did not have diet data, and a further N=464 had diet data outside the acceptable range (800-8000 kcals / day for men, 600-6000 kcals / day for women) leaving a sample of N=5201.
- Sample descriptives are available in Table 2

Table 2: Sample Descriptives

Characteristic	Exam N = 5,201[1]
Age (y)	62.20 (10.34)
Gender	
Female	2,745 / 5,201 (53%)
Male	2,456 / 5,201 (47%)
Race or ethnicity	
Non-Hispanic White	2,133 / 5,201 (41%)
Chinese American	588 / 5,201 (11%)
Black/African-American	1,277 / 5,201 (25%)
Hispanic	1,203 / 5,201 (23%)
BMI (kg/m^2^)	28.27 (5.39)
Physical activity (MET-min/week)	5,791.54 (5,989.79)
Caloric intake (kcals/day)	1,709.01 (797.38)
Kidney function (egfr)	80.93 (18.73)
Systolic blood pressure (mmHg)	126.31 (21.41)
HDL-cholesterol (mg/dL)	50.97 (14.89)
Smoking history (lifetime pack years)	11.15 (20.58)
Diabetes status	
Normoglycemia/IFG	4,551 / 5,196 (88%)
Diabetes (treated or untreated)	645 / 5,196 (12%)
Total cholesterol (mg/dL)	194.42 (35.66)
Takes hypertentsion medicine	
No	3,272 / 5,200 (63%)
Yes	1,928 / 5,200 (37%)
Takes lipid-lowering medicine	
No	4,341 / 5,192 (84%)
Yes	851 / 5,192 (16%)
Experienced a hard CVD event	

No	4,486 / 5,201 (86%)
Yes	715 / 5,201 (14%)
Mean follow-up time (days)	5,457.60 (2,156.42)
Mean follow-up time (years)	14.95 (5.91)

[1]Mean (SD); n / N (%)