MESA-MIND-Longitudinal-Proteomics-cSVD

LekkiWood (LekkiWood@Gmail.com)

2025-12-10

Table of contents

Summary

Step 1: Cleaning and Formatting Proteins

Input file names

Raw metabolite tables

- 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).

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
- This file included data on the following numbers of participants, stratified by exam:

Exam	N_Pps
1	5949
5	3917
6	2873

Step 2: Format Phenotypes

Input files

- Covariates from E1: MESAe1FinalLabel02092016.dta
- Covariates from E5: MESAe5_FinalLabel_20140613.dta
- Covariates from E6: MESAe6_FinalLabel_20220513.dta
- Afib info: SHARe_MesaEventsThruYear2020_AF_DS.txt
- ApoE info: MESA_ApoE_03102014.sas7bdat
- Incident CVD: MESAEvThru2020AllCohort_20241120.dta
- Microbleeds: MESAe6as253as301_BMRICMB_08052025.csv
- Perivascular spaces: MESAe6as253as301_BMRIPVS_20250310.csv
- White matter hyperintensities: MESAe6anyFIRST_BMRIWMHVol_20240422.csv
- Intracranial volumes: SHARe_AncilMesaAF_BMRIROIVol_DS.txt
- Fractional anisotropy: mesae6anyfirst_bmriTotalFAMUSE_20250828.csv
- White matter hyperintensities: MESAe6anyFIRST_BMRIWMHVol_20240422.csv

Formatting

Outcomes:

- Microbleeds were coded as 0/1, where 0= no microbleeds (value: 0) and 1 = presence of microbleeds (all non-zero values except missing). This yielded the following frequencies:

Var1	Freq
0	1908
1	1054

Then, those images with a low image quality (value = 4; N=50) were recoded to missing, yielding the following frequencies:

Var1	Freq
0	1908
1	1054

- Perivascular spaces (variable: epvs_wholebrain_vol) were recoded to missing where the variable ‘pvs_exclude’ was coded as 1 (N=39).

- White matter hyperintensities (variable: wm_wmh) were divided by 1000 to convert to ml (following Rizwan's code), and those where the variable wmh_exclude had a code of 1 (N=64) were set to missing.
- Fractional anisotropy (variable: wmfa) was coded to missing where the variable fa_exclude had a value of 1 (N=152).

Covariates

True time invariant covariates

- Race/ethnicity, gender, and highest education level were all taken from exam 1 data. ApoE information was taken from its own dataset (above).
    - ApoE was coded 0/1/2 where 0= no e4 isoform (codes 22, 23, 33), 1 = e4 isoform (24, 34, 44), and 2 = no isoform data. The ApoE variable was formatted as a factor.
    - Gender was coded such that female = 0 and male = 1.
    - Education was recoded 0/1, such that 0 = less than high school (codes: 0: NO SCHOOLING / 1: GRADES 1-8 / 2: GRADES: 9-11) and 1= high school or more (all other codes, excluding missing).
    - Race/ethnicity was recoded retaining the original MESA coded whereby 1=White American; 2= Chinese American, 3=Black, African-American, and 4 = Hispanic. Race/ethnicity was coded as a factor variable.

Pseudo-time invariant covariates

- Although some variables are technically time invariant, where they were included due to their effects on MRI data, since MRI data are only measured at one exam for this analysis (exam 6), these covariates were always taken from exam 6.
- These 'pseudo time invariant covariates' were: atrial fibrillation, myocardial infarction, congestive heart failure, LDL, systolic blood pressure, hypertension medication, and site (since site seems to affect MRI more than proteins??).
    - Afib, MI, and CHF were coded 0/1, such that 0= no diagnosis and 1= afib diagnosis. Missing data was left as missing (this is different to Rizwan who coded missing data as no diagnosis).
    - Site was arbitrarily coded as 0=Wake forest, 1= Columbia, 2=Johns Hopkins, 3=University of Minnesota, 4=Northwestern, 5=UCLA

Time varying covariates

- The following covariates were taken from the exam when the proteins were used, as these were seen to affect proteins more in the short term than they affect MRI (?): kidney function (egfr), BMI , cigarette smoking (never/former/current; coded as ordinal), diabetes status.
  - Smoking was harmonized and coded such that 0 = never smoker, 1= past smoker, 2=current smoker.
  - Diabetes has harmonized and coded such that 0= no diabetes (including impaired fasting glucose), and 1 = diabetes (treated and untreated).
- The following covariates were taken from exam 6: age (age6c), kidney function (egfr; cepgfr6c), BMI (bmi6c), systolic blood pressure (sbp6c), LDL (ldl6), site(site6c), the use of hypertension medication (htnmed6c; coded 0= no, and 1= yes), cigarette smoking (cig6c), diabetes status (dm036t)

Step 3: Sample descriptives

- There were N=6036 MESA participants with protein information, equating to N=5948 at exam 1, N=3917 at exam 5, and N=2872 at exam 6.
- Of those with protein data, N=1307 participants had at least one MRI outcome included in this analysis (after QC above), equating to N=1272 at exam 1, N=1256 at exam 5, and N=1270 at exam 6.
- Of those with protein data and at least one MRI outcome, N=1242 participants had no missing covariate data and so were included in this analysis, equating to N=1238 at exam 1, N=1212 at exam 5, and N=1240 at exam 6.

Table 1: Sample Descriptives

Characteristic	**N**	**1** N = 1,238	**5** N = 1,212	**6** N = 1,240	**p-value**
___Age (y)___	3,690	55 (50, 63)	65 (60, 72)	71 (66, 78)	<0.001
___Gender___	3,690	NA	NA	NA	>0.9
Female	NA	648 (52%)	635 (52%)	648 (52%)	NA
Male	NA	590 (48%)	577 (48%)	592 (48%)	NA
___Field Center___	3,690	NA	NA	NA	>0.9
Wake Forest	NA	216 (17%)	212 (17%)	216 (17%)	NA
Columbia	NA	200 (16%)	196 (16%)	201 (16%)	NA
Johns Hopkins	NA	159 (13%)	154 (13%)	159 (13%)	NA
Minnesota	NA	236 (19%)	227 (19%)	237 (19%)	NA
Northwestern	NA	220 (18%)	219 (18%)	220 (18%)	NA
UCLA	NA	207 (17%)	204 (17%)	207 (17%)	NA
___Highest education level___	3,690	NA	NA	NA	>0.9
Up to and including high school	NA	129 (10%)	126 (10%)	129 (10%)	NA
More than high school	NA	1,109 (90%)	1,086 (90%)	1,111 (90%)	NA
___Race or ethnicity___	3,690	NA	NA	NA	>0.9
Non-Hispanic White	NA	524 (42%)	516 (43%)	523 (42%)	NA
Chinese American	NA	150 (12%)	151 (12%)	151 (12%)	NA
Black/African-American	NA	326 (26%)	315 (26%)	327 (26%)	NA
Hispanic	NA	238 (19%)	230 (19%)	239 (19%)	NA
___BMI (kg/m^2^)___	3,690	27.3 (24.2, 30.6)	27.7 (24.6, 31.3)	27.7 (24.7, 31.3)	0.061
___Smoking status___	3,690	NA	NA	NA	<0.001
Never	NA	670 (54%)	584 (48%)	600 (48%)	NA
Former	NA	439 (35%)	545 (45%)	571 (46%)	NA
Current	NA	129 (10%)	83 (6.8%)	69 (5.6%)	NA
___LDL levels___	3,690	104 (82, 130)	104 (82, 130)	104 (82, 130)	>0.9
___systolic blood pressure___	3,690	124 (112, 139)	124 (112, 139)	124 (112, 139)	>0.9
___Diabetes status___	3,690	NA	NA	NA	<0.001
Normoglycemia/IFG	NA	1,162 (94%)	1,025 (85%)	980 (79%)	NA
Diabetes (treated or untreated)	NA	76 (6.1%)	187 (15%)	260 (21%)	NA

Table 1: Sample Descriptives *(continued)*

Characteristic	**N**	**1** N = 1,238	**5** N = 1,212	**6** N = 1,240	**p-value**
___Atrial fibrillation___	3,690	NA	NA	NA	>0.9
No	NA	1,058 (85%)	1,033 (85%)	1,059 (85%)	NA
Yes	NA	180 (15%)	179 (15%)	181 (15%)	NA
___Myocardial Infarction___	3,690	NA	NA	NA	>0.9
No	NA	1,202 (97%)	1,177 (97%)	1,204 (97%)	NA
Yes	NA	36 (2.9%)	35 (2.9%)	36 (2.9%)	NA
___Coronary Heart Failure___	3,690	NA	NA	NA	>0.9
No	NA	1,217 (98%)	1,191 (98%)	1,219 (98%)	NA
Yes	NA	21 (1.7%)	21 (1.7%)	21 (1.7%)	NA
___ApoeE information___	3,690	NA	NA	NA	>0.9
No E4 isoform	NA	883 (71%)	867 (72%)	886 (71%)	NA
E4 isoform	NA	335 (27%)	325 (27%)	334 (27%)	NA
No ApoE data	NA	20 (1.6%)	20 (1.7%)	20 (1.6%)	NA
___Kidney function (egfr)___	3,690	81 (71, 92)	85 (72, 95)	78 (65, 89)	<0.001
___fractional anisotropy___	3,538	0.398 (0.378, 0.415)	0.398 (0.379, 0.415)	0.398 (0.378, 0.415)	>0.9
___white matter hyperintensities___	3,658	3 (1, 7)	3 (1, 7)	3 (1, 7)	>0.9
___Enlarged perivascular spaces___	3,120	3,073 (2,071, 4,665)	3,046 (2,060, 4,663)	3,073 (2,071, 4,666)	>0.9
___Presence of microbleeds?___	2,962	NA	NA	NA	>0.9
No	NA	639 (64%)	630 (64%)	639 (64%)	NA
Yes	NA	352 (36%)	348 (36%)	354 (36%)	NA

add icv to table

Notes

##Footnotes

[1] This is a reproducible file for many runs, containing many data checks. Values of 0 or NULL are expected, and just indicate no problem with the data.

Session Info

For reproducibility

```
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.5.2 (2025-10-31)
 os       Linux Mint 21
 system   x86_64, linux-gnu
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       America/Chicago
 date     2025-12-10
 pandoc   3.2 @ /usr/lib/rstudio-server/bin/quarto/bin/tools/x86_64/ (via rmarkdown)
 quarto   1.8.25 @ /usr/local/bin/quarto

- Packages -------------------------------------------------------------------
 package     * version date (UTC) lib source
 backports     1.5.0   2024-05-23 [1] CRAN (R 4.5.0)
 base64url     1.4     2018-05-14 [1] CRAN (R 4.5.1)
 broom         1.0.8   2025-03-28 [1] CRAN (R 4.5.0)
 callr         3.7.6   2024-03-25 [1] CRAN (R 4.5.1)
 cards         0.7.1   2025-12-02 [1] CRAN (R 4.5.2)
 cardx         0.3.1   2025-12-04 [1] CRAN (R 4.5.2)
 cli           3.6.5   2025-04-23 [1] CRAN (R 4.5.2)
 codetools     0.2-20  2024-03-31 [4] CRAN (R 4.5.0)
 colorspace    2.1-1   2024-07-26 [1] CRAN (R 4.5.0)
 data.table    1.17.8  2025-07-10 [1] CRAN (R 4.5.1)
 digest        0.6.37  2024-08-19 [1] CRAN (R 4.5.1)
 dplyr         1.1.4   2023-11-17 [1] CRAN (R 4.5.0)
 evaluate      1.0.5   2025-08-27 [1] CRAN (R 4.5.1)
 fastmap       1.2.0   2024-05-15 [1] CRAN (R 4.5.0)
 fs            1.6.6   2025-04-12 [1] CRAN (R 4.5.0)
```

```
generics      0.1.4   2025-05-09 [1] CRAN (R 4.5.1)
glue          1.8.0   2024-09-30 [1] CRAN (R 4.5.0)
gt            1.1.0   2025-09-23 [1] CRAN (R 4.5.2)
gtsummary     2.5.0   2025-12-05 [1] CRAN (R 4.5.2)
htmltools     0.5.8.1 2024-04-04 [1] CRAN (R 4.5.0)
igraph        2.1.4   2025-01-23 [1] CRAN (R 4.5.0)
jsonlite      2.0.0   2025-03-27 [1] CRAN (R 4.5.0)
kableExtra    1.4.0   2024-01-24 [1] CRAN (R 4.5.2)
knitr         1.50    2025-03-16 [1] CRAN (R 4.5.2)
later         1.4.2   2025-04-08 [1] CRAN (R 4.5.0)
lifecycle     1.0.4   2023-11-07 [1] CRAN (R 4.5.0)
magrittr      2.0.4   2025-09-12 [1] CRAN (R 4.5.1)
munsell       0.5.1   2024-04-01 [1] CRAN (R 4.5.0)
pillar        1.11.1  2025-09-17 [1] CRAN (R 4.5.1)
pkgconfig     2.0.3   2019-09-22 [1] CRAN (R 4.5.0)
prettyunits   1.2.0   2023-09-24 [1] CRAN (R 4.5.0)
processx      3.8.6   2025-02-21 [2] CRAN (R 4.5.1)
ps            1.9.1   2025-04-12 [1] CRAN (R 4.5.0)
purrr         1.1.0   2025-07-10 [1] CRAN (R 4.5.1)
quarto        1.5.1   2025-09-04 [1] CRAN (R 4.5.2)
R6            2.6.1   2025-02-15 [1] CRAN (R 4.5.0)
Rcpp          1.0.14  2025-01-12 [1] CRAN (R 4.5.0)
rlang         1.1.6   2025-04-11 [1] CRAN (R 4.5.0)
rmarkdown     2.29    2024-11-04 [1] CRAN (R 4.5.0)
rstudioapi    0.17.1  2024-10-22 [1] CRAN (R 4.5.0)
scales        1.3.0   2023-11-28 [1] CRAN (R 4.5.0)
secretbase    1.0.5   2025-03-04 [1] CRAN (R 4.5.1)
sessioninfo   1.2.3   2025-02-05 [1] CRAN (R 4.5.1)
stringi       1.8.7   2025-03-27 [1] CRAN (R 4.5.0)
stringr       1.5.1   2023-11-14 [1] CRAN (R 4.5.0)
svglite       2.2.2   2025-10-21 [2] CRAN (R 4.5.1)
systemfonts   1.3.1   2025-10-01 [1] CRAN (R 4.5.2)
targets       1.11.4  2025-09-13 [1] CRAN (R 4.5.1)
textshaping   1.0.4   2025-10-10 [2] CRAN (R 4.5.1)
tibble        3.3.0   2025-06-08 [1] CRAN (R 4.5.1)
tidyr         1.3.1   2024-01-24 [1] CRAN (R 4.5.0)
tidyselect    1.2.1   2024-03-11 [1] CRAN (R 4.5.0)
vctrs         0.6.5   2023-12-01 [1] CRAN (R 4.5.0)
viridisLite   0.4.2   2023-05-02 [1] CRAN (R 4.5.0)
withr         3.0.2   2024-10-28 [1] CRAN (R 4.5.0)
xfun          0.53    2025-08-19 [1] CRAN (R 4.5.1)
xml2          1.4.0   2025-08-20 [1] CRAN (R 4.5.1)
yaml          2.3.10  2024-07-26 [1] CRAN (R 4.5.0)
```

```
 [1] /home/awood/R/x86_64-pc-linux-gnu-library/4.5
 [2] /usr/local/lib/R/site-library
 [3] /usr/lib/R/site-library
 [4] /usr/lib/R/library

------------------------------------------------------------------------------
```