MESA-MIND-Longitudinal-Proteomics-cSVD

LekkiWood (LekkiWood@Gmail.com)

2025-12-12

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). Then, those images with a low image quality (value = 4; N=0) were recoded to missing.

Var1	Freq
0	1730
1	886

- Perivascular spaces (variable: epvs_wholebrain_vol) were recoded to missing where the variable ‘pvs_exclude’ was coded as 1 (N=0).
- 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 were set to missing.
- Fractional anisotropy (variable: wmfa) was coded to missing where the variable fa_exclude had a value of 1.

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=1429 MESA participants with at least one MRI outcome after the exclusions above.
- Of those with MRI data, N=1429 participants had protein data, equating to N=0 at exam 1, N=0 at exam 5, and N=1429 at exam 6.
- Of those with protein data and at least one MRI outcome, N=0 participants had no missing covariate data and so were included in this analysis, equating to N=0 at exam 1, N=0 at exam 5, and N=0 at exam 6.

	Exam		
Characteristic	**1** N = 938[1]	**5** N = 926[1]	**6** N = 940[1]
Age (y)	56.77 (8.22)	66.19 (8.09)	72.47 (8.04)
Gender			
Female	492 / 938 (52%)	486 / 926 (52%)	493 / 940 (52%)
Male	446 / 938 (48%)	440 / 926 (48%)	447 / 940 (48%)
Field Center			
Wake Forest	168 / 837 (20%)	166 / 827 (20%)	169 / 839 (20%)
Columbia	110 / 837 (13%)	107 / 827 (13%)	110 / 839 (13%)
Johns Hopkins	185 / 837 (22%)	183 / 827 (22%)	186 / 839 (22%)
Minnesota	170 / 837 (20%)	170 / 827 (21%)	170 / 839 (20%)
Northwestern	0 / 837 (0%)	0 / 827 (0%)	0 / 839 (0%)
UCLA	204 / 837 (24%)	201 / 827 (24%)	204 / 839 (24%)
Highest education level			
Up to and including high school	105 / 938 (11%)	103 / 926 (11%)	105 / 940 (11%)
More than high school	833 / 938 (89%)	823 / 926 (89%)	835 / 940 (89%)
Race or ethnicity			
Non-Hispanic White	380 / 938 (41%)	377 / 926 (41%)	380 / 940 (40%)
Chinese American	137 / 938 (15%)	138 / 926 (15%)	138 / 940 (15%)
Black/African-American	231 / 938 (25%)	224 / 926 (24%)	231 / 940 (25%)
Hispanic	190 / 938 (20%)	187 / 926 (20%)	191 / 940 (20%)
BMI (kg/m^2^)	27.62 (4.97)	28.08 (5.19)	28.02 (5.33)
Smoking status			
Never	0 / 840 (0%)	0 / 866 (0%)	0 / 884 (0%)
Former	496 / 840 (59%)	441 / 866 (51%)	448 / 884 (51%)
Current	344 / 840 (41%)	425 / 866 (49%)	436 / 884 (49%)
LDL levels	107.55 (35.85)	107.40 (35.35)	107.50 (35.83)
systolic blood pressure	126.69 (20.49)	126.76 (20.49)	126.76 (20.53)
Diabetes status			
Normoglycemia/IFG	875 / 938 (93%)	780 / 926 (84%)	736 / 940 (78%)

Diabetes (treated or untreated)	63 / 938 (6.7%)	146 / 926 (16%)	204 / 940 (22%)
Takes hypertentsion medicine			
No	391 / 938 (42%)	386 / 926 (42%)	392 / 940 (42%)
Yes	547 / 938 (58%)	540 / 926 (58%)	548 / 940 (58%)
Atrial fibrillation			
No	795 / 938 (85%)	783 / 926 (85%)	796 / 940 (85%)
Yes	143 / 938 (15%)	143 / 926 (15%)	144 / 940 (15%)
Myocardial Infarction			
No	911 / 938 (97%)	899 / 926 (97%)	913 / 940 (97%)
Yes	27 / 938 (2.9%)	27 / 926 (2.9%)	27 / 940 (2.9%)
Coronary Heart Failure			
No	919 / 938 (98%)	907 / 926 (98%)	921 / 940 (98%)
Yes	19 / 938 (2.0%)	19 / 926 (2.1%)	19 / 940 (2.0%)
ApoeE information			
No E4 isoform	669 / 938 (71%)	663 / 926 (72%)	671 / 940 (71%)
E4 isoform	259 / 938 (28%)	253 / 926 (27%)	259 / 940 (28%)
No ApoE data	10 / 938 (1.1%)	10 / 926 (1.1%)	10 / 940 (1.1%)
Kidney function (egfr)	82.44 (15.28)	82.29 (19.57)	76.72 (19.60)
Intracranial volume	1,360,159.33 (145,347.03)	1,360,711.30 (144,918.81)	1,360,215.07 (145,244.17)
fractional anisotropy	0.39 (0.03)	0.39 (0.03)	0.39 (0.03)
white matter hyperintensities	6.79 (10.39)	6.70 (10.24)	6.80 (10.39)
Enlarged perivascular spaces	3,574.55 (2,228.27)	3,564.42 (2,230.86)	3,573.19 (2,226.09)
Presence of microbleeds?			
No	579 / 875 (66%)	572 / 864 (66%)	579 / 877 (66%)
Yes	296 / 875 (34%)	292 / 864 (34%)	298 / 877 (34%)

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

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     2025-12-12
 pandoc   3.2 @ /usr/lib/rstudio-server/bin/quarto/bin/tools/x86_64/ (via rmarkdown)
 quarto   1.8.26 @ /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)
 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)
 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)
 commonmark     1.9.5   2025-03-17 [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)
 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)
 litedown      0.7     2025-04-08 [2] CRAN (R 4.5.1)
 magrittr      2.0.4   2025-09-12 [1] CRAN (R 4.5.1)
 markdown      2.0     2025-03-23 [2] CRAN (R 4.5.1)
 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)
 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)
 targets       1.11.4  2025-09-13 [1] 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)
 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

──────────────────────────────────────────────────────────────────────────────
```