MESA-MIND-Longitudinal-Proteomics-cSVD

LekkiWood (LekkiWood@Gmail.com)

2025-12-18

Table of contents

Important Notes - read me first

- 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: 2025-12-18. 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)

Summary

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: 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	1974
1	1107

- 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)

Missingness

The missingness for each variable those who have at least one MRI measure can be found in Table 2

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=1307 participants had protein data, equating to N=1272 at exam 1, N=1256 at exam 5, and N=1270 at exam 6

- Sample descriptives are available in Table 3

Table 2: Missingness for those who have at least one MRI measure

Variable	Frequency (N)	Proportion (%)
icv	307	23.9
mb_present	252	19.6
epvs	197	15.3
fa	52	4.05
ldl	34	2.65
egfr	26	2.02
diabetes	25	1.95
wmh	19	1.48
htnmeds	11	0.857
E4	10	0.779
AFprevalent	4	0.312
edu	2	0.156
smoking	2	0.156
sbp	2	0.156
BMI	1	0.0779
MIprevalent	1	0.0779
CHFprevalent	1	0.0779
idno	0	0
sidno	0	0
Exam	0	0
time	0	0
BL_age	0	0
age	0	0
gender	0	0
site	0	0
race	0	0

Table 3: Sample Descriptives

Characteristic	Exam		
	1 N = 1,307[1]	**5** N = 1,281[1]	**6** N = 1,284[1]
Age (y)	56.40 (7.96)	65.84 (7.83)	72.20 (7.79)
Gender			
Female	689 / 1,307 (53%)	676 / 1,281 (53%)	676 / 1,284 (53%)
Male	618 / 1,307 (47%)	605 / 1,281 (47%)	608 / 1,284 (47%)
Field Center			
Wake Forest	204 / 1,045 (20%)	199 / 1,024 (19%)	204 / 1,045 (20%)
Columbia	165 / 1,045 (16%)	161 / 1,024 (16%)	165 / 1,045 (16%)
Johns Hopkins	239 / 1,045 (23%)	232 / 1,024 (23%)	239 / 1,045 (23%)
Minnesota	228 / 1,045 (22%)	227 / 1,024 (22%)	228 / 1,045 (22%)
Northwestern	0 / 1,045 (0%)	0 / 1,024 (0%)	0 / 1,045 (0%)
UCLA	209 / 1,045 (20%)	205 / 1,024 (20%)	209 / 1,045 (20%)
Highest education level			
Up to and including high school	133 / 1,305 (10%)	129 / 1,279 (10%)	130 / 1,282 (10%)
More than high school	1,172 / 1,305 (90%)	1,150 / 1,279 (90%)	1,152 / 1,282 (90%)
Race or ethnicity			
Non-Hispanic White	553 / 1,307 (42%)	545 / 1,281 (43%)	542 / 1,284 (42%)
Chinese American	157 / 1,307 (12%)	156 / 1,281 (12%)	157 / 1,284 (12%)
Black/African-American	350 / 1,307 (27%)	340 / 1,281 (27%)	345 / 1,284 (27%)
Hispanic	247 / 1,307 (19%)	240 / 1,281 (19%)	240 / 1,284 (19%)
BMI (kg/m^2^)	27.95 (5.21)	28.41 (5.36)	28.35 (5.44)
Smoking status			
Never	0 / 1,162 (0%)	0 / 1,183 (0%)	0 / 1,207 (0%)
Former	699 / 1,162 (60%)	609 / 1,183 (51%)	615 / 1,207 (51%)
Current	463 / 1,162 (40%)	574 / 1,183 (49%)	592 / 1,207 (49%)
LDL levels	107.96 (35.28)	107.96 (34.98)	107.96 (35.28)
systolic blood pressure	126.34 (20.35)	126.37 (20.37)	126.34 (20.35)
Diabetes status			

Normoglycemia/IFG	1,219 / 1,303 (94%)	1,076 / 1,280 (84%)	994 / 1,259 (79%)
Diabetes (treated or untreated)	84 / 1,303 (6.4%)	204 / 1,280 (16%)	265 / 1,259 (21%)
Takes hypertentsion medicine			
No	537 / 1,273 (42%)	527 / 1,248 (42%)	537 / 1,273 (42%)
Yes	736 / 1,273 (58%)	721 / 1,248 (58%)	736 / 1,273 (58%)
Atrial fibrillation			
No	1,095 / 1,280 (86%)	1,073 / 1,255 (85%)	1,095 / 1,280 (86%)
Yes	185 / 1,280 (14%)	182 / 1,255 (15%)	185 / 1,280 (14%)
Myocardial Infarction			
No	1,245 / 1,283 (97%)	1,220 / 1,258 (97%)	1,245 / 1,283 (97%)
Yes	38 / 1,283 (3.0%)	38 / 1,258 (3.0%)	38 / 1,283 (3.0%)
Coronary Heart Failure			
No	1,261 / 1,283 (98%)	1,236 / 1,258 (98%)	1,261 / 1,283 (98%)
Yes	22 / 1,283 (1.7%)	22 / 1,258 (1.7%)	22 / 1,283 (1.7%)
ApoeE information			
No E4 isoform	916 / 1,274 (72%)	899 / 1,249 (72%)	916 / 1,274 (72%)
E4 isoform	348 / 1,274 (27%)	340 / 1,249 (27%)	348 / 1,274 (27%)
No ApoE data	10 / 1,274 (0.8%)	10 / 1,249 (0.8%)	10 / 1,274 (0.8%)
Kidney function (egfr)	82.72 (15.47)	82.46 (19.44)	76.86 (19.53)
Intracranial volume	1,360,304.37 (145,625.45)	1,361,045.01 (145,456.35)	1,360,304.37 (145,625.45)
Fractional anisotropy	0.40 (0.03)	0.40 (0.03)	0.40 (0.03)
White matter hyperintensities	6.74 (10.15)	6.62 (9.89)	6.71 (10.16)
Enlarged perivascular spaces	3,701.67 (2,374.34)	3,702.34 (2,383.48)	3,701.67 (2,374.34)
Presence of microbleeds?			
No	661 / 1,032 (64%)	652 / 1,017 (64%)	661 / 1,032 (64%)
Yes	371 / 1,032 (36%)	365 / 1,017 (36%)	371 / 1,032 (36%)

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

Step 5: Cross sectional Protein-Wide Association Studies (PWAS)

- This section includes PWAS where proteins were only used at one time point (exam 1 or exam 6), even if this was not the same time point as when the MRI data were used.

- All estimates are standardized.

White matter hyperintensity (WMH)

Model Specification

- The associations between WMH and proteins were analyzed via linear regression
- For WMH the numeric covariates included icv, age, egfr, BMI, sbp, ldl. [2]
- For WMH the factor covariates included gender, race, site, edu, htnmeds, smoking, E4, AFprevalent, diabetes, MIprevalent, CHFprevalent. [2]
- WMH was transformed using an inverse normal transformation (with blom constant).

WMH: Exam 6 proteins

- The E6 proteins -> E6 MRI for WMH included data from N=933 participants.

- A total of N=0 exam 6 proteins were significantly associated with WMH at an FDR corrected P<.05.

WMH: Exam 1 proteins

- The E1 proteins -> E6 MRI for WMH included data from N=913 participants.

- A total of N=1 exam 1 proteins were significantly associated with WMH at an FDR corrected P<.05 (see Table 4 and Figure 1)

Enlarged Perivascular Spaces (EPVS)

Model Specification

- The associations between EPVS and proteins were analyzed via linear regression
- For EPVS the numeric covariates included icv, age, egfr, BMI, sbp, ldl. [2]
- For EPVS the factor covariates included gender, race, site, edu, htnmeds, smoking, E4, AFprevalent, diabetes, MIprevalent, CHFprevalent. [2]
- EPVS was transformed using an inverse normal transformation (with blom constant).

EPVS: Exam 6 proteins

- The E6 proteins -> E6 MRI for EPVS included data from N=926 participants.

- A total of N=0 exam 1 proteins were significantly associated with EPVS at an FDR corrected P<.05

EPVS: Exam 1 proteins

- The E1 proteins -> E6 MRI for EPVS included data from N=906 participants.

- A total of N=0 exam 1 proteins were significantly associated with EPVS at an FDR corrected P<.05

Fractional Anisotropy (FA)

Model Specification

- The associations between FA and proteins were analyzed via linear regression
- For FA the numeric covariates included age, egfr, BMI, sbp, ldl. [2]
- For FA the factor covariates included gender, race, site, edu, htnmeds, smoking, E4, AFprevalent, diabetes, MIprevalent, CHFprevalent. [2]
- FA was transformed using an inverse normal transformation (with blom constant).

FA: Exam 6 proteins

- The E6 proteins -> E6 MRI for FA included data from N=1170 participants.

- A total of N=2 exam 6 proteins were significantly associated with FA at an FDR corrected P<.05 (see Table 5 and Figure 2).

FA: Exam 1 proteins

- The E1 proteins -> E6 MRI for FA included data from N=1146 participants.

- A total of N=391 exam 1 proteins were significantly associated with FA at an FDR corrected P<.05.

Microbleeds (MB)

Model Specification

- The associations between MB and proteins were analyzed via logistic regression (MB coded as 'present' vs. 'absent' as above)
- For MB the numeric covariates included age, egfr, BMI, sbp, ldl. [2]
- For MB the factor covariates included gender, race, site, edu, htnmeds, smoking, E4, AFprevalent, diabetes, MIprevalent, CHFprevalent. [2]

MB: Exam 6 proteins

- The E6 proteins -> E6 MRI for MB included data from N= 975 participants.

- A total of N=0 exam 6 proteins were significantly associated with MB at an FDR corrected P<.05.

MB: Exam 1 proteins

- The E1 proteins -> E6 MRI for FA included data from N=949 participants.

- A total of N=0 exam 1 proteins were significantly associated with MB at an FDR correct P<.05.

Table 4: Associations Between Exam 1 Proteins and White Matter Hyperintensity Reaching an FDR-corrected P<.05 Equivalent

Protein	N	Beta	SE	P	P-FDR	Olink ID	Protein CV	UniProt ID
MYL3	913	0.14	0.03	1.60×10^{-5}	4.69×10^{-2}	OID30251	22.36	P08590

Table 5: Associations Between Exam 6 Proteins and Fractional Anisotropy Reaching an FDR-corrected P<.05 Equivalent

Protein	N	Beta	SE	P	P-FDR	Olink ID	Protein CV	UniProt ID
OMG	1170	0.12	0.03	7.30×10^{-6}	2.13×10^{-2}	OID21352	71.71	P23515
HAVCR1	1170	-0.12	0.03	2.06×10^{-5}	3.01×10^{-2}	OID21422	95.89	Q96D42

Models control for icv,age,egfr,BMI,sbp,ldl
gender,race,site,edu,htnmeds,smoking,E4,AFprevalent,diabetes,MIprevalent,CHFprevalen
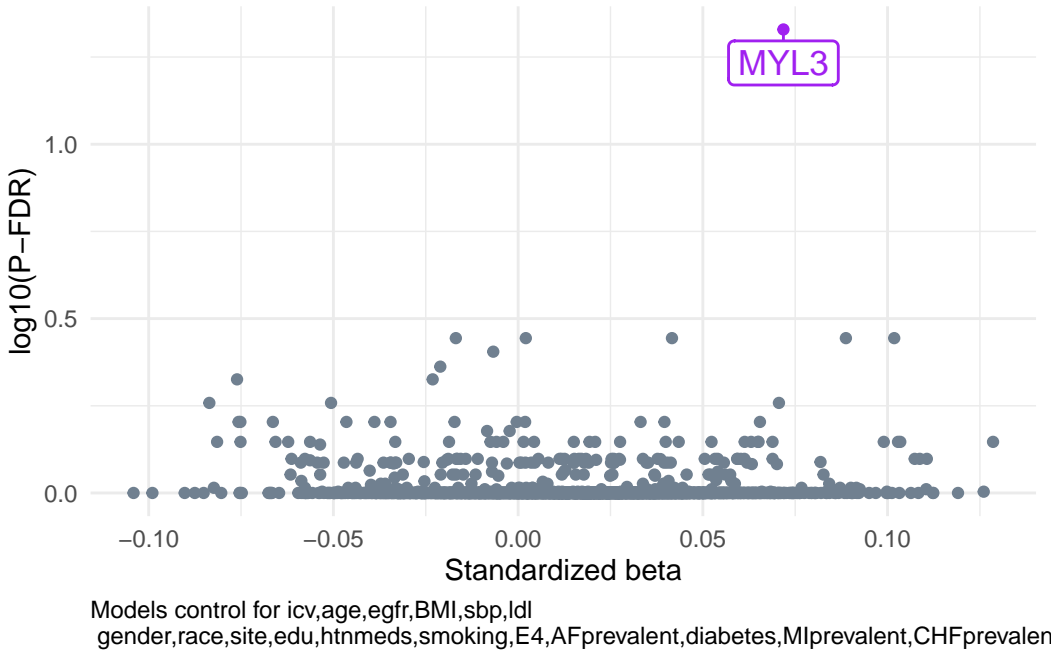
Figure 1: Associations Between Exam 1 Proteins and White Matter Hyperintensity
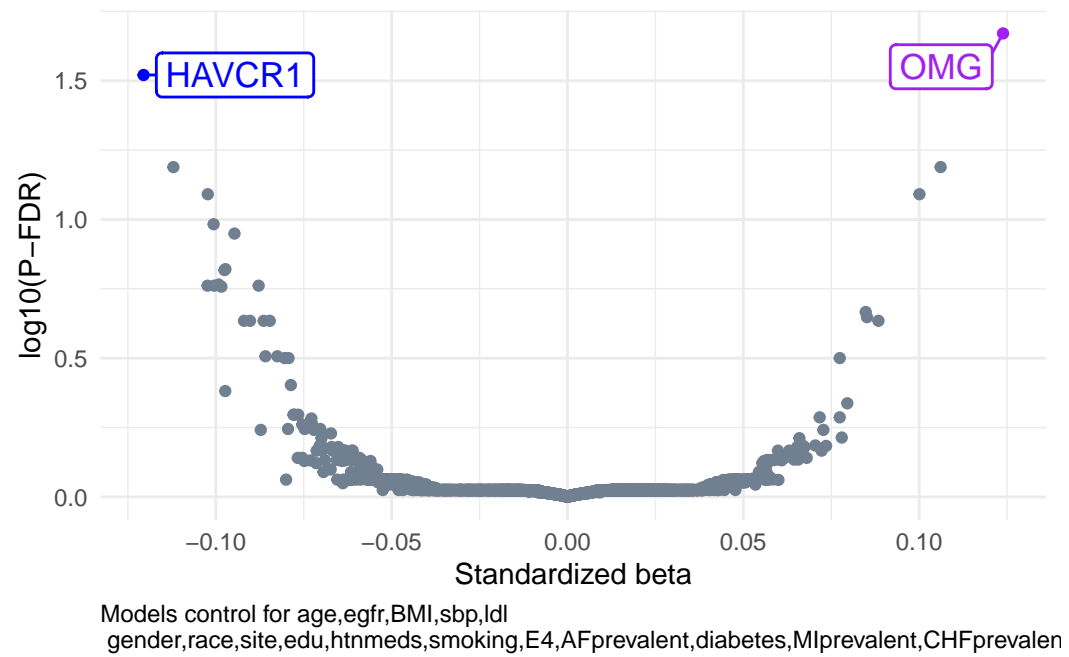
Figure 2: Associations Between Exam 6 Proteins and Fractional Anisotropy

\newpage

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.

[2] These are intentionally written as variable names to avoid reproducibility errors.

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     2025-12-18
 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)
 colorspace     2.1-1   2024-07-26 [1] 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)
farver         2.1.2    2024-05-13 [1] CRAN (R 4.5.0)
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)
ggplot2        3.5.2    2025-04-09 [1] CRAN (R 4.5.0)
ggrepel        0.9.6    2024-09-07 [1] CRAN (R 4.5.0)
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)
gtable         0.3.6    2024-10-25 [1] CRAN (R 4.5.0)
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)
labeling       0.4.3    2023-08-29 [1] CRAN (R 4.5.0)
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)
munsell        0.5.1    2024-04-01 [1] CRAN (R 4.5.0)
naniar         1.1.0    2024-03-05 [1] CRAN (R 4.5.2)
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)
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)
visdat       0.6.0   2023-02-02 [1] CRAN (R 4.5.2)
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

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