Supplementary Material

A leakage-controlled stress-test framework for unsupervised cardiometabolic biomarker phenotyping in NHANES

Krishna Sai Pokala

Supplementary scope

This supplement provides the technical audit trail for the preprint reporting package. Raw NHANES source files are not redistributed. The repository and submission package provide source code, aggregate tables, figures, and manuscript materials generated from the final v2-plus results object.

S1. Leakage-control roles and variable domains

Domain	Primary clustering variables	Held-out post-hoc or sensitivity role
Age	Adult age	Cohort criterion and descriptive context
Adiposity	Body mass index; waist circumference	Obesity and central-obesity indicators
Blood pressure	Mean systolic and diastolic blood pressure	Hypertension indicator
Glycemia	HbA1c; glucose	Diabetes and prediabetes indicators
Lipids	Triglycerides; HDL cholesterol; LDL cholesterol	Dyslipidemia and metabolic-syndrome components
Renal	eGFR; urinary albumin-to-creatinine ratio	Albuminuria and CKD-risk indicators
Inflammation / hematologic	High-sensitivity C-reactive protein; white blood cell count	Descriptive inflammatory and hematologic context
Demographic, socioeconomic, and survey variables	Excluded from primary clustering	Sensitivity analysis and population-oriented description

S2. Final seed diagnostics

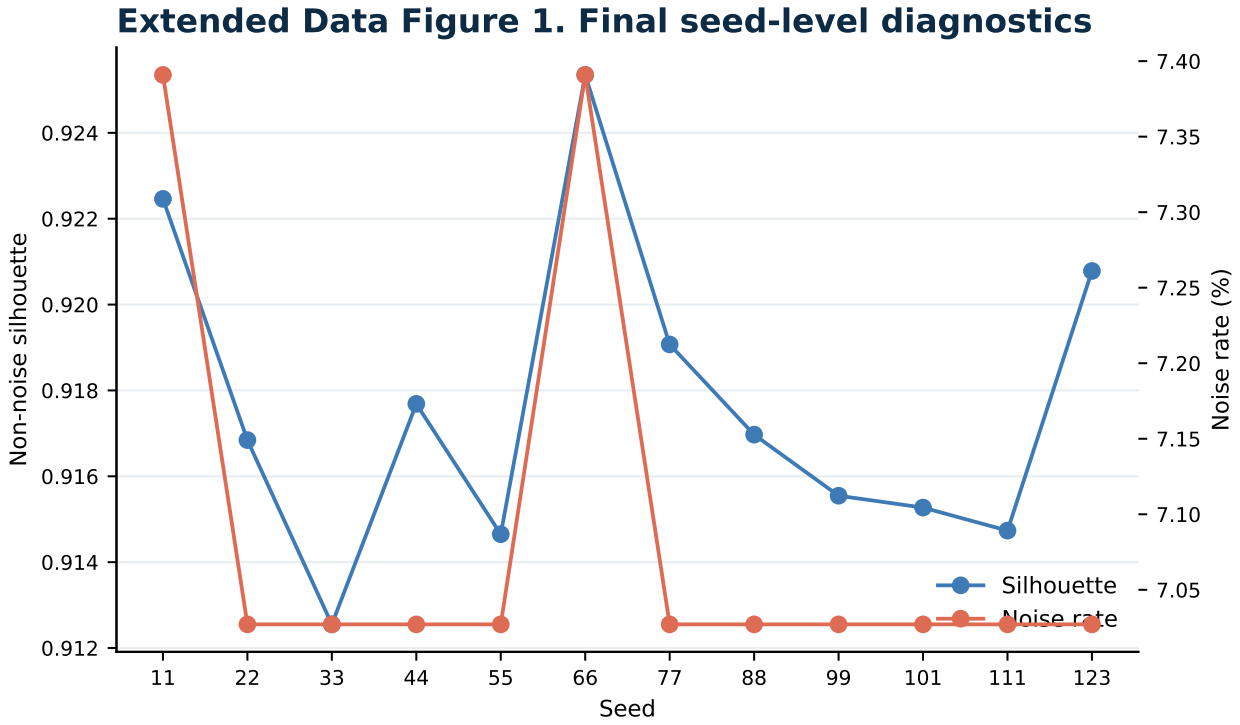

Figure 1: Final seed-level diagnostics across the 12 final seeds.

Seed	Clusters	Noise rate	Mean membership	Non-noise silhouette
11	3	0.0739	0.846	0.922
22	3	0.0703	0.849	0.917
33	3	0.0703	0.837	0.913
44	3	0.0703	0.852	0.918
55	3	0.0703	0.842	0.915
66	3	0.0739	0.857	0.925
77	3	0.0703	0.834	0.919
88	3	0.0703	0.854	0.917
99	3	0.0703	0.844	0.916
101	3	0.0703	0.844	0.915
111	3	0.0703	0.850	0.915
123	3	0.0703	0.843	0.921

S3. Triglycerides variable harmonization and missingness

Supplementary Table S3. Triglycerides harmonization and missingness in the discovery cycle.

Item	Value
Selected 2021–2023 triglycerides column	LBXTLG
Alias set checked	LBXTLG; LBXTR; LBXSTR
Missingness	2,969 participants (49.1%)
Primary-feature status	Included as an objective lipid feature; interpreted with caution

The triglycerides variable was included after alias resolution to LBXTLG. Because missingness was substantial, lipid-domain findings were interpreted with caution, and the limitation is stated explicitly in the manuscript and supplement. This table makes the harmonization decision and missingness level visible for reader and reviewer assessment.

S4. Feature-block ablation summary

Ablation	Mean ARI versus full objective reference
Numeric-only objective	0.995
Remove liver/inflammation	0.984
Remove renal	0.975
Remove glycemic	0.969
Remove hematologic	0.966
Remove age	0.966
Remove adiposity	0.935
Remove blood pressure	0.922
Add demographics	0.650
Remove lipids	0.263

S5. Temporal replication best matches

Discovery-to-replication match	Profile correlation	Euclidean distance
D1 lower-glycemic central-adiposity -> R1	0.909	0.208
D0 lipid-prediabetes -> R3	0.861	0.121
D2 higher BP/renal-risk -> R1	0.714	0.208

S6. Machine-readable tables

The submission package includes machine-readable CSV files for final phenotype groups, selected post-hoc enrichment signals, model/stability audit, phenotype profiles, full post-hoc enrichment, ablation summaries, and temporal replication matches. These tables are intended to support reviewer audit and downstream reproduction of aggregate reporting artifacts.