Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

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

Krishna Sai Pokala

ok50250@umbc.edu

University of Maryland, Baltimore County https://orcid.org/0009-0001-3012-4018

Research Article

Keywords: NHANES, cardiometabolic phenotyping, unsupervised learning, HDBSCAN, UMAP, clustering stability, leakage control, biomarker phenotypes, metabolic risk, reproducible research

Posted Date: June 3rd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9889895/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

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

An exploratory reproducibility study using public US survey data

Krishna Sai Pokala[1]
[1]Independent Researcher, Health Informatics, United States
Corresponding author: Krishna Sai Pokala, ok50250@umbc.edu
Repository: https://github.com/KrishnaSaiPokala/nhanes-hdbscan-phenotyping

Preprint version: this manuscript has not been certified by peer review. Findings should be interpreted as exploratory and descriptive.

Abstract

Background: Threshold-based cardiometabolic labels are indispensable for clinical communication, but they compress continuous multivariable physiology into a limited set of cut points. Unsupervised phenotyping can describe latent biomarker structure, but biomedical clustering requires rigorous leakage control, stability testing, robustness analysis, and replication before interpretation. **Objective:** To evaluate whether a leakage-controlled unsupervised workflow can reproducibly describe cardiometabolic biomarker-density tendencies among US adults using public NHANES data. **Methods:** The discovery analysis used 6,048 adults from NHANES August 2021–August 2023. Temporal replication used NHANES 2017–March 2020. The primary feature space contained objective examination and laboratory biomarkers spanning age, adiposity, blood pressure, glycemia, lipids, renal function, inflammation, and hematologic domains. Disease labels, self-reported diagnoses, demographic and socioeconomic variables, and survey weights were excluded from primary clustering and reserved for post-hoc characterization or sensitivity analyses. Biomarkers were preprocessed, represented using TruncatedSVD and UMAP, and clustered using HDBSCAN. Multi-seed stability, feature-block ablation, post-hoc enrichment, and temporal replication were evaluated. **Results:** The final discovery solution identified three non-noise biomarker-density phenotypes plus an algorithmic noise/outlier group. Mean pairwise adjusted Rand index was 0.9935, mean pairwise normalized mutual information was 0.9885, mean representation-space non-noise silhouette was 0.9168, and mean noise rate was 7.10%. The largest phenotype showed lipid-prediabetes metabolic-risk enrichment; a second phenotype showed lower-glycemic central adiposity; and a smaller phenotype showed higher blood-pressure/renal-risk tendency. Lipid feature-block removal was the most destabilizing ablation. Best temporal-replication profile correlations ranged from 0.714 to 0.909. **Conclusions:** A leakage-controlled SVD–UMAP–HDBSCAN workflow provided a reproducible descriptive atlas of exploratory cardiometabolic biomarker-density tendencies in NHANES adults. These findings are descriptive and hypothesis-generating; they are not diagnostic classes, causal subtypes, treatment recommendations, or a clinical decision tool.

Keywords: NHANES; cardiometabolic phenotyping; HDBSCAN; UMAP; unsupervised learning; clustering stability; biomedical informatics; computational epidemiology; reproducible research.

Key messages

- A leakage-controlled workflow using objective NHANES cardiometabolic biomarkers supported a reproducible density-based descriptive solution with three non-noise groups plus an algorithmic noise/outlier group.
- Disease labels, self-reported diagnoses, survey weights, and demographic or socioeconomic variables were

excluded from primary clustering, reducing circularity in post-hoc disease-burden interpretation.

- Multi-seed stability, feature-block ablation, post-hoc enrichment, and temporal replication were treated as central validation steps rather than optional checks.
- The findings are exploratory and descriptive rather than diagnostic, causal, or treatment-directing.

1 Introduction

Cardiometabolic risk is often organized through diagnostic and risk-factor categories such as obesity, metabolic syndrome, dyslipidemia, hypertension, diabetes, and chronic kidney disease risk. These categories are clinically useful because they translate complex biology into interpretable decision points. They are also incomplete. Individuals assigned to the same clinical category can differ substantially in adiposity, lipid burden, glycemic status, blood pressure, renal markers, inflammatory activity, and hematologic profile. Conversely, individuals who do not meet conventional diagnostic thresholds may still occupy adverse regions of multivariable cardiometabolic biomarker space.

Metabolic syndrome criteria illustrate this tension. Adult Treatment Panel III and later harmonized definitions formalized the co-occurrence of central adiposity, elevated triglycerides, reduced high-density lipoprotein cholesterol, elevated blood pressure, and impaired glucose regulation as a threshold-based risk construct[1, 2]. These definitions remain valuable for clinical and epidemiologic use, but they were designed to classify risk using interpretable cut points rather than to discover structure in continuous biomarker distributions. A related issue appears in obesity research: metabolically healthier and metabolically adverse obesity phenotypes show that adiposity and cardiometabolic burden do not always align along a single axis, while also cautioning against treating apparently lower-risk obesity as benign or durable[3, 4].

Data-driven phenotyping offers a complementary way to describe heterogeneity. In adult-onset diabetes, clustering studies have identified subgroups with differing clinical characteristics and complication profiles[5, 6]. That precedent is informative but also highlights a methodological hazard: unsupervised clusters are not biological subtypes simply because they are visually separable or clinically interpretable. Clusters can reflect preprocessing choices, missingness, scaling, dimensionality reduction, hyperparameters, or algorithm-specific geometry. In biomedical applications, a single clustering result is therefore insufficient. Stability, robustness, leakage control, interpretability, and replication must be treated as design requirements[7, 8].

NHANES provides an unusually useful setting for this problem because it combines public-use interview, examination, and laboratory data from a national US population sample[9, 10]. The August 2021–August 2023 cycle provides a recent discovery cohort, and NHANES 2017–March 2020 provides a temporally distinct pre-pandemic comparison period[11, 12]. This study asks whether a leakage-controlled unsupervised workflow can reproducibly describe density-defined cardiometabolic biomarker tendencies among US adults after explicit ablation and temporal replication checks. The intent is descriptive population phenotyping and workflow stress-testing, not diagnostic classification or clinical decision support.

2 Methods

2.1 Study design and data sources

This exploratory computational epidemiology study used public-use NHANES data. The discovery analysis used NHANES August 2021–August 2023. Temporal replication used NHANES 2017–March 2020 pre-pandemic data. Adults aged 20 years or older were eligible if they had sufficient objective cardiometabolic biomarker information for the aggregate phenotyping analysis. The final discovery cohort included 6,048 adults. Raw

NHANES source files were not redistributed; the accompanying repository and submission package contain code, aggregate derived tables, figures, and reporting artifacts.

2.2 Feature space and leakage-control design

The primary feature space included objective biomarkers spanning age, adiposity, blood pressure, glycemia, lipids, renal function, inflammation, and hematologic markers. Disease labels, self-reported diagnoses, demographic variables, socioeconomic variables, and survey weights were excluded from primary clustering and reserved for post-hoc interpretation or sensitivity analysis. Diabetes and prediabetes indicators were based on standard glycemic thresholds[16]; hypertension was summarized using contemporary blood-pressure thresholds[17]; chronic kidney disease risk and albuminuria indicators were based on eGFR and urinary albumin-to-creatinine ratio categories[18]; and eGFR used the 2021 CKD-EPI creatinine equation without a race coefficient[19]. Obesity, central obesity, dyslipidemia, and metabolic-syndrome component burden were treated as descriptive post-hoc constructs.

2.3 Preprocessing, representation learning, and clustering

Continuous biomarkers were numerically coerced, screened for implausible values, transformed where clinically and statistically appropriate, imputed within the preprocessing pipeline, and scaled before representation learning. The biomarker matrix was reduced using TruncatedSVD with 75 requested components, then embedded using UMAP with 10 components and 50 neighbors[13]. HDBSCAN was fit in the learned representation using minimum cluster size 300 and minimum samples 50[14, 15]. The algorithmic noise group was interpreted as observations not assigned to dense non-noise regions, not as a clinical diagnosis.

2.4 Model selection and validation analyses

The final model was selected using structural quality, multi-seed stability, interpretability, and robustness considerations. Disease-burden enrichment was not used for model selection. Multi-seed stability was evaluated by refitting the full pipeline across 12 final seeds and comparing participant-level labels using adjusted Rand index and normalized mutual information[20, 21]. Non-noise silhouette was used as a diagnostic in the learned representation[22]; it was not interpreted as evidence of clinical separability. Feature-block ablation removed one prespecified biomarker domain at a time and compared the perturbed solution with the full objective-biomarker reference using adjusted Rand index. Temporal replication compared discovery phenotype profiles with NHANES 2017–March 2020 replication profiles using profile correlation and Euclidean distance. Reporting was prepared with attention to transparent observational reporting and reproducible data-management principles[23, 24].

3 Results

3.1 Final phenotype solution and stability

The final discovery solution identified three non-noise HDBSCAN phenotypes plus an algorithmic noise/outlier group (Figure 2; Table 1). The largest phenotype comprised 2,998 adults (49.6%); the second comprised 2,258 adults (37.3%); the smallest non-noise phenotype comprised 345 adults (5.7%); and the algorithmic noise/outlier group included 447 adults (7.4%). Across the final seed set, the number of non-noise phenotypes was stable, the noise fraction remained narrow, and participant-level agreement was high. Mean pairwise adjusted Rand index

was 0.9935 and mean pairwise normalized mutual information was 0.9885. The mean non-noise silhouette in the learned representation was 0.9168, and the mean noise rate was 7.10%. These diagnostics support internal reproducibility of the selected representation-and-density solution, while leaving clinical interpretation to the post-hoc analyses.

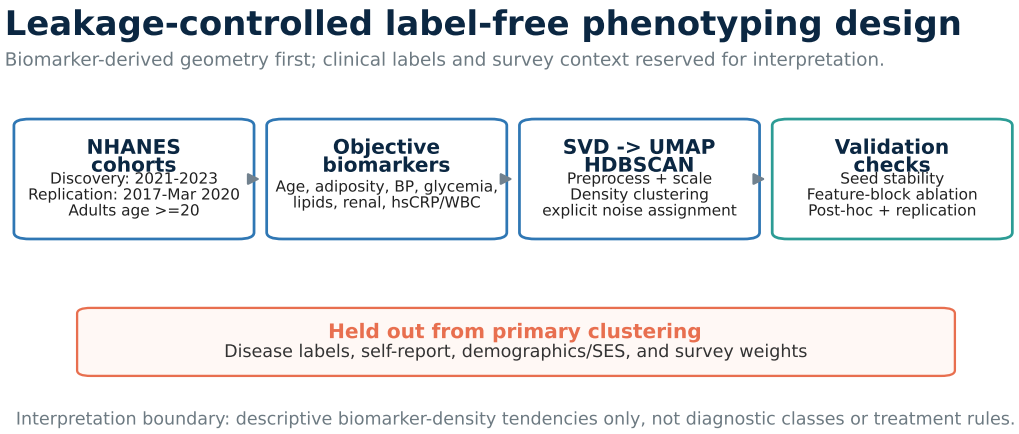


Figure 1: Study design and leakage-control architecture. Objective examination and laboratory biomarkers were used to define the unsupervised feature geometry. Disease labels, self-reported diagnoses, demographic and socioeconomic variables, and survey weights were held out from primary clustering and used only after the phenotype solution was frozen.

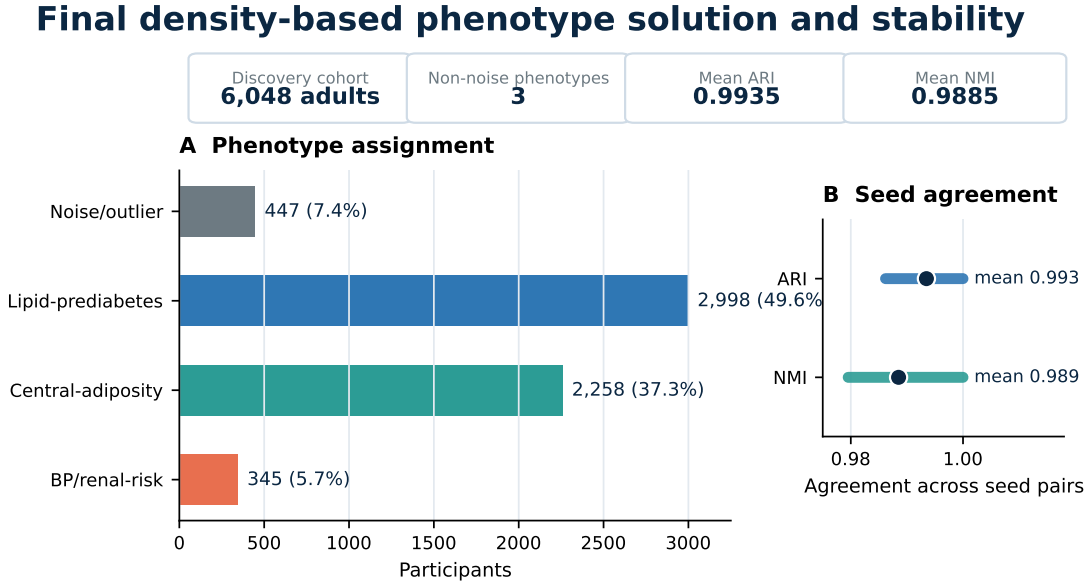


Figure 2: Final density-based phenotype solution and stability. Panel A shows final phenotype group sizes. Panel B summarizes pairwise agreement across final seed pairs; selected model diagnostics are shown in the header badges and noted in the text. Silhouette is a learned-representation diagnostic and should not be interpreted as clinical separability.

Table 1: Final phenotype groups in the discovery analysis.

Label	N	Percent	Descriptive interpretation	Interpretation boundary
-1	447	7.4%	Algorithmic noise/outlier group	Observations not assigned to dense non-noise regions
0	2,998	49.6%	Lipid-prediabetes metabolic-risk tendency	Descriptive enrichment pattern, not a diagnostic class
1	2,258	37.3%	Lower-glycemic central-adiposity tendency	Central adiposity with comparatively lower glycemic burden
2	345	5.7%	Higher blood-pressure/renal-risk tendency	Smaller dense region with BP/renal-risk signals

3.2 Biomarker phenotype atlas

The phenotype atlas showed coherent differences across cardiometabolic domains (Figure 3). The largest phenotype was interpreted as a lipid-prediabetes metabolic-risk tendency on the basis of its biomarker profile and post-hoc enrichment pattern. A second phenotype showed central-adiposity tendency with comparatively lower glycemic burden. The smallest non-noise phenotype showed a higher blood-pressure/renal-risk tendency. The algorithmic noise/outlier group was retained as an explicit model output rather than forced into a phenotype. These labels summarize relative biomarker profiles and enrichment gradients; they should be read as descriptive population patterns, not diagnostic categories.

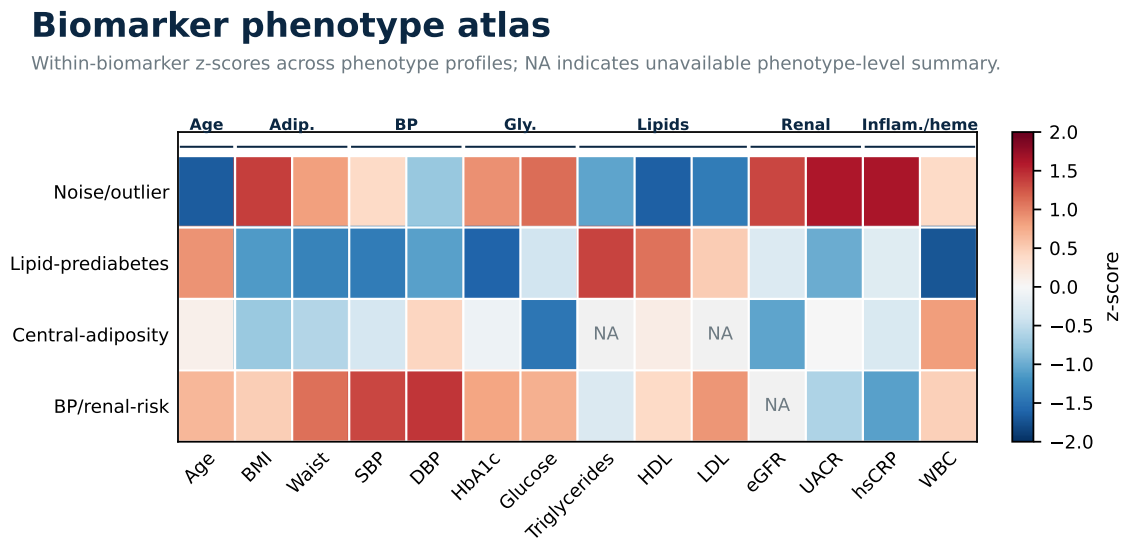


Figure 3: Biomarker phenotype atlas. Rows show final phenotype groups and columns show phenotype-level biomarker profiles grouped by domain. Values are within-biomarker z-scores across phenotype profiles and are intended for descriptive interpretation, not diagnostic thresholding. NA indicates unavailable phenotype-level summary values for that biomarker.

3.3 Post-hoc cardiometabolic burden gradients

Post-hoc enrichment analyses showed coherent cardiometabolic burden gradients after clustering (Figure 4; Table 2). The higher blood-pressure/renal-risk phenotype showed the strongest albuminuria enrichment (weighted estimate 0.160 versus overall 0.101; lift 1.58), and also showed elevated diabetes, hypertension, cardiovascular self-report, and CKD-risk signals. The lipid-prediabetes metabolic-risk phenotype showed enrichment for biomarker-defined prediabetes (0.499 versus 0.370; lift 1.35), dyslipidemia, diabetes, and metabolic-syndrome

component burden. The lower-glycemic central-adiposity phenotype showed central-obesity enrichment but lower diabetes and prediabetes burden than the lipid-prediabetes phenotype. These enrichment results were descriptive and were not used for clustering or model selection.

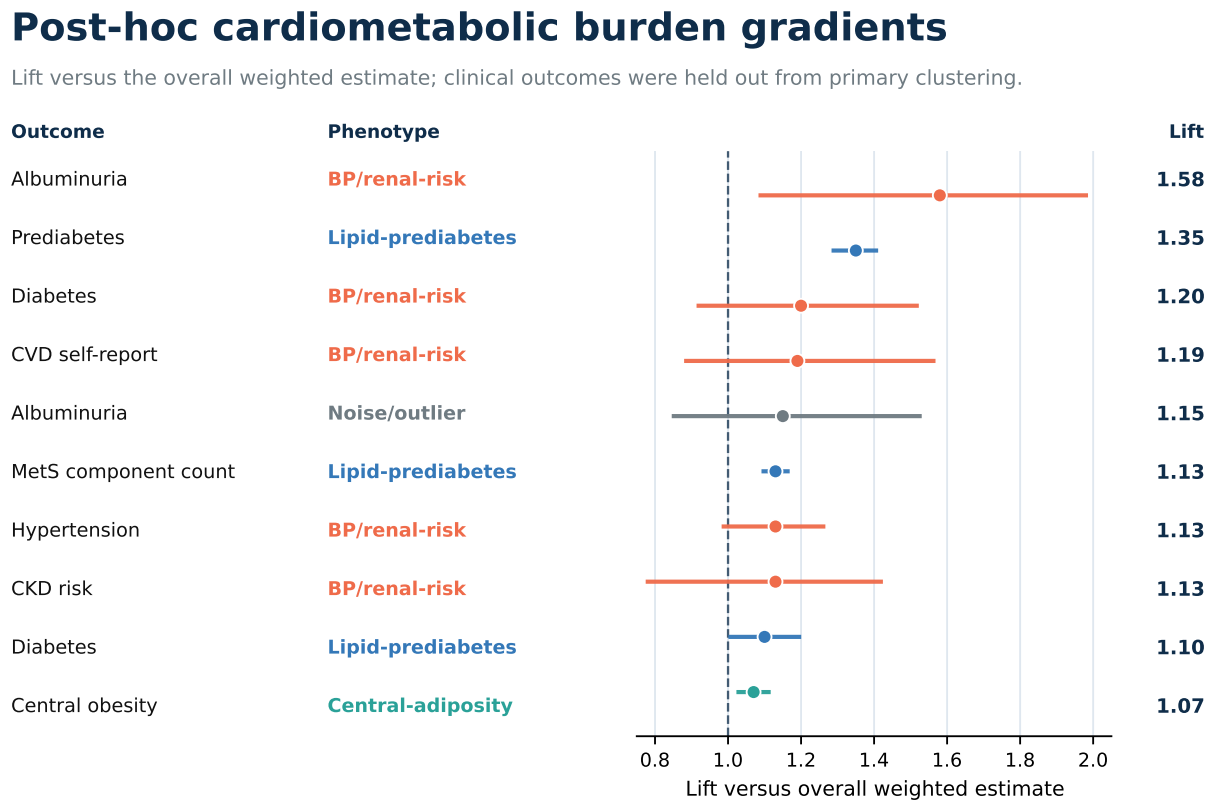


Figure 4: Post-hoc cardiometabolic burden gradients. Points show lift versus the overall weighted estimate for selected enrichment signals; intervals approximate 95% confidence intervals on the lift scale. Disease-burden variables were held out from primary clustering.

Table 2: Selected post-hoc disease-burden enrichment signals. The 95% CI is shown on the weighted-estimate scale.

Phenotype	Outcome	Weighted estimate (95% CI)	Overall	Lift
Higher blood-pressure/renal-risk	Albuminuria	0.160 (0.110–0.200)	0.101	1.58
Lipid-prediabetes metabolic-risk	Prediabetes biomarker	0.499 (0.477–0.520)	0.370	1.35
Higher blood-pressure/renal-risk	Diabetes any	0.179 (0.137–0.226)	0.149	1.20
Higher blood-pressure/renal-risk	CVD self-report	0.114 (0.085–0.150)	0.096	1.19
Algorithmic noise/outlier group	Albuminuria	0.117 (0.086–0.154)	0.101	1.15
Lipid-prediabetes metabolic-risk	Metabolic syndrome count	1.976 (1.917–2.032)	1.747	1.13
Higher blood-pressure/renal-risk	Hypertension any	0.573 (0.500–0.638)	0.506	1.13
Higher blood-pressure/renal-risk	CKD risk	0.160 (0.110–0.200)	0.141	1.13
Lipid-prediabetes metabolic-risk	Diabetes any	0.164 (0.150–0.178)	0.149	1.10
Lower-glycemic central-adiposity	Central obesity	0.598 (0.574–0.620)	0.558	1.07

3.4 Robustness and temporal transportability

Feature-block ablation identified lipid-domain removal as the most destabilizing perturbation (mean adjusted Rand index versus the full objective-biomarker reference solution 0.263; Figure 5). Removing age, glycemic, renal, liver/inflammation, hematologic, adiposity, or blood-pressure blocks preserved substantially higher sim-

ilarity to the reference solution. Adding demographic variables as a sensitivity analysis changed the cluster geometry and produced lower similarity to the objective-biomarker reference, supporting the decision to exclude demographic variables from primary clustering. Temporal replication in NHANES 2017–March 2020 supported partial profile transportability rather than exact phenotype identity, with best discovery-to-replication profile correlations from 0.714 to 0.909.

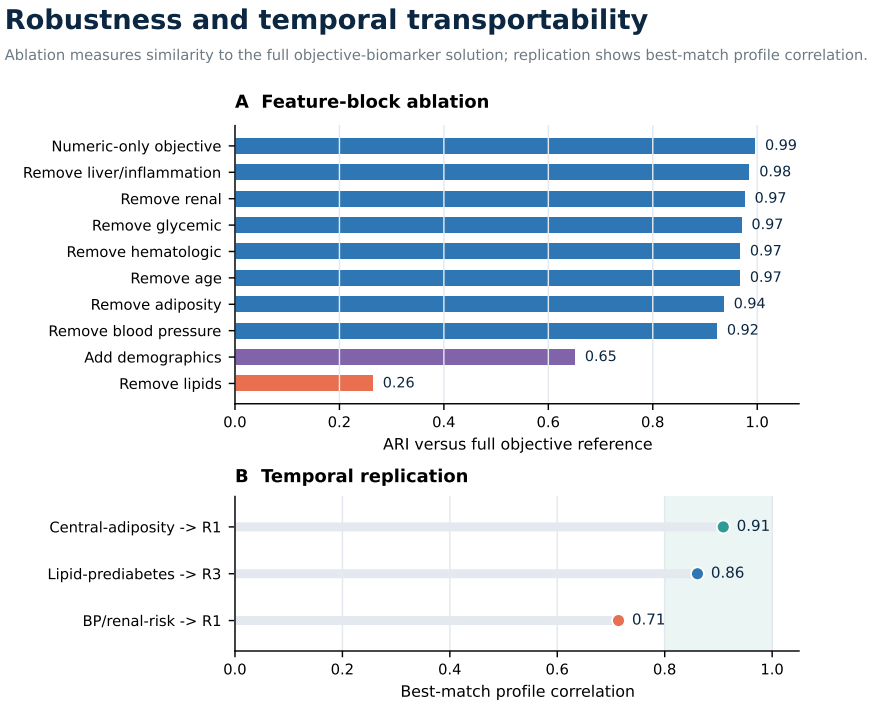


Figure 5: Robustness and temporal transportability. Panel A shows feature-block ablation robustness, ranked by adjusted Rand index versus the full objective-biomarker reference solution. Panel B shows best-match temporal replication profile correlations. Replication was interpreted as partial profile transportability rather than exact phenotype identity.

4 Discussion

This study identified stable density-based cardiometabolic biomarker phenotype tendencies among US adults using objective NHANES examination and laboratory measurements. The analysis deliberately separated phenotype discovery from clinical interpretation: disease labels, self-reported diagnoses, survey weights, demographic variables, and socioeconomic variables were excluded from primary clustering. Despite that separation, the learned groups showed coherent post-hoc burden gradients. The largest phenotype aligned with lipid-prediabetes metabolic-risk enrichment; a second phenotype captured central adiposity with comparatively lower glycemic burden; and a smaller dense region showed higher blood-pressure/renal-risk tendency with albuminuria, hypertension, diabetes, cardiovascular self-report, and CKD-risk signals.

The findings complement, rather than replace, conventional cardiometabolic risk frameworks. Metabolic syndrome criteria summarize coordinated abnormalities in adiposity, lipids, blood pressure, and glucose using clinically interpretable thresholds[1, 2]. The present analysis instead treated these domains as continuous multivariable biomarker structure and evaluated threshold-defined disease-related variables only after clustering. The central-adiposity phenotype should not be described as metabolically healthy obesity. A more cautious interpretation is that adiposity and glycemic burden can be partially dissociated in a national biomarker sample, consistent with prior evidence that cardiometabolic risk is structured by overlapping physiologic domains rather than by a

single diagnostic axis[3, 4].

The analysis also illustrates why leakage control is central in biomedical clustering. If disease labels, self-reported diagnoses, or demographic variables are used to define clusters, subsequent disease-burden differences can become circular. Here, clinical and survey-context variables were held out from the primary feature space and reintroduced only after the biomarker-derived solution was fixed. The demographic sensitivity analysis changed the learned geometry, reinforcing the decision to keep the primary solution biomarker-based. Feature-block ablation added a second interpretive layer: lipid removal was the major destabilizing perturbation, whereas most other single-domain removals preserved substantially higher similarity to the reference solution. This pattern suggests that lipid-related measurements were central to the density structure observed in the discovery cohort.

The study should be interpreted as an exploratory computational epidemiology analysis rather than a clinical subtype study. HDBSCAN noise is an algorithmic assignment, not a disease state. The phenotype names are intentionally modest and refer to relative biomarker-density tendencies, not to diagnoses, mechanisms, treatment strata, or prospective risk categories. Temporal replication supported partial profile transportability rather than exact phenotype identity, especially for the smaller blood-pressure/renal-risk phenotype. That result is useful because it shows that some profile structure persisted across survey periods, but it also constrains the strength of the claim.

Several limitations should guide interpretation. NHANES is cross-sectional for this analysis; the results do not establish causality, disease progression, treatment response, or prospective risk. UMAP-HDBSCAN workflows can be sensitive to preprocessing, dimensionality reduction, and hyperparameters even when seed stability is high. The reported silhouette is a learned-representation diagnostic and should not be read as evidence of clinical separability. Triglycerides were correctly resolved to LBXTLG for the 2021–2023 discovery cycle, but missingness was high (2,969 participants; 49.1%), and that limitation is emphasized in the supplementary audit. Post-hoc enrichment analyses are descriptive and may be influenced by residual confounding, measurement differences, missingness, and survey-design choices. External cohort validation, longitudinal outcome evaluation, and sensitivity analyses using alternative representation spaces remain necessary before any clinical interpretation. Accordingly, the study is best read as a reproducible descriptive atlas and stress-test framework rather than as evidence for clinical subtype assignment.

5 Conclusion

A leakage-controlled SVD–UMAP–HDBSCAN workflow provided a reproducible descriptive atlas of exploratory cardiometabolic biomarker-density tendencies among US adults in NHANES. The solution was internally stable across seeds, showed coherent post-hoc cardiometabolic burden gradients despite excluding disease labels from cluster formation, was most sensitive to removal of lipid-domain features, and demonstrated partial temporal profile transportability in a pre-pandemic NHANES comparison period. The main contribution is not a new clinical classification system, but an auditable framework for describing multivariable cardiometabolic biomarker structure without using diagnostic labels to create the groups. The findings support further external and longitudinal study of objective biomarker-derived cardiometabolic heterogeneity while remaining descriptive and hypothesis-generating rather than diagnostic, causal, or treatment-directing.

Ethics statement

This analysis used de-identified public-use NHANES data. No individual-level identifiable data are redistributed in the manuscript or accompanying repository.

Data availability

NHANES public-use data are available from the National Center for Health Statistics. Discovery analyses used NHANES August 2021–August 2023 public-use files, and temporal replication used NHANES 2017–March 2020 pre-pandemic public-use files. Raw NHANES source files are not redistributed. Aggregate derived tables and figures are included in the submission package and public repository.

Code availability

Code and aggregate reporting artifacts are available in the public project repository (github.com/KrishnaSaiPokala/nhanes-hdbscan-phenotyping). Users should obtain raw NHANES source files directly from NCHS for full local reconstruction.

Funding

No external funding was received.

Competing interests

The author declares no competing interests.

Author contributions

Krishna Sai Pokala conceived the study, designed the analysis, developed the software, interpreted the results, and prepared the manuscript.

Acknowledgments

The author thanks the NHANES participants and the National Center for Health Statistics for maintaining public-use data resources.

References

[1] Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. Executive summary of the third report of the National Cholesterol Education Program Expert Panel on Detection, Evaluation, and Treatment of High Blood Cholesterol in Adults. *JAMA* **285**, 2486–2497 (2001). doi:10.1001/jama.285.19.2486.

[2] Alberti, K. G. M. M. *et al.* Harmonizing the metabolic syndrome: a joint interim statement. *Circulation* **120**, 1640–1645 (2009). doi:10.1161/CIRCULATIONAHA.109.192644.

[3] Blüher, M. Metabolically healthy obesity. *Endocrine Reviews* **41**, bnaa004 (2020). doi:10.1210/endrev/bnaa004.

[4] Kramer, C. K., Zinman, B. & Retnakaran, R. Are metabolically healthy overweight and obesity benign? *Annals of Internal Medicine* **159**, 758–769 (2013). doi:10.7326/0003-4819-159-11-201312030-00008.

[5] Ahlqvist, E. *et al.* Novel subgroups of adult-onset diabetes and their association with outcomes. *Lancet Diabetes Endocrinol.* **6**, 361–369 (2018). doi:10.1016/S2213-8587(18)30051-2.

[6] Zaharia, O. P. *et al.* Risk of diabetes-associated diseases in subgroups of patients with recent-onset diabetes. *Lancet Diabetes Endocrinol.* **7**, 684–694 (2019). doi:10.1016/S2213-8587(19)30187-1.

[7] von Luxburg, U. Clustering stability: an overview. *Foundations and Trends in Machine Learning* **2**, 235–274 (2010). doi:10.1561/2200000008.

[8] Hennig, C. What are the true clusters? *Pattern Recognition Letters* **64**, 53–62 (2015). doi:10.1016/j.patrec.2015.04.009.

[9] National Center for Health Statistics. National Health and Nutrition Examination Survey. Centers for Disease Control and Prevention. https://www.cdc.gov/nchs/nhanes/.

[10] National Center for Health Statistics. NHANES tutorials and analytic guidelines. Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/tutorials/.

[11] National Center for Health Statistics. NHANES August 2021–August 2023 questionnaires, datasets, and related documentation. Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2021-2023.

[12] National Center for Health Statistics. NHANES 2017–March 2020 pre-pandemic questionnaires, datasets, and related documentation. Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020.

[13] McInnes, L., Healy, J. & Melville, J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426 (2018).

[14] Campello, R. J. G. B., Moulavi, D. & Sander, J. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, 160–172 (2013). doi:10.1007/978-3-642-37456-2_14.

[15] McInnes, L., Healy, J. & Astels, S. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* **2**, 205 (2017). doi:10.21105/joss.00205.

[16] American Diabetes Association Professional Practice Committee. Diagnosis and classification of diabetes: Standards of Care in Diabetes–2024. *Diabetes Care* **47**, S20–S42 (2024). doi:10.2337/dc24-S002.

[17] Whelton, P. K. *et al.* 2017 ACC/AHA guideline for high blood pressure in adults. *Hypertension* **71**, e13–e115 (2018). doi:10.1161/HYP.0000000000000065.

[18] Kidney Disease: Improving Global Outcomes CKD Work Group. KDIGO 2024 clinical practice guideline for the evaluation and management of chronic kidney disease. *Kidney International* **105**, S117–S314 (2024). doi:10.1016/j.kint.2023.10.018.

[19] Inker, L. A. *et al.* New creatinine- and cystatin C-based equations to estimate GFR without race. *N. Engl. J. Med.* **385**, 1737–1749 (2021). doi:10.1056/NEJMoa2102953.

[20] Hubert, L. & Arabie, P. Comparing partitions. *Journal of Classification* **2**, 193–218 (1985). doi:10.1007/BF01908075.

[21] Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison. *J. Mach. Learn. Res.* **11**, 2837–2854 (2010).

[22] Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987). doi:10.1016/0377-0427(87)90125-7.

[23] von Elm, E. *et al.* The Strengthening the Reporting of Observational Studies in Epidemiology statement. *PLoS Medicine* **4**, e296 (2007). doi:10.1371/journal.pmed.0040296.

[24] Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* **3**, 160018 (2016). doi:10.1038/sdata.2016.18.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- PokalaNHANESHDBSCANSupplementaryResearchSquareKSP.pdf