



Unsupervised Phenotype Clustering for Cardio metabolic Risk Stratification Using NHANES 1999-2006 Data

Zhomart Ghiran *

Department of Epidemiology and Biostatistics, University of Toronto, Toronto, Canada

DESCRIPTION

Cardio metabolic health is influenced by a complex combination of body composition, glucose regulation, lipid metabolism, blood pressure, lifestyle and demographic characteristics. Traditional approaches often evaluate these factors separately or classify individuals according to predefined clinical categories. Such methods can overlook meaningful combinations of characteristics that occur naturally within a population. Unsupervised clustering provides an alternative framework by allowing groups of individuals with similar health profiles to emerge from the available data without requiring predefined phenotype labels. When applied to nationally representative health information such as the National Health And Nutrition Examination Survey (NHANES), this approach can provide a broader perspective on patterns associated with cardio metabolic risk. NHANES 1999-2006 contains extensive information on anthropometric measurements, laboratory biomarkers, blood pressure, health behaviours and demographic characteristics.

Variables such as body mass index, waist circumference, fasting glucose, triglycerides, high-density lipoprotein cholesterol, systolic blood pressure and other metabolic indicators can be considered together to describe individual health profiles. Rather than treating each measurement as an isolated predictor, clustering techniques can identify combinations that repeatedly occur among different groups of participants. Unsupervised clustering begins by transforming relevant variables into a form that allows meaningful comparison. Standardization is particularly important because clinical measurements may use very different scales. For example, glucose concentration and waist circumference cannot be compared directly without appropriate normalization. After pre-processing, clustering algorithms can assess similarities between individuals and assign participants to groups according to their overall profiles. Approaches such as hierarchical clustering, k-means clustering,

Gaussian mixture models and density-based methods can produce different representations of population structure. Selecting an appropriate method requires consideration of sample characteristics, variable distributions and the desired interpretation of the resulting phenotypes.

One potential outcome is the identification of a relatively metabolically favorable group characterized by lower adiposity, healthier lipid levels, normal glucose regulation and favorable blood pressure. Another cluster may display substantial central adiposity with elevated triglycerides and impaired glucose regulation, reflecting a pattern associated with increased metabolic vulnerability. Additional groups may demonstrate elevated blood pressure with comparatively moderate changes in other biomarkers, while another phenotype could combine several adverse characteristics simultaneously. These patterns are important because people with similar values for one individual measurement may have substantially different overall cardio metabolic profiles. The interpretation of clusters should extend beyond simple descriptions of average values. Cardio metabolic risk is multidimensional and the combination of abnormalities may provide greater information than any single marker. For instance, central adiposity accompanied by elevated triglycerides and reduced high-density lipoprotein cholesterol may indicate a different risk pattern from isolated obesity. Likewise, elevated glucose in an individual with otherwise favourable metabolic characteristics may represent a different phenotype from simultaneous abnormalities in glucose, blood pressure and lipid metabolism.

NHANES data also make it possible to consider demographic and behavioral characteristics alongside physiological measurements. Age, sex, physical activity, smoking behavior, dietary indicators and socioeconomic variables may differ across identified clusters. These differences can help clarify why certain cardio metabolic profiles occur more frequently in particular population groups. However, associations should be interpreted

Correspondence to: Zhomart Ghiran, Department of Epidemiology and Biostatistics, University of Toronto, Toronto, Canada, Email: zhomart.ghiran@utoronto.ca

Received: 01-Dec-2025, Manuscript No. JDMGP-25-32429; **Editor assigned:** 03-Dec-2025, PreQC No JDMGP-25-32429; **Reviewed:** 17-Dec-2025, QC No. JDMGP-25-32429; **Revised:** 23-Dec-2025, Manuscript No. JDMGP-25-32429; **Published:** 31-Dec-2025, DOI: 10.35248/2153-0602.25.16.400

Citation: Ghiran Z (2025). Unsupervised Phenotype Clustering for Cardio metabolic Risk Stratification Using NHANES 1999-2006 Data. J Data Mining Genomics Proteomics. 16:400.

Copyright: © 2025 Ghiran Z. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

cautiously because clustering describes patterns rather than proving that one characteristic directly causes another. A major advantage of unsupervised classification is its ability to reveal heterogeneity that conventional threshold-based categories may conceal. Clinical cutoffs are useful for diagnosis and treatment decisions, but biological risk often exists along a continuum. Two individuals positioned on opposite sides of a threshold may have very similar underlying profiles, whereas two individuals meeting the same diagnostic category may have substantially different combinations of metabolic abnormalities. Cluster-based phenotyping can complement conventional categories by highlighting these multidimensional patterns.

The reliability of the resulting phenotypes depends strongly on variable selection, missing-data handling, scaling procedures and the chosen clustering algorithm. Sensitivity assessments using alternative methods or parameter settings can determine whether the identified groups remain reasonably consistent. Because NHANES uses a complex sampling design, appropriate weighting and consideration of the survey structure are also important when describing population-level patterns. External validation with other population datasets would further strengthen confidence in the generalizability of the phenotypes. These approaches may have practical value for cardio metabolic risk assessment and prevention.