Original Article
http://mjiri.iums.ac.ir
Medical Journal of the Islamic Republic of Iran (MJIRI)

Med J Islam Repub Iran. 2026 (17 Aug);40.88. https://doi.org/10.47176/mjiri.40.88

Latent Class Analysis for Clustering Iranian Individuals Based on Cardiovascular Risk Factors

Hajar Nikbakht[1], Seyyed Amir Yasin Ahmadi[2, 3], Abbas Aghaei[1], Mahdi Chinichian[4]*

Received: 20 Mar 2026 Revised: 10 Aug 2026 Accepted: 13 Aug 2026 Published: 17 Aug 2026

Abstract

Background: Cardiovascular disease (CVD) remains a major public health concern, and conventional risk-stratification tools, such as WHO/ISH (IraPEN), may not adequately capture the heterogeneity of risk profiles within the population. This study aimed to identify the underlying patterns of CVD risk factors among employees undergoing routine occupational health assessments in Iran.

Methods: In this cross-sectional secondary analysis, data from 1,281 adults aged 40–79 years attending a governmental occupational health clinic in 2024 were analyzed. Core cardiovascular risk variables—including sex, diabetes, smoking, systolic blood pressure, and total cholesterol—were used. Latent class analysis (LCA) employing a five-class model was conducted.

Results: Five latent classes of cardiovascular risk were identified, reflecting heterogeneous conditions. These classes included high-risk males without DM (population share: 0.1824), hyperlipidemic non-smoking females (population share: 0.1487), low-risk individuals (population share: 0.3011), high-risk males without hypertension (population share: 0.36), and high-risk males without hyperlipidemia (population share: 0.0079).

Conclusion: The LCA-based classification identified five distinct and internally homogeneous cardiovascular risk profiles that captured sex-specific and metabolically clustered patterns of risk. This approach reveals heterogeneity not reflected in the conventional IraPEN five-tier risk groups.

Keywords: Cardiovascular Diseases, Risk Assessment, Atherosclerosis, Primary Prevention, Occupational Health

Cite this article as: Nikbakht H, Ahmadi SAY, Aghaei A, Chinichian M. Latent Class Analysis for Clustering Iranian Individuals Based on Cardiovascular Risk Factors. *Med J Islam Repub Iran.* 2026 (17 Aug);40:88. https://doi.org/10.47176/mjiri.40.88

Introduction

The burden of cardiovascular diseases (CVD) in Iran represents a significant public health challenge, characterized by high prevalence and severe complications. Studies indicate that lifestyle factors—including poor nutrition, physical inactivity, and tobacco use—play a critical role in exacerbating this burden. Furthermore, population aging and demographic shifts are expected to intensify the impact of these diseases. Given these challenges, strategic planning for the prevention and management of CVD is essential. Due to population growth and aging, the CVD burden in Iran is projected to rise significantly between 2005 and 2025, even in the absence of worsening epidemiological trends. Notably, the highest disease burden is expected to persist among the working-age population (under 65 years). While population aging is an inevitable phenomenon, premature CVD can be mitigated through

Corresponding author: *Dr Mahdi Chinichian, chinichian.m@iums.ac.ir*

1. Preventive Medicine and Public Health Research Center, Psychosocial Health Research Institute, Department of Community and Family Medicine, School of Medicine, Iran University of Medical Sciences, Tehran, Iran
2. Clinical Cancer Research Center, Milad General Hospital, Tehran, Iran
3. Department of Epidemiology, School of Public Health, Iran University of Medical Sciences, Tehran, Iran
4. Occupational Medicine Research Center, Iran University of Medical Sciences, Tehran, Iran

Key Messages:

↑What is "already known" in this topic:

Cardiovascular disease represents a significant public health challenge in Iran and globally. National programs, such as IraPEN, depend on conventional tools for stratifying cardiovascular risk. Standard prediction instruments, including the ACC/AHA ASCVD Risk Estimator, are commonly employed to inform primary prevention decisions; however, they may not adequately account for the heterogeneity and clustering of risk factors among individuals.

→What this article adds:

This study employs latent class analysis (LCA) to derive five distinct latent profiles of cardiovascular risk factors, rather than relying solely on predefined categorical risk groups such as those utilized in IraPEN. The identified latent classes encompass both low- and high-risk profiles, featuring sex-specific and metabolically clustered patterns, and demonstrate that individuals may be classified as high cardiovascular risk even in the absence of all traditional IraPEN "red-flag" factors.

coherent health policies and structured planning, such as early intervention, systematic screening, and the early identification of high-risk groups (1).

Over the past four decades, Iran has experienced significant socioeconomic transformations, resulting in a marked increase in CVD prevalence and an epidemiological transition in which CVD has supplanted infectious diseases as the primary cause of mortality. Currently, CVD accounts for 46% of all deaths and constitutes a substantial portion of the overall disease burden, primarily driven by ischemic heart disease and stroke. Contributing factors to this epidemic include socioeconomic shifts, dietary changes, reduced physical activity, urbanization, increased life expectancy, and inadequate access to specialized healthcare. In response, the Iranian Ministry of Health has prioritized strategies for CVD prevention and control, focusing on health promotion, early diagnosis, and the management of both acute and chronic cases (2).

Estimating the costs associated with disease prevention programs is of paramount importance in developing countries. Previous evaluations of the cost-effectiveness of the WHO/ISH (IraPEN) Tool program have demonstrated that the most economically viable strategies involve preventive treatments targeting high-risk individuals. The IraPEN program in Iran, along with similar initiatives in low- and middle-income countries (LMICs), has proven to be highly cost-effective (3). This significance is further underscored by the substantial financial burden of CVD, which these programs can help alleviate (4).

Currently, the IraPEN risk stratification is based on variables such as age, sex, diabetes, smoking status, systolic blood pressure, and cholesterol levels. Based on these criteria, individuals aged 40 to 80 are categorized into five risk groups—ranging from low risk (<10%) to very high risk (>40%)—regarding the 10-year probability of fatal and non-fatal cardiovascular and cerebrovascular events. Despite the implementation of IraPEN, it has faced criticisms, most notably the lack of periodic updates. Furthermore, there is a recognized need to categorize individuals based on the actual distribution of risk patterns. Latent Class Analysis (LCA) is a statistical methodology that identifies underlying patterns based on the distribution of manifest variables. The present study aims to utilize LCA to investigate the patterns of CVD risk factors among employees of the Ministry of Health. LCA offers a data-driven alternative to predefined risk categories by identifying naturally occurring subgroups of individuals with similar risk profiles. Unlike the fixed five-level IraPEN stratification, LCA can capture complex combinations of risk factors and reveal more homogeneous and clinically meaningful groups. This approach may therefore provide a more practical and informative framework for identifying high-risk populations and guiding targeted preventive interventions.

Methods

Study design

This study was conducted as a secondary analysis of previously collected occupational health data using a cross-sectional design. The analysis relied entirely on information obtained during routine occupational health examinations, which are regularly conducted to monitor employees' health status and identify potential risk factors for chronic diseases. Since the data had already been collected as part of standard workplace health assessments, the present study did not involve any additional interventions, measurements, or direct data collection from participants. Consequently, the research was limited to analyzing existing records, ensuring that no extra procedures or burdens were imposed on the individuals whose data were included in the study.

Participants

The study population consisted of adults aged between 40 and 79 years who attended the Occupational Medicine Clinic of Iran University of Medical Sciences in 2024 as part of their mandatory occupational health assessments. These evaluations are routinely conducted to ensure compliance with national occupational health regulations and to monitor workers' physical well-being. For the purposes of this research, available clinical records were systematically reviewed, and any files missing key variables required for cardiovascular risk estimation were excluded from the analysis to maintain data accuracy and methodological rigor. Ethical approval for the study was obtained from the Ethics Committee of Iran University of Medical Sciences under protocol number IR.IUMS.FMD.REC.1403.1206.

To safeguard participant privacy, all identifying information was removed prior to data processing, ensuring that confidentiality was rigorously upheld. Since the study relied solely on pre-existing clinical data and involved no direct interaction with participants, it did not introduce additional burdens, procedures, or financial costs. Attendance at the occupational health clinic was obligatory within the national regulatory framework; therefore, inclusion in this research did not disrupt or modify the routine health-care pathway of the individuals whose records were analyzed.

Variables

Cardiovascular risk for each individual was assessed using two standardized and widely accepted risk-prediction instruments. One of these tools was the Atherosclerotic Cardiovascular Disease Risk Estimator (ASCVD), a model developed by the American College of Cardiology and the American Heart Association to estimate the 10-year probability of both fatal and non-fatal cardiovascular events among adults aged 40 to 79 years. The ASCVD risk calculation incorporates several key demographic and clinical variables known to influence cardiovascular outcomes. These include sex as a fundamental demographic determinant; systolic blood pressure and the use of antihypertensive medications to assess blood pressure control status; diabetes status as a major metabolic risk factor; smoking status as a critical lifestyle determinant; and total cholesterol level. Together, these variables facilitate a comprehensive estimation of an individual's underlying cardiovascular risk profile. To enable a direct comparison between the two classification approaches (LCA and Ira-

[DOI: 10.47176/mjiri.40.88]

PEN), the manifest variables used in the LCA were selected to correspond precisely to the cardiovascular risk variables assessed within the IraPEN risk stratification framework. For LCA, blood pressure and cholesterol were entered as yes/no categories rather than raw numbers, and the exact cut-off values (SBP > 140 mmHg, cholesterol > 195 mg/dl) were based on IraPEN.

Bias and missing data

This study, based on a secondary analysis of previously collected data and lacking random sampling, must consider the potential limitation of selection bias. The study population comprised individuals who attended routine occupational health examinations, which may restrict the sample's representativeness relative to the broader general population. The existing dataset utilized for secondary analysis did not contain any missing data.

Statistical analysis

LCA was employed to identify distinct underlying patterns of cardiovascular risk factors based on the distribution of manifest variables. The manifest variables included sex, diabetes status, smoking status, systolic blood pressure, and total cholesterol. LCA models with three, four, five, and six-class solutions were estimated with 30 number of iterations using the poLCA package in R programming language version 4.4.2. Model selection was guided by a combination of statistical fit indices and the substantive interpretability of the classes. Model fit and the optimal number of classes were evaluated using standard information criteria—Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC)—as well as likelihood-based goodness-of-fit statistics, including the likelihood ratio chi-square statistic (G^2) and the Pearson chi-square goodness-of-fit statistic (X^2).

Results

The data of a total of 1,281 individuals were studied, with the age range of 40 to 77 years. Approximately 72.37% of the individuals were male, and 5.23% were diabetic. Furthermore, 19.05% were cigarette smokers, while 9.13% of the individuals had systolic blood pressure (SBP) exceeding 140 mmHg. Additionally, 33.41% of the individuals had total cholesterol levels greater than 195 mg/dl.

LCA was conducted four times, evaluating three, four, five, and six classes. Based on AIC and BIC, the three-class model was identified as the optimal model, whereas the six-class model was deemed superior according to G2 and X2. Nevertheless, the five-class model was ultimately selected due to a significant change in G2 and X2 compared to the four-class model (Table 1).

Table 1. Comparison of the models based on information criteria and fitting indices

Number of classes	AIC	BIC	G2	X2
3	5560	5648	17.22	19.23
4	5567	5686	12.27	15.98
5	5571	5720	4.03	4.50
6	5579	5760	0.57	0.52

G2: Likelihood ratio/deviance statistic, X2: Chi-square goodness of fit

Table 2. Conditional probabilities of the imported variables among the classes

Classes: names	Male sex	Having DM	Smoking	High SBP	High cholesterol
Class 1: High-risk males without DM	1.0000	0.0000	0.3001	0.4130	0.3636
Class 2: Hyperlipidemic non-smoker females	0.2634	0.0562	0.0000	0.0354	0.7836
Class 3: Low-risk individuals	0.4670	0.0298	0.0202	0.0096	0.0550
Class 4: High-risk males without hypertension	0.9824	0.0757	0.3536	0.0000	0.3743
Class 5: High-risk males without hyperlipidemia*	1.0000	0.9843	0.3022	1.0000	0.0000

* This class should be interpreted cautiously due to its very small size.

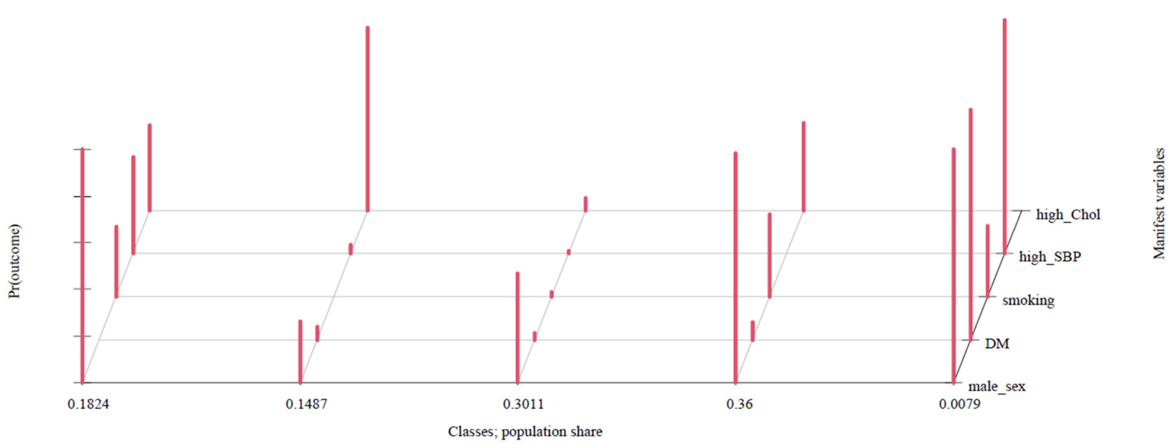

Figure 1. Comparison of conditional item probabilities (or risk factor prevalence) across the five identified latent classes. The horizontal axis represents Class 1 to Class 5 (from left to right).

[DOI: 10.47176/mjiri.40.88]

According to the five-class model, the five classes were as follows (Table 2, Figure 1):

Class 1: High-risk males without DM (population share: 0.1824)

Class 2: Hyperlipidemic non-smoking females (population share: 0.1487)

Class 3: Low-risk individuals (population share: 0.3011)

Class 4: High-risk males without hypertension (population share: 0.36)

Class 5: High-risk males without hyperlipidemia (population share: 0.0079).

The association between the probability of belonging to each class and the ASCVD score was examined using simple beta regression. Consequently, classes 2 and 3 were classified as low risk, while the remaining classes were categorized as high-risk (Table 3, Figure 2).

According to the IraPEN classification results, out of these 1,281 individuals, there were 1,273 low risk individuals. Only there were six intermediate risks, one moderately high risk and one high risk individuals. The classes of LCA attributed to the eight non-low risk individuals are shown (Table 4). Accordingly, there were no individuals in low-risk class 3.

Discussion

The present study utilized LCA to identify underlying patterns of cardiovascular risk factors among employees undergoing occupational health evaluations. The findings highlight the heterogeneity of cardiovascular risk profiles within the population and demonstrate how multiple risk factors may cluster together to form distinct subgroups. Identifying such patterns can provide valuable insights for targeted preventive strategies and more efficient risk stratification in clinical and public health settings.

According to the IraPEN classification, only eight out of 1,281 individuals were categorized as low risk. In contrast, based on the LCA results, approximately 45% of individuals were assigned to low-risk classes (i.e., Classes 2 and 3). This suggests that IraPEN, compared with LCA, may underestimate individuals' risk of cardiovascular events. Although the smallest LCA class (Class 5) had a very low prevalence (0.0079), it was still larger than the smallest IraPEN category (the very high-risk class), which had a prevalence of zero. Therefore, the small size of Class 5 in the LCA does not necessarily indicate that this class is inappropriate. The five-class structure of the LCA allowed it to be compared with the five IraPEN categories. However, the descriptive comparison indicated poor agreement between the two systems. Overall, our LCA-based system appears to be more appropriate and informative than IraPEN for estimating cardiovascular risk.

The application of LCA in cardiovascular research has gained increasing attention in recent years. For instance, Bu et al. (2021) utilized LCA in the United Kingdom to explore the relationship between loneliness, social isolation, and modifiable cardiovascular risk factors among adults aged over 50 years. Analyzing data from 8,218 participants in the English Longitudinal Study of Ageing, the authors identified four latent groups: low-risk, high-risk, clinical-risk, and lifestyle-risk classes. Their findings showed that loneliness was associated with a higher likelihood of belonging to the high-risk and lifestyle-risk groups, underscoring the influence of psychosocial factors in shaping cardiovascular risk patterns (5).

Tegegne et al. (2022) similarly investigated the clustering of lifestyle risk behaviors and their association with cardiovascular disease among British adults using data from the UK Biobank. Their latent class analysis revealed that the accumulation of unhealthy behaviors, such as smoking, physical inactivity, poor diet, excessive alcohol consumption, and prolonged sedentary behaviors substantially increased the risk of cardiovascular disease. Individuals exhibiting multiple concurrent risk behaviors were significantly more likely to develop CVD, highlighting the importance of addressing clustered lifestyle factors in prevention programs (6).

Other studies have also used LCA to classify risk domains in populations at high cardiovascular risk. Zhang et al. (2020) conducted a study in China to identify the major risk domains among individuals at high risk of CVD. By analyzing data from 2,951 residents, the authors identified four principal risk groups characterized by dyslipidemia, abnormal body weight, impaired respiratory function, and abnormal sleep-related breathing. Significant differences

Table 3. The association of the probability of being in each class with ASCVD score using beta regression

Probability of being in	Odds ratio (95% CI)	*P* value	Result
Class 1	1.075 (1.050-1.101)	<0.001	High risk class
Class 2	0.828 (0.807-0.849)	<0.001	Low risk class
Class 3	0.784 (0.762-0.806)	<0.001	Low risk class
Class 4	1.020 (1.002-1.038)	0.031	High risk class
Class 5	1.016 (1.001-1.031)	0.037	High risk class

Table 4. Comparison of our clustering system with IraPEN classification in individuals who are not at low risk in IraPEN

IraPEN category	Probability of being in class 1	Probability of being in class 2	Probability of being in class 3	Probability of being in class 4	Probability of being in class 5
Intermediate risk	0	0.986*	0.014	0	0
Intermediate risk	0.946*	0.051	0.003	0	0
Intermediate risk	0	0.003	0.007	0	0.990*
Moderately high risk	0	0	0.013	0.987*	0
Intermediate risk	0.946*	0.051	0.003	0	0
Intermediate risk	0	0	0.013	0.987*	0
Intermediate risk	0.955*	0.008	0.034	0	0.002
High risk	0.999*	0	0.001	0	0

* Final attributed class

[DOI: 10.47176/mjiri.40.88]

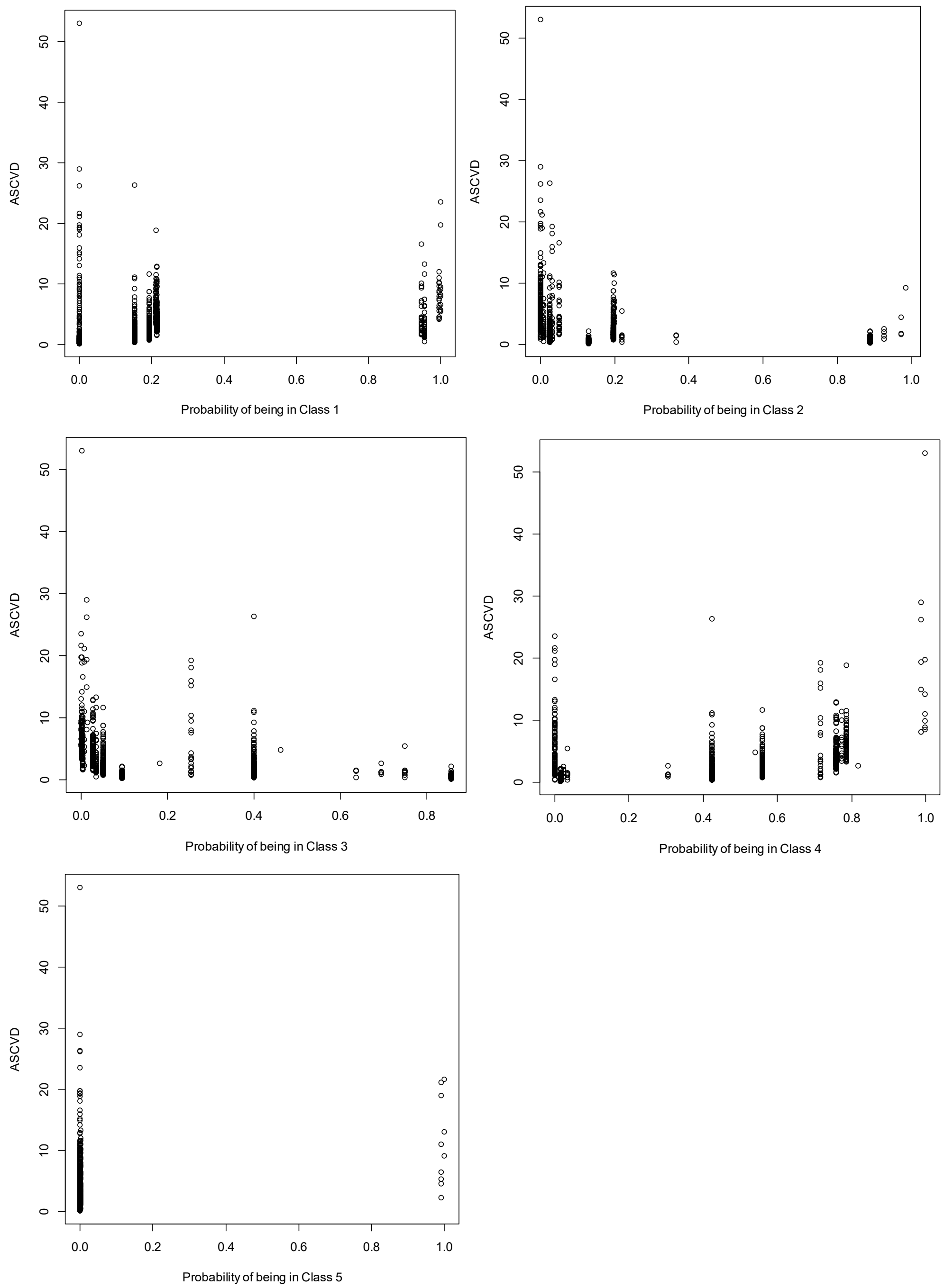

Figure 2. Scatter plot for the association of the probability of being in each class with ASCVD score

in age, sex distribution, and medical history were observed among these classes, indicating that risk factor clustering varies substantially across subpopulations (7).

In addition to risk factor clustering, LCA has been utilized to characterize broader socio-clinical phenotypes in patients with established CVD. Kundi et al. (2024) employed LCA to identify socio-clinical subgroups among patients with atherosclerotic cardiovascular disease (ASCVD) using data from the U.S. National Health Interview Survey linked to mortality records. Three major classes were identified, each with distinct demographic and comorbidity profiles, and membership in certain classes was associated with a significantly higher long-term mortality risk. These findings illustrate how the identification of latent subgroups can improve understanding of heterogeneity in cardiovascular outcomes (8).

[DOI: 10.47176/mjiri.40.88]

Furthermore, Chen et al. (2022) investigated the relationship between various patterns of physical activity and cardiovascular disease risk among Chinese adults. Utilizing LCA, three distinct physical activity patterns were identified, and their associations with the Framingham Risk Score were examined. The results indicated that higher levels of occupational or leisure-time physical activity were associated with a lower cardiovascular risk, whereas more sedentary patterns were linked to an increased risk, particularly among men (9).

Overall, previous studies consistently demonstrate that LCA is a valuable methodological approach for identifying hidden patterns of cardiovascular risk factors, behaviors, and socio-clinical characteristics. The application of this method in the present study contributes to the growing body of evidence that emphasizes the importance of recognizing heterogeneous risk profiles when designing prevention strategies and risk assessment frameworks.

A primary limitation of this study is its reliance on existing secondary data, which restricts control over measurement accuracy, data collection procedures, and the ability to rectify potential recording errors. This constraint may reduce the internal validity of the findings. Furthermore, since the data were originally collected within the routine framework of occupational health assessments, certain behavioral or social determinants of CVD could not be evaluated. Nevertheless, the relatively large sample size and the homogeneity of the study population—employees assessed under standardized health monitoring protocols—help mitigate some of these limitations and enhance the reliability of the latent class patterns identified.

Conclusion

In conclusion, this study demonstrated that cardiovascular risk factors within the study population cluster into distinct patterns rather than occurring independently. Using latent class analysis, five underlying classes of cardiovascular risk were identified, reflecting various combinations of demographic and clinical risk factors. These classes included groups characterized by relatively low risk, sex-specific risk patterns, and clusters dominated by metabolic and behavioral risk factors such as diabetes, hypertension, dyslipidemia, and smoking. The identification of these heterogeneous risk profiles underscores the complexity of cardiovascular risk distribution within the population and suggests that conventional single-factor approaches may not fully capture this variability. Recognizing such latent subgroups can facilitate more targeted prevention strategies and improve cardiovascular risk stratification in occupational and public health settings. In summary, the LCA-based classification revealed five distinct and internally homogeneous cardiovascular risk profiles that captured sex-specific and metabolically clustered patterns of risk, thereby uncovering heterogeneity not reflected in the conventional IraPEN five-class risk groups and may offer a complementary framework for future research.

Acknowledgment
None.

Conflict of Interests
The authors declare that they have no competing interests.

Authors' Contributions
HN: design and conceptualization, critical revision; SAYA: design and conceptualization, formal analysis and interpretation, drafting; AA: data collection, critical revision; and MC: design and conceptualization, critical revision.

All the authors approved the final version of the manuscript.

Ethical Considerations
Ethical approval for the study was obtained from the Ethics Committee of Iran University of Medical Sciences under protocol number IR.IUMS.FMD.REC.1403.1206. To protect participant privacy, all identifying information was removed prior to data processing, thereby ensuring that confidentiality was rigorously maintained. Since the study relied solely on pre-existing clinical data and did not involve direct interaction with participants, it did not impose any additional burdens, procedures, or financial costs.

Funding Support
Iran University of Medical Sciences with grant number 31705.

Data Availability
Data is available upon a reasonable request.

AI Use Statement
We acknowledge the use of ChatGPT 5.2 (OpenAI) for minor language editing and improving the clarity of the manuscript during the revision of the draft. The tool was used only for linguistic refinement and did not contribute to the study design, data analysis, interpretation of results, or the generation of scientific content.

References

1. Sadeghi M, Haghdoost AA, Bahrampour A, Dehghani M. Modeling the burden of cardiovascular diseases in Iran from 2005 to 2025: the impact of demographic changes. Iran J Public Health. 2017 Apr;46(4):506.
2. Sarrafzadegan N, Mohammmadifard N. Cardiovascular disease in Iran in the last 40 years: prevalence, mortality, morbidity, challenges and strategies for cardiovascular prevention. Arch Iran Med. 2019 Apr 1;22(4):204-10.
3. Mokhtari M, Khalil D, Farzadfar F, Daroudi R, Asadi-Lari M. The burden of cardiovascular disease attributable to modifiable risk factors and cost-effectiveness analysis of IraPEN program in the general population of Iran. Med J Islam Repub Iran. 2022;36(1):e73. Doi: 10.47176/mjiri.36.73
4. Alipour V, Zandian H, Yazdi-Feyzabadi V, Avesta L, Moghadam TZ. Economic burden of cardiovascular diseases before and after Iran's health transformation plan: evidence from a referral hospital of Iran. Cost Eff Resour Alloc. 2021 Dec;19:1-0. Doi: 10.1186/s12962-020-00250-8
5. Bu F, Steptoe A, Fancourt D. Relationship between loneliness, social isolation and modifiable risk factors for cardiovascular disease: a latent

[DOI: 10.47176/mjiri.40.88]

class analysis. J Epidemiol Community Health. 2021 Aug 1;75(8):749-54. Doi: 10.1136/jech-2020-215539

6. Tegegne TK, Islam SM, Maddison R. Effects of lifestyle risk behaviour clustering on cardiovascular disease among UK adults: latent class analysis with distal outcomes. Sci Rep. 2022 Oct 17;12(1):17349. Doi: 10.1038/s41598-022-22469-6
7. Zhang ZH, Zhang Q, Li CH. Classification of main risk domains in population at high-risk of cardiovascular diseases: a latent class analysis. Chin J Pub Health. 2020;36(8):1196-9. Doi: 10.11847/zgggws1123678
8. Kundi H, Hagan K, Yahya T, Sharma G, Al-Kindi S, Javed Z, Nasir K. Socioclinical Phenotyping in Patients With Atherosclerotic Cardiovascular Disease: A Latent Class Analysis. Am J Cardiol. 2024 Jan 15;211:57-63. Doi: 10.1016/j.amjcard.2023.10.067
9. Chen C, Liu J, Lu S, Ding G, Wang J, Qin Y, Wang Z, Wang X, Zhang Z, Xiang Q. Study on the associations of physical activity types and cardiovascular diseases among Chinese population using latent class analysis method. Sci Rep. 2022 May 16;12(1):8081. Doi: 10.1038/s41598-022-12182-9

[DOI: 10.47176/mjiri.40.88]