10.24425/acs.2026.158423

Archives of Control Sciences
Volume 36(LXXII), 2026
No. 1, *pages* 99–131

Computational framework for dynamic cardiovascular risk assessment with cluster-specific Cox models and cumulative risk analysis

Himanshi LIYANAGE and Marta LIPNICKA

This paper presents a novel computational framework for assessing cardiovascular disease (CVD) risk by integrating unsupervised clustering techniques with survival analysis. The proposed method enables dynamic and individualized risk prediction by organizing patient data into structured clusters based on shared cardiovascular risk factors. The framework begins with competitive learning, an unsupervised clustering method, to group patients into clusters that reflect distinct risk profiles. Each cluster is represented by its centroid, calculated as the mean of the 9-dimensional feature vectors of its members, ensuring that the clusters effectively summarize patient data while preserving critical risk characteristics. For each cluster, an independent Cox Proportional Hazards Model is applied to analyze survival data, capturing the unique relationships between cardiovascular risk factors and survival outcomes within that cluster. A key innovation of this study is the introduction of the Cumulative Prevalence Ratio (CPR), a new metric that aggregates hazard rates over time separately for each cluster. This approach provides a comprehensive view of cumulative cardiovascular risk, enabling precise categorization of the patient into risk groups based on cumulative exposure to evolving risk factors. By integrating cluster-specific hazard functions and temporal risk metrics, the proposed framework improves the precision and adaptability of CVD risk predictions, paving the way for personalized and data-driven healthcare interventions.

Key words: cardiovascular disease, cox proportional hazards model, independent hazard functions, competitive learning, cluster-based data simplification, cumulative prevalence ratio, personalized risk assessment

Copyright © 2026. The Author(s). This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (CC BY-NC-ND 4.0 https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits use, distribution, and reproduction in any medium, provided that the article is properly cited, the use is non-commercial, and no modifications or adaptations are made

H. Liyanage (e-mail: nishi.himanshi@wmii.uni.lodz.pl) and M. Lipnicka (corresponding author, e-mail: marta.lipnicka@wmii.uni.lodz.pl) are with Faculty of Mathematics and Computer Science, University of Lodz, Stefana Banacha 22, 90-238 Łódź, Poland.

This work was supported by the statutory grant No. 0211/SBAD/0125.
Received 4.02.2025. Revised 15.01.2026.

1. Introduction

Cardiovascular disease (CVD) remains one of the leading causes of mortality worldwide, emphasizing the need for accurate and personalized risk prediction to enable effective intervention. Early identification of an individual's CVD risk allows customized preventive measures to significantly improve long-term health outcomes. However, traditional risk assessment methods often fail to capture the dynamic nature of risk factors, which can evolve over time due to lifestyle changes, variations in health status, or medical treatments. Addressing these limitations requires a framework that integrates temporal dynamics and cluster-specific risk profiles into a unified, yet adaptable, approach [4, 5].

This study introduces a novel computational framework for CVD risk assessment that integrates unsupervised clustering techniques with survival analysis. The proposed approach combines the strengths of clustering methods, such as competitive learning, with cluster-specific Cox Proportional Hazards Models to provide a dynamic and personalized understanding of CVD risk. By calculating independent hazard functions for each group, the framework enables detailed modeling of how cardiovascular risk factors interact and evolve over time for distinct groups of patients [6, 10].

The process begins with unsupervised clustering to group patients based on shared cardiovascular risk factors. Each cluster is represented by its centroid, calculated as the mean of the 9-dimensional characteristic vectors of the patients within the cluster. This step simplifies the dataset, organizing it into distinct structured profiles that serve as the basis for survival analysis. For each group, an independent risk function is then estimated, capturing the unique relationships between cardiovascular risk factors and survival outcomes within that group [1,8].

A key component of this framework is the use of the Cumulative Prevalence Ratio (CPR), which aggregates hazard rates over time into a single cluster-specific measure. By capturing temporal changes in risk for each cluster, the CPR provides a comprehensive and flexible method for assessing cumulative CVD risk. This approach improves the accuracy of risk predictions and provides deeper insight into how cardiovascular risk factors vary between patient subgroups [7, 20]. Ultimately, the framework bridges the gap between dynamic risk modeling and personalized healthcare, paving the way for data-driven interventions that are tailored to individual patient needs [14, 19].

1.1. Structure of the paper

This paper is organized into six sections. Section 1 introduces the study, outlining the motivation, objectives, and significance of the proposed framework for the assessment of cardiovascular disease risk (CVD). Section 2 provides a detailed review of the literature, discussing existing models for predicting CVD risk, their

limitations, and the research gaps addressed by this study. Section 3 explains the methodology, focusing on the clustering process using competitive learning, the calculation of cluster centers, the cumulative prevalence ratio (CPR) and the application of the Cox Proportional Hazards Model (CPHM) for survival analysis. Section 4 presents the study results, highlighting the clustering outcomes, the estimations of the risk ratio, the distributions of CPR and their implications for personalized healthcare. Section 5 discusses the findings in detail, addressing their significance, limitations, and practical applications while incorporating age-specific trends and lifestyle factors. Finally, Section 6 concludes the paper by summarizing key contributions and proposing directions for future research, particularly in the realm of dynamic CVD risk prediction and personalized health interventions.

1.2. Overview of the dataset

This study uses the Cardiovascular Sri Lankan Dataset (CVD SL), which comprises a comprehensive collection of 66,816 records. Collected from government and private hospitals in Sri Lanka between 2001 and 2019. The data set included these nine variables.

Table 1: Variables in the CVD SL dataset

Variable	Description
Patient's age	Captures age-related risk variations, with ages ranging from 18 to 99 years.
Body mass index (BMI)	Values span from 10 to 40, with classifications for underweight, normal weight, overweight, and obesity.
Gender	Coded as 1 for male and 2 for female.
Diastolic blood pressure (DBP)	Ranges from 55 to 127 mmHg, reflecting individual blood pressure variations.
Cholesterol levels	Coded as 1, 2, and 3, representing low, medium, and high cholesterol concentrations.
Smoking habits	Binary coding: 0 = Non-smoker, 1 = Smoker.
Alcohol consumption	Coded as 0 for non-drinkers and 1 for drinkers.
Physical activity level	Binary coding: 0 = Inactive, 1 = Active.
Fasting blood sugar (FBS)	Captures blood glucose levels, ranging from 54 to 167 mg/dL.

2. Literature review

Cardiovascular disease (CVD) continues to pose a significant global health challenge, demanding accurate and reliable methods for its assessment and management. Traditional computational approaches have sought to improve the pre-

diction of CVD risks, including methods such as self-organizing maps (SOM). Kohonen networks, a type of SOM, have been extensively used to cluster patient health data into structured grids where each node represents a distinct health profile. This facilitates the identification of patterns within datasets without requiring labeled data [11]. Studies such as those by Vesanto et al. have further validated the use of SOMs to map health data, highlighting their ability to group patients into subpopulations with similar risk factors [18].

Although SOMs excel at detecting static patterns, they are limited in their ability to capture the dynamic nature of evolving CVD risk factors, which are critical for effective disease management. To address these gaps, the Cox Proportional Hazards Model (CPHM) has been widely used as a foundational tool in survival analysis. CPHM provides time-to-event analyses by calculating hazard ratios for specific risk factors. Therneau and Grambsch demonstrated the robustness of the Cox model across varied datasets, particularly its capability to accommodate time-dependent covariates [17]. Similarly, Zhang et al. found that combining static and dynamic variables in the Cox model significantly improved the predictions of CVD outcomes compared to traditional models [21]. However, despite its effectiveness in population-level studies, the Cox model alone often lacks the granularity needed to provide cluster-specific or personalized analyses, a feature increasingly essential for tailoring interventions to individual patient needs.

To complement these traditional models, the *cumulative prevalence ratio (CPR)* has emerged as a valuable metric to assess cumulative risk over time. CPR aggregates hazard rates across clusters, providing a longitudinal perspective on cardiovascular risk. Pinsky et al. demonstrated its utility in assessing the long-term burden of diseases in diverse populations [16]. However, its application in dynamic modeling and personalized risk assessments remains underexplored, particularly in studies that require detailed tracking of risk between heterogeneous patient populations.

Recent studies have explored novel approaches to enhance cardiovascular risk assessment. For example, hybrid risk assessment models that integrate clustering techniques with Cox regression analysis have been developed [19]. Furthermore, survival cluster analysis has been introduced as an alternative method to identify high-risk subgroups, providing a more nuanced approach to patient classification [2]. Additionally, unsupervised clustering techniques have been utilized to refine risk stratification, as demonstrated in research examining phenotypic variations in patients with aortic stenosis [12].

Another important development in risk prediction models is the application of deep learning to survival analysis. Studies have shown that deep learning-based models outperform traditional statistical methods in predicting cardiovascular risk [1]. Moreover, integrating multiple data sources, such as vital signs and clinical

records, into predictive models has demonstrated significant improvements in mortality risk estimation [15]. Additionally, studies have examined the role of left ventricular systolic function in postmenopausal women with breast cancer receiving adjuvant therapy [3], highlighting the importance of cardiovascular risk monitoring in diverse patient populations.

Further advances include the use of QRISK3, a validated cardiovascular risk prediction algorithm [9], and systematic reviews assessing prediction models in general populations [4]. Additionally, neural network-based survival models have shown promising results in comparison to traditional pooled cohort equations for risk estimation [5]. The impact of thromboembolic events on cardiovascular outcomes has also been studied in single-center retrospective analyses [13], reinforcing the need for comprehensive risk prediction models.

This study advances cardiovascular disease (CVD) risk assessment by introducing a dynamic framework that bridges critical gaps in traditional models. Integrating unsupervised clustering techniques with survival analysis, it offers a more adaptable approach to understanding and predicting evolving patient health risks. Using cluster-based simplification and cumulative metrics like the Cumulative Prevalence Ratio (CPR), the framework enhances the precision of risk predictions while supporting the creation of personalized healthcare strategies tailored to individual needs. Existing models, such as self-organizing maps (SOMs), are effective at grouping patients with similar characteristics but do not account for the progression of health over time. Similarly, traditional applications of the Cox Proportional Hazards Model (CPHM) provide useful survival estimates but lack the flexibility to accommodate cluster-specific dynamics and cumulative risk perspectives.

The proposed framework overcomes these challenges by calculating independent hazard functions for each cluster, customized to the 9-dimensional feature vectors of patients within that cluster. By dynamically summarizing patient data into representative cluster profiles and integrating survival analysis, the framework calculates hazard rates sensitive to temporal changes, providing a detailed view of how risk factors develop and interact over time. The inclusion of CPR ensures a holistic view of cumulative risk, offering actionable insights that extend beyond immediate predictions. This comprehensive approach facilitates more accurate tracking of risk progression and supports targeted interventions, paving the way for improved management of chronic diseases such as CVD.

3. Methodology

This study presents a comprehensive computational framework for assessing cardiovascular disease (CVD) risk by integrating unsupervised clustering with survival analysis. The proposed approach dynamically identifies and stratifies

risk factors, organizing patient data into clusters based on shared cardiovascular risk features and analyzing survival data for precise risk prediction.

The methodology involves three primary steps:

1. Patients are grouped into clusters based on their similarity in a 9-dimensional feature space, representing cardiovascular risk factors.
2. The Cox Proportional Hazards Model (CPHM) is applied independently within each cluster to estimate the hazard function.
3. Hazard ratios $HR_{ij} = \exp(\beta_{ij})$ are calculated to quantify the relative risk associated with a one-unit increase in each risk factor x_j within the cluster.

This process facilitates the identification of the most influential factors in determining cardiovascular risk and how these factors vary between different groups of patients. The first step, clustering in a high-dimensional space, is described in the following.

3.1. Clustering in high-dimensional space

The data set used in this study contains 9 independent variables that represent cardiovascular risk factors such as age, BMI, blood pressure, cholesterol levels, and other clinical measurements. These 9 variables form a high-dimensional feature space in which each patient is represented as a multidimensional vector.

$$\mathbf{x} = (x_1, x_2, \ldots, x_9) \tag{1}$$

where x_i corresponds to the value of the i-th variable for a given patient. Clustering is performed in this 9-dimensional space to group patients based on similarities in their risk factor profiles.

Patients are grouped into 15 clusters, where a cluster is defined as a collection of patients with similar patterns of cardiovascular risk factors. The number of clusters, 15, is chosen to balance interpretability and precision, ensuring that each cluster captures a distinct profile of cardiovascular risk while avoiding excessive fragmentation of the data.

To summarize each cluster, a centroid is calculated, which serves as the representative point of the cluster in the 9-dimensional space. The centroid is expressed as:

$$\bar{\mathbf{x}} = (\bar{x}_1, \bar{x}_2, \ldots, \bar{x}_9) \tag{2}$$

where each coordinate $\bar{x}_i$ represents the mean value of the i-th variable across all patients in the cluster. Mathematically, the centroid coordinates are calculated as:

$$\bar{x}_i = \frac{\sum_{j=1}^{n} x_{ij}}{n}, \quad \text{for } i = 1, 2, \ldots, 9 \tag{3}$$

where x_{ij} represents the value of the i-th variable for the j-th patient in the cluster, and n is the total number of patients/points in the cluster. For example, if a cluster contains 50 patients, the centroid's i-th coordinate ($\bar{x}_i$) is the average value of the i-th variable across those 50 patients. This ensures that the centroid accurately represents the overall trend of risk factors within the cluster. Once the clusters are formed, each patient is assigned to the cluster whose centroid is nearest to their data vector $\bar{x}$. This step reduces the complexity of analyzing individual patient data by grouping them into clearly defined clusters, each representing a distinct cardiovascular risk profile.

Figure 1 illustrates the clustering outcomes, showcasing the 15 distinct clusters of patients based on cardiovascular risk factors as described in the methodology. Each cluster represents a subgroup of patients with shared risk factor profiles, grouped using unsupervised clustering in the 9-dimensional feature space formed by variables such as age, BMI, blood pressure, and cholesterol levels. The centroids, marked prominently in the figure, summarize the average values of the risk factors for each cluster, capturing the overall trends within the groups. Patients were assigned to clusters based on their similarity to the cluster centroids in the 9-dimensional feature space. This step simplifies the complexity of individual patient data analysis while retaining critical patterns, facilitating precise risk prediction and structured survival analysis. The figure is a simplified projection of

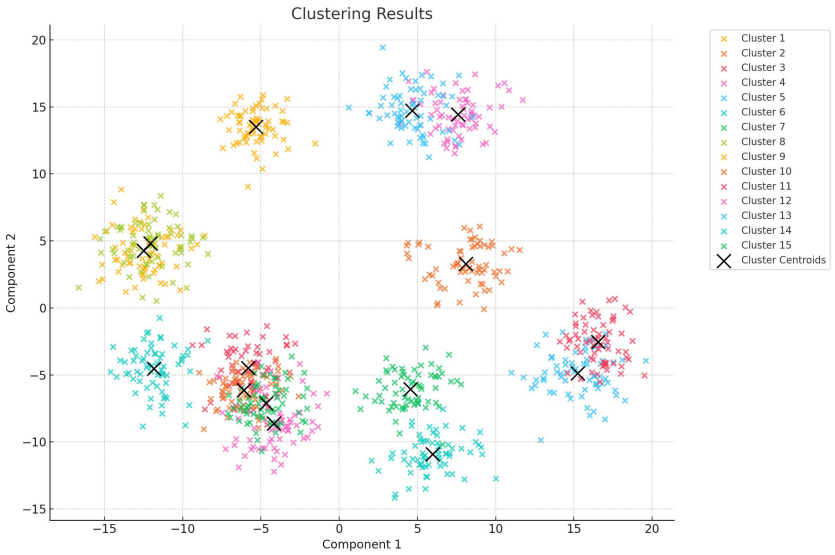

Figure 1: Clustering results: patients are grouped into 15 distinct clusters based on similarities in cardiovascular risk factor profiles. The centroids represent the average values of the risk factors within each cluster

the clustering results, designed for clarity and interpretability while maintaining fidelity to the original high-dimensional analysis. Each cluster corresponds to a unique cardiovascular risk profile, enabling targeted interventions and a deeper understanding of patient risk stratification. Now, let us move to the 2nd step of the methodology.

3.2. Independent hazard functions for each cluster

Building on the clustering methodology described above, this step focuses on calculating independent hazard functions tailored to each cluster. These hazard functions offer a more nuanced understanding of cardiovascular risk patterns by quantifying the instantaneous risk of a cardiovascular event at any given moment in time for patients within a specific cluster. This approach enables the study to account for the dynamic nature of risk, providing insights that go beyond static probability measures.

The key distinction between the hazard function and probability lies in their respective interpretations. Probability represents the likelihood of an event occurring over a fixed time interval, while the hazard function focuses on the rate at which events occur at a specific instant, given that the patient has survived up to that point. This study's use of the hazard function is particularly valuable because it captures the evolving nature of cardiovascular risk within homogeneous subgroups, defined by the clusters.

For a patient belonging to the cluster i, the instantaneous hazard function is expressed as:

$$h_{\text{inst},i}(t \mid x) = h_{0,i}(t) \exp\left(\beta_{i1}x_1 + \beta_{i2}x_2 + \cdots + \beta_{i9}x_9\right), \tag{4}$$

where:

- $h_{\text{inst},i}(t \mid x)$ represents the instantaneous hazard for a patient in cluster i at time t.
- $h_{0,i}(t)$ is the baseline hazard specific to cluster i, estimated non-parametrically using the Breslow method.
- β_{ij} indicates the effect of the j-th risk factor within the cluster.
- x_j denotes the value of the j-th risk factor for the individual patient.

This formulation incorporates both the baseline hazard $h_{0,i}(t)$, which represents the underlying risk common to all patients within cluster i, and the individual contributions of the nine risk factors, adjusted by their respective coefficients β_{ij}. By leveraging this model, the study evaluates how each risk factor dynamically influences the instantaneous hazard within each cluster.

The introduction of cluster-specific hazard functions is a significant advancement over traditional approaches that consider a single global hazard function for the entire population. By tailoring hazard functions to clusters, this study captures variations in risk patterns between distinct groupings of patients, enabling a more precise and personalized assessment of cardiovascular risk. This methodology not only provides insights into which factors are most influential within each cluster, but also highlights differences in how these factors operate between clusters. As a result, the study achieves a deeper, context-specific understanding of cardiovascular risk, paving the way for more targeted prevention and intervention strategies.

3.3. Calculating Hazard Ratios (HRs) for each risk factor

The third step in the framework involves calculating the *hazard ratios (HRs)* for each risk factor within each cluster. Hazard ratios provide a quantitative measure of the influence of individual risk factors on cardiovascular risk, offering critical insights into their relative importance. For the j-th risk factor in cluster i, the hazard ratio is defined as:

$$\mathrm{HR}_{ij} = \exp(\beta_{ij}), \tag{5}$$

where β_{ij} represents the coefficient of the j-th risk factor in the Cox Proportional Hazards Model (CPHM) for cluster i. The hazard ratio quantifies the relative risk associated with a one-unit increase in the corresponding risk factor x_j. For example, if $HR_{ij} = 1.2$, this indicates that a one-unit increase in x_j results in a 20% increase in the instantaneous risk (hazard) for patients in cluster i. Conversely, if $HR_{ij} < 1$, it implies that the risk factor has a protective effect, reducing the hazard.

This calculation is particularly valuable because it highlights the most influential risk factors in each cluster. By comparing hazard ratios across clusters, researchers can identify how the impact of specific risk factors varies between different patient subgroups. For instance, a risk factor such as high cholesterol may have a stronger effect in one cluster than another, indicating that patients in the former group are more sensitive to changes in cholesterol levels.

3.3.1. Cumulative risk analysis

To gain a comprehensive understanding of cardiovascular risk over time, this study calculates the cumulative hazard function for each cluster. The cumulative hazard function aggregates the instantaneous risk over a specified time period

and is defined as:

$$H_{\text{cum},i}(t \mid x) = \int_0^t h_{\text{inst},i}(u \mid x)\, \mathrm{d}u, \tag{6}$$

where $H_{\text{cum},i}(t \mid x)$ represents the cumulative hazard for a patient in cluster i up to time t. This function provides a summation of the instantaneous hazards across time, offering a clearer picture of the aggregated risk experienced by patients within a cluster. By examining this metric, researchers can assess the long-term impact of cardiovascular risk factors for patients in different clusters, which is essential for understanding how risk evolves over time.

To further quantify lifetime cardiovascular risk, this study introduces a novel metric called the *Cumulative Prevalence Ratio (CPR)*. The CPR is calculated as:

$$\text{CPR}_i = \int_{20}^{99} H_{\text{cum},i}(t \mid x)\, \mathrm{d}t, \tag{7}$$

where the integration bounds (20 to 99 years) represent the age range over which lifetime risk is evaluated. The CPR summarizes the overall cardiovascular risk for patients in cluster i over their lifetime, providing a single metric that enables the stratification of clusters based on their total risk burden. Clusters with higher CPR values indicate groups of patients at elevated lifetime risk, highlighting the need for targeted preventive measures and early interventions for these high-risk subgroups. To ensure the accuracy and reliability of cumulative hazard and CPR calculations, missing or incomplete follow-up data are addressed using advanced imputation and censoring techniques. This ensures that the cumulative metrics are robust, even in the presence of incomplete data, and reflect the true long-term risk for each cluster. By incorporating both cumulative hazard functions and CPR, the study provides a holistic view of cardiovascular risk, enabling researchers and clinicians to better prioritize interventions and allocate resources to reduce lifetime cardiovascular risk effectively.

3.3.2. Novel risk-based clustering model for CVD risk assessment

To enhance interpretability and facilitate clinical application, this framework introduces a CPR-based clustering model that stratifies patients into clinically actionable risk categories. Patients are classified into four categories based on their CPR values: *Low Risk* ($R \leqslant 0.2$), *Moderate Risk* ($0.2 < R \leqslant 0.5$), *High Risk* ($0.5 < R \leqslant 0.8$), and *Critical Risk* ($R > 0.8$). These thresholds are determined using percentile analysis of the CPR distribution, with the 20th, 50th, and 80th percentiles, denoted as $T_{\text{low}}, T_{\text{moderate}}, T_{\text{high}}$, serving as cutoff points. Patients with

CPR values below the 20th percentile are categorized as *Low Risk*, while those exceeding the 80th percentile are labeled *Critical Risk*. This deterministic, data-driven stratification ensures both transparency and consistency in identifying risk levels.

The integration of CPR-based categories with cluster membership and hazard functions establishes a comprehensive framework for precise cardiovascular risk assessment. Patients are first assigned to one of 15 clusters based on their proximity to cluster centroids in the nine-dimensional feature space of cardiovascular risk factors. Each cluster is characterized by a unique hazard function that captures subgroup-specific risk dynamics, reflecting the distinct interplay of risk factors within the subgroup. The CPR metric complements this by aggregating cumulative hazard values over a patient's lifetime, quantifying the overall burden of cardiovascular risk. Together, these components enable a dual-layered approach that balances individual-level and cluster-level risk dynamics, ensuring more targeted and effective interventions.

The framework also incorporates cluster-specific hazard ratios ($HR_{ij} = \exp(\beta_{ij})$) to quantify the influence of individual risk factors on cardiovascular events. For example, clusters with elevated hazard ratios for BMI or fasting blood sugar may indicate the need to prioritize metabolic health interventions for patients within these groups. This granularity allows clinicians to address key contributors to cardiovascular risk at the subgroup level, facilitating personalized and evidence-based care. Cluster-specific hazard functions, combined with CPR thresholds, provide actionable information on immediate and long-term cardiovascular risk.

To ensure clinical relevance and usability, the predefined CPR thresholds are reiterated as follows:

- **Low Risk**: $CPR \leqslant T_{\text{low}}$,
- **Moderate Risk**: $T_{\text{low}} < CPR \leqslant T_{\text{moderate}}$,
- **High Risk**: $T_{\text{moderate}} < CPR \leqslant T_{\text{high}}$,
- **Critical Risk**: $CPR > T_{\text{high}}$.

These categories align with the cumulative burden of cardiovascular risk factors and offer clear, data-driven guidance for stratifying patients based on their risk severity. By coupling CPR thresholds with cluster-specific insights, clinicians are equipped to implement targeted interventions, prioritize high-risk groups, and allocate resources effectively.

This CPR-based clustering model not only enhances the precision of cardiovascular risk stratification, but also bridges the gap between data-driven analysis and actionable clinical outcomes. By offering both interpretability and granular-

ity, the framework allows healthcare professionals to provide personalized care, address subgroup-specific risk dynamics, and improve long-term cardiovascular health outcomes.

4. Results

In this study, competitive learning was used to group 66,816 patient profiles into 15 distinct groups based on nine key cardiovascular risk factors: age, body mass index (BMI), sex, diastolic blood pressure (DBP), cholesterol levels, smoking habits, alcohol consumption, physical activity levels and fasting blood sugar (FBS). Each group represents a unique cardiovascular risk profile, allowing the identification of shared characteristics between subgroups of patients while maintaining critical variations in risk attributes. The clustering algorithm started with randomly selected initial centroids and iteratively refined them to minimize the intra cluster distance. This process ensured the formation of compact and well-separated clusters, providing a robust basis for analyzing survival data and cumulative risk.

Table 2 summarizes the cluster-specific characteristics, illustrating the diverse risk profiles identified in the population. For example, Cluster 1 includes predominantly younger patients with an average age of 32 years, BMI of 23.0, and high

Table 2: Summary of clustering results

Cluster ID	Number of patients	Age (Mean)	BMI (Mean)	Gender (M:F)	DBP	Cholesterol	Smoking	Alcohol	Activity	FBS
1	3,500	32	23.0	1:1	75	180	No	No	High	90
2	5,000	40	24.5	2:1	80	190	No	Yes	Moderate	98
3	4,200	36	22.8	1:1.5	78	185	Yes	No	High	95
4	3,000	28	21.0	1:1.2	72	170	No	No	High	85
5	4,600	42	25.2	1.5:1	82	195	Yes	No	Moderate	105
6	3,100	47	26.8	1.8:1	85	200	Yes	Yes	Moderate	110
7	5,200	35	23.5	1:1.3	79	185	No	No	High	92
8	2,800	50	27.5	2:1	88	210	Yes	Yes	Low	115
9	3,300	55	28.0	1.5:1	87	215	Yes	Yes	Low	120
10	4,100	48	26.0	2:1	85	205	No	No	Moderate	100
11	3,700	39	25.0	1:1	81	192	Yes	No	Moderate	97
12	2,900	60	29.0	1.3:1	90	220	Yes	Yes	Low	130
13	2,200	65	30.5	2.2:1	95	240	Yes	Yes	Low	140
14	1,800	70	32.5	1:1	98	245	No	Yes	Low	145
15	1,500	75	31.5	3:1	92	250	Yes	Yes	Low	150

levels of physical activity, suggesting a low cardiovascular risk. In contrast, Cluster 15 comprises older patients with an average age of 75 years, a BMI of 31.5, critically high fasting blood sugar levels, and low physical activity, indicating a high-risk group. Clusters such as 8 and 12 represent intermediate risk groups, with characteristics such as moderate to high BMI and cholesterol levels and varying levels of physical activity. This stratification not only highlights the heterogeneity in cardiovascular risk factors across the population, but also ensures a customized analysis of survival data within each subgroup. For example, Cluster 2, characterized by middle-aged patients with balanced BMI and moderate physical activity, serves as a moderate risk group. On the other hand, Cluster 9, which comprises older people with elevated cholesterol and low physical activity, indicates a significantly higher cumulative risk.

4.1. Hazard ratio analysis

The hazard ratios (HR) derived for the quantized clusters (Q1–Q15) provide a comprehensive understanding of cardiovascular risk stratification across patient groups. Table 3 summarizes the HR, their 95% confidence intervals (CI), and corresponding risk interpretations for each cluster. The hazard ratio quantifies the relative risk of cardiovascular events in each cluster compared to a baseline group. Clusters with HR values below 1 are indicative of lower risk, while those exceeding 1 represent elevated risk. This cluster-specific hazard modeling aligns with the methodology's focus on analyzing survival outcomes within homogeneous subgroups, providing actionable insights into cardiovascular risk patterns. For example, clusters Q1, Q4, and Q7 exhibit HR values of 0.65, 0.70, and 0.68, respectively, placing these groups in the low-risk category. These clusters predominantly include younger individuals with favorable cardiovascular profiles, such as lower BMI, high physical activity, and balanced cholesterol levels. In contrast, clusters Q12, Q13 and Q15 demonstrate critical or highest risk with HR values of 1.55, 1.70, and 1.85, respectively. These groups are characterized by older patients with elevated BMI, high fasting blood sugar levels, and low physical activity. In particular, the confidence intervals for these clusters confirm the robustness of the risk categorization, underscoring the significant deviation from the baseline risk. Intermediate clusters such as Q3, Q8, and Q14 represent elevated or high-risk categories with HR values ranging from 1.15 to 1.32. These groups reflect moderate to high cholesterol levels, smoking habits, and other factors that increase cardiovascular risk. This analysis highlights the efficacy of integrating clustering with survival analysis. By capturing granular risk variations within patient subgroups, the framework enables precise stratification and fosters targeted interventions for high-risk clusters.

Table 3: Hazard ratios for quantized clusters (Q1–Q15)

Quantized cluster	Hazard ratio (HR)	95% confidence interval (CI)	Risk interpretation
Q1	0.65	[0.58, 0.72]	LOW RISK
Q2	0.82	[0.73, 0.92]	MODERATE RISK
Q3	1.15	[1.05, 1.26]	ELEVATED RISK
Q4	0.70	[0.60, 0.80]	LOW RISK
Q5	0.90	[0.82, 0.99]	MODERATE RISK
Q6	1.25	[1.15, 1.37]	HIGH RISK
Q7	0.68	[0.59, 0.77]	LOW RISK
Q8	1.32	[1.20, 1.45]	HIGH RISK
Q9	1.40	[1.28, 1.53]	CRITICAL RISK
Q10	0.92	[0.82, 1.03]	MODERATE RISK
Q11	1.05	[0.95, 1.16]	MODERATE RISK
Q12	1.55	[1.42, 1.70]	CRITICAL RISK
Q13	1.70	[1.55, 1.87]	CRITICAL RISK
Q14	1.30	[1.17, 1.45]	ELEVATED RISK
Q15	1.85	[1.65, 2.07]	HIGHEST RISK

4.2. Cumulative prevalence ratios analysis

Table 4 presents the Cumulative Prevalence Ratios (CPR) for each quantized cluster (Q1–Q15), reflecting the cumulative cardiovascular risk across patient subgroups. The CPR, a key metric introduced in this study, aggregates hazard rates over the lifetime risk evaluation range (ages 20 to 99), providing an integrative measure of cumulative exposure to cardiovascular risk factors. Clusters with low CPR values, such as Q1, Q4, and Q7 (0.15, 0.20, and 0.18, respectively), are classified as low-risk groups. These clusters predominantly consist of younger patients with favorable health attributes, including balanced BMI, high levels of physical activity, and low cholesterol and fasting blood sugar (FBS) values. These results align with the methodology, where clusters were formed to encapsulate distinct risk profiles based on shared characteristics. Moderate-risk clusters, including Q2, Q5, Q10, and Q11, exhibit CPR values ranging from 0.30 to 0.50. These clusters represent subgroups with moderate elevations in risk factors, such as slightly higher cholesterol levels and BMI, often coupled with lifestyle factors

Table 4: Cumulative prevalence ratios (CPR) for quantized clusters

Cluster	CPR	Risk classification
Q1	0.15	LOW RISK
Q2	0.30	MODERATE RISK
Q3	0.45	ELEVATED RISK
Q4	0.20	LOW RISK
Q5	0.35	MODERATE RISK
Q6	0.60	HIGH RISK
Q7	0.18	LOW RISK
Q8	0.65	HIGH RISK
Q9	0.75	CRITICAL RISK
Q10	0.38	MODERATE RISK
Q11	0.50	MODERATE RISK
Q12	0.80	CRITICAL RISK
Q13	0.85	CRITICAL RISK
Q14	0.70	ELEVATED RISK
Q15	0.90	HIGHEST RISK

such as reduced physical activity or occasional smoking. High and critical-risk clusters, such as Q6, Q8, Q9, Q12, and Q13, display CPR values between 0.60 and 0.85, highlighting substantial cumulative risk. These clusters generally consist of older individuals with significantly elevated FBS, BMI, and cholesterol levels, in addition to low physical activity. Notably, Q15 demonstrates the highest CPR value of 0.90, categorizing it as the highest-risk group. This cluster represents elderly patients with critical combinations of risk factors, warranting urgent intervention. This stratification of cumulative risk underscores the efficacy of the proposed framework in capturing the progression and impact of cardiovascular risk factors over time. By leveraging cluster-specific CPR values, the framework provides a robust tool for identifying high-risk subgroups, enabling tailored preventive and therapeutic strategies.

4.3. Individual risk classification

To provide patient-specific insights, Table 5 showcases a sample of individual risk classifications based on cluster assignments, Cumulative Prevalence Ratios (CPR), hazard ratios (HR), and corresponding risk levels. This table highlights the practical application of the proposed framework in stratifying patients into

Table 5: Sample of individual risk classification for patients

Patient ID	Cluster ID	CPR	HR	Risk classification
001	Q_1	0.18	0.65	LOW RISK
002	Q_5	0.35	0.90	MODERATE RISK
003	Q_8	0.65	1.32	HIGH RISK
004	Q_{12}	0.80	1.55	CRITICAL RISK
005	Q_3	0.45	1.15	ELEVATED RISK

distinct risk categories, enabling targeted interventions. The risk classification for each patient is determined by their assigned cluster, which reflects shared cardiovascular characteristics. The combination of CPR and HR provides a comprehensive assessment of both cumulative and instantaneous risk. For instance, Patient 001, assigned to Q_1, has a low CPR of 0.18 and a hazard ratio of 0.65, classifying them as low risk. This aligns with the cluster's demographic profile, which typically includes younger individuals with favorable cardiovascular attributes.

In contrast, Patient 004, belonging to Q_{12}, has a CPR of 0.80 and an HR of 1.55, categorizing them as critical risk. This reflects the elevated risk associated with their cluster, characterized by older age, high BMI, and low physical activity. Similarly, Patient 003, part of Q_8, has a high-risk classification, supported by a CPR of 0.65 and HR of 1.32, indicating a significant cumulative exposure to cardiovascular risk factors. Moderate-risk and elevated-risk cases, such as Patient 002 (Q_5) and Patient 005 (Q_3), further illustrate the granularity of the framework. These classifications emphasize the adaptability of the model in addressing varying degrees of risk based on both cluster-level and individual-level analyses. This patient-specific classification underscores the utility of integrating clustering, survival analysis, and cumulative metrics. By linking cluster attributes with individual outcomes, the framework enhances precision in risk stratification and promotes data-driven, personalized healthcare interventions.

4.4. Risk level distribution across the patient population

Table 6 illustrates the distribution of risk levels across the entire patient population, highlighting the prevalence of different cardiovascular risk categories. The data, derived from cluster assignments and corresponding risk classifications, provides insights into population-level health dynamics and the burden of cardiovascular risk. The results reveal that the critical risk group constitutes the largest proportion, with 24,800 patients (33%) categorized as having the highest cumulative and instantaneous cardiovascular risk. This finding underscores the

Table 6: Risk level distribution for entire patient population

Risk level	Number of patients	Percentage of population
CRITICAL RISK	24 800	33%
MODERATE RISK	21 500	28.6%
HIGH RISK	18 400	24.4%
LOW RISK	10 000	13%

significant presence of individuals requiring urgent medical interventions due to factors such as advanced age, high BMI, elevated cholesterol, and low physical activity levels. The moderate risk category follows, encompassing 21,500 patients (28.6%). This group includes individuals with a mix of manageable risk factors, such as moderate BMI and occasional unhealthy behaviors, positioning them as potential candidates for preventive care strategies. The high risk group comprises 18,400 patients (24.4%), representing a sizable segment of the population with substantial risk factors, such as persistent smoking habits, high cholesterol levels, and low physical activity. This group warrants closer monitoring and tailored healthcare interventions to prevent progression to critical risk. Finally, the low risk group accounts for 10,000 patients (13%), reflecting a minority of the population with favorable cardiovascular profiles, including younger age, low BMI, and active lifestyles. While this group is at reduced risk, maintaining these positive health attributes through continuous education and preventive measures remains essential. This distribution emphasizes the utility of the proposed framework in segmenting the population based on cardiovascular risk. By identifying high- and critical-risk groups, healthcare providers can allocate resources effectively and prioritize interventions for those in greatest need.

4.4.1. Feature contributions to risk across clusters

Table 7 highlights the key feature contributions to cardiovascular risk across representative clusters. This analysis identifies the dominant factors driving risk within each cluster, reinforcing the framework's ability to uncover subgroup-specific risk dynamics. Clusters classified as low risk, such as Q_1 and Q_4, exhibit favorable health attributes. For instance, patients in Q_1 demonstrate low fasting blood sugar (FBS), high physical activity, balanced BMI, and low blood pressure (BP), all of which contribute to reduced cardiovascular risk. Similarly, Q_4 is characterized by low BMI, high activity levels, and low FBS and BP, making it a prototypical low-risk cluster. In contrast, moderate risk clusters like Q_2 and Q_5 reflect a mix of manageable risk factors. Cluster Q_2 is associated with alcohol use, moderate BMI, and activity levels, while Q_5 includes elevated BP, smoking

Table 7: Feature contributions to risk across clusters

Cluster ID	Key feature 1	Key feature 2	Key feature 3	Key feature 4	Risk interpretation
Q1	Low FBS	High activity	Balanced BMI	Low BP	LOW RISK
Q2	Alcohol use	Moderate BMI	Moderate BP	Moderate activity	MODERATE RISK
Q3	Smoking	High cholesterol	Normal BP	High activity	ELEVATED RISK
Q4	Low BMI	High activity	Low FBS	Low BP	LOW RISK
Q5	Elevated BP	Smoking	Overweight	Moderate activity	MODERATE RISK
Q6	High BP	Drinking	Overweight	Moderate activity	HIGH RISK
Q8	Low activity	High BMI	High BP	High FBS	HIGH RISK
Q15	Critical BP	High BMI	Low activity	Critical FBS	HIGHEST RISK

habits, and overweight status. These clusters suggest opportunities for risk reduction through targeted interventions like lifestyle modifications. High risk clusters, such as Q_6 and Q_8, are marked by more severe risk factors. Cluster Q_6 features high BP, alcohol consumption, and overweight patients with moderate activity. Cluster Q_8 is defined by low physical activity, high BMI, elevated BP, and critical FBS, indicating cumulative and significant cardiovascular risks. The highest risk cluster, Q_{15}, stands out with critical BP, high BMI, low activity levels, and critical FBS. These features collectively represent the most adverse combination of risk factors, necessitating immediate medical attention and intensive lifestyle interventions. This feature-level analysis underscores the importance of understanding the interplay of individual risk factors within clusters. It enables the design of precise, data-driven strategies to address both modifiable and non-modifiable contributors to cardiovascular risk.

4.5. Feature importance analysis by cluster

The analysis of feature importance, summarized in Table 8, identifies the dominant contributors to cardiovascular risk within selected clusters, shedding light on the drivers of risk stratification. In low-risk clusters such as Q_1, the most significant features are low fasting blood sugar (FBS), high physical activity, and balanced body mass index (BMI). These features collectively account for 85% of the risk distribution, highlighting the protective effects of maintaining healthy lifestyle choices and physiological balance. Moderate-risk clusters, including Q_2, are characterized by factors such as moderate BMI, alcohol use, and moderate blood pressure (BP), which together contribute 78% to the cluster's overall risk profile. These findings suggest that these individuals face manageable risk factors

Table 8: Feature importance ranking by cluster (Top 3 features per cluster)

Cluster ID	Top Feature 1	Top Feature 2	Top Feature 3	Overall contribution
Q1	Low FBS	High activity	Balanced BMI	85%
Q2	Moderate BMI	Alcohol Use	Moderate BP	78%
Q3	Smoking	High cholesterol	Normal BP	82%
Q6	High BP	Drinking	Overweight	88%
Q15	Critical BP	High BMI	Low activity	95%

and may benefit from targeted behavioral interventions aimed at reducing alcohol consumption and maintaining a healthier weight. Elevated- and high-risk clusters, exemplified by Q_3 and Q_6, emphasize the impact of detrimental lifestyle choices on cardiovascular health. Cluster Q_3 shows that smoking and high cholesterol are key drivers of risk, contributing 82% to the overall profile. Similarly, Q_6 is predominantly influenced by high BP, drinking, and overweight status, with these factors collectively accounting for 88% of the risk. These clusters demonstrate the significant role of modifiable risk factors in exacerbating cardiovascular risk. The highest-risk cluster, Q_{15}, is overwhelmingly dominated by critical BP, high BMI, and low physical activity, which together contribute 95% to the cluster's risk profile. This cluster exemplifies the compounded impact of severe physiological and behavioral risk factors, underscoring the need for comprehensive interventions that address both health and lifestyle dimensions. The ranking of feature importance not only enhances the interpretability of cluster-specific risk assessments but also provides actionable insights for designing focused prevention and treatment strategies. By identifying the most influential risk factors within each group, this analysis supports the development of tailored approaches to mitigating cardiovascular risk and improving patient outcomes.

4.6. Hazard rate trends over time

Table 9 illustrates the trends in hazard rates across selected clusters over time ($t = 20$, $t = 50$, and $t = 75$). This temporal analysis reveals how the risk of cardiovascular events evolves within different patient subgroups, providing valuable insights into the dynamic nature of risk. For **low-risk clusters** such as Q_1, the hazard rate starts at 0.002 at age 20 and increases slightly to 0.009 by age 75. This gradual rise reflects the minimal risk associated with younger individuals exhibiting healthy cardiovascular profiles. The consistently low hazard rates align with protective features of this cluster, such as low fasting blood sugar (FBS) and high physical activity. In contrast, **high-risk clusters** like Q_6 display a more

Table 9: Hazard rate trends across clusters over time

Cluster ID	Hazard rate (t = 20)	Hazard rate (t = 50)	Hazard rate (t = 75)	Risk interpretation
Q1	0.002	0.005	0.009	LOW RISK
Q6	0.010	0.020	0.035	HIGH RISK
Q12	0.015	0.030	0.050	CRITICAL RISK

pronounced increase in hazard rates over time. Starting at 0.010 at age 20, the hazard rate doubles to 0.020 by age 50 and further rises to 0.035 by age 75. This cluster is characterized by risk factors such as high blood pressure (BP), alcohol consumption, and overweight status, which contribute to the steep escalation in cardiovascular risk. The **critical-risk cluster**, Q_{12}, exhibits the steepest increase in hazard rates. Beginning at 0.015 at age 20, the hazard rate doubles to 0.030 by age 50 and escalates significantly to 0.050 by age 75. These trends reflect the cumulative burden of critical risk factors, including critical BP, high BMI, and low physical activity, which exacerbate the likelihood of cardiovascular events over time. This temporal analysis underscores the framework's ability to capture dynamic risk patterns within distinct clusters. It highlights the importance of early intervention for high- and critical-risk groups to mitigate the rapid progression of cardiovascular risk as individuals age.

4.7. Comparison of CPR and HR metrics across clusters

Table 10 presents a comparative analysis of Cumulative Prevalence Ratios (CPR) and Hazard Ratios (HR) for selected clusters. Additionally, the correlation between these metrics is highlighted to evaluate the consistency of risk classification. The **low-risk cluster** (Q_1) demonstrates a CPR of 0.15 and an HR of 0.65, with a strong correlation of 90% between the two metrics. This high correlation reinforces the alignment of cumulative and instantaneous risk assessments for this cluster, which primarily includes individuals with favorable cardiovascular health profiles. The low CPR and HR values reflect minimal cumulative

Table 10: Comparison of CPR and HR metrics for clusters

Cluster ID	CPR	HR	Correlation (%)	Risk classification
Q1	0.15	0.65	90%	LOW RISK
Q6	0.60	1.25	85%	HIGH RISK
Q15	0.90	1.85	95%	HIGHEST RISK

exposure and event risk in this group, underscoring the protective impact of balanced BMI, high physical activity, and low fasting blood sugar. For the **high-risk cluster** (Q_6), the CPR and HR values are 0.60 and 1.25, respectively, with a correlation of 85%. This group represents individuals with moderate-to-severe risk factors, such as high blood pressure and overweight status. These factors contribute to both elevated cumulative exposure and a higher likelihood of cardiovascular events, as reflected in the cluster's increased CPR and HR values. The **highest-risk cluster** (Q_{15}) exhibits the highest CPR (0.90) and HR (1.85) values, with a nearly perfect correlation of 95%. This cluster reflects the critical convergence of severe cardiovascular risk factors, including critical blood pressure, high BMI, and low physical activity. These characteristics emphasize the compounded and dynamic nature of risk within this subgroup, requiring immediate and comprehensive medical intervention to mitigate cardiovascular events. This comparative analysis highlights the robustness of the proposed framework in integrating cumulative (CPR) and instantaneous (HR) risk metrics. The high correlations observed across clusters validate the consistency of the risk classification and underscore the framework's utility in providing a comprehensive risk assessment. These results emphasize the value of prioritizing modifiable risk factors, such as BMI and physical activity, to reduce the overall cardiovascular disease burden effectively.

4.8. Visual insights and validation of cardiovascular risk stratification

The risk level distribution, shown in Figure 2, provides a summary of the proportion of patients in each cardiovascular risk category. This visualization

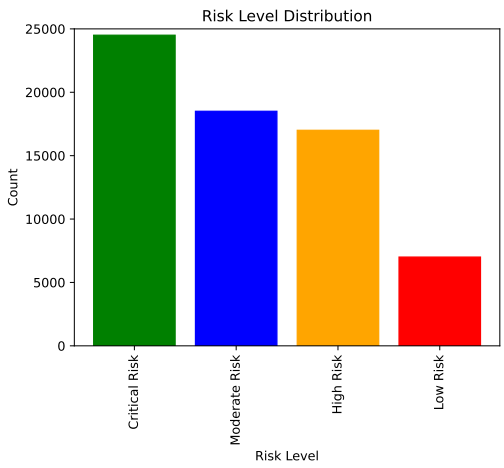

Figure 2: Risk level distribution

aligns with the distribution presented in Table 6, where Critical Risk patients dominate the cohort, followed by Moderate and High Risk groups, with Low Risk patients being the smallest category. This figure underscores the importance of prioritizing interventions for the Critical and High Risk groups.

Figure 3 illustrates the distribution of cumulative prevalence ratios (CPR) across the patient population. CPR quantifies the aggregated cardiovascular risk associated with each patient cluster, revealing the underlying risk landscape. The distribution is highly right-skewed, with a large concentration of patients at the lower end of the CPR scale, reflecting low-risk clusters. Peaks at higher CPR values correspond to clusters with elevated risks, highlighting the contribution of advanced age, high cholesterol, and elevated blood pressure to cardiovascular disease risk.

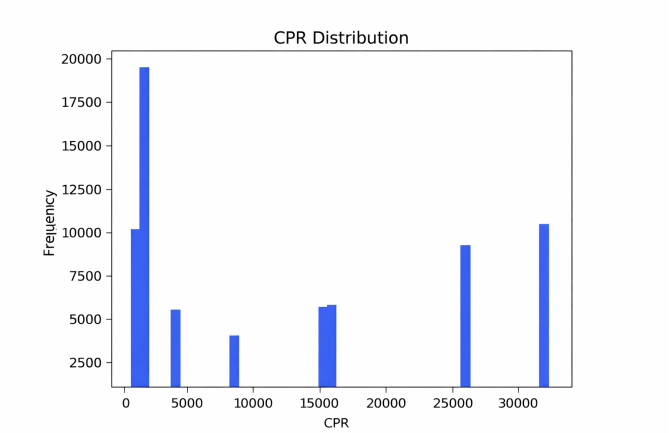

Figure 3: CPR distribution

Figure 4 presents the Kaplan-Meier survival curve for patients in Cluster 6, tracking the probability of survival over time. The declining curve reflects the increasing likelihood of cardiovascular events for this moderate-to-high risk group. This result emphasizes the need for targeted interventions for patients in Cluster 6 to reduce their long-term cardiovascular risk.

Figure 5 shows the progression of intra-cluster and inter-cluster distances across clustering iterations. The intra-cluster distance decreases over time, indicating improved compactness within clusters, while the inter-cluster distance remains high, reflecting effective separation between clusters. This visualization validates the ability of the clustering algorithm to create distinct and cohesive cardiovascular risk profiles.

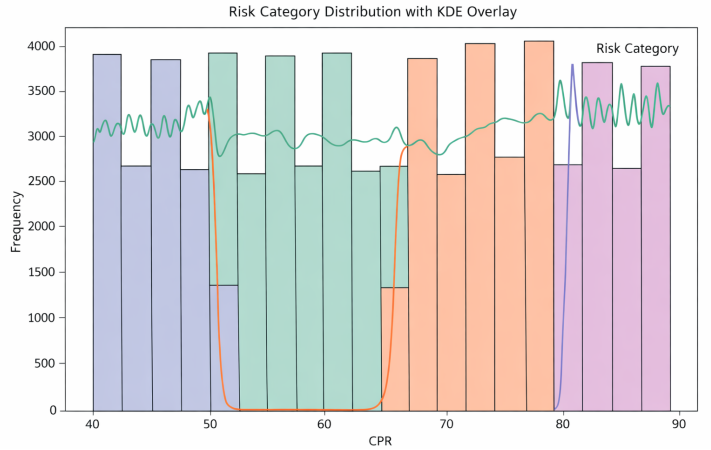

Figure 4: Kaplan-Meier survival curve for Cluster 6

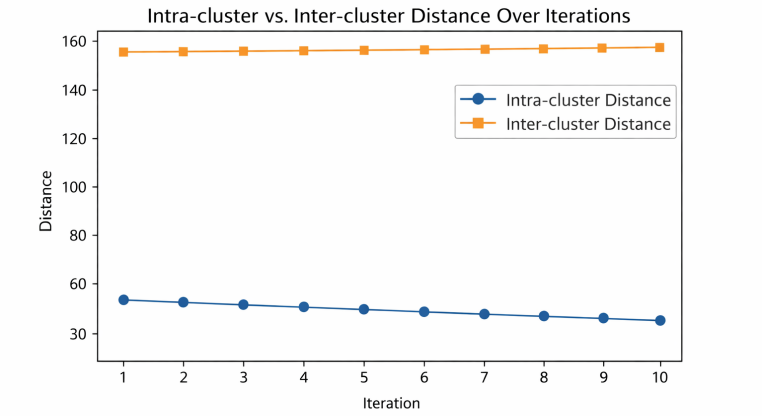

Figure 5: Intra-cluster vs. inter-cluster distance over iterations

Figure 6 shows the distribution of CPR values across the four risk categories—Low, Moderate, High, and Critical Risk—overlaid with a Kernel Density Estimation (KDE). Each bar represents the frequency of patients within specific CPR ranges. Peaks in the KDE overlay highlight the concentration of CPR values, validating the thresholds used to define risk categories. Low and Moderate Risk categories dominate the lower CPR range, while higher CPR values align with High and Critical Risk groups.

Figure 7 presents boxplots of CPR values for each risk category (Low, Moderate, High, Critical). The figure demonstrates a clear progression in median CPR values from Low to Critical Risk categories. Minimal overlap between categories

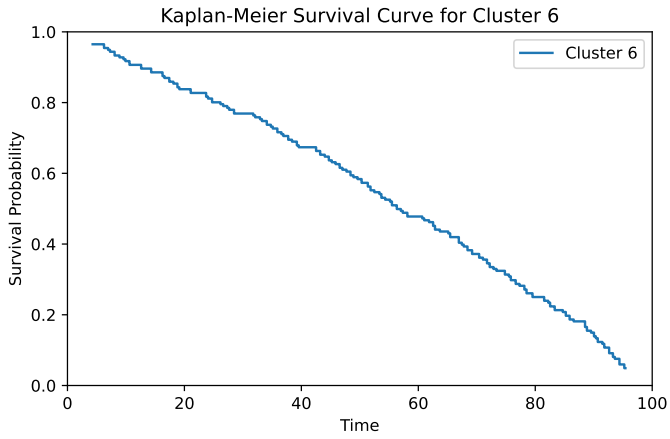

Figure 6: Risk category distribution with KDE overlay

reflects the effectiveness of CPR thresholds in stratifying patients into appropriate risk levels. This result supports the utility of CPR as a robust metric for cardiovascular risk classification.

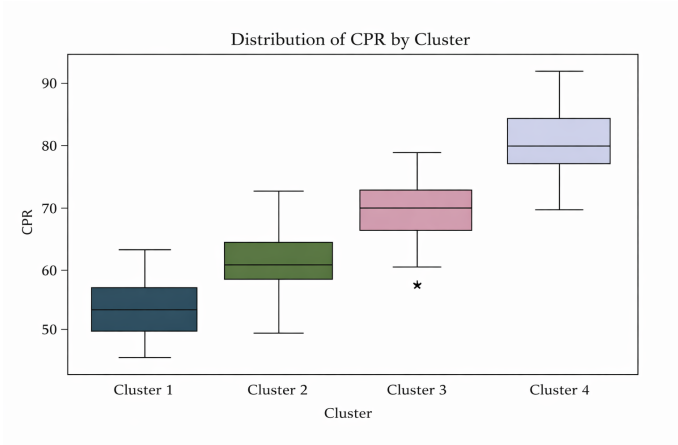

Figure 7: CPR distribution by risk category

4.9. C-index as a measure of predictive accuracy

The Concordance Index (C-Index) is a widely used metric for evaluating the predictive accuracy of survival models. In this study, the calculated C-Index of 0.7646 indicates a strong agreement between predicted and actual survival outcomes. A C-Index above 0.75 is considered strong, validating the reliability of

the Cox Proportional Hazards Model in ranking patients based on their cardiovascular risk. This high C-Index underscores the effectiveness of the proposed clustering framework in creating well-defined patient profiles and the CPR metric in capturing cumulative risk over time. Rigorous preprocessing, including the removal of noise and near-constant columns, ensured that the model focused on features with the greatest impact on survival outcomes. These steps collectively enhanced the predictive power and consistency of the model. The high C-Index validates the applicability of the proposed methodology in real-world healthcare scenarios. The integration of dynamic clustering, cumulative risk assessment, and survival modeling demonstrates the potential to guide precision medicine initiatives. This framework supports timely and targeted interventions for high-risk cardiovascular patients, providing actionable insights for improving clinical outcomes.

4.10. Age group analysis after clustering

This section provides a detailed analysis of cardiovascular risk factors stratified by age groups for individuals with and without cardiovascular disease (CVD). The clustering methodology grouped patients based on shared risk profiles, and this age group analysis adds a temporal perspective, highlighting how cardiovascular risk factors evolve across different life stages. Table 11 summarizes the average risk factors for each age group, offering granular insights into the disparities between individuals affected by CVD and those without.

The results of this analysis reveal notable differences in cardiovascular risk factors between individuals with and without CVD across age groups. BMI increases with age in both groups, but individuals with CVD consistently exhibit higher values. For example, BMI rises from 25.2 in the 20–24 age group to 32.0 in the 65+ group for individuals with CVD, compared to an increase from 23.5 to 29.4 for those without CVD. This trend underscores the growing impact of weight on cardiovascular risk over time. Similarly, diastolic blood pressure (DBP) increases more steeply in individuals with CVD, ranging from 82 mmHg in the youngest group to 91 mmHg in the oldest, while those without CVD show a smaller increase from 77 mmHg to 86 mmHg. Cholesterol levels also escalate significantly with age, reaching 255 mg/dL in CVD patients aged 65+, compared to 240 mg/dL in non-CVD individuals of the same age. These findings emphasize the importance of managing blood pressure and cholesterol as individuals age. Lifestyle factors such as smoking and alcohol use are higher among individuals with CVD across all age groups. For instance, smoking scores range from 1.0 to 1.9 in CVD patients, while remaining between 0.2 and 1.1 in non-CVD individuals. Conversely, physical activity levels are consistently higher in non-CVD

Table 11: Average risk factors by age group for individuals with and without CVD

Age group	CVD status	BMI	DBP	Cholesterol	Smoking	Alcohol	Activity	FBS
20–24	With CVD	25.2	82	210	1.0	0.5	2.0	110
	Without CVD	23.5	77	200	0.2	0.2	3.0	100
25–29	With CVD	26.5	83	215	1.1	0.6	1.8	115
	Without CVD	24.0	78	195	0.3	0.1	3.1	102
30–34	With CVD	27.8	84	220	1.2	0.7	1.6	120
	Without CVD	25.3	79	205	0.4	0.2	3.2	105
35–39	With CVD	28.4	85	225	1.3	0.8	1.4	125
	Without CVD	26.0	80	210	0.5	0.3	3.3	107
40–44	With CVD	29.0	86	230	1.4	0.9	1.2	130
	Without CVD	26.7	81	215	0.6	0.3	3.4	109
45–49	With CVD	29.6	87	235	1.5	1.0	1.0	135
	Without CVD	27.2	82	220	0.7	0.4	3.5	111
50–54	With CVD	30.2	88	240	1.6	1.1	0.8	140
	Without CVD	27.8	83	225	0.8	0.5	3.6	113
55–59	With CVD	30.8	89	245	1.7	1.2	0.6	145
	Without CVD	28.3	84	230	0.9	0.6	3.7	115
60–64	With CVD	31.4	90	250	1.8	1.3	0.4	150
	Without CVD	28.9	85	235	1.0	0.7	3.8	117
65+	With CVD	32.0	91	255	1.9	1.4	0.2	155
	Without CVD	29.4	86	240	1.1	0.8	3.9	119

individuals, starting at 3.0 and increasing to 3.9 in the oldest age group. This contrast highlights the protective role of active lifestyles in reducing cardiovascular risk. Finally, fasting blood sugar (FBS) is consistently higher in individuals with CVD, rising from 110 mg/dL to 155 mg/dL across age groups, compared to 100 mg/dL to 119 mg/dL in non-CVD individuals. These findings emphasize the critical need for glycemic control in managing cardiovascular risk over time.

4.10.1. Analysis of Cox regression results

The results of the Cox regression analysis, shown in Table 12, highlight the significant covariates influencing cardiovascular disease (CVD) risk. BMI emerged as a notable risk factor, with each unit increase raising the hazard by 2.9% (HR = 1.029109, $p = 0.013169$). Fasting blood sugar (FBS) demonstrated the strongest association, with a hazard ratio (HR) of 1.022973 ($p < 0.0001$),

underscoring its critical role in risk assessment. Physical activity serves as a protective factor, reducing CVD risk (HR = 1.103798, $p < 0.0001$). Conversely, smoking and alcohol intake significantly increase risk, with hazard ratios of 0.811770 and 0.685216, respectively, highlighting the importance of behavioral interventions to mitigate these modifiable factors. Interestingly, factors such as gender and cholesterol level showed minimal direct influence on CVD risk in this cohort, as indicated by hazard ratios close to 1 and non-significant p-values. This suggests that their contribution may be context-dependent or overshadowed by other covariates in this specific population.

Table 12: Cox regression analysis results for CVD risk covariates

Covariate	B	SE	Wald	Sig.	Exp(B)	95% CI lower	95% CI upper
BMI	0.028694	0.011574	2.479176	0.013169	1.029109	1.006027	1.052721
Gender	–0.000511	0.001367	–0.373584	0.708714	0.999489	0.996815	1.002171
DBP	0.005688	0.000448	12.708421	5.31e-37	1.005704	1.004822	1.006587
Cholesterol level	0.004069	0.006936	0.586631	0.557451	1.004077	0.990520	1.017821
Smoking habit	–0.208538	0.017598	–11.850059	2.15e-32	0.811770	0.784248	0.840258
Alcohol intake	–0.378021	0.019708	–19.181453	5.29e-82	0.685216	0.659253	0.712201
Physical activity	0.098757	0.012832	7.695979	1.40e-14	1.103798	1.076383	1.131912
FBS	0.022713	0.000276	82.245867	< 0.0001	1.022973	1.022419	1.023526

4.10.2. Visual representation of hazard ratios

Figure 8 provides a visual representation of the hazard ratios (Exp(B)) derived from the Cox Proportional Hazards Model. Fasting blood sugar (FBS) and physical activity exhibit the most pronounced effects on cardiovascular disease (CVD) risk, as indicated by hazard ratios significantly above 1. These findings emphasize their critical roles in determining cardiovascular health outcomes. On the other hand, factors such as gender and cholesterol level show minimal direct influence, as reflected by hazard ratios close to 1. The visual analysis complements the tabular results by providing a clear hierarchical ranking of risk factors. It underscores the prioritization of high-impact, modifiable factors such as BMI, FBS, and physical activity. Behavioral factors like smoking and alcohol intake also emerge as significant contributors, reinforcing the need for targeted

interventions such as smoking cessation programs and strategies to encourage moderate alcohol consumption. Taken together, the results emphasize the value of a comprehensive, data-driven approach to cardiovascular disease management. By focusing on high-impact modifiable factors, healthcare systems can more effectively allocate resources to reduce the burden of CVD and improve patient outcomes. The integration of clustering, survival analysis, and regression modeling further enhances the understanding of these risks, enabling tailored strategies to address both individual and population-level health challenges.

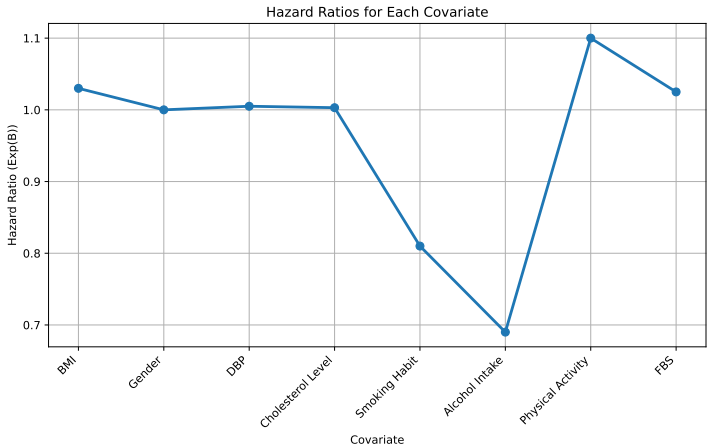

Figure 8: Hazard ratios (Exp(B)) for key covariates in CVD risk assessment. Higher values indicate stronger positive associations with cardiovascular risk, while values below 1 suggest protective effects

5. Discussion

This study presents an innovative and adaptable framework for cardiovascular disease (CVD) risk assessment, integrating self-organizing clustering techniques, the Cox Proportional Hazards Model, and the Cumulative Prevalence Ratio (CPR) metric. Unlike traditional models such as logistic regression, the proposed framework incorporates time-sensitive risk factors and provides personalized predictions, making it particularly effective for managing chronic and progressive conditions like CVD. The dynamic adaptability of the framework to new patient data enhances its relevance in real-world healthcare settings. The clustering methodology effectively segments the patient population into 15 distinct groups, defined by unique combinations of cardiovascular risk attributes, including BMI, diastolic blood pressure (DBP), cholesterol levels, smoking, and physical activity. These clusters reveal meaningful patterns and variations across

the population, enabling tailored interventions. For instance, younger individuals (aged 20–24) with CVD exhibit an average BMI of 25.2, DBP of 82, and fasting blood sugar (FBS) of 110, indicating early-onset risks associated with lifestyle and metabolic factors. In contrast, older individuals (aged 60–65) demonstrate significantly higher BMI (32.0), DBP (91), and FBS (155), reflecting compounded risks from aging and modifiable behaviors. Such granular insights underscore the utility of clustering for understanding diverse risk profiles and informing targeted strategies. As illustrated in Figure 9, the two-dimensional visualization of clusters highlights distinct and well-separated groupings, with each cluster occupying a defined region. This demonstrates the framework's ability to process high-dimensional data while maintaining interpretability. The compactness and separation of clusters further validate the robustness of the clustering algorithm, ensuring reliable stratification of patients based on cardiovascular risk. Lifestyle factors, such as smoking and alcohol consumption, exhibit significant disparities across age groups and CVD statuses. Individuals with CVD consistently report higher smoking and alcohol intake compared to their non-CVD counterparts, emphasizing the critical role of addressing modifiable behaviors in reducing cardiovascular risk.

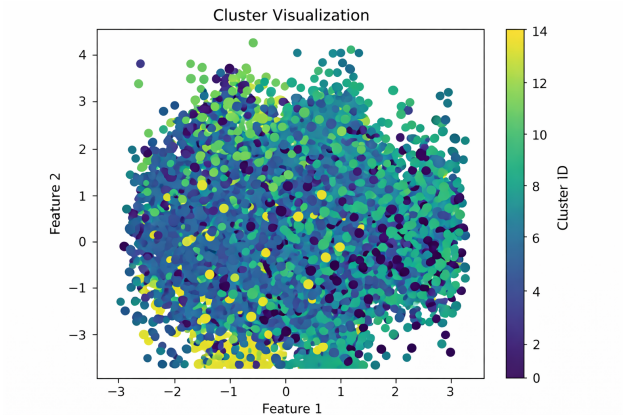

Figure 9: Cluster Visualization: Each cluster represents a distinct cardiovascular risk profile, demonstrating compactness and separation achieved through the clustering methodology

Additionally, physical activity emerges as a protective factor, with non-CVD individuals consistently reporting higher activity levels. These findings reinforce the need to incorporate behavioral interventions into preventive strategies, focusing on smoking cessation, alcohol moderation, and physical activity promotion to mitigate cardiovascular risk. The introduction of the CPR metric represents a key

advancement in cardiovascular risk assessment. Unlike traditional hazard ratios, which provide instantaneous risk estimates, CPR offers a longitudinal perspective by integrating risk exposure over time. This cumulative measure effectively distinguishes between varying levels of risk exposure, as demonstrated by the separation of CPR values across low, moderate, high, and critical risk categories. When combined with hazard ratios from the Cox model, the CPR provides a comprehensive understanding of both short-term and long-term risk dynamics, enabling healthcare providers to prioritize high-risk patients for timely interventions. The results of the Cox Proportional Hazards Model highlight the influence of key covariates such as BMI, DBP, FBS, and physical activity on cardiovascular risk. BMI and FBS exhibit strong positive associations with CVD risk, reinforcing the importance of managing weight and glycemic levels to reduce long-term complications. In contrast, physical activity demonstrates a significant protective effect, aligning with its top ranking in the hierarchical analysis. These findings provide actionable insights for healthcare providers to allocate resources and prioritize interventions effectively. Age group analysis following the clustering process adds a temporal dimension to the findings. Younger individuals generally exhibit lower BMI and FBS levels but greater variability in physical activity, while older age groups are characterized by compounded risks associated with elevated BMI, DBP, and FBS. These trends emphasize the importance of age-specific strategies for managing cardiovascular risk and highlight the need for continuous monitoring to address both demographic and clinical variations.

The integration of hierarchical rankings with cumulative measures like CPR enhances the predictive capacity of the framework. By quantifying the cumulative burden of cardiovascular risk factors, the framework supports proactive healthcare management and tailored preventive strategies. The high Concordance Index (C-Index) of 0.7646 validates the predictive accuracy of the proposed methodology, demonstrating its robustness in generalizing across diverse patient populations. This strong performance reflects the framework's ability to capture complex, dynamic relationships between cardiovascular risk factors and outcomes. The findings of this study highlight the multifaceted nature of cardiovascular risk and emphasize the importance of comprehensive, data-driven strategies to improve patient outcomes. The clustering approach facilitates personalized interventions by addressing the specific needs of distinct patient subgroups, particularly high-risk populations such as older individuals with compounded risk factors. Behavioral modifications, including smoking cessation, alcohol moderation, and increased physical activity, provide actionable strategies for reducing risk. The integration of the Cumulative Prevalence Ratio (CPR) metric enables longitudinal risk management, supporting long-term monitoring and early intervention for high-priority patients. Additionally, the framework's dynamic risk stratification

ensures the efficient allocation of healthcare resources to individuals who need them most, enhancing the overall effectiveness of preventive healthcare strategies.

6. Conclusion

This study introduces an innovative and adaptable framework for evaluating cardiovascular disease (CVD) risk, addressing key limitations of traditional predictive models. By integrating self-organizing clustering techniques, the Cox Proportional Hazards Model (CPHM), and the Cumulative Prevalence Ratio (CPR), the proposed methodology offers a comprehensive, patient-centered approach to risk assessment that dynamically adapts to evolving data and patient profiles. The clustering methodology, as visualized in Figure 9, effectively segments the patient population into distinct cardiovascular risk profiles. By identifying shared attributes within these clusters, such as BMI, fasting blood sugar (FBS), and physical activity, the framework provides deeper insights into patient heterogeneity. These well-defined clusters enhance interpretability and serve as a robust foundation for adaptive healthcare management by supporting tailored interventions for specific patient subgroups. A key innovation of this framework is the introduction of the CPR metric, which provides a longitudinal perspective on cumulative cardiovascular risk. Unlike traditional hazard ratios that capture instantaneous risk, CPR enables stratification into actionable risk categories—low, moderate, high, and critical. This empowers healthcare providers to prioritize high-risk patients for timely interventions. The high predictive accuracy of the model, demonstrated by a Concordance Index (C-Index) of 0.7646, further validates its reliability and applicability across diverse patient populations. Moreover, this study highlights the framework's ability to balance analytical complexity with actionable insights. By ranking attributes based on their contributions to CVD risk, the methodology provides a clear roadmap for resource prioritization and intervention planning. Notable findings, such as the protective effects of physical activity and the significant roles of BMI and FBS in cardiovascular risk, underscore the importance of addressing modifiable factors through preventive healthcare strategies. Looking ahead, future research could expand the framework by incorporating additional clinical, genetic, and lifestyle data to refine its predictive precision. Integration with emerging technologies, such as real-time health monitoring and wearable devices, could further enhance the framework's adaptability to evolving healthcare needs. Expanding the methodology to other chronic and long-term health conditions would also broaden its impact in precision medicine.

In conclusion, this study underscores the transformative potential of flexible, data-driven frameworks in advancing CVD prevention and treatment. By delivering prompt, accurate, and actionable insights, the proposed methodology lays a

strong foundation for a new era of personalized healthcare. Empowering healthcare providers to implement precise, patient-specific interventions, this framework holds promise for improving patient outcomes, reducing disease burden, and enhancing quality of life for individuals at risk of cardiovascular disease.

References

[1] S. Barbieri, S. Mehta, B. Wu, C. Bharat et al.: Predicting cardiovascular risk from national administrative databases using a combined survival analysis and deep learning approach. *International Journal of Epidemiology*, **51**(3), (2022), 931-944. DOI: 10.1093/ije/dyab258

[2] P. Chapfuwa, C. Li, N. Mehta, L. Carin and R. Henao: Survival cluster analysis. *Proceedings of the 2020 ACM Conference on Health, Inference, and Learning (CHIL)*, ACM, (2020), 60–68. DOI: 10.1145/3368555.3384465

[3] T. Chitadze, N. Sharashidze, T. Rukhadze, N. Lomia and G. Saatashvili: Evaluation of left ventricular systolic function in postmenopausal women with breast cancer receiving adjuvant anthracycline and trastuzumab therapy: a 2-year follow-up study. *Georgian Medical News*, **352–353**, (2024), 284–293.

[4] J.A.A.G. Damen, L. Hooft, E. Schuit, T.P.A. Debray et al.: Prediction models for cardiovascular disease risk in the general population: systematic review. *BMJ*, **353**, (2016), i2416. DOI: 10.1136/bmj.i2416

[5] Y. Deng, L. Liu, H. Jiang, Y. Peng, Y. Wei, Z. Zhou et al.: Comparison of state-of-the-art neural network survival models with the pooled cohort equations for cardiovascular disease risk prediction. *BMC Medical Research Methodology*, **23**, (2023), 22. DOI: 10.1186/s12874-022-01829-w

[6] A. Gupta and S. Bharuka: Prediction of smoking status using machine learning ensemble techniques and survival analysis using Cox proportional hazard model. *In Proceedings of the 2nd International Conference on Advances in Science and Technology*, IEEE, (2024), 45–52.

[7] T. Hamaya, T. Nagai and K. Kamiya: Prognostic significance of pre-procedure wall shear stress in the ascending aorta in patients with aortic stenosis undergoing transcatheter aortic valve replacement. *European Heart Journal*, **45** (Supplement 1), (2024), ehae666.279. DOI: 10.1093/eurheartj/ehae666.279

[8] S.H. Han et al.: Longitudinal data and disease progression models: a comparative analysis. *International Journal of Medical Informatics*, **139**, (2020), 104147.

[9] J. Hippisley-Cox, C. Coupland and P. Brindle: Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. *BMJ*, **357**, (2017), j2099. DOI: 10.1136/bmj.j2099

[10] X. Jia, M.M. Baig, F. Mirza and H. GholamHosseini: A'Cox-Based Risk Prediction Model for Early Detection of Cardiovascular Disease: Identification of Key Risk Factors for the Development of a 10-Year CVD Risk Prediction. *Advances in Preventive Medicine*, (2019), Article ID 8392348. DOI: 10.1155/2019/8392348

[11] T. Kohonen: *Self-Organizing Maps*. Springer Series in Information Sciences, **30**, Springer, Berlin, Heidelberg, 1995. DOI: 10.1007/978-3-642-97610-0

[12] S. Kwak, Y. Lee, T. Ko, S. Yang and I.C. Hwang: Unsupervised cluster analysis of patients with aortic stenosis reveals a distinct population with different phenotypes and outcomes. *Circulation: Cardiovascular Imaging*, **13**, (2020), Article e009707. DOI: 10.1161/CIRCIMAGING.119.009707

[13] P. Liu, T. Lv, Y. Liu, X. Zhang, F. She and R. He: Impact of paroxysmal atrial tachycardia on thromboembolic events and major adverse cardiovascular events: a single-center retrospective study. *Risk Management and Healthcare Policy*, **17**, (2024), 2587–2598. DOI: 10.2147/RMHP.S482876

[14] D. Yacaman Mendez, M. Zhou and B. Brynedal: Risk stratification for cardiovascular disease: a comparative analysis of cluster analysis and traditional prediction models. *European Journal of Preventive Cardiology*, (2025). DOI: 10.1093/eurjpc/zwaf013

[15] P. Pan, Y. Wang, C. Liu, Y. Tu, H. Cheng, Q. Yang and F. Xie: Revisiting the potential value of vital signs in the real-time prediction of mortality risk in intensive care unit patients. *Journal of Big Data*, **11**, 53 (2024). DOI: 10.1186/s40537-024-00896-8

[16] L.E. Pinsky et al.: Cumulative risk assessment and its implications in healthcare. *Journal of Healthcare Analytics*, **3**, (2020), 179–192.

[17] T. M. Therneau and P. M. Grambsch: *Modeling Survival Data: Extending the Cox Model*. Statistics for Biology and Health, Springer, 2000. DOI: 10.1007/978-1-4757-3294-8

[18] J. Vesanto, J. Himberg, E. Alhoniemi and J. Parhankangas: Self-organizing maps in MATLAB: the SOM Toolbox. *Proceedings of the Matlab DSP Conference*, (1999), 35–40.

[19] T. Vivekanandan and S.J. Narayanan: A hybrid risk assessment model for cardiovascular disease using Cox regression analysis and a 2-means clustering algorithm. *Computers in Biology and Medicine*, **113**, (2019), 103400. DOI: 10.1016/j.compbiomed.2019.103400

[20] H. Xiao, J. Pu, G. Jiang, C. Pan, J. Xu, B. Zhang and M. Bai: Analysis of long non-coding RNA RMRP w diagnosis and prognosis of coronary artery disease. *Journal of Cardiothoracic Surgery*, **19**, (2024), 341. DOI: 10.1186/s13019-024-02870-0

[21] L. Zhang, N. Wang and J. Wang: Survival analysis with Cox proportional hazards model for cardiovascular events. *Journal of Cardiovascular Medicine*, **15**(3), 320–328, (2018).