

Optimization of medical check-up patient segmentation for personalized healthcare services

Ferdiansyah, Hendra Tjahyadi

Graduate Informatics Program, Faculty of Artificial Intelligence and Data Science, Universitas Pelita Harapan, Jakarta, Indonesia

Article Info

Article history:

Received Dec 12, 2025

Revised Aug 3, 2026

Accepted Aug 12, 2026

Keywords:

Clustering

Hierarchical clustering

K-means

Medical check-up

Patient segmentation

ABSTRACT

Medical check-up (MCU) services commonly apply uniform examination pathways that do not reflect individual variations in demographic and lifestyle risk factors. This study aims to develop an unsupervised learning model that enables early personalization of MCU recommendations using non-laboratory pre-screening data. A dataset of 11,061 participants was processed following the cross-industry standard process for data mining (CRISP-DM) framework, with median imputation, outlier removal, and min–max scaling applied to enhance data quality. K-means and agglomerative hierarchical clustering (AHC) were evaluated using Silhouette score and Davies–Bouldin index, and cluster characteristics were examined based on age, body mass index (BMI), sex, smoking, alcohol consumption, exercise, and family medical history. K-means with four clusters provided the most stable segmentation (Silhouette 0.366; Davies–Bouldin index 1.21), producing clinically interpretable profiles ranging from high-risk groups marked by smoking and elevated BMI to lower-risk individuals with active lifestyles. These findings demonstrate that simple pre-screening variables can support meaningful early risk stratification and guide personalized MCU pathways. The study also outlines an integration approach for embedding the clustering model into electronic medical record (EMR) systems to enable real-time deployment within clinical workflows.

This is an open access article under the [CC BY-SA](#) license.



Corresponding Author:

Hendra Tjahyadi

Graduate Informatics Program, Faculty of Artificial Intelligence and Data Science

Universitas Pelita Harapan, Jakarta, Indonesia

Email: hendra.tjahyadi@uph.edu

1. INTRODUCTION

Efforts to improve healthcare efficiency increasingly demand personalized preventive strategies based on early risk identification. In current practice, demographic and lifestyle information collected during early medical check-up (MCU) registration often remains underutilized for clinical decision-making. For instance, Miki *et al.* [1] demonstrated through latent class analysis of 12,168 Japanese workers that distinct health behavior clusters associated with non-communicable disease (NCD) risk were identifiable, yet rarely leveraged for preventive stratification. Similarly, Chang *et al.* [2] showed that lifestyle clustering patterns linked to metabolic syndrome (MetS) risk in Taiwan were not systematically translated into personalized interventions, while Slagter *et al.* [3] highlighted strong associations between smoking, obesity, and MetS components that are frequently overlooked in routine screening practices. Consistent findings in adolescent populations further confirm that smoking-related and behavioral risk factors contribute significantly to cardiometabolic risk trajectories; however, such information is similarly underutilized in routine early risk identification and preventive stratification processes [4], [5]. Collectively, these studies suggest that variables such as body mass index (BMI), age, smoking behavior, and lifestyle habits contain

clinically meaningful signals that remain underexploited at the pre-screening stage. Prior research has demonstrated that machine learning-based segmentation effectively identifies health-risk groups within large populations. Manzini *et al.* [6] used longitudinal deep learning clustering on 11,028 type 2 diabetes mellitus (T2DM) patients' health records to derive multiple phenotypic trajectories with differential clinical evolution. Pahlevani *et al.* [7] applied a cluster-based algorithm to optimize home health care (HHC) planning in Australia, improving resource allocation for high-risk patients. Yang *et al.* [8] employed clustering techniques on medical data for smart healthcare systems, identifying meaningful patient risk trends, while Ginting *et al.* [9] demonstrated how cluster analysis of high hospital users in Singapore revealed distinct utilization and mortality patterns. At a broader system level, Wang *et al.* [10] emphasized the role of big data analytics in enabling healthcare organizations to uncover obesity-related risks and support preventive decision-making.

Beyond post-diagnostic datasets, a growing body of literature highlights the potential of clustering approaches based on non-laboratory, demographic, and lifestyle-related variables. Gesteiro *et al.* [11] reviewed early identification strategies for MetS and proposed pre-MetS status based on predictors such as BMI and age. Hewner *et al.* [12] demonstrated that combining nursing knowledge with machine learning enables the identification of high-need primary care patients, with smoking and physical inactivity emerging as dominant risk indicators. Similarly, Yan *et al.* [13] showed that inactivity and alcohol consumption are strongly associated with healthcare utilization patterns in segmented multiethnic populations, while Brommels [14] discussed patient segmentation based on self-management ability, emphasizing the impact of lifestyle factors on chronic care pathways. Yuill and Kunz [15] further demonstrated how machine learning-based population segmentation supports personalized risk prediction using demographic and behavioral attributes.

Recent studies have provided additional evidence supporting the feasibility of non-laboratory risk stratification using unsupervised machine learning. Coral *et al.* [16] demonstrated that machine learning-based subclassification of obesity using anthropometric profiles significantly improves prediction of cardiometabolic complications, reinforcing BMI as a core clustering variable. Wu *et al.* [17] showed that integrating lifestyle factors into interpretable machine learning models enhances detection of chronic disease risks. In the Indonesian context, Estiko *et al.* [18] reported that hypertension screening using non-laboratory risk factors and machine learning is effective at the population level, supporting early risk identification before laboratory testing. Complementarily, Hu *et al.* [19] applied unsupervised clustering to hospital-scale electronic health record (EHR) data for cardiovascular disease detection and found that meaningful cluster structures are associated with reduced mortality risk and shorter hospital stays.

Epidemiological and population-based clustering studies further corroborate strong interrelationships among demographic, anthropometric, and behavioral risk factors. Biswas *et al.* [20] identified distinct clusters of metabolic and behavioral cardiovascular risk factors across South and Southeast Asia, emphasizing regional relevance. The authors [3], [5] reported consistent associations between smoking behavior and MetS components in adult and adolescent populations, respectively. Beuken *et al.* [21] further demonstrated that obesity-related subpopulations identified through machine learning exhibit significantly different metabolic risk profiles even without laboratory parameters, while Sheng *et al.* [22] showed that older adults tend to cluster into sedentary behavior-dominated risk profiles.

Nevertheless, most existing clustering studies rely on post-diagnostic, laboratory-intensive, or multimodal EHR data, limiting their applicability for early MCU screening. Hierarchical clustering has been shown to capture clinically meaningful patient hierarchies and long-term outcomes in longitudinal settings [6] and in specific high-risk populations such as hemodialysis patients [23]. More complex approaches, including deep clustering and multimodal architectures, have also been proposed for EHR data [24], [25]. Despite their strong predictive performance, these methods often present challenges in interpretability and implementation complexity, reducing their suitability for routine MCU deployment in resource-constrained settings.

In light of these findings, there remains a clear research gap in applying unsupervised learning methods to support early risk stratification using non-laboratory pre-screening data in MCU services, particularly in the Indonesian context. Therefore, this study aims to develop and evaluate an unsupervised machine learning-based clustering model that utilizes demographic and lifestyle variables collected during registration to generate early patient risk segments before laboratory testing. The resulting risk segments are intended to support more personalized MCU pathways while preserving the simplicity and feasibility required for routine clinical practice.

The remainder of this paper is organized as follows. Section 2 outlines the research framework, data characteristics, and preprocessing procedures. Section 3 presents the clustering results, evaluates cluster quality, and provides clinical interpretations of the identified risk groups. Section 4 concludes the paper with key findings and directions for future research.

2. METHOD

2.1. Research framework

This study adopts the cross-industry standard process for data mining (CRISP-DM) to ensure experimental validity and reproducibility. Previous studies have demonstrated the effectiveness of clustering-based segmentation in healthcare contexts [8], [9]. In the context of MCU services, the CRISP-DM framework provides a structured linkage between technical modeling decisions and clinical workflow requirements.

As illustrated in Figure 1, the workflow follows the CRISP-DM framework consisting of six iterative stages. The process begins with business understanding, which defines the objectives of the MCU segmentation and its expected clinical impact. This is followed by data understanding, where demographic and lifestyle variables such as age, BMI, smoking, alcohol consumption, and family medical history are explored to identify potential relationships and data quality issues. The data preparation phase includes handling missing values, outlier removal, and normalization to ensure that all variables are ready for clustering. The modeling stage implements both k-means and hierarchical clustering to segment patients into distinct risk groups. Next, evaluation is conducted using the Silhouette score and Davies–Bouldin index to measure cluster performance. Finally, the deployment phase involves integrating the segmentation model into the electronic medical record (EMR)/MCU system, allowing real-time identification of patient risk groups during registration.

2.2. Data characteristics

Data were collected from MCU services at a clinic in Jakarta, consisting of 11,061 patient entries covering a wide age range (2–85 years). Variables used for segmentation include age, BMI, sex, smoking, alcohol consumption, exercise, and family medical history. These variables were selected because they are routinely available during early registration and reflect established demographic and lifestyle risk indicators.

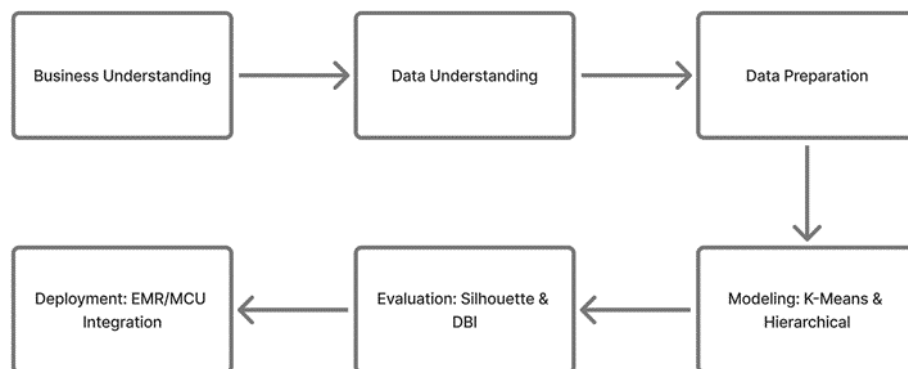


Figure 1. CRISP-DM workflow

Table 1 shows that numerical characteristics exhibit wide variability across features. Patient age ranged from 2–85 years (mean 39.28 ± 11.33), with nearly half in the 30–47 age group. BMI ranged from $13.2\text{--}59.4\text{ kg/m}^2$ (mean 25.86 ± 4.80), indicating a general overweight tendency. Basal metabolic rate (BMR) displayed a wide and skewed range (10.88–2,599 kcal) after removing two clearly erroneous outliers. The revised distribution still indicates substantial variability across patients, supporting the necessity of preprocessing steps such as outlier handling before clustering.

Table 1. Descriptive statistics of numerical variables

Variable	n	Mean	SD	Min	Q1	Median	Q3	Max
Age (years)	10,748	39.28	11.33	2	30	39	47	85
BMI (kg/m ²)	9,075	25.86	4.80	13.2	22.6	25.4	28.4	59.4
BMR (kcal)	9,016	1470.64	266.66	10.88	1269.75	1467.50	1649.00	2599.00

From a lifestyle perspective, approximately 17.3% of patients were reported as active smokers, 6.5% consumed alcohol, and 51.1% engaged in regular physical activity. Family medical history was encoded as a binary variable (FMH_bin), where a value of 1 indicates the presence of one or more relevant

family health conditions. Based on the processed data, about 51.7% of patients fell into this category. A summary of the binary variables is presented in Table 2.

Table 2. Binary variable statistics

Variable	n	Mean	Yes (1) %
Smoking	11.061	0.173	17.29
Alcohol	11.061	0.065	6.52
Exercise	11.061	0.511	51.07
FMH bin	11.061	0.517	51.73

2.3. Data preprocessing

As shown in Figure 2, the preprocessing pipeline ensures that all relevant demographic and lifestyle variables, such as age, BMI, and behavioral indicators, are properly prepared for clustering. The steps include handling missing values, detecting outliers, and applying normalization to standardize feature scales. Median imputation was applied to numerical variables to address missing entries while preserving the distribution characteristics of the dataset. This approach is suitable for right-skewed clinical variables such as BMI and BMR, where the mean may be influenced by extreme values. Outlier handling using the interquartile range (IQR) method was then performed for BMI and BMR to remove clearly erroneous or physiologically implausible values, ensuring that distance-based clustering methods are not distorted by extreme observations.

Categorical variables, including sex and family medical history, were encoded numerically to allow integration with other features in the clustering pipeline. Binary encoding was selected to maintain simplicity and interpretability in downstream analysis. Finally, all numerical variables were normalized to a 0–1 range using min–max scaling. This step is essential to prevent variables with larger numeric ranges, such as age or BMI, from dominating similarity calculations in k-means or hierarchical clustering. After preprocessing, six variables, namely age, BMI, sex, smoking, alcohol consumption, and exercise, as shown in Table 3, are retained for model development, while BMR was excluded due to its high variability and strong correlation with age and BMI, which could introduce redundancy in clustering outcomes.

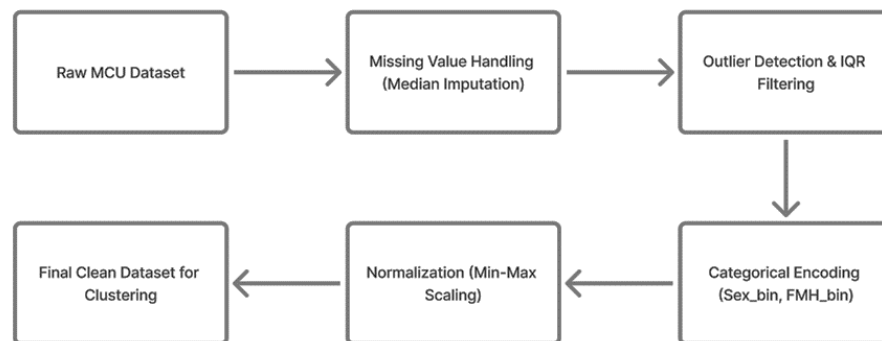


Figure 2. Data preprocessing pipeline

Table 3. Final variables for clustering

Variable type	Variable name	Data type	Description
Demographic	Age	Numeric	Patient age (years)
Demographic	BMI	Numeric	Body mass index
Demographic	Sex_bin	Binary	Male = 1, female = 0
Lifestyle	Smoking	Binary	1 = smoker
Lifestyle	Alcohol	Binary	1 = alcohol consumer
Lifestyle	Exercise	Binary	1 = performs regular exercise

2.4. System architecture and experimental scenario

The proposed system adopts a dual-stream clustering architecture designed to evaluate segmentation consistency across algorithms while maintaining computational efficiency. As shown in Figure 3, the cleaned and normalized dataset is fed in parallel into two clustering modules: k-means and agglomerative hierarchical clustering (AHC). This configuration enables cross-validation of cluster structures and ensures that group formations are not dependent on a single algorithmic assumption.

K-means was selected for its scalability and suitability for large datasets, particularly in environments where rapid turnaround is required for operational deployment in MCU workflows. AHC with Ward linkage was included to capture potential hierarchical relationships within the population and to verify the stability of the segmentation structure produced by k-means. Both algorithms operate on the same preprocessed feature set consisting of age, BMI, sex, smoking, alcohol consumption, exercise, and family medical history.

To determine the optimal number of clusters, an exploratory analysis was conducted using k values ranging from 3 to 6. The choice of four clusters was supported by the Elbow pattern, separation trends observed in the Silhouette score, and clinical interpretability of the resulting groups. During parameter tuning, metrics were computed on a randomly sampled subset of 1,000 observations to reduce computational overhead and allow multiple iterative evaluations. The final model was then retrained on the full dataset to ensure robustness and generalizability. Cluster quality was assessed using two complementary evaluation metrics: Silhouette score, which measures cohesion and separation, and Davies–Bouldin index, which evaluates compactness and cluster distinctiveness. These metrics collectively provide a quantitative basis for selecting the most stable clustering solution and validating the consistency of patterns observed across the two algorithms.

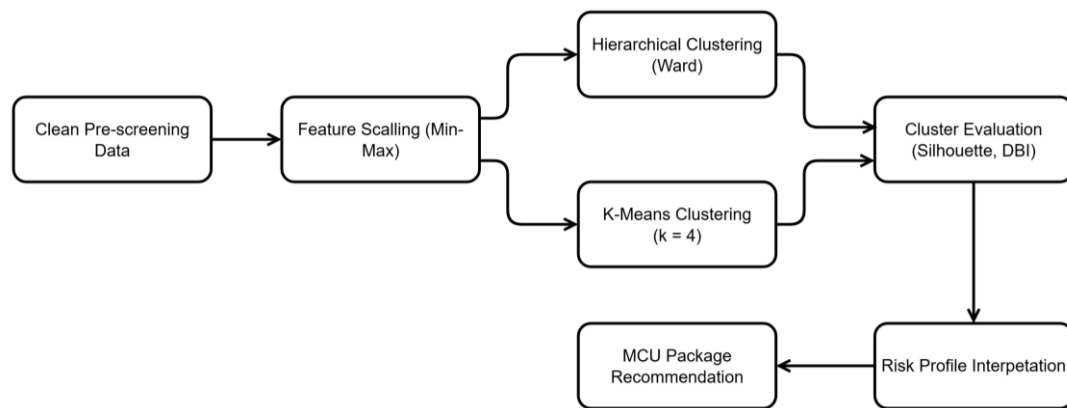


Figure 3. Patient segmentation system architecture

3. RESULTS AND DISCUSSION

3.1. Exploratory data analysis

Distribution visualizations shown in Figure 4 indicate the distribution characteristics of the MCU patients, with age presented in Figure 4(a) and BMI in Figure 4(b). The BMI distribution exhibits a right-skewed BMI distribution, validating the decision to use median imputation and aligning with urban Asian population patterns [1], [9]. Although the dataset includes a small number of patients under 20 years of age, their inclusion does not affect clustering outcomes because missing lifestyle variables (smoking and alcohol) were imputed as non-risk values. This preprocessing decision ensures that age-related data sparsity does not introduce bias in subsequent clustering and evaluation steps.

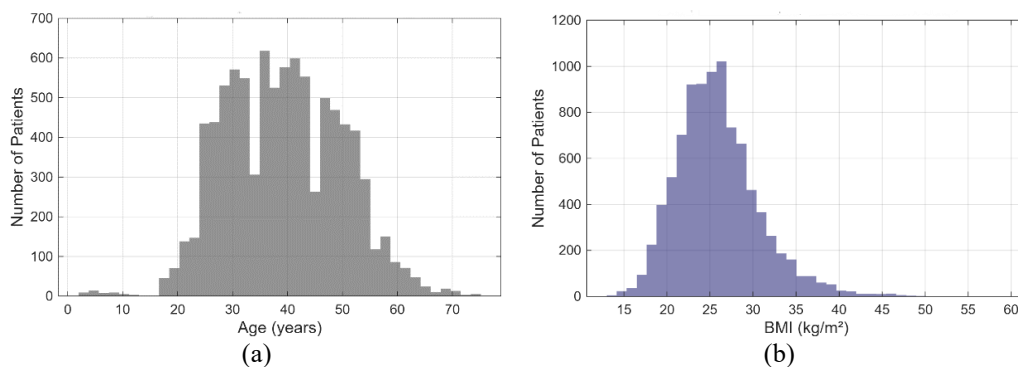


Figure 4. Distribution of MCU patients for (a) age and (b) BMI

3.2. Cluster quality evaluation

Previous comparative studies indicate that k-means often provides sharper separation for health data. For instance, Pahlevani *et al.* [7] presented a mixed-integer linear programming model for HHC routing and scheduling problems in an Australian case study, incorporating a cluster-based algorithm with stochastic simulation and hybrid simulated annealing to achieve fair and balanced workloads, thereby demonstrating the efficiency of clustering-based approaches in optimizing patient segmentation and resource allocation. Similarly, Yang *et al.* [8] utilized clustering techniques on medical data to design a smart healthcare system by analyzing similarities and differences among patient clusters to enhance personalized health education. In a related study, Melnykova *et al.* [26] emphasized methodological rigor by formalizing preprocessing pipelines and introducing an ensemble clustering algorithm that improved cluster separability and reproducibility, as evidenced by higher Dunn index, lower Hopkins statistics, and expert-validated accuracy.

Evaluation results in the present study align with findings reported in the clinical segmentation literature, supporting the validity of clustering-based patient stratification approaches. Consistent with prior work in [8], [26], the clustering performance in this study indicates that pre-screening MCU data contain inherent structures that can be effectively captured using unsupervised learning methods. Notably, this study extends existing evidence by demonstrating that such structures can be identified using non-laboratory registration data within routine MCU services.

To quantitatively evaluate the quality of the clustering results, two internal validation metrics, namely, Silhouette score and Davies–Bouldin index, were employed in this study. The comparative performance of k-means and AHC (Ward linkage) based on these metrics is explicitly summarized in Table 4. This quantitative evaluation provides an objective basis for assessing cluster cohesion and separation, as well as for selecting the most appropriate clustering method for patient risk stratification.

On a randomly selected subset of 1,000 observations, k-means with four clusters achieved a Silhouette score of 0.366 and a Davies–Bouldin index of 1.21, indicating relatively compact clusters with good separation. In comparison, AHC (Ward linkage) produced a Silhouette score of 0.350 and a Davies–Bouldin index of 1.34, reflecting slightly lower clustering quality. Although both algorithms were able to capture consistent natural groupings within the MCU pre-screening data, the results indicate that k-means provides marginally superior performance in terms of internal validation criteria. Scatter plots of age–BMI in Figure 5 further illustrate clear separation among the four identified clusters. Cluster distinctions become more pronounced when age and BMI are considered alongside lifestyle and family medical history variables, providing visual support for the quantitative evaluation results and reinforcing the robustness of the selected clustering solution.

Table 4. Clustering evaluation results

Model	Number of clusters	Silhouette score	Davies–Bouldin index
K-means	4	0.366	1.21
Hierarchical (Ward)	4	0.350	1.34

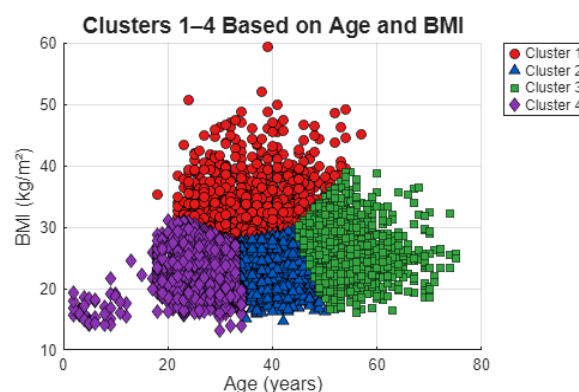


Figure 5. Scatter plots of age-BMI

3.3. Risk profiling and clinical interpretation

Overall, the clustering results reveal clear and clinically relevant groupings within the MCU population. The consistent separation across demographic and lifestyle patterns demonstrates that early risk

stratification is achievable even without laboratory input. K-means exhibited slightly superior performance compared to hierarchical methods, supporting its suitability for real-time deployment in MCU workflows. These findings emphasize the practical value of lightweight clustering models in enhancing personalized healthcare delivery. Following the selection of the optimal clustering model, each cluster was systematically profiled to understand its clinical and behavioral characteristics in greater detail. Table 5 presents the statistical distribution of key demographic, anthropometric, lifestyle, and family medical history variables across the four identified clusters. This cluster-level profiling enables meaningful interpretation of patient risk segments and provides a foundation for personalized MCU recommendations.

Table 5. Cluster profiles (K-means, $k = 4$)

Cluster	n	Mean age	Mean BMI	Male %	Smoking %	Alcohol %	Exercise %	Family medical history %
1	2,294	36.99	25.93	0	5.4	8.5	66.3	100
2	2,831	39.08	26.77	100	43.6	11.7	84.3	100
3	2,211	43.07	24.79	79.5	0.05	1.5	0	26.6
4	1,738	37.27	25.68	77.1	31.8	9.3	100	0

Cluster profiles were analyzed using the mean and proportion of key demographic and lifestyle variables. This approach enables systematic comparison across clusters in terms of age, anthropometric characteristics, and behavioral risk factors. Such descriptive profiling supports clinically interpretable differentiation of patient risk segments derived from the clustering results.

- i) Cluster 1—moderate risk (female): this cluster is composed entirely of women with a positive family medical history. The authors [1], [8] show that genetic predisposition remains a major metabolic risk predictor even among individuals with relatively healthy lifestyles. The profile of this cluster indicates that hereditary factors may contribute to elevated long-term risk despite favorable behavioral characteristics at the time of MCU registration.
- ii) Cluster 2—high risk: this cluster is characterized by higher BMI values combined with a high prevalence of smoking and alcohol consumption. This pattern aligns with findings by Slagter *et al.* [3], who link smoking and obesity with elevated metabolic risk, as well as Tyler *et al.* [4], who highlight the need for intensive monitoring for patients exhibiting multiple lifestyle-related risk factors. The convergence of anthropometric and behavioral risks suggests that individuals in this group represent the most vulnerable segment within the MCU population.
- iii) Cluster 3—mixed risk (elderly): this cluster is dominated by older adults with low levels of physical activity. These characteristics are consistent with epidemiological studies by Miki *et al.* [1], which associate sedentary behavior with increased metabolic risk in adult and working populations. Similarly, Gesteiro *et al.* [11] identify physical inactivity as a key predictor of MetS development across age groups, supporting the mixed but age-dependent risk profile observed in this cluster.
- iv) Cluster 4—low–moderate risk: individuals in this cluster exhibit high levels of regular physical activity and near-normal BMI values. This profile aligns with findings by Miki *et al.* [1], which associate active lifestyles with lower NCD. In addition, Chang *et al.* [2] highlight the protective role of physical activity against MetS, reinforcing the relatively favorable risk status of this group.

The overall distribution of the four clusters in the multidimensional feature space is visualized in Figure 6, which highlights the relative positioning of patient segments and the separation boundaries formed by demographic and lifestyle variables. This visualization complements the tabular cluster profiles by illustrating how risk groups are structured within the feature space. Together, the tabular and visual representations provide a coherent interpretation of clustering outcomes.

3.4. System implementation implications

The model is designed to operate at patient registration. As shown in Figure 7, pre-screening data flow into a segmentation engine that produces real-time risk labels. This approach is consistent with clustering-based risk stratification systems in earlier population-health studies [26], [27].

This approach mirrors clustering-based risk stratification systems in earlier population-health studies and decision-support EMR integrations, such as Aldhyani *et al.* [27], who applied soft clustering for chronic disease diagnosis using EMR data, demonstrating practical integration of clustering models into clinical workflows; Low *et al.* [28], who validated a 4-years segmentation approach on utilization and mortality, supporting EMR stability for longitudinal data; and Si *et al.* [29], who reviewed deep representation learning from EHRs, reinforcing EMRs' role in embedding clustering outputs for personalized recommendations. The system is further supported by literature on digital health competencies among healthcare professionals [30] and segmentation stability in longitudinal datasets [28]. Comparative clustering analyses

also emphasize the importance of selecting appropriate methods for clinical contexts [31]. Deep representation learning studies [29] further reinforce the relevance of embedding clustering outputs into EMR-driven recommendation systems, enabling personalized add-on examinations such as lipid panels or cardiac assessments depending on risk groups.

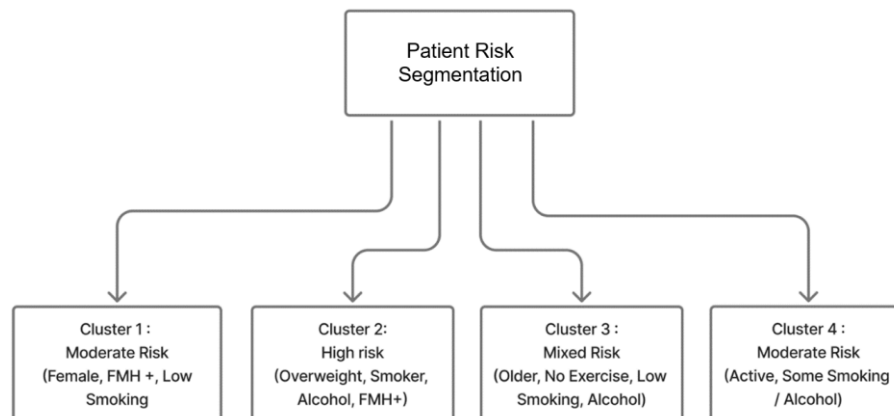


Figure 6. Risk stratification map

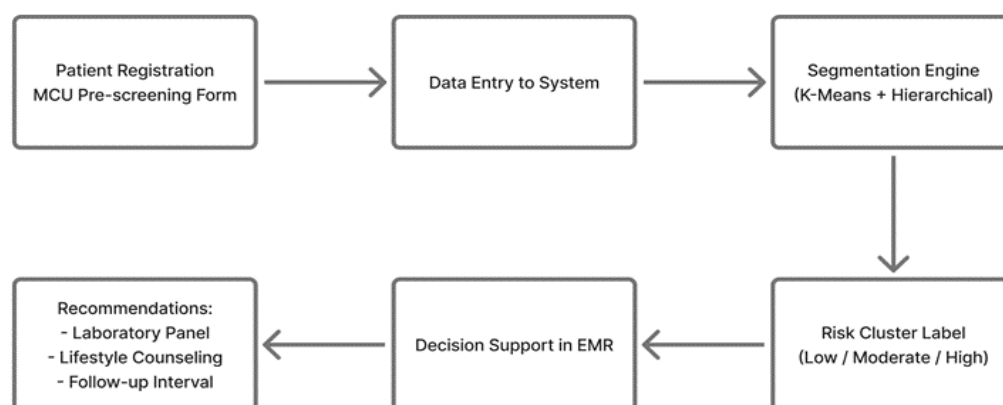


Figure 7. EMR integration scheme

4. CONCLUSION

This study demonstrates that unsupervised clustering based on non-laboratory pre-screening variables can effectively stratify MCU patients into clinically meaningful risk groups. Using demographic and lifestyle information routinely collected at registration, the proposed approach identifies distinct patient segments without reliance on laboratory test results. Among the evaluated methods, k-means with four clusters exhibited slightly superior internal validation performance compared to AHC, indicating better cluster cohesion and separation. The findings highlight the feasibility of integrating lightweight clustering models into early MCU workflows to support data-driven risk stratification and personalized preventive care. The proposed segmentation framework can be embedded within EMR systems to enable real-time identification of patient risk profiles at registration. Future research may extend this work by validating the model across broader populations and incorporating additional clinical variables to further refine risk characterization.

ACKNOWLEDGMENTS

The authors gratefully acknowledge the support of Universitas Pelita Harapan in providing the facilities and institutional resources required for this research.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Ferdiansyah	✓	✓	✓	✓	✓	✓	✓	✓	✓		✓			
Hendra Tjahyadi	✓	✓		✓	✓	✓	✓	✓	✓	✓		✓		✓

C : Conceptualization

M : Methodology

So : Software

Va : Validation

Fo : Formal analysis

I : Investigation

R : Resources

D : Data Curation

O : Writing - Original Draft

E : Writing - Review & Editing

Vi : Visualization

Su : Supervision

P : Project administration

Fu : Funding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data supporting the findings of this study were obtained from Kyoai Medical Services and are available from the corresponding author, [HT], upon reasonable request and subject to institutional approval. Restrictions apply to the availability of these data due to privacy and confidentiality regulations of the data owner; therefore, the data are not publicly available.

REFERENCES

- [1] T. Miki, K. Yamamoto, M. Kanai, K. Takeyama, M. Iwatake, and Y. Hagiwara, "Identifying clusters of health behaviors in a Japanese working population at risk for non-communicable diseases: a latent class analysis of 12,168 individuals," *SSM - Population Health*, vol. 24, Dec. 2023, doi: 10.1016/j.ssmph.2023.101539.
- [2] Y. H. Chang, C. Y. Li, H. P. Ma, C. Y. Wu, Y. Y. Jou, and C. B. Lee, "Clustering of lifestyle behaviors and their association with risk of metabolic syndrome among adults in Taiwan: nationwide cross-sectional study," *JMIR Public Health and Surveillance*, vol. 11, Sep. 2025, doi: 10.2196/73114.
- [3] S. N. Slagter *et al.*, "Associations between smoking, components of metabolic syndrome and lipoprotein particle size," *BMC Medicine*, vol. 11, no. 1, Dec. 2013, doi: 10.1186/1741-7015-11-195.
- [4] N. S. Tyler *et al.*, "An artificial intelligence decision support system for the management of type 1 diabetes," *Nature Metabolism*, vol. 2, no. 7, pp. 612–619, Jun. 2020, doi: 10.1038/s42255-020-0212-y.
- [5] N. Thangiah, K. Chinna, T. T. Su, M. Y. Jalaludin, N. Al-Sadat, and H. A. Majid, "Clustering and tracking the stability of biological CVD risk factors in adolescents: the Malaysian health and adolescents longitudinal research team study (MyHeARTs)," *Frontiers in Public Health*, vol. 8, Mar. 2020, doi: 10.3389/fpubh.2020.00069.
- [6] E. Manzini *et al.*, "Longitudinal deep learning clustering of type 2 diabetes mellitus trajectories using routinely collected health records," *Journal of Biomedical Informatics*, vol. 135, Nov. 2022, doi: 10.1016/j.jbi.2022.104218.
- [7] D. Pahlevani, B. Abbasi, J. W. Hearne, and A. Eberhard, "A cluster-based algorithm for home health care planning: a case study in Australia," *Transportation Research Part E: Logistics and Transportation Review*, vol. 166, Oct. 2022, doi: 10.1016/j.tre.2022.102878.
- [8] W. C. Yang, J. P. Lai, Y. H. Liu, Y. L. Lin, H. P. Hou, and P. F. Pai, "Using medical data and clustering techniques for a smart healthcare system," *Electronics*, vol. 13, no. 1, Dec. 2024, doi: 10.3390/electronics13010140.
- [9] M. L. Ginting, Y. H. Ang, S. H. Ho, G. Sum, and C. H. Wong, "Understanding the characteristics of high users of hospital services in Singapore and their associations with healthcare utilisation and mortality: a cluster analysis," *PLoS ONE*, vol. 18, no. 7 July, Jul. 2023, doi: 10.1371/journal.pone.0288441.
- [10] Y. Wang, L. A. Kung, and T. A. Byrd, "Big data analytics: understanding its capabilities and potential benefits for healthcare organizations," *Technological Forecasting and Social Change*, vol. 126, pp. 3–13, Jan. 2018, doi: 10.1016/j.techfore.2015.12.019.
- [11] E. Gesteiro, A. Megía, A. G.-Grau, S. F.-Veledo, J. Vendrell, and M. G.-Gross, "Early identification of metabolic syndrome risk: a review of reviews and proposal for defining pre-metabolic syndrome status," *Nutrition, Metabolism and Cardiovascular Diseases*, vol. 31, no. 9, pp. 2557–2574, Aug. 2021, doi: 10.1016/j.numecd.2021.05.022.
- [12] S. Hewner, E. Smith, and S. S. Sullivan, "Identifying high-need primary care patients using nursing knowledge and machine learning methods," *Applied Clinical Informatics*, vol. 14, no. 3, pp. 408–417, May 2023, doi: 10.1055/a-2048-7343.
- [13] S. Yan, Y. H. Kwan, J. Thumboo, and L. L. Low, "Characteristics and health care utilization of different segments of a multiethnic Asian population in Singapore," *JAMA Network Open*, vol. 2, no. 9, Sep. 2019, doi: 10.1001/jamanetworkopen.2019.10878.
- [14] M. Brommels, "Patient segmentation: adjust the production logic to the medical knowledge applied and the patient's ability to self-manage—a discussion paper," *Frontiers in Public Health*, vol. 8, May 2020, doi: 10.3389/fpubh.2020.00195.

- [15] W. Yuill and H. Kunz, "Using machine learning to improve personalised prediction: a data-driven approach to segment and stratify populations for healthcare," *Studies in Health Technology and Informatics*, vol. 289, pp. 29–32, Jan. 2022, doi: 10.3233/SHTI210851.
- [16] D. E. Coral *et al.*, "Subclassification of obesity for precision prediction of cardiometabolic diseases," *Nature Medicine*, vol. 31, no. 2, pp. 534–543, Feb. 2025, doi: 10.1038/s41591-024-03299-7.
- [17] Y. Wu, D. Dong, L. Zhu, Z. Luo, Y. Liu, and X. Xie, "Interpretable machine learning models for detecting peripheral neuropathy and lower extremity arterial disease in diabetics: an analysis of critical shared and unique risk factors," *BMC Medical Informatics and Decision Making*, vol. 24, no. 1, Jul. 2024, doi: 10.1186/s12911-024-02595-z.
- [18] R. I. Estiko, B. Widyantoro, D. A. Juzar, R. M. Yusup, I. F. Rakhmat, and E. Rijanto, "Screening hypertension using non-laboratory risk factors with machine learning: a retrospective cross-sectional study in Indonesia," *BMJ Open*, vol. 15, no. 8, Aug. 2025, doi: 10.1136/bmjopen-2024-092364.
- [19] Y. Hu *et al.*, "Detecting cardiovascular diseases using unsupervised machine learning clustering based on electronic medical records," *BMC Medical Research Methodology*, vol. 24, no. 1, Dec. 2024, doi: 10.1186/s12874-024-02422-z.
- [20] T. Biswas *et al.*, "Clustering of metabolic and behavioural risk factors for cardiovascular diseases among the adult population in South and Southeast Asia: findings from WHO STEPS data," *The Lancet Regional Health - Southeast Asia*, vol. 12, May 2023, doi: 10.1016/j.lansea.2023.100164.
- [21] M. J. M. Beuken *et al.*, "Identification of clusters in a population with obesity using machine learning: secondary analysis of the Maastricht study," *JMIR Medical Informatics*, vol. 13, Feb. 2025, doi: 10.2196/64479.
- [22] Y. Sheng, R. Bond, R. Jaiswal, J. Dinsmore, and J. Doyle, "Augmenting k-means clustering with qualitative data to discover the engagement patterns of older adults with multimorbidity when using digital health technologies: proof-of-concept trial," *Journal of Medical Internet Research*, vol. 26, no. 1, Mar. 2024, doi: 10.2196/46287.
- [23] Y. Komaru, T. Yoshida, Y. Hamasaki, M. Nangaku, and K. Doi, "Hierarchical clustering analysis for predicting 1-year mortality after starting hemodialysis," *Kidney International Reports*, vol. 5, no. 8, pp. 1188–1195, Aug. 2020, doi: 10.1016/j.ekir.2020.05.007.
- [24] I. Kowsar, S. B. Rabbani, K. F. B. Akhter, and M. D. Samad, "Deep clustering of electronic health records tabular data for clinical interpretation," in *5th IEEE International Conference on Telecommunications and Photonics, ICTP 2023 - e-Proceedings*, Dec. 2023, pp. 1–5, doi: 10.1109/ICTP60248.2023.10490723.
- [25] M. Bampa, I. Miliou, B. Jovanovic, and P. Papapetrou, "M-ClustEHR: a multimodal clustering approach for electronic health records," *Artificial Intelligence in Medicine*, vol. 154, Aug. 2024, doi: 10.1016/j.artmed.2024.102905.
- [26] N. Melnykova, N. Shakhovska, M. Gregus, V. Melnykov, M. Zakharchuk, and O. Vovk, "Data-driven analytics for personalized medical decision making," *Mathematics*, vol. 8, no. 8, Jul. 2020, doi: 10.3390/MATH8081211.
- [27] T. H. H. Aldhyani, A. S. Alshebami, and M. Y. Alzahrani, "Soft clustering for enhancing the diagnosis of chronic diseases over machine learning algorithms," *Journal of Healthcare Engineering*, vol. 2020, pp. 1–16, Mar. 2020, doi: 10.1155/2020/4984967.
- [28] L. L. Low, S. Yan, Y. H. Kwan, C. S. Tan, and J. Thumboo, "Assessing the validity of a data-driven segmentation approach: a 4 year longitudinal study of healthcare utilization and mortality," *PLoS ONE*, vol. 13, no. 4, Apr. 2018, doi: 10.1371/journal.pone.0195243.
- [29] Y. Si *et al.*, "Deep representation learning of patient data from electronic health records (EHR): a systematic review," *Journal of Biomedical Informatics*, vol. 115, Mar. 2021, doi: 10.1016/j.jbi.2020.103671.
- [30] G. Jimenez *et al.*, "Digital health competencies for primary healthcare professionals: a scoping review," *International Journal of Medical Informatics*, vol. 143, Nov. 2020, doi: 10.1016/j.ijmedinf.2020.104260.
- [31] A. Aljohani, "Optimizing patient stratification in healthcare: a comparative analysis of clustering algorithms for EHR data," *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00568-8.

BIOGRAPHIES OF AUTHORS



Ferdiansyah    is a graduate candidate of the Informatics Graduate Program from Universitas Pelita Harapan. He earned his bachelor's degree in Accounting from Perbanas Institute in Indonesia. He can be contacted at email: ferdi.9986@gmail.com.



Hendra Tjahyadi    is an associate professor of the Graduate Informatics Program at Universitas Pelita Harapan, Indonesia. He earned his bachelor's degree in Electrical Engineering from Universitas Kristen Maranatha, master's degree in Instrumentation and Control from Institut Teknologi Bandung, and Ph.D. in Control Engineering from the School of Engineering, Flinders University. His research interests are in adaptive control, signal processing, and artificial intelligence. He can be contacted at email: hendra.tjahyadi@uph.edu.