Unsupervised Machine Learning with Cluster Analysis in Patients Discharged after an Acute Coronary Syndrome: Insights from a 23,270-Patient Study

Tanya Mohammadi, PhD[a], Fabrizio D'Ascenzo, MD[b], Martino Pepe, MD, PhD[c], Stefano Bonsignore Zanghì[d], Marco Bernardi, MD[e], Luigi Spadafora, MD[e], Giacomo Frati, MD, MSc[f,g], Mariangela Peruzzi, MD, PhD[f,h], Gaetano Maria De Ferrari, MD[b], and Giuseppe Biondi-Zoccai, MD, MStat[f,h,*]

Characterization and management of patients admitted for acute coronary syndromes (ACS) remain challenging, and it is unclear whether currently available clinical and procedural features can suffice to inform adequate decision making. We aimed to explore the presence of specific subsets among patients with ACS. The details on patients discharged after ACS were obtained by querying an extensive multicenter registry and detailing patient features, as well as management details. The clinical outcomes included fatal and nonfatal cardiovascular events at 1-year follow-up. After missing data imputation, 2 unsupervised machine learning approaches (k-means and Clustering Large Applications [CLARA]) were used to generate separate clusters with different features. Bivariate- and multivariable-adjusted analyses were performed to compare the different clusters for clinical outcomes. A total of 23,270 patients were included, with 12,930 cases (56%) of ST-elevation myocardial infarction (STEMI). K-means clustering identified 2 main clusters: a first 1 including 21,998 patients (95%) and a second 1 including 1,282 subjects (5%), with equal distribution for STEMI. CLARA generated 2 main clusters: a first 1 including 11,268 patients (48%) and a second 1 with 12,002 subjects (52%). Notably, the STEMI distribution was significantly different in the CLARA-generated clusters. The clinical outcomes were significantly different across clusters, irrespective of the originating algorithm, including death reinfarction and major bleeding, as well as their composite. In conclusion, unsupervised machine learning can be leveraged to explore the patterns in ACS, potentially highlighting specific patient subsets to improve risk stratification and management.

(Am J Cardiol 2023;193:44−51)

Acute coronary syndromes (ACS), that is, and ST-elevation myocardial infarction (STEMI), unstable angina, and non-STEMI, comprising non−ST-elevation ACS, share atherothrombosis as the underlying pathophysiologic mechanism.[1−4] This etiologic premise calls into question the rigid distinction between unstable angina, NSTEMI, and STEMI. Modern artificial intelligence methods can be leveraged to appraise clustering features.[5,6] The Prediction of Adverse Events following an Acute Coronary Syndrome (PRAISE) study has recently reported on supervised machine learning to improve the risk prognostication in patients with ACS.[7] We aimed to further expand the evidence base stemming from the PRAISE dataset by exploring the clustering features among patients with ACS and leveraging available clinical and procedural features by means of established unsupervised machine learning methods, with the ultimate aim of possibly improving decision making.

Methods

The details on the PRAISE study have already been reported in detail in *The Lancet*.[7] Briefly, the data on adult patients with ACS were obtained by pooling several international registries, including the BleeMACS, FRASER, RENAMI, and SECURITY studies, with hospitalization occurring between 2003 and 2019. The diagnosis of ACS was based on contemporary guidelines, with patient management as per guideline-informed institutional practice. The outcomes of interest, collected up to 2 years of follow-up,

[a]School of Mathematics, Statistics, and Computer Science, College of Science, University of Tehran, Tehran, Iran; [b]Division of Cardiology, Cardiovascular and Thoracic Department, Città della Salute e della Scienza, Turin, Italy; [c]Division of Cardiology, Department of Emergency and Organ Transplantation, University of Bari, Bari, Italy; [d]Faculty of Pharmacy and Medicine, Sapienza University of Rome, Latina, Italy; [e]Department of Clinical, Internal Medicine, Anesthesiology and Cardiovascular Sciences, Sapienza University of Rome, Italy; [f]Department of Medical-Surgical Sciences and Biotechnologies, Sapienza University of Rome, Latina, Italy; [g]IRCCS NEUROMED, Pozzilli, Italy; and [h]Mediterranea Cardiocentro, Napoli, Italy. Manuscript received September 11, 2022; revised manuscript received and accepted January 29, 2023.

Authors Tanya Mohammadi and Fabrizio D'Ascenzo contributed equally as joint first authors.

Authors Gaetano Maria De Ferrari and Giuseppe Biondi-Zoccai contributed equally as joint last joint authors.

Funding: none.

See page 51 for disclosure information.

*Corresponding author: Tel: +3907736551; fax: +3907736551.

E-mail address: giuseppe.biondizoccai@uniroma1.it (G. Biondi-Zoccai).

https://doi.org/10.1016/j.amjcard.2023.01.048

were death, myocardial infarction (defined according to each study methods), and major bleeding (defined according to the Bleeding Academic Research Consortium as type 3 or 5).[8]

In keeping with best practice for unsupervised machine learning, the PRAISE dataset was first investigated for missing data and the presence of correlated variables. The correlated features were identified with a heatmap of Pearson correlation coefficients. A single variable from each pair of highly correlated variables was included in the process of clustering. The missing data were imputed and hierarchical clustering of variables was performed. The results of variable clustering were assessed with the Rand similarity index. A representative variable with the highest squared loading on the first principal component was selected for recognized clusters. Furthermore, the clustering tendency of the dataset was assessed with Hopkins statistic, and the number of patient clusters was determined using a majority voting ensemble method.

Next, the Euclidean distance was used for calculating the dissimilarity matrix, and then, the k-means and the Clustering Large Applications (CLARA) algorithms were run. K-means is a commonly used algorithm for partition clustering and the CLARA algorithm is an extension of the Partitioning Around Medoids clustering method capable of dealing with large datasets. Then, the results of clustering were illustrated using principal components, and their validity was investigated.

The results are presented as mean (standard deviation) for continuous and as absolute numbers (%) for categorical variables. Clusters were then compared in the nonimputed dataset using frequentist methods. Specifically, the means of continuous variables were compared using independent *t* tests. The chi-square test with Yates continuity correction was used for comparing the categorical variables between the clusters. The statistical significance was set at 2-tailed $\alpha = 0.05$. All data analyses were performed with R version 4.0.2 for Windows (R Foundation for Statistical Computing, Vienna, Austria). Notably, we used a variety of R packages for the analysis, including NbClust.

Results

In total, 23,270 patients (5,466 women [23.5%]) were included. The characteristics of the total sample are presented in Table 1. The mean (standard deviation), minimum, and maximum age was 63.4 (12.5), 23, and 100 years, respectively. Age, serum creatinine, estimated glomerular filtration rate, hemoglobin, and left ventricular ejection fraction had been recorded as continuous, and other features had been recorded as binary variables. Primary explorations showed that 8% of the data were missing in the dataset. Figure 1 illustrates the percentages and patterns of missing data for each 46 clinically informative features.

To ensure proper imputation and clustering, the dataset was investigated with respect to the presence of correlated variables. We considered pairs of variables as highly correlated if the absolute value of their correlations was more than 0.5. The heatmap of the Pearson correlation coefficients showed some highly correlated features within the data (Figure 2). We excluded variables that were a subset of other features from the clustering process. For example, the diabetes type was included, whereas noninsulin and insulin-dependent diabetes variables were excluded (Figure 2). Overall, the remaining variables included only 9.4% missing data. Considering this small total percentage of missing data and the large numbers of variables and sample size, the missing values were imputed 1 time with the predictive

Table 1
Comparisons of the clusters using the non-imputed original dataset, focusing on presenting features (n = 23,270)

Characteristic	All (N = 23,270)	k-means		p Value	CLARA		p Value
		k1 (n_1=21,988)	K2 (n_2=1,282)		C1(n_1=11,268)	C2 (n_2=12,002)	
Age (year)	63.42 (12.50)	63.29 (12.45)	65.62 (13.15)	<0.001	66.37 (11.68)	60.65 (12.62)	<0.001
Female (%)	5,466 (23.49)	5,038 (22.91)	428 (33.39)	<0.001	2,845 (25.25)	2,621 (21.84)	<0.001
Hypertension (%)	12,734 (54.72)	11,986 (54.51)	748 (58.35)	0.008	9,318 (82.69)	3,416 (28.46)	<0.001
Dyslipidemia (%)	12,148 (52.20)	11,670 (53.07)	478 (37.29)	<0.001	6,769 (60.07)	5,379 (44.82)	<0.001
Non-insulin-dependent diabetes (%)	5,518 (23.72)	5,460 (24.84)	58 (4.52)	<0.001	3,425 (30.40)	2,093 (17.45)	<0.001
Insulin-dependent diabetes (%)	344 (1.48)	139 (0.63)	205 (15.99)	<0.001	205 (1.82)	139 (1.16)	<0.001
Peripheral artery disease	1,606 (7.65)	1,503 (7.63)	103 (8.04)	0.627	982 (9.83)	624 (5.67)	<0.001
Prior myocardial infarction (%)	3,349 (14.40)	3,213 (14.62)	136 (10.61)	<0.001	2,220 (19.71)	1,129 (9.41)	<0.001
Prior PCI (%)	3,312 (14.23)	3,181 (14.47)	131 (10.22)	<0.001	2,175 (19.30)	1,137 (9.47)	<0.001
Prior CABG (%)	777 (3.34)	725 (3.30)	52 (4.06)	0.165	657 (5.83)	120 (1.00)	<0.001
Prior stroke (%)	1,309 (5.63)	1,255 (5.71)	54 (4.21)	0.028	761 (6.75)	548 (4.57)	<0.001
Prior bleeding (%)	980 (4.62)	964 (4.59)	16 (8.33)	0.022	583 (5.68)	397 (3.63)	<0.001
Malignancy (%)	1,188 (5.57)	1,128 (5.62)	60 (4.68)	0.175	663 (6.43)	525 (4.75)	<0.001
STEMI (%)	12,930 (55.57)	12,191 (55.44)	739 (57.64)	0.130	2,972 (26.38)	9,958 (82.97)	<0.001
NSTEACS (%)	9,424 (43.25)	8,881 (43.31)	543 (42.36)	0.523	7,649 (72.02)	1,775 (15.89)	<0.001
Creatinine (mg/dL)	97.85 (55.64)	97.60 (54.73)	101.99 (68.93)	0.027	103.90 (70.06)	92.29 (36.97)	<0.001
eGFR (mL/min)	90.20 (39.19)	90.29 (39.27)	80.34 (28.62)	<0.001	85.25 (33.38)	94.88 (43.48)	<0.001
Hemoglobin (g/dL)	1,397.42 (168.46)	1,398.52 (167.31)	1,381.01 (184.02)	0.001	1,378.97 (172.26)	1,415.06 (162.80)	<0.001
LVEF (%)	52.27 (10.94)	52.63 (10.84)	47.06 (11.11)	<0.001	53.28 (11.06)	51.25 (10.72)	<0.001

C1 and C2 = Clusters 1 and 2 identified by the CLARA algorithm; CABG = Coronary Artery Bypass Graft surgery; eGFR (MDRD): estimated Glomerular Filtration Rate (Modification of Diet in Renal Disease); k1 and k2 = Clusters 1 and 2 identified by the k-means algorithm; LVEF = Left Ventricular Ejection Fraction; NSTEACS = Non-ST-segment Elevation Acute Coronary Syndrome; PCI = Percutaneous Coronary Intervention; STEMI = ST-segment Elevation Myocardial Infarction.

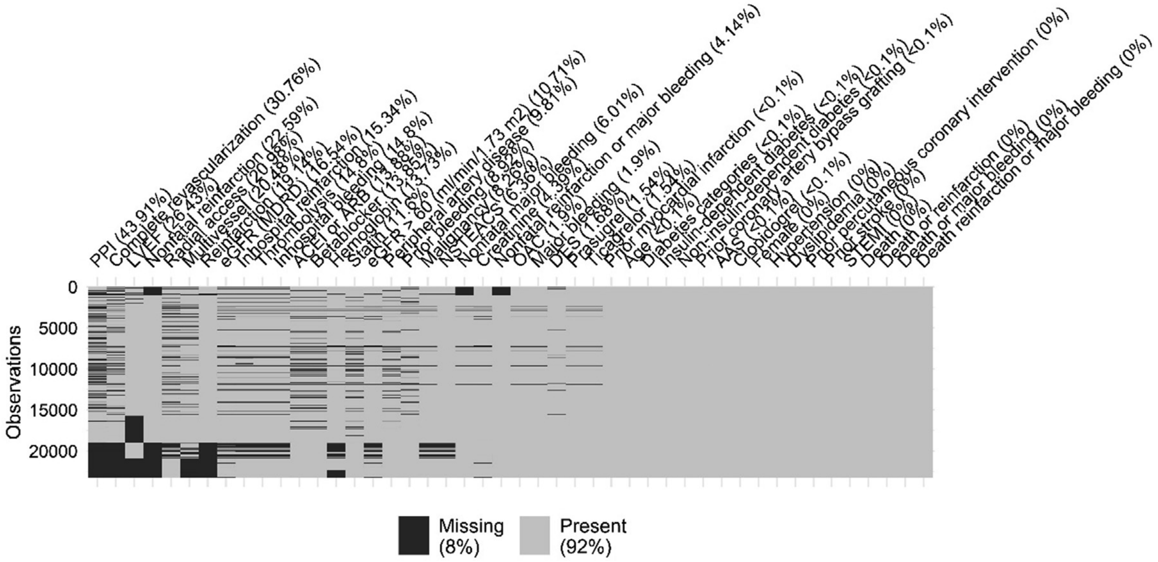

Figure 1. Missing data description in the whole dataset. AAS = anabolic-androgenic steroids; ACE = angiotensin-converting enzyme inhibitors; ARB = angiotensin receptor blockers; eGFR (MDRD) = estimated glomerular filtration rate (modification of Diet in Renal Disease), LVEF = left ventricular ejection fraction; NSTEACS = non-ST-elevation acute coronary syndrome; OAC = oral anticoagulant, PPI = proton-pump inhibitor.

mean matching method and the maximum iteration of 10. Then, we carried out hierarchical clustering of variables and achieved the maximum stability of 1.0 based on the mean corrected Rand indexes with 27 clusters and 25 bootstrap samples. Next, for each cluster, a representative variable with the highest squared loadings on the first principal component was selected. We repeated the procedure on the 27 remaining variables to get the maximum stability of 0.94 with 15 variables. At the end of the process, we incorporated hypertension, peripheral arterial disease, prior coronary artery bypass graft surgery, major bleeding, malignancy, STEMI, serum creatinine, thrombolysis, multivessel coronary artery disease, in-hospital reinfarction, anabolic-androgenic steroids, clopidogrel, and oral anticoagulation use into the clustering algorithms.

The data were scaled, and the clustering tendency was confirmed with the Hopkins statistic of 0.984. Figure 3 illustrates the result of principal component analysis. Because of the high demand for computational resources, we used a random sample containing 50% of patients to

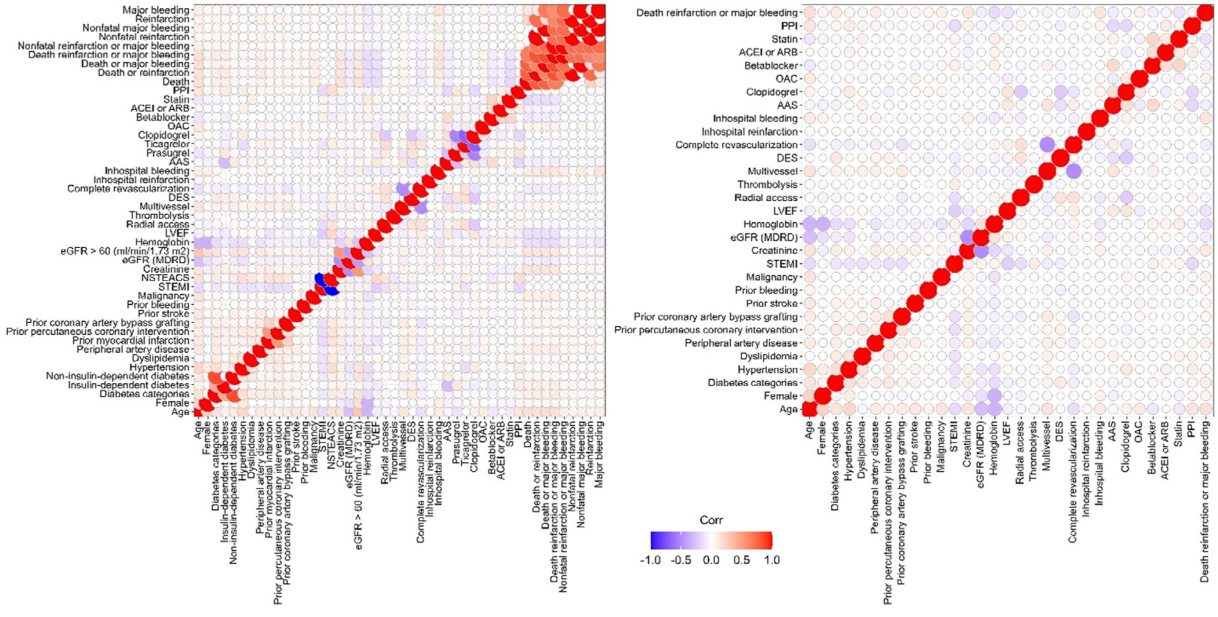

Figure 2. Heatmaps of Pearson’s correlation coefficients before (left panel) and after (right panel) excluding variables. AAS = anabolic-androgenic steroids; ACE = angiotensin-converting enzyme inhibitors; ARB = angiotensin receptor blockers; eGFR (MDRD) = estimated glomerular filtration rate (modification of Diet in Renal Disease); LVEF = left ventricular ejection fraction; NSTEACS = non-ST-elevation acute coronary syndrome; OAC = oral anticoagulant; PPI = proton-pump inhibitor.

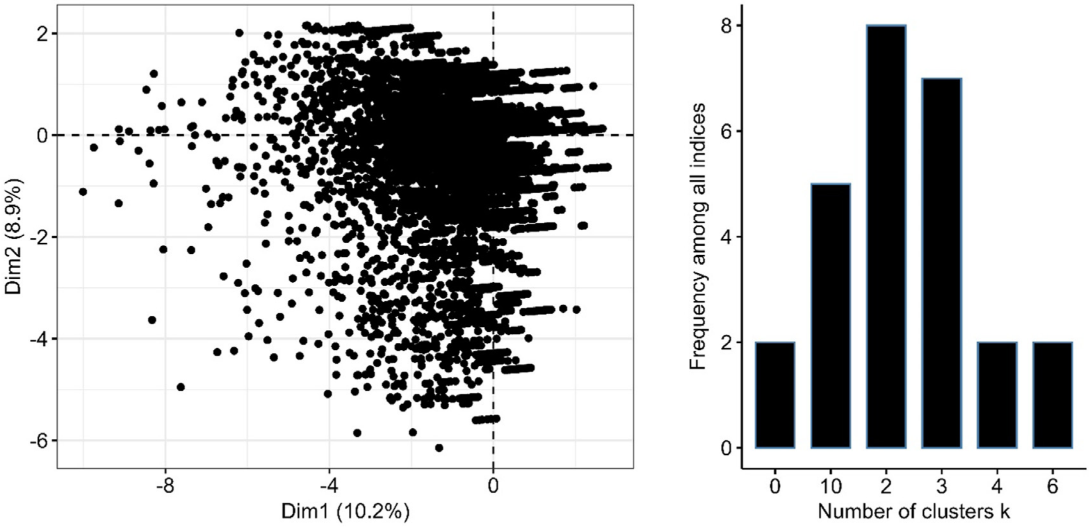

Figure 3. *(Left panel)* scatter diagram of 2 dimensions (Dim1 and Dim2) resulting from principal component analysis of the data. *(Right panel)* majority voting for the optimum number of clusters.

find the number of clusters. The number of clusters was determined using a majority voting ensemble method (Figure 3). Overall, we decided to set the number of clusters to 2 for further analysis. For the whole dataset, we used a k-means clustering algorithm with 25 random samples and the maximum iteration of 5 (Figure 4). The Euclidean distance was used for calculating the dissimilarity matrix. The algorithm clustered the dataset rows into 2 groups of 21,988 and 1,282 patients. Further investigations with a scaled 80% random sample of data (17,600 patients from cluster 1 and 1,016 from cluster 2) showed that within-cluster average distances were 4.71 and 5.48, the between-clusters average distance was 7.05, Dunn 2 index was 1.29, and average silhouette width was 0.34. Also, we applied the CLARA clustering algorithm to the whole data set using 100 samples (Figure 4) and the dissimilarity matrix of Euclidean distances. The CLARA clustered the rows into 2 groups of 11,268 and 12,002 patients. The investigations with a scaled 80% random sample of data (9,008 patients from cluster 1 and 9,608 from cluster 2) showed that the within-cluster average distances were 5.00 and 4.46, the between-clusters average distance was 5.20, the Dunn 2 index was 1.04, and the average silhouette width was 0.10.

Tables 1−4 show the results of comparing the 2 clusters using k-means and the CLARA algorithms. We called the clusters k1 and C1 for cluster 1 and k2 and C2 for cluster 2, identified by k-means and the CLARA, respectively. For the k-means clustering, there was no statistically significant difference in peripheral artery disease, previous coronary artery bypass graft surgery, malignancy, STEMI, non−ST-segment elevation acute coronary syndrome, thrombolysis, in-hospital reinfarction, and angiotensin-converting enzyme inhibitors or angiotensin receptor blocker use. Other comparisons yielded highly significant results. For the outcome

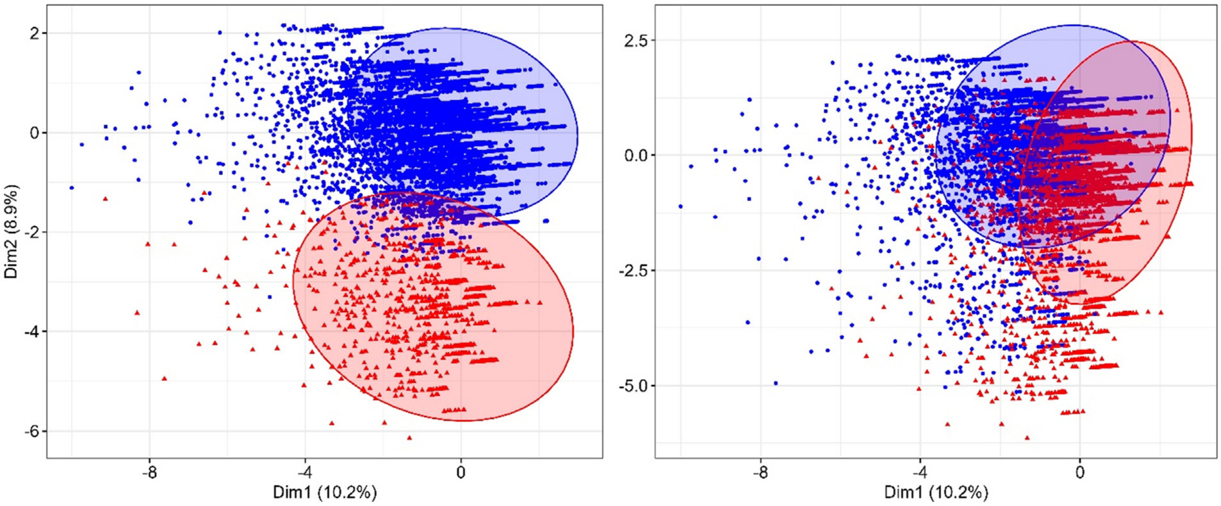

Figure 4. Clusters within the dataset using k-means *(left panel)* and the CLARA *(right panel)* algorithms. Observations are represented by points using principal components. Multivariate normal distributions have been assumed for drawing the 95% concentration ellipses.

Table 2
Comparisons of the clusters using the non-imputed original dataset, focusing on in-hospital management and outcomes (n=23,270)

Characteristic	All (N = 23,270)	k-means		p Value	CLARA		p Value
		k1 (n_1 = 21,988)	k2 (n_2 = 1,282)		C1(n_1 = 11,268)	C2 (n_2 = 12,002)	
Thrombolysis (%)	296 (1.49)	296 (1.51)	0 (0.00)	0.155	80 (0.83)	216 (2.12)	<0.001
Radial access (%)	9,204 (50.06)	9,158 (50.32)	46 (24.47)	<0.001	4,496 (50.82)	4,708 (49.35)	0.048
Multivessel disease (%)	8,668 (46.84)	8,476 (48.35)	192 (19.73)	<0.001	6,425 (74.96)	2,243 (22.58)	<0.001
Drug-eluting stent (%)	11,578 (50.61)	10,743 (49.14)	835 (82.19)	<0.001	5,850 (53.17)	5,728 (48.23)	<0.001
Complete revascularization (%)	10,124 (62.83)	9,683 (63.94)	441 (45.51)	<0.001	3,462 (48.88)	6,662 (73.77)	<0.001
Reinfarction (%)	258 (1.31)	253 (1.30)	5 (2.59)	0.209	147 (1.54)	111 (1.10)	0.008
Bleeding (%)	1,060 (5.35)	1,038 (5.29)	22 (11.40)	<0.001	540 (5.59)	520 (5.11)	0.144
Aspirin (%)	21,986 (94.49)	21,986 (100.00)	0 (0.00)	<0.001	10,811 (95.96)	11,175 (93.11)	<0.001
Clopidogrel (%)	15,075 (64.79)	14,667 (66.71)	408 (31.83)	<0.001	7,549 (67.01)	7,526 (62.71)	<0.001
Prasugrel (%)	3,281 (14.32)	2,588 (11.96)	693 (54.10)	<0.001	1,231 (11.18)	2,050 (17.23)	<0.001
Ticagrelor (%)	4,686 (20.45)	4,663 (21.56)	23 (1.80)	<0.001	2,353 (21.36)	2,333 (19.61)	0.001
Oral anticoagulant (%)	1,084 (4.75)	924 (4.29)	160 (12.49)	<0.001	581 (5.30)	503 (4.24)	<0.001
ACEI or ARB (%)	15,245 (76.07)	14,286 (76.15)	959 (74.86)	0.558	7,408 (77.58)	7,837 (74.70)	<0.001
Betablocker (%)	15,932 (79.47)	15,283 (81.44)	649 (50.66)	<0.001	7,553 (79.09)	8,379 (79.82)	0.210
Statin (%)	19,113 (92.92)	18,013 (93.38)	1,100 (85.87)	<0.001	9,010 (91.82)	10,103 (93.92)	<0.001
Proton-pump inhibitor (%)	7,859 (60.21)	6,802 (57.47)	1,057 (86.85)	<0.001	3,783 (66.09)	4,076 (55.62)	<0.001

ACEI = Angiotensin-Converting Enzyme Inhibitors; ARB = Angiotensin Receptor Blockers; C1 and C2 = Clusters 1 and 2 identified by the CLARA algorithm; k1 and k2 = Clusters 1 and 2 identified by the k-means algorithm.

Table 3
Comparisons of the clusters using the non-imputed original dataset, focusing on 2-year outcomes (n=23,270)

Characteristic	All (N = 23,270)	k-means		p Value	CLARA		p Value
		k1 (n_1 = 21,988)	k2 (n_2 = 1,282)		C1(n_1 = 11,268)	C2 (n_2 = 12,002)	
Death (%)	963 (4.14)	840 (3.82)	123 (9.59)	<0.001	538 (4.77)	425 (3.54)	<0.001
Reinfarction (%)	739 (3.93)	650 (3.70)	89 (7.17)	<0.001	402 (4.47)	337 (3.43)	<0.001
Nonfatal reinfarction (%)	548 (3.04)	482 (2.85)	66 (5.90)	<0.001	320 (3.75)	228 (2.40)	<0.001
Major bleeding (%)	724 (3.17)	647 (3.00)	77 (6.01)	<0.001	391 (3.57)	333 (2.80)	0.001
Nonfatal major bleeding (%)	596 (2.72)	548 (2.65)	48 (4.15)	0.003	321 (3.08)	275 (2.40)	0.002
Death or reinfarction (%)	1,511 (6.49)	1,322 (6.01)	189 (14.74)	<0.001	858 (7.61)	653 (5.44)	<0.001
Death or major bleeding (%)	1,559 (6.70)	1,388 (6.31)	171 (13.34)	<0.001	859 (7.62)	700 (5.83)	<0.001
Death, reinfarction, or major bleeding (%)	2,072 (8.90)	1,838 (8.36)	234 (18.25)	<0.001	1,158 (10.28)	914 (7.62)	<0.001
Nonfatal reinfarction or major bleeding (%)	1,109 (4.97)	998 (4.72)	111 (9.58)	<0.001	620 (5.78)	489 (4.22)	<0.001

C1 and C2 = Clusters 1 and 2 identified by the CLARA algorithm; k1 and k2 = Clusters 1 and 2 identified by the k-means algorithm.

variable death, the difference between the 2 clusters was statistically significant (chi-square test with Yates continuity correction $\chi^2_{(1)}$ = 100.360, p <0.001, unadjusted odds ratio [95% confidence interval] = 2.67 [2.17 to 3.26], p <0.001]). Other outcome variables were also significantly different between k1 and k2. This confirmed that we faced 2 clinical entities with respect to mortality in patients with the acute coronary syndrome. For the CLARA algorithm, there were no statistically significant differences in β-blocker use and in-hospital bleeding between C1 and C2.

Table 4
Comparisons of the clusters using the non-imputed original dataset, focusing on 2-year outcomes (n=23,270), using a 4-level clustering approach

Characteristic	4-level clustering group (k=k-means, C=CLARA)				p Value
	k1 and C2 (n = 11,175)	k1 and C1 (n = 10,813)	k2 and C2 (n = 827)	k2 and C1 (n = 455)	
Death (%)	351 (3.1%)	489 (4.5%)	74 (9.0%)	49 (10.8%)	<0.001
Reinfarction (%)	285 (3.2%)	365 (4.3%)	52 (6.5%)	37 (8.4%)	<0.001
Nonfatal reinfarction (%)	189 (2.2%)	293 (3.6%)	39 (5.4%)	27 (6.9%)	<0.0001
Major bleeding (%)	283 (2.6%)	364 (3.5%)	50 (6.1%)	27 (6.0%)	<0.001
Nonfatal major bleeding (%)	246 (2.3%)	302 (3.0%)	29 (3.9%)	19 (4.7%)	<0.001
Death or reinfarction (%)	540 (4.8%)	782 (7.2%)	113 (13.7%)	76 (16.7%)	<0.001
Death or major bleeding (%)	597 (5.3%)	791 (7.3%)	103 (12.5%)	68 (15.0%)	<0.001
Death, reinfarction, or major bleeding (%)	774 (6.9%)	1,064 (9.8%)	140 (16.9%)	94 (20.7%)	<0.001
Nonfatal reinfarction or major bleeding (%)	423 (3.9%)	575 (5.6%)	66 (8.8%)	45 (11.1%)	<0.001

C1 and C2: Clusters 1 and 2 identified by the CLARA algorithm; k1 and K2: Clusters 1 and 2 identified by the k-means algorithm.

Table 5
Multivariable analysis of the clusters using the non-imputed original dataset, focusing on 2-year outcomes (n=23,270), using a 4-level clustering approach (similar results were obtained when using repeatedly 2-level clustering according to k-means and CLARA)

Characteristic	4-level clustering group (k=k-means, C=CLARA)*				Lowest p Value
	k1 and C2 (n = 11,175)	k1 and C1 (n = 10,813)	k2 and C2 (n = 827)	k2 and C1 (n = 455)	-
Death (%)	Reference	0.99 (0.73-1.35)	3.50 (0.66-18.52)	0.69 (0.20-2.42)	0.141
Reinfarction (%)	Reference	1.01 (0.73-1.40)	Not estimable	1.46 (0.43-4.94)	0.545
Nonfatal reinfarction (%)	Reference	1.08 (0.76-1.54)	Not estimable	2.04 (0.60-7.00)	0.255
Major bleeding (%)	Reference	1.37 (1.02-1.83)	2.23 (0.28-17.68)	Not estimable	0.039
Nonfatal major bleeding (%)	Reference	1.42 (1.04-1.95)	3.46 (0.43-27.53)	Not estimable	0.027
Death or reinfarction (%)	Reference	1.04 (0.82-1.32)	2.25 (0.46-11.05)	1.10 (0.44-2.77)	0.318
Death or major bleeding (%)	Reference	1.21 (0.97-1.51)	3.41 (0.87-13.39)	0.50 (0.15-1.68)	0.079
Death, reinfarction, or major bleeding (%)	Reference	1.17 (0.97-1.43)	2.57 (0.67-9.85)	0.85 (0.34-2.09)	0.717
Nonfatal reinfarction or major bleeding (%)	Reference	1.26 (0.99-1.60)	1.94 (0.25-15.20)	0.98 (0.29-3.28)	0.059

C1 and C2 = Clusters 1 and 2 identified by the CLARA algorithm; k1 and K2 = Clusters 1 and 2 identified by the k-means algorithm.

* adjusting for age, diabetes mellitus, peripheral artery disease, prior bleeding, malignancy, estimated glomerular filtration rate, hemoglobin concentration, left ventricular ejection fraction, multivessel disease, drug-eluting stent implantation, and discharge therapy with angiotensin-converting enzyme inhibitors, angiotensin receptor blockers, and statins, and reported as odds ratio (95% confidence interval).

For radial access, the result of the chi-square test was significant (p = 0.048); however, because of the large sample size and increased likelihood of type I error, we considered the difference as practically nonsignificant. Other comparisons yielded highly significant results.

For the outcome variable death, the difference between the 2 clusters was statistically significant (chi-square test with Yates continuity correction $\chi^2_{(1)}$ = 21.980, p <0.001, unadjusted odds ratio (95% confidence interval) = 0.73 [0.64 to 0.83], p <0.001; Table 4). Other outcome variables were also significantly different between C1 and C2. This showed that the 2 clusters are different with respect to mortality.

To describe the resulting clusters more practically, we developed 2 models for predicting cluster membership using logistic regression with 10-fold cross-validation. No synthetic data were used for cluster profiling. First, the predictors with severe imbalance (less than 10% in 1 of their levels) were excluded. Second, the highly correlated pairs of features were removed by selecting 1 variable from each pair. Third, the cases with missing data were filtered out. For the outcome variable, k-means cluster membership, the class imbalance was eliminated by random down-sampling of the majority class. Furthermore, for the k1 cluster (n = 21,988) and for the k2 (n = 1,282), 7.1% and 13.1% of predictor data were missing, respectively. This additional inequality in missing data affected the process of model development for k-means. The resulting models suggested good predictive ability for both models (Supplementary Table 1). However, the Hosmer-Lemeshow test indicated a poor fitness for the k-means model. Also, the Brier calibration score was not satisfactory for the k-means algorithm. The Akaike information criterion was smaller for k-means; whereas, Nagelkerke R^2 was larger for the CLARA. Both models showed favorable sensitivity, specificity, predicted values, and C-statistic. The model specification for k-means suggested that higher age, female gender, STEMI, and particularly being treated with a drug-eluting stent (DES) were associated with the high-risk cluster (Supplementary Table 2). Conversely, patients with dyslipidemia, previous myocardial infarction, and particularly noninsulin-dependent diabetes were more categorized in the low-risk cluster. For the CLARA algorithm, the model was highly affected by STEMI. Other variables, except for female gender, were more associated with the low-risk cluster.

We finally appraised the incremental predictive role of cluster assignment, distinguishing 4 different categories based on the k-means and CLARA clustering approaches (Table 5). This 4-tier classification system proved the limited incremental prognostic accuracy in comparison to other clinical features (p >0.05) for death, reinfarction, or their composite. However, it proved significantly associated with the risk of major bleeding, even at extensive multivariable adjustment (p = 0.039 for fatal or nonfatal bleeding, p = 0.027 for nonfatal bleeding).

Discussion

This study originally provides insights on the pros and cons of machine learning techniques in profiling patients with unstable coronary artery disease. In particular, we found that unsupervised machine learning can promisingly be leveraged to explore patterns in such individuals, potentially highlighting specific patient subsets to improve risk stratification and management.

Indeed, we conducted this study to find if there are clinically meaningful clusters in patients with ACS beyond the current descriptions. Our study showed that there are 2 clinical entities among those patients. We used 2 known algorithms for clustering our data: k-means and the CLARA. In general, k-means clustering showed that patients in k2 were at 2 to 3 times greater risk for poor outcomes than k1. The mean glomerular filtration rate and the percentage of glomerular filtration rate more than 60 ml/min were higher in k1. Noninsulin-dependent diabetes, dyslipidemia, prior coronary artery problems (particularly multivessel disease), radial access, thrombolysis, and complete revascularization were more frequent in k1. They all used anabolic-androgenic steroids and also used ticagrelor, clopidogrel, and β-blockers more than patients in k2. The percentage of women was higher in k2. Patients in k2 had insulin-dependent diabetes, a previous history of major bleeding, in-hospital events, and DES use more frequently. They used prasugrel, oral anticoagulation, β-blockers, and proton-pump inhibitors more commonly. None of the patients in k2 used anabolic-androgenic steroids.

The CLARA showed that patients in C1 were more prone to poor outcomes. Overall, most of the patients in C1 had non-ST-segment elevation acute coronary syndrome. The percentage of women was higher in C1, and commonly, they were older patients with arterial hypertension, lower hemoglobin concentration, and increased serum creatinine. Although the mean glomerular filtration rate was lower in C1, the percentage of glomerular filtration rate more than 60 ml/min was higher than in C2. Noninsulin-dependent diabetes, dyslipidemia, previous coronary artery problems, particularly multivessel disease, were also more frequent in C1. They used proton-pump inhibitors more than C2. Patients in C2 had STEMI, complete revascularization, thrombolysis, and prasugrel use more frequently and their mean hemoglobin and glomerular filtration rate were higher than patients in C1.

Furthermore, model performance and specification were not favorable for both models. For the k-means model, this might be due to class imbalance and missing data, and for the CLARA model, it might be because of the crucial impact of infarction type on the whole model. In turn, the missing data in the high-risk k-means cluster might be explained as the deficiency of information that is commonly is seen in patients experiencing poor outcomes. Reassessing these concepts in a further study using another separate dataset would delineate the effects of clustering in evaluating the prognosis of patients with ACSs.

The idea of applying unsupervised machine learning, including cluster analysis, to distinguish the specific patient subsets with cardiovascular disease is not novel, and indeed, it holds the promise of proving potentially beneficial to guide prevention, diagnosis, risk stratification, prognosis, management, and rehabilitation.[9] However, most reports on the use of this powerful analytic techniques focus on stable coronary artery disease, heart failure, COVID-19, or features that only indirectly affect patient outcomes (e.g., climate).[10−14] For instance, unsupervised machine learning analyses aimed at improving the characterization of patients with heart failure have been recently reported, albeit leveraging samples of limited size.[15,16] Jani and colleagues have appraised the diagnostic and prognostic value of specific echocardiographic features in patients hospitalized for COVID-19, despite leveraging a relatively small dataset of 176 patients.[11] Koo et al[12] and Testa et al[13] have instead independently appraised the environmental features that may pose, in general terms and in specific patient subsets, an increased risk of acute myocardial infarction or stroke. Notably, no recent and adequately sized work has been previously published focusing on unsupervised learning among patients with ACSs, and this holds even truer when focusing on the features encompassing also those available at discharge.

A number of important considerations should be perused in detail to avoid misinterpreting our present findings.[17] First, a key first novelty of our work is, on 1 hand, the focus on patients at discharge after an ACS rather than at admission. Accordingly, in-hospital events can be viewed as pieces of clinical history (despite them being recent) capable of poignantly characterizing the patients.[18] The second strength of our work is the reliance on an unsupervised machine learning approach, which considers all patient features possibly relevant to define specific subsets, such that patients belonging to each subgroup are much more homogeneous among themselves rather than in comparison to other subgroups. This approach exploits information-rich variables, irrespective of their pathophysiologic or clinical premises, including but limiting itself to baseline features, procedural details, and short-term outcomes. Indeed, the reliance on proxy variables, which are associated with other features, despite not being mechanistically linked to them, is not a rare occurrence in cardiovascular research. For instance, it has been reported that ear lobe creases are associated with increased cardiovascular risk, but evidently, this is simply due to a correlation phenomenon.[19]

Second, the key rationale of unsupervised machine learning is indeed the lack of any a priori relation between a variable and a pathophysiologically relevant mechanistic proxy or a clinical outcome. Indeed, we know, for instance, that dyslipidemia is associated with adverse outcomes thanks to the hundreds of studies detailing on this association and consider it, by default, relevant. The cluster analysis instead leverages a specific variable, such as insulin-dependent diabetes mellitus, only if it is capable of discriminating the different subtypes of patients within the dataset at the end.[20] Thus, the lack of explicit focus on pathophysiologic credibility or clinical impactfulness is not a weakness per se but rather an explicit property of the approach we have showcased in our study.

Third, several potential explanations for the unexpected higher prevalence of apparently detrimental risk factors, such as dyslipidemia, previous myocardial infarction, and noninsulin-dependent diabetes, in clusters exhibiting a lower risk of events at long-term. Indeed, the unsupervised machine learning perspective uses each patient feature, such as hypertension, as a marker to characterize patients, and thus, it may indeed occur that patients with dyslipidemia could have a lower prevalence of other more severe risk factors (this could apply for instance to noninsulin-dependent diabetes because this feature would compete by definition with the presence of insulin-dependent diabetes). Pathophysiologically, another potential explanation is that some risk factors, such as dyslipidemia, could be, at least in part, treatable and thus suitable for prognostic improvement. Indeed, an ACS in a patient without dyslipidemia would likely have a better long-term prognosis than an ACS in a patient with dyslipidemia. However, an ACS in a patient with dyslipidemia who gets an optimally intensive lipid lowering therapy could eventually display a better outcome than an ACS in a patient without dyslipidemia. Focusing instead on management strategies, such as the apparent association between DESs and adverse outcomes, we may speculate that this being an observational study spanning a long period of time, DESs were selectively used in patients with more complex lesions and diffuse disease (e.g., left main disease, bifurcation lesions, chronic total occlusions, and so forth).

Despite our work novelty and the meaningful sample size, our work has many drawbacks. First, it should be viewed as a case study highlighting the potential pros of novel approaches at characterizing patients with ACSs rather than a conclusive piece demonstrating a specific hypothesis.[17] Second, there is no algorithm for unsupervised machine learning, which is inherently superior to others, and thus, our methodologic framework and the accompanying results can be viewed as relevant and valid but do not exclude other analytic approaches and conclusions. Third, any unsupervised machine learning algorithm may provide spuriously precise results because it cannot in any way, despite its refined features, eliminate issues in the originating dataset. Thus, it is crucial to pay attentive scrutiny to the PRAISE dataset and its strengths and limitations when interpreting our own results.[7] Indeed, additional details on risk baseline features, clinical variables upon presentation, and in-hospital therapies, as well as complications, would have provided additional insights on the population features and possibly impacted on the results of the analysis. However, the PRAISE study was designed with a specific pragmatic data collection protocol, and such additional variables were not collected, as clearly stated in the main study.[7] Finally, the fact that most of the predictive input from clusters was lost at the multivariable analysis suggest that the cluster analysis cannot discover any hidden and secret feature capable of distinguishing patients according to their prognosis but rather that it can be viewed as a different yet potentially more succinct way at disentangling complex data patterns.

Unsupervised machine learning may provide crucial insights in the comparative features of patients with ACSs, which can prove useful for clinicians to improve decision making, and to researchers to open new avenues for investigation.

Disclosures

Giuseppe Biondi-Zoccai has consulted for Amarin, Balmed, Cardionovum, Crannmedical, Endocore Lab, Eukon, Innovheart, Guidotti, Meditrial, Microport, Opsens Medical, Replycare, Teleflex, Terumo, and Translumina. All other authors have no conflicts of interest to declare.

Supplementary materials

Supplementary material associated with this article can be found in the online version at https://doi.org/10.1016/j.amjcard.2023.01.048.

1. GBD 2013. Mortality and causes of death collaborators. Global, regional, and national age-sex specific all-cause and cause-specific mortality for 240 causes of death, 1990-2013: a systematic analysis for the Global Burden of Disease Study 2013. *Lancet* 2013;385:117–171.
2. Saglietto A, Manfredi R, Elia E, D'Ascenzo F, DE Ferrari GM, Biondi-Zoccai G, Munzel T. Cardiovascular disease burden: Italian and global perspectives. *Minerva Cardiol Angiol* 2021;69:231–240.
3. Mitsis A, Gragnano F. Myocardial infarction with and without ST-segment elevation: a contemporary reappraisal of similarities and differences. *Curr Cardiol Rev* 2021;17:e230421189013.
4. Khandkar C, Madhavan MV, Weaver JC, Celermajer DS, Karimi Galougahi K. Atherothrombosis in acute coronary syndromes-from mechanistic insights to targeted therapies. *Cells* 2021;10:865.
5. Arfat Y, Mittone G, Esposito R, Cantalupo B, DE Ferrari GM, Aldinucci M. Machine learning for cardiology. *Minerva Cardiol Angiol* 2022;70:75–91.
6. Guo CY, Yang YC, Chen YH. The optimal machine learning-based missing data imputation for the cox proportional hazard model. *Front Public Health* 2021;9:680054.
7. D'Ascenzo F, De Filippo O, Gallone G, Mittone G, Deriu MA, Iannaccone M, Ariza-Solé A, Liebetrau C, Manzano-Fernández S, Quadri G, Kinnaird T, Campo G, Simao Henriques JP, Hughes JM, Dominguez-Rodriguez A, Aldinucci M, Morbiducci U, Patti G, Raposeiras-Roubin S, Abu-Assi E, De Ferrari GM, PRAISE study group. Machine learning-based prediction of adverse events following an acute coronary syndrome (PRAISE): a modelling study of pooled datasets. *Lancet* 2021;397:199–207.
8. Mehran R, Rao SV, Bhatt DL, Gibson CM, Caixeta A, Eikelboom J, Kaul S, Wiviott SD, Menon V, Nikolsky E, Serebruany V, Valgimigli M, Vranckx P, Taggart D, Sabik JF, Cutlip DE, Krucoff MW, Ohman EM, Steg PG, White H. Standardized bleeding definitions for cardiovascular clinical trials: a consensus report from the Bleeding Academic Research Consortium. *Circulation* 2011;123:2736–2747.
9. Alonso-Betanzos A, Bolón-Canedo V. Big-data analysis, cluster analysis, and machine-learning approaches. *Adv Exp Med Biol* 2018;1065:607–626.
10. Burrello J, Gallone G, Burrello A, Jahier Pagliari D, Ploumen EH, Iannaccone M, De Luca L, Zocca P, Patti G, Cerrato E, Wojakowski W, Venuti G, De Filippo O, Mattesini A, Ryan N, Helft G, Muscoli S, Kan J, Sheiban I, Parma R, Trabattoni D, Giammaria M, Truffa A, Piroli F, Imori Y, Cortese B, Omedè P, Conrotto F, Chen SL, Escaned J, Buiten RA, Von Birgelen C, Mulatero P, De Ferrari GM, Monticone S, D'Ascenzo F. Prediction of all-cause mortality following percutaneous coronary intervention in bifurcation lesions using machine learning algorithms. *J Pers Med* 2022;12:990.
11. Jani V, Kapoor K, Meyer J, Lu J, Goerlich E, Metkus TS, Madrazo JA, Michos E, Wu K, Bavaro N, Kutty S, Hays AG, Mukherjee M. Unsupervised machine learning demonstrates the prognostic value of TAPSE/PASP ratio among hospitalized patients with COVID-19. *Echocardiography* 2022;39:1198–1208.
12. Koo GPY, Zheng H, Pek PP, Hughes F, Lim SL, Yeo JW, Ong MEH, Ho AFW. Clustering of environmental parameters and the risk of acute myocardial infarction. *Int J Environ Res Public Health* 2022;19:8476.
13. Testa A, Biondi-Zoccai G, Anticoli S, Pezzella FR, Mangiardi M, DI Giosa A, Marchegiani G, Frati G, Sciarretta S, Perrotta A, Peruzzi M, Cavarretta E, Gaspardone A, Mariano E, Federici M, Montone RA, Dei Giudici A, Versaci B, Versaci F. Cluster analysis of weather and pollution features and its role in predicting acute cardiac or cerebrovascular events. *Minerva Med* 2022;113:825–832.
14. Bose EL, Clermont G, Chen L, Dubrawski AW, Ren D, Hoffman LA, Pinsky MR, Hravnak M. Cardiorespiratory instability in monitored step-down unit patients: using cluster analysis to identify patterns of change. *J Clin Monit Comput* 2018;32:117–126.
15. Horiuchi Y, Tanimoto S, Latif AHMM, Urayama KY, Aoki J, Yahagi K, Okuno T, Sato Y, Tanaka T, Koseki K, Komiyama K, Nakajima H, Hara K, Tanabe K. Identifying novel phenotypes of acute heart failure using cluster analysis of clinical variables. *Int J Cardiol* 2018;262:57–63.
16. Urban S, Błaziak M, Jura M, Iwanek G, Zdanowicz A, Guzik M, Borkowski A, Gajewski P, Biegus J, Siennicka A, Pondel M, Berka P, Ponikowski P, Zymliński R. Novel phenotyping for acute heart failure-unsupervised machine learning-based approach. *Biomedicines* 2022;10:1514.
17. Cleophas TJ, Zwinderman AH. *Machine Learning in Medicine − A Complete Overview*. 2nd ed. Cham: Springer; 2020.
18. Marquis-Gravel G, Dalgaard F, Jones AD, Lokhnygina Y, James SK, Harrington RA, Wallentin L, Steg PG, Lopes RD, Storey RF, Goodman SG, Mahaffey KW, Tricoci P, White HD, Armstrong PW, Ohman EM, Alexander JH, Roe MT. Post-discharge bleeding and mortality following acute coronary syndromes with or without PCI. *J Am Coll Cardiol* 2020;76:162–171.
19. Cumberland GD, Riddick L, Vinson R. Earlobe creases and coronary atherosclerosis. The view from forensic pathology. *Am J Forensic Med Pathol* 1987;8:9–11.
20. Sanders WE Jr, Burton T, Khosousi A, Ramchandani S. Machine learning: at the heart of failure diagnosis. *Curr Opin Cardiol* 2021;36:227–233.