Computational and Structural Biotechnology Journal 23 (2024) 287–294

Contents lists available at ScienceDirect

Computational and Structural Biotechnology Journal

journal homepage: www.elsevier.com/locate/csbj

Research Article

Unsupervised machine learning for risk stratification and identification of relevant subgroups of ascending aorta dimensions using cardiac CT and clinical data

Mario Zanfardino [a], Bruna Punzo [a,*], Erica Maffei [b], Luca Saba [c], Eduardo Bossone [d], Stefano Nistri [e], Ludovico La Grutta [f], Monica Franzese [a], Carlo Cavaliere [a], Filippo Cademartiri [b]

[a] IRCCS Synlab SDN, Naples, 80143, Italy
[b] Department of Imaging, Fondazione Monasterio/CNR, Pisa, 56124, Italy
[c] Department of Radiology, University Hospital of Cagliari, Cagliari, 09042, Italy
[d] Department of Public Health, University of Naples Federico II, Naples, 80131, Italy
[e] Department of Cardiology, CMSR-GHC, Vicenza, Italy
[f] University of Palermo, Palermo, Italy

ARTICLE INFO

Keywords:
Aortic dimensions
Computed tomography coronary angiography
Unsupervised learning
Clusterization
Coronary artery disease

ABSTRACT

The potential of precision population health lies in its capacity to utilize robust patient data for customized prevention and care targeted at specific groups. Machine learning has the potential to automatically identify clinically relevant subgroups of individuals, considering heterogeneous data sources. This study aimed to assess whether unsupervised machine learning (UML) techniques could interpret different clinical data to uncover clinically significant subgroups of patients suspected of coronary artery disease and identify different ranges of aorta dimensions in the different identified subgroups. We employed a random forest-based cluster analysis, utilizing 14 variables from 1170 (717 men/453 women) participants. The unsupervised clustering approach successfully identified four distinct subgroups of individuals with specific clinical characteristics, and this allows us to interpret and assess different ranges of aorta dimensions for each cluster. By employing flexible UML algorithms, we can effectively process heterogeneous patient data and gain deeper insights into clinical interpretation and risk assessment.

1. Introduction

Precise and consistent measurements of aortic diameters are critical for diagnosing, categorizing, and monitoring aortic pathologies, as well as determining the appropriate timing for follow-up and selecting candidates for surgery [1]. In more detail, the dimensions of the ascending aorta are clinically relevant because they play a crucial role in assessing the risk of aortic pathologies [2–6]. Conditions of interest related to the ascending aorta include: Aortic Aneurysms (early detection and monitoring of aortic aneurysms are vital to prevent potentially fatal complications), Aortic Dissection (a medical emergency, requiring immediate treatment), Aortic Valve Disease (the assessment of aortic di-

Statement of significance:	
Problem	Define “normal” ascending aorta and ranges of aortic dimensions is critical for diagnosing and monitoring aortic pathologies and there are no reliable reference values. Moreover, the size of the ascending aorta in isolation is not sufficient to define normal vs. borderline vs. diseased
What is Already Known	The segment of the ascending aorta is still challenging to identify and link to clear phenotypes. The tools for assessment, the available values, the population for the study, and so forth often limit our capabilities to comprehend the basic phenotypes and the early features that may lead to significant clinical issues
What This Paper Adds	Our study aimed to identify subgroups of “normal” ascending aorta and ranges of aortic dimensions that could represent distinct clinical profiles by applying a machine learning method on computed tomography (CT) data blended with clinical characteristics and cardiovascular risk factors data

* Corresponding author.
E-mail address: bruna.punzo@synlab.it (B. Punzo).

https://doi.org/10.1016/j.csbj.2023.11.021
Received 14 September 2023; Received in revised form 10 November 2023; Accepted 10 November 2023
Available online 29 November 2023

mensions is essential in planning surgical interventions for valve repair or replacement), Connective Tissue Disorders (conditions like Marfan syndrome and Ehlers-Danlos syndrome can lead to aortic root enlargement, increasing the risk of aortic dissection). Then, monitoring the dimensions of the ascending aorta is critical to detect and manage aortic pathologies, aneurysms, dissections, and related conditions. Early intervention can prevent life-threatening complications, making accurate measurement and assessment a vital aspect of patient care. Over the past decades, non-invasive imaging of the thoracic aorta had remarkable advancements, involving different imaging techniques including trans-thoracic echocardiography (TTE), trans-esophageal echocardiography, magnetic resonance imaging (MRI), computed tomography (CT), and conventional catheter angiography. Among these, CT is the most robust, reliable, and accurate non-invasive technique widely used in clinical practice both in acute and stable patients' settings; CT also allows the concomitant assessment of thoracic aorta and coronary arteries [7–9]. The normal diameters and ranges for thoracic aortic size have been established historically using ultrasound [10–17], and more recently reassessed also with CT and MR Angiography [18–21].

Despite the burden of data available, it is still difficult to identify clear phenotypes in the segment of the ascending aorta where most relevant diseases occur; the tools for assessment, the numbers available, the population for the study, and so forth were often limiting our capabilities to comprehend the basic phenotypes and the early features that may lead to significant clinical issues. The size of the ascending aorta in isolation is clearly not sufficient to define normal vs. borderline vs. diseased.

Utilizing methods that capture and integrate a broader range of patient characteristics, without the need for hundreds of variables in a single patient, becomes crucial for gaining a better understanding of prognosis and effectively targeting intensive risk-reduction therapies.

UML may identify complex patterns and interactions among data without any pre-existing labels. Clustering is a frequently used UML technique that groups observations into clusters based on shared characteristics. Moreover, Random Forest (RF), which is not a native clustering technique, could be used to create distance metrics that feed into traditional clustering methods such as K-means.

The purposes of the study were to apply an RF UML method on CT data blended with clinical characteristics and cardiovascular risk factors to identify subgroups of "normal" ascending aorta and ranges of aortic dimensions that could represent distinct clinical profiles in a large cohort of consecutive patients undergoing CT coronary angiography (CTCA) for suspected obstructive coronary artery disease (CAD). The identified clusters of patients may become useful to customize interpretation, and care, uncover patterns in population health, and offer a more precise risk assessment. In more detail, we aim to define clinical profiles based on ascending aorta dimensions and risk factors without categorically specifying what is pathological. This approach allows us to identify various clinical and physical characteristics among subgroups without establishing an absolute criterion for pathology. In other words, we seek to define clinical profiles based on a range of aortic dimensions and risk factors, thus enabling the assessment of variations in clinical presentation among these subgroups. We acknowledge that the concept of "normal" or "pathological" can vary depending on multiple individual and clinical factors. Therefore, we do not intend to establish a static definition of what is pathological but rather to identify how different patient characteristics are reflected in aortic dimensions and risk factors.

2. Materials and methods

2.1. Dataset

A total of 1170 (Table 1) consecutive patients were evaluated retrospectively for suspected CAD. These patients met the usual inclusion criteria for CTCA with 64 slice-CT equipment (i.e. stable heart rate with sinus rhythm and the ability to hold their breath for at least 12 seconds). The patients included 717 men and 453 women (age: 62.70 ± 12.80 years). All relevant patient demographics and clinical data were prospectively collected from medical records, including age, gender, weight, height, and cardiovascular risk factors such as family history of aortic disease, smoking status, diabetes, dyslipidemia, hypertension, and obesity. Dyslipidemia, diabetes, and hypertension were defined according to current guidelines. All patients underwent CTCA for coronary artery assessment, which always included the ascending aorta. Exclusion criteria included known aortic disease, bicuspid aortic valve, previous coronary revascularization, previous acute myocardial infarction or severe heart failure, severe renal impairment, atrial fibrillation, thyroid disorders, unstable clinical condition, known allergy to iodinated contrast agents and pregnancy.

2.2. Scan protocol and image evaluation

The scans were conducted using a 64-slice CT scanner. Prior to the angiographic study, an unenhanced acquisition was performed to assess the distribution and amount of coronary calcium. The CTCA was performed following the intravenous administration of 80 to 100 mL of high iodine concentration contrast agent at a rate of 4 to 5 mL/s, followed by a 40 to 50 mL saline chaser at the same rate. A bolus-tracking technique was employed to synchronize the contrast arrival in the coronary arteries and initiate the scan. The parameters were set as follows: retrospective ECG-gating with modulation dose, collimation 32×2×0.6 mm, gantry rotation time 330 ms, feed/rotation 3.84 mm, effective slice width 0.75 mm, increment 0.4 mm, medium-smooth B30f reconstruction kernel, kV 120, and mAs 700 to 900 (depending on the patient's features). The temporal windows for ECG-gated retrospective reconstructions were established at the end-diastolic and end-systolic phases. Two experienced operators analyzed the image data in agreement using an offline workstation. Aortic diameters and areas were measured using inner-to-inner and intimal lumen contour techniques respectively, on diastolic datasets at conventional and reproducible anatomic landmarks perpendicular to the vessel axis. The aortic root (AoR), sinotubular junction (STJ), and tubular ascending aorta (AAo) were measured at the maximum diameter, narrowest level in the transition of AoR to the ascending aorta, and at the level of the right pulmonary artery respectively. For details on the dataset and on image acquisition methods refer to our previous publication [22].

2.3. Machine learning analysis

Our exploratory analysis is based on a UML method, Random Forest [23], applied to suspected coronary artery disease patients. All analyses were performed, on males and females separately, in R v4.2.1. Data were cleaned through several steps of data preparation starting from 717/453 (Males/Females) patients and 19 features (age, sex, weight, height, BMI - Body Mass Index, BSA -Body Surface Area, n° of risk factors, smoking status, familiarity, diabetes, dyslipidemia, hypertension, obesity, AoR area and diameter, STJ area and diameter, and AAo area and diameter). The removal of highly correlated variables (features with correlation >0.9 were removed using "stats" v4.2.1 package using Pearson correlation coefficient) resulted in the following 14 remaining features: age, height, BMI, BSA, n° of risk factors, smoking status, familiarity, diabetes, hypertension, cholesterol, obesity, AoR diameter, STJ diameter, AAo diameter. All available variables were heterogeneous - continuous for weight, height, BMI and BSA and dichotomous for smoking status, familiarity, diabetes, dyslipidemia, hypertension, and obesity. Outliers were removed using the interquartile range (IQR) method on features of interest (values of AoR, STJ, and AAo). All values lower than the first quarter and higher than the third were eliminated. After the process described a population of 649 males and 421 females was available for further analysis. On both datasets, we applied the Ran-

Table 1
Baseline characteristics of sample data. SD = Standard Deviation; n = number of samples.

	Female (n = 421)	Male (n = 649)	Total (n = 1070)	p.value
Age				< 0.001
- Mean (SD)	65.145 (11.945)	60.649 (12.972)	62.418 (12.763)	
- Range	24.000 - 90.000	19.000 - 94.000	19.000 - 94.000	
Height				< 0.001
- Mean (SD)	161.259 (6.260)	173.445 (7.195)	168.650 (9.069)	
- Range	140.000 - 180.000	145.000 - 203.000	140.000 - 203.000	
BMI				< 0.001
- Mean (SD)	26.119 (4.920)	27.281 (3.837)	26.824 (4.331)	
- Range	13.560 - 50.781	17.066 - 42.608	13.560 - 50.781	
BSA				< 0.001
- Mean (SD)	1.737 (0.183)	1.985 (0.182)	1.887 (0.219)	
- Range	1.196 - 2.404	1.329 - 2.589	1.196 - 2.589	
n° Risk Factor				0.106944444444444
0	35 (8.3%)	68 (10.5%)	103 (9.6%)	
-1	84 (20.0%)	129 (19.9%)	213 (19.9%)	
-2	130 (30.9%)	170 (26.2%)	300 (28.0%)	
-3	114 (27.1%)	179 (27.6%)	293 (27.4%)	
-4	47 (11.2%)	66 (10.2%)	113 (10.6%)	
-5	10 (2.4%)	30 (4.6%)	40 (3.7%)	
-6	1 (0.2%)	7 (1.1%)	8 (0.7%)	
Familiarity				0.00208333333333333
- No	195 (46.3%)	361 (55.6%)	556 (52.0%)	
- Yes	226 (53.7%)	288 (44.4%)	514 (48.0%)	
Smoking				< 0.001
- No	325 (77.2%)	402 (61.9%)	727 (67.9%)	
- Yes	96 (22.8%)	247 (38.1%)	343 (32.1%)	
Diabetes Mellitus				0.127777777777778
- No	363 (86.2%)	540 (83.2%)	903 (84.4%)	
- Yes	58 (13.8%)	109 (16.8%)	167 (15.6%)	
Hypertension				0.0729166666666667
- No	153 (36.3%)	268 (41.3%)	421 (39.3%)	
- Yes	268 (63.7%)	381 (58.7%)	649 (60.7%)	
Dyslipidemia				0.197222222222222
- No	211 (50.1%)	347 (53.5%)	558 (52.1%)	
- Yes	210 (49.9%)	302 (46.5%)	512 (47.9%)	
Obesity				0.09375
- No	349 (82.9%)	514 (79.2%)	863 (80.7%)	
- Yes	72 (17.1%)	135 (20.8%)	207 (19.3%)	

dom Forest function from “Random Forest” (R package v4.7.1) to build a proximity matrix that contains a similarity measure. The algorithm was used in unsupervised mode by setting the outcome variable y = NULL and to select the best results, we fine-tuned Random Forest using different values of maxnodes parameter. The algorithm generates a proximity matrix that gives a rough estimate of the distance between samples based on the proportion of times the samples end up in the same leaf node; Using this matrix as input we performed a normal clustering procedure with PAM (Partitioning Around Medoids). “cluster” R package v2.1.4 with the “pam” function was used for clustering the RF results data into 2 clusters “around medoids” (an alternative and more robust version of K-means, PAM is an algorithm that searches for k representative objects in a data set (k medoids) and then assigns each object to the closest medoid in order to create clusters) [24]. In addition, to explore which features were most informative about the different identified clusters we used the variable importance obtained from RF. The latter was normalized (feature scaling using Min-max normalization) to obtain a number between 0 and 1. Subsequently, for each sample, we calculated a score by summing the value of the variable multiplied by the weight obtained from the above normalization (we excluded variables with a weight less than 0.4, because variables below this threshold we found that they did not significantly change the score value). A “best-threshold” was calculated using the “coords” function from “pROC” R package v1.18 on the comparison of data from

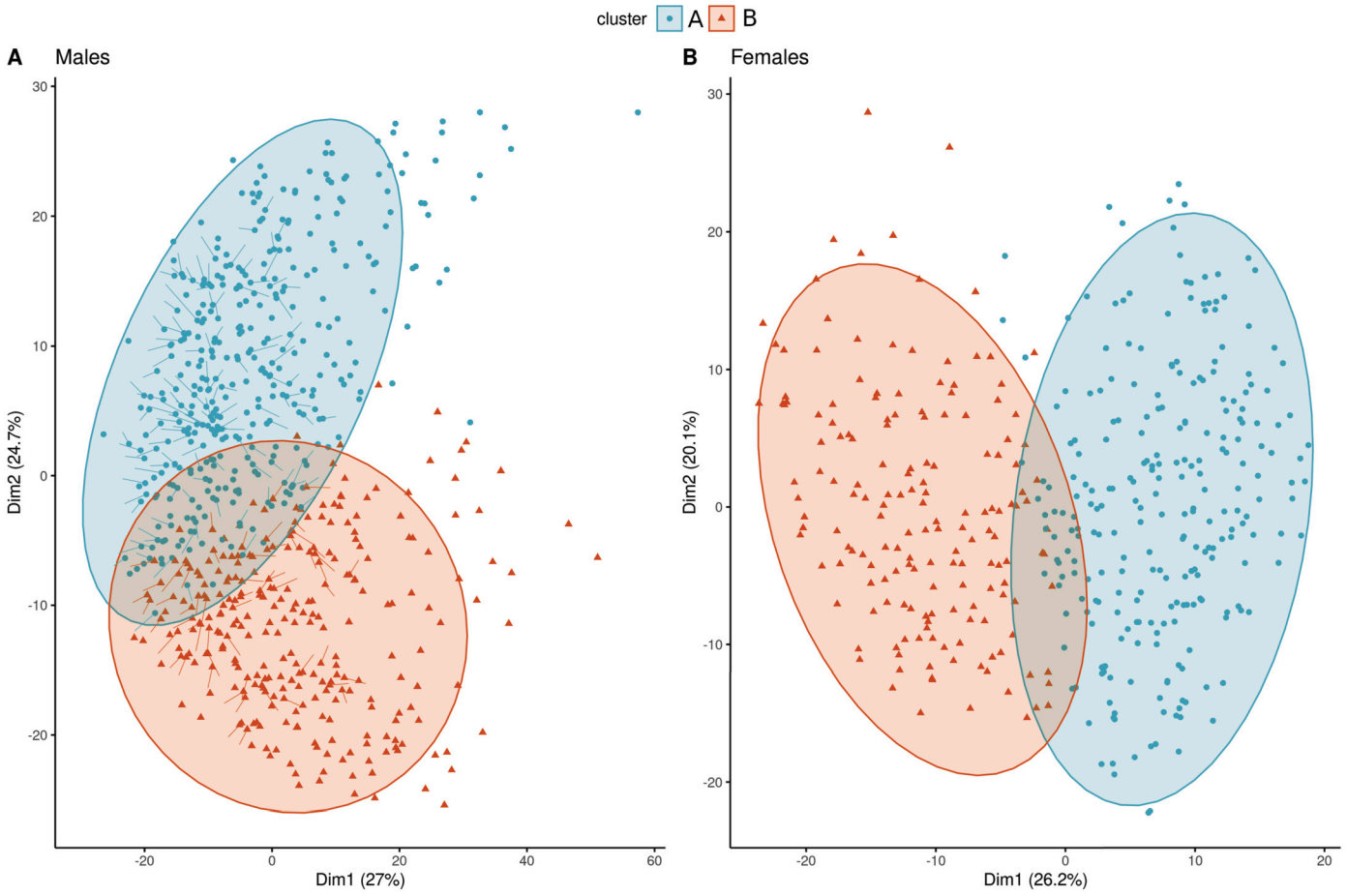

Fig. 1. Cluster subgroups identified by unsupervised clustering. Plot showing 4 groups (Males A, Males B, Females A, Females B) of individuals identified by Random Forest and PAM (Partitioning Around Medoids) analysis.

random forest clustering and the score values. The classification was evaluated using the following metrics: AUC, Accuracy, Sensitivity, and Specificity. All metrics were calculated by the confusion Matrix function from "ModelMetrics" R package v1.2.2.2.

2.4. Statistical analysis

All characteristics from the identified clusters were compared. Wilcoxon tests were performed to evaluate statistical significance of AoR, STJ, AAo values in the two clusters. Differences in categorical variables were compared by chi-squared test while a group F-test was used to compare numeric variables.

3. Results

The workflow of all analyses was represented in the graphical abstract. Performing unsupervised RF-based clustering on males and females separately we identified 2 patient subgroups for males and 2 subgroups for females. In the two resulting clusters (Fig. 1), and in both males and females, the distribution of all features of interest (AoR, STJ, AAo) have a significant difference ($p < 0.05$, wilcoxon test) (Supplementary figure 1). Moreover, there are several significant differences in the distribution of some variables (BMI, number of Risk Factors, and age) in the two clusters in both males and females (Table 3 and Supplementary Table 1). As shown in Supplementary figure 2 there was a clear difference between low and high BMI distribution across the clusters. Also, the different distribution of a number of risk factors is especially evident in male clusters. Regarding age groups, there was a clear difference, especially in male individuals with age <50 years. The most variation in the distribution of data in the clusters can be described by variable importance analyses in which the greatest weight is given by BMI, BSA and the three features of interest (Fig. 2).

3.1. Features importance analysis

Using variable importance, we defined a score that allows us to assign a new sample in one of the identified clusters. The building of the formula which allows the calculation of the score and the threshold of score used to assign a sample to one cluster rather than another, were described in methods. In more detail, the threshold used for cluster assignment was calculated by answering the following question: what is the best score (the output of the formula based on feature importance) value above which a patient is classified in one cluster or another in accordance with the clustering obtained from random forest and pam? Agreement was assessed using AUC, accuracy, sensitivity, and specificity. When we evaluate a new patient, we calculate his score using the formula derived from the feature importance and based on that value we assign him to one of the two clusters by comparing it with the threshold, which given the obtained values of AUC and accuracy, is the threshold that we give the greatest probability of accurately assigning the patient to the clusters.

The weights from variable importance (Fig. 2), and thus those used in the formula, depending on the parameters used to run the random forest algorithm because using different parameters we get different results. For this reason, the validity of the formula is assessed through the metrics shown in Table 2, calculated from the confusion matrix between clusters assigned by RF and clusters assigned by the scores. We

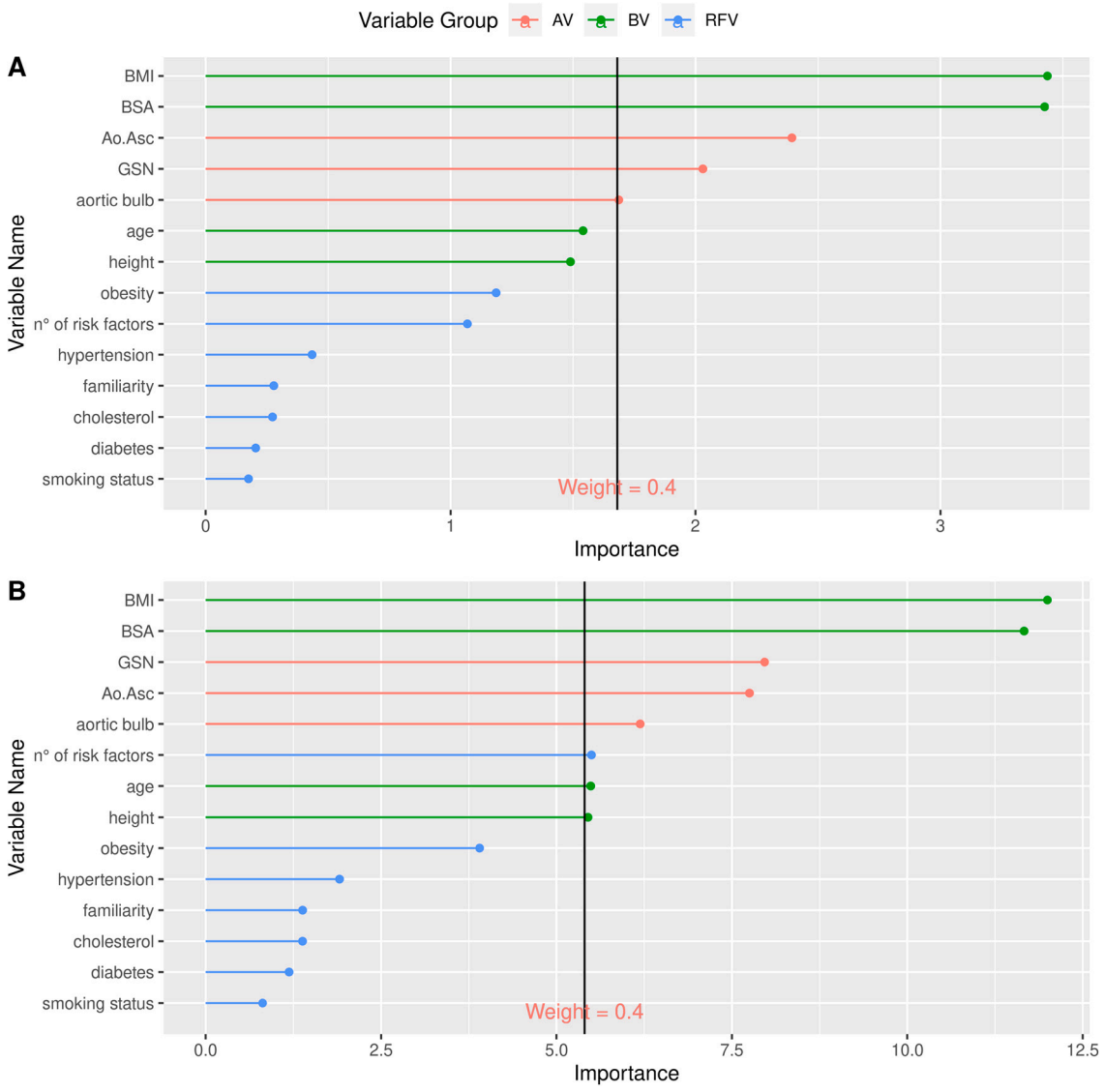

Fig. 2. Comparison of variable importance in males (A) and Females (B) clusterization (from Random Forest model). We used 0.4 as the normalized weight threshold to select features used in the score definition. There is a difference on the x-axis between males and females because they have been analyzed separately and two different models (with two different feature importance) were built. AV, Aorta Variables; BV, Basic variables; RFV, Risk Factor Variables.

Table 2
Performances of formula-based cluster prediction.

	Males	Females	Mean
Best Formula	(A*0.40)+(H*0.45)+(BMI*0,94)+(BSA*0,92) +(RF*0,40)+(B*0,43)+(G*1)+(AO*0,71)	(A*0,41)+(H*0,41)+(BMI*1)+(BSA*0,97)+ (RF*0,41)+(B*0,48)+(G*063)+(AO*0,62)	
AUC	0.9	0.91	0.905
Accuracy	0.83	0.85	0.84
Sensitivity	0.82	0.88	0.85
Specificity	0.84	0.81	0.825
Precision	0.89	0.87	0.88
Recall	0.82	0.88	0.85

choose the best formulas (Table 2) based on AUC, accuracy, sensitivity, and specificity and tuning RF on “maxnodes” parameters, while we have set “mtry” as the square root of the number of variables and “ntree” defined by the algorithm. Therefore, the resulting model consists of 500 numbers of trees, 4 variables tried at each split, and 4 terminal nodes (“maxnodes”), in the case of males. Instead, in the case of females, the resulting model consists of 500 numbers of trees, 3 variables tried at each split, and 15 terminal nodes.

3.2. *Analyses of clusters from male patients*

In male clusters, cluster B contains most of the samples with “high” BMI (>= 30), while cluster A contains most of the samples with “low” BMI (<25). In more detail, cluster B shows 45% of samples with high BMI and 42% with medium BMI (87% of samples show medium-high BMI) while cluster A shows 3% of samples with high BMI and 58% with medium BMI (97% of samples show medium-low BMI). The same re-

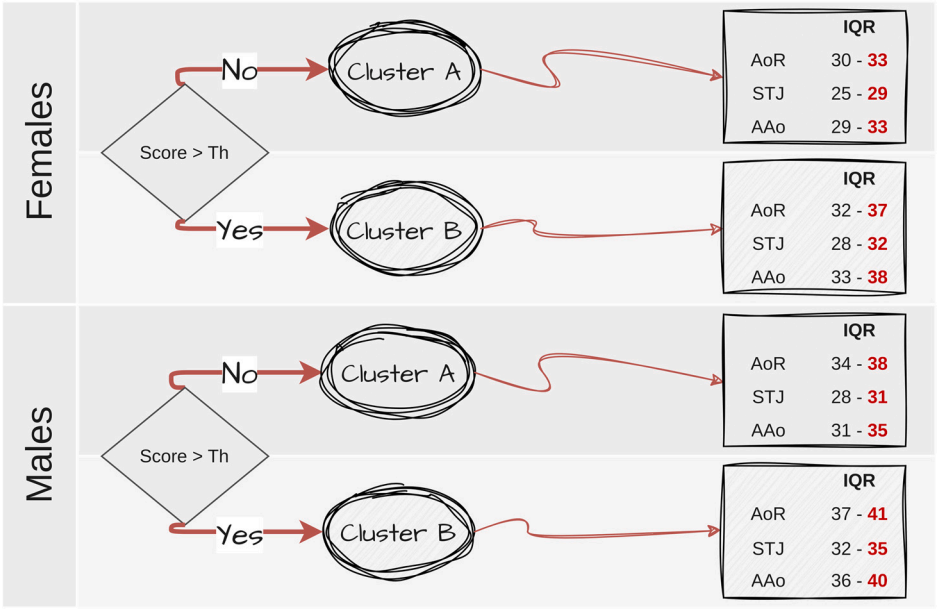

Fig. 3. Schematic representation of different ranges of aorta dimensions assigned to each cluster. A novel sample can be assigned to a cluster using their score compared to Th (threshold obtained from random forest feature importance analyses).

Table 3
Differences in the distribution of numeric variables in the two clusters in both males and female. In the red square we show interqurtile range of our variables of interest. BMI: Body Mass Index; BSA: Body Surface Area; AoR: aortic root; STJ: Sinotubular Junction: AAo: tubular ascending aorta.

Variable	Sex	Mean	Sd	Median	Min	Pctile[25]	Pctile[75]	Max	Mean	Sd	Median	Min	Pctile[25]	Pctile[75]	Max	Test
Cluster		A							B							
Age	M	59	13.2	60	19	49	68	94	62.8	12.3	65	25	56	72	90	F=13.895***
	F	64.1	12	66	24	57	73	89	66.7	11.8	67.5	24	59.8	76	90	F=4.649**
Height	M	172.4	6.9	172	145	168	176.2	203	174.9	7.3	175	155	170	180	203	F=20.418***
	F	160.5	5.5	160	140	157	165	175	162.4	7.1	163	145	157	168	180	F=9.556***
BMI	M	25.6	2.6	25.7	17.1	23.9	27.4	31.2	29.6	4	29.7	18.3	27.1	32.1	42.6	F=231.134***
	F	23.7	3	23.6	13.6	21.6	25.7	31.2	29.8	5	29.5	17.8	26.4	32.8	50.8	F=246.738***
BSA	M	1.9	0.1	1.9	1.3	1.8	2	2.4	2.1	0.2	2.1	1.6	2	2.2	2.6	F=216.736***
	F	1.6	0.1	1.7	1.2	1.6	1.7	2	1.9	0.2	1.9	1.3	1.8	2	2.4	F=255.151***
n° of risk factor	M	1.7	1.2	2	0	1	2.2	5	3	1.2	3	0	2	4	6	F=193.305***
	F	2	1.1	2	0	1	3	4	2.6	1.3	3	0	2	3.2	6	F=28.569***
AoR	M	35.9	3.2	36	28	34	38	46	39.1	3.7	39	30	37	41	47	F=145.767***
	F	31.6	2.6	32	25	30	33	40	34.5	3	35	26	32	37	41	F=113.51***
STJ	M	29.5	2.9	30	22	27.8	31	39	33.5	3.2	33	26	32	35	41	F=274.18***
	F	27.1	2.4	27	21	25	29	34	30.1	3.1	30	22	28	32	38	F=128.946***
AAo	M	33	3.5	33	23	31	35	45	37.9	4	38	28	36	40	47	F=272.077***
	F	31.4	3.2	31	23	29	33	42	35.6	3.9	35	25	33	38	45	F=141.212***

sults are also found in BSA where there is a significant difference, in this value, between the two clusters. Moreover, samples in cluster B tend to have a greater number of risk factors (68% of samples with RF >=3, compared to 25% in cluster A). It is important to note that for all mentioned variables there is a significant difference in distribution values between the two clusters while there are no significant differences in any of the risk factors taken individually.

3.3. Analyses of clusters from female patients

As well as in males, in females, cluster B contains most of the samples with “high” BMI (>= 30) while cluster A most of the samples with “low” BMI (<25). In more detail, cluster B shows 45% of samples with high BMI and 39% with medium BMI (84% of samples show medium-high BMI) while cluster A shows 2% of samples with high BMI and 31% with medium BMI (98% of samples show medium-low BMI). Also in this case, there is a significant difference in BSA values, and samples in cluster B tend to have a greater number of risk factors (54% of samples with RF >=3, compared to 32% in cluster A). Moreover, just as in males, there is a significant difference in BMI, BSA, and n° of risk factor distribution between the two clusters and no significant differences in each single risk factor.

3.4. Quartile-based ranges of normality

From the results listed above, it was evident that the clustering obtained by means of a UML method made it possible to identify 4 populations with significant differences in terms of physical characteristics and, above all, with different aorta-related size ranges (Table 3 and ST1). These results could then be exploited to assess in an individual-specific manner the cases in which a certain value (e.g., the diameter of the aorta) could be an alarm bell or even identify a pathological value. Indeed, once the individual patient has been assigned to one of the clusters and using his/her physical and clinical characteristics (thus calculating the score defined above), we could assess differently and specifically whether aorta-related values are associated with a specific clinical profile. To do this, we could use the interquartile ranges for each cluster and differentiate by gender (Fig. 3). Thus, a value of 35 in the diameter of a male’s aorta could be evaluated as a clinical profile that could be pathological or at least represent an alarm if the subject belongs to Cluster A. But, at the same time, the same value could be normal if he belongs to Cluster B, and this is because, in the two different cases, the patient presents different physical characteristics that are correlated to the under-study dimensions aorta related.

4. Discussion

In this study, based on individuals without known aortic or coronary artery disease, we applied a UML method to identify subgroups with distinct demographic and clinical (aorta dimensions and risk factors) characteristics. We demonstrate that unsupervised learning algorithms may be used to handle clinical data with heterogeneous characteristics and generate classifications with varying risks for coronary artery disease. Our approach's progressive nature lies in the absence of specifications regarding data partitioning based on type or expertise, opting instead for agnostic methodologies to handle highly heterogeneous data types. Through this method, we achieved the identification of four patient groups (2 for males and 2 for females) that exhibit distinct phenotypic and clinical characteristics. In more detail, individuals in cluster B of males had a medium-high BMI and a number of risk factors greater than 2 (>=3) while individuals in cluster A had a medium-low BMI and a number of risk factors less than 3. In the same way, for females, individuals of cluster B had a medium-high BMI and number of risk factors greater than 2 while cluster A individuals had a medium-low BMI and number of risk factors less than 3. These results agree with the feature importance analysis that identified BMI and aortic sizes as variables that contribute the most to cluster definition. Furthermore, by analyzing the distributions of values related to aortic size, we can assign different ranges of values to patients with different characteristics. In fact, while an aorta diameter value might be alarming for a patient with certain characteristics, this might not be true for a different patient. Subdividing individuals into different clusters based on their clinical characteristics allows us to have different ranges that can be interpreted differently. Collectively, these findings indicate that unsupervised clustering can potentially aid in integrating diverse patient data, enabling a more comprehensive understanding of distinct disease risk trajectories. While the full implementation of machine learning in clinical practice is still underway, our data provide evidence of its capacity to identify clinically significant subgroups. Indeed, these data-driven methodologies have the potential to facilitate the development of automated clinical scoring systems and generate meaningful clinical insights from existing healthcare data repositories. When interpreting our findings, it is essential to consider various limitations. Primarily, the generalizability of the current study may be constrained due to the inclusion of patients exclusively from one center, where study enrollment was contingent on the requirement for coronary CT angiography; this entails at least a significant clinical suspicion, even though in this case the population is medium-low pre-test probability of coronary obstructive disease. Additionally, it is important to acknowledge that the observed phenotypic differences within our cohort were likely influenced by population structure, encompassing factors such as social and genetic variation associated with geographic distribution. As a result, it is crucial to evaluate the effectiveness of our clustering algorithm in more diverse populations and datasets with varying structures to ascertain its reliability and generalizability. Finally, note that some of these detected differences could be related to differences in the number of samples under 50 years of age between the two clusters (Males: 28% in cluster A compared to 14% in cluster B; Females: 15% in cluster A compared to 8% in cluster B).

5. Conclusion

A UML approach was applied to healthy patients suspected of coronary artery disease. Four clusters were identified from this specific population. Identification of subgroups and their demographic and clinical characteristics may be useful to anticipate treatment strategies and probable outcomes for groups of these patients. UML may allow multidimensional data to be organized and to make sense of the data in terms of a range of normality.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Ministry of Health under contract "Ricerca Corrente RRC-2022-23680785". The authors are grateful and acknowledge the funding source.

Appendix A. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csbj.2023.11.021.

References

[1] Muraru D, Maffessanti F, Kocabay G. Ascending aorta diameters measured by echocardiography using both leading edge-to-leading edge and inner edge-to-inner edge conventions in healthy volunteers. Eur Heart J Cardiovasc Imaging 2014;15:415–22.
[2] Parachuri K, Salhab V. Aortic size distribution in the general population: explaining the size paradox in aortic dissection. Cardiology 2015;4:265–72.
[3] Ceja-Rodriguez A, Realyvasquez M. Differences in aortic diameter measurements with intravascular ultrasound and computed tomography after blunt traumatic aortic injury. Ann Vasc Surg 2018;50:148–53.
[4] Lembrança M, Teivelis L. Thoracic aortic size in Brazilian smokers: measures using low-dose chest computed tomography anatomical and epidemiological assessment. Ann Vasc Surg 2021;20:76:e2315.
[5] Wang M, Desai TKM. Thoracic aortic aneurysm: optimal surveillance and treatment. Clevel Clin J Med 2020;9:557–68.
[6] Takaoka HO, Kitahara H. Utility of computed tomography in cases of aortic valve stenosis before and after transcatheter aortic valve implantation. Cardiovasc Interv Ther 2020;1:72–84.
[7] Runza G, Fattouch K, Cademartiri F, La Fata A, Damiani L, La Grutta L, et al. ECG-gated multidetector computed tomography for the assessment of the postoperative ascending aorta. Radiol Med 2009;114:705–17.
[8] Li Y, Fan Z, Xu L, Yang L, Xin H, Zhang N, et al. Prospective ECG-gated 320-row CT angiography of the whole aorta and coronary arteries. Eur Radiol 2012;22:2432–40.
[9] Roos JE, Willmann JK, Weishaupt D, Lachat M, Marincek B, Hilfiker PR. Thoracic aorta: motion artifact reduction with retrospective and prospective electrocardiography-assisted multi-detector row CT. Radiology 2002;222:271–7.
[10] Erbel R, Aboyans V, Boileau C, Bossone E, Bartolomeo RD, Eggebrecht H, et al. 2014 ESC guidelines on the diagnosis and treatment of aortic diseases: document covering acute and chronic aortic diseases of the thoracic and abdominal aorta of the adult. The Task Force for the Diagnosis and Treatment of Aortic Diseases of the European Society of Cardiology (ESC). Eur Heart J 2014;35:2873–926.
[11] Vriz O, Driussi C, Bettio M, Ferrara F, D'Andrea A, Bossone E. Finding community structure in very large networks. Am J Cardiol 2013;112:1224–9.
[12] Burman ED, Keegan J, Kilner PJ. Aortic root measurement by cardiovascular magnetic resonance: specification of planes and lines of measurement and corresponding normal values. Circ Cardiovasc imaging 2008;1:104–13.
[13] Devereux RB, de Simone G, Arnett DK, Best LG, Boerwinkle E, Howard BV, et al. Normal limits in relation to age, body size and gender of two-dimensional echocardiographic aortic root dimensions in persons ≥15 years of age. Am J Cardiol 2012;110:1189–94.
[14] Hager A, Kaemmerer H, Rapp-Bernhardt U, Blücher S, Rapp K, Bernhardt TM, et al. Diameters of the thoracic aorta throughout life as measured with helical computed tomography. J Thorac Cardiovasc Surg 2002;123:1060–6.
[15] Lin FY, Devereux RB, Roman MJ, Meng J, Jow VM, Jacobs A, et al. Assessment of the thoracic aorta by multidetector computed tomography: age- and sex-specific reference values in adults without evident cardiovascular disease. J Cardiovasc Comput Tomogr 2008;2:298–308.
[16] Roman MJ, Devereux RB, Kramer-Fox R, O'Loughlin J. Two-dimensional echocardiographic aortic root dimensions in normal children and adults. Am J Cardiol 1986;64:507–12.
[17] Vasan RS, Larson MG, Benjamin EJ, Levy D. Echocardiographic reference values for aortic root size: the Framingham heart study. J Am Soc Echocardiogr 1995;8:793–800.
[18] Rogers IS, Massaro JM, Truong QA, Mahabadi AA, Kriegel MF, Fox CS, et al. Distribution, determinants, and normal reference values of thoracic and abdominal aortic diameters by computed tomography (from the Framingham heart study). Am J Cardiol 2013;111:1510–6.

[19] McComb BL, Munden RF, Duan F, Jain AA, Tuite C, Chiles C. Normative reference values of thoracic aortic diameter in American College of Radiology Imaging Network (ACRIN 6654) arm of National Lung Screening Trial. Clin Imaging 2016;40:936–43.
[20] Mao SS, Ahmadi N, Shah B, Beckmann D, Chen A, Ngo L, Flores FR, Gao Y, Budoff MJ. Normal thoracic aorta diameter on cardiac computed tomography in healthy asymptomatic adults: impact of age and gender. Acad Radiol 2008;15:827–34.
[21] Nevsky G, Jacobs JE, Lim RP, Donnino R, Babb JS, Srichai MB. Sex-specific normalized reference values of heart and great vessel dimensions in cardiac CT angiography. Am J Roentgenol 2011;196:788–94.
[22] Forte E, Punzo B, Salvatore M, Maffei E, Nistri S, Cavaliere C, et al. Low correlation between biometric parameters, cardiovascular risk factors and aortic dimensions by computed tomography coronary angiography. Medicine 2020;99:e21891.
[23] Breiman L. Random forests. Machine learning. Mach Learn 2001;45:5–32.
[24] Reynolds AP, Richards G, de la Iglesia B, Rayward-Smith VJ. Clustering rules: a comparison of partitioning and hierarchical clustering algorithms. J Math Model Algorithms 2006;5:475–504.