ECG-Based Unsupervised Clustering in Coronary Artery Disease Associates with Ventricular Arrhythmia

Josseline Madrid[1], Patricia B Munroe[2], Stefan van Duijvenboden[3], Julia Ramírez[1], Ana Mincholé[1]

[1]Aragon Institute of Engineering Research (I3A), Zaragoza, Spain
[2]Queen Mary University of London, London, United Kingdom
[3]University of Oxford, Oxford, United Kingdom

Abstract

Aims: *Coronary Artery Disease (CAD) is one of the main causes of life-threatening ventricular arrhythmias (LTVAs) leading to sudden cardiac death. CAD slows ventricular conduction across individuals, manifesting as heterogeneous QRS morphologies. This study aimed to identify distinct clusters of CAD individuals based on QRS morphology using unsupervised learning, and investigate their association with LTVA risk.*

Methods: *An average heartbeat was derived from 10-second electrocardiograms (ECGs, lead I) from 1,458 individuals diagnosed with CAD in the UK Biobank study. An unsupervised clustering algorithm based on 3-nearest neighbours was used to classify each individual, then we evaluated the association of each cluster with LTVA risk.*

Results: *There were a total of 65 LTVA events in the population. The algorithm distinguished 3 distinct clusters of QRS-related morphological features, which significantly differed in terms of LTVA events rate. Cluster 2, characterized by the lowest QRS amplitudes and widest QRS complexes, was strongly associated with LTVA risk.*

Conclusions*: Our analysis has identified CAD individuals at risk of LTVA using the QRS morphology. The identified cluster could be used to tailor care and provide refined risk assessment in CAD individuals to apply specific prevention measures.*

1. Introduction

Sudden cardiac death (SCD) is a leading cause of CVD mortality, becoming a public health problem accounting for an estimated 15% – 20% of all deaths [1]. Life-threatening ventricular arrhythmias (LTVAs) can be a precursor of SCD, in 80% of cases SCD occurs in patients with underlying coronary artery disease (CAD) in people over 50 years old [1,2].

The surface electrocardiogram (ECG) offers a rapid assessment of the underlying cardiac electrophysiology in a low-cost and non-invasive way. In particular, the QRS complex morphologies on the ECG reflect the ventricular conduction velocity that is reduced in the presence of CAD, and is associated with higher LTVA risk [3].

Machine learning techniques have been widely used in the literature as affordable approaches to diagnose CAD [4]. Recent studies based on unsupervised clustering algorithms have demonstrated to be able to interpret heterogeneous clinical data to discover clinically important CAD subgroups with distinct clinical trajectories (i.e., myocardial infarction, stroke, and mortality)[5] and identify risk phenotypes of CAD in patients undergoing single-photon emission computed tomography (SPECT) myocardial perfusion imaging (MPI)[6]. Despite these advances in risk stratification in CAD, significant challenges remain.

Current non-invasive SCD risk stratification strategies are only based on the left ventricular ejection fraction (LVEF) and the presence and severity of heart failure symptoms to predict SCD risk in patients with CAD [2]. Risk stratification studies specific of LTVA in CAD based on ECG morphology have not yet been performed. The risk assessment of LTVA in CAD based on the ECG morphology could be easily scaled for population-level screening. The aim of this study was to identify distinct groups of CAD individuals based on QRS morphology through the application of unsupervised learning techniques, and to investigate their association with LTVA risk.

2. Methods

2.1. UK Biobank Study cohort

The UK Biobank study is a large-scale biomedical cohort that contains up-to-date health information from half a million participants from the United Kingdom [7]. Our study population consisted of 1,458 individuals from the Imaging study diagnosed with CAD in the UK Biobank

ISSN: 2325-887X DOI: 10.22489/CinC.2023.169

study at the time of the ECG acquisition. CAD was defined according to the WHO International Classification of Diseases (ICD) as ICD-9 410 to 412, or ICD-10 I21 to I24 and I25.2 [8]. LTVA events were defined by the ICD-10 codes as I47.2, I49.0, I46.0, I46.1, I46.9, I47.0 or Classification of Interventions and Procedures codes (OPCS) K576, K641, X503 or X504

LTVA risk was defined as LTVA mortality or admission to hospital with a LTVA diagnosis 6-months before or after the CAD diagnosis. The available information included collections of 10-second ECGS (lead I) recorded at rest and health electronic records for each subject considered in the study.

2.2. Signal Preprocessing and QRS-waves characterization

Preprocessing of the ECG signals involved baseline wander removal through cubic splines interpolation, low pass filtering at 40 Hz to remove electric and muscle noise, and removal of ectopic beats. An average heartbeat was derived from the filtered ECG signal. Also, average heartbeats with high signal-to-noise ratio were dismissed. A single-lead wavelet-based delineator[9] was used to locate QRS-waves onset, peak and end timings.

After preprocessing, the characterization of ECG waveforms was performed by extracting a vector of features. QRS morphology was mathematically characterized by a combination of Hermite functions[10]. We considered four Hermite functions to recover most of the QRS energy due to high QRS heterogeneity among each individual. This was confirmed by visual inspection of the reconstruction. The reconstruction error and the width of the Hermite functions were included as parameters in the model. Also, standard QRS biomarkers were considered, such as QRS amplitude, up and down slopes [11] and duration. Initially, ten QRS-related morphological features were considered for this model, as represented in Figure 1.

2.3. Identification of Clusters using QRS Biomarkers

Prior to performing the clustering of ECG heartbeats, feature selection techniques were applied [12]. This step facilitates the learning task and reduces problems of multicollinearity. Multicollinearity undermines the statistical significance of an independent variable [13]. In this study, a filter type feature selection algorithm based on the correlation between each pair of features was implemented. The correlation threshold was set to be larger than 0.8.

Then, a k-means clustering algorithm based on 3-nearest neighbors was used to classify each individual into 3 distinct clusters. The distance between neighbors was evaluated using the Euclidean distance. The clustering analysis was performed blindly to clinical data.

2.4 Statistical Analysis

Statistical analysis was performed using Matlab (version R2022b). Statistical nonparametric tests (chi-square test) were performed to evaluate the association of each of the clusters with LTVA risk. The Kruskal Wallis statistical test was used to compare differences in association with LTVA risk across all clusters.

The Wilcoxon rank sum test was used to compare the distance within each cluster's centroid for subjects who had a LTVA event versus those who hadn't. Statistical significance was assumed when $P < 0.05$.

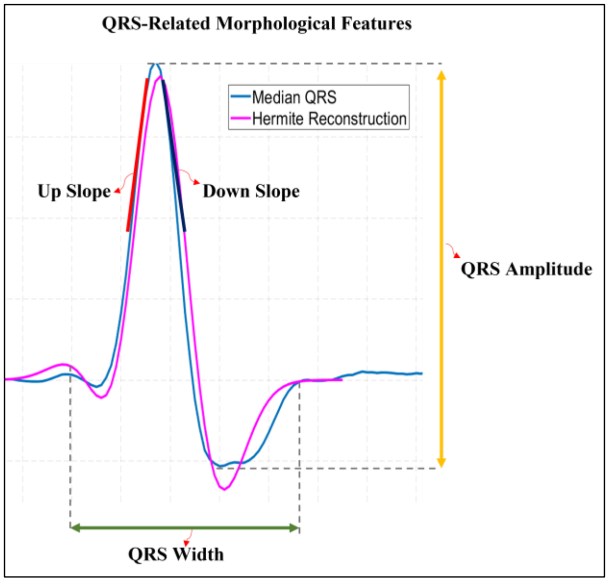

Figure 1. Representation of the QRS-related features extracted to perform the unsupervised clustering model.

3. Results

From the 1,458 CAD individuals included in this study (median age of 70 years [IQR 9] and 84% male), there were a total of 65 LTVA events (4.46%) in the population. There were no demographic differences between individuals who suffered an LTVA event and those who did not.

The final model included eight QRS-related morphological features, i.e., amplitude, width, Hermite's coefficients, Hermite's reconstruction error and Hermite's function width. Upward and downward slopes were removed from the model due to a high correlation with QRS amplitude and width.

The unsupervised clustering algorithm identified 3 distinct clusters of QRS-related morphological features in CAD, which significantly differed in terms of LTVA

Table 1. Median and Interquartile Range (IQR) results for the features in each cluster.

Features	All		Cluster 1		Cluster 2		Cluster 3		p-value
	Median	*[IQR]*	*Median*	*[IQR]*	*Median*	*[IQR]*	*Median*	*[IQR]*	*p-value*
Amplitude	824.02	[386.48]	1291.74	[235.97]	592.24	[177.19]	903.53	[180.68]	< 0.001
Width	84.00	[28]	84.00	[26]	86.00	[32]	82.00	[24]	< 0.001
Hermite Width	12.81	[2.40]	13.12	[2.36]	12.73	[2.71]	12.73	[2.07]	0.06
Hermite Error	0.02	[0.01]	0.02	[0.01]	0.03	[0.02]	0.02	[0.01]	< 0.001
Hermite Coef. 1	2.58	[0.63]	2.62	[0.57]	2.56	[0.77]	2.56	[0.53]	0.24
Hermite Coef. 2	0.35	[0.32]	0.39	[0.26]	0.31	[0.34]	0.36	[0.30]	< 0.001
Hermite Coef. 3	-0.57	[0.63]	-0.57	[0.53]	-0.54	[0.81]	-0.59	[0.55]	< 0.001
Hermite Coef. 4	-0.09	[0.41]	-0.04	[0.35]	-0.12	[0.46]	-0.08	[0.38]	0.42

events rate as shown in Figure 2. Cluster 2 showed the highest rate of LTVA events, compared to the other two (6.39%, P = 0.004, Figure 2). Cluster 3 exhibited the lowest rate of LTVA events (3.22%, P = 0.04). Kruskal Wallis statistical test demonstrated significant differences in association with LTVA risk across all clusters (P = 0.02).

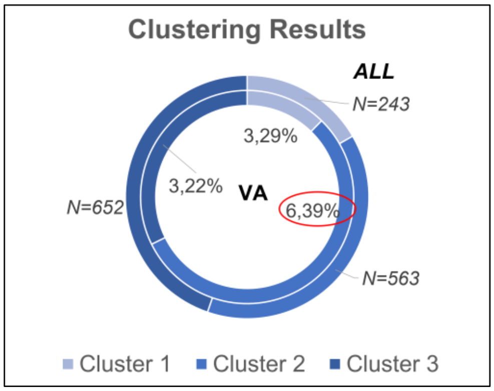

Figure 2. The unsupervised clustering algorithm identified three distinct clusters based on QRS-related morphological features.

A median heart beat was calculated from the individuals in each cluster. Differences in amplitude and width of QRS complexes are observed in the representative median beat for each cluster in Figure 3. Cluster 2 (which exhibited the highest rate of LTVA events) was mainly characterized by lower QRS amplitude, and a wider QRS than clusters 1 and 3. Cluster 3 exhibited narrower QRS complexes as shown in Table 1.

QRS amplitude showed the most significant differences among the clusters ($P < 0.005$), being the lowest in Cluster 2 (median 592.24 μV) compared to Cluster 1 and Cluster 3 (median 1291.74 μV and 903.53 μV, respectively). As well as significant differences in the reconstruction error of the Hermite functions and the duration of the QRS complex. Also, differences in morphological variations were described by Hermite's function coefficients 2 and 3 which are related to higher variability in Q and S waves. Figure 4 offers a graphical representation of these differences among clusters.

The distribution of sex and age revealed similar ratios in the three clusters, suggesting that the differences among clusters were not determined by these main cardiovascular risk factors.

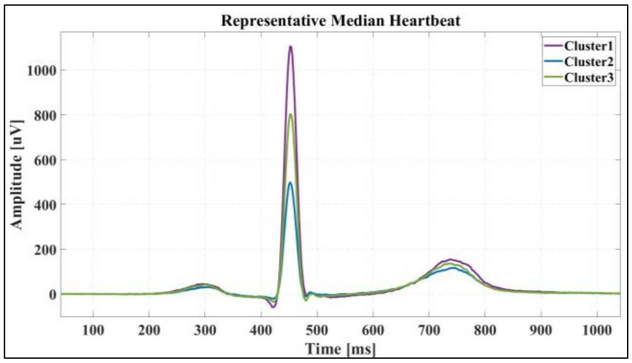

Figure 3. Median beat representative of each cluster obtained by the 3-nearest neighbors clustering algorithm

4. Discussion and Conclusions

Our analysis has identified in an unsupervised manner a cluster of individuals (cluster 2) with CAD at risk of LTVA using the QRS morphology. This cluster exhibited the lowest QRS amplitudes and widest QRS complexes which are associated with slowed ventricular conduction and high risk of SCD [1,14]. In accordance to the previous findings, this cluster had the highest rate of LTVA events demonstrating a strong association with LTVA risk.

Unsupervised learning techniques are able to identify hidden ECG morphological patterns to provide a refined risk assessment. The importance of unsupervised learning algorithms relies on the ability to cluster unlabeled data according to associations within the data. Therefore, unsupervised learning algorithms have become a useful tool to explore disease associations in clinical data, where the outcome is unknown.

Identification of CAD individuals at risk of LTVA through unsupervised techniques offers an early and reliable measure of SCD risk allowing physicians to apply specific prevention measures among groups of individuals. The ECG-based unsupervised clustering study is a useful method to infer LTVA risk. Given that the ECG is a low-cost, widely available non-invasive tool, this method could be scaled for non-invasive population-level screening.

The QRS and T-wave morphologies on the ECG reflect the ventricular conduction velocity and dispersion of repolarization, respectively, keeping key information for early screening of SCD in a non-invasive manner. Further studies will investigate the contribution of additional LTVA risk factors in CAD.

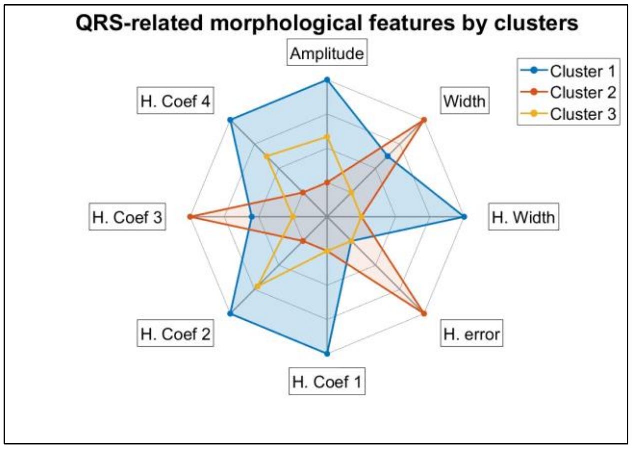

Figure 4. Comparison of the QRS-related morphological features in each cluster.

Acknowledgments

This work was supported by projects PID 2021-128972OA-I00 and LMP141_21 funded by the Spanish Ministry of Science and Innovation (MCIN) and by Gobierno de Aragón, and by fellowship RYC2021-031413-I from MCIN.

References

1. Zipes DP, Wellens HJJ. Sudden Cardiac Death. Circulation [Internet]. 1998;98:2334–51. Available from: https://doi.org/10.1161/01.CIR.98.21.2334

2. Hayashi M, Shimizu W, Albert CM. The Spectrum of Epidemiology Underlying Sudden Cardiac Death. Circ Res [Internet]. 2015;116:1887–906. Available from: https://doi.org/10.1161/CIRCRESAHA.116.304521

3. Sörnmo L, Laguna P. Bioelectrical Signal Processing in Cardiac and Neurological Applications. Bioelectrical Signal Processing in Cardiac and Neurological Applications [Internet]. 2005 [cited 2023 May 23]; Available from: http://www.sciencedirect.com:5070/book/9780124375529/bioelectrical-signal-processing-in-cardiac-and-neurological-applications

4. Alizadehsani R, Abdar M, Roshanzamir M, Khosravi A, Kebria PM, Khozeimeh F, et al. Machine Learning-Based Coronary Artery Disease Diagnosis: A comprehensive review. Comput Biol Med [Internet]. 2019;111:103346. Available from: https://www.sciencedirect.com/science/article/pii/S001048251930215X

5. Flores AM, Schuler A, Eberhard AV, Olin JW, Cooke JP, Leeper NJ, et al. Unsupervised Learning for Automated Detection of Coronary Artery Disease Subgroups. J Am Heart Assoc [Internet]. 2021;10:e021976. Available from: https://doi.org/10.1161/JAHA.121.021976

6. Williams MC, Bednarski BP, Pieszko K, Miller RJH, Kwiecinski J, Shanbhag A, et al. Unsupervised Learning to Characterize Patients with Known Coronary Artery Disease Undergoing Myocardial Perfusion Imaging. Eur J Nucl Med Mol Imaging [Internet]. 2023;50:2656–68. Available from: https://doi.org/10.1007/s00259-023-06218-z

7. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. PLoS Med [Internet]. 2015 [cited 2023 May 22];12:e1001779. Available from: https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1001779

8. World Health Organization. International Statistical Classification of Diseases and Related Health Problems. World Health Organization; 2011.

9. Martínez JP, Almeida R, Olmos S, Rocha AP, Laguna P. A Wavelet-Based ECG Delineator Evaluation on Standard Databases. IEEE Trans Biomed Eng. 2004;51:570–81.

10. SÖrnmo L, BÖRJESSON PO, Nygårds ME, Pahlm O. A Method for Evaluation of QRS Shape Features Using a Mathematical Model for the ECG. IEEE Trans Biomed Eng. 1981;BME-28:713–7.

11. Pueyo E, Sornmo L, Laguna P. QRS Slopes for Detection and Characterization of Myocardial Ischemia. IEEE Trans Biomed Eng. 2008;55:468–77.

12. Nezamabadi K, Sardaripour N, Haghi B, Forouzanfar M. Unsupervised ECG Analysis: A Review. IEEE Rev Biomed Eng. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 208–24.

13. Allen MP. The problem of multicollinearity. In: Allen MP, editor. Understanding Regression Analysis [Internet]. Boston, MA: Springer US; 1997. p. 176–80. Available from: https://doi.org/10.1007/978-0-585-25657-3_37

14. Das M, Suszko AM, Nayyar S, Viswanathan K, Spears DA, Tomlinson G, et al. Automated Quantification of Low-Amplitude Abnormal QRS Peaks From High-Resolution ECG Recordings Predicts Arrhythmic Events in Patients With Cardiomyopathy. Circ Arrhythm Electrophysiol [Internet]. 2017;10:e004874. Available from: https://doi.org/10.1161/CIRCEP.116.004874

Address for correspondence:
Josseline Madrid, jmadrid@unizar.es
Instituto de Investigación en Ingeniería de Aragón (I3A), Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.