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]Biomedical Signal Interpretation and Computational Simulation (BSICoS)
Instituto de Investigación en Ingeniería de Aragón (I3A)
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.
Tel. +34-976762707, e-mail: jmadrid@unizar.es
[2]Queen Mary University of London, London, United Kingdom.
[3]University of Oxford, Oxford, United Kingdom.

Abstract

Coronary Artery Disease (CAD) is a leading cause of life-threatening ventricular arrhythmias (LTVAs). This study aimed to identify distinct clusters of CAD individuals based on QRS morphology using a 3-nearest neighbors clustering algorithm. Cluster 1, characterized by the lowest QRS amplitudes and widest QRS complexes, was strongly associated with LTVA risk.

Introduction

Life-threatening ventricular arrhythmias (LTVA) are an important cause of morbidity in coronary artery disease (CAD). 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 which is associated with higher LTVA risk.[6] The presence of CAD slows ventricular conduction heterogeneously across individuals, manifesting as different QRS morphologies. The aim of this study was to identify distinct groups of CAD individuals based on QRS morphology through the application of unsupervised learning techniques.

Methods

UK Biobank Study cohort

The UK Biobank study is a large-scale biomedical cohort which contains up to date health information from half a million participants from United Kingdom.[7] Our study population consisted of 1,458 individuals diagnosed with CAD in the UK Biobank study at the time of enrollment. 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 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-seconds I-lead ECGs recorded at rest, health electronic records and follow-up data for each subject considered in the study.

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. Average heartbeats with high signal-to-noise ratio were dismissed. A single-lead wavelet-based delineator[2] was used to locate QRS-waves delineation marks.

After preprocessing, the characterization of ECG waveforms was performed extracting a vector of features. QRS morphology was mathematically characterized by a combination of Hermite functions[5]. We considered six 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. Also, standard QRS biomarkers were considered in this study, such as QRS amplitude, slopes,[4] and duration.

Identification of Clusters using QRS Biomarkers

Prior to performing the clustering of ECG heartbeats, feature extraction and dimensionality reduction techniques were applied.[3] This step facilitates the learning task and reduces problems of multicollinearity. Multicollinearity undermines the

statistical significance of an independent variable. [1]In this study, a filter type feature selection algorithm based on the correlation between each feature was implemented. The correlation threshold was set to be larger than 0.8. Then, a k-means clustering algorithm based on 3-nearest neighbours 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.

Statistical Analysis

Statistical nonparametric tests (chi-square test) were performed to evaluate the association of each of the clusters versus the others 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 evaluate the centroid distances within each cluster between subjects who had a LTVA event versus those who hadn't. Statistical significance was assumed when $P < 0.05$.

Results

There were a total of 65 LTVA events in the population. The unsupervised algorithm identified 3 distinct clusters of QRS-related morphological features in CAD, which significantly differed in terms of LTVA events rate (Table 1). Cluster 1 showed the highest rate of LTVA events (6.38%). It was mainly characterized by lower QRS amplitude, flatter down slopes, and a wider QRS than clusters 2 and 3. These characteristics are observed in the representative median beat for each cluster in Figure 1. QRS amplitude showed the most significant differences among the clusters, being the lowest in Cluster 1 (median 592.42) compared to Cluster 2 and Cluster 3 (904.82 and 1292.35, respectively). Also, significant differences in morphological variations were described by Hermite basis 2, 3 and 5.

Table 1. Results of K-means Clustering

Clusters (N=1458)	LTVA Ratio	*P*
Cluster 1 (N=564)	6.38 %	0.005*
Cluster 2 (N=652)	3.22 %	0.049
Cluster 3 (N= 242)	3.31 %	0.342

Conclusion

Our analysis has identified in an unsupervised manner three distinct clusters of CAD individuals using the QRS morphology. The cluster with the lowest QRS amplitudes and widest QRS complexes was strongly associated with LTVA risk. Further studies will investigate the contribution of additional LTVA risk factors in CAD.

References

[1] ALLEN, M.P. The problem of multicollinearity. In ALLEN, M.P.Ed. *Understanding Regression Analysis* [online]. Boston, MA: Springer US, 1997. s. 176–180. ISBN 978-0-585-25657-3.

[2] MARTÍNEZ, J.P. et al. A Wavelet-Based ECG Delineator Evaluation on Standard Databases. In *IEEE Transactions on Biomedical Engineering*. 2004. Vol. 51, no. 4, s. 570–581.

[3] NEZAMABADI, K. et al. [s.l.]: Institute of Electrical and Electronics Engineers Inc., 2023.

[4] PUEYO, E. et al. QRS Slopes for Detection and Characterization of Myocardial Ischemia. In *IEEE transactions on bio-medical engineering*. 2008. Vol. 55, s. 468–477.

[5] SÖRNMO, L. et al. A Method for Evaluation of QRS Shape Features Using a Mathematical Model for the ECG. In *IEEE Transactions on Biomedical Engineering*. 1981. Vol. BME-28, no. 10, s. 713–717. [cit. 2023-05-23].

[6] SÖRNMO, L. - LAGUNA, P. Bioelectrical Signal Processing in Cardiac and Neurological Applications. In *Bioelectrical Signal Processing in Cardiac and Neurological Applications* [online]. 2005. [cit. 2023-05-23].

[7] SUDLOW, C. et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. In *PLOS Medicine* [online]. 2015. Vol. 12, no. 3, s. e1001779. [cit. 2023-05-22].

[8] WORLD HEALTH ORGANIZATION. *International statistical classification of diseases and related health problems*. [s.l.]: World Health Organization, 2011. ISBN 9789241549165.

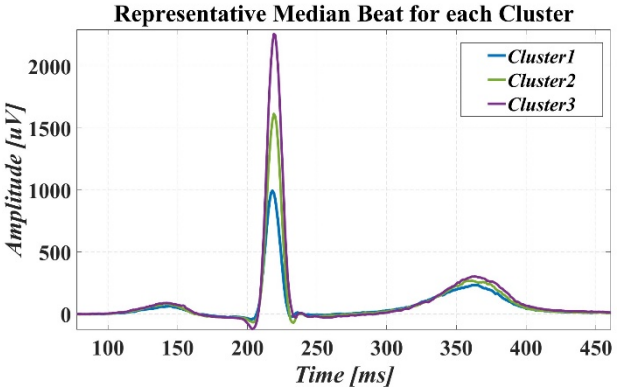

Figure 1. Median beat representative for each cluster obtained by the 3-nearest neighbors clustering algorithm