Unsupervised ECG Clustering Reveals Distinct Associations with Cardiac Magnetic Resonance Features

Josseline Madrid[1], Mihir M. Sanghvi[2], William J Young[2], Stefan van Duijvenboden[3], Patricia B Munroe[2], Julia Ramírez[1, 2, 4], Ana Mincholé[1, 4]

[1]IIS, I3A, University of Zaragoza, Zaragoza, Spain
[2] William Harvey Research Institute, Barts and the London Faculty of Medicine and Dentistry, Queen Mary University of London, London, United Kingdom
[3]University of Oxford, Oxford, United Kingdom
[4]Centro de Investigación Biomédica en Red – Biomateriales, Bioingeniería y Nanomedicina

Abstract

Exploring the association between the electrocardiogram (ECG) and cardiac magnetic resonance (CMR)-derived features may enhance our understanding of cardiovascular physiology. We aimed to identify clusters of individuals without diagnosed cardiovascular disease (CVD) based on their ECG phenotypes in an unsupervised manner and evaluate their cardiac anatomical differences through CMR.

Spatial and single-lead ECG markers were calculated from 10-second 12-lead ECGs from 51,974 UK Biobank individuals without diagnosed CVD. A k-means clustering model grouped individual ECG phenotypes into k clusters. Statistical analyses were conducted to assess ECG, demographic and CMR differences across clusters.

Three distinct ECG-based clusters were identified (N1=19,470, N2=22,256, N3=8,997), with significant differences in ECG morphology and CMR-derived anatomical features. The most discriminative ECG features involved ventricular repolarization in precordial leads (i.e., T- and ST-segment amplitude). Cluster-specific electro-anatomical alignment was stronger in Cluster 3.

Our findings show that ECG phenotyping through unsupervised clustering can reveal anatomical cardiac differences. Future work will evaluate the association with incident risk of each of these clusters.

1. Introduction

Cardiac magnetic resonance (CMR) is considered the gold standard for evaluating cardiac morphology and function[1], providing detailed insights into ventricular volumes, myocardial mass, and atrial remodeling. However, it's complexity and cost limit its use for large-scale screening. In contrast, the electrocardiogram (ECG) is inexpensive, widely available and can reflect structural and functional abnormalities[2].

Previous studies have explored associations between individual ECG features and structural measurements derived from CMR, such as QRS duration prolongation [3], Sokolow-Lyon voltage criterion[3] and QRS complex fragmentation[4]. For instance, Q waves have been linked to infarct size and location, while ST-segment elevation correlates with transmural ischemia and myocardial area at risk [5,6]. These findings suggest that ECG patterns provide insight into regional myocardial remodeling and disease.

ECG-based unsupervised clustering has mainly been used to reveal novel phenotypic subgroups in specific disease populations [7], including patients with coronary artery disease [8,9] and hypertrophic cardiomyopathy [10], some of which show distinct CMR profiles and increased cardiovascular risk [7–11]. However, such approaches have not yet been explored in the general population. This could aid in detecting subclinical cardiac variation and support early, low-cost screening of asymptomatic individuals.

We hypothesized that there are subgroups of individuals without diagnosed cardiovascular diseases (CVD) who have distinct ECG-based phenotypes, and each are exhibiting significant differences in cardiac anatomy. In this study, we applied unsupervised clustering to identify ECG-based phenotypic clusters and evaluate their anatomical characteristics through CMR features.

2. Methods

2.1. UK Biobank Cohort

The UK Biobank (UKB) is a large-scale cohort of individuals from the United Kingdom [11]. Our study population included 51,974 individuals without diagnosed CVD, who participated in the UKB CMR Imaging study and had a 10-second 12-lead ECG recorded at rest. This work was conducted under application number 2964.

2.2. ECG biomarkers

ECG signal pre-processing and the computation of median heartbeats per lead were performed following the methodology described in [9]. A set of biomarkers was extracted from each median heartbeat in the eight independent leads (I, II, V1–V6), including both standard and advanced ECG features [9]. In addition, P-wave morphology was characterized using Hermite functions[9], applying two different basis functions for P- and T-wave reconstruction and four for the QRS complex. Features such as reconstruction error and waveform width were additionally included as markers. A total of 29 ECG-related markers were obtained from each median heartbeat per lead. Beyond single-lead features, we derived 8 spatial features, including QT dispersion, QRS-T angle [12] and P-wave loop characteristics [13]. Additionally, the RR-interval was included, making a total of 241 biomarkers. Signal processing analyses were performed using MATLAB (version R2022b).

2.3. Identification of Clusters

After removing ECG features that had a strong Spearman correlation (r>0.8) with multiple other features and those with missing data (>10%). Then, missing values were imputed using k-nearest neighbors'[14], and to account for potential confounding, the remaining ECG features were adjusted for age, sex, and body mass index (BMI) using multivariable linear regression models. The resulting residuals were standardized and used in the subsequent analyses.

The optimal number of clusters was determined using a grid search approach, evaluating the elbow method for k-means clustering algorithm across 2 to 10 clusters. The optimal number of clusters 'k' was determined by selecting the value that minimized the sum of squared errors distances. Finally, a k-means clustering algorithm was employed to categorize individuals into k clusters based on their ECG features. Clustering analysis were performed using MATLAB (version R2022b).

2.4. Statistical Analyses

We compared ECG, cardiovascular risk factors (age, sex, smoking status, alcohol consumption, BMI, systolic and diastolic blood pressure [SBP, DBP]) and ventricular CMR[1] features across each cluster. To compare continuous variables, we applied the Kruskal Wallis test, reported as median [interquartile range (IQR)]. Categorical features were analyzed using the Chi-square test, described as numbers [percentages].

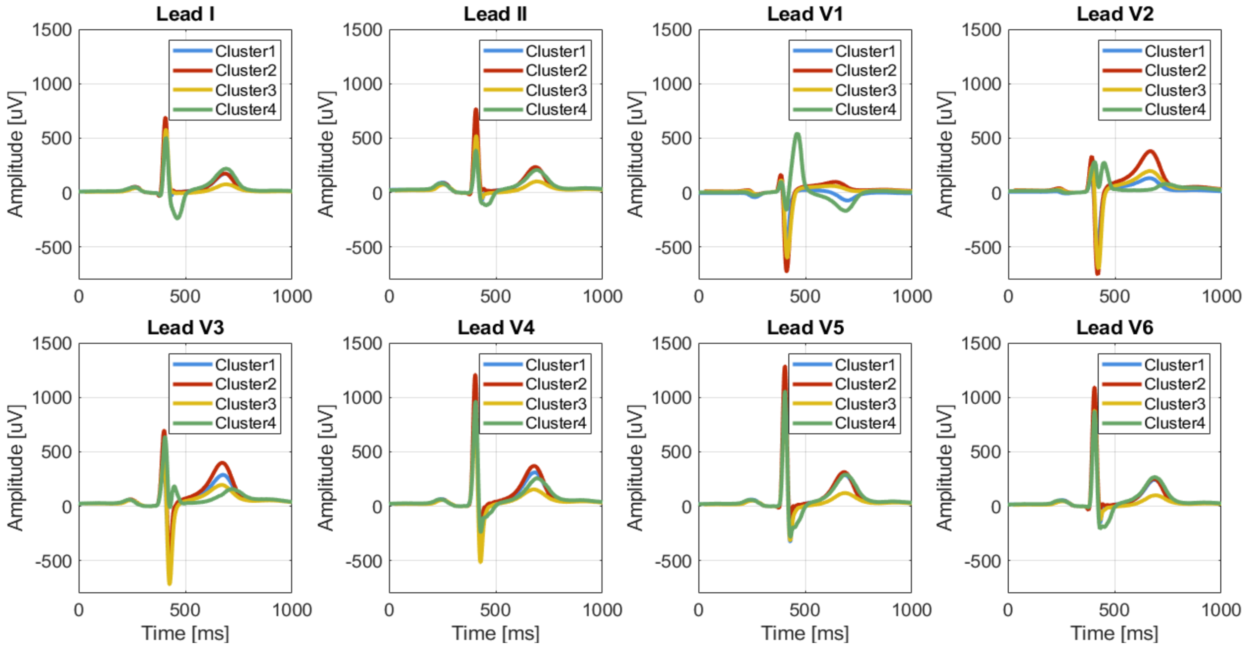
The contribution of ECG features to the clustering process was assessed using a random forest model with 500 trees. The most representative ECG features identified by the random forest model were further investigated to assess their relationship with cardiac anatomical parameters derived from CMR within each cluster. To do so, multivariable linear regression models were fitted separately for each cluster, allowing exploration of subgroup-specific associations. For each model, we report coefficient of determination (R^2), Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE). The Chow test was used to determine whether the relationships between ECG and CMR features differed significantly across clusters by testing for structural breaks in the regression models. Specifically, we assessed whether the regression coefficients for each cluster were statistically different, using the first cluster as reference. P-values were adjusted using Bonferroni correction.

Table 1. Cardiovascular risk factors and CMR characteristics in the study population and in each cluster.

Characteristic	**All (N=51,974)**		**Cluster 1 (N=19,470)**		**Cluster 2 (N=22,256)**		**Cluster 3 (N=8,997)**		**Bonferroni corrected *P* Value**
Cardiovascular risk factor									
Male sex, no. [%]	23021	45.4%	8675	44.6%	10192	45.8%	4154	46.2%	*0.01*
Age, yr	65	11.0	64	12.0	65	11.0	65	12.0	< ***0.001***
BMI, kg/m^2	25.8	5.4	25.56	5.4	25.8	5.3	26.2	5.6	< ***0.001***
SBP, mmHg	139	25.5	137	25.0	139	25.5	142	26.0	< ***0.001***
DBP, mmHg	78.5	13.5	78.5	14.0	78.5	13.5	80.5	14.0	< ***0.001***
Diabetes, no. [%]	2284	4.5%	788	4.1%	943	4.2%	553	6.2%	< ***0.001***
Smoker, no. [%]	1755	3.5%	692	3.6%	770	3.5%	293	3.3%	*0.45*
Alcohol, no. [%]	8468	16.7%	3400	17.5%	3534	15.9%	1534	17.1%	< ***0.001***
CMR									
LVEDV, ml	141.2	44.6	140.8	43.6	140.9	44.8	142.9	46.0	< ***0.001***
LVESV, ml	56.2	23.5	56.0	23.0	56.3	23.5	56.5	24.6	*0.20*
LVM, g	81.0	31.3	79.8	30.4	80.9	31.1	83.6	33.9	< ***0.001***
LVMVR, g/ml	0.6	0.1	0.6	0.1	0.6	0.1	0.6	0.1	< ***0.001***
RVEDV, ml	149.7	51.0	151.2	50.8	149.1	51.8	148.3	50.2	< ***0.001***
RVESV, ml	63.3	28.2	64.5	28.6	62.6	28.4	62.2	27.5	< ***0.001***
WT, mm	9.2	2.1	9.1	2.1	9.1	2.0	9.4	2.2	< ***0.001***

BMI: body mass index, SBP: systolic blood pressure, DBP: diastolic blood pressure, CMR: cardiac magnetic resonance, LVM: left ventricular mass, LVMVR: left ventricular mass to volume ratio, LV: left ventricular, RV: right ventricular, EDV: end-diastolic volume, ESV: end-systolic volume, WT: wall thickness.

Figure 1. Median ECG representing each cluster for each independent lead.



3. Results

The study population exhibited a median age of 65 [12] years and a balanced gender distribution (45.39% males, **Table 1**). For each individual, a total of 241 standard and advanced ECG features were calculated. After applying feature selection, 187 adjusted ECG features were input in a k-means clustering algorithm (k=4), resulting in 4 clusters with distinct ECG phenotypes. Cluster 1 included 19,470 individuals; cluster 2: 22,256; cluster 3: 8,997 and cluster 4: 1,253.

Clusters 1-3 had a balanced gender distribution (~45% males), whereas cluster 4 had a higher proportion of males (70.5%). Moreover, individuals in cluster 4 were, on average, five years older compared to clusters 1-3. Clusters 3 and 4 exhibited higher BMI (~26.3 [5.5] kg/m^2), higher prevalence of diabetes (6.2% and 9.0%, respectively) and higher SBP and DBP (~143 [25] mmHg and 80 [14] mmHg), compared to clusters 1 and 2.

Figure 1 displays the median heartbeat of each independent lead across the identified clusters. Cluster 4 demonstrated clear morphological abnormalities, potentially representing underdiagnosed CVD, and was therefore excluded from further analyses. Cluster 3 had the shortest RR interval 1006 [226] ms, the highest QT dispersion (68 [68] ms), and the widest QRS-T angle (41.39 [59.34] ^{o}).

Random forest analyses highlighted several ECG features as the most important in determining cluster membership, including T-wave amplitude (lead V2), T-wave Hermite basis function 1 (lead V1), ST-segment amplitude (lead V1), TMV index (lead V6), and QRS amplitude (lead V1). These ECG features, adjusted for age, sex and BMI and represented as residuals, were subsequently taken forward into multivariable regression analyses to assess their association with CMR-derived anatomical parameters.

Analysis of CMR features showed that Cluster 3 had the highest left ventricular end-diastolic volume (LVEDV, 142.9 [46.0] ml), left ventricular end-systolic volume (LVESV, 56.5 [24.6] ml), left ventricular mass (LVM, 83.6 [33.9] g/m^2) and wall thickness (WT, 9.4 [2.2]mm, **Table 1**). Cluster 1 exhibited the highest right ventricular end-diastolic volume (RVEDV, 151.2 [50.8] ml), and right ventricular end-systolic volume (RVESV, 64.5 [28.6] ml).

Multivariable linear regression analyses regarding the contribution of CMR features in determining the ECG features revealed that few CMR features were significantly associated with specific ECG features, and these associations were cluster-dependent (**Table 2**). However, the models had limited explanatory power, having a higher R² in cluster 3. No association was found with LVEDV and left ventricular mass to volume ratio LVMVR.

4. Discussion and Conclusions

The main finding of this study is the identification of three distinct ECG-based clusters among a population of over 51,000 individuals without diagnosed CVD in the UKB Imaging study, using unsupervised clustering and evaluating the degree of electro-anatomical alignment within each cluster. These clusters showed significant differences in ECG morphology and CMR derived anatomical features.

Table 2. Contribution of anatomical CMR features within each cluster to determine the five most important adjusted-ECG features.

ECG feature	Cluster	R^2	MAE	RMSE	*P* Value
T wave amplitude Lead V2	Cluster 1	0.01	0.47	0.63	
	Cluster 2	0.03	0.74	0.95	<0.001
	Cluster 3	0.08	0.70	0.95	<0.001
ST amplitude Lead V1	Cluster 1	0.01	0.51	0.67	
	Cluster 2	0.02	0.61	0.83	<0.001
	Cluster 3	0.10	0.80	1.15	<0.001
TMV Lead V6	Cluster 1	0.01	0.34	0.49	
	Cluster 2	0.01	0.34	0.46	0.01
	Cluster 3	0.03	1.11	1.67	<0.001
QRS Amplitude Lead V1	Cluster 1	0.02	0.54	0.70	
	Cluster 2	0.03	0.75	0.98	<0.001
	Cluster 3	0.09	0.87	1.14	<0.001
T-wave's Hermite Base 1 Lead V1	Cluster 1	0.01	0.51	0.71	
	Cluster 2	0.01	0.81	0.91	<0.001
	Cluster 3	0.02	0.90	1.00	<0.001

R^2: coefficient of determination, MAE: mean absolute error, RMSE: root mean squared error.

Individuals in cluster 3 showed greater dispersion of ventricular repolarization and associated with higher left ventricular volumes, ejection fractions, myocardial mass, and increased wall thickness. Cluster 1, in contrast, was characterized by lower QRS and T-wave amplitudes and higher right ventricular volumes, while cluster 2 had higher ST-segment deviation but intermediate CMR features.

The ECG features that most strongly distinguished the clusters were primarily related to ventricular repolarization (T-wave amplitude, ST-segment, TMV index) particularly in the precordial leads. This highlights the importance of incorporating full 12-lead ECG data when exploring cardiac phenotypes. Furthermore, abnormalities in ventricular repolarization have been previously associated with an increased arrhythmic risk[9,16]. Therefore, the presence of distinct repolarization patterns across clusters may not only reflect underlying structural variation but also carry potential prognostic implications.

The degree to which these ECG features could be explained by structural CMR markers varied across clusters, with Cluster 3 demonstrating the strongest electro-anatomical alignment. This may suggest that individuals in Cluster 3 exhibit patterns of electromechanical remodeling, possibly reflecting early or subclinical stages of cardiovascular adaptation or, alternatively, a more efficient and physiologically integrated cardiac phenotype. Considering this, further studies should determine whether such alignment reflects beneficial adaptation or emerging risk.

Among the limitations, the identified clusters represent descriptive, hypothesis-generating phenotypes, and the predominance of White-European ancestry in the UKB cohort limits the generalizability of the findings.

Future work should explore the longitudinal implications of these clusters, assess their prognostic value, and investigate the integration of ECG phenotypes with other clinical and imaging data to enhance cardiovascular risk prediction. Furthermore, validation in external cohorts should be performed.

Acknowledgments

This work was supported by projects PID 2021-128972OA-I00, CNS2022-135899, CNS2023-143599, PID2023-148975OB-I00 and TED2021-130459B-I00, funded by the Spanish Ministry of Science and Innovation (MCIN) and by Gobierno de Aragón, and by fellowships RYC2019-027420-I and RYC2021-031413-I from MCIN and IT 2/24 from Ibercaja-CAI Research Stay Program. MMS acknowledges his British Heart Foundation (BHF) Clinical Research Training Fellowship (FS/CRTF/22/24353). WJY acknowledges the National Institute for Health and Care Research (NIHR) Integrated Academic Training program, which support his Academic Clinical Lectureship post. PBM acknowledges the support of the National Institute for Health and Care Research Barts Biomedical Research Centre (NIHR203330).

References

1. Petersen SE, et al. Reference ranges for cardiac structure and function using cardiovascular magnetic resonance (CMR) in Caucasians from the UK Biobank population cohort. Journal of Cardiovascular Magnetic Resonance. 2017;19.
2. Kamel H, et al. Association between left atrial abnormality on ECG and vascular brain injury on MRI in the Cardiovascular Health Study. Stroke. 2015; 46:711–6.
3. Brar S, et al. Correlation of ECG and cardiac MRI for assessment of ventricular hypertrophy and dilatation in adults with repaired tetralogy of Fallot. International Journal of Cardiology Congenital Heart Disease. 2024; 16:100508.
4. Xie J, et al. Relationship Between Fragmented QRS Complex and Left Ventricular Fibrosis and Function in Patients with Danon Disease. Front Cardiovasc Med. 2022;9.
5. Allencherril J, et al. Correlation of anteroseptal ST elevation with myocardial infarction territories through cardiovascular magnetic resonance imaging. J Electrocardiol. 2018; 51:563–8.
6. Rinta-Kiikka I, et al. Correlation of Electrocardiogram and Regional Cardiac Magnetic Resonance Imaging Findings in ST-Elevation Myocardial Infarction: A Literature Review. Annals of Noninvasive Electrocardiology. Blackwell Publishing Inc.; 2014. p. 509–23.
7. Nezamabadi K, et al. Unsupervised ECG Analysis: A Review. IEEE Rev Biomed Eng. Institute of Electrical and Electronics Engineers Inc.; 2023. p. 208–24.
8. Madrid J, et al. ECG-Based Unsupervised Clustering in Coronary Artery Disease Associates with Ventricular Arrhythmia. Comput Cardiol (2010). IEEE Computer Society; 2023.
9. Madrid J, et al. Unsupervised clustering of single-lead electrocardiograms associates with prevalent and incident heart failure in coronary artery disease. European Heart Journal - Digital Health, 2025.
10. Lyon A, et al. Distinct ECG Phenotypes Identified in Hypertrophic Cardiomyopathy Using Machine Learning Associate With Arrhythmic Risk Markers. Front Physiol. 2018;9.
11. 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. PLoS Med. 2015; 12: e1001779.
12. Young WJ, et al. A Method to Minimise the Impact of ECG Marker Inaccuracies on the Spatial QRS-T angle: Evaluation on 1,512 Manually Annotated ECGs. Biomed Signal Process Control. 2021;64:102305.
13. Ortigosa N, et al. Characterization of Changes in P-Wave VCG Loops Following Pulmonary-Vein Isolation. Sensors (Basel). 2021
14. Murti DMP, et al. K-Nearest Neighbor (K-NN) based Missing Data Imputation. Proceeding - 2019 5th International Conference on Science in Information Technology: ICSITech 2019. 2019;83–8.

Address for correspondence:
Josseline Madrid, jmadrid@unizar.es
Universidad de Zaragoza, Mariano Esquillor s/n, 50018, Zaragoza, Spain.