ECG Based Unsupervised Clustering Model Predicts Heart Failure and Cardiovascular Diseases in Individuals without Cardiovascular Disease

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

[1]Aragon Institute of Engineering Research (I3A), 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

Abstract

Early detection of cardiovascular risk in the general population is challenging. We hypothesize that individuals without known cardiovascular disease (CVD) but at high risk may share electrocardiographic (ECG) features suitable for non-invasive risk stratification. This study aimed to identify clusters of individuals without CVD based on ECG morphology using unsupervised clustering and assess their association with incident CVD risk.

A median heartbeat was derived from 10-second 12-lead resting ECGs from 56,251 individuals without prevalent CVD in the UK Biobank. An unsupervised model classified individuals into k distinct clusters based on their ECG features. Survival analysis assessed the association of each cluster with incident cardiac risk (4-years follow-up).

The model distinguished 2 clusters with varying ECG features, which significantly differed in terms of HF and CVD rate. Cluster 2 (N = 25,097) included the highest rate of heart failure (HF) (0.5%, $p<0.001$) and CVD (4.2%, $p<0.001$). These individuals exhibited ECG features that are associated with higher cardiac risk.

Our analysis identified a group of individuals at risk of HF and CVDs using 10-second ECGs enabling fast and noninvasive risk assessment in apparently healthy populations.

1. Introduction

Cardiovascular disease (CVD) is the main cause of mortality worldwide, accounting for 17.9 million deaths each year[1]. Early detection of CVD is crucial to reduce the burden of these diseases on individuals and healthcare systems. Despite progress in technology, prediction of CVD is still a challenge. Current risk stratification strategies such as Framingham Risk Score[2] and the Pooled Cohort Equations[3] consider only traditional risk factors to predict CVD risk in a 10-year span. The inclusion of electrocardiographic (ECG) information that associate with higher CVD risk could improve risk stratification and facilitate timely initiation of primary prevention therapies.

In supervised statistical modelling, ECG features (e.g., T-peak-to-T-end [Tpe], T-wave morphological variations, heart rate variability) are associated with CVD risk [4–9]. In contrast, unsupervised models are useful to reveal hidden patterns in the ECG data, exploring complex interactions between ECG features, and possibly identifying new risk profiles[10]. Specifically, unsupervised clustering models have proven to be effective in discerning subgroups of individuals with different ECG features identifying associations with higher risk of arrhythmia in hypertrophic cardiomyopathy and coronary artery disease[11,12].

We hypothesize that individuals without known CVD but at a higher CVD risk might share ECG features that can be used to optimize risk stratification. This study aimed to identify distinct clusters of individuals without prevalent CVD based on their ECG morphology using unsupervised machine learning and investigate their association with incident CVD risk.

2. Methods

2.1. The study population

Our study population consisted of 56,251 individuals without known prevalent CVD from the Imaging cohort in the UK Biobank study [13]. The available information for individuals in this cohort included collections of 10-second 12-lead ECGs recorded at rest, as well as health electronic records. The UK Biobank study received approval from the North West Multi-Centre Research Ethics Committee[14], and this work was conducted under application number 8256.

2.2. Definition of cardiovascular diseases

Diseases were defined by the WHO International Classification of Diseases and Related Health Outcomes,

Table 1. Baseline and ECG characteristics for all individuals and clusters in the study

Characteristic	All (N=56,251)		Cluster 1 (N=31,154)		Cluster 2 (N=25,097)		Bonferroni corrected P Value
Demographic and clinical							
Male sex, no. [%]	25796	[45.86%]	16068	[51.58%]	9728	[38.76%]	***< 0.001***
Age, yr	65	[12]	64	[12]	66	[11]	***< 0.001***
BMI, kg/m^2	25.81	[5.41]	25.07	[4.82]	26.89	[5.85]	***< 0.001***
SBP, mmHg	139	[25.5]	137	[25]	142	[25.5]	***< 0.001***
DBP, mmHg	79	[14]	77.5	[13.5]	80	[13.5]	***< 0.001***
Diabetes, no. [%]	2633	[4.68%]	1064	[3.42%]	1569	[6.25%]	***< 0.001***
Smoker, no. [%]	1926	[3.42%]	1129	[3.62%]	797	[3.18%]	0.07
Alcohol, no. [%]	9316	[16.56%]	5202	[16.70%]	4114	[16.39%]	6.69
LVEF, %	56	[7]	56	[7]	56	[8]	0.53
Association with CV events							
AF events, no. [%]	690	[1.23%]	345	[1.11%]	345	[1.37%]	0.06
HF events, no. [%]	238	[0.42%]	104	[0.33%]	134	[0.53%]	***< 0.001***
VA events, no. [%]	72	[0.13%]	39	[0.13%]	33	[0.13%]	11.7
CVD events, no. [%]	2123	[3.77%]	1078	[3.46%]	1045	[4.16%]	***< 0.001***

Tenth Revision codes. Atrial fibrillation (AF) was identified using codes I48.0 to I48.4 and I48.9. Heart failure (HF) was defined by codes I50.0, I50.1, I13.0, I13.2, and I50.9. Ventricular arrhythmias (VA) were defined by codes I47.0, I47.2, I49.0, and I46.0. CVDs were identified using codes I21.0 to I21.4, I21.9, I22.0, I22.1, I22.8, I22.9, I23.0 to I23.6, I23.8, I24.0, I24.8, I24.9, I25.1 to I25.6, I25.8, I25.9, I40.0, I40.1, I40.8, I40.9, I41.1, I412, I41.8, I42.0 to I42.9, I43.0 to I43.2, I43.8, I44.1, I44.2, I47.2, I49.0, I46.0, I46.1, I46.9, I47.0, I48.0 to I48.4, I48.9, I49.9, I50.0, I50.1, I50.9, I51.4, I13.0, I13.2, I50.9, I61.0 to I61.6, I61.8, I61.9, I63.0 to I63.5, I63.8, I63.9, I70.0 to I70.8, I71.0 to I71.9, I74.0 to I74.5, I74.8, I74.9.

2.3. ECG Signal Processing

To remove high frequency noise, the ECG signals were low pass filtered at 40Hz, followed by removal of baseline wander using cubic splines interpolation. Using only sinus beats, a median heartbeat was calculated and a wavelet-based delineator[15] was used to locate the ECG waves onsets, peaks and end timings.

For each lead, parameters related to amplitude, duration and morphology of the QRS and T waveforms were calculated. Each waveform morphology was mathematically characterized by a combination of Hermite functions[16]. For the QRS-complex we considered four Hermite functions, and two for the T wave. The reconstruction error and the width of the Hermite functions were included as parameters in the model. To further characterize the QRS-complex we calculated the upward and downward slopes as in [17]. Additionally, we calculated the T-wave's morphological differences with respect to a normal reference (TMV index) as described in [18]. A total of 22 ECG-related parameters were obtained from each lead's median heartbeat. Additionally, the RR-interval was included as a feature in this model.

2.4. Identification of Clusters using ECG Biomarkers

To refine the set of features, highly correlated features ($r>0.8$) were filtered out to reduce redundancy. Then, principal components analysis was performed to reduce dimensionality in the standardized features. To determine the optimal number of clusters, a grid search from 2 to 10 clusters was conducted using the gap statistic. The gap statistic compares the within-cluster variation in the data to the expected variation under a reference null distribution[19] identifying the optimal number of clusters as the one with the greatest difference between the observed and expected variation, this indicates the final clustering structure is most distinct from random noise. Then it was applied to a k-means clustering algorithm to categorize the ECG features into k distinct clusters. Clustering analysis were performed using Matlab (version R2022b).

2.5. Statistical Analyses

The Wilcoxon rank sum test was used to compare continuous variables, presented as median [interquartile range (IQR)], across clusters, while the Fisher's exact test was employed to compare categorical variables, presented as numbers [percentages]. The variables compared across clusters include: the ECG parameters, demographic

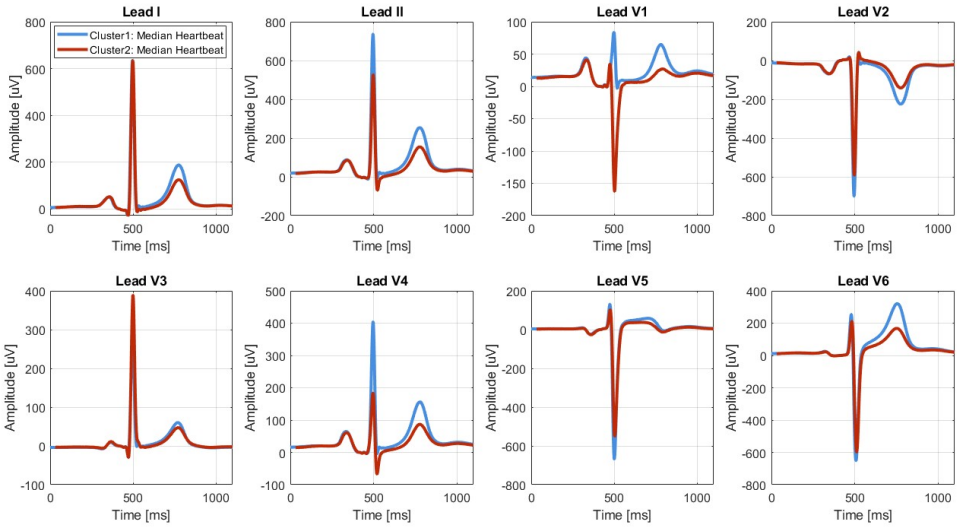

Figure 1. Representative median ECG of each cluster for each independent lead.

information (age, sex, smoking status and alcohol consumption), and clinical features such as body mass index (BMI), left ventricular ejection fraction (LVEF), systolic and diastolic blood pressure (SBP and DBP). The P-values were adjusted using Bonferroni correction.

The association of cluster membership and incident risk of AF, HF, VA and CVD was assessed using Cox-proportional hazards models. Multivariable Cox models were adjusted for covariates including age, sex, smoking status, alcohol consumption, BMI, SBP and DBP, excluding those who had high number of missing data (>10%). Hazard ratio (HR), 95% confidence interval (CI) and P-values were reported for each model.

3. Results

The median age of all 56,251 individuals included in this study was 65 [12] years and 45.86% were male. Further information about demographic and clinical characteristics are found in Table 1. From the initial 177 ECG features (across all 12 leads), 135 remained after the feature selection process. These standardized features were used to derive 82 principal components. The gap statistic indicated the optimal number of clusters was 2. The 2-means clustering algorithm identified 2 distinct clusters of ECG morphological features.

The distribution of males was higher in cluster 1 with 51.58% and lower in cluster 2 with 38.76%. Individuals in cluster 1 were slightly younger with 64[12] years, compared to cluster 2 with 66[11] years. Clinical features in cluster 2 had a higher BMI (26.89 [5.85] kg/m^2), rate of diabetes (6.25%) and systolic blood pressure (142 [25.5] mmHg), compared to cluster 1.

The representative median ECG of each cluster for each lead is shown in Figure 1. Cluster 2 was characterized by shorter RR intervals compared to cluster 1, 988 [198] ms vs 1072 [198] ms respectively. Regarding differences in depolarization across all leads, cluster 2 exhibited wider QRS complexes and flatter upward slopes. Regarding differences in repolarization, individuals in cluster 2 had longer QTc and Tpec intervals and higher TMV indices compared to cluster 1. Additionally, individuals in cluster 1 were characterized by higher T-wave's amplitude. Finally, the larger second Hermite coefficient in the reconstruction of the T-wave in cluster 2 indicates more biphasic T-waves in leads I, II, and V4-V6 in this cluster.

During the 4-year follow-up, there were a total of 690 diagnoses of AF, 238 of HF, 72 of VA and 2,123 of CVD. The identified clusters differed significantly in terms of cardiovascular events rate. Cluster 2 had the highest rate of incident HF 0.53% and CVD 4.16%.

Univariable Cox models showed that cluster 2 was significantly associated with risk of incident HF (HR: 1.70, [1.32 – 2.21], $P < 0.001$) and CVD (HR: 1.29, [1.18 – 1.41], $P < 0.001$). After adjustment for covariates the association remained significant for HF (HR: 1.41, [1.07 – 1.85], $P = 0.01$) and CVD (HR: 1.18, [1.07 – 1.29], $P < 0.001$). SBP and DBP were not considered in the multivariate model due to high number of missing data.

4. Discussion and Conclusions

The main finding of this study is the identification of two distinct clusters of individuals in a large general population cohort using 12 lead, 10-seconds ECGs. Among these, cluster 2 was identified as having a significantly higher risk of incident HF and CVDs. This study highlights the utility of unsupervised clustering models in effectively distinguishing groups of individuals based on ECG features.

Individuals in cluster 2 were characterized by wider QRS complexes, flatter upward slopes, higher TMV

indices, prolonged QTc and Tpec intervals. All of these characteristics have been associated in the literature with higher cardiac risk[17,18,20–22]. Although these values remained within 'healthy' ranges, their elevation may indicate increased variability in the heart's electrical activity which may create conditions that favor the development of adverse cardiac outcomes. Moreover, this cluster showed the highest rate of HF and CVD, and was significantly associated with an increased risk in both univariate and multivariate Cox models independent of traditional risk factors.

Cardiac ion channel and myocardial structural abnormalities are reflected by the ECG waveform, yet the majority of CVD risk stratification tools ignore this information. This study shows that unsupervised ECG-based risk stratification in the general population can capture ECG abnormalities that may group individuals according to CVD risk, with potential application for improved management of CVD. Given that the ECG is a low cost and non-invasive tool, this approach could be implemented for screening in the general population.

This study focused on ECG features related to the QRS complex and T-wave to characterize the ventricular depolarization and repolarization. However, future studies could incorporate information from the P-wave, to identify a cluster with potential association with atrial fibrillation. Additionally, integrating spatial features could further enhance its performance.

Acknowledgments

This work was supported by projects PID 2021-128972OA-I00, CNS2022-135899, CNS2023-143599 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.

References

1. World Health Organization 2024. World Health Statistics 2024.

2. 2018 Prevention Guidelines Tool CV Risk Calculator.

3. Goff DC, Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Gibbons R, et al. 2013 ACC/AHA Guideline on the Assessment of Cardiovascular Risk: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol. 2014;63:2935–59.

4. Niu J, Deng C, Zheng R, Xu M, Lu J, Wang T, et al. The Association and Predictive Ability of ECG Abnormalities with Cardiovascular Diseases: A Prospective Analysis. Glob Heart. 2020;

5. Erelund S, Karp K, Wiklund U, Hörnsten R, Arvidsson S. Are ECG Changes in Heart-Healthy Individuals of Various Ages Related to Cardiac Disease 20 Years Later? Ups J Med Sci. 2021;126:6064.

6. Maron BJ, Friedman RA, Kligfield P, Levine BD, Viskin S, Chaitman BR, et al. Assessment of the 12-Lead ECG as a Screening Test for Detection of Cardiovascular Disease in Healthy General Populations of Young People (12–25 Years of Age). Circulation. 2014;130:1303–34.

7. Terho HK, Tikkanen JT, Kenttä T V, Junttila JM, Aro AL, Anttonen O, et al. Electrocardiogram As a Predictor of Sudden Cardiac Death in Middle-Aged Subjects Without a Known Cardiac Disease. IJC Heart & Vasculature. 2018;20:50–5.

8. Ramírez J, van Duijvenboden S, Aung N, Laguna P, Pueyo E, Tinker A, et al. Cardiovascular Predictive Value and Genetic Basis of Ventricular Repolarization Dynamics. Circ Arrhythm Electrophysiol.

9. Ramírez J, Duijvenboden S van, Young WJ, Orini M, Jones AR, Lambiase PD, et al. Analysing Electrocardiographic Traits and Predicting Cardiac Risk in UK Biobank. 2021;10:204800402110236.

10. 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.

11. Lyon A, Ariga R, Mincholé A, Mahmod M, Ormondroyd E, Laguna P, et al. Distinct ECG Phenotypes Identified in Hypertrophic Cardiomyopathy Using Machine Learning Associate with Arrhythmic Risk Markers. Front Physiol. 2018;9.

12. Madrid J, Munroe PB, Van Duijvenboden S, Ramirez J, Minchole A. ECG-Based Unsupervised Clustering in Coronary Artery Disease Associates with Ventricular Arrhythmia. Comput Cardiol (2010). IEEE Computer Society; 2023.

13. 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.

14. Ramírez J, van Duijvenboden S, Young WJ, Orini M, Lambiase PD, Munroe PB, et al. Common Genetic Variants Modulate the Electrocardiographic Tpeak-to-Tend Interval. Am J Hum Genet.

15. 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.

16. 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.

17. Pueyo E, Sornmo L, Laguna P. QRS Slopes for Detection and Characterization of Myocardial Ischemia. IEEE Trans Biomed Eng.

18. Ramírez J, Kiviniemi A, van Duijvenboden S, Tinker A, Lambiase PD, Junttila J, et al. ECG T-Wave Morphologic Variations Predict Ventricular Arrhythmic Risk in Low-and Moderate-Risk Populations. J Am Heart Assoc. 2022;11.

19. Tibshirani R, Walther G, Hastie T. Estimating the Number of Clusters in a Data Set Via the Gap Statistic. J R Stat Soc Series B Stat Methodol. 2001; 63:411–23.

20. Locati ET, Bagliani G, Padeletti L. Normal Ventricular Repolarization and QT Interval: Ionic Background, Modifiers, and Measurements. Card Electrophysiol Clin [Internet]. 2017;9:487–513.

21. Lund LH, Jurga J, Edner M, Benson L, Dahlström U, Linde C, et al. Prevalence, Correlates, and Prognostic Significance of QRS Prolongation in Heart Failure with Reduced and Preserved Ejection Fraction.

22. Erikssen G, Liestøl K, Gullestad L, Haugaa KH, Bendz B, Amlie JP. The Terminal Part of the QT Interval (T peak to T end): A Predictor of Mortality after Acute Myocardial Infarction. Annals of Noninvasive Electrocardiology. 2012;17:85–94.

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.