Artificial intelligence-enabled electrocardiogram for mortality and cardiovascular risk estimation: a model development and validation study

Arunashis Sau, Libor Pastika, Ewa Sieliwonczyk, Konstantinos Patlatzoglou, Antônio H Ribeiro, Kathryn A McGurk, Boroumand Zeidaabadi, Henry Zhang, Krzysztof Macierzanka, Danilo Mandic, Ester Sabino, Luana Giatti, Sandhi M Barreto, Lidyane do Valle Camelo, Ioanna Tzoulaki, Declan P O'Regan, Nicholas S Peters, James S Ware, Antonio Luiz P Ribeiro, Daniel B Kramer, Jonathan W Waks, Fu Siong Ng

Summary

Background Artificial intelligence (AI)-enabled electrocardiography (ECG) can be used to predict risk of future disease and mortality but has not yet been adopted into clinical practice. Existing model predictions do not have actionability at an individual patient level, explainability, or biological plausibi. We sought to address these limitations of previous AI-ECG approaches by developing the AI-ECG risk estimator (AIRE) platform.

Methods The AIRE platform was developed in a secondary care dataset (Beth Israel Deaconess Medical Center [BIDMC]) of 1 163 401 ECGs from 189 539 patients with deep learning and a discrete-time survival model to create a patient-specific survival curve with a single ECG. Therefore, AIRE predicts not only risk of mortality, but also time-to-mortality. AIRE was validated in five diverse, transnational cohorts from the USA, Brazil, and the UK (UK Biobank [UKB]), including volunteers, primary care patients, and secondary care patients.

Findings AIRE accurately predicts risk of all-cause mortality (BIDMC C-index 0·775, 95% CI 0·773–0·776; C-indices on external validation datasets 0·638–0·773), future ventricular arrhythmia (BIDMC C-index 0·760, 95% CI 0·756–0·763; UKB C-index 0·719, 95% CI 0·635–0·803), future atherosclerotic cardiovascular disease (0·696, 0·694–0·698; 0·643, 0·624–0·662), and future heart failure (0·787, 0·785–0·789; 0·768, 0·733–0·802). Through phenome-wide and genome-wide association studies, we identified candidate biological pathways for the prediction of increased risk, including changes in cardiac structure and function, and genes associated with cardiac structure, biological ageing, and metabolic syndrome.

Interpretation AIRE is an actionable, explainable, and biologically plausible AI-ECG risk estimation platform that has the potential for use worldwide across a wide range of clinical contexts for short-term and long-term risk estimation.

Funding British Heart Foundation, National Institute for Health and Care Research, and Medical Research Council.

Introduction

The electrocardiogram (ECG) has been a fundamental tool in clinical medicine for over a century. With the advent of artificial intelligence (AI), the potential applications of the ECG have substantially expanded, including both diagnostic and predictive capabilities.[1–3] From 2020 onwards, studies have shown the remarkable predictive capabilities of AI-ECG models, not only in terms of predicting mortality, but also cardiac diseases.[4–7]

Existing mortality prediction models are limited by prediction of survival at one or a small number of set timepoints and do not provide information on specific actionable pathways. A high-risk prediction is unhelpful to a clinician if there is no accompanying information on how to improve the survival trajectory of their patient. To make AI-ECG predictions more actionable, considering time-to-event predictions and specific predictions for diseases with established preventive and disease modifying treatments is essential.

Furthermore, the adoption of AI into clinical practice is limited by concerns regarding explainability and biological plausibility. Just as knowledge of drug action mechanisms are important for physicians to have confidence in their application, biological plausibility of AI predictions ensures their credibility and acceptance. To address these limitations of existing risk prediction models, we aimed to develop and perform transnational validation on an AI-ECG risk prediction platform that is not only accurate, but also actionable, explainable, and biologically plausible.

Methods

Study design and cohorts

In this study, we first developed the AI-ECG risk estimation (AIRE) model for the prediction of all-cause mortality. We subsequently developed seven additional submodels. The eight models together are referred to as the AIRE platform. A model development and validation

de Medicina, and Telehealth
Center and Cardiology Service,
Hospital das Clínicas,
Universidade Federal de Minas
Gerais, Belo Horizonte, Brazil
(Prof A L P Ribeiro PhD);
Richard A and Susan F Smith
Center for Outcomes Research
in Cardiology (D B Kramer) and
Harvard-Thorndike
Electrophysiology Institute
(J W Waks MD), Beth Israel
Deaconess Medical Center,
Harvard Medical School,
Boston, MA, USA; Department
of Cardiology, Chelsea and
Westminster Hospital NHS
Foundation Trust, London, UK
(F S Ng)

Correspondence to:
Dr Fu Siong Ng, National Heart
and Lung Institute, Imperial
College London,
London W12 0NN, UK
f.ng@imperial.ac.uk

Research in context

Evidence before this study

From 2019 onwards, studies have shown artificial intelligence (AI) models can accurately diagnose prevalent disease from electrocardiograms (ECGs). Novel AI biomarkers, such as AI-ECG-derived age, can predict future health risks. However, most models give survival probabilities at fixed timepoints rather than personalised risk trajectories over time. We searched PubMed without language restrictions for studies on AI-ECG for mortality prediction. Our search from database inception to Nov 1, 2023, used the search terms ("artificial intelligence" OR "deep learning" OR "neural networks") AND ("ECG" OR "electrocardiogram") AND ("mortality prediction"). We identified 142 articles, 12 of which used ECGs alone to predict mortality at one or a small number of specific timepoints; none predicted long-term time-to-mortality with ECGs. Additional actionable predictions, biological plausibility, detailed explainability, and integration with existing risk factors have also not previously been described.

Added value of this study

Our AI-ECG risk estimation (AIRE) platform can accurately forecast short-term and long-term mortality risk from a single ECG. Additionally, for the first time, we show significantly improved risk prediction of various future events in comparison to existing risk metrics, showing the potential of AI-ECG in individualised risk assessment with and without existing clinical, biochemical, and imaging parameters. Additionally, the platform predicts future atherosclerotic cardiovascular events, heart failure, and ventricular arrhythmia, providing actionable insights. Detailed explainability analyses and biological exploration provide novel insights into AI-ECG predictions and could aid clinician confidence in model predictions.

Implications of all the available evidence

AIRE has been validated in five diverse, transnational cohorts from the USA, Brazil, and the UK. AIRE can improve risk stratification across a wide range of clinical contexts, including primary and secondary care, short-term and long-term risk prediction, and at population and disease-specific levels. Clinicians could act on AIRE's predictions to provide targeted, personalised, and earlier intervention.

flow chart is shown in figure 1. This study complies with all relevant ethical regulations; further details are provided in the appendix (pp 2–3).

We studied five intentionally diverse cohorts from a wide range of patient groups (appendix pp 2–3). Although each cohort consisted of individuals from a specific subset of the population (ie, primary care, secondary care, cardiomyopathy, and volunteers), these cohorts combined can be considered representative of a wide range of patient groups and volunteers. The Beth Israel Deaconess Medical Center (BIDMC) cohort is a secondary care dataset composed of routinely collected data from Boston, MA, USA. The São Paulo-Minas Gerais Tropical Medicine Research Center (SaMi-Trop) is a cohort of patients with chronic Chagas cardiomyopathy.[8] The Clinical Outcomes in Digital Electrocardiography (CODE) cohort is a Brazilian database of ECGs recorded in primary care[9] containing ECGs of both 10 s and 7 s duration. The subset of this dataset with only 10 s ECGs is referred to as CODE-10s. The Longitudinal Study of Adult Health (ELSA-Brasil) cohort consists of Brazilian public servants.[10] The UK Biobank (UKB) is a longitudinal study of volunteers.[11]

Platform development

We developed the AIRE platform using the BIDMC cohort as the derivation dataset. As leads III, aVL, aVR, and aVF are linear combinations of leads I and II, these leads were not used for model development or evaluation. To confirm eight leads were not inferior to 12 leads, we trained a model with 12 leads, which had no better performance than the model with eight leads (appendix

p 7). For mortality endpoints, ECGs without paired life status at 30 days were excluded. 50% of the data were used for training, 10% for validation, and 40% for internal test. Data were split by patient identification number and stratified by presence of ECGs with paired 5-year life status. Therefore, ECGs from the same patient could only be in one of training, validation, or internal test sets. ECGs were treated independently and multiple ECGs per patient (if available) were used. We used a previously described convolutional neural network architecture based on residual blocks[12] and modified the architecture, such that the final layer accommodated a discrete-time survival approach.[13] The discrete-time survival approach allowed the model to account for both time to outcome (mortality) and censorship (ie, loss to follow-up). Further details including ECG pre-processing are in the appendix (p 4). The single-lead (lead I) model, AIRE-1L, was developed with the same method; we used the same BIDMC data split but with just lead I as the model input. The TRIPOD guidelines were used.

Using the CODE dataset, we finetuned the model to be more representative of a primary care population (AIRE-primary care) with a low risk of adverse events. We used 75% of the CODE dataset for finetuning, with 5% used as a validation set. The final 20% was used for internal validation of AIRE-primary care. Data were split by patient identification number.

We also developed five other subsequent models by fine-tuning the AIRE model separately for cardiovascular death (AIRE-CV death), non-cardiovascular death (AIRE-NCV death), ventricular arrhythmia (AIRE-VA), atherosclerotic cardiovascular disease (AIRE-ASCVD),

and heart failure (AIRE-HF). These models were trained in the BIDMC dataset. We used the same splits as were used for training the original model. We performed fine-tuning by loading the previous model and training using a low learning rate without freezing any layers. Diverse external validation was performed to explore validity in diverse cohorts and ensure the models were not overfit to the development cohort. Internal validation and external validation datasets are shown in figure 1 and the appendix (p 5).

We compared AIRE-primary care and AIRE-ASCVD to the Stanford Estimator of ECG Risk (SEER), which had a similar goal of predicting cardiovascular mortality and atherosclerotic cardiovascular disease.[7] The model code and weights were downloaded and the performance evaluated in the UKB, as this was a dataset external to AIRE-primary care, AIRE-ASCVD, and SEER with cause of death and atherosclerotic cardiovascular disease event data available.

Statistical analysis

In the test set, we generated predictions for all ECGs for the primary analyses of all-cause mortality. Sensitivity analyses, including a single random ECG per patient, were also performed. For analyses requiring a single predictor value, the probability of survival at 5 years was used. Risk quartiles (low, intermediate-low, intermediate-high, and high) were defined based on values in the validation set. Given the diverse populations and event rates evaluated, risk quartiles were redefined in each dataset. In each case, if categorical risk levels were required, 5% of the dataset was used to define the quartiles and evaluation was performed in the remaining 95%. Kaplan–Meier curves comparing the risk quartiles were plotted and statistical significance assessed with the log-rank test. Sensitivity analyses of the original validation set quartiles were also performed. If binary predictions were required for metrics, Youden's index was used to identify the threshold that maximised the sensitivity and specificity in the validation set.

Cox models were fit with the test dataset comparing demographics, clinical variables, imaging parameters, and AIRE platform predictions. For the Cox models incorporating AIRE platform predictions, all model outputs (ie, predicted probabilities of death at each timepoint) were used as inputs, as well as age, sex, heart rate, PR interval, QRS duration, and QTc interval. These models are designated AIRE-Cox for the AIRE model and AIRE-CV death-Cox, AIRE-ASCVD-Cox, AIRE-VA-Cox, and AIRE-HF-Cox for the other models. Comparator Cox model components are stated for each analysis; they included left ventricular ejection fraction (LVEF), ECG parameters (ie, heart rate, PR interval, QRS duration, and QTc interval), and cardiovascular risk factors (diabetes, hypertension, smoking history, hyperlipidaemia, and ethnicity). Atherosclerotic cardiovascular disease risk factors were systolic blood pressure, total cholesterol,
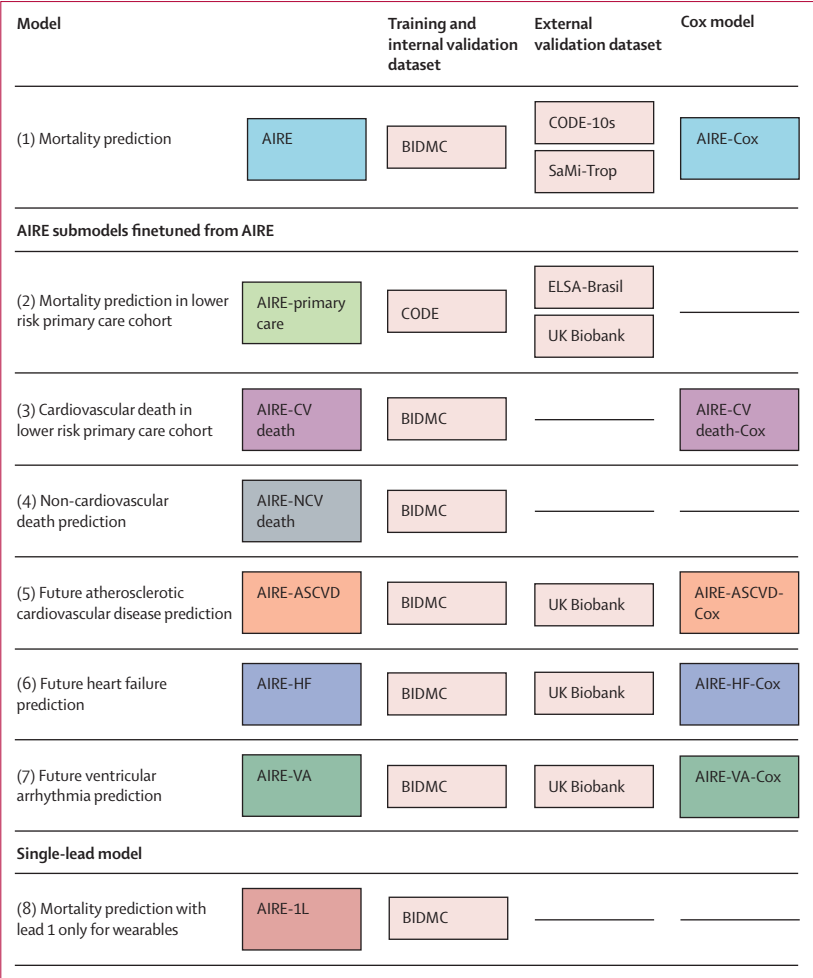
Figure 1: Overview of models in the AIRE platform
Schematic depicting all eight models in the AIRE platform, training datasets, and validation datasets. AIRE and AIRE-primary care were trained for all-cause mortality; the remaining models were trained for the outcomes they are named after. CODE-10s denotes the subset of CODE with ECGs of 10 s duration. Cox models, if applicable, include the AIRE model, age, sex, and ECG parameters. AI=artificial intelligence. AIRE=AI-ECG risk estimator. ASCVD=atherosclerotic cardiovascular disease. AIRE-1L=AIRE using lead I only. BIDMC=Beth Israel Deaconess Medical Center. CODE=Clinical Outcomes in Digital Electrocardiography. CV=cardiovascular. ECG=electrocardiography. ELSA-Brasil=Brazilian Longitudinal Study of Adult Health. HF=heart failure. NCV=non-cardiovascular. SaMi-Trop=São Paulo-Minas Gerais Tropical Medicine Research Center. VA=ventricular arrhythmia.

HDL cholesterol, hypertension, smoking history, diabetes, and ethnicity. 10-year atherosclerotic cardiovascular disease risk was assessed with the pooled cohort equation. Atherosclerotic Risk in Communities (ARIC) heart failure risk factors were BMI, systolic blood pressure, prevalent atherosclerotic cardiovascular disease, diabetes, smoking history, previous myocardial infarction, hypertension, and ethnicity. For the Cox model comparisons, complete case analysis was used. Although the clinical, demographic, and imaging data were unlikely to be missing at random, they are likely to be available for the patient groups in whom these predictions are likely to be relevant. Nested Cox models were compared with the likelihood ratio test, while non-nested Cox models were compared with the partial likelihood ratio test. Statistical

analyses were performed with R (version 4.2.0) statistical package or Python (version 3.9). Primary analysis used all ECGs that met inclusion criteria for each analysis; sensitivity analyses used a single ECG per participant.

Diagnostic and imaging data

ICD-9 and ICD-10 codes were used to define the presence or absence of disease in the BIDMC and UKB cohorts. Cardiovascular death in the BIDMC cohort was defined as mortality occurring within 30 days of a diagnostic code for acute myocardial infarction, ischaemic stroke, intracranial haemorrhage, sudden cardiac death, or heart failure.[7,14] In the UKB, cause of death was ascertained based on the ICD-10 code stated as the primary cause of death. Diagnostic codes were not available in the SaMi-Trop, CODE, and ELSA-Brasil datasets. Echocardiograms within 60 days of an ECG were linked and used for analyses incorporating echocardiographic parameters. Medication usage was not available in the BIDMC cohort, therefore ICD-9 and ICD-10 codes consistent with a diagnosis of hypertension were used to code for antihypertensive medication use for calculation of the pooled cohort equation. Blood results and blood pressure readings taken within 180 days of the ECG were averaged. Sensitivity analyses were performed with 90-day and 30-day results. Normal ECG definition is described in the appendix (p 5).

Explainability and biological plausibility

To understand the ECG morphologies associated with predicted survival, we used three approaches. Median beats were extracted with the BRAVEHEART ECG analysis software.[15] First, we trained a variational autoencoder using median ECG beats (appendix p 5). In preliminary analyses, models based on only the variational autoencoder latent features were inferior to the supervised deep learning approach (appendix p 7), therefore the variational autoencoder was used for explainability only, and not used for AIRE model training or any of the prediction models described in this study. Variational autoencoder latent features were input into a linear regression with predicted survival as the output. The top three most important features as assessed by the *t* value were visualised by latent traversal. Second, using the median beats, we calculated the average waveform from the 10 000 ECGs with the lowest and highest AIRE predicted mortality. The mean and SD of these waveforms was then plotted. Third, gradient-weighted class activation maps were plotted with tf-keras-vis version 0.8.7.[16]

To understand the biology underlying AIRE predictions, we performed phenome-wide association studies (PheWAS). We conducted PheWAS analysis in the UKB, which contains data from more than 3000 phenotypes derived from patient measurements, surveys, and investigations. Univariate correlation was performed to investigate the association between ECG predicted

survival and phenotypes adjusted for age, sex, and age². We also investigated the association of predicted survival with continuous echo traits in the BIDMC dataset. Left ventricular trabeculation was calculated as previously described.[17] Deep learning-derived brain age was calculated as previously described.[18] Further methods including for genome-wide association study (GWAS) are described in the appendix (p 6).

Role of the funding source

The funders of the study were not involved in the study design, data collection, analysis, interpretation, or reporting.

Results

AIRE accurately predicts mortality across diverse timepoints. In the BIDMC cohort, 1 163 401 ECGs were available from 189 539 participants. Mean follow-up was 5·46 years (SD 5·81) on a per ECG basis and 3·41 years (4·08) taking a random ECG per participant. 34 851 (18·4%) of 189 539 participants died during follow-up (appendix p 14). AIRE produces patient-specific survival curves from only a single ECG and can predict time-to-death (figure 2A and B). Figure 2C shows the evolution of AIRE-predicted survival based on multiple ECGs performed over several years of follow-up.

In the hold-out test set, AIRE predicted all-cause mortality with a concordance-index of 0·775 (95% CI 0·773–0·776). Detailed performance metrics are shown in the appendix (pp 15–17). Figure 3A shows the marked separation of survival curves of risk quartiles in the test set. The appendix (p 18) shows age-adjusted and sex-adjusted hazard ratios for high-risk versus low-risk quartiles for all cohorts. When considering only ECGs labelled as normal by cardiologists, there remained a significant difference in mortality between patients at high risk and low risk (figure 3B). Importantly, AIRE had similar performance in both men and women and in major ethnic groups (figure 4A and appendix p 19). Further results are reported in the appendix (p 7).

AIRE is superior to demographic data and traditional risk factors for mortality prediction. We compared the ability of AIRE to predict mortality against use of demographic data, risk factors, and risk scores in the BIDMC test set (figure 4B). AIRE-Cox had a significantly higher C-index than all other parameters combined (0·794 [95% CI 0·792–0·795] vs 0·759 [0·758–0·761], p<0·0001). In the BIDMC test set, AIRE-CV death predicted cardiovascular death with a C-index of 0·832 (95% CI 0·831–0·834). AIRE-CV death-Cox had a significantly higher C-index for prediction of cardiovascular death than all other parameters combined (0·844 [95% CI 0·839–0·849] vs 0·795 [0·789–0·801], p<0·0001; figure 4C). Finally, AIRE-NCV death predicted non-cardiovascular death with a C-index of 0·749 (95% CI 0·747–0·751). The appendix (p 8) has sensitivity analyses of a single random ECG per patient.
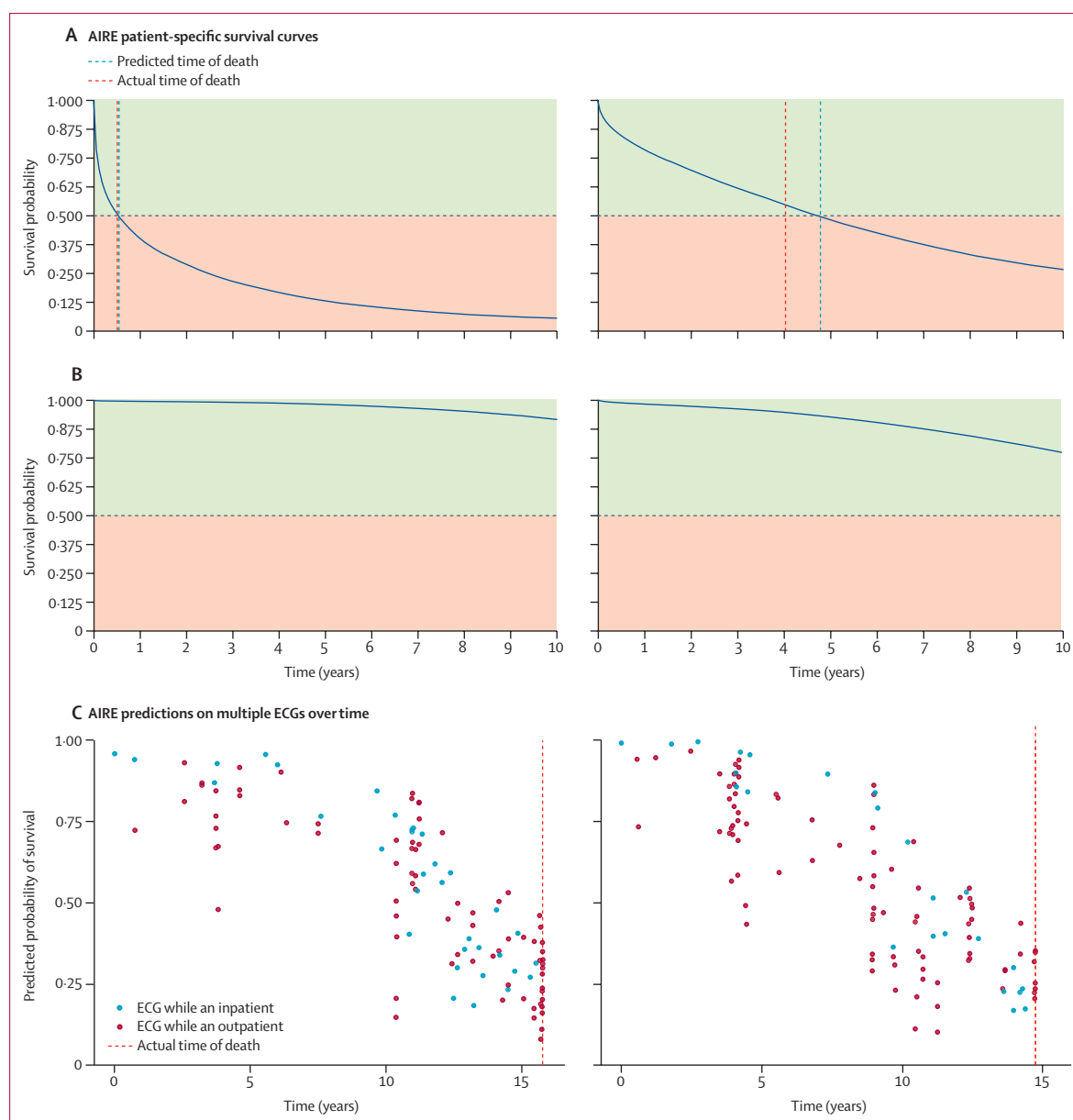
Figure 2: Example patient-specific survival predictions

AIRE outputs patient-specific survival curves. Two examples are shown for patients who died during follow-up (A), and two examples are shown of patients who survived through the follow-up period (B). Dashed red lines indicate the date of death and dashed blue lines indicate AIRE-predicted date of death. (C) Two examples of patients with multiple ECGs during the study period are shown. Each panel contains data from one patient. Each dot is a survival prediction from a single ECG. Blue dots indicate inpatient ECGs, and red dots are outpatient ECGs. AIRE-predicted survival trends downward over time and predicted probability of survival is particularly low before actual time of death (red dashed line). Inpatient episodes show temporary falls in predicted survival, and outpatient predictions are more stable over time. AI=artificial intelligence. AIRE=AI-ECG risk estimator. ECG=electrocardiography.

AIRE predicts mortality in transnational external datasets. We tested if AIRE was applicable to a wide range of settings, from volunteers to primary care to patients with cardiomyopathy. First, we evaluated the performance of AIRE in the SaMi-Trop cohort of patients with chronic Chagas cardiomyopathy.[8] The appendix has the dataset demographics for all cohorts (p 14) and results summary (pp 18, 24). The C-index was 0·773 (95% CI 0·733–0·813; figure 3C, appendix p 9).

The CODE cohort is a Brazilian database of ECGs recorded in primary care.[9] To investigate the external validity of AIRE, we first evaluated the performance of AIRE without any fine-tuning. As AIRE was trained exclusively on 10s ECGs, we evaluated the model on the 10s subset (CODE-10s); AIRE had a C-index of 0·762
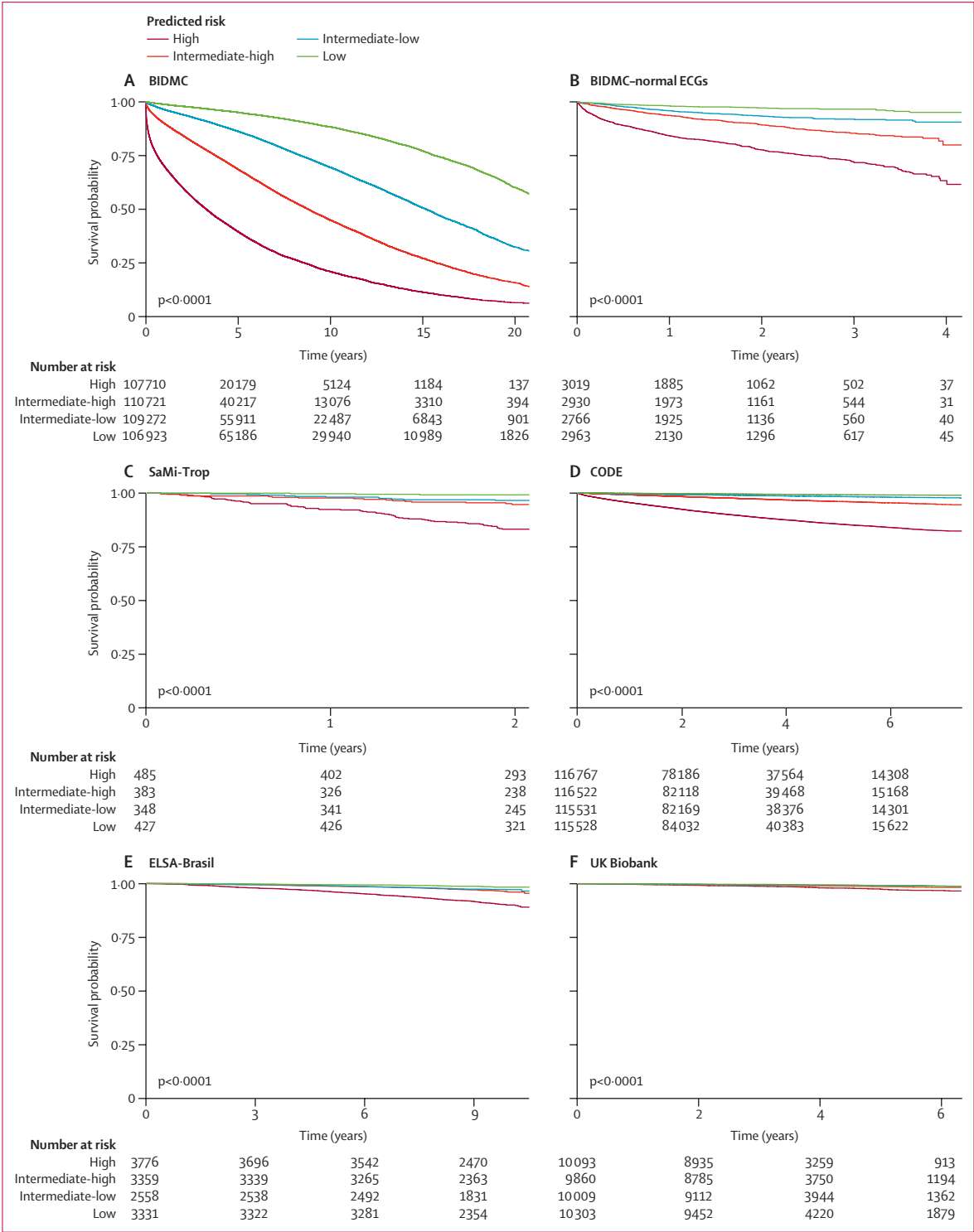
***Figure 3:* Mortality prediction Kaplan–Meier curves**
Kaplan–Meier curves of AIRE-predicted all-cause mortality by risk quartile in (A) the whole BIDMC test set, (B) BIDMC normal ECGs, and (C) SaMi-Trop cohort of patients with Chagas disease. AIRE-primary care predictions are shown for (D) CODE cohort of primary care patients in Brazil, (E) ELSA-Brasil volunteer cohort, and (F) UK Biobank volunteer cohort. AI=artificial intelligence. AIRE=AI-ECG risk estimator. BIDMC=Beth Israel Deaconess Medical Center. CODE=Clinical Outcomes in Digital Electrocardiography. ECG=electrocardiography. ELSA-Brazil=Brazilian Longitudinal Study of Adult Health. SaMi-Trop=São Paulo-Minas Gerais Tropical Medicine Research Center.

(95% CI 0·759–0·765) for all-cause mortality prediction. AIRE-primary care more accurately predicted mortality with an improved C-index of 0·802 (95% CI 0·799–0·805; figure 3D) than AIRE. When considering only the ECGs labelled as normal (26766 ECGs from 21897 people), there was a significant difference in mortality between high-risk and low-risk participants based on model predictions (appendix p 18).

Further evaluation of AIRE-primary care was performed in another independent external dataset, ELSA-Brasil (n=13739), a volunteer cohort of civil servants from Brazil.[10] The C-index was 0·713 (95% CI 0·691–0·735; figure 3E). When considering normal ECGs only, there was a significant difference in mortality between participants at high and low risk (appendix p 18).

Finally, we also evaluated the performance of AIRE-primary care in the UKB, a healthy volunteer population (n=42 386) with only 526 (1·2%) deaths during follow-up. The C-index was 0·638 (95% CI 0·608–0·668; figure 3F) for all-cause mortality. As cause of death was available in the UKB, we also examined the ability of AIRE-primary care to predict cardiovascular death. C-index for cardiovascular death was 0·695 (0·636–0·754). In 2023, Hughes and colleagues reported SEER with the similar goal of predicting cardiovascular mortality.[7] AIRE-primary care was superior to SEER at predicting cardiovascular death in the UKB (SEER C-index 0·572 [95% CI 0·514–0·630], p<0·0001).

AIRE-ASCVD was able to predict future atherosclerotic cardiovascular disease in participants without known atherosclerotic cardiovascular disease (C-index 0·696 [95% CI 0·694–0·698]; 227 588 ECGs from 56 598 participants). Detailed performance metrics are shown in the appendix (pp 20–22). We externally validated these findings in the UKB, a healthy volunteer population. AIRE-ASCVD had reduced performance in predicting atherosclerotic cardiovascular disease (C-index 0·643 [95% CI 0·624–0·662]), but was

significantly better than the SEER model (SEER C-index 0·547 [95% CI 0·527–0·567]; p<0·0001).[7]

We compared AIRE-ASCVD to other risk parameters, including the pooled cohort equation and atherosclerotic cardiovascular disease risk factors, in a subset of outpatients (4580 ECGs from 2926 participants) in the BIDMC test set with appropriate available data. AIRE-ASCVD-Cox had a significantly higher C-index than all
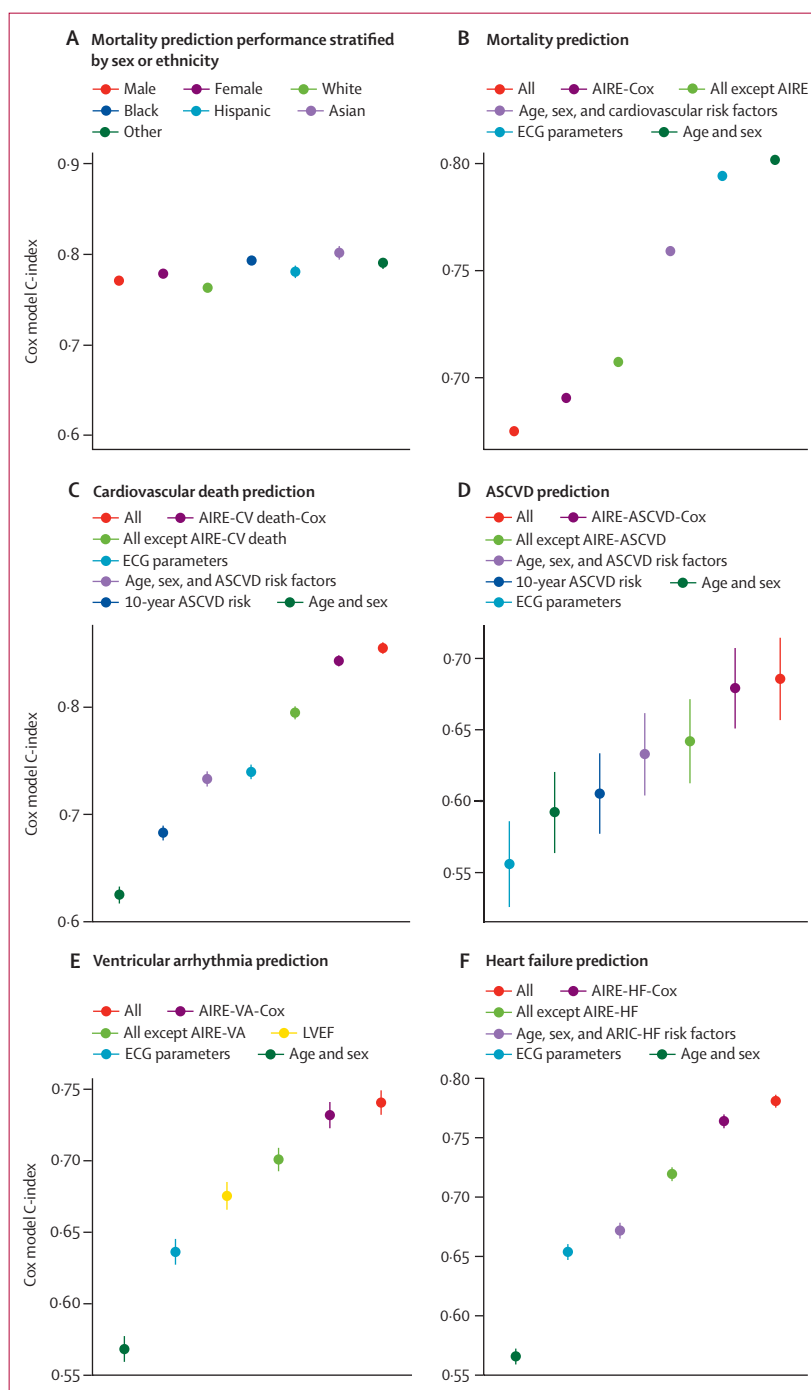
***Figure 4:* Mortality and disease prediction performance**
(A) Comparison of AIRE performance across sex and major ethnic groups; AIRE performs similarly across all demographic groups. Using Cox models, AIRE was compared with existing risk factors and ECG parameters. In these Cox models, age, sex, and ECG parameters were incorporated with AIRE to create AIRE-Cox. In all comparisons, AIRE-Cox had a significantly higher C-index than all comparators for both (B) all-cause mortality and (C) cardiovascular mortality. We also evaluated disease specific models, (D) AIRE-ASCVD, (E) AIRE-VA, and (F) AIRE-HF. ECG parameters: heart rate, PR interval, QRS duration, and QTc interval. Cardiovascular risk factors: diabetes, hypertension, smoking history, hyperlipidaemia, and ethnicity. Atherosclerotic cardiovascular disease risk factors: systolic blood pressure, total cholesterol, HDL cholesterol, hypertension, smoking history, diabetes, and ethnicity. 10-year atherosclerotic cardiovascular disease risk assessed with the pooled cohort equation. Atherosclerotic Risk in Communities heart failure risk factors: BMI, systolic blood pressure, prevalent atherosclerotic cardiovascular disease, diabetes, smoking history, previous myocardial infarction, hypertension, and ethnicity. AI=artificial intelligence. AIRE=AI-ECG risk estimator. ASCVD=atherosclerotic cardiovascular disease. CV=cardiovascular. ECG=electrocardiography. HF=heart failure. LVEF= left ventricular ejection fraction. VA=ventricular arrhythmia.
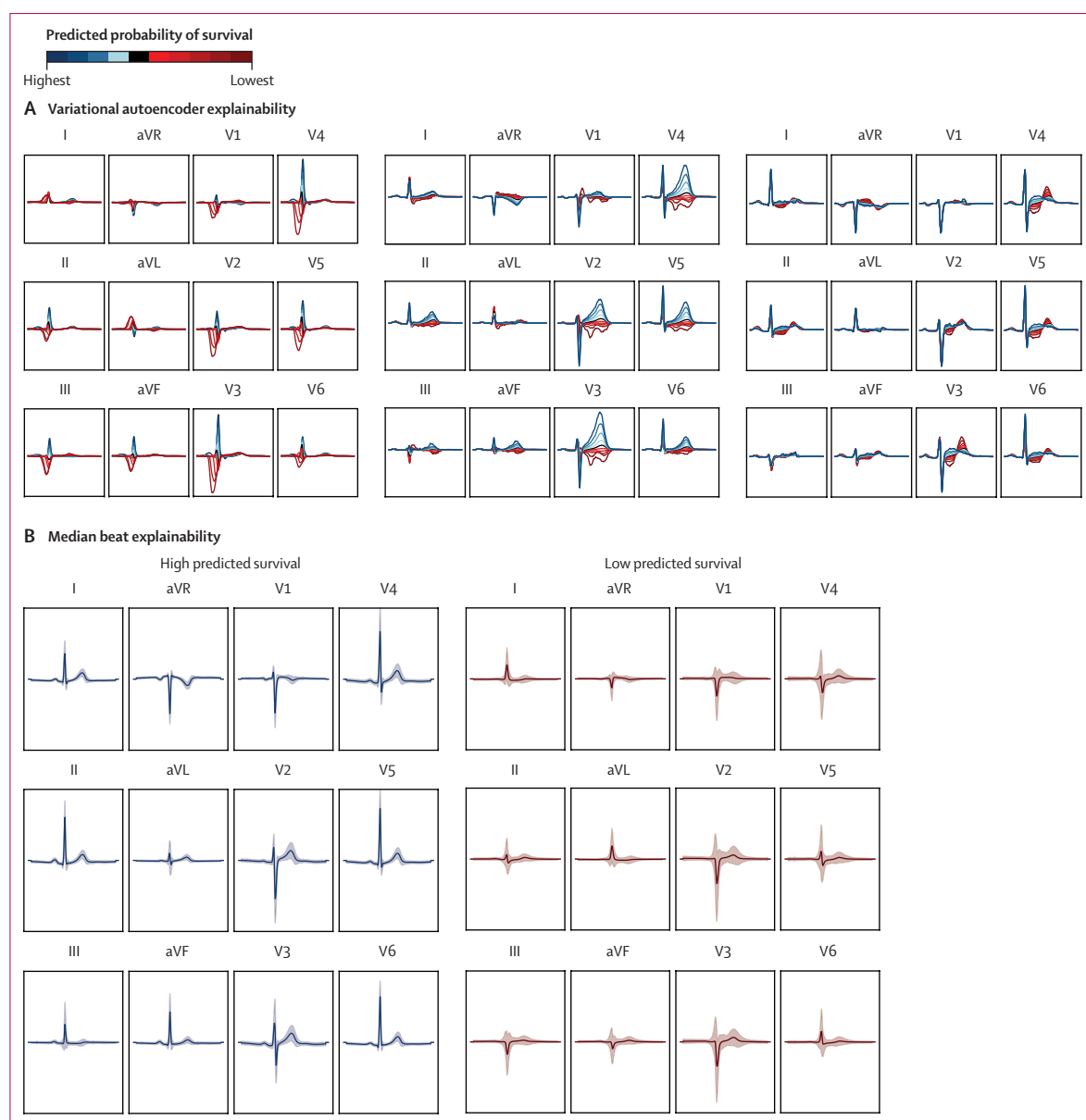
Figure 5: AIRE model explainability—BIDMC test set
Two explainability approaches to understand ECG morphologies associated with AIRE predictions are shown. (A) A variational autoencoder was used to identify the most important morphological features in AIRE-predicted mortality; each subpanel shows one of three latent features, identifying the importance of a broad QRS complex in a left bundle morphology and biphasic and inverted T waves. (B) Mean with SD (shaded region) ECG median waveforms for the 10 000 highest and lowest predicted survival ECGs from the BIDMC test set. This analysis identified poor R wave progression, low QRS amplitude, and T wave flattening or inversion as important features in AIRE-predicted survival. AI=artificial intelligence. AIRE=AI-ECG risk estimator. BIDMC=Beth Israel Deaconess Medical Center. ECG=electrocardiography.

other factors combined (0·679 [95% CI 0·651–0·708] vs 0·642 [0·613–0·672], p<0·0001; figure 4D). The appendix (p 10) shows sensitivity analyses using a single random ECG per participant for all three actionable prediction endpoints. Numerical results for all Cox models are shown in the appendix (p 23).

AIRE-VA was able to accurately predict future ventricular arrhythmia (C-index 0·760 [95% CI 0·756–0·763], 393 203 ECGs from 62 443 participants) in participants without a previous history of ventricular arrhythmia. Compared with other conventional risk parameters, including ECG parameters and LVEF (101 935 ECGs from 21 093 participants), AIRE-VA-Cox had a significantly higher C-index than all other factors combined (0·722 [95% CI 0·716–0·728] vs 0·699 [0·693–0·704], p<0·0001; figure 4E). The appendix (p 7) describes performance in subgroups of LVEF less than 50% and dilated cardiomyopathy. In the UKB (34 400 participants

with both ECG and cardiac MRI [CMRI] and without previous ventricular arrhythmia, 44 events), AIRE-VA had similar performance in predicting first occurrence of ventricular arrhythmia (0·719 [95% CI 0·635–0·803]) and performed at least equivalent to LVEF from CMRI (C-index 0·595 [95% CI 0·494–0·697], p=0·11).

AIRE-HF was able to accurately predict future heart failure in participants without a previous history of heart failure (C-index 0·787 [95% CI 0·785–0·789], 310 200 ECGs from 61747 unique participants). In a subset of patients with the available data, we compared AIRE-HF-Cox to other conventional risk parameters in Cox models (36 486 ECGs from 12 288 subjects), including heart failure risk factors identified in the ARIC study.[19] AIRE-HF had a significantly higher C-index than all other factors combined (0·761 [95% CI 0·755–0·767] vs 0·716 [0·710–0·722], p<0·0001; figure 4F). In the external validation cohort, UKB, AIRE-HF-Cox had similar performance at predicting future heart failure (C-index 0·768 [0·733–0·802]).

We trained a single-lead version of AIRE using lead I only (AIRE-1L). The performance of AIRE-1L (C-index 0·751 [95% CI 0·750–752]; appendix p 11) was only slightly inferior in discrimination compared with the 8-lead AIRE model.

Using a variational autoencoder, we found features of QRS morphology, particularly broader and more left bundle-branch block morphologies, inverted and biphasic T waves, and ST segment changes were identified as the most significant morphological features associated with high predicted mortality (figure 5A). In a second approach, using median beats in the BIDMC test set, we found poor precordial R wave progression, low QRS amplitude, and T wave flattening or inversion as important features in AIRE-predicted survival (figure 5B). The third approach, gradient-weighted class activation maps, supports the contribution of the QRS complex and T waves to AIRE predictions (appendix p 13).

We performed a genome-wide association study to identify genetic loci associated with high-risk AIRE predictions (figure 6A, appendix p 25). We found significant
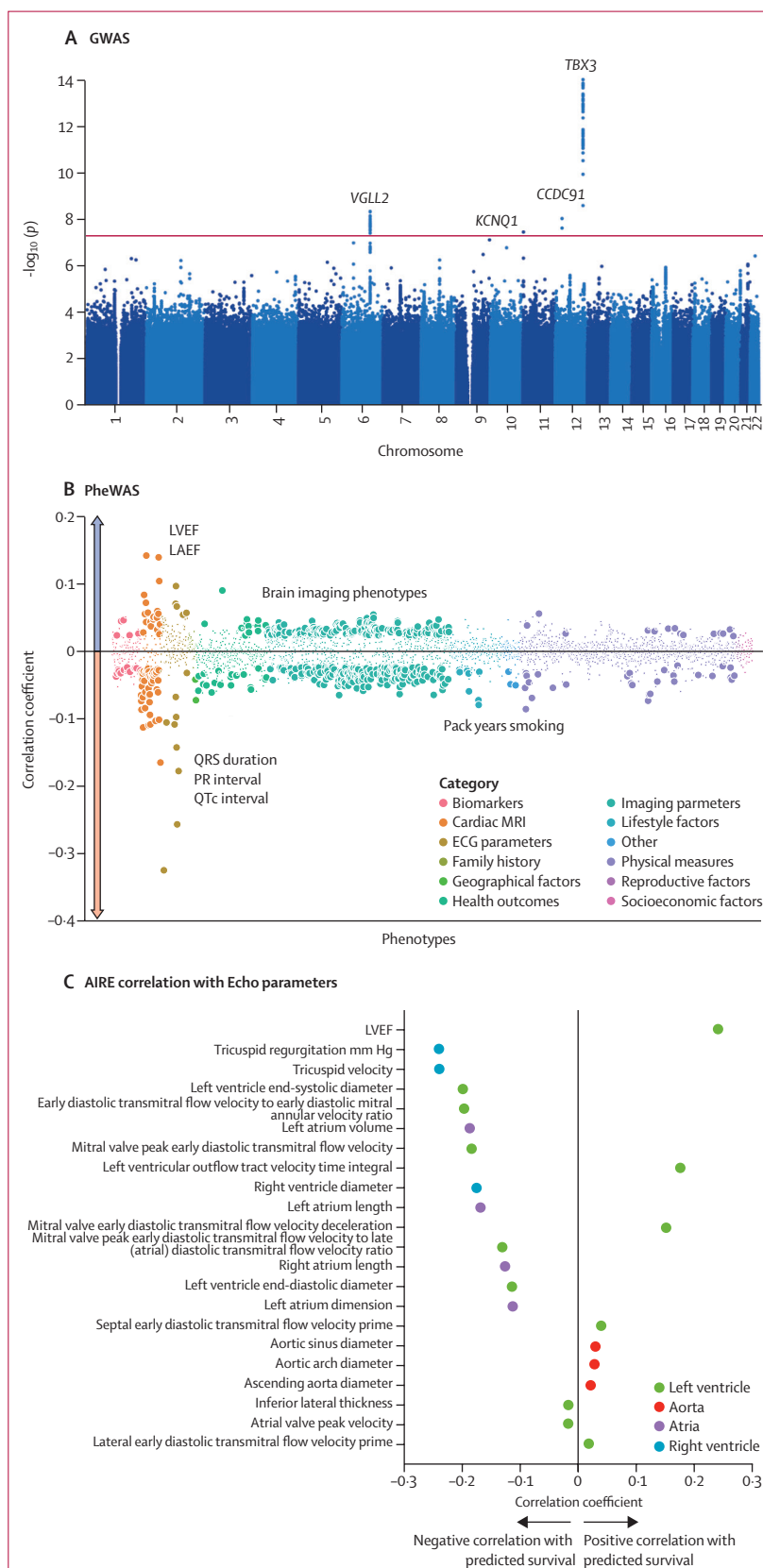
Figure 6: **Exploration of underlying biology of AIRE predictions through PheWAS and GWAS**

Three approaches to explore biological associations of AIRE predictions are shown. (A) GWAS in the UK Biobank. Manhattan plot of genomic loci associated with predicted survival is shown. Nearest genes to significant single nucleotide polymorphisms are shown. The red line depicts the genome-wide significant threshold (p<5×10⁻⁸). (B) PheWAS in the UK Biobank showing phenotypic associations with AIRE predictions. Cardiac associations include LVEF and atrial and right ventricular phenotypes. Non-cardiac associations included brain phenotypes such as total volume of white matter hyperintensities and pack years of smoking. (C) Association of AIRE-predicted survival with echocardiographic parameters in the BIDMC test set. LVEF was positively associated with AIRE-predicted survival, whereas chamber dilatation and tricuspid regurgitation velocity were negatively associated. AI=artificial intelligence. AIRE=AI-ECG risk estimator. ECG=electrocardiography. GWAS=genome-wide association study. LAEF=left atrium ejection fraction. LVEF=left ventricle ejection fraction. PheWAS=phenome-wide association study.

loci adjacent to *TBX3*, *VGLL2*, *CCDC91*, and *KCNQ1*. The identified genes reinforce our ECG explainability analyses, which includes our variational autoencoder analysis. *TBX3* has been associated with QRS duration, QRS voltage, and QRS-T angle,[20,21] abnormalities of which were shown to be associated with high-risk predictions on our analysis of variational autoencoder latent features. *KCNQ1* is associated with long QT syndrome 1 and QT interval.[22] *VGLL2* has also been associated with ECG morphologies.[23] These genes were also associated with non-ECG traits. *TBX3* has been associated with blood pressure,[24] myocardial mass,[20] and trabecular development.[17] *VGLL2* has been associated with blood pressure, atrial fibrillation, BMI, and AI-ECG derived delta-age.[23–27] *KCNQ1* is also associated with metabolic syndrome phenotypes.[28] Finally, *CCDC91* is associated with BMI.[28]

We performed a PheWAS in the UKB (figure 6B). In particular, CMRI associations included reduced LVEF, more positive (ie, abnormal) global longitudinal strain, increased left ventricular mass, and increased left atrial size, which were correlated with reduced AIRE-predicted survival. We also specifically examined the association between predicted survival and left ventricular trabeculation and found a significant negative correlation (appendix p 26). In BIDMC echocardiographic analyses (figure 6C), we found LVEF was positively correlated with predicted survival whereas left atrial volume, and surrogate measures of pulmonary pressure (tricuspid regurgitation velocity), and right ventricular diameter were both negatively correlated with predicted survival. Finally, investigation of non-cardiac imaging phenotypes identified associations with multiple brain imaging phenotypes including the total volume of white matter hyperintensities and deep learning-derived brain age (appendix p 27).[18]

Discussion

We describe, for the first time, an actionable, explainable, and biologically plausible mortality and risk prediction AI-ECG platform of eight AI-ECG models. Importantly, our platform was externally validated across ethnically and demographically diverse transnational cohorts.

This study substantially extends the work of others on mortality protection with ECG. Raghunath and colleagues described the use of deep learning for mortality prediction,[5] and in 2023, Sun and colleagues built upon this work.[6] Our study has several considerable differences from these previous publications. First, the use of survival neural network architecture provides our model with the ability to predict time of death without being constrained to a small number of timepoints. Additionally, this architecture allows us to use training data from participants who were censored (time of death not known), allowing the use of more real-world data, which is almost always censored, for model training. Furthermore, by comparing our model to existing clinical risk factors and imaging parameters, we have shown the

significant additive value of our model beyond traditional approaches. Finally, we performed external validation across diverse populations, showing the wide applicability of our model platform.

Model positive predictive value (PPV) was reduced in the volunteer cohorts, which is unsurprising, given the low event rate in these populations and the dependence of PPV on disease prevalence or, in this case, incidence.[29] Our findings are similar to other AI-ECG studies in this area, which have reported similar PPVs in specific cohorts.[7,30] We propose the AIRE platform would be best used in different ways depending on the clinical population. In low-risk populations, the high negative predictive value would allow confident reassurance of individuals at lowest risk, while in higher risk populations, the high positive predictive value could allow clinicians to identify patients at increased risk of adverse events.

The AIRE platform can also predict future cardiovascular events, such as atherosclerotic cardiovascular disease, heart failure, and ventricular arrhythmia, in addition to predicting mortality. Atherosclerotic cardiovascular disease prediction is used extensively in international guidelines for decision making around lipid lowering therapies.[31] In this study, we have shown that AIRE-ASCVD provides additional information that could improve atherosclerotic cardiovascular disease risk prediction, and it is superior and additive to the existing pooled cohort equation. AIRE-primary care was superior to SEER, an AI-ECG model described in 2023, for cardiovascular death, and AIRE-ASCVD was superior to SEER for atherosclerotic cardiovascular disease prediction.

Similarly, predicting future ventricular arrhythmia is a particularly important endpoint, as there are clear preventive and therapeutic options. Guidelines advocate LVEF as the primary factor in determining eligibility for a primary prevention ICD. In our study, we showed that AIRE-VA is a better predictor of future ventricular tachycardia and ventricular fibrillation than LVEF and could therefore potentially be incorporated into this decision-making framework.

Finally, predicting future heart failure is important given the high number of unplanned hospital admissions due to undiagnosed heart failure.[32] Through prediction of heart failure with AIRE-HF, early clinical assessment, echocardiography, and institution of appropriate therapies might reduce adverse events, particularly in heart failure with reduced ejection fraction.

Extending risk prediction to the single-lead ECG is particularly important given the rapidly increasing number of single-lead devices, including consumer products.[33] Our study highlights the excellent performance of AIRE at mortality prediction on only a single lead, which could be particularly applicable for inpatient cardiac monitoring, where frequent AIRE predictions could be employed for early detection of risk. Single-lead AIRE models could also be used for remote monitoring in outpatient settings with the use of wearable devices,

for example in patients with chronic diseases such as heart failure, in which high-risk predictions could trigger pre-emptive treatments to prevent hospital admissions.

A substantial challenge in deep learning is explainability of the model predictions. Our ECG explainability findings are in line with previous studies that highlight these features as being prognostically important[34–36] and are reinforced by our GWAS findings, with consistent associations with QRS duration, voltage, QRS-T angle, and QT interval. We identified phenotypic associations with cardiac chamber structure and function, including trabeculation and myocardial mass. These plausible biological pathways were reinforced by GWAS associations with *TBX3*. *VGLL2* has also been described in relation to AI-ECG derived delta-age, which is a marker of accelerated biological ageing.[9] We also identified AIRE-predicted survival as inversely correlated with deep learning-derived brain-age. Finally, we identified variants in *KCNQ1* and *CCDC91* that suggest AIRE might capture metabolic risk as an additional mechanism. These findings suggest AIRE-predicted survival is a biomarker of overall health, including biological age and the presence of clinical and subclinical disease.

There are limitations to the accuracy and granularity of ICD diagnostic codes that are used in this study to ascertain disease status. In particular, ventricular arrhythmias as reported by ICD codes are not necessarily sustained or haemodynamically significant and, therefore, patients predicted to have these events would not necessarily benefit from an implantable cardioverter defibrillator. There are drawbacks to each of the explainability methods presented in this work, and there is no perfect explainability technique for deep learning.[37] Although deep learning models are not fully explainable, our aim is that, by providing multiple complementary explainability analyses, the reader can gain some insight into the ECG morphologies associated with high risk. The ultimate goal of developing AIRE is to use it to guide treatment decisions for patients by integrating it fully into an electronic health records system or medical device, although the work presented is only one of many steps towards that goal. Other future steps required would include testing AIRE prospectively in clinical studies and obtaining the appropriate regulatory approvals.

In conclusion, we describe the AIRE platform, an actionable, explainable, and biologically plausible AI-ECG risk estimation platform that has the potential for use worldwide across a wide range of clinical contexts, including primary and secondary care, for short-term and long-term risk prediction at population and disease-specific levels.

Contributors

Declaration of interests

Data sharing

Acknowledgments

References

1   Sau A, Ibrahim S, Ahmed A, et al. Artificial intelligence-enabled electrocardiogram to distinguish cavotricuspid isthmus dependence from other atrial tachycardia mechanisms. *Eur Heart J Digit Health* 2022; **3**: 405–14.

2   Sau A, Ibrahim S, Kramer DB, et al. Artificial intelligence-enabled electrocardiogram to distinguish atrioventricular re-entrant tachycardia from atrioventricular nodal re-entrant tachycardia. *Cardiovasc Digit Health J* 2023; **4**: 60–67.

3   Attia ZI, Kapa S, Lopez-Jimenez F, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. *Nat Med* 2019; **25**: 70–74.

4   Sau A, Ribeiro AH, McGurk KA, et al. Prognostic significance and associations of neural network-derived electrocardiographic features. *Circ Cardiovasc Qual Outcomes* (in press).

5   Raghunath S, Ulloa Cerna AE, Jing L, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nat Med* 2020; **26**: 886–91.

6   Sun W, Kalmady SV, Sepehrvand N, et al. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. *NPJ Digit Med* 2023; **6**: 21.

7   Hughes JW, Tooley J, Torres Soto J, et al. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. *NPJ Digit Med* 2023; **6:** 169.

8   Cardoso CS, Sabino EC, Oliveira CDL, et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. *BMJ Open* 2016; **6:** 5e011181.

9   Lima EM, Ribeiro AH, Paixão GMM, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nat Commun* 2021; **12:** 5117.

10   Schmidt MI, Duncan BB, Mill JG, et al. Cohort profile: longitudinal study of adult health (ELSA-Brasil). *Int J Epidemiol* 2015; **44:** 68–75.

11   Sudlow C, Gallacher J, Allen N, 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:** 3e1001779.

12   Ribeiro AH, Ribeiro MH, Paixao GMM, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020; **11:** 1760.

13   Gensheimer MF, Narasimhan B. A scalable discrete-time survival model for neural networks. *PeerJ* 2019; **7:** e6257.

14   Suchard MA, Schuemie MJ, Krumholz HM, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *Lancet* 2019; **394:** 1816–26.

15   Stabenau HF, Waks JW. BRAVEHEART: open-source software for automated electrocardiographic and vectorcardiographic analysis. *Comput Methods Programs Biomed* 2023; **242:** 107798.

16   Kubota Y. tf-keras-vis. 2020. https://github.com/keisen/tf-keras-vis (accessed July 18, 2024).

17   Meyer HV, Dawes TJW, Serrani M, et al. Genetic and functional insights into the fractal structure of the heart. *Nature* 2020; **584:** 589–94.

18   Jonsson BA, Bjornsdottir G, Thorgeirsson TE, et al. Brain age prediction using deep learning uncovers associated sequence variants. *Nat Commun* 2019; **10:** 5409.

19   Agarwal SK, Chambless LE, Ballantyne CM, et al. Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) study. *Circ Heart Fail* 2012; **5:** 422–29.

20   van der Harst P, van Setten J, Verweij N, et al. 52 genetic loci influencing myocardial mass. *J Am Coll Cardiol* 2016; **68:** 1435–48.

21   Young WJ, Haessler J, Benjamins JW, et al. Genetic architecture of spatial electrical biomarkers for cardiac arrhythmia and relationship with cardiovascular disease. *Nat Commun* 2023; **14:** 1411.

22   Jespersen T, Grunnet M, Olesen S-P. The KCNQ1 potassium channel: from gene to physiological function. *Physiology (Bethesda)* 2005; **20:** 408–16.

23   Verweij N, Benjamins JW, Morley MP, et al. The genetic makeup of the electrocardiogram. *Cell Syst* 2020; **11:** 229–38.e5.

24   Zhu X, Zhu L, Wang H, Cooper RS, Chakravarti A. Genome-wide pleiotropy analysis identifies novel blood pressure variants and improves its polygenic risk scores. *Genet Epidemiol* 2022; **46:** 105–21.

25   Sakaue S, Kanai M, Tanigawa Y, et al. A cross-population atlas of genetic associations for 220 human phenotypes. *Nat Genet* 2021; **53:** 1415–24.

26   Christakoudi S, Evangelou E, Riboli E, Tsilidis KK. GWAS of allometric body-shape indices in UK Biobank identifies loci suggesting associations with morphogenesis, organogenesis, adrenal cell renewal and cancer. *Sci Rep* 2021; **11:** 10688.

27   Libiseller-Egger J, Phelan JE, Attia ZI, et al. Deep learning-derived cardiovascular age shares a genetic basis with other cardiac phenotypes. *Sci Rep* 2022; **12:** 22625.

28   Graff M, Scott RA, Justice AE, et al. Genome-wide physical activity interactions in adiposity - a meta-analysis of 200 452 adults. *PLoS Genet* 2017; **13:** 4e1006528.

29   Altman DG, Bland JM. Diagnostic tests 2: predictive values. *BMJ* 1994; **309:** 102.

30   Khurshid S, Friedman S, Reeder C, et al. ECG-based deep learning and clinical risk factors to predict atrial fibrillation. *Circulation* 2022; **145:** 122–33.

31   Visseren FLJ, Mach F, Smulders YM, et al. 2021 ESC guidelines on cardiovascular disease prevention in clinical practice. *Eur Heart J* 2021; **42:** 3227–337.

32   Bachtiger P, Kelshiker MA, Petri CF, et al. Survival and health economic outcomes in heart failure diagnosed at hospital admission versus community settings: a propensity-matched analysis. *BMJ Health Care Inform* 2023; **30:** 1.

33   Bachtiger P, Petri CF, Scott FE, et al. Point-of-care screening for heart failure with reduced ejection fraction using artificial intelligence during ECG-enabled stethoscope examination in London, UK: a prospective, observational, multicentre study. *Lancet Digit Health* 2022; **4:** e117–25.

34   Istolahti T, Lyytikäinen LP, Huhtala H, et al. The prognostic significance of T-wave inversion according to ECG lead group during long-term follow-up in the general population. *Ann Noninvasive Electrocardiol* 2021; **26:** 1e12799.

35   Schröder LC, Holkeri A, Eranti A, et al. Poor R-wave progression as a predictor of sudden cardiac death in the general population and subjects with coronary artery disease. *Heart Rhythm* 2022; **19:** 952–59.

36   Imanishi R, Seto S, Ichimaru S, Nakashima E, Yano K, Akahoshi M. Prognostic significance of incident complete left bundle branch block observed over a 40-year period. *Am J Cardiol* 2006; **98:** 644–48.

37   Ghassemi M, Oakden-Rayner L, Beam AL. The false hope of current approaches to explainable artificial intelligence in health care. *Lancet Digit Health* 2021; **3:** e745–50.