Artificial intelligence-enhanced electrocardiography for cardiovascular risk prediction

A thesis submitted to Imperial College London for the degree of

Doctor of Philosophy

Dr Arunashis Sau

July 2024

National Heart and Lung Institute

Imperial College London

Statement of Originality

In accordance with Imperial College London regulations, I declare that all the work within this thesis is my own, unless specifically stated otherwise in the text.

Dr Arunashis Sau

July 2024

Copyright Declaration

The copyright of this thesis rests with the author. Unless otherwise indicated, its contents are licensed under a Creative Commons Attribution-Non Commercial 4.0 International Licence (CC BY-NC). Under this licence, you may copy and redistribute the material in any medium or format. You may also create and distribute modified versions of the work. This is on the condition that: you credit the author and do not use it, or any derivative works, for a commercial purpose. When reusing or sharing this work, ensure you make the licence terms clear to others by naming the licence and linking to the licence text. Where a work has been adapted, you should indicate that the work has been changed and describe those changes. Please seek permission from the copyright holder for uses of this work that are not included in this licence or permitted under UK Copyright Law.

To my family

Acknowledgements

I would like to thank several individuals who have helped me in completing this thesis. First, my supervisors Dr Fu Siong Ng, Prof Nicholas S Peters and Prof Danilo Mandic who have provided endless guidance and support. Foremost, Fu Siong, who has been an extraordinary mentor over the past five years, teaching me invaluable lessons in research, clinical electrophysiology, leadership and more. His willingness to support my ideas and passion projects has been incredibly meaningful to me and has allowed the work in this thesis to come to fruition – thank you for taking a chance. I am immensely grateful for the British Heart Foundation for their generous funding, without which this research could not have been performed.

I am grateful to several others for their specific contributions to my work. I would like to thank Dr Libor Pastika for help in data wrangling and being a much-needed sounding board when solving challenging problems. I would like to thank Dr Nikesh Bajaj, Dr Konstantinos Patlatzoglou and Dr Ewa Sieliwonczyk for their help in developing some of the explainability methods described in this thesis. Specifically, Nikesh helped develop the explainability approach in Chapter 3, while Konstantinos and Ewa helped develop the Variational Autoencoder and median based explainability approaches.

I would like to thank Dr Kathryn McGurk, Dr Ewa Sieliwonczyk and Prof James S Ware for helping with the Genome-wide association studies described in this thesis.

I am immensely grateful for my collaborators around the world who have entrusted me with vast amounts of data, without which none of this would be possible – I hope you have gained a fraction of what I have from our interactions. I am particularly grateful to Dr Antonio Ribeiro and Prof Antonio Ribeiro for their collaborative spirit that helped start this all and to Dr Jonathan Waks and Dr Daniel Kramer for their incredible support in creating the unparalleled BIDMC dataset for my work, and for the regular thought-provoking discussions.

Finally, I would like to thank my family for their tireless support. Your belief in my aspirations and ability has been a constant source of strength and motivation that I deeply value.

Abstract

The electrocardiogram (ECG) is an ubiquitous and low-cost investigation that records the electrical activity of the heart non-invasively. Through the advent of artificial intelligence (AI), previously unheralded insights are being discovered through the application of AI to the ECG (AI-ECG). In this thesis I explore the application of AI-ECG for cardiovascular risk prediction. Additionally, I explore the application of detailed phenotyping and genotyping methods to AI-ECG derived traits, including Phenome- and Genome-wide association studies. First, using an unsupervised machine learning approach, I demonstrated how neural-network derived ECG features derived from a simple rhythm identification task could be used to identify phenogroups with prognostic significance. Second, I developed a novel supervised machine learning based mortality prediction model, using a discrete-time survival approach. This mortality prediction model forms the basis of an AI-ECG risk estimation platform (AIRE), capable of predicting not only mortality, but also the future occurrence of important cardiovascular diseases: atherosclerotic cardiovascular disease, heart failure, ventricular arrhythmia, complete heart block and atrial fibrillation. I demonstrate the exceptional performance of AIRE and compare it to existing clinical, imaging, biochemical and genetic risk factors. Finally, I explore how AI-ECG traits can be thought of as digital biomarkers, and describe the finding that AI-ECG sex prediction can be used to develop a continuous sex score that can capture the continuum of risk and sex phenotypes within each sex. I found

sex score discordance was a novel AI-ECG biomarker capable of identifying females with disproportionately elevated cardiovascular risk.

Table of Contents

Statement of Originality 2
Copyright Declaration 3
Acknowledgements 5
Abstract 7
List of Figures 14
List of Tables 17
List of Abbreviations 18
Publications 22
Awards 28
Presentations 29
Chapter 1 – Introduction ***32***
The ECG: unchanged for over 80 years 32
Artificial intelligence 35
What is artificial intelligence 35
Traditional programming vs machine learning 36
Traditional AI vs Generative AI 38
Supervised vs. Unsupervised machine learning 40
Traditional machine learning vs. Deep learning 41
Explainability 42
Artificial intelligence-enabled electrocardiography (AI-ECG) 44
Deep learning to improve ECG interpretation 45
Deep learning to harness 'hidden' insights from the ECG 45
Disease phenotyping 47
Clinical translation 48
Cardiovascular risk prediction - state of the art 49
The burden of cardiovascular disease 49
Why predict risk of cardiovascular diseases? 50
Mortality prediction 52
Atrial fibrillation 52
AI-ECG for cardiovascular risk prediction 54
Why AI-ECG? 54
Cardiovascular disease 54
Arrhythmia 54
Mortality 55
Biological aging 56
Biological exploration 56
Digital biomarkers and OMICs 56
Phenome-wide association 57
Genome-wide association study 58
Scope of thesis 58
Hypotheses 58
Chapter 2 - Methods ***60***
Introduction 60

The digital electrocardiogram **60**
ECG pre-processing **61**
Cohorts **62**
Ethical approvals **65**
Normal ECG definition **66**
AI-ECG models **66**
Anatomy of a neural network 66
Training a neural network 70
AI-ECG model architectures 75
Classification and Regression 79
Deep learning survival analysis 80
Training AI-ECG models 82
Model optimisation 83
Transfer learning 84
BIDMC cohort data splits 84
Survival and statistical analyses **86**
Explainability **87**
Saliency mapping 87
Median beats 88
Inherently explainable deep learning architectures 89
Visualisation of average beats 92
Correlation with standard ECG parameters 93
Biological exploration **93**
Phenome-wide association study (PheWAS) 93
I used two PheWAS approaches in this thesis, I have termed them continuous PheWAS and Disease PheWAS, these are described in detail below. 93
Continuous PheWAS 93
Disease PheWAS 94
Genome-wide association study 94
Chapter 3 - AI-ECG phenogroups predict adverse outcomes ***96***
Introduction **96**
Methods **98**
Neural network-derived ECG features from CODE dataset 98
Unsupervised machine learning for clustering into phenogroups 98
Survival analysis in derivation and external testing cohorts 99
Phenome-wide association study 100
GWAS 100
Model explainability 101
Results **102**
Derivation cohort 103
External testing 113
Further analysis of Phenogroup C 119
Phenome-wide association study 120
Genome-wide association study 123
Model explainability 125
Discussion **127**
Hypothesis driven vs. data driven research 127
Prognostic significance of ECG phenogroups 128
Disease PheWAS 129
UK Biobank PheWAS 129
Genome wide-association study 130

Comparison to previous machine learning approaches for ECGs 131
Potential clinical applications of NN-derived ECG features 132
Limitations 133
Conclusion 134

Chapter 4 - Supervised AI-ECG for mortality prediction 135

Introduction 135
From phenotypes to personalised predictions 135
Actionable predictions 136
Explainability and biological plausibility 136
AI-ECG risk estimation platform (AIRE) 137

Methods 137
Cohorts 139
AI-ECG risk estimation platform development 139
Fine-tuned sub-models 140
Comparison of AIRE with other models 140
Survival and statistical analyses 141
Diagnostic and imaging data 142
Explainability 143
PheWAS 144
GWAS 144

Results 145
AIRE accurately predicts mortality across diverse timepoints 145
AIRE on normal ECGs 156
AIRE is superior to traditional risk factors for mortality prediction 156
AIRE performance in high-risk disease groups 158
..... 160
Single-lead ECG model performance 161
AIRE predicts mortality in transnational external datasets 162
AIRE outperforms AI-ECG models of biological aging 165
Explainable ECG morphology associates with adverse prognosis 166
Biological plausibility: Genetic associations of AIRE-predicted survival 168
Biological plausibility: Phenotypic associations of AIRE-predicted survival 169

Discussion 174
Importance of time-to-event for actionability 174
Actionable risk prediction in high-risk disease groups 175
Single-lead ECG applications 176
Explainability and biological plausibility 176
AIRE-predicted survival is a biomarker of overall health 177
Reduced performance in volunteer populations 178
Limitations 178
Conclusion 179

Chapter 5 - AI-ECG for prediction of diverse cardiovascular outcomes 180

Introduction 180
Expanding AIRE 180
Current risk prediction methods 181
Aims 181

Methods 182
Study overview 182
Cohorts 184
Disease definitions 184
Model fine-tuning 185
Comparison with previous AI-ECG model 185
Comparator clinical risk scores 186
Polygenic risk score 186

Survival analyses 186
Results 187
Atherosclerotic cardiovascular disease 187
Ventricular arrhythmia 188
Heart failure 191
Complete heart block 192
Atrial fibrillation 194
Cross-Task Model Performance Evaluation 198
Discussion 201
Future ASCVD prediction 201
Future VA prediction 202
Future HF prediction 202
Future CHB prediction 203
Future AF prediction 203
Limitations 205
Conclusion 206
Chapter 6 - AI-ECG as a biomarker of ECG sex phenotype 207
Introduction 207
Methods 209
Model development 209
Sex discordance score 209
Explainability 210
Normal ECG and healthy subject definition 210
Survival and statistical analyses 211
Phenome-wide association studies 211
Genome-wide association study 212
Results 213
AI-ECG accurately classified sex from the ECG 213
Explainable ECG morphologies are associated with AI-ECG sex predictions 216
Sex discordance score reveals female specific cardiovascular risk 219
Sex discordance score is associated with differential phenotypes in males and females 224
Sex discordance score is associated with greater cardiovascular disease in females 228
Sex discordance score has genetic associations 233
Discussion 235
AI-ECG predictions as a biomarker 235
AI-ECG is highly accurate at classifying sex, particularly in lower-risk populations 235
Sex discordance score is associated with differential phenotypes in males and females in lower-risk populations 236
Biological exploration provides insights into mechanisms of adverse CV prognosis conferred by sex misclassification in females but not males 237
AI-ECG may help overcome sex biases in identification and treatment of CV disease 238
Limitations 239
Conclusion 239
Chapter 7 - Conclusions 241
Summary of findings 241
Data-driven research 242
Clinical relevance 243
AI-ECG predictions as a biomarker 244
Limitations 245
Future directions 246

Expanding AIRE 246
Applying AI-ECG to images of ECGs 247
Integration of clinical, imaging, biochemical and longitudinal data 247
Randomised trials of AI-ECG versus usual care 248
Conclusions 248

List of Figures

Figure 1-1 - A timeline of the major milestones in the development of the ECG. 34

Figure 1-2 Schematic depicting the main difference between traditional programming and machine learning 37

Figure 1-3 Diagrammatic summary of artificial intelligence approaches 40

Figure 2-1 Schematic diagram of a convolutional neural network 69

Figure 2-2 Flow of data in a machine learning study 71

Figure 2-3 Example typical training and tuning loss curves 74

Figure 2-4 Schematic of the ResNet architecture used in this thesis.....76

Figure 2-5 Variational autoencoder schematic diagram.....90

Figure 2-6 Variational Autoencoder reconstruction performance.....92

Figure 3-1 Data analysis pipeline 102

Figure 3-2 Survival analysis in the derivation dataset 112

Figure 3-3 Survival analysis in the five external testing datasets.....115

Figure 3-4 Survival analysis in the five external testing datasets – sensitivity analysis 118

Figure 3-5 Phenome-wide association study 122

Figure 3-6 UK Biobank ECG associations 123

Figure 3-7 Genome-wide association study 124

Figure 3-8 Genome-wide association study secondary analysis 125

Figure 3-9 Model explainability 126

Figure 4-1 AIRE platform overview 138

Figure 4-2 Example subject-specific survival predictions.....149

Figure 4-3 Mortality prediction Kaplan-Meier curves.....150
Figure 4-4 Time-dependent receiver operating characteristic curve of AIRE predictions over 10 years.....152
Figure 4-5 Mortality prediction performance.....155
Figure 4-6 Mortality prediction in subgroups of severe aortic stenosis and primary pulmonary hypertension.....159
Figure 4-7 Mortality prediction sensitivity analyses.....160
Figure 4-8 Single lead mortality prediction.....161
Figure 4-9 Mortality prediction Kaplan-Meier curves – external testing.....164
Figure 4-10 AIRE model explainability - BIDMC Test set.....167
Figure 4-11 Phenome and Genome-wide association studies.....172
Figure 5-1 Extended AIRE platform overview.....183
Figure 5-2 AIRE-ASCVD performance.....188
Figure 5-3 AIRE-VA performance.....190
Figure 5-4 AIRE-HF performance.....192
Figure 5-5 AIRE-CHB performance.....193
Figure 5-6 AIRE-AF Kaplan-Meier curves.....195
Figure 5-7 AIRE-AF Performance.....196
Figure 5-8 AIRE-AF UK Biobank.....198
Figure 5-9 Cross-Task Model Performance Heatmap.....199
Figure 6-1 AI-ECG sex prediction performance metrics.....215
Figure 6-2 ECG morphologies associated with AI-ECG sex prediction.....218
Figure 6-3 AI-ECG sex prediction correlations with ECG parameters.....219
Figure 6-4 Sex discordance score survival analysis in unselected BIDMC test set.220

Figure 6-5 Association of sex discordance score with all-cause, cardiovascular, and non-cardiovascular mortality 223
Figure 6-6 Association of sex discordance score with echocardiographic parameters 225
Figure 6-7 Association of sex discordance score with cardiac MRI parameters.....226
Figure 6-8 UK Biobank Phenome-wide association study (PheWAS) 228
Figure 6-9 Disease phenome-wide association study.....230
Figure 6-10 Genome-wide association study of sex discordance score 234

List of Tables

Table 2-1 77

Table 3-1 104

Table 3-2 108

Table 4-1 146

Table 4-2 151

Table 4-3 153

Table 4-4 157

Table 4-5 163

Table 4-6 170

Table 4-7 171

Table 6-1 214

Table 6-2 231

Table 6-3 232

List of Abbreviations

AF	Atrial fibrillation
AGI	Artificial general intelligence
AHRE	Atrial high-rate episode
AI	Artificial intelligence
AI-ECG	Artificial intelligence-enabled electrocardiography
AIRE	AI-ECG risk estimation platform
ARIC-HF	Atherosclerotic Risk in Communities - Heart Failure
ASCVD	Atherosclerotic cardiovascular disease
AUROC	Area under the receiver operating characteristics curve
AV	Atrioventricular
BIDMC	Beth Israel Deaconess Medical Center
BMC	Bone mineral content
BMI	Body mass index
BRAVEHEART	Beth Israel Analysis of Vectors of the Heart
BSA	Body surface area
CHB	Complete heart block
CIED	Cardiac implantable electronic device
CNN	Convolutional neural network
CODE	Clinical Outcomes in Digital Electrocardiography
COVID-19	Coronavirus disease 2019
CRT	Cardiac resynchronisation therapy

CV	Cardiovascular
CVD	Cardiovascular disease
DCM	Dilated cardiomyopathy
DCM	Dilated cardiomyopathy
ECG	Electrocardiogram
EDV	end-diastolic volume
ELSA-Brasil	Longitudinal Study of Adult Health
eQTL	expression quantitative trait loci
ESV	end-systolic volume
FRS	Framingham Risk Score
Grad-CAM	Gradient-weighted Class Activation Mapping
GWAS	Genome-wide association studies
HDF5	Hierarchical Data Format 5
HR	Hazard ratio
ICD	Implantable cardioverter defibrillator
ICD	International Classification of Diseases
ILR	Implantable loop recorder
IMT	intima-media thickness
KL	Kullback-Leibler
LAEF	LA ejection fraction
LAEF	LA ejection fraction
LBBB	Left bundle block
LLM	Large language model

LLM	Large language model
LVEF	left ventricular ejection fraction
LVSD	Left ventricular systolic dysfunction
MACE	Major adverse cardiovascular event
MI	myocardial infarction
ML	Machine learning
MLP	Multilayer perceptron
MRI	Magnetic resonance imaging
MV	mitral valve
NN	Neural-network
NRI	Net reclassification index
OR	Odds ratio
PCA	Principle component analysis
PCE	Pooled Cohort Equation
PheWAS	Phenome-wide association studies
PRS	Polygenic risk scores
PVC	Premature ventricular contraction
RAM	Random-access memory
RBBB	Right bundle branch block
ReLu	Rectified linear unit
ResNet	Residual Neural Network
RNN	Recurrent neural network
RVEF	Right ventricular ejection fraction

SaMi-Trop	São Paulo-Minas Gerais Tropical Medicine Research Center
SCD	Sudden cardiac death
SCD	Sudden cardiac death
SEER	Stanford Estimator of ECG Risk
SNP	Single-nucleotide polymorphism
SR	Sinus rhythm
SV	stroke volume
TNMG	Telehealth Network of Minas Gerais
UK	United Kingdom
UKB	UK Biobank
VA	Ventricular arrhythmia
VAE	Variational autoencoder

Publications

For period of PhD research October 2021 to July 2024

Five additional first author manuscripts currently undergoing peer review

First Author Original Article Publications

---

* Joint first authorship

1. **A Sau**, … FS Ng. **Artificial intelligence–enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform.** Lancet Digital Health (accepted, in press)

2. L Pastika*, **A Sau*, …,** FS Ng. **Artificial Intelligence-enhanced Electrocardiography Derived Body Mass Index as a Predictor of Future Cardiometabolic Disease.** npj Digital Medicine. 7, 167 (2024). https://doi.org/10.1038/s41746-024-01170-0

3. **A Sau*** & S Kapadia* … FS Ng. **Temporal Trends and Lesion Sets for Persistent Atrial Fibrillation Ablation: A Meta-Analysis With Trial Sequential Analysis and Meta-Regression**

Circulation Arrhythmia Electrophysiology. 2023 Sep;16(9):e011861. doi: 10.1161/CIRCEP.123.011861

4. **A Sau*** & A Ahmed* … FS Ng. **Machine learning-derived cycle length variability metrics predict spontaneously terminating ventricular tachycardia in implantable cardioverter defibrillator recipients** European Heart Journal - Digital Health, 2023, ztad064, doi.org/10.1093/ehjdh/ztad064

5. X Shi* & **A Sau***, … FS Ng. **Information theory-based direct causality measure to assess cardiac fibrillation dynamics.** Journal of The Royal Society Interface. 2023 Oct;20(207):20230443. doi: 10.1098/rsif.2023.0443

6. **A Sau**, … FS Ng. **Artificial intelligence–enabled electrocardiogram to distinguish atrioventricular re-entrant tachycardia from atrioventricular nodal re-entrant tachycardia**. Cardiovascular Digital Health Journal. 2023 Jan 31.

7. **A Sau**, … , FS Ng. **Artificial intelligence-enabled electrocardiogram to distinguish cavotricuspid isthmus dependence from other atrial tachycardia mechanisms.** European Heart Journal-Digital Health. 2022 Sept 3;3 https://doi.org/10.1093/ehjdh/ztac042

8. **A Sau**, … F S Ng. **Prognostic Significance of Ventricular Arrhythmias in 13 444 Patients With Acute Coronary Syndrome: A Retrospective Cohort Study Based on Routine Clinical Data (NIHR Health Informatics Collaborative VA-ACS Study).** J Am Heart Assoc. 2022 Mar 15;11(6):e024260. doi: 10.1161/JAHA.121.024260. PMID: 35258317

Other publications

Original Article

1. JW Waks, … **A Sau, …** AE Buxton. **The Spatial Ventricular Gradient is Associated with Inducibility of Ventricular Arrhythmias During Electrophysiology Study. Heart Rhythm**

   Invited Review

2. **A Sau**, FS Ng. **The emerging role of artificial intelligence enabled electrocardiograms in healthcare**.

   BMJ Medicine 2023 Jul 31;2(1):e000193. doi: 10.1136/bmjmed-2022-000193

   Original Article

3. HF Stabenau, **A Sau** …, JW Waks. **Limits of the spatial ventricular gradient and QRST angles in patients with normal electrocardiograms and no known cardiovascular disease stratified by age, sex, and race**

   Journal of Cardiovascular Electrophysiology. 2023 Nov;34(11):2305-2315. doi: 10.1111/jce.16062

   Invited Editorial

4. **A Sau**, … S Hudson. **British Cardiovascular Society/British Heart Foundation/British Atherosclerosis Society/British Society for Cardiovascular Research Young Investigator Award 2023**

   Heart. 2023 Sep 28;109(20):1571-1573. doi: 10.1136/heartjnl-2023-323063

   Invited Editorial

5. **A Sau**, L Pastika, FS Ng. **Atrial fibrillation phenotypes: the route to personalised care?**

   Heart. 2023 Nov 10;109(23):1727-1728. doi: 10.1136/heartjnl-2023-322812

Letter

6. **A Sau**, FS Ng. **Response to letter by Saumarez et al. entitled 'Regarding the editorial by Sau and Ng. "Hypertrophic cardiomyopathy risk stratification based on clinical or dynamic electrophysiological features: two sides of the same coin"'**.

   Europace. 2023 Jul 4;25(7):euad174. doi: 10.1093/europace/euad174

   Invited Editorial

7. **A Sau**, FS Ng. **Hypertrophic cardiomyopathy risk stratification based on clinical or dynamic electrophysiological features: two sides of the same coin**, EP Europace, 2023; euad072

   Original article

8. H Wu, KHK Patel, … **A Sau**, … FS Ng. **A fully-automated paper ECG digitisation algorithm using deep learning**. Sci Rep 12, 20963 (2022).

9. *Invited narrative review*

   KHK Patel, RK Reddy, **A Sau**, … FS Ng. **Obesity as a risk factor for cardiac arrhythmias.** BMJ Medicine. 2022;1:e000308. doi: 10.1136/bmjmed-2022-000308

10. *Original article*

    X Li, X Shi, B S Handa, **A Sau**, … FS Ng. **Classification of Fibrillation Organisation Using Electrocardiograms to Guide Mechanism-Directed Treatments.** Front Physiol. 2021 Nov 11;12:712454. doi: 10.3389/fphys.2021.712454. PMID: 34858198

11. *Original article*

    M Falkenberg, … **A Sau**, … K Christensen. **Identifying locations susceptible to micro-anatomical reentry using a spatial network**

representation of atrial fibre maps. PLoS One. 2022 Jun 23;17(6):e0267166. doi: 10.1371/journal.pone.0267166

Awards

2024

Finalist: Heart Rhythm Society, Young Investigator Award

First Prize: Royal Society of Medicine President's Prize

First Prize: European Heart Rhythm Association – e-Cardiology Award

First Prize: British Heart Rhythm Association – Young Investigator Award

2023

Finalist: American Heart Association Elizabeth Barrett-Connor Early Career Research Award

First Prize: Heart Rhythm Congress Basic Science Young Investigator Award

Third Prize: Royal Society of Medicine President's Prize

First Prize: British Cardiovascular Society/British Heart Foundation 2023 Young Investigator Award

Imperial Clinical Academic Training Office - Oral Presentation Award

2022

First Prize: European Heart Rhythm Association - e-Cardiology award

Presentations

1. Artificial intelligence-enabled electrocardiography for hypertension diagnosis, prediction and risk stratification
   (Finalist, Royal Society of Medicine Cardiology President's prize 2024, London)

2. Artificial intelligence–enabled electrocardiography for mortality and ventricular arrhythmia prediction
   (*First Prize, e-Cardiology award, Oral presentation, European Heart Rhythm Association 2024, Berlin*)

3. Artificial intelligence–enabled electrocardiography for mortality and ventricular arrhythmia risk estimation: An actionable, explainable and biologically plausible platform
   (*First prize, Young Investigator Award, Oral presentation*, British *Heart Rhythm Society 2024, Manchester)*

4. Neural Network-Derived Electrocardiographic Features Predict Future Cardiovascular Disease and All-Cause Mortality
   (*Finalist, Elizabeth Barrett-Connor Early Career Research Award, Oral presentation, American Heart Association 2023, Philadelphia*)

5. Neural network-derived electrocardiographic features predict all-cause mortality and have important phenotypic and genotypic associations

(Third prize, Royal Society of Medicine Cardiology President's prize 2023, London)

6. Neural Network-derived Electrocardiographic Features Predict Mortality And Future Malignant Arrhythmia And Conduction Disease
   (*First prize, Young Investigator Award, Oral presentation*, *Heart Rhythm Congress 2023, Birmingham)*

7. Exploring The Prognostic Significance And Important Phenotypic And Genotypic Associations Of Neural Network-derived Electrocardiographic Features
   (*First prize, Young Investigator Award, Oral presentation, British Cardiovascular Society 2023, Manchester)*

8. Unsupervised Machine Learning Identifies Prognostically Significant Phenogroups from Neural Network-Derived ECG features
   (*Oral presentation, Heart Rhythm Society 2023, New Orleans, USA)*

9. Temporal trends and lesion sets for persistent atrial fibrillation ablation – a meta-analysis with trial sequential analysis and meta-regression
   (*Poster presentation, Heart Rhythm Society 2023, New Orleans, USA)*

10. Artificial intelligence-enabled electrocardiogram to distinguish atrioventricular re-entrant tachycardia from atrioventricular nodal re-entrant tachycardia
    (Oral presentation, Heart Rhythm Congress 2022, Birmingham)

11. Artificial intelligence-enabled electrocardiogram to distinguish cavotricuspid isthmus dependence from other atrial tachycardia mechanisms *(Oral presentation, Heart Rhythm Congress 2022, Birmingham)*

12. Classification of organised atrial arrhythmias using explainable artificial intelligence
(*First Prize, e-Cardiology award, Oral presentation, European Heart Rhythm Association 2022, Copenhagen*)

Chapter 1 – Introduction

Parts of the work in this chapter of the thesis have appeared in publication

A Sau & FS Ng. The emerging role of artificial intelligence enabled electrocardiograms in healthcare.

BMJ Medicine 2023 Jul 31;2(1):e000193. doi: 10.1136/bmjmed-2022-000193

The ECG: unchanged for over 80 years

The electrocardiogram (ECG) was first performed in a human by Sir Augustus Waller in 1887, a British Physiologist of St Mary's Medical School, London (1). He showed electrical activity preceding ventricular contraction. Dr. Willem Einthoven, a Dutch physiologist, built on Waller's work, and described five deflections which he termed ABCDE, this soon changed to the nomenclature we use today of PQRST, the change inspired by conventions set by Descartes (2). The early 20th century saw the first clinical uses of the three lead ECG, including the first description of atrial fibrillation by Sir Thomas Lewis (University College Hospital, London) in 1909 (3).

Dr. Frank N. Wilson (University of Michigan), in 1934 first described the eponymous 'central terminal' (4) which allowed the invention of the unipolar precordial leads. Finally in 1942 Dr. Emanuel Goldberger (Lincoln Hospital, New York) described the augmented unipolar leads, concluding the major developments in the ECG. The recommendations for the standardised 12-lead ECG were published in 1954 by the

American Heart Association (5). **Figure 1-1** depicts the major milestones in the development of the ECG.

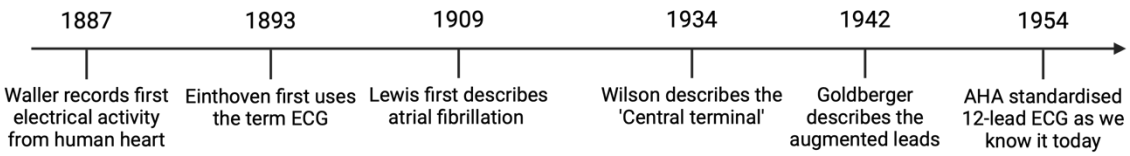


Figure 1-1 - A timeline of the major milestones in the development of the ECG. Adapted from AlGhatrif & Lindsay 2012 (1). AHA: American Heart Association

Since the invention of the ECG, millions of ECG recordings have been taken and used to guide clinical care, including detecting arrhythmias, diagnosing heart attacks and identifying inherited cardiac conditions. Traditionally, clinicians are taught to interpret the ECG by understanding its constituent waves, P, QRS and T, intervals and segments. Basic rule based computer algorithms have existed for decades, but are rarely useful in clinical practice (6), due to their limited performance. Additionally as they are specifically programmed around recognised clinical and physiological parameters, they do not take advantage of the entirety of information available to us in the ECG. As clinicians reading ECGs, we can learn over the course of a career, from seeing thousands of ECGs, the nuances of ECG interpretation and correlation with clinical diagnoses. However, even the most experienced ECG enthusiast is unlikely to be able to see and assimilate more than a few tens of thousands of ECGs in their lifetime.

Artificial intelligence

What is artificial intelligence

Artificial intelligence (AI) as a field has experienced exponential growth and interest in the last 10 years, and with good reason. In the field of healthcare, clinicians are promised previously un-heralded insights, diagnostics, and predictive power at their fingertips. AI additionally holds promise in streamlining and reducing administrative burden for clinicians (7).

AI and machine learning are terms often used interchangeably. AI is a branch of computer science dealing with machines that may have the cognitive capabilities that would ordinarily be expected of humans. Machine learning is a subbranch of AI and deals with the ability of machines to learn from data, rather than explicitly being programmed to accomplish a certain task.

During the course of this PhD, the release of ChatGPT has catapulted AI into the public domain (8), for the first time opening up an AI powered chatbot to the public. With it, the public interest in AI has skyrocketed, and alongside the interest, significant concerns have been raised. In particular, some have raised concerns around artificial general intelligence (AGI). The concept of AGI is a (currently) hypothetical type of AI that would have the ability to accomplish any intellectual task that a human can perform (9), and more. Some consider AGI to pose an existential threat to humanity (10). Although some argue that the promises of AI are far from the reality, my personal belief is that AI will radically change society - for better and worse. The goal of this thesis is to describe one way AI might change medicine for the better.

Traditional programming vs machine learning

The difference between traditional programming and machine learning is depicted in **Figure 1-2**. Traditional programming has been used for decades and is usually rule based and deterministic. Deterministic models/programs produce the same output every time for a given input and program. Stochastic models on the other hand incorporate randomness and therefore a given input can result in any number of varying outputs with each run of the model. A simple example of traditional programming can be found in computerised ECG algorithms where after measurement of QRS duration, a duration of more than 120ms is deemed abnormal and consistent with some form of intraventricular conduction delay. It has subsequently been observed that prolonged QRS duration is associated with increased risk of death (11).

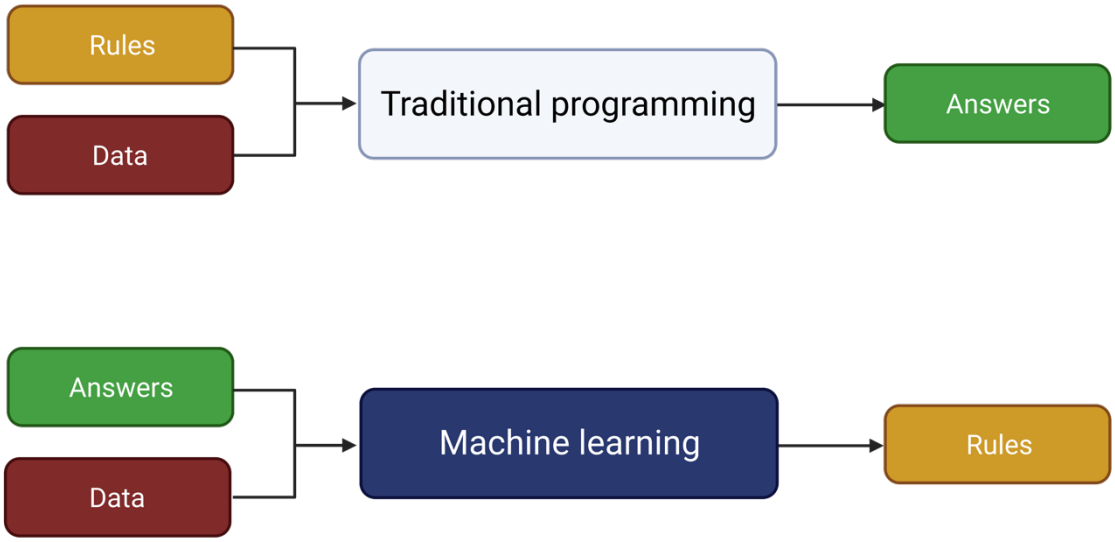


Figure 1-2 Schematic depicting the main difference between traditional programming and machine learning

Machine learning on the other hand inverts this paradigm. Data is matched with labels (at least in the case of supervised machine learning, unsupervised machine learning is discussed below), a training architecture is defined, and the 'rules' are learnt by the model, by identifying what features of the data are linked to the labels. Regression models in conventional statistics have followed a similar paradigm for decades. The distinction between statistics and machine learning is a complex topic. Machine learning is built on a statistical framework, but machine learning in general has an enhanced capability to identify complex, non-linear, relationships compared to conventional statistics.

Traditional AI vs Generative AI

ChatGPT and other large language models (LLMs), are now what most of the general public would regard as AI. LLMs are a type of generative AI, a type of AI that can create something new. The training data informs the distribution on which the model bases the new outputs. Traditional AI on the other hand is used to analyse data and make predictions. In general, traditional AI focuses on performing a singular or handful of tasks, while the capabilities of generative AI models tend to be broader. Most of the concerns regarding AI relate to the capabilities of generative AI. Indeed, the theoretical risks are far greater with generative AI approaches, including fraud and crime in general.

Generative vs traditional AI is an important distinction, as in general, the AI-ECG models described in this thesis are traditional AI models (with one exception, the variational autoencoder (VAE)).

Figure 1-3 depicts some important subtypes of AI, particularly in relation to the ECG.

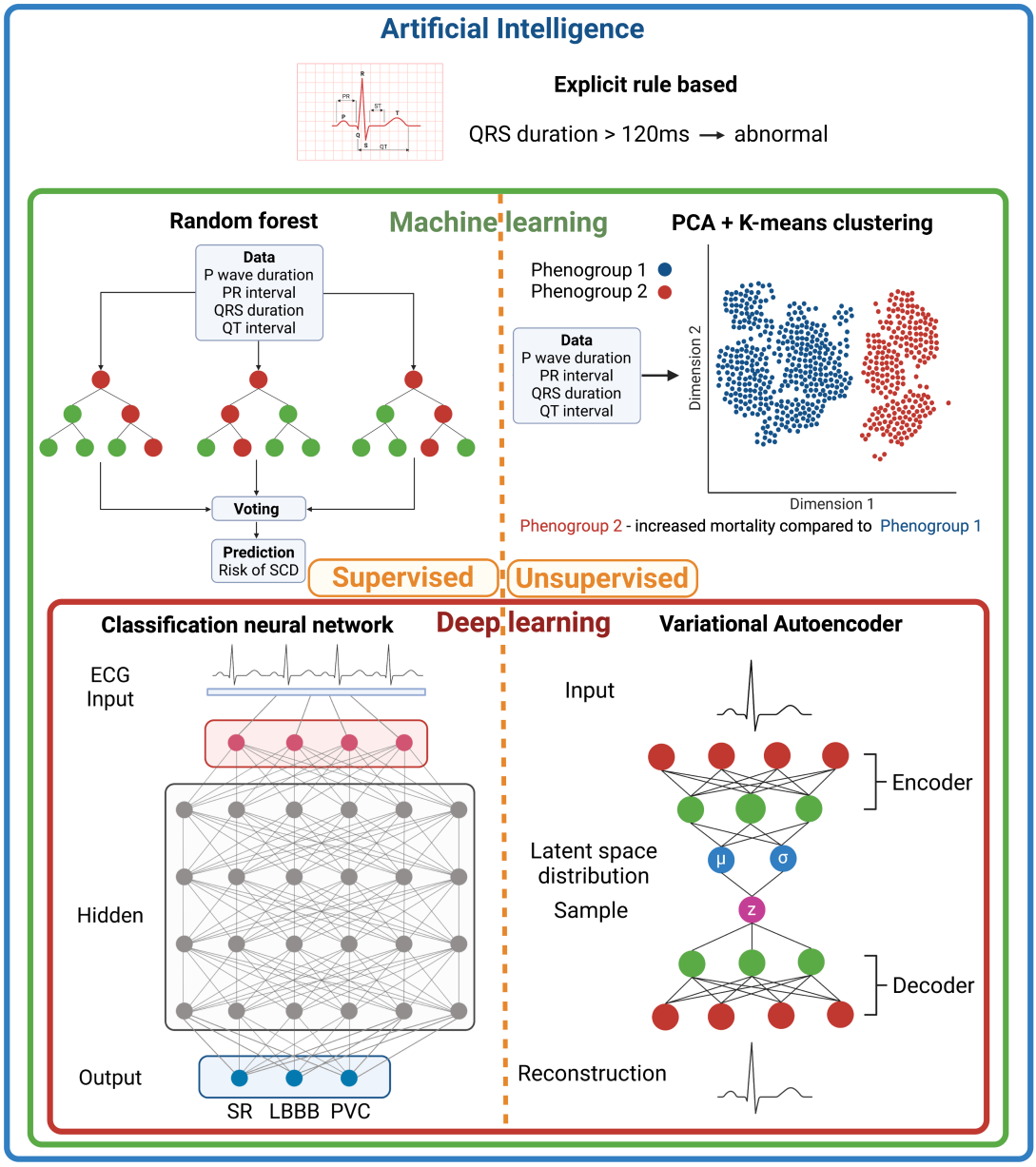

Artificial Intelligence
Explicit rule based
QRS duration > 120ms → abnormal
Machine learning
Random forest
Data
P wave duration
PR interval
QRS duration
QT interval
Voting
Prediction
Risk of SCD
PCA + K-means clustering
Phenogroup 1
Phenogroup 2
Data
P wave duration
PR interval
QRS duration
QT interval
Dimension 2
Dimension 1
Phenogroup 2 - increased mortality compared to Phenogroup 1
Supervised
Unsupervised
Deep learning
Classification neural network
ECG Input
Hidden
Output
SR LBBB PVC
Variational Autoencoder
Input
Encoder
Latent space distribution
μ
σ
Sample
z
Decoder
Reconstruction

Figure 1-3 Diagrammatic summary of artificial intelligence approaches

Examples of how artificial intelligence, machine learning and deep learning can be applied to the electrocardiogram. Random forest is an ensemble of several uncorrelated decision trees, each using a random subset of features. Each tree makes a classification prediction, the prediction with the most votes is the model's final prediction. Principle component analysis (PCA) is a form of linear dimensionality reduction that can be applied for example to ECG measurements. It can be combined with clustering algorithms such as K-means to find clusters of patients who may have difference clinical features and outcomes. Neural networks are multi-layered non-linear models that can be used to make connections within the input data and determine complex relationships between input and label to make classification decisions. A variational autoencoder is a form of unsupervised deep learning where the network learns to compress the input into a latent space distribution before reconstructing and aiming to return to the same signal as the input. They are commonly used for non-linear dimensionality reduction or removal of noise from signals. A single layer autoencoder with linear activation functions is almost equivalent to PCA.

SCD: sudden cardiac death, SR: sinus rhythm, LBBB: left bundle block. PVC: premature ventricular contraction

Supervised vs. Unsupervised machine learning

In the domain of traditional AI, by far the most common type of machine learning is supervised machine learning. This involves providing labelled data and allowing the

model to identify important features within the data that relate the input data to the label. For example, an ECG could be labelled with an abnormality such as left bundle branch block or another diagnosis such as left ventricular dysfunction. The machine learning model is given the data and the 'answers' (i.e. the labels) and then 'learns' the rules required to map the data to the corresponding label.

Unsupervised learning on the other hand does not use labelled data, instead the goal of learning is to identify common features in the input data, these could be used for example to cluster data into similar groups. In the healthcare setting most commonly this would be employed to cluster patients into similar phenogroups that may share clinical characteristics, prevalent disease, or response to a particular treatment.

Traditional machine learning vs. Deep learning

Traditional machine learning (not to be confused with traditional AI) involves human input for feature engineering, for example a human may determine QRS duration to be an important feature for predicting left ventricular impairment. After feature engineering, a variety of algorithms may be used to allow learning from these features and to then make intelligent classification or regression decisions.

Deep learning on the other hand, takes the entire input (e.g. the whole 12 lead ECG as a time series) without any human input for feature engineering, and derives features from the data. This has significant advantages in that features that may not be apparent to humans can be 'learnt' by the algorithm.

For example, if training a machine learning model to identify vehicle type, a feature used in traditional machine learning might be the number of wheels, cars have four wheels while a motorbike has two. While using deep learning, the model might 'learn'

the same feature - number of wheels, or might use another feature, such as the presence of windows in the car.

Deep learning architectures are built from models called neural networks. One definition is that a deep learning model has 3 or more layers. Neural networks are structured with multiple layers, most of which are ‘hidden’ layers (that is, between in the input and output layers) that are used to make connections within the input data and determine complex, non-linear, relationships between input and label.

Deep learning models however typically require much larger datasets to learn from, and due to the algorithm complexity have significantly longer model training times which are often very demanding on computational resources. Additionally, interpreting why the algorithm has made a particular decision is significantly more challenging for deep learning architectures, and can result in algorithms that function as a ‘black box’ (12).

Explainability

One major limitation and potential concern with any machine learning model, and in particular with deep learning models, is the challenge in explaining the reasons for making classification decisions. There are different schools of thought on this issue. On one hand, the major concern regarding a lack of interpretability is that the model may be using erroneous factors to make decisions. A famous story is that of huskies and wolf (13). A group of researchers trained a model on images of huskies and wolves in order to differentiate between the two species of animal. The model had excellent performance, however on reviewing cases that the model got wrong, they found that if a husky was not in the snow it was identified as a wolf, and a wolf in the

snow was identified as a husky. This was because in the dataset the model had been trained on, all images of huskies had a snowy background, and images of wolves did not. Challenges such as this are very real in the field of AI.

A review of studies using AI models to detect COVID-19 from a chest radiograph showed that many models used erroneous features (14). For example, parts of the image outside the lung fields such as the 'portable' marker that would generally reflect a more unwell patient, needing a portable study. These more unwell patients in general had COVID during the study period.

There are several methods to partially overcome these limitations. Saliency mapping is a commonly used technique to identify parts of the image or signal most important for the prediction decision (15), however there is no perfect method and some methods have important limitations (16).

The other school of thought is that interpretability is not required for adoption of AI models into clinical practice. McCoy et al argue that in striving for explainability in our AI models we may limit machines to the level of reasoning that humans are capable of, and therefore deprive the world of the unique problem-solving capabilities of AI (17). Furthermore, it can be argued that we hold AI to a higher bar than other tools used by clinicians. There are several examples of medical therapies and procedures where the exact mechanism of benefit is not clearly understood, but are used extensively following favourable data from clinical trials (18, 19). Lastly, human clinical decision making and intuition may often escape explainability (20), so it could be argued that AI should not be held to this potentially unrealistic target.

The answer to this question is likely a compromise between these two extremes, some explainability in AI is important, if for no other reason than to allow humans to more easily accept AI into clinical workflows. Requiring complete understanding of

the AI model however is probably an unrealistic goal that may stifle innovation if pursued rigorously.

As with any novel medical tool, testing AI models in clinical trials to show safety and efficacy is paramount to meeting the standards of evidence-based medicine and integrating AI successfully into clinical practice.

Artificial intelligence-enabled electrocardiography (AI-ECG)

Millions of ECGs are recorded around the world every year, some institutions have been storing these ECGs digitally for decades. In parallel, neural networks were developed and eventually the compute power required to develop neural networks using hundreds of thousands of ECGs allowed the first AI-ECG models to be developed. ECGs are particularly well suited for the application of AI because of the very large numbers of ECGs that are performed, the ease of storing them and the fact that they are high-dimensional and therefore have an untapped information content greater than for example tabular data. Although medical imaging is another high-dimensional data type, very large (>1 million) medical image datasets are difficult to curate store and use.

Artificial intelligence is a broad term as discussed above, however in the field of Artificial intelligence-enabled electrocardiography (AI-ECG), AI is in general taken to refer to deep learning. Deep learning-ECG could be considered a more accurate terminology, however, is not commonly accepted. Therefore, for the remainder of this thesis I will use the nomenclature of AI-ECG, as in the literature.

Deep learning to improve ECG interpretation

ECG machines have had computer interpretation for many years; therefore, it may not be immediately apparent what deep learning adds to this. Traditional computerised ECG interpretation uses a set of criteria developed by humans to make classification decisions based on the ECG input. For example, the computer interpretation on an ECG machine is programmed to measure PR interval, and where the PR interval is greater than 200ms it will make the interpretation of 1st degree heart block, in line with human defined criteria.

In contrast, the deep learning approach is able 'learn' from thousands of input ECGs and develop its own set of complex criteria for each diagnosis of interest. The latter approach can be thought of as being analogous to how a clinician may develop skills in ECG reading over a career of years of ECG interpretation and clinical correlation. In contrast, a deep learning model can accrue this 'experience' in a matter of hours if provided with the appropriate data.

Smith et al have shown that a deep learning approach to ECG interpretation can be significantly more accurate than traditional computerised interpretation (21). In their study, the deep learning model in particular had a significantly improved positive predictive value. Others have shown AI-ECG interpretation can outperform skilled humans (22, 23)

Deep learning to harness 'hidden' insights from the ECG

The recent exponential increase in deep learning applications in healthcare and beyond are largely due to two main factors, the availability of large healthcare datasets and the advances in processing power that allows very large deep learning

algorithms to be trained in a reasonable time frame. This has allowed AI-ECG models to be trained on very large datasets of more than 2 million ECGs (23). In particular this data and processing revolution has allowed models to be trained for tasks that would ordinarily be more suited to other modalities of investigation. Thereby unlocking the potential to not only emulate human performance, but to carry out tasks that even an expert clinician cannot do.

Given the relative ease and low cost of performing an ECG, an important potential role for AI-ECG is detection of subclinical disease. The first major example of this is detection of left ventricular systolic dysfunction (24). Attia et al described an AI-ECG model with impressive accuracy in predicting asymptomatic left ventricular systolic dysfunction (LVSD) (accuracy 85.7%, area under the receiver operating characteristics curve (AUROC) 0.93). Remarkably, patients with 'false' positive ECGs had a four-fold higher rate of incident LV dysfunction in the future. Other important examples include aortic stenosis detection (25) (AUROC 0.85), and identification of occult atrial fibrillation (AF) from the sinus rhythm ECG (AUROC 0.87) (26). These AI-ECG models are able to identify features that are both apparent to humans and novel features not previously described for by humans human (27).

The ubiquitous ECG also provides a fantastic opportunity for detection of rarer diseases that may otherwise go undetected. Ko et al describe the detection of hypertrophic cardiomyopathy with impressive performance (28)(sensitivity 87% and specificity 90%, AUROC 0.96). DNNs have also been applied to ECGs for amyloidosis (AUROC 0.91) (29) and peripartum cardiomyopathy detection (30) (AUROCs of 0.92 - LVEF <35%), 0.89 - LVEF < 45%), and 0.87 - LVEF < 50%). Application of deep learning to screening for rare disease could be particularly relevant in primary care and could trigger referral on for further testing. It is important

to however confirm the utility of this approach in prospective clinical trials, as with any screening investigation.

Disease phenotyping

Supervised machine learning is by far the most common form of machine learning applied to ECGs. However, there is also significant potential in using the ECG for unsupervised learning to identify important disease phenotypes that may not be apparent from human inspection of clinical data and the ECG. The application of unsupervised machine learning to the ECG is an emerging field, and is powered by the ease of acquiring an ECG. After disease phenotypes are identified, detailed exploration of the phenotypes is needed to understand how they might be useful clinically. This is an important distinction to supervised models, which typically have a clear and focussed clinical application before model development.

Feeny et al applied unsupervised machine learning to the QRS waveforms of patients who had cardiac resynchronisation therapy (CRT) devices implanted for heart failure. They identified two prognostically important subgroups, with significance beyond the commonly used parameters of QRS duration and left bundle branch block. Their findings indicate unsupervised machine learning could be used to better identify patients who would benefit from CRT (31). While Porumb et al describe using unsupervised machine learning in the form of a convolutional autoencoder to predict nocturnal hypoglycaemia (32). A convolutional autoencoder attempts to compress the input signal into a latent representation of features (in this case 20 or 50 neurons) and then decompress to create a signal as close to the original as possible. Through this technique the latent representation provides

features that describe the input signal and can then be used for clustering or for a supervised ML task. The algorithm correctly identified clusters of ECGs based on glucose levels with 90% accuracy.

Clinical translation

Applications of AI-ECG have been extensively studied in retrospective datasets, however prospective evaluation in clinical environments is important prior to clinical adoption. Yao et al described the first randomised clinical trial of AI-ECG vs. usual care (33). They applied the AI-ECG algorithm for LVSD detection previously described (24) in a cluster randomised trial of 120 primary care teams. A total of 22641 patients were recruited. Overall, the AI-ECG group had a 32% increased rate of low ejection fraction diagnosis (Odds ratio 1.32 (1.01-1.61), absolute diagnosis 2.1% vs. 1.6%). Importantly echocardiogram utilisation was not significantly different between the groups, indicating that AI-ECG may be a cost-effective tool.

Bachtiger et al report the application of a DNN for left ventricular dysfunction screening embedded within an ECG-enabled stethoscope (34). Using the ECG-enable stethoscope (Eko DUO; Eko Health, Oakland, CA, USA) participants had a single lead ECG recorded in the four standard anatomical positions for cardiac auscultation and one handheld location. Using the pulmonary location, which had the best single position performance, the AUROC was 0.85 for detecting left ventricular ejection fraction of less than 40%.

These examples demonstrate how quickly AI-ECG models can go from initial development (2019) to clinical evaluation studies (2021-2022), and show the great potential of AI-ECG in improving care for patients.

Cardiovascular risk prediction - state of the art

The burden of cardiovascular disease

Cardiovascular disease has varying definitions that can be limited to diseases of the cardiac and non-cardiac vascular system - including ischaemic heart disease, stroke and peripheral vascular disease. Broader definitions include all diseases of the heart and vascular system, including, for example, non-ischaemic heart failure and arrhythmias (35). For this thesis the second, broader definition will be used. Atherosclerotic cardiovascular disease (ASCVD) will be the term used to discuss the first, narrower definition.

ASCVD is one of the most common causes of death worldwide, and continues increase (36). Ischaemic heart disease and stroke resulted in 18.6 million deaths in 2019, having increased substantially (54%) since 1990 (37). As a result, treatment of cardiovascular disease makes up a significant proportion of healthcare spending, for example in the US this was $212 billion in 1996 and had risen to $320 billion in 2016 (38). This is due to a variety of factors, including obesity, a more sedentary lifestyle, dietary changes and widening health inequalities (39). Unfortunately, the burden of global ASCVD continues to increase substantially.

Non-atherosclerotic cardiovascular diseases also have significantly increasing burden. Atrial fibrillation (AF) is the most common cardiac arrhythmia with an estimated prevalence of 37,564 million cases (0.51% of the world's population). The prevalence of AF has increased by 33% in the last 20 years and has been called an epidemic by some (40). In clinical practice the most significant sequelae of AF are symptoms and embolic strokes.

Heart failure has also been called an epidemic (41) with 1-2% of UK healthcare expenditure being directed to managing it in 2017. As a result, significant resources have been directed at the diagnosis and treatment of heart failure. Early diagnosis remains key to improving outcomes.

Why predict risk of cardiovascular diseases?

Unsurprisingly, significant efforts have been made to predict the development of cardiovascular diseases as this may help facilitate early detection and diagnosis, early medical or interventional therapy and may help encourage lifestyle changes. Prediction of incident disease at the core of several major paradigms in cardiology:

1) Prediction of incident ASCVD to guide lipid lowering therapy (principally statins)

2) Predicting incident stroke (and bleeding) in patients with atrial fibrillation to guide prescription of anticoagulation

3) Predicting life-threatening ventricular arrhythmias to guide implantable cardioverter defibrillator (ICD) implantation

Currently risk prediction algorithms in clinical use are based on clinical history, biomarkers and imaging parameters integrated using traditional statistics. These may be integrated, using for example a Cox proportional hazards model (42-44), may consist of simple scoring system such as CHADsVASc (45) or HASBLED (46), or may use a single parameter such as left ventricular ejection fraction. These scores have seen significant clinical uptake due to their ease of use. However, the major disadvantage is their limited performance. The Framingham Heart Study resulted in the Framingham Risk Score (FRS) which was the first major CV risk prediction tool (47). Various iterations of the American Heart Association guidelines have used

versions of the FRS to guide preventive therapy for ASCVD, including the Pooled Cohort Equation (PCE). A major limitation however is the limited performance in other cohorts, for example analyses suggest ASCVD risk may be overestimated by over 50% in contemporary European populations (48), overestimation was also observed in contemporary multi-ethnic American populations (49).

Since the seminal trials of anticoagulation in AF (50), patients at elevated risk of stroke have been recommended anticoagulation, previously with warfarin and more recently with direct oral anticoagulants (DOACs) (51). Currently the most common risk score used for stroke prediction in AF is the CHADsVASc score (45).

Ventricular arrhythmia, although relatively rare in comparison to ASCVD and AF, is more frequently imminently life threatening. In patients with a history of haemodynamically compromising sustained ventricular arrhythmia, an implantable cardioverter defibrillator (ICD) is recommended (52). However sudden cardiac death (SCD) in subjects not previously known to have cardiovascular disease remains a significant problem. Additionally, in subjects with established cardiovascular diseases, prediction of life-threatening ventricular arrhythmias is challenging. In subjects with heart failure, the European and American guidelines advocate consideration of ICD implantation where the left ventricular ejection fraction is <35%. However this leaves a significant number of individuals unprotected, who may go on to have SCD (53), while also leading to ICD implantation in subjects who will never have life-threatening ventricular arrhythmias. As ICDs can have significant risks of complications in the long-term, more precise risk stratification strategies are needed.

Mortality prediction

Prevention of death, along with preservation of quality of life, is a major goal of clinical medicine. Prediction of death therefore stands to be potentially the most important application of risk prediction. Cohorts that may benefit from mortality prediction vary greatly, from acutely unwell patients such as those admitted to the intensive care unit to the general population (54). Inpatient cohorts in whom risk prediction can guide the level of care (e.g. ward-level vs. high-dependency vs intensive care) or specific interventions stand to benefit the most. The role of mortality prediction in an unselected general population cohort is less clear, those identified at high risk could benefit from a series of basic investigations which may include bloods tests and imaging such as echocardiography.

Risk prediction in specific, generally high risk, disease groups is of significant interest. There are two main reasons why mortality prediction in this context is potentially more clinically relevant. First, patients with high-risk diseases are likely to be at highest risk of death from that disease and not other competing risks. Second, often that high risk disease has a specific intervention. Examples of high-risk diseases in cardiology could include severe aortic stenosis and pulmonary arterial hypertension, both of which have had mortality prediction efforts (55, 56).

Atrial fibrillation

Risk prediction and screening for AF is one potential strategy to increase AF detection and reduce the incidence of stroke due to AF. Interest in the area has been driven by the increasing burden of AF related strokes (57) as well as availability and uptake of mobile and digital solutions to AF detection and prediction. The rising

incidence of AF has led to significant increased associated morbidity, mortality and associated costs of healthcare (58). Often AF is asymptomatic, and the first presentation may be with a catastrophic stroke. AF related strokes are also known to be amongst the most disabling (59). Anticoagulation has long been established to be effective in reducing the risk of stroke (50), and is generally advisable with an elevated CHADsVASc score (45).

Another challenge in AF detection is that often AF is paroxysmal and therefore may not be apparent during short healthcare contacts. Longer term monitoring is cumbersome and costly, and may only pick up 9% of cases (60). Mobile health solutions such as the Apple Watch or Alivecor KardiaMobile (61) may be a more effective way to detect AF, however may not be accessible to all demographics, particularly older subjects that are more likely to have undetected AF and are more likely to derive benefit from anticoagulation.

Another challenge in AF care is the disconnect between clinical and subclinical AF. The seminal trials of anticoagulation for stroke prevention used a clinical diagnosis of AF to enrol patients. Clinical AF is generally either symptomatic, or long-lasting enough that it is detected during short clinical encounters or 10 second ECGs. In contrast, subclinical AF may be short-lived and asymptomatic, and detected only during prolonged monitoring or via a cardiac implantable electronic device (CIED). Recent trials of anticoagulation in CIED detected atrial high-rate episodes (AHRE) have shown small or no benefits to anticoagulation in this context (62), but with increased risks of bleeding. AF screening studies have also demonstrated only small benefits from anticoagulation in screening detected AF (63). There therefore remains an unknown question of how best to address subclinical AF detection, and who, if anyone, would benefit from anticoagulation in subclinical AF.

AI-ECG for cardiovascular risk prediction

Why AI-ECG?

The electrocardiogram (ECG) is widely for the assessment of cardiovascular disease. The ECG captures information relating to a wide range of cardiac pathology, from cardiac ischaemia to cardiomyopathy and channelopathies to conduction system disease. Several specific ECG features have been associated with adverse long-term prognosis, including left bundle branch block (LBBB) (64), prolongation of QRS duration (11), and ECG changes consistent with left ventricular hypertrophy (65). However, these crude ECG features are based on human interpretation or ECG measurements and there may be other ECG features, not easily discernible by humans, that may be as (or more) clinically and prognostically meaningful.

Cardiovascular disease

An obvious and important area for improvement is prediction of ASCVD, as described above. Hughes et al have described the prediction of cardiovascular death using AI-ECG (66), while Awasthi et al. applied AI-ECG for the identification of occult coronary artery disease.

Arrhythmia

A major potential use of AI-ECG for cardiovascular risk prediction is the identification of occult atrial fibrillation (AF) and/or prediction of future AF development. The

seminal publication in this field was Attia et al who described, using AI-ECG applied to a sinus rhythm ECG, to identify subjects who would go on to be diagnosed with clinical AF within 30 days (26). Subsequently others have developed models to predict the future development of AF several years after a sinus rhythm ECG (67), these models have been combined with clinical risk scores such as CHARGE-AF (68) to provide further improvements.

Prediction of ventricular arrhythmias and sudden cardiac death has also been attempted. Kwon et al have described the use of the ECG to predict sudden cardiac arrest within 24hrs (69). Their model had an AUROC of 0.948 in an external dataset using 12 leads and 0.925 using a single lead (lead I). The latter has important implications for the potential for use in a wearable device or implantable loop recorder (ILR). The 24hr time horizon however is challenging and makes intervention challenging.

Previous small studies have demonstrated the potential application of AI-ECG to predict ventricular arrhythmias with a longer time horizon (70, 71). Sammani et al (70) describe the application of a variational autoencoder based VT/VF prediction model in dilated cardiomyopathy, with a C-index of 0.67, in a total of 695 subjects.

Mortality

Preventing death is one of two major goals of clinical medicine (the second being maintaining quality of life), and is an important endpoint when evaluating clinical interventions. Raghunath et al first described the ability of AI-ECG to predict death (72), with Sun et al having incrementally added to this goal (73).

Biological aging

Frailty and accelerated aging are associated with mortality (74). There are many ways to measure frailty and biological age, including clinical scores, biomarkers and imaging (74). AI-ECG has also been used to measure biological aging and was found to be associated with mortality (75). An AI-ECG model was trained to identify chronological age, which initially seems of minimal value. However, the concept introduced here is that AI-ECG age is a marker of biological or 'heart' age. Delta-age is the difference between chronological age and AI-ECG age, where AI-ECG age is higher than chronological age, this can be considered a marker of accelerated biological aging, and was found to be associated with adverse prognosis. A similar concept has been described with respect to BMI (76).

Biological exploration

Digital biomarkers and OMICs

While, in general, the main goal of AI-ECG models has been to identify prevalent or predict future disease, a relatively untapped application of AI-ECG traits is as a digital, 'omics' biomarker. Parallels can be drawn with the relatively recent field of radiomics. Radiomics has been defined as the 'extraction and analysis of large amounts of advanced quantitative imaging features with high throughput from medical images' (77).

A digital biomarker on the other hand can be defined as ‘a digital biomarker to be a characteristic or set of characteristics, collected from digital health technologies, that is measured as an indicator of normal biological processes, pathogenic processes, or responses to an exposure or intervention, including therapeutic interventions.’ (78). AI-ECG traits could be considered an ECG biomarker or ‘ECG-omics’ traits. For this thesis I will refer to these traits as ECG biomarkers.

Phenome-wide association

Radiomics traits have been explored with phenome-wide association studies (PheWAS) to provide a detailed exploration of biological associations (79, 80). PheWAS was first described, and is still most commonly used in the context of single-nucleotide polymorphisms (SNP) identified through genome-wide association studies (GWAS). PheWAS is essentially a GWAS in reverse and involves taking the trait of interest (which may be a SNP or biomarker) and looking for associations with that trait in a dataset of detailed phenotype data (81). The specific methodology underlying PheWAS is described in the methods chapter.

Although type 2 error can be a significant problem with PheWAS studies, they provide an interesting, hypothesis generating approach that can take advantage of the detailed datasets available, and may identify novel hypotheses not previously considered.

Another benefit to PheWAS is as a way to further explain AI-ECG predictions and describe their biological underpinnings.

Genome-wide association study

Most commonly, GWAS aims to evaluate the associations of a trait with common SNPs. These results can be used for biological exploration and may provide hypothesis generating evidence for disease biology and/or identify drug targets. Additionally, the results of a GWAS can be used to develop genomic risk prediction scores such as polygenic risk scores (PRS).
Traditionally GWAS has been used in relation to clinical diseases, however recently GWAS has been applied to imaging and ECG traits (82, 83).

Scope of thesis

The main aims of this theses are to explore how AI-ECG can be used to predict risk of adverse cardiovascular events. Additionally, I aimed to explore the biology around AI-ECG predictions, and better understand the phenotypic and genetic associations with these digital traits.

Hypotheses

The main hypotheses addressed in this thesis were:

1. AI-ECG can be used to predict risk of death
2. AI-ECG can be used to predict risk of cardiovascular diseases
3. AI-ECG predictions can be used as a biomarker for prevalent and future disease

4. AI-ECG predictions can be explained and have biologically plausible associations

In order to address these hypotheses, there are several specific aims described in this thesis.

1. Develop AI-ECG models using unsupervised machine learning
2. Develop AI-ECG models in a supervised way to predict mortality and cardiovascular diseases
3. Use multiple strategies to explain AI-ECG predictions
4. Explore the biological plausibility of AI-ECG predictions
5. Explore the meaning of AI-ECG predictions as a biomarker of clinical and subclinical disease

Chapter 2 - Methods

Parts of the work in this chapter of the thesis have appeared in the pre-prints:

Sau et al, Neural network-derived electrocardiographic features have prognostic significance and important phenotypic and genotypic associations medRxiv, https://www.medrxiv.org/content/10.1101/2023.06.15.23291428v1

and

Sau et al, Artificial intelligence–enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform, Lancet Digital Health (accepted, in press)

Introduction

This chapter will provide details of the cohorts and broad methods applied to all analyses and explanations for the principles underlying the main analyses. Specific methods for each chapter will be covered in the respective methods sections for each results chapter.

The digital electrocardiogram

In the UK, ECGs are generally printed onto paper with a background grid, at a standardised ‘paper speed’ (25mm/sec) and voltage calibration (10mm = 1mV).

Paper ECGs are not well suited to big data analysis techniques, given the massive workload that scanning large numbers of ECGs would be. Additionally, the process of printing, storing (often in poor conditions) and scanning ECGs may lead to information loss or distortion. Clinical ECGs may also (unfortunately) often have writing or drawing on the tracing itself. The ideal solution is to use natively digital ECGs, which already exist in an appropriate format for machine learning. Unfortunately, many institutions around the world do not store ECGs in a digital format, this is an important limitation to the findings in this thesis that is discussed in Chapter 7.

Digital ECG formats vary, but largely consist of lists of numbers at a given sampling rate (often 500hz) at a given scale (often 1mV) for each of the 12 leads of the ECG. It is important to note that leads III, aVL, aVR, aVF are linear combinations of leads I and II and therefore do not provide additional independent information content beyond the other 8 leads of the ECG. As including these additional leads can increase model training time without providing any benefits, it is common in the AI-ECG literature to omit these leads during model training and inference. The baseline AI-ECG model used in chapter 3 was initially trained using all 12 leads, therefore the full 12 leads are used in this model, however the remaining models use only the 8 independent ECG leads.

ECG pre-processing

In order to remove high frequency noise and low frequency baseline wander, ECGs were pre-processed with a bandpass filter 0.5 to 100hz, a notch filter at 60hz. My

prior work and the work of others (23, 84, 85) has found downsampling the ECG signal can retain model performance with the advantage of reducing model training time and improving compatibility with ECGs recorded at a lower sampling rate. ECGs from all datasets were therefore re-sampled to 400hz. Zero padding (in order to create an input size that was a power of 2, a convention in machine learning) resulted in a signal with 4096 samples for each lead for a 10s recording.

Cohorts

A common paradigm in machine learning is that models are only as good as the data they are built on. Additionally, external testing of machine learning models is key to ensure they generalise to other datasets/populations. In order to explore the hypotheses described above, large ECG datasets were needed. I therefore sought collaborators around the world to establish what I believe is one of the world's largest international digital ECG cohorts. The complete dataset is made up of six diverse cohorts, each with important advantages and limitations. Access to these datasets evolved over the course of the research period, this evolution, in part, guided selection of derivation and external testing cohorts for each subsequent analysis.

(I) The CODE Cohort

The Clinical Outcomes in Digital Electrocardiography (CODE) cohort is a database of 2,322,513 ECG records from 1,676,384 different patients of 811 counties in the state of Minas Gerais/Brazil from the Telehealth Network of Minas Gerais (TNMG). The

cohort is linked to public mortality databases. Patients over 16 years old with a valid ECG performed from 2010 to 2017 were included. Clinical data, including medical diagnoses, were self-reported. In a 15% stratified sample of the original cohort (CODE-15) an ECG was labelled “normal” according to conventional clinical reporting and based on automated interval measurements (75). The population reflects a primary care setting.

(II) The SaMi-Trop Cohort

The SaMi-Trop cohort is a prospective cohort of 1,631 patients with chronic Chagas cardiomyopathy and has been previously described in detail (86). Briefly, the inclusion criteria were: self-reported Chagas disease and aged 19 years or more. Digital ECGs were performed in 2011-2012 by TNMG. 83% of this cohort had abnormal ECGs (87).

(III) ELSA-Brasil Cohort

ELSA-Brasil is a cohort study of 15,105 Brazilian public servants (i.e. volunteers) aged 35 to 74 at enrolment. All active or retired employees of six participating institutions were eligible for the study. The full inclusion criteria and protocol have been previously described (88). 13,739 subjects had ECG and outcome data available for analysis.

(IV) The UK Biobank Cohort

The UK Biobank is longitudinal study of over 500,000 volunteers aged 40-69 at the time of enrolment in 2006-2010 (89). At baseline assessment participants provided information on health and lifestyle via questionnaire, had physical measures taken (including height, weight, and blood pressure) and donated samples of blood urine and saliva. A subgroup of participants were invited back for subsequent visits for additional investigations, including for detailed studies including cardiac magnetic

resonance imaging (MRI), brain MRI and digital ECGs. 42,386 subjects with digital ECGs taken at the instance 2 visit were available for analysis at the start of my research period. Subsequent more ECGs were conducted which reflects the different n numbers in this cohort in different analyses.

There is evidence of healthy volunteer selection bias (90). Outcomes were linked to cancer and death registry data, hospital admissions and primary care records. Detailed phenotyping using the cardiac MRI data has been previously described (80, 91).

(V) The Whitehall II Cohort

The Whitehall II cohort has been previously described (92). Briefly, British civil servants were enrolled into this cohort and had repeated calls for medical investigations. Between 2007 and 2009 participants (n=5,066) had digital 12-lead ECGs performed, for 5 minutes. The first 10s of this recording was considered in our analysis. Causes of death were ascertained based on the information contained in the death certificate. The novel ECG parameters micro-QRS fragmentation and QRS-T angle have been previous applied to this cohort, the same values were used in our analysis (93). Due to dataset usage restrictions, this cohort was available for use only in Chapter 3.

(VI) The Beth Israel Deaconess Medical Center (BIDMC) cohort

The BIDMC cohort is a dataset comprised of routinely collected data from Beth Israel Deaconess Medical Center (BIDMC), Boston, USA. BIDMC is a large urban care hospital catering for secondary to quaternary level care. Subject over 16 years old with a valid ECG performed from 2014 to 2023 were included in the cohort. Prior ECGs back to 2000 were included for these subjects. Mortality was determined via the Massachusetts Department of Public Health and/or review of the BIDMC

electronic medical record, while diagnostic International Classification of Diseases (ICD) codes were used to determine disease status. Subjects were censored at time of death or last in-person hospital contact. 1,163,401 ECGs were available from 189,539 subjects. Mean follow-up period was 5.46±5.81 years on a per ECG basis

Ethical approvals

The CODE study was approved by the Research Ethics Committee of the Universidade Federal de Minas Gerais, protocol 49368496317.7.0000.5149. The Whitehall II study was approved by the Joint University College London/University College London Hospitals Committees on the Ethics of Human Research. The UK Biobank has approval from the North West Multi-Centre Research Ethics Committee as a Research Tissue Bank (application ID 48666). The Longitudinal Study of Adult Health (ELSA-Brasil) was approved by the Research Ethics Committees of the participating institutions and by the National Committee for Research Ethics (CONEP 976/2006) of the Ministry of Health. The São Paulo-Minas Gerais Tropical Medicine Research Center (SaMi-Trop) study was approved by the Brazilian National Institutional Review Board (CONEP), No. 179.685/2012. For the BIDMC cohort ethics review and approval was provided by the Beth Israel Deaconess Medical Center Committee on Clinical Investigations, IRB protocol # 2023P000042.

Normal ECG definition

I evaluated the performance of the AI-ECG models in clinician reported normal ECGs. In the BIDMC dataset a subset of ECGs had Cardiologist reports. Normal ECGs in BIDMC were determined by searching for ‘normal ecg’ in the free text reports, a whole word match was required in order to exclude ‘abnormal ecg’. ECGs with the phrase ‘otherwise’ were also excluded from the normal definition. Additionally, I filtered by heart rate (60-100 bpm), PR interval (less than 200ms), QRS duration (less than 120ms) and QTc interval (less than 470ms). Normal ECGs in ELSA-Brasil and CODE were defined as previously described (94).

AI-ECG models

Anatomy of a neural network

Deep learning allows the development of highly sophisticated models capable of ‘learning’ from very large amounts of data. As the majority of deep learning is supervised in nature, that will be the focus when describing the concepts. Neural networks, inspired by the structure of the human brain, consist of multiple layers, or neurons, that lie between the input data and the output label. The ’magic’ of neural networks lies in their ability to ‘learn’ complex, non-linear patterns from data. Broadly a neural network consists of interconnected layers of artificial neurons, through which data flows. The weights and biases of the neurons dictates how data is processed as it flows.

Gradient descent is the means by which this flow of data leads to an updating of the neuron weights, such that with each pass of the data, the neural network iteratively improves the output predictions to more closely match the target labels. The aim of gradient descent is to minimise the loss function. The loss function compares the target (labels) with the predictions during model training, the loss is the difference between the target and the predictions. With (typically) each mini batch of data, model weights are updated in the direction of steepest descent (the direction which minimises the loss), to guide the model to the optimal solution.

A simple form of neural network is a multilayer perceptron (MLP). An MLP consists of an input layer, a number of ‘hidden’ layers and the output layer. The hidden layers in an MLP are fully connected, meaning each neuron in one layer connects to every node in the subsequent layer. MLPs are well suited to tabular data, although traditional machine learning approaches (including random forests and gradient boosting methods) also perform well on this datatype. The multiple layers and connections between them allow the computation of complex relationships.

Although multiple layers allow complex relationships, without activation functions these relationships are bound by linearity, no matter how many layers there are. Non-linearity is commonly introduced into neural networks through activation functions such as the ubiquitous rectified linear unit (ReLu). This simple function outputs the input directly if positive or zero if negative. This surprisingly simple function has been shown to be extremely effective in neural networks. Sigmoid is another commonly used activation function, often used in the final classification layer. The sigmoid output shape creates a continuous output in the range 0 to 1. Despite the potential of

the MLP, they are very inefficient for high dimensional data such as images, due to the very large number of parameters introduced by their fully connected layer structure. They also disregard spatial information and therefore are vulnerable to small shifts in the image input.

Convolutional layers solve some of these issues and are now commonly used for image processing. The convolutional neural network was originally developed for computer vision and excel at identifying local features that can be combined into increasing complexity with deeper layers (95). Convolutional layers can find and match patterns across the whole image and therefore are not vulnerable to translational shifts. As their weights are shared across the whole local region, they typically have far fewer parameters and therefore much shorter training times for a given input size. Multiple CNN layers allow earlier layers to learn simple features like edges, while deeper layers can learn more complex shapes that consist of the features learnt by earlier layers. Pooling layers (e.g. max pooling) are typically used to reduce spatial dimensions and reduce model complexity, while aiding the model to remain effective when features in the image change location. Finally, a fully connected layer (also known as a Dense layer) at the model head (final layers producing an output) would be typically used to combine the feature maps into an output value. A schematic diagram of a CNN is shown in Figure 2-1.

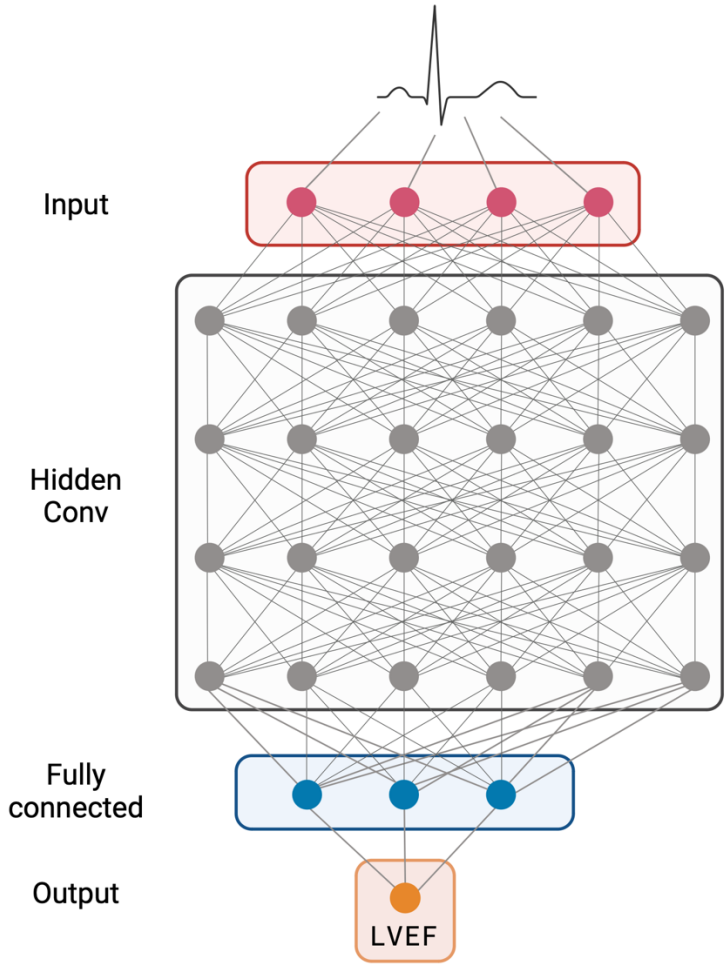


Figure 2-1 Schematic diagram of a convolutional neural network

Trained with ECG inputs for the supervised classification task of left ventricular ejection fraction (LVEF) determination. The input here is an ECG signal, this is passed to hidden convolutional layers, then a fully connected layer and finally the output, LVEF: left ventricular ejection fraction.

The reduced number of parameters per convolutional layer allows the creation of very deep neural networks. However, in this context vanishing gradients can be a challenge. As networks become deeper, the gradients used in updating the network

weights become very small. The Residual Neural Network (ResNet) was developed to combat the problem of vanishing gradients (96). Through the use of skip connections (also known as residual connections), the gradient information can bypass a block and be directly added to the output of a deeper layer. Gradients can therefore flow more easily through the network allowing very deep networks to be trained.

Training a neural network

Training a neural network broadly consists of four steps:

1. Data preparation
2. Model architecture design
3. Model training
4. Testing/external testing

I will address these steps in general terms, describing a supervised machine learning approach:

1. **Data preparation**

Data is first pre-processed and cleaned to ensure accuracy. This is discussed in more detail in a later section with specific reference to ECG pre-processing. Where supervised machine learning is used, data is split into training, tuning and test sets (**Figure 2-2**). The training dataset is used to develop the model. During training the tuning set serves several purposes, first it can be used for architecture selection and hyperparameter tuning. Secondly, it can be used to decide how long to train the model for, the number of epochs (one epoch is the passing of the entire training

dataset into the model), and select the model iteration with the best performance. After the optimal model is selected using the tuning set, the internal test set is used to evaluate the performance of the model in a dataset with the same demographics as the training dataset. It is important to avoid data leakage - this occurs when some information about the test set data 'leaks' into the training or tuning datasets. In the field of medicine one crucial way to prevent data leakage is by ensuring each subject has data allocated to only one of the training, tuning and test sets.

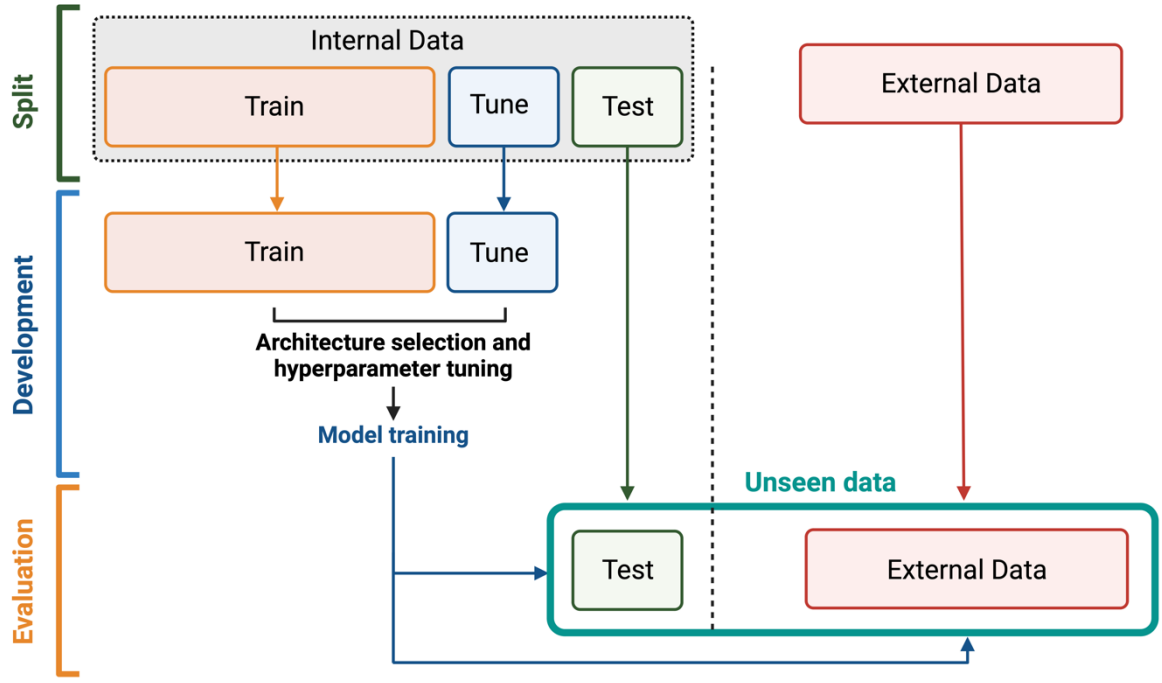


Figure 2-2 Flow of data in a machine learning study

An internal (development) dataset is typically split into training (Train), tuning (Tune) and testing (Test) sets. Training and tuning sets are used for model development before evaluation on the unseen internal test set. An external dataset is ideally utilised to ensure the model generalises well to a different population.

2. **Model architecture design**

Depending on the type of input (tabular data, free text, time-series) different neural network types such as an MLP, CNN or recurrent neural network (RNN) would be appropriate. After selection of the type of model, the detailed architecture design including selection of activation functions and number of layers is performed. This may be performed from scratch or using previously described architectures. Broadly, deeper networks have the capacity to learn more complex features that may improve performance, but increases the risk of overfitting. Overfitting is when the model effectively ‘memorises’ the input dataset and therefore cannot perform well on unseen data (i.e. it cannot generalise). A larger network with more parameters has a greater capacity to ‘memorise’ and therefore overfit, this problem is also worsened if the dataset is small and therefore easier to capture within the models parameters. The overfitting problem is less of an issue when dealing with very large datasets as in this thesis.

There are other measures in architecture design that can be used to reduce overfitting. These include batch normalisation and dropout. Batch normalisation transforms the inputs to the layer, so they have a mean of zero and unit variance. This introduces some noise, helping to reduce overfitting. Dropout involves randomly ignoring or ‘dropping’ a number of layer outputs, this also introduces noise into the model training process and makes ‘memorisation’ more difficult.

3. **Model training**

After steps 1 and 2, model training can begin. Here the training data along with the labels is fed through the neural network in batches. In each batch the network adjusts its weights through backpropagation and gradient descent to iteratively improve the model performance. The loss function compares the target (labels) with the predictions during model training, the loss is the difference between the target and the predictions. Therefore, the lower the loss, the better the model performance. The goal of model training is to minimise the loss function. After the data flows through the model (forward pass) using the current weights, the final output is compared to the ground truth using the loss function. Then the backward pass (hence backpropagation) allows the optimiser to adjust each weight in the network to help minimise the error. This iterative process is performed for the full dataset in batches. One complete run of the training dataset is called an epoch. Typically, several epochs (up to 50 or more) are needed for model training. The learning rate determines how quickly the optimiser should make changes in the networks weights. Typically, during model training the loss of the training dataset is compared to another dataset, the tuning dataset to help overcome overfitting. Tuning set loss is typically monitored, when this begins to rise, I.e. the model starts overfitting to the training dataset, training can be stopped, usually with a tolerance of a certain number of epochs (**Figure 2-3**). This is known as Early Stopping. Typically, the model with the best tuning set loss would be used for evaluation on the test set.

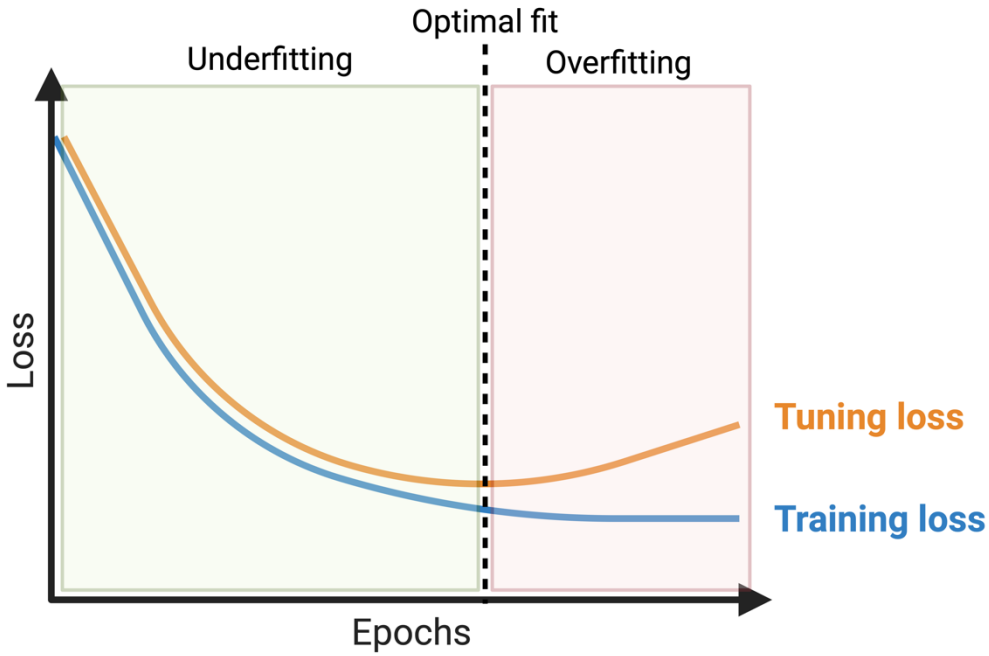


Figure 2-3 Example typical training and tuning loss curves

The lower the loss the closer the model fit to the data. Typically training and tuning loss fall in tandem up until the optimal fit, following which the tuning loss starts to rise – indicating overfitting.

4. **Testing/external testing**

   Finally, the performance of the model should be evaluated on an independent, unseen dataset. This is generally performed in an internal and (ideally) external dataset. The internal test set should be entirely independent of model training (i.e. it should not be the tuning dataset, and should not be used to select the final model). The external test dataset allows assessment of the model's ability to generalise to different data distributions.

Hyperparameters

Hyperparameters are usually determined by a human and dictate elements of model training. Examples of hyperparameters include the number of layers, batch size, learning rate, dropout. They may select based on experience, experimentation or systematic ‘hyperparameter tuning’.

AI-ECG model architectures

AI-ECG models are typically convolutional neural networks (CNN). The most common AI application in healthcare involves making a clinical diagnosis, this is a classification task.

Digital ECGs are stored as 1-dimensional time-series, therefore it may seem more appropriate to treat the data as a time-series and use architectures designed for this, such as recurrent neural networks including long short-term memory networks. However, although the ECG could be thought of the time-series, the goal in interpreting an ECG is not to predict the next part of the ECG, e.g. predicting the QRS complex from the p wave, as would be common for time-series analysis. Instead interpretation of the ECG is more about morphology of each component of the ECG, and therefore networks more suited to image analysis, like CNNs are more commonly used (97).

There are several types of CNN architecture design that have been applied for AI-ECG including simple CNNs (24), Residual Networks (ResNet) (23) and DenseNet (98).

Another problem with deeper networks can be vanishing gradients, as discussed above the residual block design of the ResNet is one way to overcome this issue (99). Work I conducted prior to this thesis compared different architectures for AI-ECG models (84, 85) and found this (relatively shallow) ResNet performed well in a wide range of AI-ECG tasks with large datasets.

The backbone architecture used throughout the thesis is based on the ResNet, but is considerably shallower than the archetypal ResNets used in computer vision (e.g. ResNet 18 and 34 (99)). The relatively shallow nature of the architecture allows rapid model training and reduces the risk of overfitting. The architecture consists of convolutional layer followed by four residual blocks (depicted in **Figure 2-4**) (23). Each residual block has two convolutional layers. Dropout layers are incorporated in each block. Stride length was 4. A model summary in tabular form is shown in **Table 2-1**.

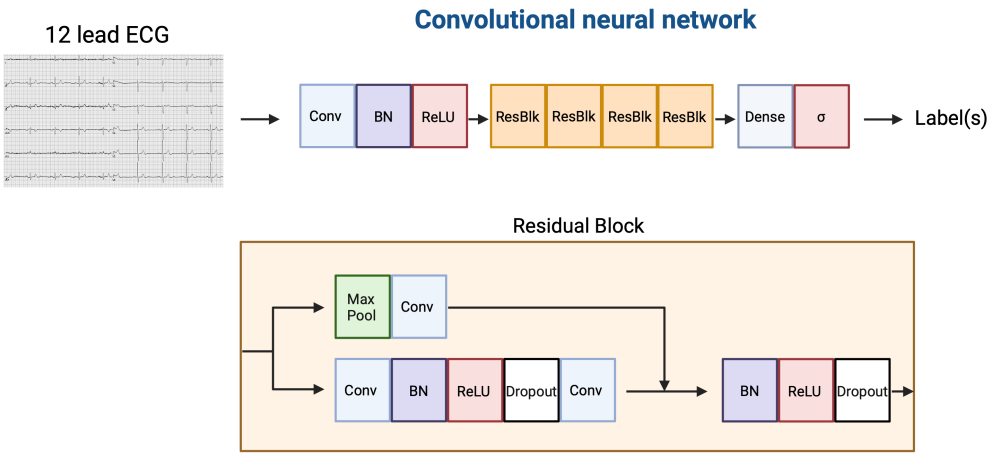


Figure 2-4 Schematic of the ResNet architecture used in this thesis

Conv: 1D convolutional layer, BN: batch normalization, ReLu: rectified linear unit, ResBlk: residual block

Adapted from Ribeiro et al 2020 (23).

Figure 2-4 gives an overview of the model architecture used in this thesis. Although it is possible that different architectures may suit different tasks, the differences are likely to be small and not clinically relevant. I therefore used this architecture for all tasks described in the thesis. In this section I will describe the role of each layer of the CNN.

Table 2-1

Model summary for the AIRE model (chapter 4)

Layer type	**Shape**
InputLayer	(None, 4096, 8)
Conv1D	(None, 4096, 64)
BatchNormalization	(None, 4096, 64)
Activation	(None, 4096, 64)
Conv1D	(None, 4096, 128)
BatchNormalization	(None, 4096, 128)
Activation	(None, 4096, 128)
Dropout	(None, 4096, 128)
MaxPooling1D	(None, 1024, 64)
Conv1D	(None, 1024, 128)
Conv1D	(None, 1024, 128)
Add	(None, 1024, 128)
BatchNormalization	(None, 1024, 128)
Activation	(None, 1024, 128)
Dropout	(None, 1024, 128)
Conv1D	(None, 1024, 196)

BatchNormalization	(None, 1024, 196)
Activation	(None, 1024, 196)
Dropout	(None, 1024, 196)
MaxPooling1D	(None, 256, 128)
Conv1D	(None, 256, 196)
Conv1D	(None, 256, 196)
Add	(None, 256, 196)
BatchNormalization	(None, 256, 196)
Activation	(None, 256, 196)
Dropout	(None, 256, 196)
Conv1D	(None, 256, 256)
BatchNormalization	(None, 256, 256)
Activation	(None, 256, 256)
Dropout	(None, 256, 256)
MaxPooling1D	(None, 64, 196)
Conv1D	(None, 64, 256)
Conv1D	(None, 64, 256)
Add	(None, 64, 256)
BatchNormalization	(None, 64, 256)
Activation	(None, 64, 256)
Dropout	(None, 64, 256)
Conv1D	(None, 64, 320)
BatchNormalization	(None, 64, 320)
Activation	(None, 64, 320)
Dropout	(None, 64, 320)
MaxPooling1D	(None, 16, 256)
Conv1D	(None, 16, 320)
Conv1D	(None, 16, 320)
Add	(None, 16, 320)

BatchNormalization	(None, 16, 320)
Activation	(None, 16, 320)
Dropout	(None, 16, 320)
Flatten	(None, 5120)
Dense	(None, 49)
Activation	(None, 49)

Classification and Regression

Classification tasks are the most commonly used for CNNs designed for clinical use and typically use a final sigmoid or SoftMax activation function. In the case of simple binary classification task (e.g. LVEF <40% or not) the CNN would have a single sigmoid activation that produces a continuous output value of between 0 and 1, similar to a logistic regression. Multiple diagnoses can be handled as mutually exclusive or not. For example, ECG diagnoses are often not mutually exclusive, e.g. left bundle branch block, sinus tachycardia and 1st degree heart block can all be present on an ECG, in this case a three-neuron sigmoid activation would be used to permit positive predictions in all three classes at the same time. If, however, the classes are mutually exclusive (e.g. LVEF <35 %, LVEF 35-45% or LVEF >45%) a SoftMax activation forces the model to ‘choose’ only one class.

A less common approach is a regression task. A simple example of this would be age prediction using an ECG, where a linear final activation might be used. This is analogous to a linear regression. Recent analysis suggests binarisation of continuous variables (such as LVEF as described above) is a suboptimal approach, and treating continuous data as a regression task yields better performing models (100).

Finally, a more complex application in healthcare is survival analysis. The traditional statistics equivalent would be a Cox proportional hazards model, where both time-to-event and censoring are incorporated. Prior work has shown where time-to-event/censorship data is available, a Cox regression may be superior to a logistic regression (101). This is intuitive and similar to example described above where for a continuous variable, a regression model is superior to a classification approach. Additionally, incorporating a survival approach allows us to use subjects lost to follow up. For example, if our end point of interest is 5-year mortality, strictly subjects without an event or completed 5 year follow up should be excluded. However, in survival analysis, these subjects can contribute to the model, up until loss to follow up. Implementation of a survival analysis approach in a neural network however is a more complex task than simple regression and classification.

Deep learning survival analysis

In the machine learning literature classification and regression tasks dominate. In clinical medicine, when aiming to predict future outcomes, we must consider both time-to-event (the event might be death or occurrence of a disease) and censoring. If we used a classification approach, we would choose one follow-up time and lose information on the time-to-event and we would not be able to account of loss to follow up, instead needing to either remove subjects that were lost to follow up or accept the potential noise introduced by including subjects in whom the true outcome is not known. This relatively unique problem that exists in clinical applications is typically solved using survival analysis in traditional statistics. Most AI-ECG models predicting

future events do not employ a survival approach (e.g. (66)), presumably due to the complexities around implementation of this type of loss function.

In order to use a survival approach for AI-ECG I implemented a recently described discrete-time survival neural network (102). Gensheimer and Narasimhan describe several advantages of their discrete-time survival approach compared to prior descriptions. This includes the ability to train models where the whole dataset will not fit in memory (as is the case with large ECG datasets) and faster training times than previous approaches. One drawback is the need to discretise follow-up time intervals, however this is not a significant issue if sufficient timepoints are used. For this thesis a 10-year time horizon was selected (as the vast majority of subjects had an event or completed follow up in this time). A non-proportional distribution of timepoints was used to encode with more granularity towards the start of follow up, where more events take place.

A custom loss function is used. Specifically, the patients' follow-up time is divided into a set of fixed intervals, and an estimated conditional hazard probability for each interval is calculated (probability of failure in that interval, given that the individual has survived at least to the beginning of the interval). For each time interval j, the loss function is defined as:

$$\sum_{i=1}^{d_j} ln\left(h_j^i\right) + \sum_{i=d_j+1}^{r_j} ln\left(1 - h_j^i\right)$$

where h_j^i is the hazard probability for individual i during time interval j, r_j is the number of individuals that have not experienced failure or censoring before the beginning of

the interval j, and d_j is the number of individuals who have during the interval j. The overall loss function is the sum of the losses for each time interval.

The input labels of the network include n (in this case n = 49) values to encode alive or dead at each timepoint, and separate n binary inputs encode censor or not at each timepoint. The output of the neural network is an n-dimensional vector, with each element representing the predicted conditional probability of surviving that time interval $\left(1 - h_j^i\right)$. An individual's probability of surviving through the end of time interval j is given by:

$$S_j = \prod_{i=1}^{j} (1 - h_i)$$

Training AI-ECG models

The ECG datasets used in this thesis include over 3 million individual ECG recordings, all anonymised prior to transfer to Imperial College London. Computational efficiency is therefore an important consideration. As discussed above, downsampling to 400hz and reduction to the 8 independent leads is an important step to reduce model training time. After pre-processing, ECGs were stored in a Hierarchical Data Format 5 (HDF5) format, which allows millions of ECG recordings to be stored in a single file, but with fast read access of specific files, without loading the entire file into random-access memory (RAM). The models in this thesis were trained using Imperial College London's high performance computing cluster. Nvidia GTX 6000 GPUs on the cluster were used. Significant advantages of this cluster included the ability to train multiple models using multiple GPUs at the

same time. I was often able to train eight or more models simultaneously. This greatly increased the efficiency of my work. The Keras framework with a TensorFlow (v2.4.1) backend was used for neural network training and inference (103, 104).

Model optimisation

Neural networks have many 'hyperparameters' that are manually configured variables that effect model architecture or training. The values of these hyperparameters can have an impact on model performance, and optimisation can be extremely time consuming. However, beyond certain key variables, such as learning rate, modification of other variables often have diminishing value. Additionally, extensive hyperparameter optimisation can lead to an overestimation of performance in the tuning set that then does not translate to the test set or external test set. This is however less of a problem with the very large datasets used in this thesis.

Throughout the course of this thesis, several models were trained. Detailed and systematic hyperparameter optimisation of models is time consuming, computationally costly, and often yields minor, if any, improvements. Learning rate is well established to be an important, perhaps the single most important (105), hyperparameter, therefore this was the sole hyper parameter optimised for the work presented in this thesis. In prior work (84, 85) I explored the optimal batch size and dropout in the ResNet architecture I used, and found appropriate values of these to be 32 and 0.2 (i.e. keep 0.8) respectively.

Transfer learning

Transfer learning and finetuning are important concepts in the field of machine learning that allow models trained for one task to be used when building a model for another task. Transfer learning involves taking the model built for one task and using it as a backbone for another task. This might involve a transfer of 'knowledge' between two highly similar tasks e.g. detection of left ventricular hypertrophy and diagnosis of hypertrophic cardiomyopathy or very distinct tasks, such as using models trained on broad image libraries and applying them to detecting skin cancer (106). Fine-tuning is the process of taking the pre-trained model and adapting it for the specific goal of the new model. This can be carried out with different approaches. One approach may involve keeping all of the pre-trained weights and just replacing the classification layer, for example if the two tasks are very similar. Another approach would be to simply use the pre-trained weights as starting point for training the new model, but allowing the new model to 'overwrite' the existing weights if appropriate. Thirdly, the weights themselves could be used with a totally different machine learning approach. Combinations of these strategies can also be employed. The key advantage of pre-training, transfer learning and fine-tuning is the ability to train models faster and improve performance of the final model. This is particularly helpful when limited data is available for final task.

BIDMC cohort data splits

The BIDMC cohort is used for training of all supervised machine learning models in this thesis. Subjects were split by patient ID, so a single subject could only have ECGs in one of train/tuning/test - avoiding data leakage. The split was stratified by

subjects with paired life status at 5 years. 50% of the data was used for model training, 10% for tuning and 40% of internal testing. The proportion of data reserved for internal test was intentionally selected to be relatively large compared to some areas of the machine learning literature. This was for a few reasons 1. The very large overall dataset means a large test split could be afforded, without significant loss of performance. 2. A smaller training dataset would mean shorter model training times. 3. A larger test set would allow detailed downstream analyses in a large number of subjects and in important smaller subgroups. The tuning set was used for saving the ‘best’ model during model training, as measured by the tuning loss. Where several models were trained, the model with the lowest tuning loss was used for evaluation of the test set. The subjects in the test set were not used for model training or tuning. The final model was used for evaluation of internal test and external test cohorts.

Survival and statistical analyses

As discussed above, survival analysis is in important statistical technique in clinical medicine to build or evaluate the performance of a model. The Cox proportional hazards model is a common way to perform survival analyses. Recent work suggests virtually all real-world clinical datasets will violate the proportional hazards assumptions if sufficiently powered and that statistical tests for the proportional hazards assumption may be unnecessary (107). Given the very large datasets used in this thesis, and in line with these recommendations, the proportional hazards assumption was not evaluated and the hazard ratio from the Cox models in this thesis, and should be interpreted as a weighted average of the true hazard ratios over the follow-up period. Throughout the thesis, p values of <0.05 are considered statistically significant, except where multiple comparisons are performed, and Bonferroni correction is performed. Statistical analyses were performed in Python or the R programming languages.

Explainability

As discussed in the introduction, complete explainability of deep learning models, like is possible in ‘simple’ regression models, is not possible due to the millions of non-linear calculations that make up a deep learning model. ‘Black box’ models however are unlikely to meet acceptance in society, particularly if used for healthcare. I therefore explored several methods of obtaining explainability during the course of this thesis.

Saliency mapping

A commonly deployed method for explainability of CNNs, is saliency mapping. Developed for computer vision (as with most aspects relating to CNNs), the basic principle behind a saliency map is that they are a way to identify which parts of an image support the classification prediction of the neural network.

To construct a saliency map, the gradient of the output of the network is calculated with respect to the input. This shows how each pixel of the input effects the class score. In this way we can understand the role each pixel plays in the network’s prediction. These gradients can then be used to construct a visual representation or heatmap, where a certain colour indicates higher pixel importance.

As briefly mention in the Introduction, an excellent example of the value of saliency mapping has been described in an analysis of chest X-ray models for the diagnosis

of Coronavirus disease 2019 (COVID-19) (14). DeGrave *et al* described how AI models for chest X-ray interpretation often focussed on non-clinical elements of the image, such as the location marker text that typically is placed in the top right of a chest X-ray image. This marker often reveals important information about the patient, for example that they may be situated in the resuscitation department at the time of the X-ray, and/or that the X-ray was taken with a portable machine. Both of these elements suggest a more unwell patient. At the time these data were collected, the majority of the most unwell patients had COVID-19 and therefore the model learnt the 'shortcut' of using the location markers to diagnose COVID-19. Of course, a model that does this is useless in clinical practice.

A challenge with saliency mapping, however, is that it does not provide further information beyond spatial location of 'important' features. When used for ECGs if a saliency map highlights, for example, the QRS complex, it cannot tell us what morphology of the QRS is important. We can only infer that the QRS is important (which is not surprising given that the QRS complex is the major component of the surface ECG).

Median beats

Although the 10s ECG contains a wealth of information, useful for clinicians and deep learning approaches, when seeking to visualise importance of ECG segments, median ECG beats are a valuable tool. Median beats were extracted using the Beth Israel Analysis of Vectors of the Heart (BRAVEHEART) ECG analysis software (108). Briefly, the R peaks of the ECG signals are identified, a CNN is used to identify QRS and T wave onset and offset and the beats are averaged to create the median.

The main advantage of this approach is the ability to identify the effects of different elements of a single heart beat on model predictions. Creating median beats allows subsequently to depict average morphologies between patients, without R wave misalignment issues. However, this may also have the negative effect of averaging out subtle signal changes, particularly if there is beat to beat variability in the morphology or presence of these changes. The impact of this on model predictive performance is mitigated as the median beats are used only for interpretability and not for model training or testing.

Inherently explainable deep learning architectures

One way to combine the advantages of deep learning with explainability is variational autoencoder (VAE). A VAE is trained by unsupervised learning (that is there are no labels). The input is the ECG, and the training target is the ECG (**Figure 2-4**). The goal of the network is to compress the ECG into a lower dimensional representation (e.g. 32 nodes or numbers, termed latent features). This type of architecture is used common in imaging fields for feature extraction (109), and can be applied to the ECG with a similar goal.

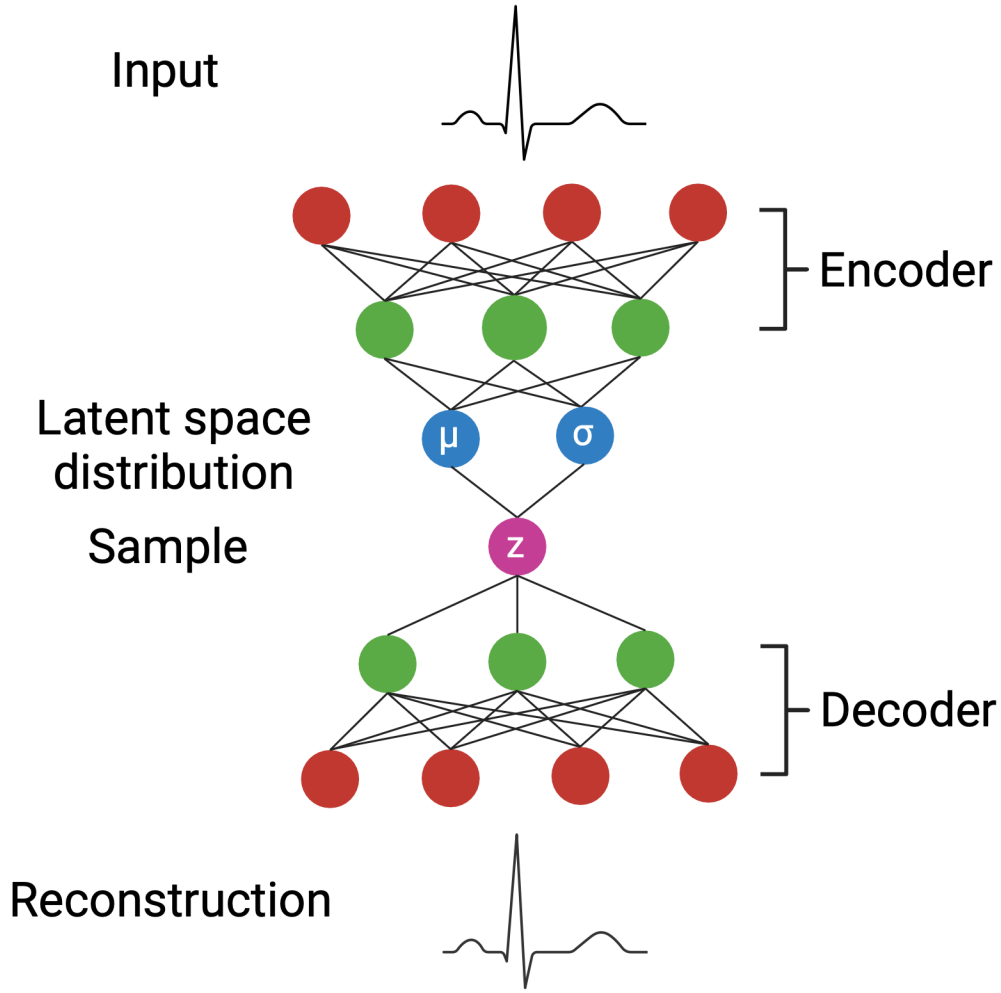


Figure 2-5 Variational autoencoder schematic diagram.

Z indicates the latent features; these features are used in downstream analyses.

The main advantage of this approach is that the values of the latent features can be directly attributed to changes in ECG morphology (**Figure 2-5**). These latent features can then be used in 'simple' regression models which allow preservation of complete explainability.

A disadvantage however is that as the VAE is trained in an unsupervised manner, and the level of dimensionality reduction is typical significant (to a relatively small number of features). Additionally, in using these latent features in a 'simple' regression model, the advantages of deep learning are partially lost and therefore the VAE + regression model approach was inferior to an end-to-end supervised CNN approach in preliminary analyses.

Another approach to using VAE features is as an explainability adjunct to an end-to-end supervised CNN. Here we simply assess the association between the VAE features and CNN predictions, allowing us to visualise the most associated VAE features with the CNN predictions. Although this form of explainability is indirect, it can provide valuable morphological insights, that are not available with saliency mapping approaches.

For this thesis, a VAE was trained using median ECG beats, this was performed in collaboration with Dr. Konstantinos Patlatzoglou and Dr. Ewa Sieliwonczyk. The VAE consists of three components: the encoder, the decoder and the latent space. The encoder and the decoder are made up of one-dimensional convolutional layers with increasing filters and decreasing kernel sizes closer to the latent space. The latent space was restricted to 30 features at a maximum, although typically only a subset of these features was used by the model for the reconstruction. The model was trained to minimise both the median ECG reconstruction loss, defined by a symmetric mean absolute percentage error function, and the Kullback-Leibler divergence (KL loss). This second term is specifically added to the VAE model to ensure that the features generated by the model are generative and disentangled. An additional β-parameter was included as a weight on the KL-term to optimize the balance between the reconstruction loss and the latent factor interpretability. We tested beta values of 0.1,

0.25, 0.5, 1, 3, 5 and 10 and defined the best model at a β-parameter of 0.25 based on the Pearson correlation between the median and its reconstruction in the tuning dataset, as well as a visual inspection of the latent vector transversals.

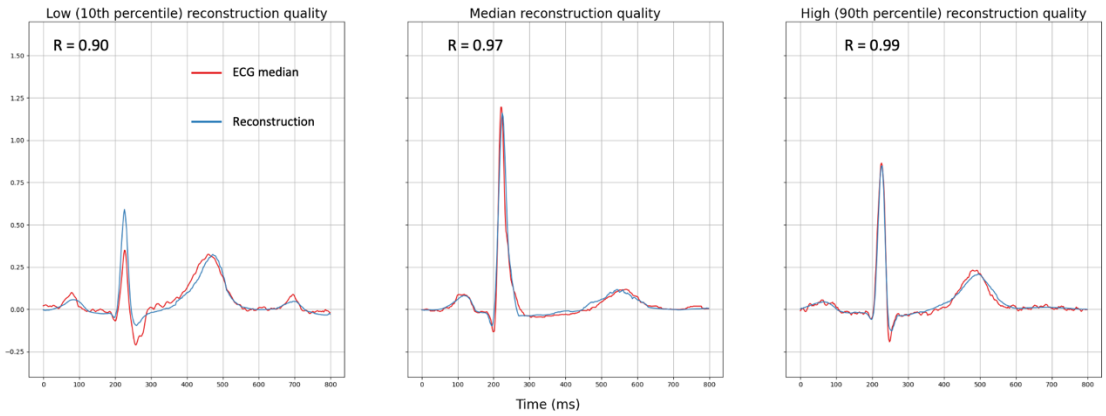


Figure 2-6 Variational Autoencoder reconstruction performance

Figure provided by Dr Ewa Sieliwonczyk. The Variational Autoencoder produces median beat reconstructions (blue) that closely match the original ECG medians (red). Example reconstructions spanning a range of reconstruction qualities are shown.

Visualisation of average beats

Another approach to visualise model predictions is simply to visualise averages of the most extreme predictions. The use of median beats is particularly advantageous here. We can simply take a sample of the ECGs with the highest and lowest predictions and average them to visualise features considered high/low risk. Interpreting the results of this method is inherently subjective, but can provide helpful additional insight for clinicians.

Correlation with standard ECG parameters

Finally, simple correlation with established ECG parameters can add quantitative metrics for assessing important aspects of model predictions, but is limited to the few established ECG parameters, including heart rate, PR interval, QRS duration and QTc interval.

Throughout the course of this thesis, I explore these methodologies to provide explainability to the AI-ECG models.

Biological exploration

As described in the introduction, I sought to explore the underlying biological associations with AI-ECG predictions, which can be thought of as digital biomarkers.

Phenome-wide association study (PheWAS)

I used two PheWAS approaches in this thesis, I have termed them continuous PheWAS and Disease PheWAS, these are described in detail below.

Continuous PheWAS

A PheWAS of continuous or ordinal categorical traits was performed in the UKB. The same approach was also applied to the echocardiographic data in the BIDMC dataset.

The UKB contains data from over 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 relevant covariates (described in the individual methods sections). Additionally, left ventricular trabeculation was calculated as previously described (110). Deep learning-derived brain age was calculated as previously described (111).

Disease PheWAS

In order to evaluate AI-ECG associations with the full spectrum of cardiac and non-cardiac diseases, I additionally performed a PheWAS of disease Phecodes, Termed disease PheWAS for the purposes of this thesis. Disease codes as defined by ICD9 and ICD10 codes converted into Phecodes in the BIDMC dataset. Phecodes were defined as previously described (112). Subsequently covariate-adjusted logistic regression is performed to evaluate the association between the AI-ECG prediction and the disease Phecodes. Bonferroni correction is applied to set the significance threshold.

Genome-wide association study

To identify genetic associations with the AI-ECG traits, I performed a genome-wide association study (GWAS) in the UKB. Where necessary, the data were normalized by rank-based inverse normal transform prior to the analysis. The GWAS analysis was adjusted for the following covariates: age at imaging visit, sex, height, body

mass index (BMI), imaging assessment centre and the first 10 genetic principal components.

Standard quality control was undertaken. Included single nucleotide polymorphisms (SNPs) had a minor allele frequency (MAF) >0.1% and an imputation INFO score of >0.4. Unrelated individuals of genetically-determined European ancestry were included as previously described (110). The GWAS was undertaken using FastGWA mixed linear model association analysis through the Genome-wide Complex Trait Analysis software using a genetic relationship matrix (GRM) to adjust for population structure (113). The Manhattan plots depict the nearest gene. Top SNPs were identified that had a P-value <5x10-8. Associations of the detected SNPs were evaluated using the NHGRI-EBI GWAS Catalog, PhenoScanner, GTEx and the GeneAtlas UK Biobank PheWAS browser.

Chapter 3 - AI-ECG phenogroups predict adverse outcomes

Introduction

Parts of the work in this chapter of the thesis have appeared in the pre-print: Sau et al Neural network-derived electrocardiographic features have prognostic significance and important phenotypic and genotypic associations medRxiv, https://www.medrxiv.org/content/10.1101/2023.06.15.23291428v1

As discussed in the introduction, supervised machine learning (ML) is by far most common form of machine learning in healthcare, and the same holds for machine learning applied to ECGs. Prior models have applied AI-ECG models to predict a label e.g., rhythm diagnoses, future atrial fibrillation or impaired left ventricular function (23, 24, 67).

Unsupervised machine learning in the field of AI-ECG has not been well studied. The main advantage of unsupervised machine learning in this context is the ability to perform an entirely data-driven exploration, without anchoring to specific hypotheses regarding disease states or specific outcome predictions. For the same reasons however, caution is needed to ensure any results of unsupervised machine learning are well evaluated and clinically relevant.

During the course of the aforementioned supervised machine learning the CNN identifies many thousands of ECG features, which are not limited to conventional ECG parameters and morphology, and applies these features to make the label prediction (114). These neural network (NN)-derived ECG features have not been extensively studied. There is the potential that they may be universal, transferable, and applicable to a range of clinically useful tasks.

I hypothesised that the ECG features learned by a CNN may have implications beyond the original task for which it was trained and may have important clinical, phenotypic, and genotypic associations that could be uncovered through unsupervised machine learning.

In this chapter, in order to test this hypothesis, I applied an existing AI-ECG model trained to classify six common ECG diagnoses relating to rhythm and conduction disease (23), these were, atrial fibrillation, sinus tachycardia, sinus bradycardia, right and left bundle branch block and 1st degree atrioventricular block. This model was selected as it had been trained for common diagnoses that involved all conventional components of the ECG. In order to use high level NN-derived ECG features, I used the features from the penultimate layer of the CNN and applied unsupervised machine learning to identify clinically distinct phenogroups. This is a form of transfer learning, as discussed in Chapter 2. Specifically in this context I use deep-learning derived features alongside traditional machine learning.

Given the critical role of external testing (115), I tested our analysis in five external datasets across three continents with diverse ancestral backgrounds. Furthermore, the model was evaluated across diverse cohorts including volunteers, primary care patients and patients with established cardiomyopathy. I subsequently explored potential biological mechanisms underlying the associations of the ECG

phenogroups with survival by evaluating their associations with a wide range of phenotypes and genetics. In particular, the findings in this chapter were pivotal to the development of follow on studies described in chapters 4 to 6.

Methods

Neural network-derived ECG features from CODE dataset

The CODE dataset was previously used to train a CNN to detect six common ECG abnormalities (first-degree AV block, RBBB, LBBB, sinus bradycardia, AF, and sinus tachycardia) (23). This model is referred to as CODE-CNN. Transfer learning was used to repurpose a machine learning model trained for one task to another. I hypothesised the ECG features learned by the CODE-CNN could be used to cluster subjects into clinically meaningful phenogroups. In order to extract features of the CODE-CNN, the final classification layer of the model was removed and the output of 5120 NN-derived ECG features is used to cluster the subjects.

Unsupervised machine learning for clustering into phenogroups

In order to derive phenogroups using the 5120 features extracted using the CODE-CNN model, K-means clustering was applied in the CODE dataset. k = 3 was selected using the elbow method (116). The sklearn package in Python was used to perform K-means clustering (117). In order to determine cluster assignment in the external testing cohorts, the 5120 features were extracted using CODE-CNN as

described above for each ECG. The K-means model (trained on the NN-derived ECG features from the CODE dataset) was then used to identify the nearest cluster centre and therefore determine the cluster assignment for each ECG. There was no retraining or calibration performed when applying the CODE-CNN and K-means models to the external testing datasets.

Survival analysis in derivation and external testing cohorts

Survival analysis for the 3 phenogroups was first performed for the CODE dataset, and then in the five external testing datasets (UK Biobank, Whitehall II, ELSA Brasil, BIDMC and SaMi-Trop). Survival analysis was repeated after excluding subjects with the six diagnoses identified by the CODE-CNN and also considering only physician-adjudicated normal ECGs in the CODE dataset. Descriptive statistics are displayed as medians (interquartile ranges) for continuous variables and numbers (percentages) for categorical variables. Kaplan-Meier plots were used to display cumulative mortality. The log rank test was used to compare survival curves.
Cox proportional hazards regression modelling was used to estimate hazard ratios for mortality while correcting for other known variables. The adjusted variables differed in each dataset depending on the data available. Statistical analyses were performed with R 4.0.0 statistical package (R Core Team, Vienna, Austria) or Python (version 3.9).

Phenome-wide association study

To better understand the biology underlying our AI-ECG model and to explore the detailed phenogroup associations, I performed phenome-wide association studies (PheWAS). In the UK Biobank phenogroup C had very limited representation and therefore was excluded from this analysis. I used two PheWAS approaches, firstly I used a Disease PheWAS to explore the association of ECG phenogroup with incident diseases and treatments. Logistic regression was performed to investigate this association. The second approach, Biobank PheWAS, used the UK Biobank that contains data from over 3000 phenotypes derived from patient measurements, surveys, and investigations. Univariate correlation was performed to investigate the association between ECG phenogroup and phenotypes as previously described (80). The phenogroup variable was encoded as 0 (phenogroup A) and 1 (phenogroup B).

GWAS

To identify genetic associations with the ECG phenogroups, a GWAS was performed. Subjects in phenogroup C were excluded from this analysis, due to the small number of subjects, as described above for PheWAS. Using logistic regression, a binary phenogroup variable was adjusted for the following covariates: age at imaging visit, sex, height, body mass index (BMI) and the first 10 genetic principal components. Further details are as described in Chapter 2.

Model explainability

In order to understand the elements of the ECG contributing most significantly to phenogroup determination I modified a commonly used technique in computer vision, Grad-CAM (118). A signal importance map for a given input of a 12-lead ECG signal was computed. First, the final layer of the original trained CNN is removed, and a new layer with a fixed operation is added. This layer takes the 5120 features from the trained CNN and computes its Euclidean distance to the centre of three phenogroup clusters, which results in three distance scores. Using softmin, these distance scores are mapped to probability scores. Using Grad-CAM, for a given input, we compute the gradient of weights of the first convolution layer with respect to the probability score for a specified phenogroup (e.g., phenogroup A). The output of the first layer for a given input is then multiplied by an average gradient of respective channels, and finally, the positive values are taken as the importance score for the saliency map. To extract the generalised saliency map for each phenogroup, 1000 ECGs from the centre of the three phenogroup clusters (approximately equidistant from all three cluster centroids) were taken. After computing saliency maps, each 10s ECG was averaged to one beat per lead by aligning beats with R-peaks. Similarly, the respective saliency map of each 10s was averaged to a single beat. A generalised saliency map for a cluster was computed by averaging (median) all the 1000 saliency maps. While plotting the generalised saliency map, single beats of all the 1000 ECGs were also averaged (mean) to one beat per lead.

Results

An overview of the analysis pipeline is provided in (**Figure 3-1**).

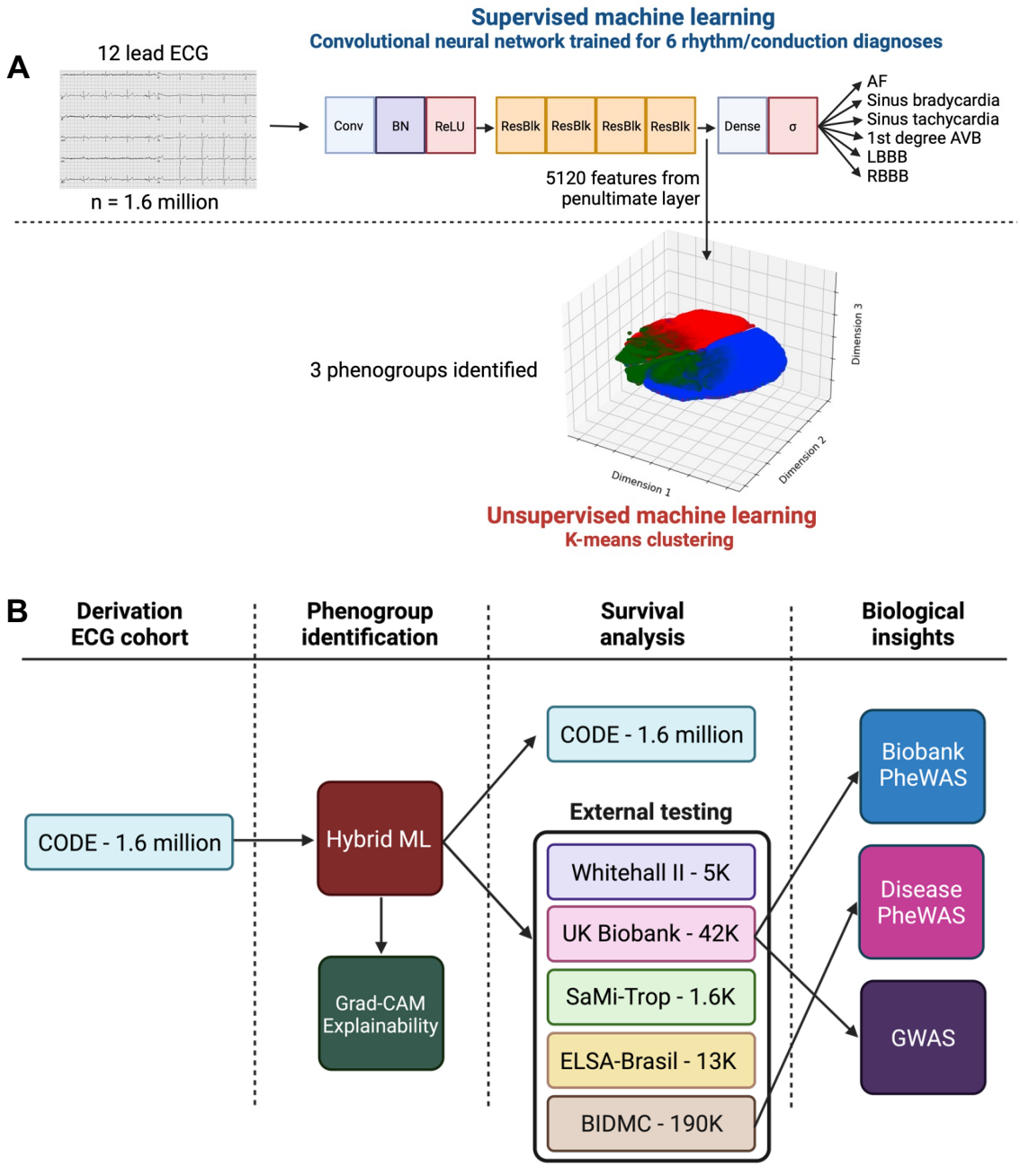


Figure 3-1 Data analysis pipeline

(A) Hybrid machine learning approach with combination of supervised and unsupervised machine learning to use NN-derived ECG features to identify phenogroups from the 12-lead ECG. 3D figure shows a dimensionally reduced representation of the phenogroups (B) Flow of data and analyses performed in this study. AF: atrial fibrillation, AVB: atrioventricular block, LBBB: left bundle branch block, RBBB: right bundle branch block, ECG: electrocardiogram, ML: machine learning, Grad-CAM: Gradient-weighted Class Activation Mapping, PheWAS: phenome-wide association study, GWAS: genome-wide association study

Derivation cohort

The Clinical Outcomes in Digital Electrocardiography (CODE) cohort is a database of ECGs obtained by the Telehealth Network of Minas Gerais (TNMG). Patients over 16 years old had digital ECGs recorded mostly in primary care facilities. Using the first recorded ECG per patient, 1,558,421 ECGs in the dataset were available. Patients were followed up for a mean of 3.68±1.87 years. 52,127 (3.3%) subjects died during the follow-up period. I used an established rhythm/conduction disease classification CNN applied to the CODE dataset to generate the NN-derived ECG features for our study (23). This model is subsequently referred to as CODE-CNN. Using a hybrid ML pipeline (**Figure 3-1**), three phenogroups were identified. The demographic distribution of these 3 phenogroups is shown in **Table 3-1**.

Table 3-1

	Phenogroup A	Phenogroup B	Phenogroup C	P value
CODE dataset				
n	450,336	481,026	627,059	
Age	48.68 (16.81)	58.20 (17.69)	48.78 (16.68)	<0.001
Sex (M)	167,730 (37.2)	238,029 (49.5)	221,283 (35.3)	<0.001
Hypertension	43,060 (9.6)	185,738 (38.6)	263,842 (42.1)	<0.001
Previous MI	999 (0.2)	5,543 (1.2)	5,062 (0.8)	<0.001
Smoker	10,290 (2.3)	34,927 (7.3)	63,598 (10.1)	<0.001
Diabetes mellitus	9,318 (2.1)	35,684 (7.4)	56,468 (9.0)	<0.001
COPD	1,137 (0.3)	4,280 (0.9)	5,849 (0.9)	<0.001
Chagas disease	1,234 (0.3)	21,187 (4.4)	12,169 (1.9)	<0.001
Hyperlipidaemia	5,007 (1.1)	22,332 (4.6)	33,251 (5.3)	<0.001
1st degree AVB	237 (0.1)	20,470 (4.3)	141 (0.0)	<0.001
RBBB	100 (0.0)	37,292 (7.8)	22 (0.0)	<0.001
LBBB	147 (0.0)	20,447 (4.3)	16 (0.0)	<0.001
Sinus tachycardia	52 (0.0)	34,201 (7.1)	116 (0.0)	<0.001
Sinus bradycardia	58 (0.0)	24,467 (5.1)	40 (0.0)	<0.001
Atrial fibrillation	8 (0.0)	20,671 (4.3)	43 (0.0)	<0.001
Whitehall II				
n	3,370	1,696	*	<0.001
Age	64.99 (5.63)	66.48 (5.89)		<0.001

Sex (M)	2,236 (66.4)	1,365 (80.4)		<0.001
Heart rate (bpm)	70.48 (9.47)	62 (13.54)		<0.001
QTc interval (ms)	423.53 (17.59)	425.59 (29.59)		<0.001
QRS duration (ms)	104.85 (8.4)	111.12 (18.28)		<0.001
UK Biobank				
n	25,723	16,466	197	
Age	63.47 (7.60)	65.22 (7.85)	60.37 (7.59)	<0.001
Sex (M)	10,558 (41.0)	9,957 (60.5)	23 (11.7)	<0.001
Heart rate (bpm)	65.67 (8.72)	56.00 (11.22)	74.21 (5.80)	<0.001
QRS duration (ms)	85.42 (9.73)	93.86 (18.10)	80.21 (8.96)	<0.001
PR interval (ms)	156.61 (19.34)	179.41 (32.44)	141.98 (12.96)	<0.001
QTc interval (ms)	423.78 (22.24)	417.31 (30.04)	433.26 (19.37)	<0.001
ELSA Brasil				
n	11,175	2,536	*	
age	51.69 (8.87)	54.83 (9.67)		<0.001
Sex (M)	4771 (42.7)	1477 (58.2)		<0.001
BMI (mean (SD))	27.01 (4.77)	27.07 (4.69)		0.625
Obesity	2571 (23.0)	572 (22.6)		0.643
Previous MI	149 (1.3)	102 (4.0)		<0.001
Heart failure	135 (1.2)	92 (3.6)		<0.001
Previous stroke	139 (1.2)	49 (1.9)		0.009

Hypertension	3,799 (34.0)	1,120 (44.2)		<0.001
Diabetes mellitus	2,218 (19.8)	505 (19.9)		0.964
Coronary artery disease	264 (2.4)	177 (7.2)		<0.001
Hyperlipidaemia	5,549 (49.7)	1,253 (49.4)		0.853
Smoker	1,461 (13.1)	337 (13.3)		0.799
Sami-TROP				
n	286	1,130	215	
Age	60.25 (12.73)	60.04 (12.86)	54.88 (11.60)	<0.001
Sex (M)	80 (28.0)	411 (36.4)	43 (20.0)	<0.001
BIDMC				
n	118602	70245	125	
Age	52.15 (16.99)	61.51 (18.59)	40.47 (16.16)	<0.001
Sex (M)	51633 (43.5)	38832 (55.3)	40 (32.0)	<0.001
Diabetes mellitus	14176 (12.0)	8644 (12.3)	9 (7.2)	0.018
Smoker	7580 (6.4)	3755 (5.3)	7 (5.6)	<0.001
Hypertension	33108 (27.9)	20004 (28.5)	23 (18.4)	0.002
Hyperlipidaemia	29238 (24.7)	16032 (22.8)	20 (16.0)	<0.001
Heart rate (bpm)	74.47 (10.88)	79.61 (27.02)	75.52 (5.70)	<0.001
PR Interval (ms)	153.46 (21.82)	167.93 (42.68)	140.40 (13.35)	<0.001
QRS duration (ms)	89.93 (10.78)	101.42 (23.69)	84.93 (7.15)	<0.001

QTc interval (ms)	426.56 (90.34)	455.13 (500.98)	420.58 (38.54)	<0.001

Survival analysis

The survival analysis demonstrated the 3 phenogroups had different mortality profiles (**Figure 3-2A**). Phenogroup B had a 2.57-fold higher risk of mortality compared to phenogroup A, while phenogroup C had a 15% lower risk (phenogroup B vs A, hazard ratio (HR) 2.57, 95% CI 2.51-2.63, p<0.001; phenogroup C vs A, HR 0.85, 95% CI 0.82-0.87, p<0.001). The prognostic significance of the phenogroups was retained, after adjusting for known covariates (age, gender, known cardiovascular comorbidities and the six ECG diagnoses on which the neural network was trained). In these adjusted analyses, phenogroup B (highest mortality) had a 1.2-fold increase in long-term mortality compared to phenogroup A (HR 1.20, 95% CI 1.17-1.23, p < 0.0001, **Table 3-2**).

Table 3-2

Cox regression models for prediction of mortality (unless otherwise specified)

	Adjusted		
Analysis	HR	95% CI	P value
CODE†			
Cluster B	1.20	1.17-1.23	<0.0001
Cluster C	0.82	0.80-0.84	<0.0001
CODE† – no ECG diagnosis cohort			
Cluster B	1.15	1.12-1.19	<0.0001
Cluster C	0.80	0.78-0.83	<0.0001
CODE† – normal ECG cohort			
Cluster B	0.94	0.80-1.10	0.42
Cluster C	0.95	0.84-1.09	0.49
UKB (MACE) ‡			
Cluster B	1.21	1.01-1.45	0.04
Whitehall II (all-cause mortality) §			
Cluster B	1.61	1.16-2.24	0.005
$\log_2$ (QRS micro-fragmentation)	1.24	0.96-1.59	0.10
Whitehall II (cardiovascular mortality) §			

Cluster B	1.72	0.92-3.18	0.09
$\log_2$ (QRS micro-fragmentation)	1.67	1.06-2.6	0.03
Whitehall II (non-cardiovascular mortality) §			
Cluster B	1.54	1.05-2.24	0.03
$\log_2$ (QRS micro-fragmentation)	1.08	0.80-1.46	0.62
ELSA-Brasil (all-cause mortality) ¶			
Cluster B	1.04	0.86-1.27	0.65
Sami-Trop ‖			
Cluster B	1.50	0.85-2.64	0.16
Cluster C	0.50	0.17-1.54	0.23
BIDMC #			
Cluster B	1.20	1.17-1.24	<0.0001
Cluster C	0.73	0.33-1.63	0.45

CODE no ECG diagnosis cohort: None of the following diagnoses on the ECG: 1st degree AVB, RBBB, LBBB, sinus tachycardia, sinus bradycardia and atrial fibrillation

† Adjusted for age, sex, hypertension, previous MI, smoking status, diabetes mellitus, COPD, chagas disease, hyperlipidaemia, 1st degree AVB, RBBB, LBBB, sinus tachycardia, sinus bradycardia and atrial fibrillation

‡ Adjusted for age, sex, previous MI, smoking status, hypertension, diabetes mellitus, hyperlipidaemia, PR interval, QRS duration, QT interval, QTc interval and heart rate

§ adjusted for age, sex, heart rate, QTc interval, QRS duration, $\log_2$ (QRS micro-fragmentation) and total cosine R to T

¶ adjusted for age, sex, hypertension, previous MI, smoking status, diabetes mellitus, BMI, obesity, heart failure, previous stroke, coronary artery disease and hyperlipidaemia

|| adjusted for age and sex

adjusted for age, sex, diabetes mellitus, smoking status, hypertension, hyperlipidaemic, heart rate, PR interval, QRS duration, QTc interval

I then evaluated the performance of the phenogroups in the group of patients without any of the six diagnoses for which CODE-CNN was originally trained. Importantly, the predictive ability of the phenogroups was retained in this group (n=1412285, **Figure 2B**, **Table 3-2**). I also analysed a subgroup of patients with physician-adjudicated normal ECGs (n = 92138). In this analysis, unadjusted phenogroup B was prognostically significant (phenogroup B vs A, HR 1.65, 95% CI 1.43-1.92, p < 0.0001) (Figure 3-2C) but lost significance after adjustment for covariates (**Table 3-2**).

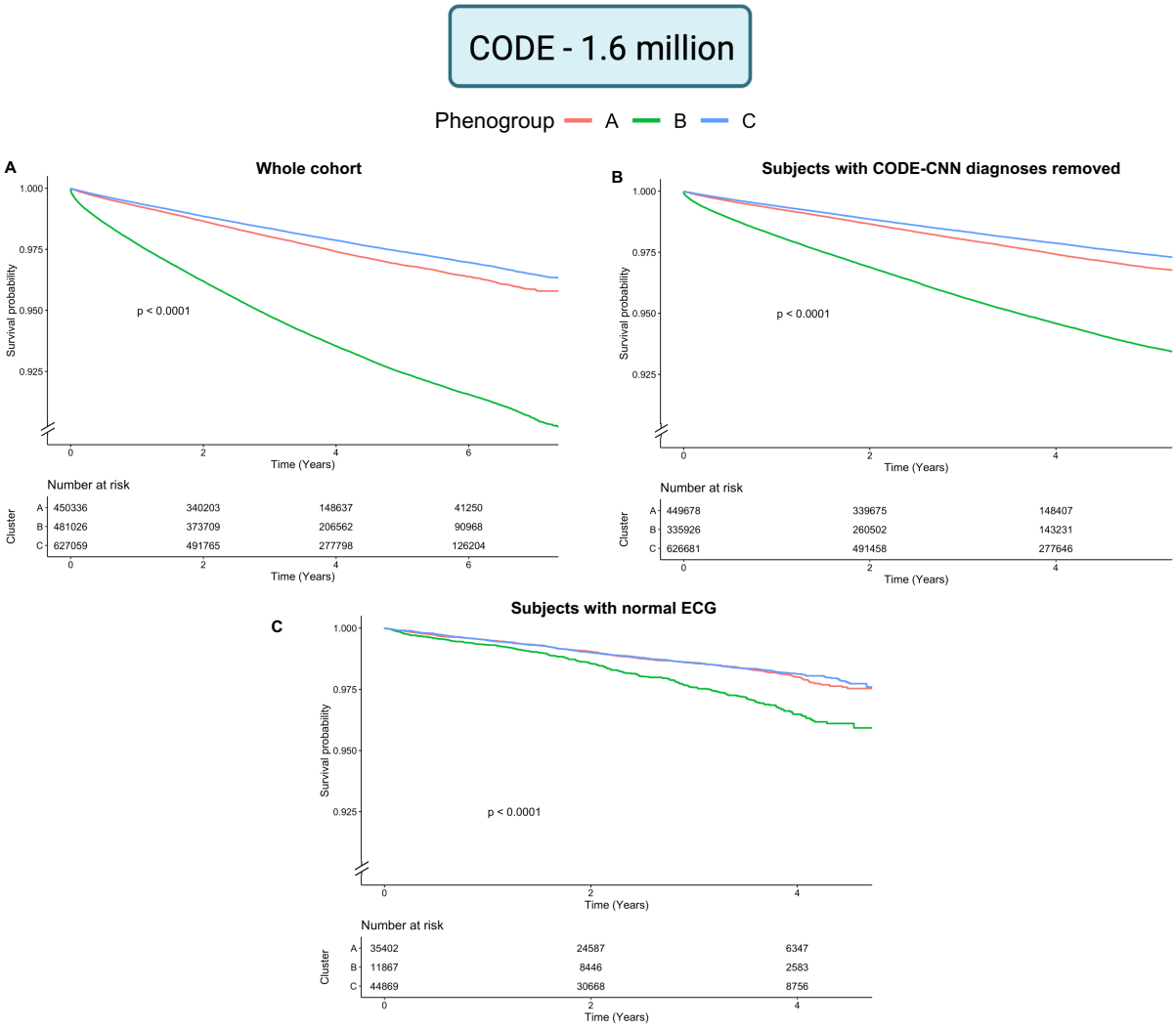


Figure 3-2 Survival analysis in the derivation dataset

The 3 phenogroups have prognostic significance, with phenogroup B having a markedly worse prognosis. Data are shown for: (A) whole Clinical Outcomes in Digital Electrocardiography (CODE) cohort, (B) CODE cohort with removal of subjects with any of the following diagnoses on the ECG: 1st degree atrioventricular block, right bundle branch block, left bundle branch block, sinus tachycardia, sinus bradycardia and atrial fibrillation. (C) Subset of the CODE cohort with normal ECGs.

External testing

I then externally tested our findings in five diverse cohorts. There were 5066 subjects available for analysis in the Whitehall II study. For this analysis events were censored at 5 years of follow-up. 163 (3.22%) subjects died during this period. In the UK Biobank, there were 42386 subjects available for analysis. Patients were censored at the time of their first event and 411 (0.97%) subjects died during the follow-up period of 3.73±1.57 years. Given the comparatively significantly lower mortality rate and the additional data available in the UK Biobank, the primary endpoint for this dataset was specified as MACE (definition described methods). 967 (2.3%) of subjects had a MACE event during follow-up. The ELSA-Brasil cohort had 13,739 subjects available for analysis. Mean follow-up was 9.35±1.28 years and 599 (4.4%) subjects died during follow up. The SaMi-Trop cohort had 1,631 subjects available for analysis. Mean follow-up period was 2.08±0.39 years and 104 (6.38%) subjects died during the follow-up period. Lastly, the BIDMC cohort consists of a secondary care population. 188972 subjects were available for analysis. Mean follow-up period was 5.46±5.81 years and 34851 (18.4%) subjects died during follow-up. The demographics distribution of these cohorts is shown in **Table 3-2**.

The prognostic significance of the phenogroups was demonstrated in all of the external cohorts (**Figure 3-3**). The BIDMC secondary care population had a significantly higher mortality rate than other cohorts and accordingly had the largest absolute risk difference between phenogroups. In the volunteer cohorts (Whitehall II, UK Biobank and ELSA-Brasil), phenogroup C had very limited representation and

therefore was excluded from analysis. This may be due to the relatively selected population enrolled into these cohorts. Cox regression models adjusted for known covariates showed the phenogroups retained their prognostic significance in the CODE, UK Biobank, Whitehall II and BIDMC cohorts, but not ELSA-Brasil. Importantly, this included adjustment for standard ECG parameters in the UK Biobank, Whitehall II and BIDMC cohorts.

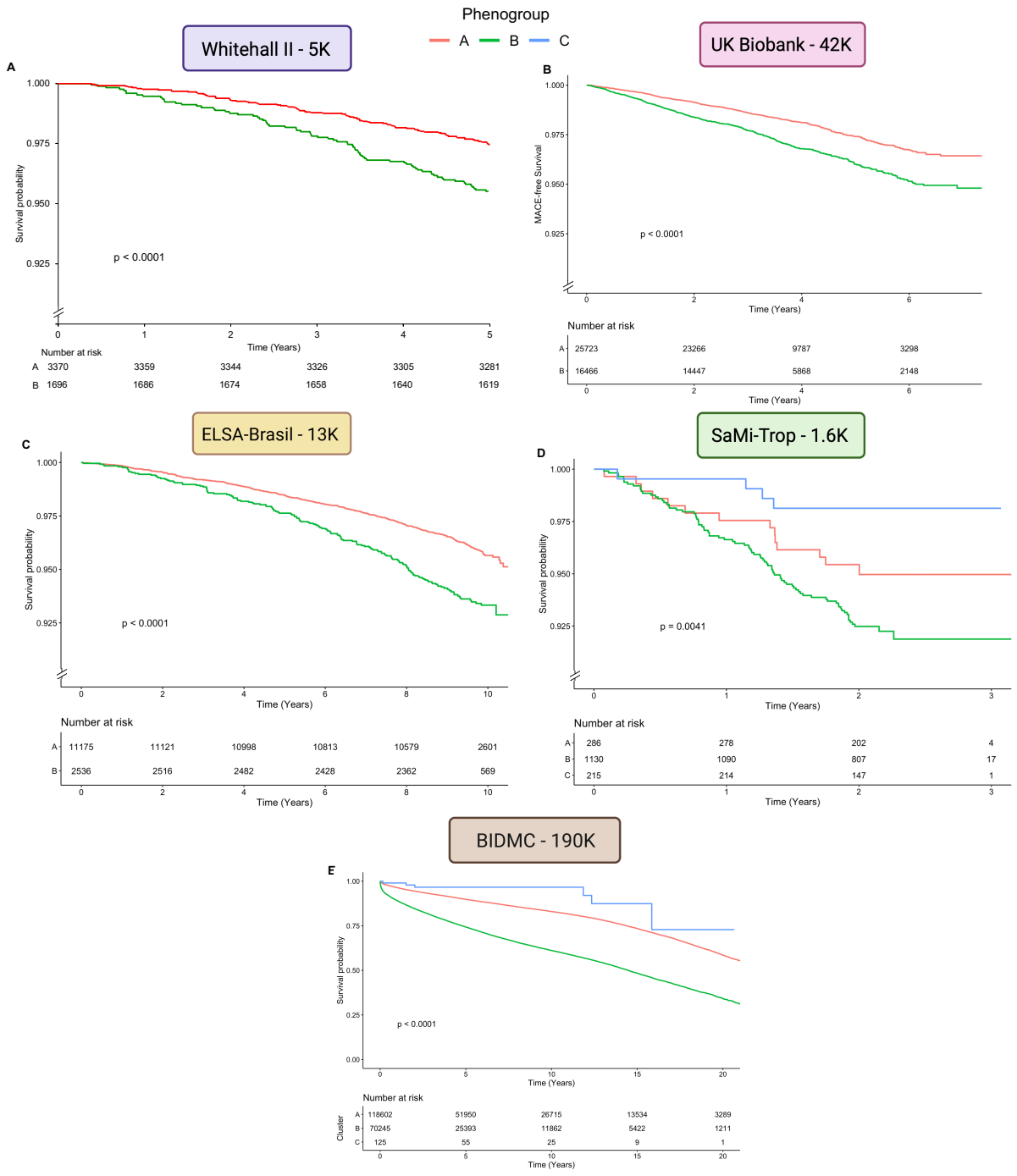


Figure 3-3 Survival analysis in the five external testing datasets

In the volunteer populations (Whitehall II, UK Biobank and Brazilian Longitudinal Study of Adult Health (ELSA-Brasil)) phenogroup C has very few subjects and therefore is excluded. Phenogroup B has a significantly higher event rate. (A) Whitehall II cohort (B) UK Biobank, survival free of major adverse cardiovascular

events is depicted (C) ELSA-Brasil cohort, (D) São Paulo-Minas Gerais Tropical Medicine Research Center (SaMi-TROP) cohort, (E) Beth Israel Deaconess Medical Center (BIDMC) cohort.

To investigate the prognostic significance of the phenogroups in a more specific disease context, the SaMi-Trop cohort was analysed. In contrast to the volunteer cohorts described above (who would be expected to largely have relatively normal ECGs), most patients in the SaMi-Trop cohort had abnormal ECGs (87). In this cohort of patients with chronic Chagas cardiomyopathy, the three phenogroups had marked prognostic significance (**Figure 3-3C**) although the Cox model adjusted for age and gender was not statistically significant (**Table 3-2**).

In BIDMC, UK Biobank and Whitehall II I additionally analysed the association of phenogroup B with cardiovascular (CV) vs non-CV death. In Whitehall II, for all-cause mortality, ECG phenogroup remained a significant predictor of all-cause mortality, after adjustment for age, heart rate, QTc, QRS duration, micro-QRS fragmentation and QRS-T angle (162 events, HR 1.61 95% CI 1.155-2.236, p = 0.005). This finding was significant for non-cardiovascular mortality (114 events, HR 1.54 95% CI 1.04-2.28, p = 0.032). Cardiovascular mortality was less common in this cohort, there was a trend towards statistical significance (48 events, HR 1.72 95% CI 0.92-3.18, p = 0.09). Phenogroup B was associated with a higher HR for CV death compared to non-CV death in both BIDMC and UK Biobank. (BIDMC CV death HR 3.22 (3.07-3.38) p < 0.0001, UK Biobank HR 2.63 (1.78-3.88), BIDMC Non-CV death HR 2.36 (2.31-2.42) p<0.0001, UKB 1.21(0.996-1.47) p 0.055).

I went on to perform survival analysis on the subset of ECGs without any of the original rhythm/conduction diagnoses for which CODE-CNN was trained (**Figure 3-**

4). Phenogroup B had a statistically significant higher mortality in the UK Biobank, Whitehall II and BIDMC cohorts. Phenogroup B trended towards higher mortality but was non-significant in the comparatively smaller cohorts of SaMi-Trop ($p = 0.05$) and ELSA-Brasil ($p = 0.11$).

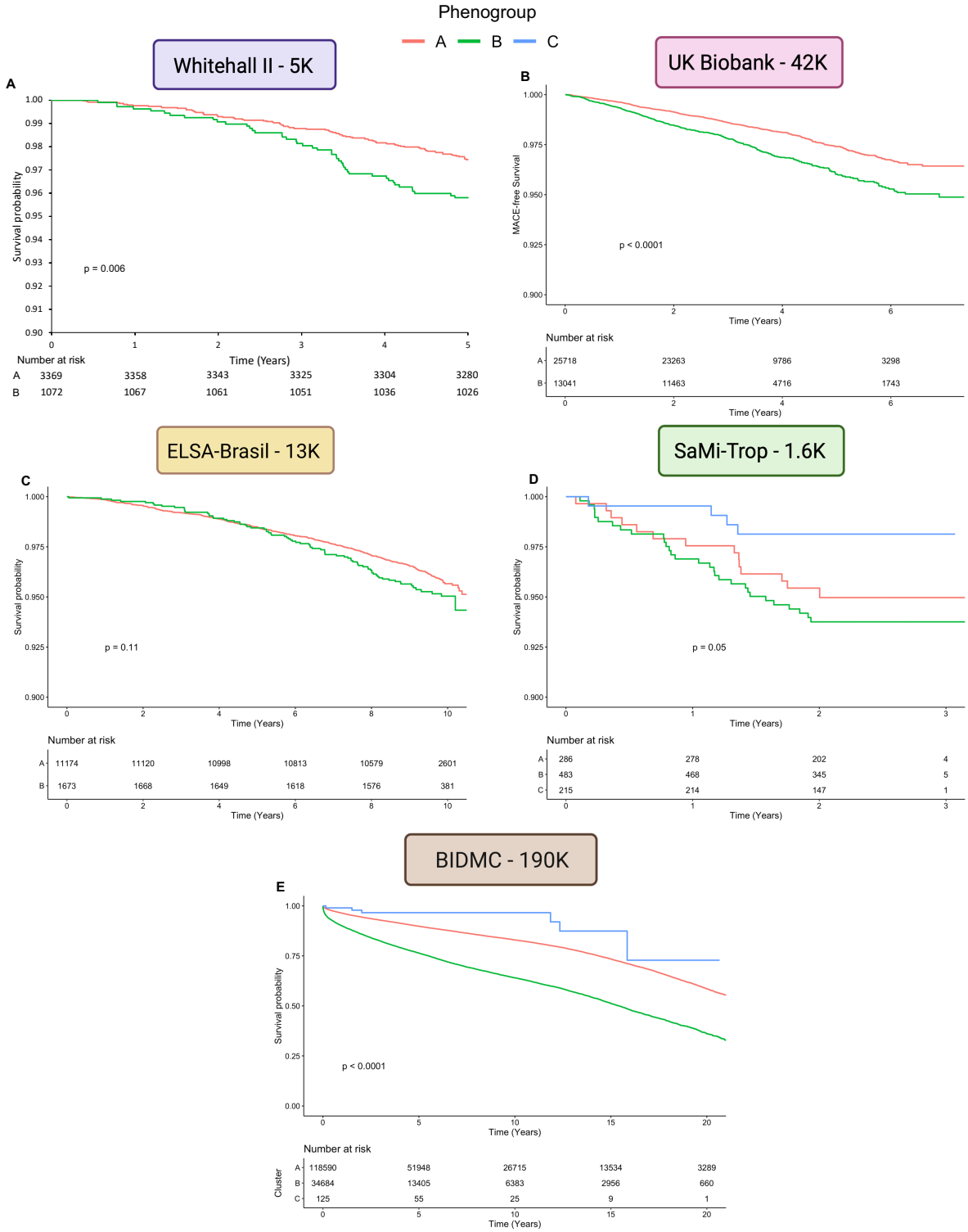


Figure 3-4 Survival analysis in the five external testing datasets – sensitivity analysis

Removal of subjects with any of the following diagnoses on the ECG: 1st degree AVB, RBBB, LBBB, sinus tachycardia, sinus bradycardia and atrial fibrillation. In the

volunteer populations (Whitehall II, UK Biobank and ELSA-Brasil) phenogroup C has very few subjects and therefore is excluded. (A) Whitehall II (B) UK Biobank (survival free of major adverse cardiovascular events is depicted) (C) ELSA-Brasil cohort (D) SaMi-TROP cohort and (E) BIDMC cohort. Phenogroup B has a significantly higher event rate.

Further analysis of Phenogroup C

In order to explore Phenogroup C, which was not well represented in all cohorts I used several approaches. First, using logistic regression I investigated the significant clinical predictors of assignment to phenogroup C vs A. The presence of hypertension (OR 6.92, $p < 0.0001$), younger age (OR 0.98 per year, $p < 0.0001$), and tobacco use (OR 3.78, $p < 0.0001$) were the strongest clinical differentiators. Data on CV risk factors was not available in Sami-TROP, however younger age was a significant predictor of Phenogroup C assignment (OR 0.97 per year, $p < 0.0001$). Phenogroup C therefore represents a younger cohort with increased cardiovascular risk (CV) factors. It is possible that the younger age counteracts the increased CV risk factors, leading to a mortality profile similar to Phenogroup A.

I then investigated the boundaries of the phenogroups in the volunteer cohorts of the UK Biobank and ELSA-Brasil, where data on cardiovascular risk factors was available. Specifically, using the distance to Phenogroup C, I reassigned the third of each cohort that was closest to the Phenogroup C centroid. I found that the reassigned Phenogroup C had a similar mortality profile to Phenogroup A (ELSA Phenogroup C vs A: Unadjusted HR 1.20 (0.997-1.45), p 0.053. UK Biobank Phenogroup C vs A: Unadjusted HR 1.17 (1.00-1.38, $p = 0.048$)), this was consistent

with the mortality profile of Phenogroup C in CODE, which was similar to Phenogroup A.

Next, using logistic regression I investigated the significant clinical predictors of assignment to reassigned Phenogroup C vs A. In these analyses, age had the opposite direction of effect, with older subjects more likely to be reassigned to Phenogroup C ((ELSA OR 1.02, $p < 0.0001$. UKB OR 1.01 per year, $p < 0.0001$). Hypertension (ELSA OR 1.20, $p = 0.37$. UKB OR 1.00, $p = 0.99$) and smoking status (ELSA OR 1.05, p 0.39. UKB OR 1.09, $p = 0.24$) were not significant predictors.

I therefore conclude that Phenogroup C was a population of younger subjects with increased cardiovascular risk factors that was not well represented in all populations. Reassignment of subjects nearest the Phenogroup C centroid showed similar survival properties to Phenogroup A, however the determinants of Phenogroup C assignment in CODE did not hold true in the reassigned Phenogroup C subjects (from Phenogroup A); suggesting these subjects should not be assigned to Phenogroup C, and their original assignments were more suitable.

Phenome-wide association study

In order to evaluate phenotypic associations with the ECG phenogroups, I performed PheWAS using two approaches. Firstly, I evaluated incident disease in the BIDMC cohort (**Figure 3-5A**). I found the high-risk phenogroup (B) was associated with a significantly higher rate of future atrial fibrillation (OR 2.89, $p < 0.00001$), ischaemic heart disease (OR 1.44, $p < 0.00001$), atrioventricular block (OR 4.88, $p < 0.00001$), cardiomyopathy (OR 2.04, $p < 0.00001$), ventricular tachycardia (OR 2.00, $p < 0.00001$) and cardiac arrest (OR 6.68, $p < 0.00001$). Secondly I investigated

phenotypes in the UK Biobank. **Figure 3-5B** shows the Manhattan plot of the univariate correlation P-values and correlation coefficients between ECG phenogroup and non-ECG phenotypes. 512 of 3142 comparisons reached the Bonferroni threshold for significance. These included increased cardiac chamber volumes, which positively correlated with phenogroup B (higher MACE event rate). Left ventricular ejection fraction (LVEF), right ventricular ejection fraction (RVEF) and measures of left ventricular strain were negatively correlated.

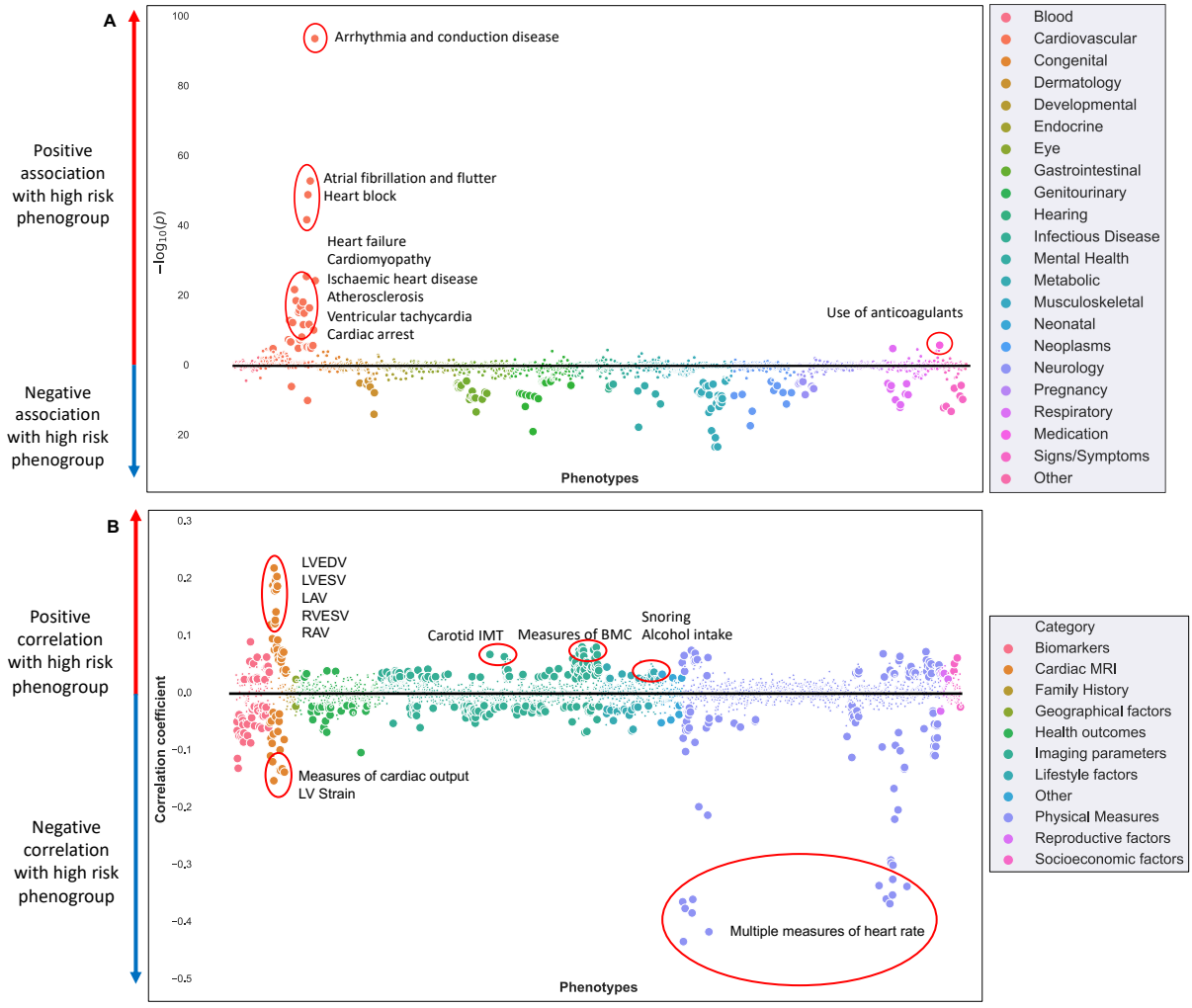


Figure 3-5 Phenome-wide association study

Manhattan plot showing negative logarithm of the univariate correlation P-value between phenotypes for BIDMC disease Phecodes, Disease PheWAS (A) and correlation coefficient for UK Biobank phenotypes, Biobank PheWAS (B). Small points depict associations not reaching statistical significance, while large points crossed the Bonferroni threshold for statistical significance. MRI: magnetic resonance imaging, EDV: end-diastolic volume, ESV: end-systolic volume, RV: right ventricle, LV: left ventricle, RAV: right atrial volume, SV: stroke volume, BMC: bone mineral content, IMT: intima-media thickness

ECG parameters also significantly correlated with phenogroup. Increased QRS duration, PR interval and QT interval, positively correlated with phenogroup B (**Figure 3-6**).

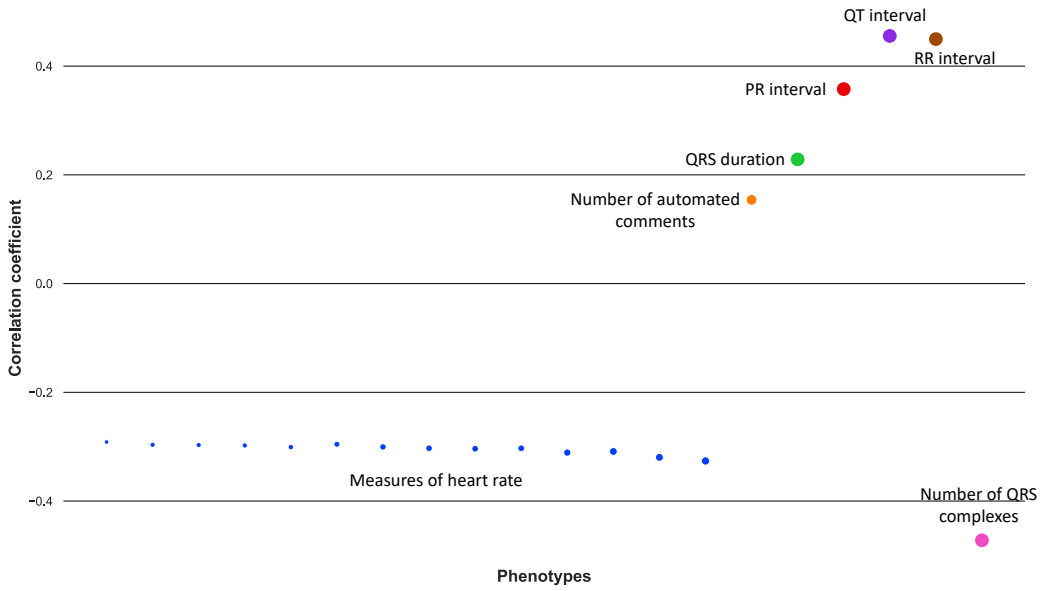


Figure 3-6 UK Biobank ECG associations

Phenome-wide association study (PheWAS) subplot depicting correlation coefficients of ECG parameter results

Genome-wide association study

To identify genotypic associations with the phenogroups, a single-trait genome-wide association study (GWAS) was conducted using 12.7 million variants in 31,119 subjects from the UK Biobank. The GWAS yielded four loci (**Figure 3-7**) which were confirmed through expression quantitative trait loci (eQTLs) for Sodium Voltage-Gated Channel Alpha Subunit 10 (*SCN10A*), Sodium Voltage-Gated Channel Alpha Subunit 5 (*SCN5A*), Caveolin 1 (*CAV1*), and Rho GTPase Activating Protein 24 (*ARHGAP24*). The lead variant for *SCN10A* has been previously significantly

associated with atrial fibrillation and flutter, and cardiac conduction (119). CAV1 has been previously associated with QRS duration, PR interval and QT interval while *SCN5A* and *SCN10A* have previously well-described roles in cardiac rhythm and conduction phenotypes (119-122). *ARHGAP24* has been previously associated with ECG parameters including PR interval, QRS duration and QT interval (123-125), however, our analysis has identified for the first time *ARHGAP24* as a gene associated with a prognostically significant phenogroup. The findings were similar when analysing only ECGs without any of the original rhythm/conduction diagnoses for which the neural network was trained (Figure 3-8)

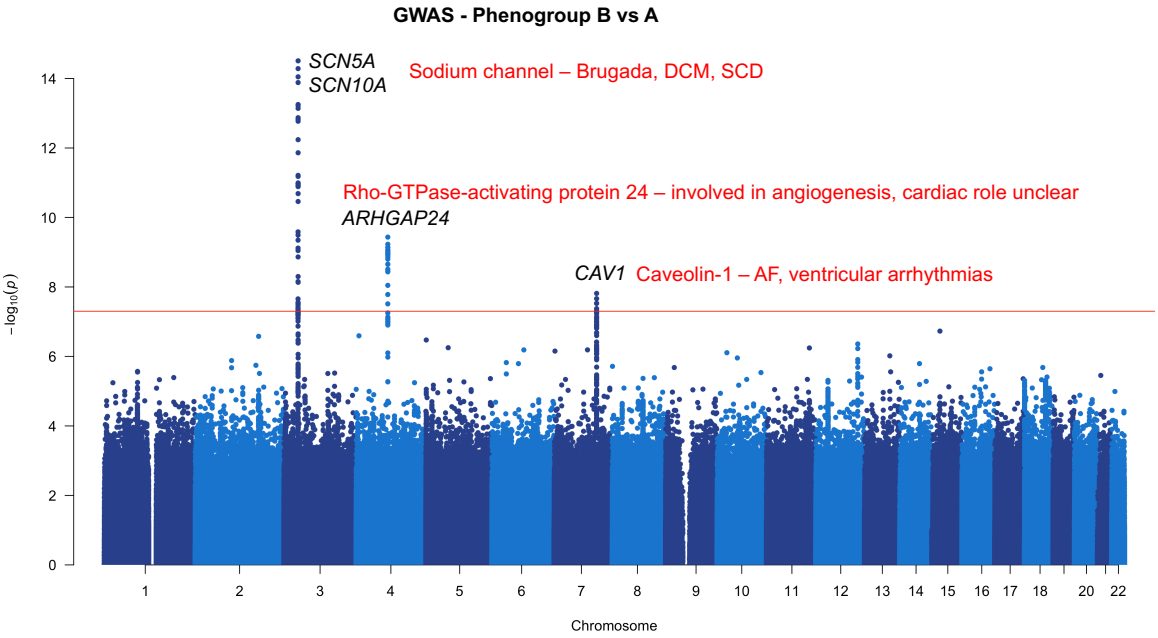


Figure 3-7 Genome-wide association study

Manhattan plots of genomic loci associated with ECG phenogroup. Nearest genes are annotated on the plot. The red line depicts the genome-wide significant threshold ($P<5 \times 10^{-8}$). DCM: dilated cardiomyopathy, SCD: sudden cardiac death, AF: atrial fibrillation

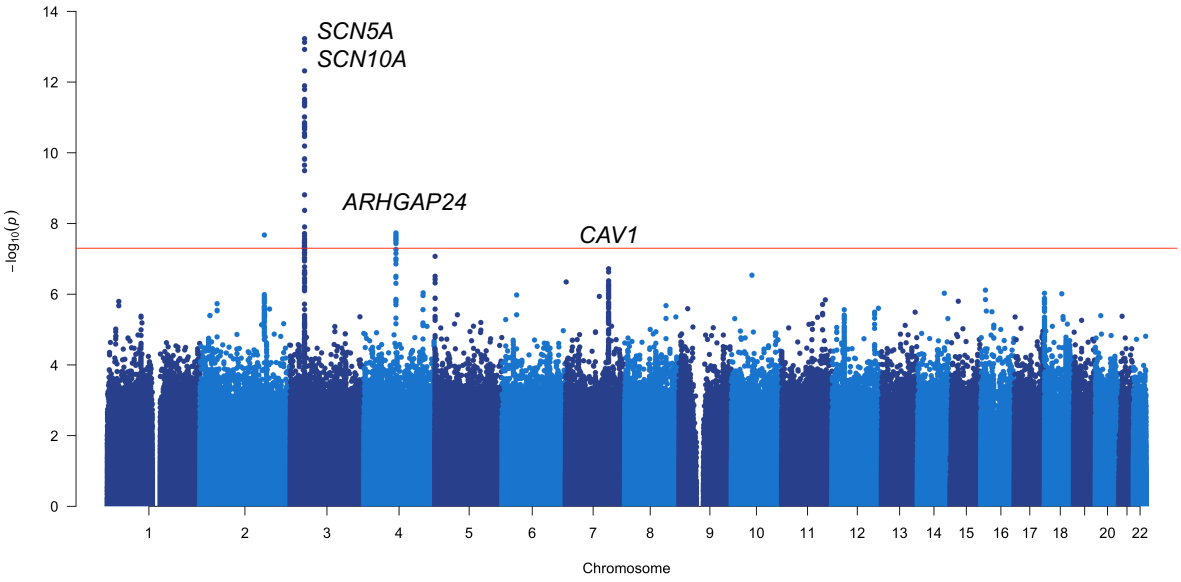


Figure 3-8 Genome-wide association study secondary analysis

Subset of patients without any of the following diagnoses on the ECG: 1st degree AVB, RBBB, LBBB, sinus tachycardia, sinus bradycardia and atrial fibrillation. Manhattan plots of genomic loci associated with ECG phenogroup. SNPs with significant associations with the phenogroups appear more significant on the Y axis (-log10(P-value)). Candidate genes are annotated on the plot. The red line depicts the genome-wide significant line (p = 5 x 10-8).

Model explainability

In order to better understand the reasons for phenogroup classification by the hybrid ML model, I used a modified Gradient-weighted Class Activation Mapping (Grad-CAM) approach (118). **Figure 3-9** shows the average saliency map for 1000 ECGs from the centre of the three phenogroup cluster centroids. The QRS complex was the

most important factor in phenogroup classification. Different parts of the QRS were most important in determining each of the phenogroups, the first part for phenogroup A, terminal part for phenogroup B and middle part for phenogroup C.

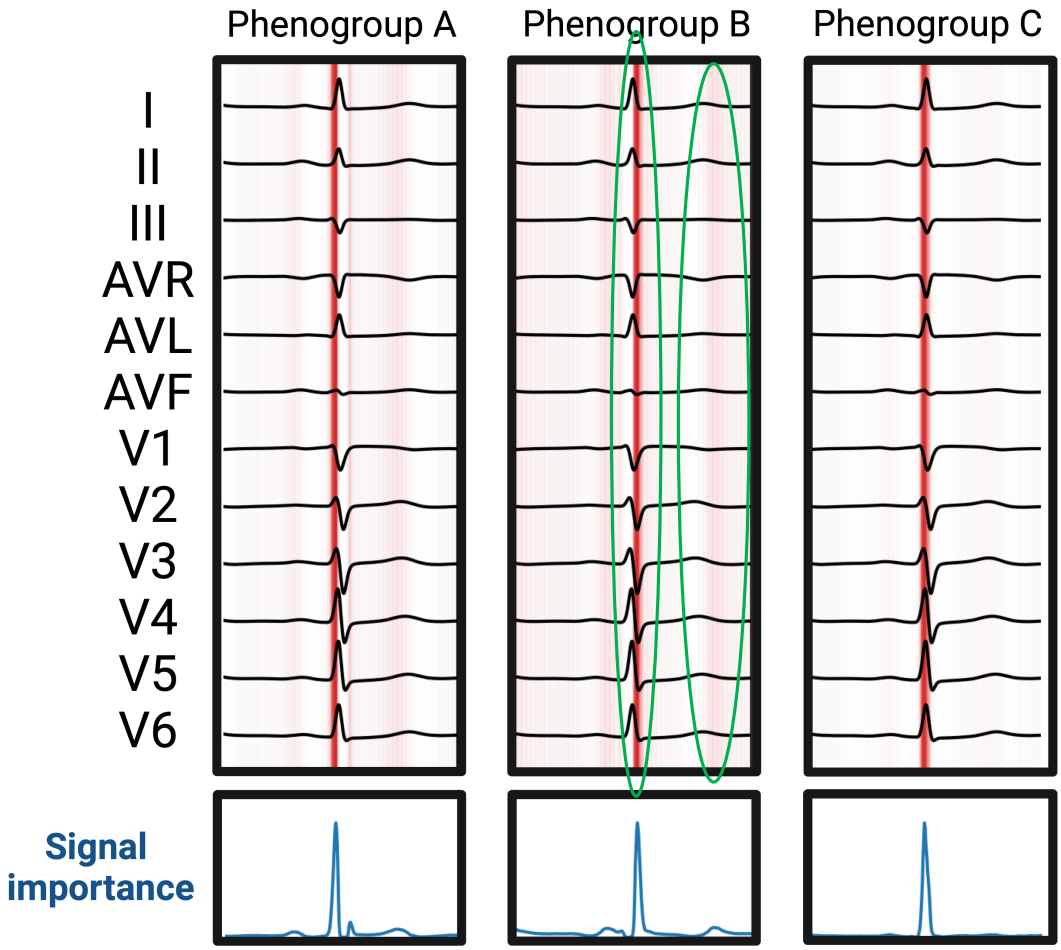


Figure 3-9 Model explainability

Gradient-weighted Class Activation Mapping (Grad-CAM) is used to generate importance maps showing the sections of the ECG signal deemed most important for phenogroup determination. Average saliency of 1000 ECGs from the centre of each cluster is shown. Areas marked with green show the terminal QRS and terminal T wave are important for identification of the high-risk phenogroup (phenogroup B).

Discussion

In this chapter, I explored the application of transfer learning to leverage NN-derived ECG features to identify ECG phenogroups with prognostic significance. I additionally explored their phenotypic and genetic associations. I undertook extensive external testing across a wide range of subjects from six datasets, these included the spectrum of healthy volunteers to patients with cardiomyopathy, normal to abnormal ECGs, North and South American to European – and showed that the prognostic significance of the phenogroups was consistent across all datasets. Using PheWAS, I provide insight into the cardiac and non-cardiac phenotypes associated with the phenogroups. GWAS identified genetic associations primarily related to cardiac conduction and arrhythmia.

Hypothesis driven vs. data driven research

A theme throughout this thesis will be the potential to unravel novel utility for the ECG through data-driven analysis. Although the primary hypothesis is that AI-ECG can be used for cardiovascular risk prediction, it is unknown what the ECG is capable of. In particular it is unclear what diseases may be identified or predicted with AI-ECG. Although supervised machine learning approaches could be applied for all diseases in turn, this may not be a very efficient approach. Unsupervised machine learning on the other hand allows a data-first approach, importantly without anchoring on prior

beliefs on what may or may not be possible with the ECG. AI-ECG advances have shown time and time again that the information content of the ECG goes far beyond what was previously thought (23, 24, 26, 126). However, application of unsupervised machine learning findings can be challenging as clinical significance may not be immediately clear.

Prognostic significance of ECG phenogroups

The first major finding of this chapter is that unsupervised machine learning applied to NN-derived ECG features can be used to derive phenogroups with clear prognostic significance. Importantly this goes beyond traditional risk factors and ECG parameters. Although the model may in part be using existing ECG markers such as QRS duration, QRS micro-fragmentation and bundle branch block morphology (11, 64, 93), in adjusted analyses, I have shown the ECG phenogroups provide additive prognostic information beyond these.

When applied to the three volunteer cohorts (UK Biobank, Whitehall II, and ELSA-Brasil) phenogroup B retained its prognostic significance despite the fact that these cohorts were demographically markedly different to the derivation cohort. In contrast to these volunteer groups, the SaMi-Trop cohort provides an insight into the prognostic significance of the ECG phenogroups in a very different context, in established disease in the form of Chagas Cardiomyopathy, where the three phenogroups show marked separation in mortality curves. This analysis shows the potential for these NN-derived features and phenogroups to be applied to a diverse range of disease specific cohorts and clinical problems, such as a broader

cardiomyopathy cohort or inherited cardiac conditions such as hypertrophic cardiomyopathy.

Disease PheWAS

The disease PheWAS was used to identify the association between the phenogroups and incident disease, and provides important insights into the potential mechanisms for the increased risk of death observed in phenogroup B. As the phenogroups were derived using the ECG and a model developed for identifying rhythm and conduction diseases, it is unsurprising that many of the identified associations relate to arrhythmia and conduction disease. However interestingly, other important associations were identified, including ischaemic heart disease, heart failure and cardiomyopathy. While supervised machine learning models have been used to identify existing heart failure , my findings here suggest the potential to additionally predict future heart failure/cardiomyopathy (24). It is well established that ECG parameters can predict heart block and ventricular arrhythmia ((127, 128), my findings here suggest machine learning could also be applied to predict these diseases. Prediction of occult/future atrial fibrillation from ECGs using deep learning has also been well described and represents an important and potentially clinically transformative tool (26, 67).

UK Biobank PheWAS

By leveraging the extremely deep phenotyping performed in the UK Biobank, I explored the potential biological mechanisms underlying the associations of the ECG

phenogroups with survival. Our PheWAS analysis has identified reduced right and left ventricular ejection fraction and abnormalities in cardiac conduction/repolarisation as potential biological mechanisms underlying the differences in prognosis of the ECG phenogroups. I report several cardiac MRI associations with the higher risk phenogroup, including lower LVEF, RVEF and left ventricular strain that have been previously associated with higher mortality (129-131). In contrast, higher BP and heart rate were negatively correlated with the higher risk phenogroup, while extensive previous literature describes that higher BP and heart rate associate with higher mortality (132, 133). Similarly, I found that the higher risk phenogroup was associated with higher BMC, in contrast to prior literature describing higher mortality with lower BMC (134). These findings may suggest the ECG phenogroups provide distinct information beyond traditional cardiovascular risk factors and non-cardiovascular risk markers.

The higher risk phenogroup was associated with increased PR interval, QRS duration and QT interval. Increased QRS duration and QT interval have been previously associated with higher morality (135, 136) however increased PR interval was not previously found to be associated with increased mortality (137). Our analyses in the UK Biobank and Whitehall II cohorts have shown that ECG phenogroup adds prognostic significance beyond basic ECG parameters.

Genome wide-association study

Phenogroups identified from NN-derived ECG features had significant genetic correlates that provide a plausible mechanistic pathway for the difference in prognosis between phenogroups. The lead variants identified here are in genes with

established roles in cardiac electrophysiology: *SCN5A*, *SCN10A* and *CAV1* (119-122). *SCN5A* and *SCN10A* encode the sodium channels Nav1.5 and Nav1.8 respectively. Nav1.5 is responsible for the initiation of the cardiac action potential via the inward sodium current. Mutations in *SCN5A* and *SCN10A* have been identified as a cause of inherited syndromes such as Brugada syndrome (138, 139) and dilated cardiomyopathy (140). A role in heart failure and sudden death in ischaemic heart disease has also been described (140, 141). *CAV1* encodes caveolin-1 which is a cytoplasmic membrane-anchored scaffolding protein that co-localises to the atria and has been linked to atrial fibrillation, ECG parameters and ventricular arrhythmias (122, 142, 143). *ARHGAP24* encodes Rho-GTPase-activating protein 24 which is a key regulator of angiogenesis and is involved in cell polarity, cell morphology and cytoskeletal organization, but does not have an established role in the heart (144). The association between *ARHGAP24* and ECG parameters has been previously described, however the mechanism is unclear (123-125). In my study, I have shown a novel association between *ARHGAP24* and a prognostically significant ECG phenogroup. My findings suggest that propensity to conduction system disease and arrhythmia may lie on the causal pathway between these variants and the increased mortality seen in phenogroup B.

Comparison to previous machine learning approaches for ECGs

Other studies have used supervised machine learning to predict mortality (72, 145). This approach broadly involves mortality labels being provided to the machine learning model, which can then ‘learn’ the important features predictive of mortality. In contrast, an unsupervised approach involves the machine learning model

clustering subjects based on features without knowledge of the mortality label. Given the ability of unsupervised or self-supervised models to 'learn' from massive amounts of data without knowledge of outcome labels, there is significant potential to apply these methods to clinical data, where outcome labels may be lacking.

Unsupervised machine learning has also previously been applied to ECGs for various applications. Recent studies have used a variational auto-encoder to evaluate a median ECG signal and compress it to 32 factors, these were then used to perform classification tasks using a supervised machine learning model or traditional statistics (146, 147). In this case, unsupervised machine learning was used only for feature extraction compared to my novel approach that successfully identifies clinically relevant phenogroups with no mortality labels used in model training. In contrast, Feeny et al applied unsupervised machine learning to the QRS waveforms themselves in patients who had cardiac resynchronisation therapy (CRT) devices implanted for heart failure and identified two prognostically important subgroups, with significance beyond the commonly used parameters of QRS duration and left bundle branch block. Their method of feature extraction used principal component analysis, compared to the deep learning approach used in my study. Our deep learning approach has the potential to identify more complex, non-linear, relationships between elements of the ECG.

Potential clinical applications of NN-derived ECG features

Using a hybrid machine learning approach, I have described for the first time a novel method to identify high-risk subjects from the ECG alone. Through robust and diverse external testing, I have demonstrated the relevance of the ECG phenogroups

in a wide range of clinical contexts. The addition of ECG phenogroup may significantly improve the performance of risk prediction algorithms, and therefore aid management decisions.

NN-derived ECG features could potentially be applied to risk stratification of a general population to identify high-risk individuals for more intensive investigation and follow-up and lower risk individuals who may be reassured. In disease-specific cohorts such as heart failure and cardiomyopathy, NN-derived ECG features may be applied to assess risk and guide more intensive medical therapy, implantable devices, or transplantation consideration. These potential applications however require further evaluation in the appropriate populations, in prospective studies.

Limitations

I have described the novel application of NN-derived ECG features for risk prediction, and externally tested my findings in a diverse population.

The volunteer cohorts of the UK Biobank, Whitehall II and ELSA-Brasil are selected populations. This may explain the absence of phenogroup C in these cohorts. This however does not detract from the phenogroup performance in these cohorts and in fact strengthens the findings given the diverse groups in which they were tested.

Detailed explainability for deep learning models is challenging. Although some information can be obtained through importance maps, we are unable to provide detailed understanding into the reasons for phenogroup classification. Despite the interesting findings in this chapter, it is important to acknowledge a key limitation of this work - the challenge of clinical implementation. An effectively binary classification of two groups without more detailed risk stratification may be difficult to act on, and

does not provide the precision that more granular approaches may yield. This is an important aspect that I aim to explore in the subsequent chapters of this thesis.

Conclusion

In this chapter I described the use of NN-derived ECG features, derived from an established AI-ECG model trained to identify 6 common diagnoses, to identify prognostically-significant phenogroups from the 12-lead ECG. I explored the biological basis underlying the difference in prognosis between the phenogroups, and identified phenotypic and genotypic associations through PheWAS and GWAS. I tested my findings in five external datasets across three continents and diverse patient populations. I conclude that NN-derived ECG features have important applications beyond the original model from which they are derived and may be transferable and applicable for risk prediction in a wide range of settings, in addition to mortality prediction.

Chapter 4 - Supervised AI-ECG for mortality prediction

Introduction

Parts of the work in this chapter of the thesis have appeared in the pre-print Sau et al, Artificial intelligence–enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform, Lancet Digital Health (accepted, in press)

From phenotypes to personalised predictions

In chapter 3 I have described how AI-ECG can be used to identify a phenogroup at high risk of cardiovascular disease and death. However, assignment of a subject to one of two or three phenogroups does not provide much detailed information on which alter management. A more personalised, continuous prediction would be helpful.

Recent studies have demonstrated the remarkable predictive capabilities of AI-ECG models for not only predicting future cardiac diseases, but also death (66, 73, 148, 149) . Novel AI-ECG biomarkers, such as AI-ECG predicted age and BMI (75, 76), have also been shown to be capable of capturing information on future risk. However, although these existing mortality prediction models have good performance metrics, they are challenging to integrate into routine clinical care. Existing mortality

prediction models are limited by prediction of survival at one or a small number of set time points, rather than providing an individualised survival prediction over time. Just as the Cox proportional hazards model is superior to logistic regression when time-to-event information is available (101), survival-based deep learning may outperform classification-based approaches.

Actionable predictions

A key facet of risk prediction is that identifying that an individual is at an elevated risk of developing disease is not helpful unless there is a medical or lifestyle change that can alter this trajectory. Another key limitation of existing models is the lack of information for clinicians on specific actionable pathways. A high-risk prediction is unhelpful to a clinician if there is no accompanying information on how to affect the survival trajectory of their patient. To make AI-ECG predictions more actionable, it is essential to consider factors such as time-to-event predictions and specific predictions for diseases with established preventive and disease modifying treatments. Disease specific predictions are addressed in Chapter 5.

Explainability and biological plausibility

Furthermore, the adoption of AI into clinical practice is significantly limited by concerns regarding explainability and biological plausibility. Just as knowledge of the mechanisms of action of drugs are important for physicians to have confidence in their application, biological plausibility of AI predictions ensures their credibility and acceptance.

AI-ECG risk estimation platform (AIRE)

To address these limitations of existing risk prediction models, I aimed to develop and test the AI-ECG risk estimation platform (AIRE), that was not only accurate, but also actionable, explainable, and biologically plausible. In this chapter I describe my approach to mortality prediction, including developing a novel approach using a discrete time survival loss function developed by Gensheimer & Narasimhan (102) in combination with a common CNN architecture design, the residual network (ResNet). The AIRE platform constitutes models for mortality prediction and several important cardiovascular diseases. This chapter focuses on mortality, while **Chapter 5** discusses the other models relating to cardiovascular diseases.

Methods

In this chapter I first developed the AI-ECG risk estimation (AIRE) model for prediction of all-cause mortality. I subsequently fine-tuned the model on a primary care population (AIRE-primary care) and created a single lead model that could be used in remote monitoring or inpatient telemetry devices. An overview of models developed in this chapter is provided in **Figure 4-1.**

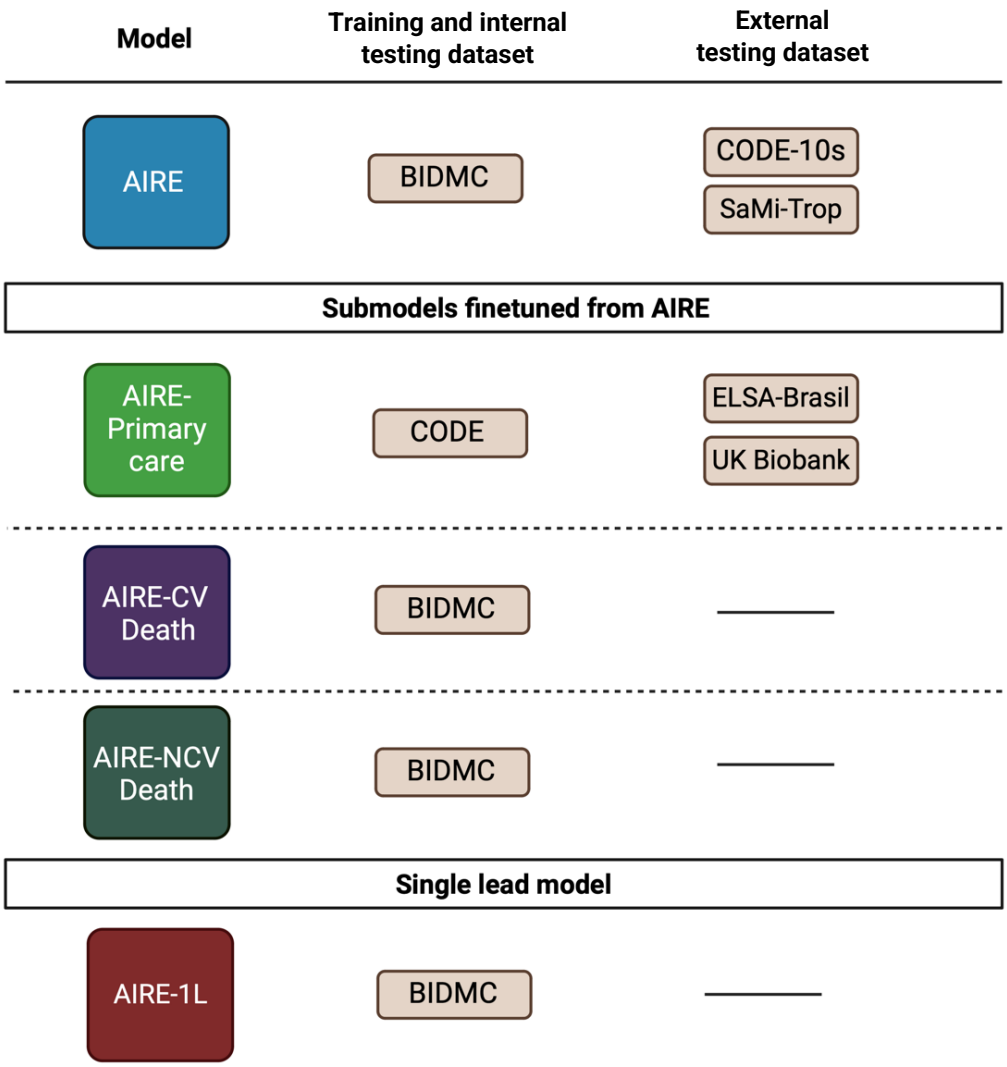


Figure 4-1 AIRE platform overview

Schematic depicting all the mortality prediction models in the AIRE platform, training datasets and testing 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 10s duration. CV: cardiovascular, NCV: non-cardiovascular, BIDMC: Beth Israel Deaconess Medical Center, CODE: Clinical Outcomes in Digital Electrocardiography, SaMi-TROP: São Paulo-Minas Gerais Tropical Medicine Research Center, ELSA-Brasil: The Brazilian Longitudinal Study of Adult Health.

Cohorts

I studied five cohorts, full details in the Methods chapter. Briefly, the Beth Israel Deaconess Medical Center (BIDMC) cohort is a secondary care dataset comprised of routinely collected data from, Boston, USA. The São Paulo-Minas Gerais Tropical Medicine Research Center (SaMi-Trop) is a cohort of patients with chronic Chagas cardiomyopathy (86). The Clinical Outcomes in Digital Electrocardiography (CODE) cohort is a Brazilian database of ECGs recorded in primary care (75) containing ECGs of both 10s and 7s duration. The subset of this dataset with only 10s ECGs is referred to as CODE-10s. The Longitudinal Study of Adult Health (ELSA-Brasil) cohort consists of Brazilian public servants (88). The UK Biobank (UKB) is longitudinal study of volunteers (89). Further details are in **Chapter 2.**

AI-ECG risk estimation platform development

I developed the AIRE platform using the BIDMC cohort as the derivation dataset. The data was split at a ratio of 50/10/40% for training, tuning and internal test, respectively as described in Chapter 2. For mortality end-points, ECGs without paired life status at 30 days were excluded as such a short duration of follow up without an event was felt to not contribute useful information for model training. Data was split by patient ID stratified by presence of ECGs with paired 5-year life status. To prevent data leakage, a single subject could have ECGs assigned to only one of training/tuning/testing datasets. I used a previously described convolutional neural network architecture based on residual blocks (23) and adapted the final layer to accommodate a discrete-time survival model (102). The discrete-time survival approach allows the model to account for both time to outcome (mortality) and

censorship (i.e., loss to follow up). Unlike other models trained to predict mortality at one, or a small number of time points, the model predicts probability of survival between discrete timepoints (49 timepoints were defined) up to 10 years and accounts for right censored data during model training.
The lead I model, AIRE-1L, was developed using the same methodology above, using the same BIDMC data split but using just lead I as the model input.

Fine-tuned sub-models

Using the CODE dataset, I finetuned the model to be more representative of a primary care population (AIRE-primary care) with a much lower risk of adverse events. I used 75% of the CODE dataset for finetuning, with 5 % as a tuning set. The final 20% was used for internal testing of AIRE-primary care. I additionally created submodels for the prediction of CV-death and Non-CV death, these models were called AIRE-CV-death and AIRE-Non-CV-death respectively. Model fine-tuning was applied by loading AIRE models weights and training at a low learning rate (0.000001-0.00001 depending on the model). The model with the lowest tuning loss was selected for evaluation on the test set.

Comparison of AIRE with other models

I compared AIRE and AIRE-primary care to the previously described AI-ECG predicted age (75). AI-ECG predicted age values are publicly available for CODE15 and SaMi-Trop. The remaining 85% of CODE was used to develop the AI-ECG age

model, and therefore was not used. For ELSA-Brasil, the model weights were downloaded (94) and used to derive AI-ECG predicted age values.

I also compared AIRE-primary care and to the recently described SEER (Stanford Estimator of ECG Risk), which had a similar goal of predicting cardiovascular mortality and ASCVD (66). The model code and weights were downloaded (150) and performance evaluated in the UKB, as this was a dataset external to AIRE-primary care and SEER, with cause of death event data available.

Survival and statistical analyses

In the test set, I generated predictions of all ECGs for the primary analyses of all-cause mortality. Sensitivity analyses including a single random ECG per subject were also performed. Model discrimination performance was reported using the C-index, and time-dependent AUROC. For analyses requiring a single predictor value (C-index and AUROC), the probability of survival at 5 years was used. Risk quartiles were defined (low, intermediate-low, intermediate-high, high) based on values in the tuning set. Given the diverse populations and event rates evaluated, risk quartiles were redefined in each dataset. In each case, where 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 using the log rank test.

Cox models were fit using the test dataset comparing demographics, clinical variables, imaging parameters and AIRE platform predictions. For the Cox models incorporating AIRE platform predictions, all model outputs (i.e., 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. Complete case analysis was used; therefore, no variables were imputed. Nested Cox models were compared with the Likelihood Ratio test, while non-nested Cox models were compared with the Partial Likelihood Ratio test (151).

Calibration is a measure of how the predicted probabilities of a model are reflected by the observed rate of each outcome. Calibration was assessed using the Brier score.

Statistical analyses were performed with R 4.2.0 statistical package (R Core Team, Vienna, Austria) or Python (version 3.9).

Diagnostic and imaging data

ICD-9 and ICD-10 codes were used to define presence/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 as previously described (66, 152). In the UKB, cause of death was ascertained based on the ICD10 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. The pooled cohort equation was calculated using the PooledCohort R package. Medication usage was not available in the BIDMC cohort, therefore ICD9 and ICD10 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 (BP)

readings taken within 180 days of the ECG were averaged. Sensitivity analyses were performed using 90 days and 30 days results.

I investigated the performance of AIRE in the high-risk disease groups of severe aortic stenosis and primary pulmonary hypertension. Severe aortic stenosis was defined based on the reported overall severity on echocardiography reports (i.e. a subjective overall assessment by the clinician undertaking the echocardiogram). Primary pulmonary hypertension was defined using ICD9 and ICD10 codes. No fine-tuning was performed when evaluating the performance of AIRE in the severe aortic stenosis and primary pulmonary hypertension disease groups.

Explainability

In order to understand the ECG morphologies associated with predicted survival I used two approaches. Median beats were extracted using the BRAVEHEART ECG analysis software as previously described (108).

First, a variational autoencoder (VAE) was trained as described in **Chapter 2**. In preliminary analyses, models based on only the VAE latent features were found to be inferior to the supervised deep learning approach described above, therefore the VAE was used for explainability only, and not used for AIRE model training or any of the prediction models described in this manuscript. VAE latent features were input into a linear regression with predicted survival as the output. The top 3 most important features as assessed by the t-value were visualised by latent traversal (146).

Second, using the median beats I calculated the average waveform from the 10,000 ECGs with the lowest and highest AIRE predicted mortality. The mean and standard deviation of these waveforms was then plotted.

PheWAS

To better understand the biology underlying AIRE and to explore the detailed mortality prediction associations, I performed PheWAS. I performed PheWAS analysis in the UKB, which contains data from over 3000 phenotypes derived from patient measurements, surveys, and investigations. Univariate correlation was performed to investigate the association between ECG phenogroup and phenotypes, adjusted for age, sex and age squared. I additionally investigated the association of predicted survival with continuous echo traits in the BIDMC dataset. Left ventricular trabeculation was calculated as previously described (110). Deep learning-derived brain age was calculated as previously described (111).

GWAS

To identify genetic associations with the ECG phenogroups, I performed a genome-wide association study (GWAS) in the UKB. As the predicted survival trait was skewed, the data were normalized by rank-based inverse normal transform prior to the analysis. The GWAS analysis was adjusted for the following covariates: age at imaging visit, sex, height, body mass index (BMI), imaging assessment centre and the first 10 genetic principal components. Remaining GWAS methods as described in Chapter 2.

Results

AIRE accurately predicts mortality across diverse timepoints

After model training, predictions for the internal testing data (BIDMC test set, 40% of total dataset) were generated. In the BIDMC cohort, 1,163,401 ECGs were available from 189,539 subjects. Mean follow-up period was 5.46±5.81 years on a per ECG basis, 3.41 (4.08) years taking a random ECG per subject. 34,851 (18.4%) subjects died during follow-up (**Table 4-1**).

Table 4-1

Dataset demographics. Data at the timepoint of a randomly selected ECG per subject is shown for the BIDMC and CODE datasets.

Categorical variables n (%), continuous variables mean (SD)

	BIDMC	**SaMi-TROP**	**CODE**	**ELSA-Brasil**	**UK Biobank**
N subjects	189539	1631	1558421	13739	42386
Age	57.68 (18.69)	59.40 (12.79)	51.66 (17.59)	52.26 (9.10)	64.14 (7.75)
Follow up (years)	3.41 (4.08)	2.08 (0.39)	3.68 (1.87)	9.35 (1.28)	3.73 (1.57)
Sex (M)	90792 (47.9)	534 (32.7)	627042 (40.2)	7483 (54.5)	20538 (48.5)
Hypertension	74409 (39.3)	-	492640 (31.6)	4924 (35.9)	6256 (14.8)
Previous MI	11788 (6.2)	-	11604 (0.7)	251 (1.8)	633 (1.5)
Smoker	23343 (12.3)	-	108815 (7.0)	1801 (13.1)	1111 (2.6)

Diabetes Mellitus	33748 (17.8)	-	101470 (6.5)	2728 (19.9)	1334 (3.1)
Hyperlipidaemia	67087 (35.4)	-	60590 (3.9)	6811 (49.6)	2985 (7.0)
Mortality	34938 (18.4)	104 (6.4)	52127 (3.3)	599 (4.4)	526 (1.2)

Using the 49 discrete time survival outputs, a subject-specific survival curves can be created from only a single ECG that predicts time-to-death. **Figure 4-2A** and **2B** demonstrate two representative subject-specific survival curves in patients who died during follow up and two curves from subjects who survived through the follow up period, each generated by AIRE from a single ECG. **Figure 4-2C** demonstrates the evolution of AIRE-predicted survival based on multiple ECGs performed over several years of follow up. ECGs nearer to the subject's time of death show falling survival probabilities, particularly shortly before the subject's death. Interestingly, when overlaying colours for inpatient and outpatient status we can observe the ability of AIRE predictions to be stable overtime (in the context of relatively stable outpatient settings) but also dynamic during inpatient admissions. Each 'dip' in AIRE predictions corresponds to an inpatient episode, during which survival probabilities are reduced. In the hold out test set, AIRE predicted all-cause mortality with a concordance-index of 0.775 (0.773-0.776). Using the predicted probability of survival at 5 years in the tuning set, quartiles of risk (low, intermediate-low, intermediate-high and high) were determined**.**

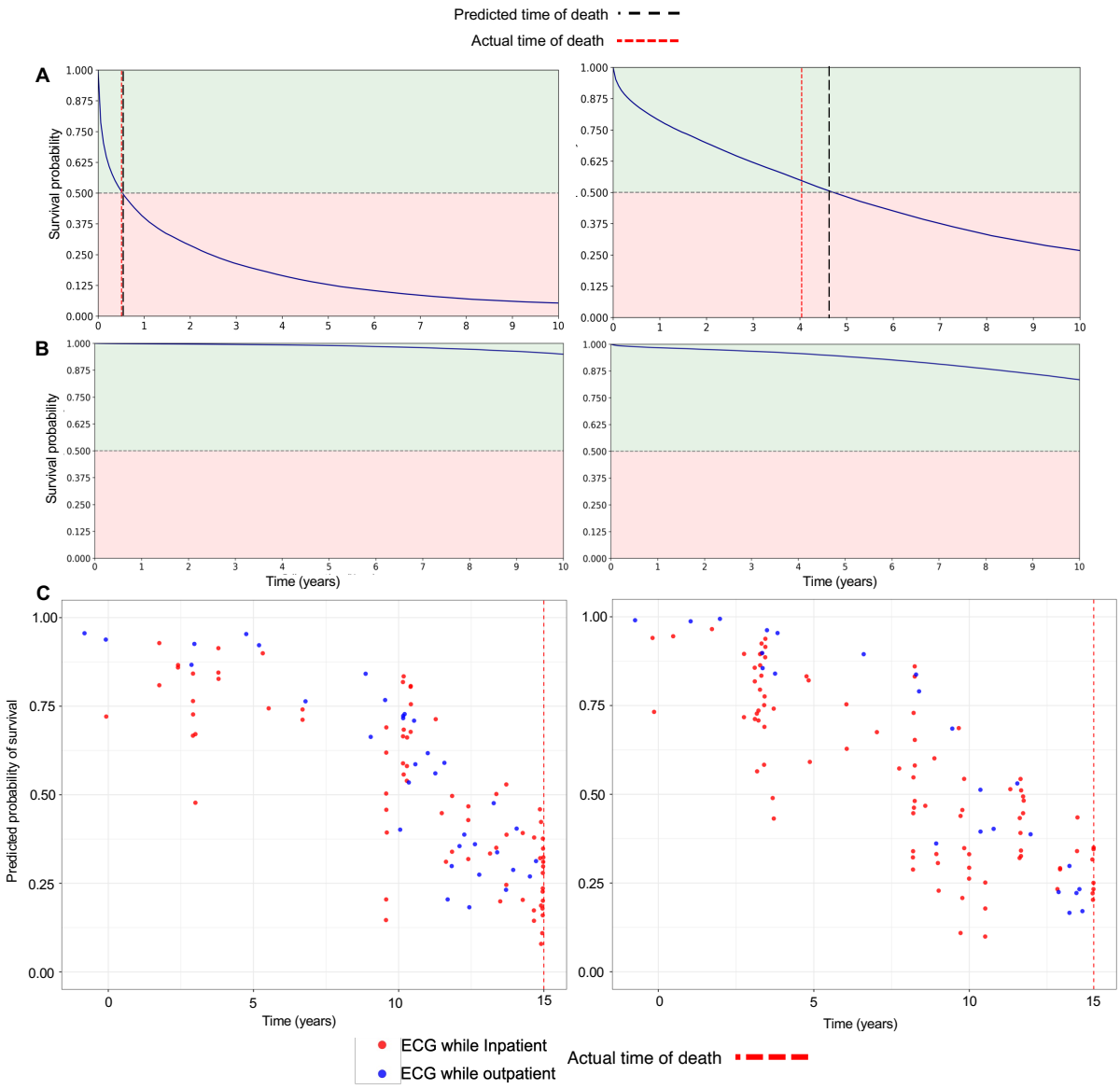


Figure 4-2 Example subject-specific survival predictions

AIRE outputs subject-specific survival curves. Two examples are shown for subjects who died during follow-up (A), and two for subjects who survived through the follow-up period (B). Dashed red lines indicate the date of death and dashed black lines indicate AIRE-predicted date of death. (C) Examples of subjects with many ECGs during the study period, each dot is a survival prediction from a single ECG. Red dots indicate inpatient ECGs, while blue are outpatient. AIRE-predicted survival trends downward over time and predicted probability of survival is particularly low prior to

actual time of death (red dashed line). Inpatient episodes show temporary falls in predicted survival, while outpatient predictions are more stable over time.

Figure 4-3A shows the marked separation of survival curves of these quartiles in the test set.

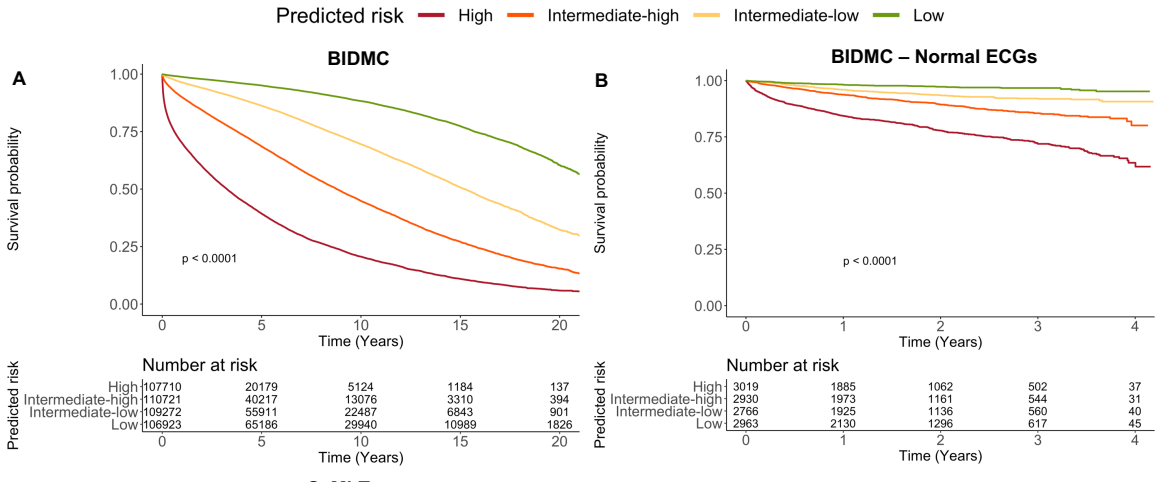


Figure 4-3 Mortality prediction Kaplan-Meier curves

Kaplan-Meier curves of AIRE-predicted all-cause mortality by risk quartile in (A) the whole BIDMC test set, (B) BIDMC normal ECGs

Table 4-2 shows age and sex adjusted hazard ratios for high-risk vs low risk quartile for all cohorts.

Table 4-2

Summary of age and sex adjusted hazard ratios (high risk quartile vs low risk)

	HR (95% CI)
BIDMC Test set	8.73 (8.54-8.91)
BIDMC Test set Normal ECGs	7.50 (5.86-9.60)
SaMi-Trop	16.59 (5.12-53.21)
CODE-10s	4.38 (4.15-4.63)
CODE Test set	5.32 (4.90-5.78)
CODE Test set Normal ECGs	5.03 (3.26 – 7.76)
ELSA-Brasil	3.15 (2.29-4.25)
ELSA-Brasil Normal ECGs	2.36 (1.49 – 3.76)
UK Biobank	1.52 (1.11 – 2.09)

In sensitivity analysis including only patients resident in Massachusetts, (for whom death status would be more accurate) model performance was unchanged (C-index 0.775 (0.773-0.776).

Time-dependent area under the ROC (AUROC) curves were calculated for the BIDMC test set and are shown in **Figure 4-4**. These demonstrate the performance of AIRE over the 10-year period evaluated (mean AUROC 0.827). This shows the short-term performance (<6 months) is high than medium to long term performance, with a slight reduction in performance after 7 years. Calibration was assessed using the Brier score as shown in **Table 4-3.** In contrast to AUROC, calibration performance was reduced at longer term periods.

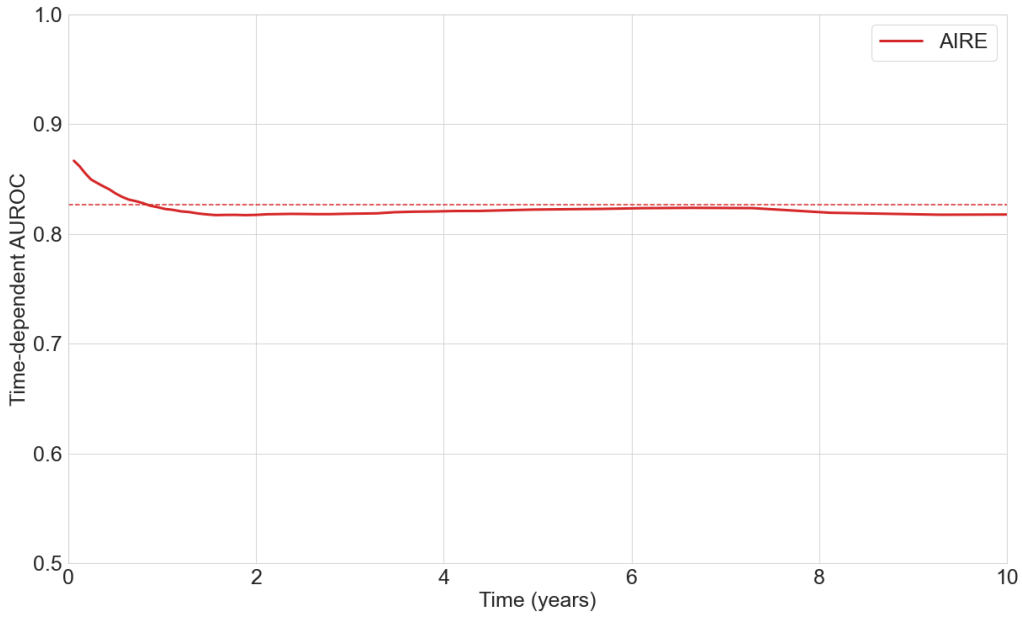


Figure 4-4 Time-dependent receiver operating characteristic curve of AIRE predictions over 10 years

Table 4-3

Model calibration at specific timepoints

Timepoint	Brier Score
1 year	0.130 (0.129-0.130)
5 years	0.160 (0.159-0.160)
10 years	0.255 (0.254-0.256)

Although the main aim of this study was to provide survival predictions at the full range of timepoints up to 10 years, I also calculated predictions at the primary endpoint of 5-years. The AUROC for predicting 5-year mortality was 0.831 (0.829 - 0.832). Using the optimal threshold as determined in the tuning set using Youden's index, the sensitivity and specificity of AIRE at predicting 5-year mortality was 0.756 and 0.747 respectively. The area under the precision recall curve was 0.722.

I additionally evaluated the performance of AIRE across important subgroups. AIRE had similar performance in both men and women and in major ethnic groups (**Figure 4-5A** and **Table 4-3**).

Table 4-4

Summary table of sex and ethnicity specific AIRE performance for mortality prediction in BIDMC Test set

	N Subjects	N ECGs	C-index
Male	30404	223128	0.770 (0.769-0.772)
Female	33701	211501	0.779 (0.777-0.780)
White	41852	293060	0.764 (0.762-0.0.766)
Black	8482	70756	0.793 (0.790-0.796)
Hispanic	3552	25998	0.781 (0.774-0.787)
Asian	3073	16095	0.802 (0.795-0.809)
Other	7146	28720	0.791 (0.785-0.796)

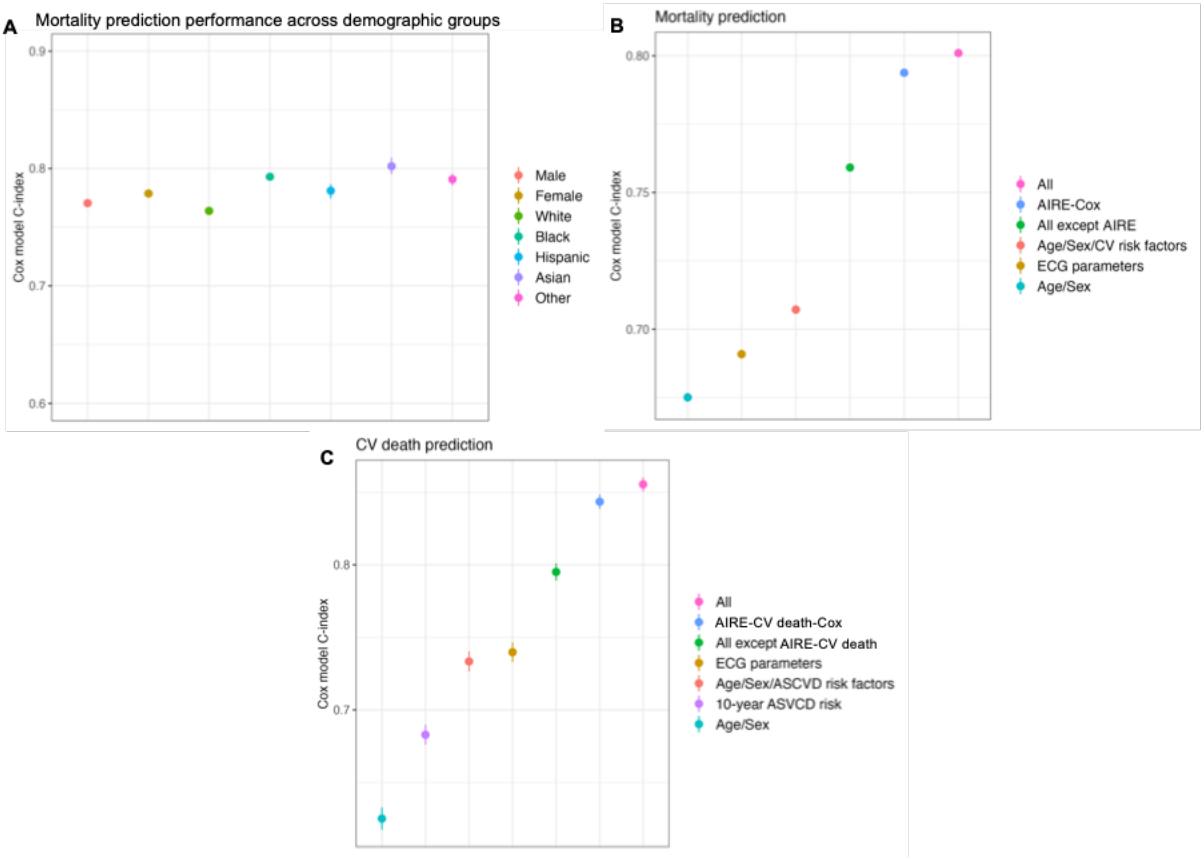


Figure 4-5 Mortality prediction performance

(A) Comparison of AIRE performance across sex and major ethnic groups, AIRE performs well 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

AIRE on normal ECGs

A subset of ECGs had detailed cardiologist interpretations available. These were more recent ECGs with shorter follow up durations compared to the whole cohort. When considering only ECGs labelled as normal by cardiologists, there remained a significant difference in mortality between high-risk and low-risk subjects (**Figure 4-3B, Table 4-2**). In the group of normal subjects, the C-index was 0.735 (0.714-0.746).

AIRE is superior to traditional risk factors for mortality prediction

I compared the ability of AIRE to predict mortality, against using clinical risk factors and risk scores in the BIDMC test set (**Figure 4-5B** and **C**). For these analyses, to mimic potential real-world applications a Cox model using AIRE, age, sex and common ECG parameters (heart rate, PR interval, QRS duration and QTc interval) was created - termed AIRE-Cox. AIRE-Cox had a significantly higher C-index than all other parameters combined (0.794 (0.792-0.795) vs 0.759 (0.758-0.761), $p < 0.0001$). Numerical results for all Cox models are shown in **Table 4-4**.
AIRE-CV Death predicted CV death with a C-index of 0.832 (0.831-0.834). In the BIDMC test set, AIRE-CV Death-Cox had a significantly higher C-index for prediction of CV death than all other parameters combined (0.844 (0.839-0.849) vs 0.795 (0.789-0.801), $p < 0.0001$) (**Figure 4-5**). Finally, AIRE-Non-CV Death predicted non-CV death with a C-index of 0.749 (0.747-0.751).

Table 4-4

Summary table of cox model results

C-index (95% CI) is reported

	Mortality	**CV Death**
Subjects (n)	112434	5807
ECGs (n)	377570	85429
Age/Sex	0.675 (0.673-0.677)	0.625 (0.617-0.633)
ECG parameters	0.691 (0.689-0.693)	0.740 (0.733-0.747)
Echo parameters	-	-
Age/Sex/ARIC-HF risk factors	-	-
10-year ASVCD risk	-	0.683 (0.676-0.690)
Age/Sex/ASCVD risk factors	-	0.733 (0.726-0.740)
Age/Sex/CV risk factors	0.707 (0.706-0.709)	-
All except AIRE	0.759 (0.758-0.761)	0.795 (0.789-0.801)
AIRE-Cox	0.794 (0.792-0.795)	0.844 (0.839-0.849)
All	0.801 (0.800-0.802)	0.855 (0.851-0.860)

AIRE performance in high-risk disease groups

For model predictions of mortality to be clinically useful, there needs to be specific interventions available to alter the trajectory of patients. Using the BIDMC test set, I investigated the performance of AIRE in two high-risk disease groups with existing risk stratification strategies and effective interventions. The data for high-risk disease group analysis was not available for the other cohorts.

For aortic stenosis, AIRE accurately predicted all-cause mortality, C-index: 0.701 (0.681- 0.721, 1293 ECGs from 449 subjects). AIRE-Cox had a significantly higher C-index compared to all other parameters combined, **Figure 4-6A**, C-index 0.709 (0.688-7.30) vs 0.683 (0.661-0.705), $p < 0.001$).

Similarly, for primary pulmonary hypertension, AIRE accurately predicted all-cause mortality, C-index 0.731 (0.724-0.738, 11741 ECGs from 789 subjects). In a subset with available recent echocardiograms (602 ECGs in 212 subjects, I evaluated AIRE in comparison and addition to other risk parameters. AIRE-Cox was superior all other parameters combined (C-index 0.753 (0.726-0.779) vs 0.700 (0.670-0.730), $p < 0.005$, **Figure 4-6B**). Figure 4-7 shows sensitivity analyses using a single random ECG per subject.

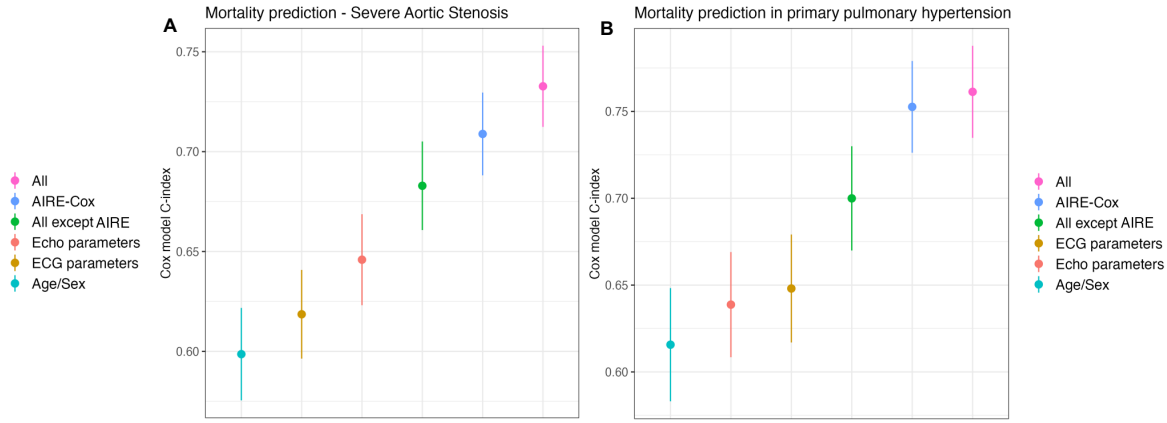


Figure 4-6 Mortality prediction in subgroups of severe aortic stenosis and primary pulmonary hypertension

AIRE was compared with existing risk factors, ECG and imaging parameters in subgroups severe aortic stenosis (A), and primary pulmonary hypertension (B). Echocardiographic parameters for severe aortic stenosis: left ventricular ejection fraction (LVEF), aortic valve area, peak gradient and mean gradient. Echocardiographic parameters for primary pulmonary hypertension: LVEF, LV end diastolic diameter, tricuspid regurgitation (TR) pressure gradient, TR severity, right ventricular (RV) function, RV diameter. ECG parameters: heart rate, PR interval, QRS duration, QTc interval.

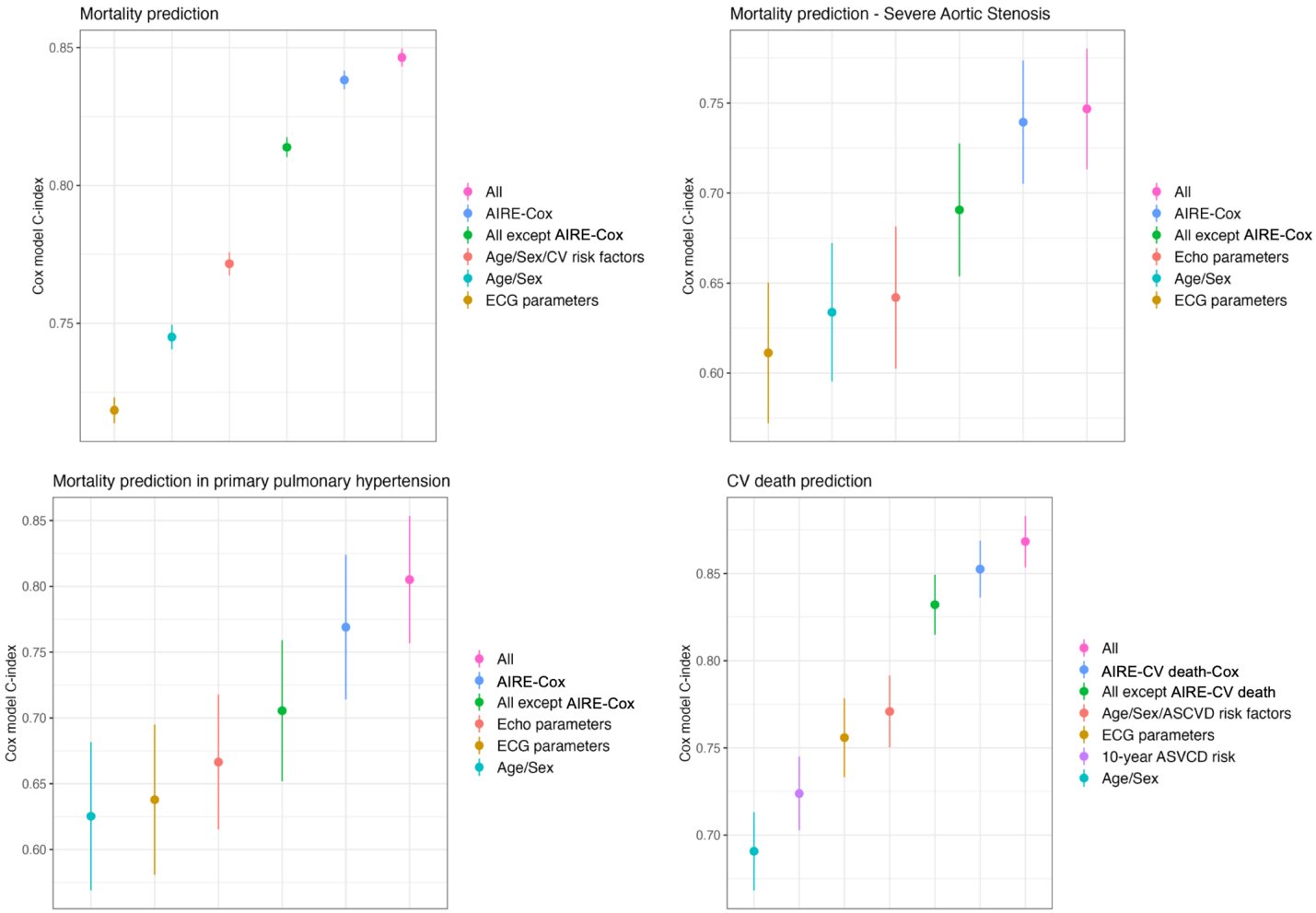


Figure 4-7 Mortality prediction sensitivity analyses

Using Cox models, AIRE was compared with existing risk factors and ECG parameters. Sensitivity analysis using one ECG per subject, selected at random

Single-lead ECG model performance

To evaluate if AIRE could be potentially used in wearable devices or inpatient telemetry units, I trained a single-lead version of AIRE using lead I only (AIRE-1L). The performance of AIRE-1L (C-index 0.751 (0.750-752) **Figure 4-8**) was only slightly inferior in discrimination compared to the 8-lead AIRE model.

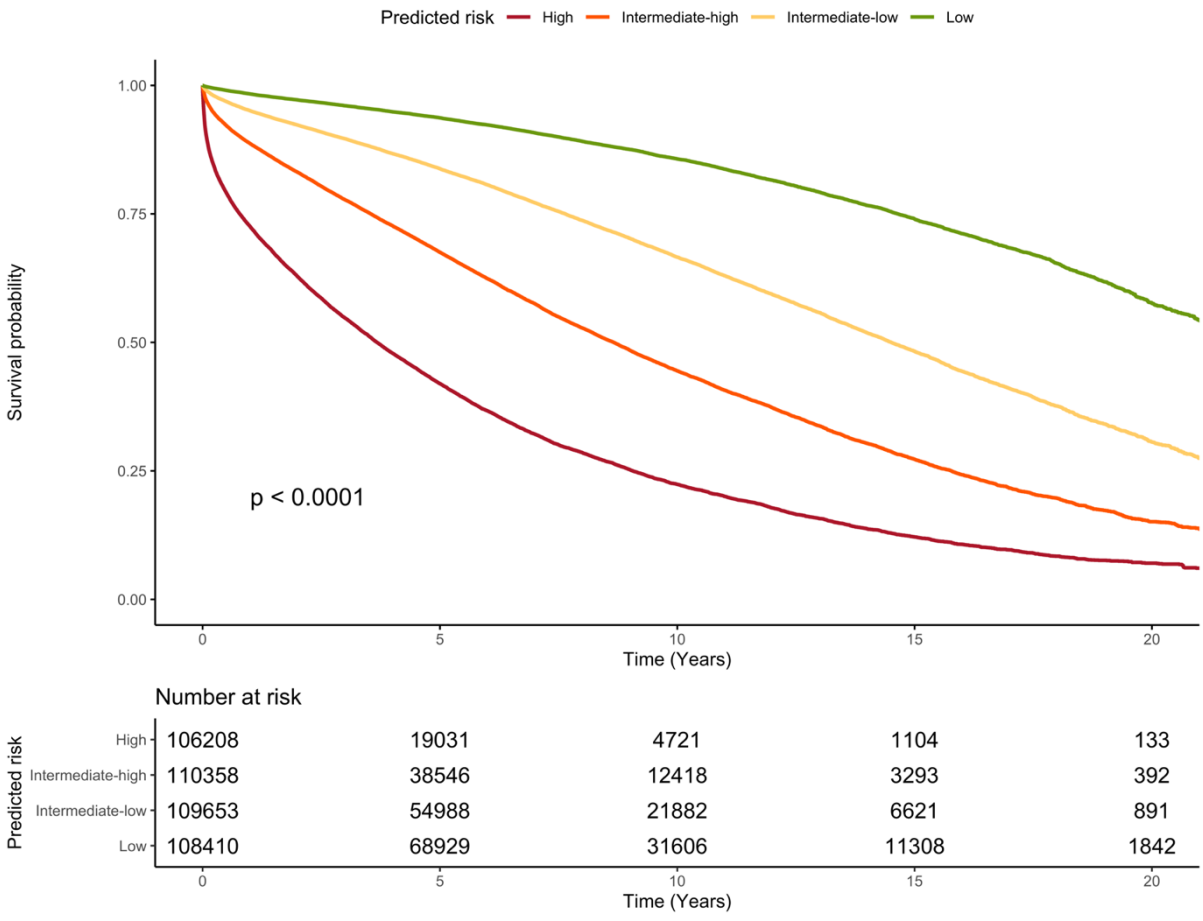


Figure 4-8 Single lead mortality prediction

BIDMC Test set, Kaplan-Meier curves of AIRE predicted all-cause mortality by risk quartile in the whole BIDMC test set using only ECG lead 1

AIRE predicts mortality in transnational external datasets

First, I evaluated the performance of AIRE in the SaMi-Trop cohort of patients with chronic Chagas cardiomyopathy (86). Dataset demographics for all cohorts are shown in **Table 4-1**. The C-index was 0.773 (0.733-0.813, quartiles shown in **Figure 4-9A**).

The CODE cohort is a Brazilian database of ECGs recorded in primary care (75). The dataset contains both 10s and 7s duration ECGs. I first evaluated the performance of AIRE without any fine-tuning. As AIRE was trained exclusively on 10s ECGs, I evaluated the model on the 10s subset (CODE-10s), AIRE had a C-index of 0.762 (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 4-9B**). When considering only the ECGs labelled as normal (26766 ECGs from 21897 subjects), there remained a significant difference in mortality between high-risk and low-risk subjects based on model predictions (**Table 4-2**).

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 (88)). The C-index was 0.713 (0.691-0.735 **Figure 4-9C**). Again, when considering normal ECGs only there was a significant difference in mortality between high and low risk subjects (**Table 4-2**).

Finally, I additionally evaluated the performance of AIRE-primary care in the UK Biobank, a relatively healthy volunteer population (n = 42386) with only 526 (1.2%) deaths during follow-up. The C-index was 0.638 (0.608-0.668 (**Figure 4-9D**)) for all-cause mortality. As cause of death was available in the UKB, I additionally examined the ability of AIRE-primary care (which was trained for all-cause mortality) to predict

CV death. C-index for CV death was 0.695 (0.636-0.754). **Table 4-5** summarises the mortality prediction results.

Table 4-5

Mortality prediction results summary table

Model performance as assessed by C-index (95% CI) is shown

AIRE: Artificial-intelligence enhanced ECG risk estimator

	BIDMC Test set	**SaMi-Trop**	**CODE-10s**	**CODE Test set**	**UK Biobank**	**ELSA-Brasil**
N ECGs	434,629	1631	958,954	464,429	42,386	13,739
N Subjects	64,105	1631	645,372	311,684	42,386	13,739
Internal/external testing	Internal	External	External	Internal	External	External
AIRE	0.775 (0.773-0.776)	0.773 (0.733-0.813)	0.762 (0.759-0.765)	-	-	-
AIRE-primary care	-	-	-	0.802 (0.799-0.805)	0.638 (0.608-0.668)	0.713 (0.691-0.735)

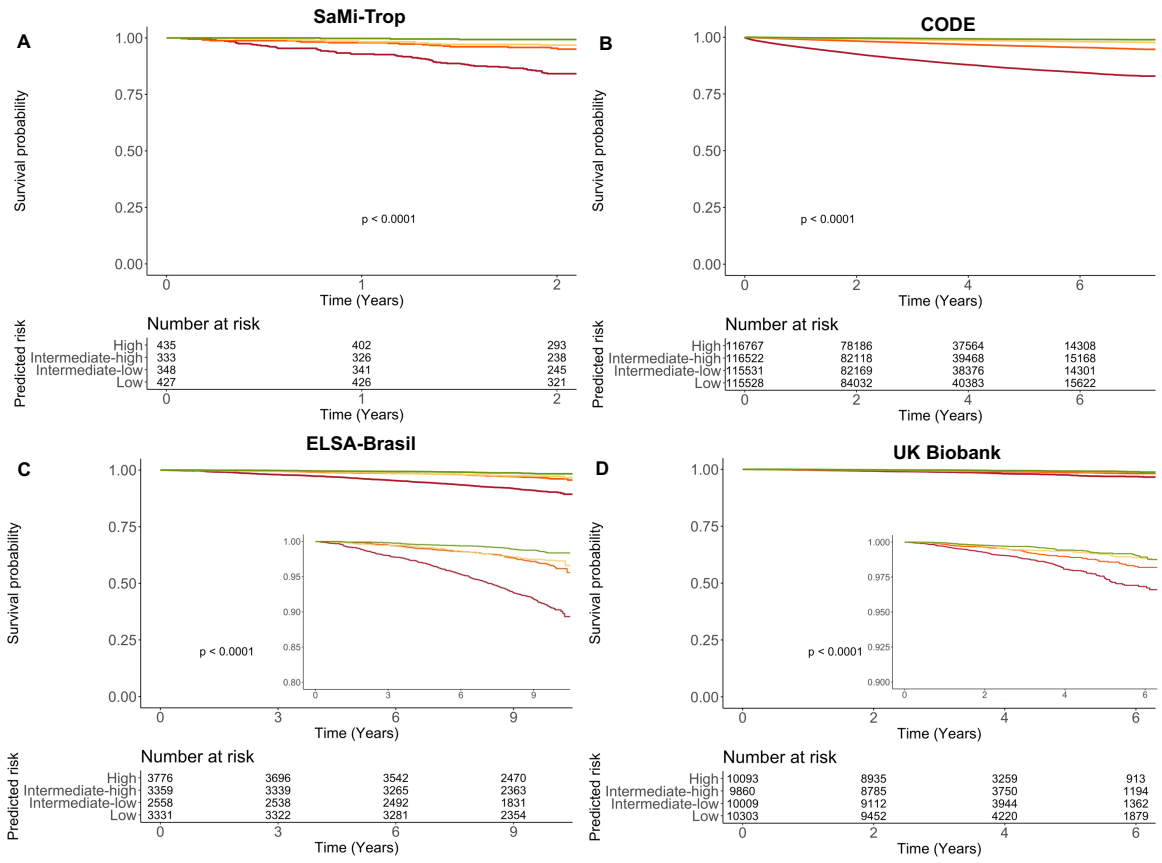


Figure 4-9 Mortality prediction Kaplan-Meier curves – external testing

Kaplan-Meier curves of AIRE-predicted all-cause mortality by risk quartile in (A) São Paulo-Minas Gerais Tropical Medicine Research Center (SaMi-TROP) cohort of subjects with Chagas disease. AIRE-primary care predictions are shown for (B) CODE cohort of primary care subjects in Brazil, (C) The Brazilian Longitudinal Study of Adult Health (ELSA-Brasil) volunteer cohort and (D) UK Biobank volunteer cohort. Sub panels show truncated y-axes for the populations with low event rates.

Hughes et al recently reported SEER (Stanford Estimator of ECG Risk) with the similar goal of predicting cardiovascular mortality (66). The model code and weights were downloaded, and performance evaluated in the UKB, as this was a dataset external to both AIRE-primary care and SEER, with cause of death available. AIRE-primary care (AIRE C-index 0.695 (0.636-0.754)) was superior to SEER at predicting CV death in the UKB (SEER C-index 0.572 (0.514-0.630) p values for comparison to AIRE-primary care <0.001).

AIRE outperforms AI-ECG models of biological aging

AI-ECG predicted age and the difference between AI-ECG predicted age and chronological age ("delta-age") have been shown to be a better markers of biological aging than chronological age (75) . I compared AIRE/AIRE-primary care predictions to AI-ECG predicted age. For the subsequent analyses, AIRE was used for SaMi-Trop and AIRE-primary care was used for CODE and ELSA-Brasil. I found AIRE/AIRE-primary care-predicted survival inversely correlated with AI-ECG age (CODE: r = -0.595, p < 0.0001, ELSA: r = -0. 449, p <0.0001, SaMi-Trop -0.479, p <0.0001) and, to a lesser extent, chronological age (CODE: r = -0.499, ELSA: -0.293, SaMi-Trop: -0.275, p < 0.0001 for all).

I compared the predictive capability of AIRE/AIRE-primary care with AI-ECG predicted age and delta-age in the CODE15 (a 15% subset of CODE), ELSA-Brasil and SaMi-Trop datasets. In age- and sex-adjusted models, AIRE/AIRE-primary care had a significantly higher C-index compared to AI-ECG predicted age or delta age in all three cohorts (CODE: 0.823 (0.813-0.833) vs 0.806 (0.796-0.816), ELSA-Brasil:

0.754 (0.734-0.774) vs 0.738 (0.718-0.758), SaMi-Trop 0.788 (0.749-0.827) vs 0.715 (0.672-0.758); $p < 0.001$ for all).

Explainable ECG morphology associates with adverse prognosis

Using a variational autoencoder, I found features of QRS morphology, particularly broader and more left bundle branch block morphologies, inverted and biphasic T waves as well as ST segment changes were identified as the most significant morphological features associated with high predicted mortality (**Figure 4-10A**).
In a second approach, using median beats the BIDMC test set I found poor precordial R wave progression, low QRS amplitude and T wave flattening/inversion as important features in AIRE-predicted survival (**Figure 4-10B**).

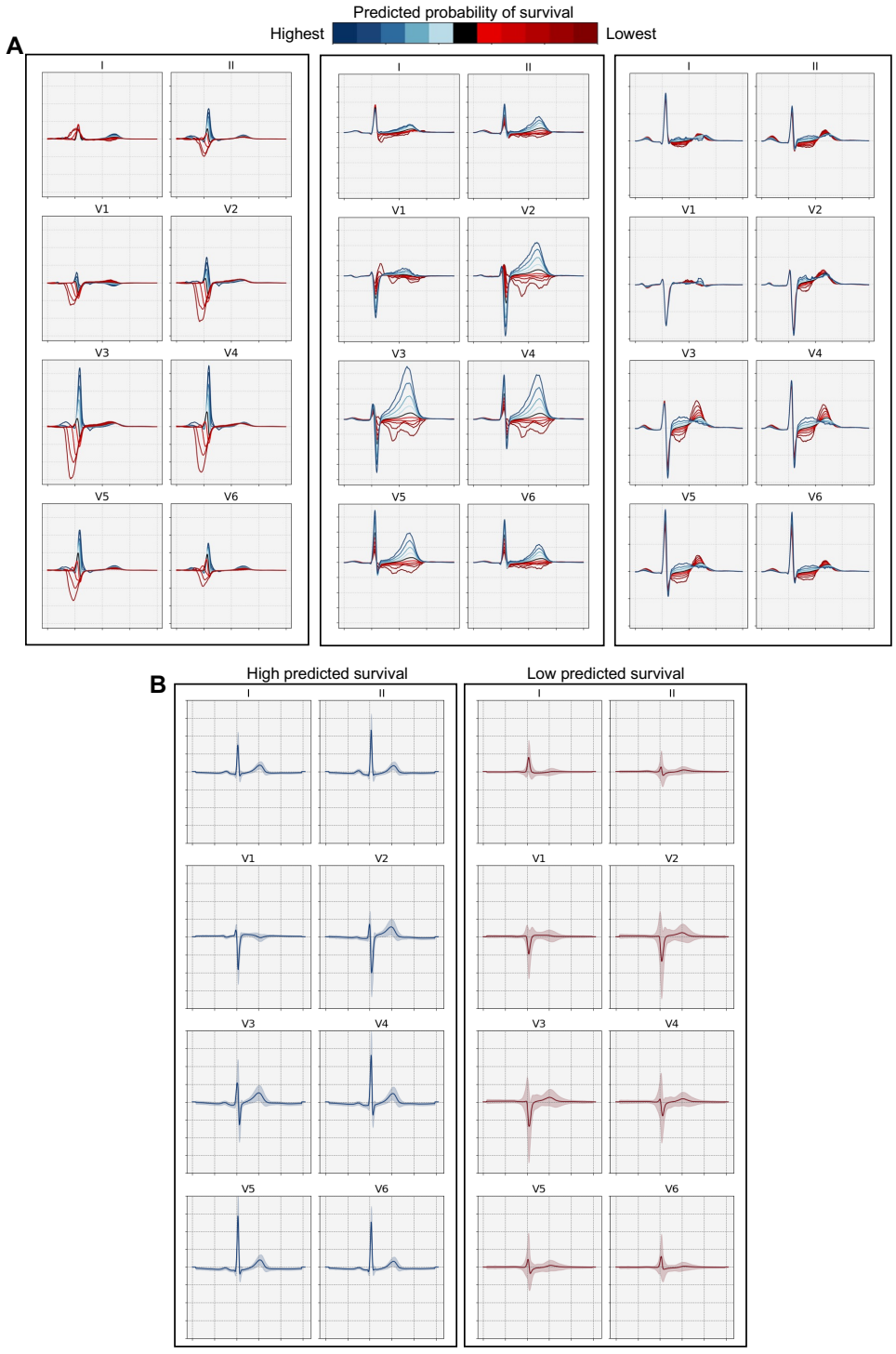


Figure 4-10 AIRE model explainability - BIDMC Test set

(A) A variational auto-encoder 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 as well as biphasic and inverted T waves. (B) Average ± standard

deviation (shaded region) ECG 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/inversion as important features in AIRE-predicted survival.

Biological plausibility: Genetic associations of AIRE-predicted survival

To investigate the underlying genetic associations with AIRE-predicted survival, we performed a genome-wide association study (**Figure 4-11A**). We found significant loci adjacent to *TBX3*, *VGLL2*, *CCDC91* and *KCNQ1*. 8% of the total phenotypic variance in predicted survival was caused by the additive effects of genetic variation. *TBX3* has been associated with blood pressure (153), ECG morphology (154), myocardial mass (155) and trabecular development (110). *VGLL2* has been associated with ECG parameters, blood pressure, atrial fibrillation and BMI (153, 154, 156, 157). *KCNQ1* (Potassium Voltage-Gated Channel Subfamily Q Member 1) encodes Kv7.1 and is most well-known for its associated with Long Qt Syndrome 1 and Jervell And Lange-Nielsen Syndrome 1 and QT interval (158) but is also associated with metabolic syndrome phenotypes (159-161). Finally, *CCDC91* (Coiled-Coil Domain Containing 91) associates with BMI (159), a variant associated with the ECG has been previously described in this region (154).

Biological plausibility: Phenotypic associations of AIRE-predicted survival

In order to investigate the biological associations with AIRE-predicted mortality, I performed a PheWAS in the UK Biobank, **Figure 4-11B**. In particular, CMR associations included reduced left ventricular (LV) ejection fraction (EF), more positive, i.e., abnormal, global longitudinal strain, increased LV mass and increased left atrial size, which were correlated with reduced AIRE-predicted survival. I additionally specifically examined the association between predicted survival and left ventricular trabeculation and found a significant negative correlation (**Table 4-6**). Important lifestyle risk factors included smoking status and alcohol intake which were both correlated with reduced predicted survival. Significant physical measures included systolic blood pressure, which was negatively correlated with predicted survival.

Table 4-6

Association of left ventricular trabeculation with predicted survival (n = 33,333)

	Beta coefficient	**R**	**p**
Predicted survival (binary)	-0.075 (0.11)	-	< 0.00001
Predicted survival (continuous)	-	-0.04	< 0.00001

In BIDMC echocardiographic analyses (**Figure 4-11C**), we found LVEF was positively correlated with predicted survival while LA volume and surrogate measures of pulmonary pressure (TR velocity) and right ventricular diameter were negatively correlated.

Finally, investigation of non-cardiac imaging phenotypes identified associations with multiple multi-model brain imaging phenotypes including the total volume of white matter hyperintensities and deep learning-derived brain age (111) (**Table 4-7**).

Table 4-7

Association of deep learning-derived brain age with predicted survival (n = 17348)

	R	**p**
Mean if Age Predictions	-0.0298972	< 0.00001
T1 Age Prediction	-0.04121074	< 0.00001
Jacobian Age Prediction	-0.02471008	0.001
Gray Matter Age Prediction	-0.02028091	0.008
White Matter Age Prediction	-0.01897039	0.012

Other non-cardiac imaging associations included predicted-survival inversely correlating with carotid intimal-media thickness, which is a marker of overall atherosclerotic burden (162).

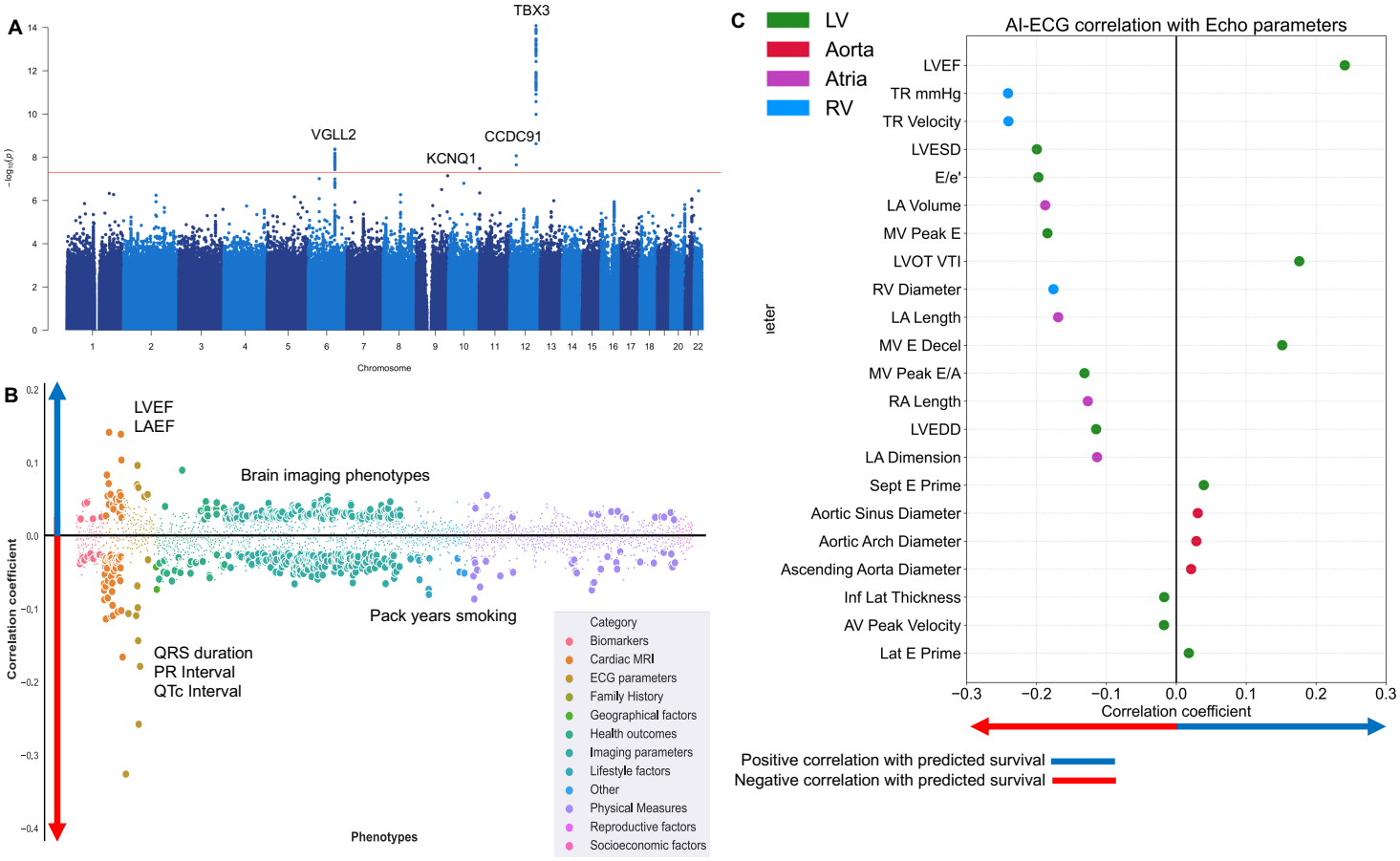


Figure 4-11 Phenome and Genome-wide association studies

(A) Genome-wide association study (GWAS) in the UK Biobank. Manhattan plots of genomic loci associated with predicted survival. Nearest genes to significant single nucleotide polymorphisms are shown. The red line depicts the genome-wide significant threshold (P<5 x 10-8). (B) Phenome-wide association study in the UK Biobank. Cardiac associations include left ventricular ejection fraction (LVEF), 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. LA: left atrium, LAEF: LA ejection fraction, TR:

tricuspid regurgitation, MV: mitral valve, LVESD: LV end-systolic diameter, LVEDD: LV end-diastolic diameter, RA: right atrium, AV: aortic valve.

Discussion

In this chapter I have developed and tested AIRE, an actionable, explainable, and biologically plausible mortality and risk prediction platform. Importantly, the platform was externally tested across ethnically and demographically diverse transnational cohorts.

Importance of time-to-event for actionability

This study substantially extends the work of others on mortality protection using the ECG. Raghunath et al previously described the use of deep learning for mortality prediction (72), while Sun et al more recently built upon this work (73). Our study has several significant differences from these previous publications. Firstly, the use of a survival neural network architecture, provides our model with the ability to predict time of death without being constrained to a small number of time points. Additionally, this allows us to use training data from subjects that were censored, for whom the time for death was not known. Specifically, I created a model that provides an individualised survival curve based on a single ECG. The ability to identify patients at risk of short-term mortality versus those at risk of long-term mortality is an important distinction. For example, patients at risk of short-term mortality may benefit from early escalation to intensive care settings, while those at risk of long-term

mortality may need more detailed outpatient, follow-up or aggressive medical or interventional treatment.

Furthermore, by comparing our model to existing clinical risk factors and demographic information I have demonstrated the significant additive value of our model beyond traditional approaches. Finally, I performed external testing across diverse populations, demonstrating the wide applicability of our model platform.

Actionable risk prediction in high-risk disease groups

By predicting risk of death in high-risk disease groups, I was able to provide potentially actionable information that could guide treatment decisions. I described how AIRE can be used to predict the risk of death in severe aortic stenosis and primary pulmonary hypertension. Severe aortic stenosis has a high mortality if left untreated. Current clinical guidelines in general advocate intervention based on echocardiographic measures and the patient's symptoms (163, 164) . However, there is increasing evidence that earlier intervention in asymptomatic individuals at high-risk of death may be of benefit, and there are randomised control trials currently investigating this approach (165, 166). Primary pulmonary hypertension also has a significant mortality rate (167). Medical therapy is available, and transplantation may be undertaken (168), therefore accurate risk stratification is paramount in guiding treatment selection for pulmonary hypertension patients. Mortality prediction using AIRE may be helpful in guiding treatments in both these conditions.

Single-lead ECG applications

Extending risk prediction to the single-lead ECG is particularly important given the rapidly increasing number of single lead devices, including consumer products (34) . Others have shown the applicability of AI-ECG algorithms to consumer single-lead products (169). Our study highlights the excellent performance of AIRE at mortality prediction on only a single lead. This could be particularly applicable for inpatient cardiac monitoring, or for remote monitoring in the outpatient setting using wearable devices, for example in patients with chronic diseases such as heart failure, where high-risk predictions could trigger pre-emptive treatments to prevent hospital admissions.

Explainability and biological plausibility

A major advantage of the deep learning approach is the ability to extract features relevant to the specific task, without anchoring on prior beliefs. However, a significant challenge is explainability of the model predictions. A significant barrier to the adoption of AI tools in clinical practice is physician reluctance to adopt technologies that are often “black boxes” (12). Explainability and biological plausibility is therefore key to improving clinician trust in AI. In this study using multiple approaches, I have explored the underlying biology behind AIRE risk predictions. Our ECG morphology findings are in line with prior studies that highlight these features as being prognostically importantly (64, 170, 171). Importantly, the significant predictive value of AIRE, even in normal ECGs, shows how deep learning can additionally make use of ECG morphological features that clinicians deem to be normal.

The accuracy of AIRE at mortality prediction raises the important question as to which biological pathways are being identified in these predictions. Using cardiac imaging data in two distinct datasets, I identified associations with cardiac chamber structure, function and measures of pulmonary pressure, that may be reflected in the ECG (172). I also identified multiple established factors, whose effects maybe reflected directly or indirectly in the ECG, including blood pressure, smoking and alcohol intake.

Previous work has described the importance of ischaemic heart disease on brain aging and outcomes (173). Our work has identified significant associations between ECG-predicted survival and brain imaging phenotypes that may reflect the combined effect of cardiovascular risk factors on cardiac and brain outcomes.

AIRE-predicted survival is a biomarker of overall health

Through GWAS, I have identified plausible biological pathways, including myocardial mass and trabecular development, that could be significant factors influencing AIRE-predicted survival. *TBX3* and *VGLL2* have been previously described in relation to AI-ECG derived delta-age, which is a marker of accelerated biological aging (75). Indeed, there was a moderate correlation between AIRE-predicted survival and AI-ECG predicted age, which was stronger than the correlation with chronological age. I also identified AIRE-predicted survival as inversely correlated with deep learning-derived brain-age. Additionally, I identified variants in *KCNQ1* and *CCDC91* that suggest AIRE may 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.

Reduced performance in volunteer populations

In our analyses, in general, AIRE-primary care had reduced performance metrics in the volunteer populations (UK Biobank, ELSA-Brasil). Our findings are consistent with other studies in that AI-ECG model performance is generally reduced in these volunteer populations (67, 75). A similar phenomenon is seen in other risk prediction models applied to the UK Biobank, such as QRISK3 (174). This may be due to differences in population characteristics, for example the UK Biobank population is older but healthier than the general population (90), and low event rates.

Limitations

AIRE demonstrated preserved model discrimination even at the 10-year timepoint, however calibration was reduced with longer time horizons. Therefore, at these longer time horizons, although the model accurately differentiates risk between subjects, the predicted and actual event rates may differ more than at shorter follow up. Although AIRE accurately predicted risk of future mortality, it is unclear if the trajectory towards death in high-risk individuals can be altered. Additionally, the AIRE model alone does not provide guidance as to what diseases may lead to death. Therefore, it may not be apparent to clinicians which actions may be beneficial to the patient. I have aimed to address these limitations in the next chapter, with the creation of disease specific AIRE submodels.

Conclusion

In conclusion, I 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- and long-term risk prediction at a population and disease-specific levels.

Chapter 5 - AI-ECG for prediction of diverse cardiovascular outcomes

Introduction

Parts of the work in this chapter of the thesis have appeared in the pre-print Sau et al, Artificial intelligence–enabled electrocardiogram for mortality and cardiovascular risk estimation: An actionable, explainable and biologically plausible platform, Lancet Digital Health (accepted, in press)

Expanding AIRE

In developing the AI-ECG risk estimation platform (AIRE), described in Chapter 4, I found the excellent performance of AI-ECG at predicting risk of death. However, in order for clinicians to be able to best alter the trajectory of subjects at high risk of death, further information around common and high-risk cardiovascular conditions would be informative. The goal of this chapter is to expand the AIRE platform, with a series of submodels that can be used to predict five important cardiovascular diseases.

Current risk prediction methods

Prediction of future disease is an important facet of clinical medicine. In cardiovascular care alone, hundreds of risk prediction scores/models have been developed, most of which are never used clinically. Some however have become ubiquitous, such as the CHADsVASc score (45). There are several major challenges with clinical risk scores that serve to limit implementation, these include limited discrimination performance, poor generalisation to external cohorts and complexity to calculate and implement.

There are several areas in cardiovascular medicine that make use of clinical risk prediction scores. Firstly, the prediction of atherosclerotic cardiovascular disease may use the Pooled Cohort Equation or more recently American Heart Association PREVENT (Predicting Risk of cardiovascular disease EVENTs), SCORE2 in Europe and QRisk3 in the UK (43, 175, 176). Prediction of future atrial fibrillation (AF) may use the CHARGE-AF score (68), although not commonly employed in clinical practice in the UK. Finally, in hypertrophic cardiomyopathy (HCM) the European Society of Cardiology (ESC) risk calculator can be used to predict sudden cardiac death. In other areas, specific singular features may be reflective of high risk, for example bifascicular block for risk of future complete heart block (CHB) or LVEF for future risk of malignant ventricular arrhythmias (VA) (177-179).

Aims

While AI-ECG cannot make up for all the challenges in clinical risk scores, I have shown the ECG captures information regarding short term and long-term risk. I hypothesised disease specific AIRE submodels could be used improve the

actionability of AIRE mortality predictions, and could be superior or at least additive to current gold standards in cardiovascular risk prediction.

I therefore aimed to develop AIRE submodels for five clinically significant and common cardiovascular diseases, ASCVD, heart failure (HF), VA, AF and CHB. Given the diverse existing approaches to AF prediction, I additionally performed a detailed study comparing AIRE-AF with existing AF risk prediction methodologies - clinical risk score, cardiac imaging and genetics (polygenic risk score).

Methods

Study overview

In this chapter I describe the development and testing of five AIRE submodels, to create the extended AIRE platform. **Figure 5-1** provides an overview.

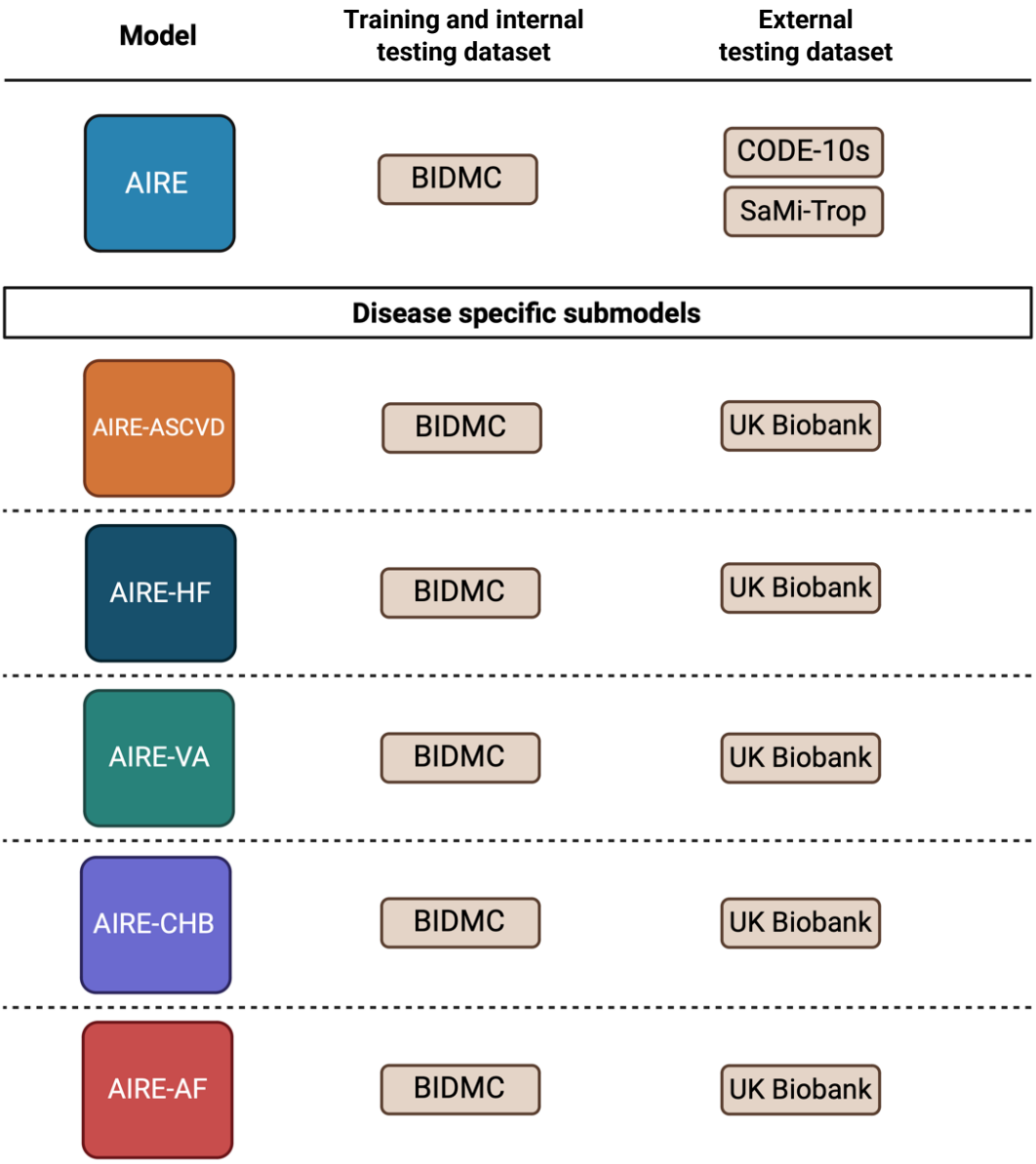


Figure 5-1 Extended AIRE platform overview

Extended, disease specific, AIRE Platform overview. Disease specific models were trained to predict future ASCVD, HF, VA, AF and CHB. Models were externally tested in the UKB.

Cohorts

This chapter used the BIDMC and UK Biobank cohorts, as described in the main methods chapter. These were the only cohorts with diagnostic codes (ICD codes) available.

Disease definitions

Diseases were defined using ICD9 and ICD10 codes in the BIDMC and UKB datasets. Atherosclerotic cardiovascular diseases (ASCVD) included coronary artery disease, myocardial infarction, stroke and peripheral vascular disease. Heart failure included the full spectrum of left ventricular ejection fraction. Ventricular arrhythmia included ventricular tachycardia and ventricular fibrillation. A limitation of ICD codes is that non-sustained ventricular tachycardia episodes cannot be distinguished from sustained, this is an important caveat to the definition of ventricular tachycardia in this study.

Atrial fibrillation diagnoses used both ECG detected atrial fibrillation and ICD code defined atrial fibrillation or atrial flutter. A previously described AI-ECG model was used to identify ECGs of AF (23). Atrial flutter was included in this definition given similarities in indication for anticoagulation. Finally, the complete heart block definition included only complete heart block and not other types of high-grade AV block. ICD codes do not clearly distinguish between Mobitz type 1 and 2, therefore neither were included.

Model fine-tuning

I developed these five subsequent models by fine-tuning the AIRE model separately. The models were named are referred to as: AIRE-VA, AIRE-ASCVD and AIRE-HF, AIRE-AF and AIRE-CHB respectively. The same patients were allocated into training, tuning and test set as was used in Chapter 4. Informal experimentation with different fine-tuning strategies demonstrated the most effective fine-tuning strategy was loading the previous model and training using a low learning rate without freezing any layers.

Subjects were coded as having prevalent disease, future disease or neither at the time of the ECG. Although the goal of these models was to predict future events, I hypothesised that information on prevalent disease would be helpful for model training. In order to include prevalent disease in the discrete-time survival model, I encoded prevalent disease at the first timepoint in the discrete-time survival labels. For the AF model, ECGs in AF were included in model training and encoded as prevalent disease at the first timepoint in the discrete-time survival labels. AF ECGs were determined using the AI-ECG model described above (23).

When evaluating model performance, subjects with prevalent disease were excluded. AF ECGs were excluded during model evaluation of the AF model. The final iteration of each model was selected based on the lowest tuning loss.

Comparison with previous AI-ECG model

As described in Chapter 4, I compared AIRE model performance with the recently described SEER (Stanford Estimator of ECG Risk) (66). In this Chapter I compare AIRE-ASCVD with SEER for future ASCVD risk estimation.

Comparator clinical risk scores

The Pooled Cohort Equation and CHARGE-AF scores were used for comparison of the ASCVD and AF models respectively. Medication information was not available in the BIDMC cohort, therefore for the anti-hypertensive medication parameter a diagnosis of hypertension was used instead. The scores were computed using the published algorithms. The ARIC-HF risk factors were used as comparator for heart failure, the model weights were not available, therefore the variables in that model were used in a Cox model (180). LVEF was the main comparator for ventricular arrhythmia prediction, while ECG morphologies and parameters were the main comparator for CHB prediction.

Polygenic risk score

Atrial fibrillation in particular is known to have a strong genetic basis. Polygenic risk scores (PRS) have been developed to quantify the risk of atrial fibrillation development. An existing AF PRS based on a meta-analysis of AF GWAS' was used in this study (181).

Survival analyses

In each case, subjects with a prevalent diagnosis of the disease of interest were excluded from survival analysis. C-indices were calculated for each model prediction. Cox models were fit to compare performance of the AIRE submodels with other clinically relevant parameters and scores. For complete heart block prediction, I additionally excluded subjects with: a prior cardiac implantable electronic device

(CIED, e.g. pacemaker or implantable cardioverter defibrillator (ICD)), diagnosis of second degree or higher heart block or ECG heart rate of less than 50 beats per minute (bpm) (to exclude current high-grade block on that ECG).

Results

Atherosclerotic cardiovascular disease

Dataset demographics are as described in Chapter 3. AIRE-ASCVD was able to predict future ASCVD in subjects without known ASCVD (C-index 0.696 (0.694-0.698) n = 227588 ECGs from 56598 subjects). I externally tested these findings in the UKB Biobank (UKB); a healthy, volunteer population. AIRE-ASCVD had reduced performance in predicting ASCVD, C-index 0.643 (0.624-0.662) but was significantly better than the SEER model (66) (SEER C-index 0.547 (0.527-0.567), p <0.0001 for comparison with AIRE).

I compared AIRE-ASCVD to other risk parameters, including the pooled cohort equation (PCE) and ASCVD risk factors (systolic blood pressure, total cholesterol, HDL cholesterol, hypertension, smoking history, diabetes mellitus and ethnicity), in a subset of outpatients (4580 ECGs from 2926 subjects) in the BIDMC test set with appropriate available data. AIRE-ASCVD-Cox had a significantly higher C-index than all other factors combined (0.679 (0.651-0.708) vs 0.642 (0.613-0.672), **Figure 5-1**, p < 0.00001).

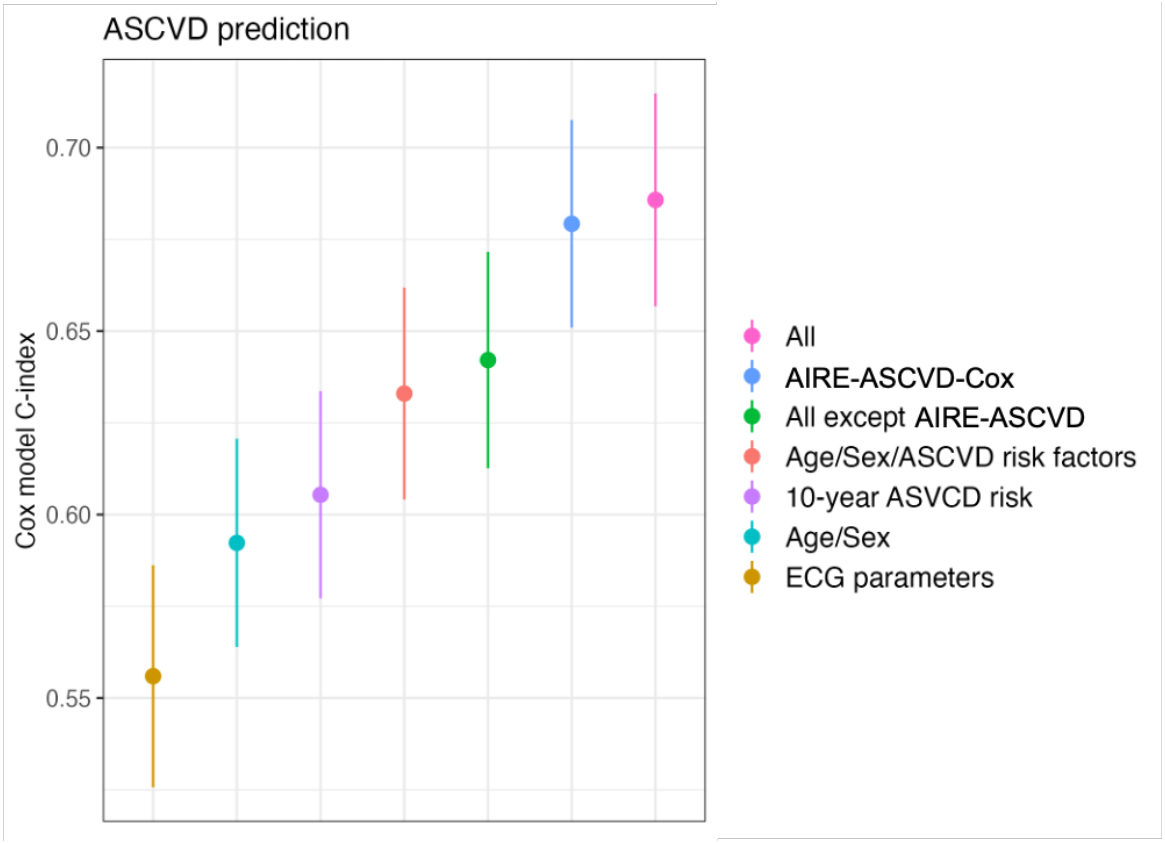


Figure 5-2 AIRE-ASCVD performance

Forest plot comparing cox model performance of AIRE-ASCVD with other risk predictors

Ventricular arrhythmia

AIRE-VA was able to accurately predict future VA (C-index 0.760 (0.756-0.763) n = 393203 ECGs from 62443 subjects) in subjects without a previous history of VA. When compared to other conventional risk parameters, including ECG parameters

and left ventricular ejection fraction, AIRE-VA-Cox, had a significantly higher C-index than all other factors combined (0.732 (0.723-0.741) vs 0.700 (0.692-0.709), **Figure 5-2**, $p < 0.0001$).

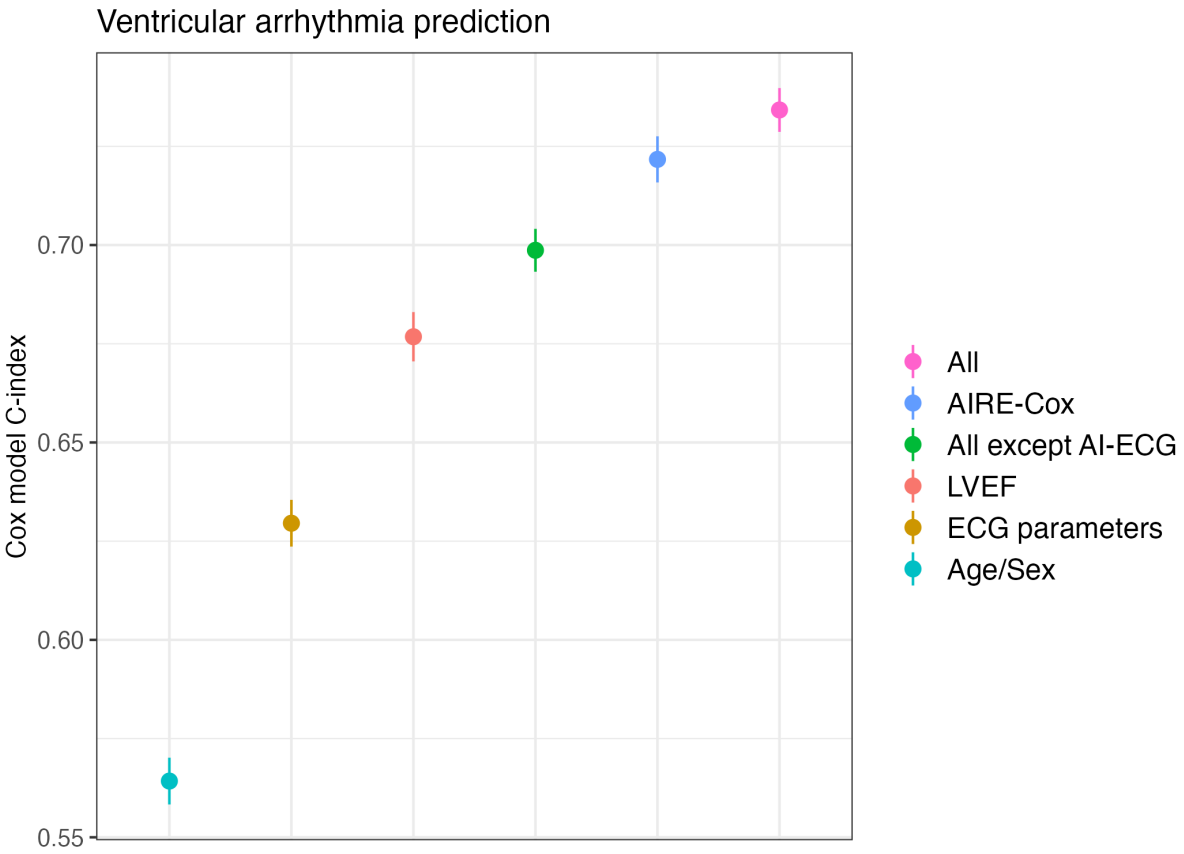


Figure 5-3 AIRE-VA performance

Forest plot comparing cox model performance of AIRE-VA with other risk predictors

In the UKB, (n = 34400 with both ECG and cardiac magnetic resonance imaging (CMR) and without previous VA, 44 events), AIRE-VA had similar performance in predicting first occurrence of VA 0.719 (0.635-0.803) and performed at least equivalent to LVEF from cardiac magnetic resonance imaging (CMR) (C-index 0.595 (0.494-0.697), p for comparison 0.11).

I went on to perform subgroup analyses in the BIDMC test set. In a subgroup of patients (11573 ECGs from 2802 subjects) with an LVEF of <50%, AIRE alone had a higher C-index than LVEF (0.649 (0.638 – 0.660) vs 0.615 (0.603 – 0.626)) in

predicting first episode of VA. In patients with dilated cardiomyopathy (DCM) (3366 ECGs from 772 subjects), AIRE alone had a C-index of 0.690 (0.669-0.711) compared to 0.608 (0.584 – 0.631) for LVEF alone. Incorporating LVEF and AIRE together in a Cox model yielded marginal improvements in C-index of 0.666 (0.656 – 0.676) and 0.694 (0.672 – 0.716) for patients with LVEF <50% or DCM respectively. I additionally investigated the performance of AIRE-VA in hypertrophic cardiomyopathy. C-index for future VA was 0.730 (0.670-0.789).

Heart failure

AIRE-HF was able to accurately predict future HF in subjects without a previous history of HF (C-index 0.787 (0.785-0.889) n = 310200, from 61747 unique subjects). In a subset of patients with the available data, I compared AIRE-HF-Cox to other conventional risk parameters in Cox models (36486 ECGs from 12288 subjects), including HF risk factors identified in the Atherosclerotic Risk in Communities (ARIC) study (body mass index, systolic blood pressure, prevalent ASCVD, diabetes mellitus, smoking history, previous myocardial infarction, hypertension and ethnicity) (180). AIRE-HF-Cox, had a significantly higher C-index than all other factors combined (0.761 (0.755-0.767) vs 0.716 (0.710-0.722), **Figure 5-4**, $p < 0.00001$). In the external testing cohort, UKB, AIRE-HF-Cox had similar performance at predicting future HF (C-index 0.768 (0.733 -0.802).

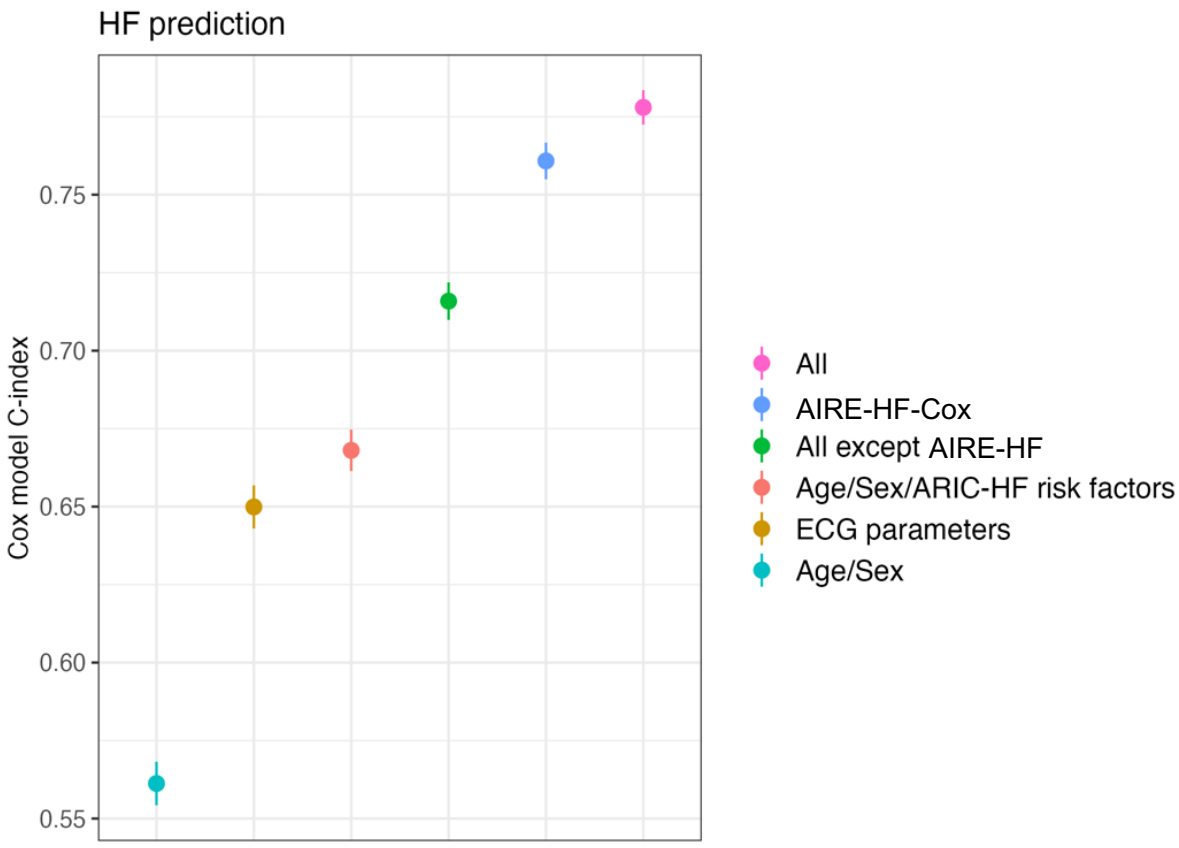


Figure 5-4 AIRE-HF performance

Forest plot comparing cox model performance of AIRE-HF with other risk predictors

Complete heart block

AIRE-CHB predicted future complete heart block with a C-index of 0.809 (0.805-0.814) on the BIDMC test set and 0.927 (0.887-0.959) on external testing in the UK Biobank. AIRE-CHB was had a significantly higher C-index than other predictors (**Figure 5-5**)

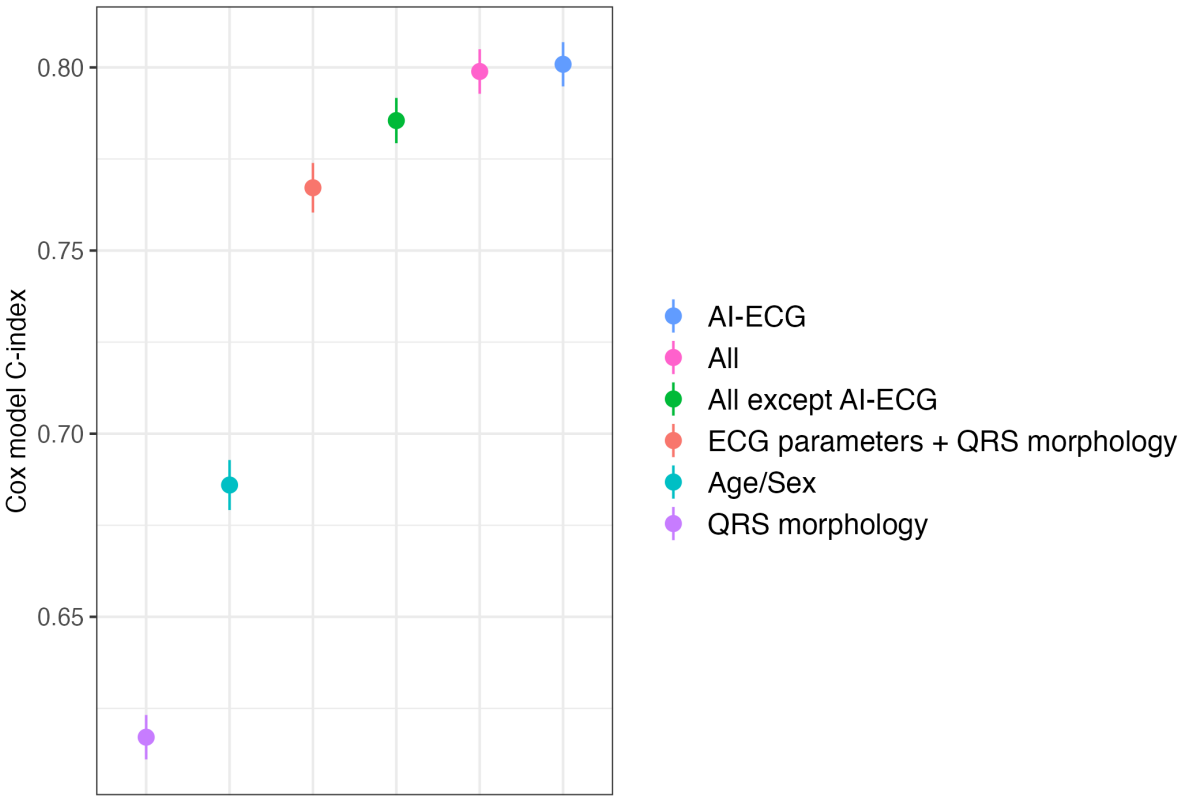


Figure 5-5 AIRE-CHB performance

Forest plot comparing cox model performance of AIRE-CHB with other risk predictors. QRS morphology: LBBB, RBBB, bifascicular/trifascicular block

I additionally investigated subjects with syncope and found in this subgroup AIRE-CHB identified a progression to CHB with a c-index of 0.805 (0.766-0.843).

For complete heart block prediction, AIRE was able to further risk stratify subjects with bifascicular block and syncope into low (1.9% 5-year CHB), intermediate (8.8%), and high risk (20.7%).

In BIDMC normal ECGs AIRE-CHB had reduced discrimination in predicting future CHB (C-index 0.713 (0.568-0.857)).

Atrial fibrillation

Given the large body of work already existing around AI-ECG for future/occult AF prediction, I performed more detailed analysis and comparative experiments for this endpoint.

AIRE-AF predicted future atrial fibrillation with a C-index of 0.753 (0.751-0.756) on the unselected BIDMC test set. On unique outpatient ECGs performance was similar (C-index 0.758 (0.750-0.765)). Furthermore, in evaluation of subjects with a normal ECG as defined by a cardiologist (full definition in Methods), the model retained performance AIRE-AF had a similar ability to predict future AF (C-index 0.725 (0.695-0.754).

On external testing in the UK Biobank the C-index was 0.716 (0.697-0.735). **Figure 5-6** shows Kaplan-Meier curves of four quartiles of risk in the BIDMC and UKB cohorts.

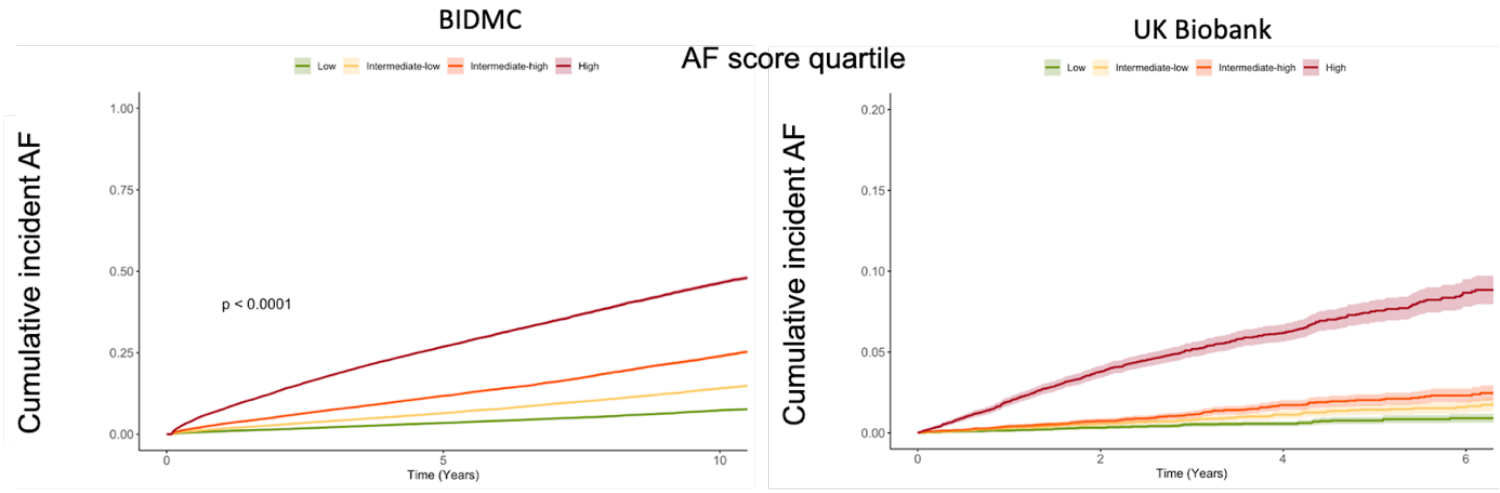


Figure 5-6 AIRE-AF Kaplan-Meier curves

Kaplan-Meier curves depicting risk of future AF in subjects without known AF. Four risk quartiles were defined using the BIDMC tuning set.

I compared AIRE-AF to the CHARGE-AF score and traditional ECG parameters and found AIRE-AF was superior to both (AIRE-CHB was had a significantly higher C-index than other predictors (**Figure 5-7**). When combined in a Cox model, AI-ECG was additive to the CHARGE-AF score (NRI 0.582 (0.537-0.626).

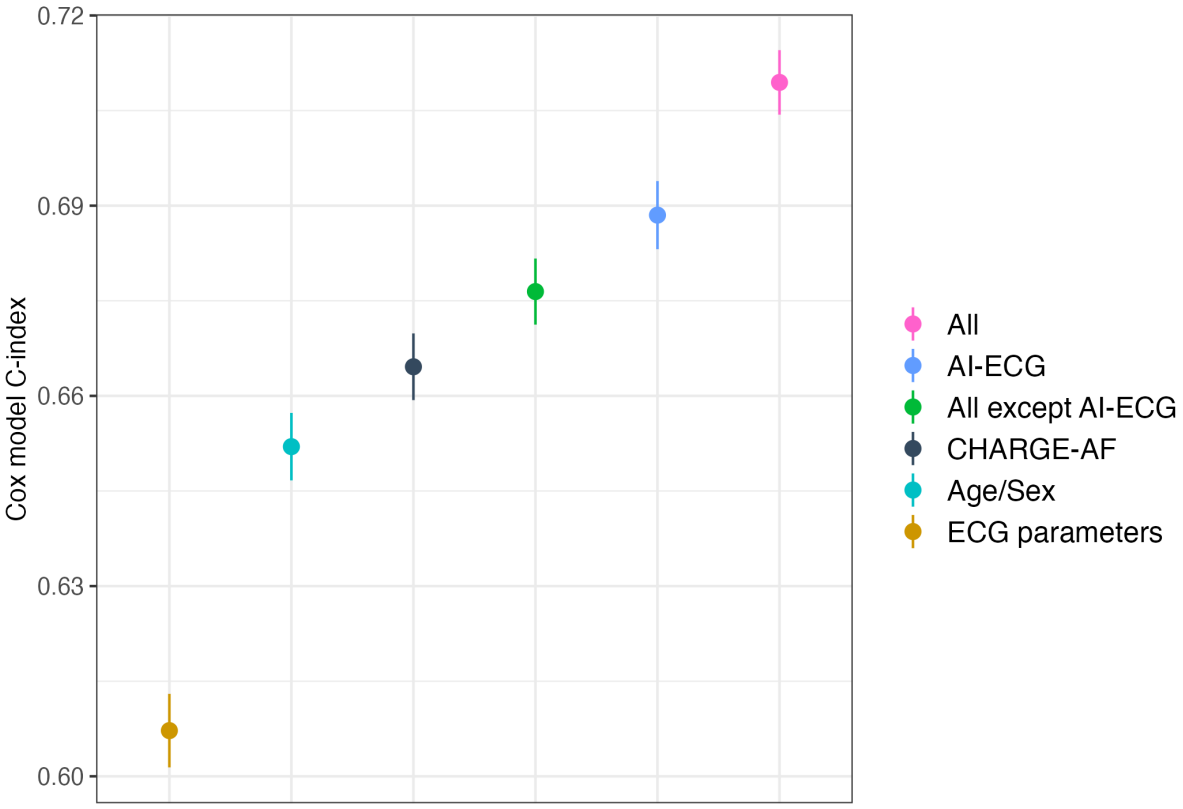


Figure 5-7 AIRE-AF Performance

Forest plot comparing cox model performance of AIRE-AF with other risk predictors

LA size as measured on echocardiography (BIDMC) was a modest predictor of future atrial fibrillation (LA dimension, C-index 0.618 (0.604-0.631), LA volume, C-index 0.646 (0.633-0.659). In Cox models, LA size had a modest additive value to AI-ECG and CHARGE-AF (NRI 0.186 ()).

In the UKB I additionally investigated the predictive ability of a previously described PRS, derived in a population external to the UKB. PRS alone had a c-index of 0.616 (0.596-0.636). In a cox model including age and sex and c-index was 0.740 (0.722-0.758).

In the UKB, where all variables were available I created Cox models to demonstrate the accuracy of four approaches: CHARGE-AF + AIRE-AF (model 1), PRS + AIRE-AF (model 2), CHARGE-AF + PRS (model 3), CHARGE-AF + AIRE-AF + PRS (model 4) (**Figure 5-8**) . I found each of the three components contributed roughly equally when combined with age and sex. Each combination of two components was also equivalent, with all three components having the best performance.

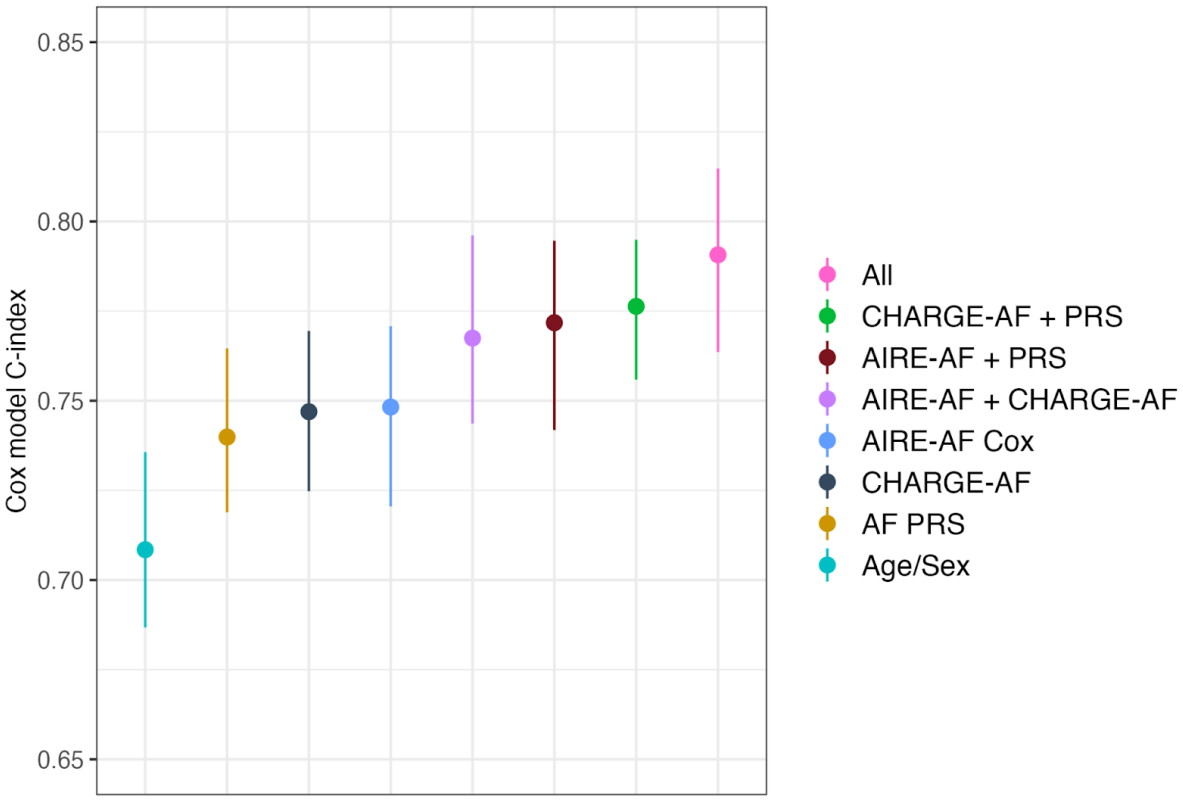


Figure 5-8 AIRE-AF UK Biobank

Forest plot comparing cox model performance of AIRE-AF in combination with other risk predictors, including PRS in the UK Biobank

Cross-Task Model Performance Evaluation

Next I investigated if each submodel was truly predicting different events, rather than just an overall marker of risk. To do this I assessed the performance of each model for each endpoint (Figure 5-9). In each case, the matched model and endpoint had the best performance compared to other pairings. Interestingly the HF model ap-

peared to be the single model with the most generalisable performance across different tasks, however the task specific models in each case were still the best performing.

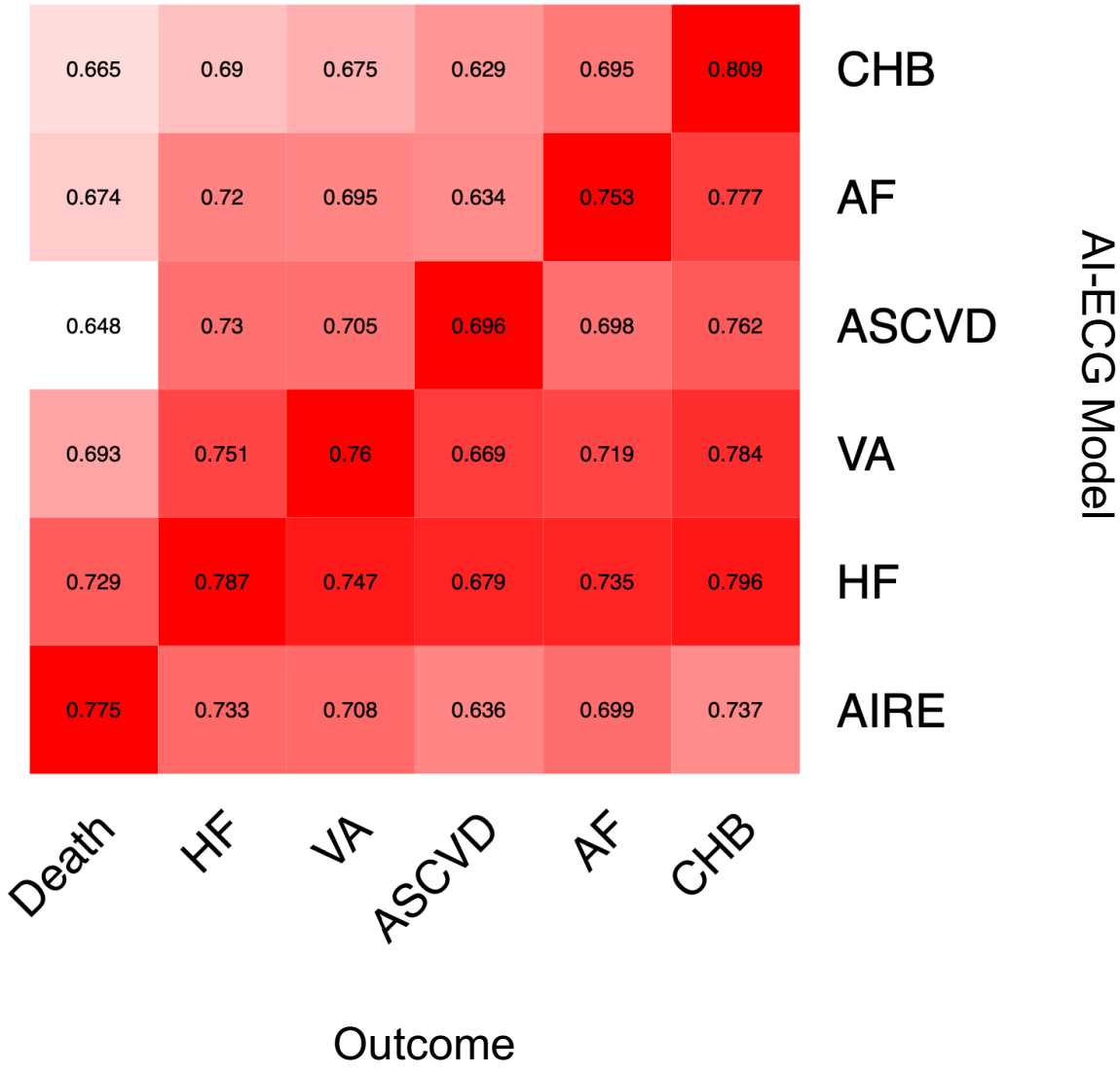


Figure 5-9 Cross-Task Model Performance Heatmap

C-index for each model/outcome pairing is shown, colours are scaled to the column maximum to show the best performing model for each outcome. In each task the outcome paired with the task-specific model performed the best.

Discussion

Using disease specific models, I have shown the AIRE platform can also predict future cardiovascular diseases such as ASCVD, HF, VA, CHB and AF, in addition to predicting mortality. The main goal of disease prediction is to provide more actionable information that can alter patient trajectories through specific interventions targeted to each disease.

Future ASCVD prediction

ASCVD prediction is currently used extensively in international guidelines for decision-making around lipid lowering therapies. A 10-year ASCVD is calculated and can be used to consider initiation of primary prevention statin therapy (182). In order to assess ASCVD risk, various scoring systems are used, including the PCE, SCORE2 and QRISK3 (182-184) . The ECG, despite being a cheap and readily available investigation, does not feature in any of these scoring systems. In this study, I have shown that AIRE-ASCVD can provide additional information that could improve ASCVD risk prediction and may be superior or at least additive to the existing PCE.

The analyses are additive to those of Hughes et al (66), who recently describe SEER, an AI-ECG model for the prediction of CV death. AIRE-ASCVD was superior to predicting ASCVD in the UKB. Additionally, the time-to-event nature of our platform provides additive actionable information as discussed above in Chapter 4. Subjects at high predicted risk of ASCVD could be offered statin therapy, or

potentially aspirin, although the recent randomised evidence for aspirin as primary prevention does not support its routine use (185).

Future VA prediction

Similarly, predicting future VA is a particularly important endpoint, as there are clear preventative, and therapeutic options. For example, in-hospital predictions of short-term high risk of VA could be used to guide the need for cardiac monitoring and potentially even prophylactic use of antiarrhythmics. Medium-term risk could be managed with a wearable cardioverter defibrillator (186), while patients at longer-term risk of ventricular arrhythmias may benefit from primary prevention implantable cardioverter defibrillator implantation. Current guidelines advocate LVEF as the primary factor in determining eligibility for a primary prevention ICD. In this study, I demonstrated that AIRE-VA is a better predictor of future VT/VF than LVEF and could therefore potentially be incorporated into this decision-making paradigm. Sub-analyses in DCM and HCM show the ability of AIRE-VA to be applied to specific disease cohorts, without even fine-tuning.

Future HF prediction

Predicting future heart failure is important given the high number of unplanned hospital admissions due to undiagnosed heart failure (187). Early clinical assessment, echocardiography and institution of appropriate therapies may reduce adverse events, particularly in HF with reduced ejection fraction (HFrEF) (188).

AIRE-HF could be incorporated in cardiology clinics, for example in patients post myocardial infarction, to guide surveillance echocardiography or even pre-emptive heart failure therapies. Prospective studies would be needed however to evaluate this approach.

Future CHB prediction

Prediction of future/occult complete heart block may be useful to guide pacemaker implantation, particularly in patients with syncope or presyncope and other features that may suggest conduction disease. These other features could include the presence of a degree of conduction disease on the baseline ECG or intermittent atrioventricular block (not meeting pacing criteria) on ECG monitoring. The current evidence base does not offer firm recommendations, particularly in the context of bifascicular block and syncope (189). Instead, in the absence of high-grade AV block, clinical judgement is the main criteria for guiding further investigation. Additionally, there is no clinical risk score to guide clinicians.

In this study I have shown the ability of AIRE-CHB to provide additional risk stratification above traditional ECG parameters and morphologies, that could be used to guide investigations and management decisions in patients with syncope at risk of bradyarrhythmias.

Future AF prediction

Interest in screening for AF has garnered significant attention in recent years. The link between clinical AF and embolic stroke is well established, with anticoagulation

therapy usually recommended if the CHADsVASc score is 1 or more (45). The aim of screening is to identify AF early in subjects with a raised CHADsVASc score, and institute anticoagulation to prevent an embolic stroke. Another important area is identifying AF in patients who have suffered a cryptogenic stroke, particularly if there are features suggesting an embolic origin.

Several methods of surveillance including wearables, personal devices and implantable loop recorders have been studied, however robust reductions in hard endpoint (i.e. stroke) have not been demonstrated (190). This maybe because screening detected AF is a distinct clinical entity compared to clinical AF. Clinical AF is sometimes detected incidentally, usually at a relatively AF high burden if picked up on a random 10s ECG, but is often diagnosed in the presence of symptoms. The seminal anticoagulation trials demonstrating a benefit of anticoagulation in reducing stroke in AF studied subjects with clinical AF. It is possible therefore that subclinical AF picked up by screening may have a lower absolute stroke risk, and therefore potentially less to gain with screening. The device detected AF trials have also shown modest benefits which support this potential difference between subclinical and clinical AF.

AI-ECG has garnered significant interest in this space as a way to target individuals with high AF risk for AF screening, rather than using an age or demographic approach that may have a low pick-up rate. In the STROKESTOP trial of 75-76 year old subjects (191), the screening arm received intermittent ECGs for 14 days and resulted a small decrease in the primary endpoint of ischaemic or haemorrhagic stroke, systemic embolism, bleeding leading to hospitalisation, and all-cause death. After screening the invited to screening arm (which included subjects who did not accept the invitation) had an AF diagnosis rate of 14% compared to 12.8% in the

control arm. Interestingly in LOOP, where the screening arm received an implantable loop recorder (ILR), although there was a 31.8% rate of AF diagnosis in the screening arm (vs. 12.2% in the control arm, there was not significant difference in the primary outcome (190). These trials support the distinction between subclinical and clinical AF in terms of modifiable stroke risk, and suggest targeted (to reduce number needed to screen) but low intensity screening (in order to detect only high burden AF) for AF may identify subjects with undetected AF that are most likely to benefit.

In this study I have shown that AIRE-AF is capable of identifying both occult AF and future AF, and therefore could potentially be used to identify individuals at high risk of AF who may benefit from screening. I have also shown that integration of clinical risk parameters and a PRS can further improve the predictive performance of AIRE-AF.

Limitations

The AIRE submodels described in this study are based on ICD codes. 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. Further evaluation in an implantable cardioverter defibrillator cohort is needed. The clinical significance of subclinical AF may not be comparable to clinical AF, further prospective studies are needed to better understand the impacts of treatment in subclinical AF, in particular if anticoagulation has a net benefit.

Conclusion

In this study, I have shown how fine-tuned submodels of AIRE can be used to accurately predict a broad spectrum of future cardiovascular diseases. These predictions may be more easily actioned given existing preventive and screening strategies for these diseases. In comparing these models to existing risk prediction scores and criteria I have shown AI-ECG is superior and additive to these existing methods, and could be integrated to optimally predict the risk of these diseases.

Chapter 6 - AI-ECG as a biomarker of ECG sex phenotype

Introduction

There are well-documented sex differences in the prevalence, diagnosis, treatment, and outcomes for cardiovascular disease (192). Females are often underserved, with increased rates of misdiagnosis and adverse events (193). Specifically for cardiovascular disease, females are less likely to receive appropriate treatments, preventive care, or aggressive risk factor modification for cardiovascular risk factors (194). This may in part be due to lack of awareness of sex-specific risk factors in women, and an inappropriate bias where clinicians perceive females to be lower risk of cardiovascular events (195).

As a result, there is growing awareness that sex is an important factor to be considered in risk stratification and management of cardiovascular disease. Sex is often included in cardiovascular risk scores as a dichotomous variable, with females often considered lower risk than males (183). However, there is emerging evidence that the biological sex phenotype may be more accurately represented as continuum (196). Recent studies have suggested that there may be a sex continuum for both cardiometabolic and psychological traits (196, 197). In the era of precision medicine, the dichotomisation of sex for purposes of risk stratification may prove a crude

approach that fails to fully capture the continuum of inter-individual variability of sex-related cardiovascular risk.

In chapter 4, I demonstrated the AI-ECG mortality prediction model performed similarly in males and females. However, I did not examine sex differences in more detail. Sex differences in the ECG are well known (198). Recently, AI-ECG models have been shown to be able to identify sex with impressive accuracy (126, 199) and sex misclassification by AI-ECG models appears to be associated with adverse prognosis (199).

However, it remains unclear if sex misclassification represents a failure of AI-ECG models to differentiate sex in the presence of gross or subtle ECG abnormalities secondary to pathology or if the “misclassifications” of sex by AI-ECG models reflect the presence of a true sex continuum of ECG phenotypes that is being detected by these models, that captures information on cardiovascular risk.

I hypothesised that the previously-described adverse prognosis associated with AI-ECG sex misclassification is due to higher-risk individuals having more ECG abnormalities, which reduces the model’s ability to pick out male- or female-specific ECG morphologies for accurate sex classification. I therefore aimed to investigate the significance of AI-ECG sex misclassification specifically in lower-risk cohorts, where gross and subclinical ECG abnormalities are far less likely, and hypothesised that a more male ECG phenotype in females is associated with adverse cardiovascular prognosis. This study is an extension of the concept relating to AI-ECG delta-age (discussed in Chapter 1) as described by Lima et al (75), where if AI-ECG predicted age is greater that chronological age, this is associated with adverse prognosis.

Methods

Model development

I developed the AI-ECG sex classification model using the BIDMC cohort as the derivation dataset. The data was split at a ratio of 50/10/40 % for training, tuning, and internal test, respectively. Data was split by patient ID stratified by presence of ECGs with paired 5-year life status. To prevent data leakage, a single subject could have ECGs assigned to only one of training/tuning/testing datasets. I used a previously described convolutional neural network architecture based on residual blocks (23) and trained for the binary task of self-reported sex classification. Male was encoded as 0 and female as 1. The final layer used a sigmoid activation. The binary cross-entropy was minimized using the Adam optimizer, dropout was set to 0.8, the learning rate was 0.00005. Models were training for up to 50 epochs and the iteration with the lowest tuning loss used on the test set. Given the excellent out-of-the-box model performance no further hyperparameter optimisation was performed.

Sex discordance score

In a manner similar to the concept of delta-age, I derived a metric to quantify the discordance between the model prediction and the true sex label. In the BIDMC test set and UKB, I generated sex predictions (continuous) of all ECGs. AI-ECG sex discordance score (subsequently referred to as sex discordance score) was calculated by taking the absolute of AI-ECG sex prediction minus self-reported sex (encoded as 0 for male, 1 for female).

Explainability

In order to understand the ECG morphologies associated with predicted sex, I used three approaches. First, we trained a variational autoencoder (VAE) using median ECG beats as described in Chapter 2. VAE latent features were input into a linear regression with AI-ECG predicted sex as the output. The top 3 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 1,000 ECGs with the lowest and highest AI-ECG sex prediction (i.e. 'most male' and 'most female' respectively). The mean and standard deviation of these waveforms was then plotted.
Third, in the UK Biobank I performed univariate correlation between AI-ECG sex prediction and ECG parameters. Using linear regression, AI-ECG sex prediction was adjusted for age, age2, height, weight, body surface area (BSA) and waist circumference.

Normal ECG and healthy subject definition

In this study, I aimed to investigate the significant of AI-ECG sex misclassification in low-risk cohorts. I selected individuals with normal ECGs and also "healthy" subjects to conduct our analyses. Normal ECG definition for BIDMC is described in Chapter 2. Clinician ECG reports were not available in the UK Biobank.
'Healthy' subjects were defined as those without hypertension, diabetes mellitus, hyperlipidaemia, prior smoking history, atherosclerotic cardiovascular disease (ASCVD), heart failure and myocardial infarction (MI) as defined by international classification of diseases 9 (ICD9) and ICD10 codes.

Survival and statistical analyses

All analyses are sex stratified. Sex discordance score quartiles were defined using the distributions in the test set, stratified by sex, for Kaplan-Meier curve visualisation. Sex discordance score quartiles were plotted, and statistical significance assessed using the log rank test. Cox model analyses used the first ECG per subject. The sex discordance score variable was standardised, so hazard ratios reflect a one standard deviation change in sex discordance score. Cox models were fit using the test dataset for BIDMC or UK Biobank datasets and adjusted for age, hypertension, hyperlipidaemia, diabetes mellitus, smoking history, heart failure, prior ASCVD and prior MI (as defined by ICD9/10 codes). Subjects were censored at the time of death, last in-person hospital contact (BIDMC) or the UKB national censoring dates.
In analyses of future disease (heart failure and MI) subjects with the prevalent disease being studied were excluded. Primary analysis was using a Cox model adjusted for the same covariates as above, while secondary analysis used a Fine and Gray model accounting for the competing risk of death.
Statistical analyses were performed with R 4.2.0 statistical package (R Core Team, Vienna, Austria) or Python (version 3.9).

Phenome-wide association studies

To better understand the biology underlying sex discordance score I performed phenome-wide association studies (PheWAS), to evaluate the association with sex discordance score with phenotypes of interest. I used two approaches.
Firstly, I used a Disease PheWAS to explore the association of sex discordance score with both prevalent and incident diseases. Logistic regression, adjusted for age

and age2, was performed to investigate this association using a random ECG per subject.

The second approach, for analysis of continuous traits, was applied to the UK Biobank that contains data from over 3000 phenotypes derived from patient measurements, surveys, and investigations. Additionally, I applied this approach to the subjects with echocardiograms within 60 days of the ECG in the BIDMC test set, in whom I investigated the association of sex discordance score with echocardiographic parameters. Univariate correlation was performed to investigate the association between sex discordance score and phenotypes or cardiac imaging parameters (cardiac magnetic resonance imaging (MRI) and echocardiography) as previously described (80). Using linear regression, sex discordance score was adjusted for age, age2, height, weight, body surface area (BSA). In the UK Biobank waist circumference was additionally included as a covariate, this data was not available for the BIDMC cohort. Cardiac MRI parameters were derived as previously described (80).

Genome-wide association study

To identify genetic associations with sex discordance score, we performed a genome-wide association study (GWAS) in the UK Biobank for males and females separately. This analysis was restricted to the autosomes. The data were normalized by rank-based inverse normal transform prior to the analysis. GWAS methods are described in detail in Chapter 2.

Results

AI-ECG accurately classified sex from the ECG

BIDMC cohort demographics are described in detail in chapters 3 and 4. In the whole BIDMC test set, AI-ECG accurately identified sex (AUROC 0.943 (0.942-0.943). Performance was reduced in higher-risk cohorts (age over 40, inpatients and patients with prevalent cardiovascular disease (CVD)) and in ECGs with abnormalities (broad QRS complex, tachycardia) (**Figure 6-1, Table 6-1**).

Table 6-1

AI-ECG model sex classification performance as reported by area under the receiver operating characteristics curve for subgroups of the BIDMC dataset. 'Healthy' subjects were defined as those without hypertension, diabetes mellitus, hyperlipidaemia, prior smoking history, atherosclerotic cardiovascular disease, heart failure and myocardial infarction (MI). Prevalent CVD was defined as prior atherosclerotic cardiovascular disease, heart failure or myocardial infarction.

Category	**n ECGs (%)**	**AUROC**	**95% CI**
Normal	13377 (2.9)	0.967	0.964 - 0.969
Broad QRS	68117 (14.6)	0.891	0.889 - 0.894
Tachycardia	54144 (11.6)	0.894	0.891 - 0.897
Healthy	94937 (20.3)	0.953	0.952 - 0.954
Prevalent CVD	238718 (51.1)	0.930	0.929 - 0.931
Outpatient	190527 (40.8)	0.963	0.962 - 0.963
Inpatient	276653 (59.2)	0.926	0.926 - 0.927
Age under 40	44944 (9.6)	0.964	0.963 - 0.966
Age over 40	422236 (90.4)	0.940	0.939 - 0.940

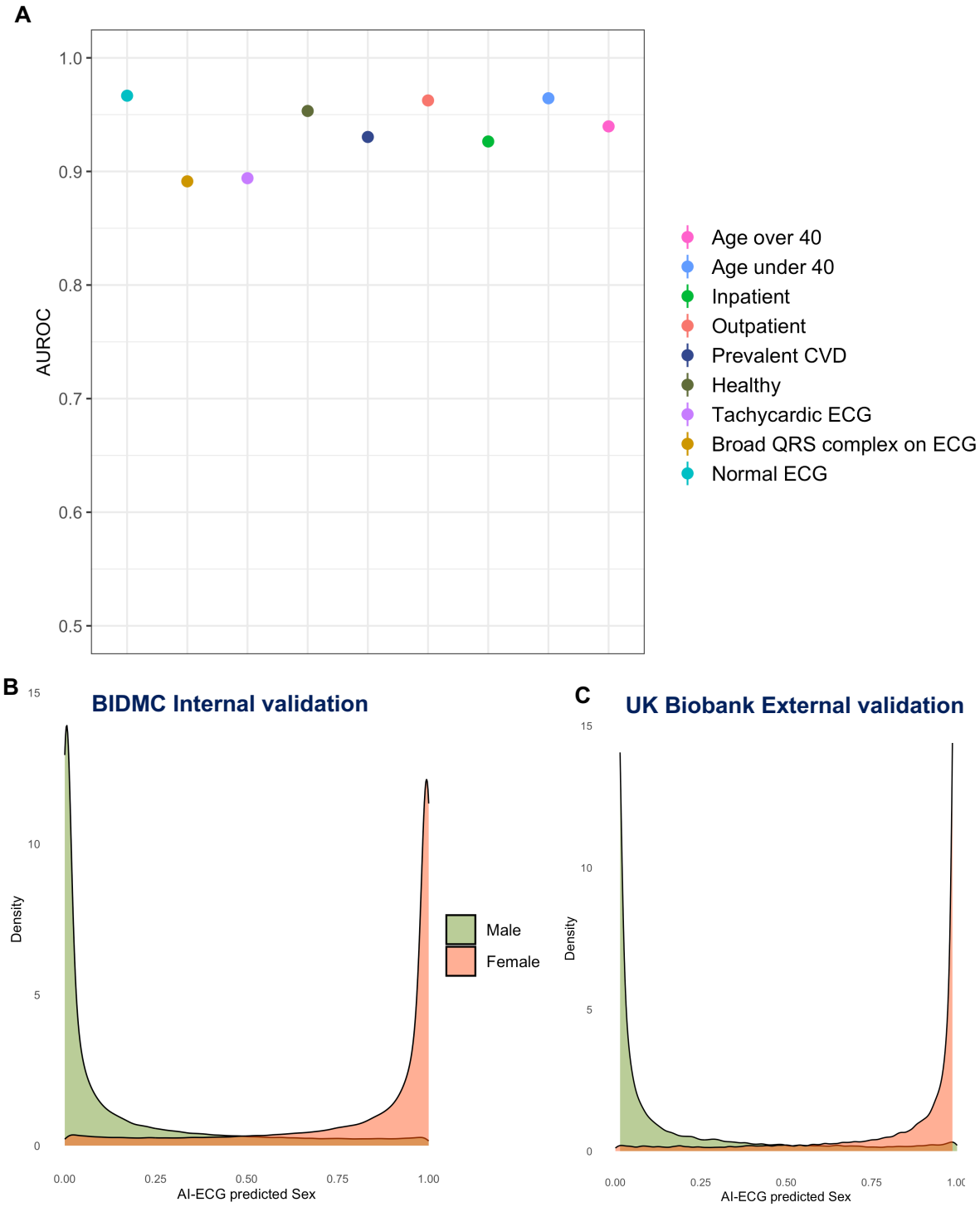


Figure 6-1 AI-ECG sex prediction performance metrics

Panel A shows classification performance as reported by area under the receiver operating characteristics curve for subgroups of the BIDMC dataset. ‘Healthy’

subjects were defined as those without hypertension, diabetes mellitus, hyperlipidaemia, prior smoking history, atherosclerotic cardiovascular disease, heart failure and myocardial infarction (MI). Prevalent CVD was defined as prior atherosclerotic cardiovascular disease, heart failure or myocardial infarction. Panels B and C show the distribution of AI-ECG sex prediction according to the biological sex in the BIDMC and UK Biobank respectively. CVD: cardiovascular disease

I externally tested the model in a cohort of volunteers from the UK Biobank. AI-ECG sex prediction performance was maintained in the UK Biobank population (AUROC 0.971 (0.969-0.972). AI-ECG sex prediction distributions in both cohorts are shown in **Figure 6-1B & C** and demonstrate excellent sex prediction accuracy, particularly at the extremes of the distributions (where sex prediction <0.2 or >0.8 accuracy was 93.1% in BIDMC and 95.5% in UKB), and predictive accuracy is not different between males and females.

Explainable ECG morphologies are associated with AI-ECG sex predictions

In order to explore the ECG morphological changes responsible for the AI-ECG sex predictions, I used a VAE to visualise the features most correlated with the AI-ECG sex predictions. **Figure 6-2A** shows the three latent features most highly correlated with AI-ECG sex predictions. In a second approach, we created median beats from the 10-second ECGs in the BIDMC test set and created average ± standard deviation representations of groups of ECGs. **Figure 6-2B** shows average representations of the 1000 ECGs with the ‘most male’ and ‘most female’ AI-ECG sex predictions.

Finally, I assessed the correlation of established ECG parameters with AI-ECG sex prediction **Figure 6-3**. These three methods highlight QRS duration, T wave morphology, QT interval, heart rate and QRS voltage as the most important factors in the derivation of AI-ECG sex prediction.

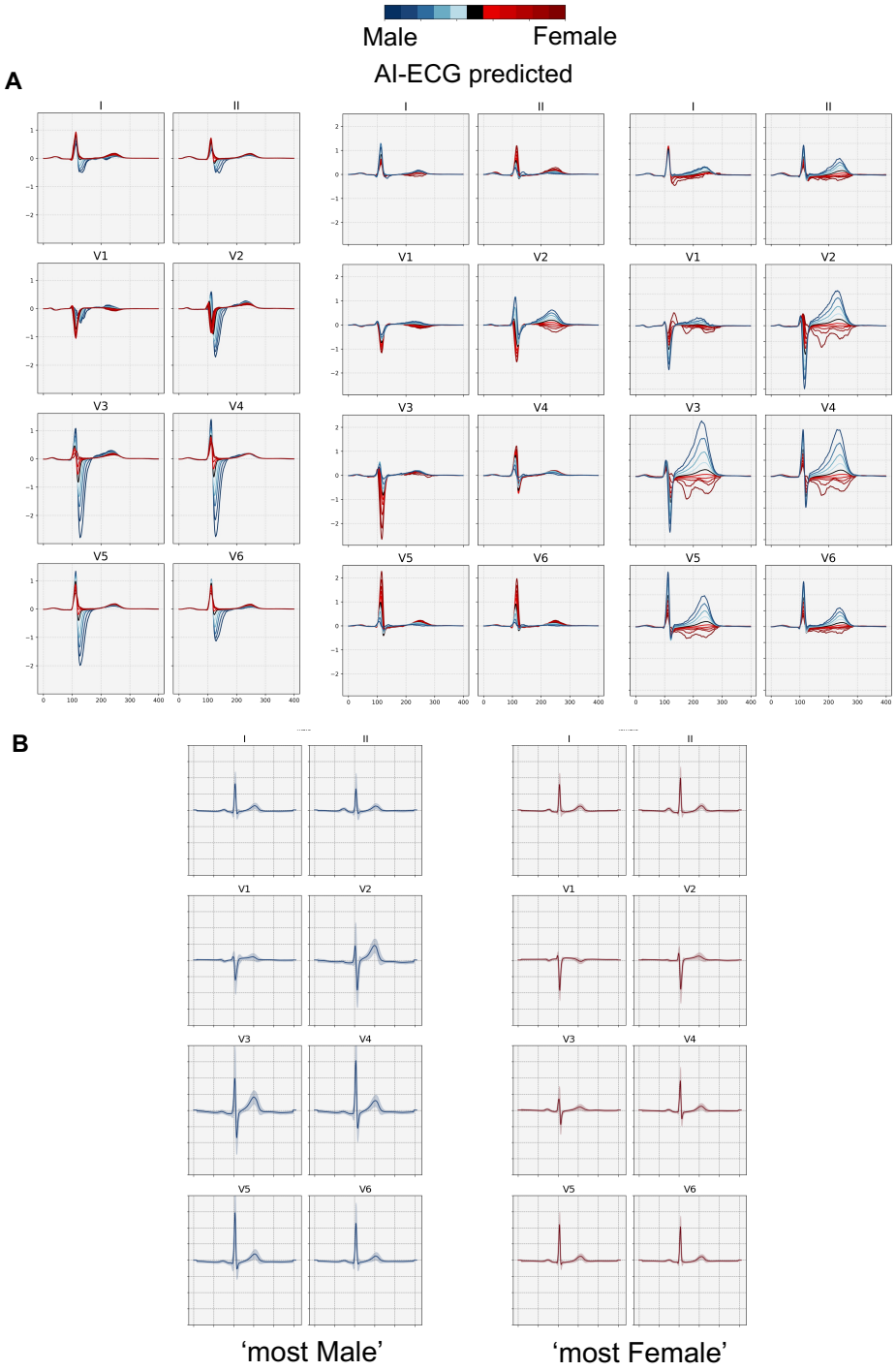


Figure 6-2 ECG morphologies associated with AI-ECG sex prediction

A) A variational auto-encoder was used to identify the most important morphological features in AI-ECG sex prediction, each subpanel shows one of three latent features. (B) Average ± standard deviation (shaded region) ECG waveforms for the 1000 highest and lowest AI-ECG sex predictions (i.e. 'most female' and 'most male') from

the BIDMC test set. QRS duration, voltage, T wave morphology and QT interval were identified as the most important morphologies associated with AI-ECG sex prediction.

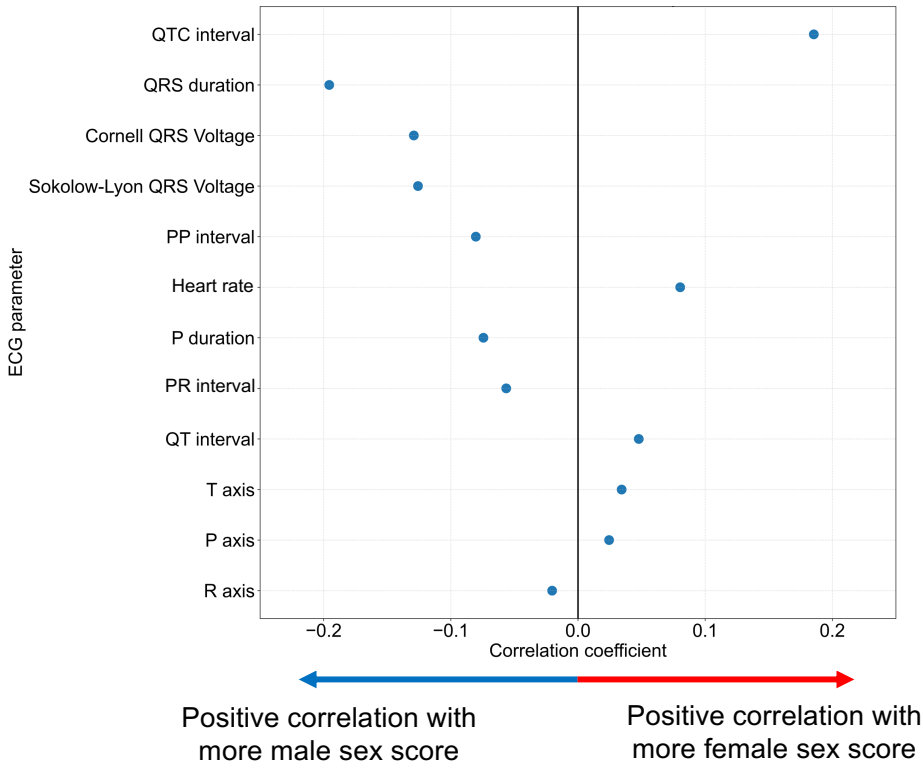


Figure 6-3 AI-ECG sex prediction correlations with ECG parameters

Univariate correlation was performed to evaluate the association of AI-ECG sex prediction with common ECG parameters. Positive correlation coefficients indicate positive correlation with 'more female' sex prediction. P value for all correlations < 0.001

Sex discordance score reveals female specific cardiovascular risk

Using the BIDMC cohort test set, I investigated the association between discordant sex predictions and all-cause mortality. Separately for males and females, quartiles of sex discordance score were identified. When analysing ECGs from the entire

cohort, which included a mixture of both high- and low-risk individuals, increased sex discordance score was associated with increased all-cause mortality in both sexes (**Figure 6-4**).

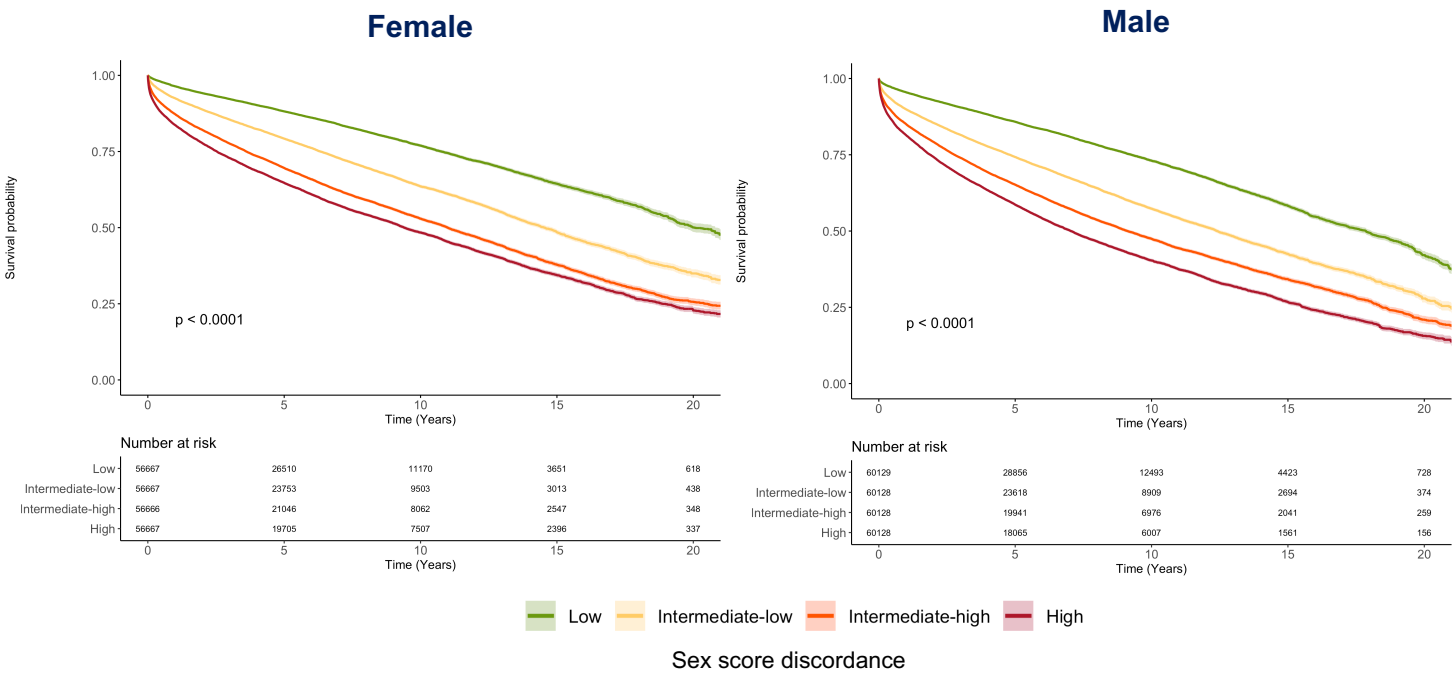


Figure 6-4 Sex discordance score survival analysis in unselected BIDMC test set

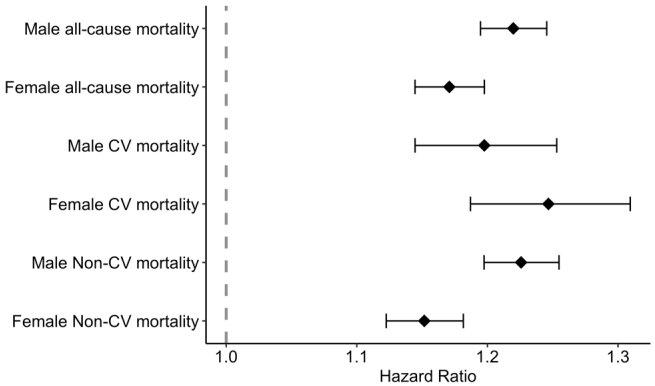
Kaplan-Meier curves showing increased all-cause mortality with higher sex discordance score in both males and females.

I further investigated this association with covariate-adjusted Cox models. Analysis was sex stratified, therefore in each case the reference group is other subjects of the same sex (i.e. in male analyses, the reference group is only other males). In the unselected BIDMC cohort, I found increased sex discordance score was associated with significantly greater all-cause, cardiovascular, and non-cardiovascular mortality, again in both sexes (**Figure 6-5A**). This may reflect the fact that the AI model is less able to accurately classify sex in higher-risk individuals with abnormal ECGs where the ECG changes relating to pathology are disproportionately greater than the subtle sex-specific ECG morphologies.

I therefore proceeded to evaluating ECGs from lower-risk populations only, which revealed contrasting findings. I evaluated the BIDMC outpatients with normal ECGs, 'healthy' individuals with normal ECGs and patients aged 40 or under. Here, increased sex discordance score was associated with an age-adjusted increased risk of CV death only in females, but not males (**Figure 6-5B-D**). In evaluating the UK Biobank cohort, which is a low-risk volunteer cohort, the same pattern was seen (**Figure 6-4E**).

A

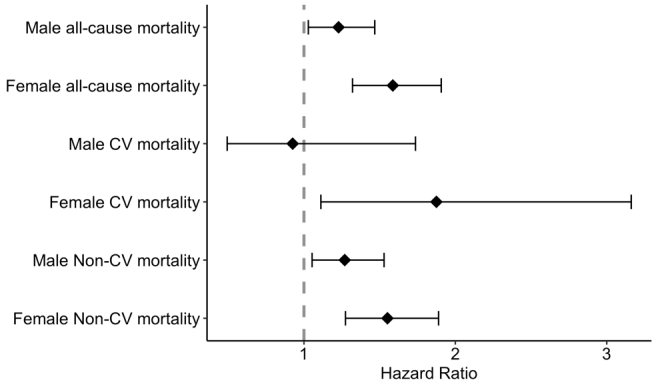
Unselected BIDMC



	HR	95% CI (%)	p value
Male all-cause mortality	1.22	1.19 to 1.25	<0.0001
Female all-cause mortality	1.17	1.14 to 1.20	<0.0001
Male CV mortality	1.20	1.14 to 1.25	<0.0001
Female CV mortality	1.25	1.19 to 1.31	<0.0001
Male Non-CV mortality	1.23	1.20 to 1.25	<0.0001
Female Non-CV mortality	1.15	1.12 to 1.18	<0.0001

B

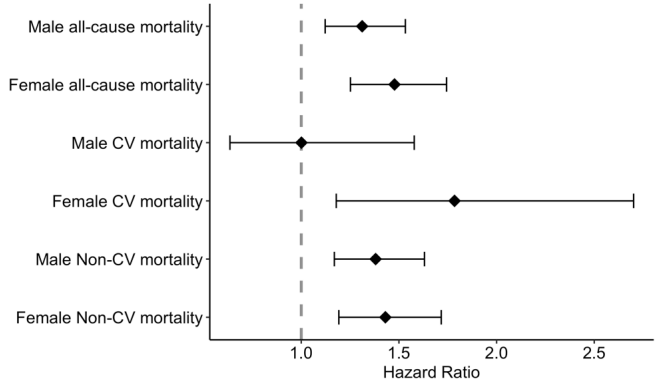
'Healthy' patients with Normal ECGs - BIDMC



	HR	95% CI (%)	p value
Male all-cause mortality	1.23	1.03 to 1.47	0.023
Female all-cause mortality	1.59	1.32 to 1.91	<0.0001
Male CV mortality	0.93	0.49 to 1.74	0.809
Female CV mortality	1.87	1.11 to 3.16	0.018
Male Non-CV mortality	1.27	1.05 to 1.53	0.012
Female Non-CV mortality	1.55	1.27 to 1.89	<0.0001

C

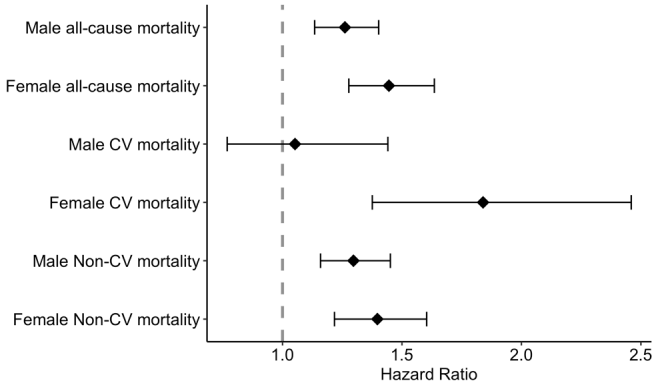
Outpatients with Normal ECGs - BIDMC



	HR	95% CI (%)	p value
Male all-cause mortality	1.31	1.12 to 1.53	0.001
Female all-cause mortality	1.48	1.25 to 1.74	<0.0001
Male CV mortality	1.00	0.63 to 1.58	0.996
Female CV mortality	1.78	1.18 to 2.70	0.006
Male Non-CV mortality	1.38	1.17 to 1.63	<0.0001
Female Non-CV mortality	1.43	1.19 to 1.72	<0.0001

D

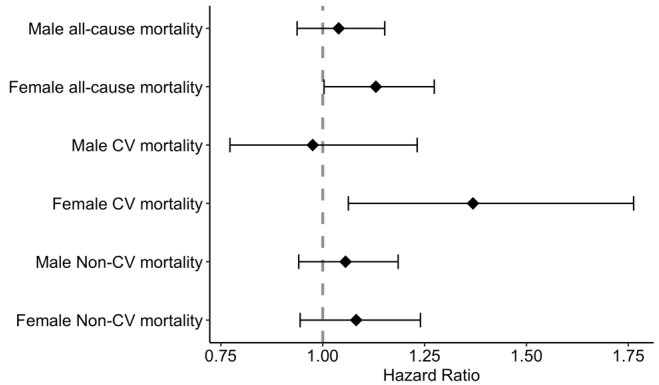
Patients aged 40 or under - BIDMC



	HR	95% CI (%)	p value
Male all-cause mortality	1.26	1.13 to 1.40	<0.0001
Female all-cause mortality	1.45	1.28 to 1.64	<0.0001
Male CV mortality	1.05	0.77 to 1.44	0.75
Female CV mortality	1.84	1.38 to 2.46	<0.0001
Male Non-CV mortality	1.30	1.16 to 1.45	<0.0001
Female Non-CV mortality	1.40	1.22 to 1.60	<0.0001

E

UK Biobank



	HR	95% CI (%)	p value
Male all-cause mortality	1.04	0.94 to 1.15	0.467
Female all-cause mortality	1.13	1.00 to 1.27	0.044
Male CV mortality	0.98	0.77 to 1.23	0.834
Female CV mortality	1.37	1.06 to 1.76	0.015
Male Non-CV mortality	1.06	0.94 to 1.18	0.354
Female Non-CV mortality	1.08	0.94 to 1.24	0.255

Figure 6-5 Association of sex discordance score with all-cause, cardiovascular, and non-cardiovascular mortality

Covariate-adjusted, sex stratified Cox models were fit to evaluate the association of sex discordance score (continuous) on all-cause, cardiovascular (CV) and non-cardiovascular mortality in the (A) unselected BIDMC cohort, (B) BIDMC 'Healthy' patients with normal ECGs (C) BIDMC outpatients with normal ECGs (D) BIDMC patients aged under 40 and (E) UK Biobank. Association between sex discordance score and CV mortality was observed in females but not in males in the latter four groups.

Sex discordance score is associated with differential phenotypes in males and females

In order to further evaluate the phenotypic differences associated with sex discordance score, I investigated correlations with echocardiographic features. I found differential phenotypic associations between the sexes. In females, greater sex discordance score was associated with reduced left ventricular ejection fraction (LVEF), and increased left ventricular (LV), right ventricular (RV), left atrial (LA) and aortic size and increased LV wall thickness, while in males, greater sex discordance score was associated with the opposite findings, in particular reduced aortic dimensions (**Figure 6-6**).

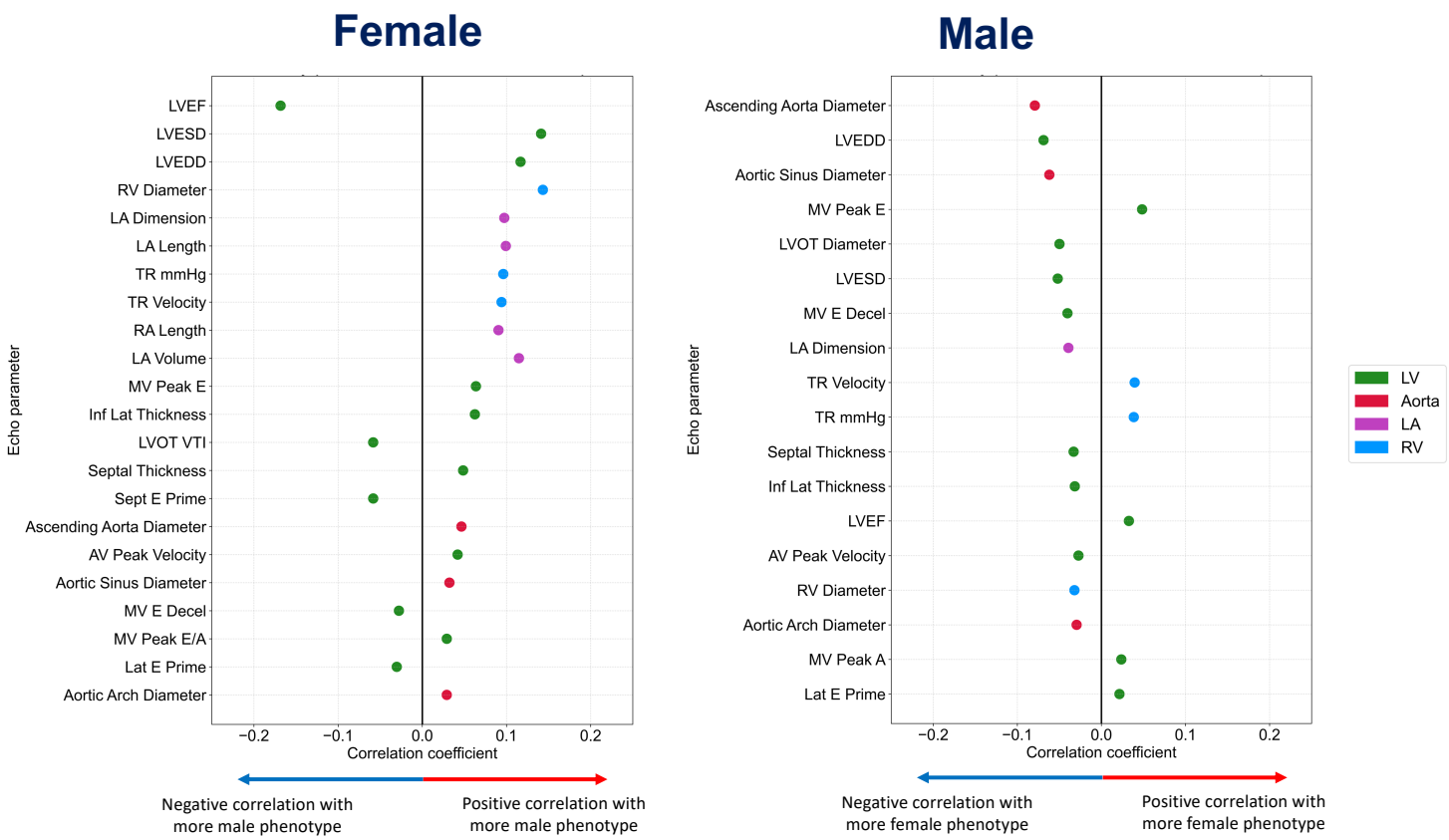


Figure 6-6 Association of sex discordance score with echocardiographic parameters

Sex stratified, univariate correlation of sex discordance score (continuous) and echocardiographic parameters was performed in the BIDMC test set. Comparisons meeting significance after Bonferroni correction are shown. LA: left atrium, TR: tricuspid regurgitation, MV: mitral valve, LVESD: LV end-systolic diameter, LVEDD: LV end-diastolic diameter, RA: right atrium, AV: aortic valve, LVOT: left ventricular outflow tract, LVEF: left ventricular ejection fraction.

I additionally performed a PheWAS using the UK Biobank. Cardiac MRI findings were similar to the echo findings described above (**Figure 6-7**).

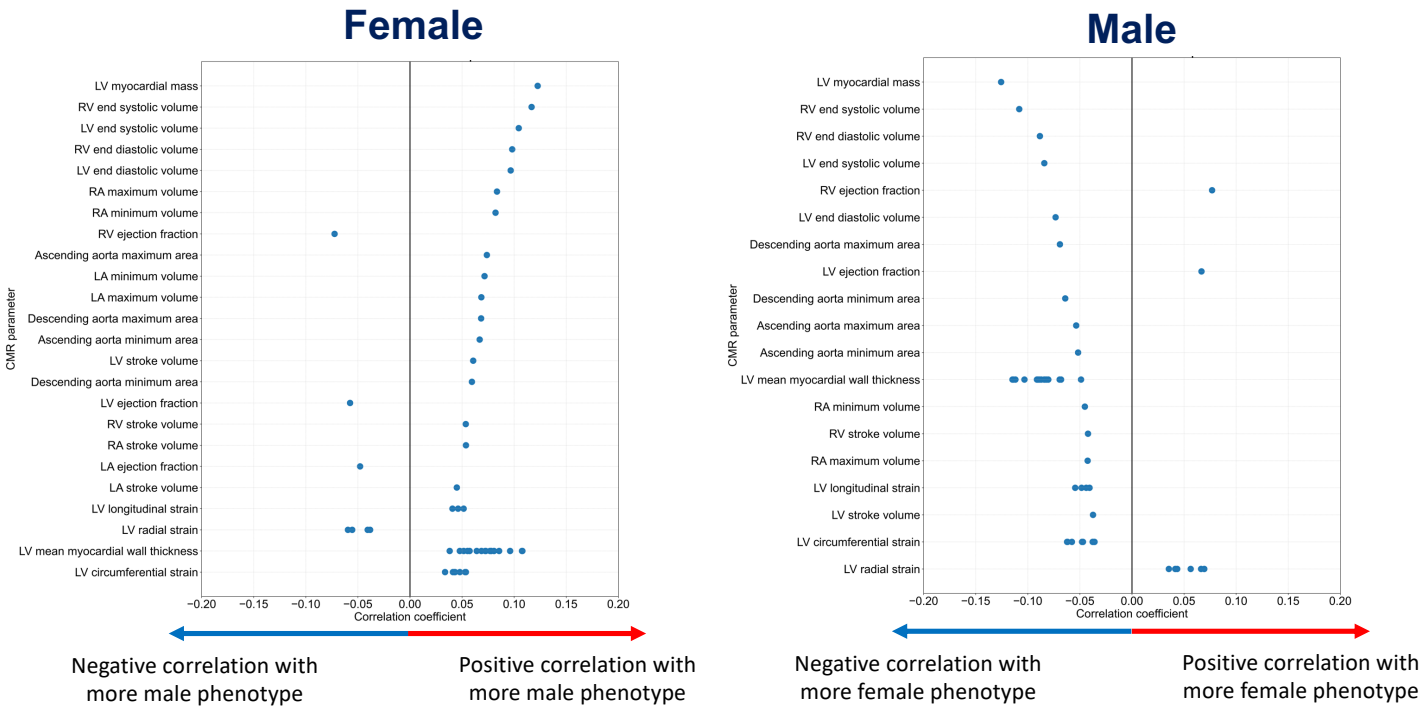


Figure 6-7 Association of sex discordance score with cardiac MRI parameters

Sex stratified, univariate correlation of sex discordance score (continuous) and cardiac MRI parameters was performed in the UK Biobank. Variables with multiple points indicate results of multiple measurements at varying locations. Comparisons meeting significance after Bonferroni correction are shown. MRI: magnetic resonance imaging, LA: left atrium, RA: right atrium, LV: left ventricle, RV: right ventricle.

Non-cardiac associations of greater sex discordance score in females included reduced body impedance (suggesting increased muscular mass) and body fat percentage, increased handgrip strength and bone area. In men these included, reduced brain volumes and bone area, increased fat mass and body impedance (**Figure 6-8**).

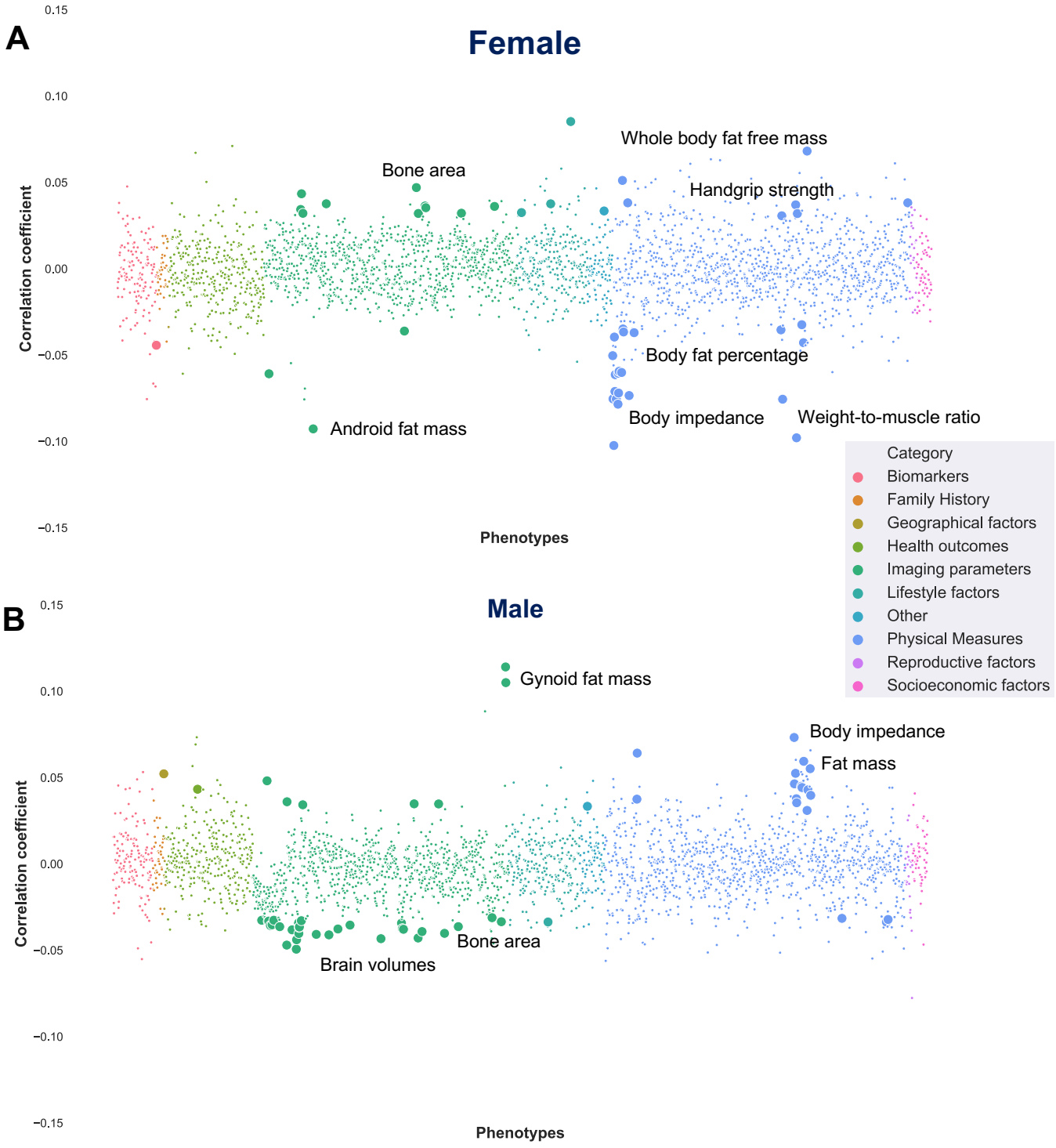


Figure 6-8 UK Biobank Phenome-wide association study (PheWAS)

Manhattan plot showing the univariate correlation P-value between AI-ECG sex score discordance and UK Biobank parameters. Larger dots indicate associations crossing the threshold for statistical significance after Bonferroni correction. (A) Females (B) Males.

Sex discordance score is associated with greater cardiovascular disease in females

In order to investigate the diseases associated with sex discordance score, I performed PheWAS of sex discordance score against 2304 Phecodes in the BIDMC test set. Again, I found differential disease associations between the sexes. The most significant associations in females with high sex discordance score were broadly cardiovascular and respiratory diseases, including heart failure, myocardial infarction (MI), cardiomyopathy and respiratory failure (**Figure 6-9**). In contrast, the strongest associations in males with high sex discordance score were broadly non-cardiac conditions including liver pathologies.

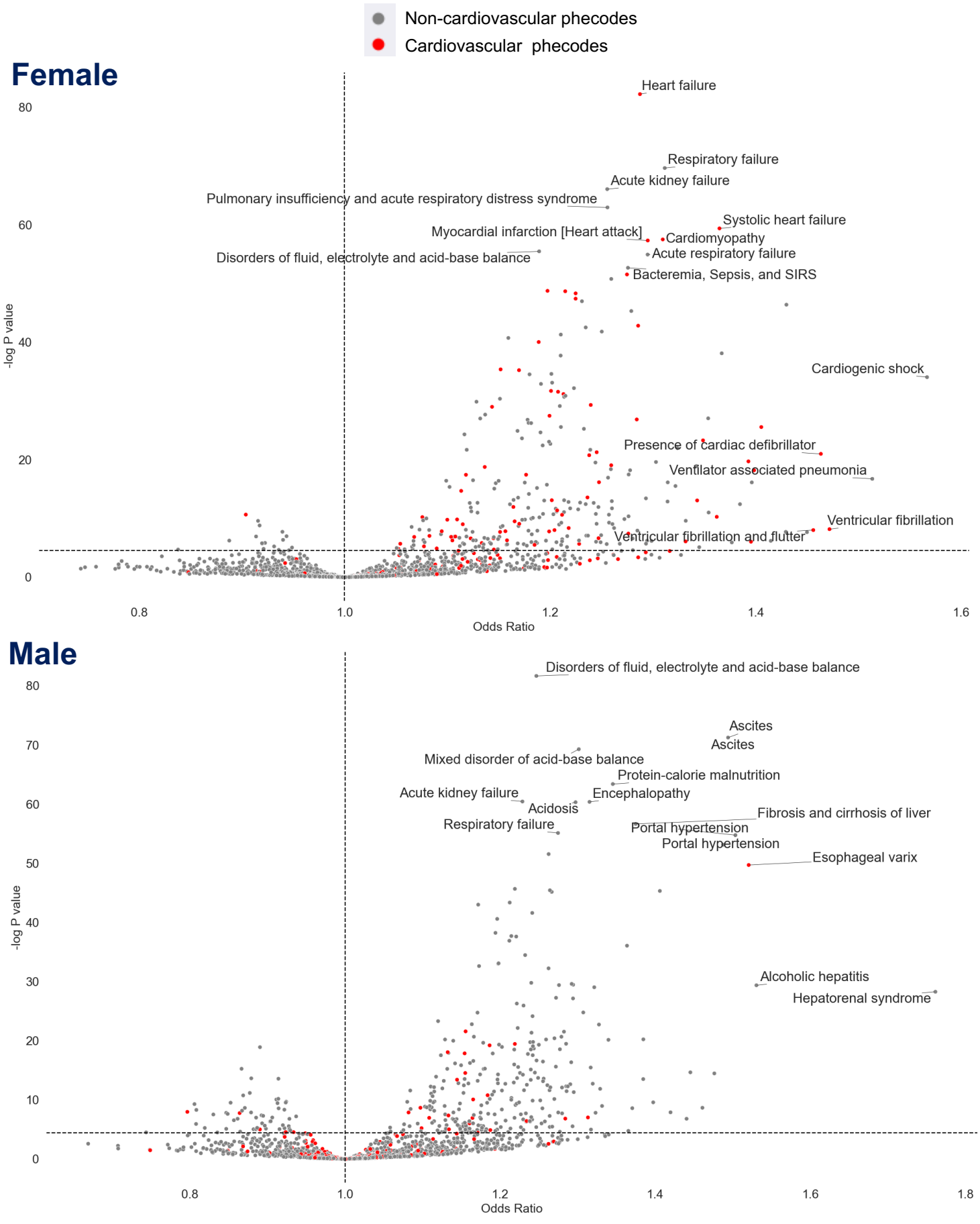

Non-cardiovascular phecodes
Cardiovascular phecodes
Female
Heart failure
Respiratory failure
Acute kidney failure
Pulmonary insufficiency and acute respiratory distress syndrome
Systolic heart failure
Myocardial infarction [Heart attack]
Cardiomyopathy
Disorders of fluid, electrolyte and acid-base balance
Acute respiratory failure
Bacteremia, Sepsis, and SIRS
Cardiogenic shock
Presence of cardiac defibrillator
Ventilator associated pneumonia
Ventricular fibrillation
Ventricular fibrillation and flutter
-log P value
Odds Ratio
Male
Disorders of fluid, electrolyte and acid-base balance
Ascites
Ascites
Mixed disorder of acid-base balance
Protein-calorie malnutrition
Acute kidney failure
Encephalopathy
Acidosis
Fibrosis and cirrhosis of liver
Respiratory failure
Portal hypertension
Portal hypertension
Esophageal varix
Alcoholic hepatitis
Hepatorenal syndrome
-log P value
Odds Ratio

Figure 6-9 Disease phenome-wide association study

Manhattan plot showing negative logarithm of the univariate correlation P-value between sex discordance score (continuous) and BIDMC disease Phecodes. Red dots indicate cardiac Phecodes, while grey dots indicate non cardiac. Horizontal dashed line indicates threshold for statistical significance after Bonferroni correction. Vertical dashed line indicates odds ratio of 1 (i.e. no association).

Next, I investigated the association of sex discordance score with future disease in BIDMC subjects with normal ECGs, again finding different associations between the sexes. In covariate-adjusted Cox models (covariates listed in Methods), sex discordance score was significantly associated with future heart failure in females but not males (**Table 6-2**). The same pattern was seen for MI (**Table 6-2**). In the UK Biobank a similar pattern was seen for future heart failure, although not meeting statistical significance (**Table 6-2**). There were no differences in future MI (**Table 6-2**). Secondary analyses using a Fine and Gray model adjusted for the competing risk of death yielded similar results (**Table 6-3**).

Table 6-2

Covariate-adjusted cox models for future heart failure/MI, hazard ratios are for standardised sex score discordance

Disease group	**HR Female**	**p value**	**HR Male**	**p value**
Future Heart failure				
BIDMC – patients with normal ECGs	1.211 (1.074-1.366)	0.002	1.126 (0.948-1.338)	0.176
BIDMC Outpatients with normal ECGs	1.194 (1.003-1.421)	0.045	1.017 (0.757-1.367)	0.909
UK Biobank	1.157 (0.972-1.376)	0.101	1.0465 (0.907-1.207)	0.532
Future MI				
BIDMC – patients with normal ECGs	1.316 (1.113-1.555)	0.001	0.967 (0.787-1.188)	0.749
BIDMC Outpatients with normal ECGs	1.162 (0.849 - 1.592)	0.348	0.958 (0.674 - 1.360)	0.809
UK Biobank	0.953 (0.740-1.229)	0.714	1. .057 (0.918-1.218	0.437

Table 6-3

Covariate-adjusted competing-risk regression (Fine-Gray) models for all-cause mortality, hazard ratios are for standardised sex score discordance

Disease group	**HR Female**	**p value**	**HR Male**	**p value**
Future Heart failure				
BIDMC – patients with normal ECGs	1.175 (1.038-1.329)	0.011	1.117 (0.944-1.322)	0.199
BIDMC Outpatients with normal ECGs	1.156 (0.969-1.380)	0.108	1.025 (0.764-1.375)	0.871
UK Biobank	1.155 (0.985-1.355)	0.076	1.045 (0.910-1.201)	0.533
Future MI				
BIDMC – patients with normal ECGs	1.315(1.127-1.535)	<0.001	0.901 (0.711-1.143)	0.392
BIDMC Outpatients with normal ECGs	1.228 (0.920-1.640)	0.165	0.889 (0.562-1.407)	0.615
UK Biobank	0.952 (0.726-1.250)	0.726	1.057 (0.918-1.217	0.443

Sex discordance score has genetic associations

To investigate if there were underlying genetic associations with sex discordance score, a genome-wide association study was performed (**Figure 6-10**). In females, we found significant loci adjacent to Insulin-like growth factor 1 receptor (*IGF1R*) and N-myc downregulated gene family member 4 (*NDRG4*), which have previously been associated with QRS duration, LV mass, and LV septal thickness, and sex hormone binding globulin, testosterone and QT interval respectively (200-203). In males, sex discordance score was associated with variants adjacent to Single stranded DNA binding protein 3 (*SSBP3*), Protein tyrosine phosphatase receptor type N2 (*PTPRN2*), IGF1R and Lipopolysaccharide induced TNF factor (*LITAF*). The associations are broadly similar to those for females, with the addition of lipid levels and bone density (204-206). In males, 11.0% of phenotypic variance was explained by genetic variation, while for females, this was 9.8%.

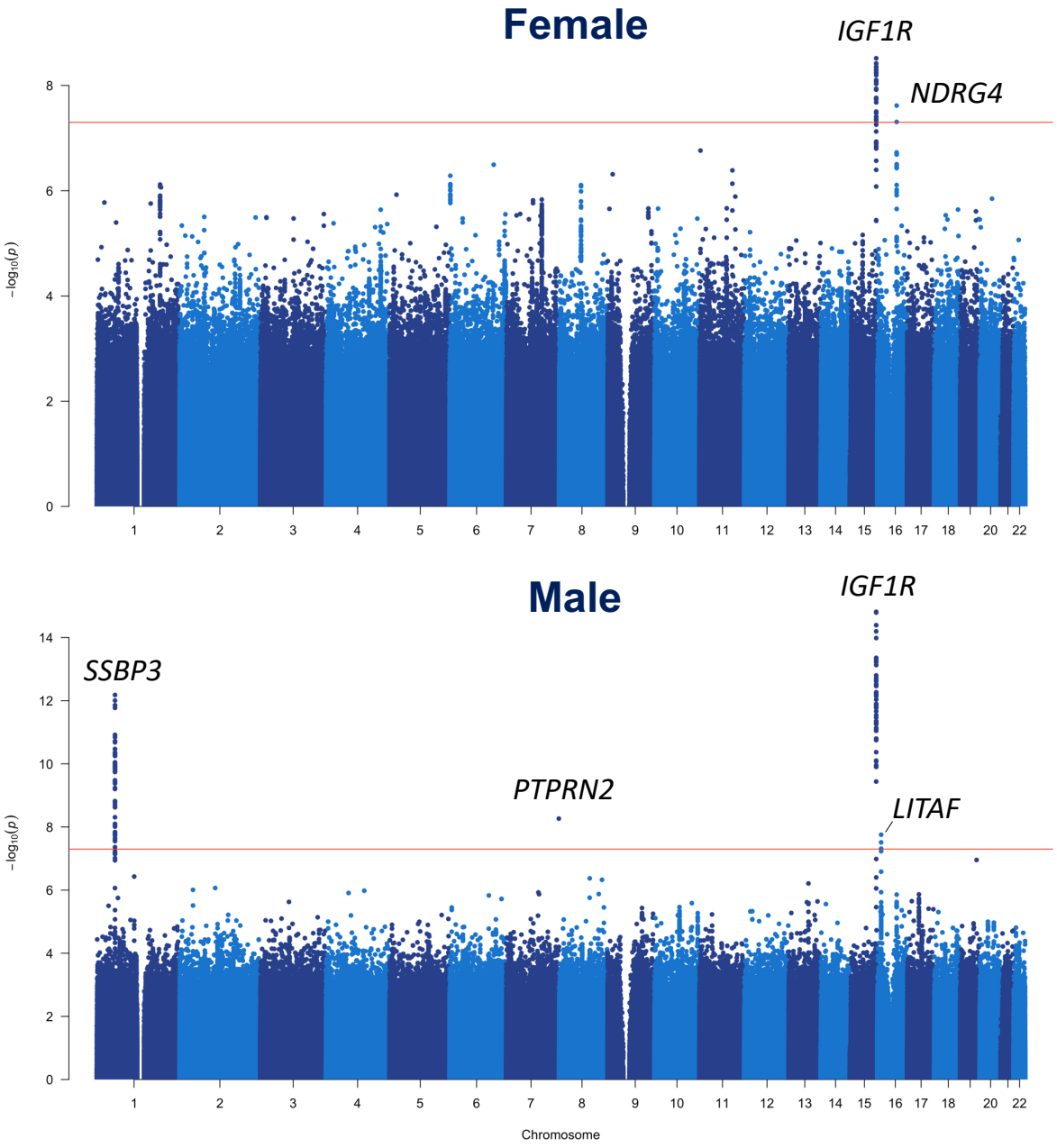


Figure 6-10 Genome-wide association study of sex discordance score

Manhattan plots of genomic loci associated with sex discordance score (continuous). Nearest genes to significant single nucleotide polymorphisms are shown. The red line depicts the genome-wide significant threshold (P<5 x 10-8).

Discussion

I describe a highly accurate AI-ECG model for sex classification, tested in two diverse populations, that identifies a novel digital biomarker of AI-ECG sex discordance score. Importantly, in lower-risk populations, discordance in AI-ECG-predicted sex and biological sex identifies higher cardiovascular risk in females but not in males.

AI-ECG predictions as a biomarker

As discussed in earlier chapters, AI-ECG predictions can be used not only for diagnosis and prediction of disease, but also as a biomarker that may inform on disease or subject phenotypes, disease trajectories and biological associations. In this chapter I have described AI-ECG predicted sex as a biomarker that can be used to identify females at elevated cardiovascular risk.

AI-ECG is highly accurate at classifying sex, particularly in lower-risk populations

I demonstrate that AI-ECG can identify sex highly accurately, as previously described (126, 199). The AI model performed well across all groups, but the model performance was greatest in lower-risk populations, where there are fewer ECG changes due to pathology. I specifically showed improved model performance in normal ECGs compared to ECGs with abnormalities. Any abnormalities on an ECG, however subtle, may obscure sex-specific ECG morphologies, thus making it more

challenging for the model to identify sex, which would explain the superior model performance in the lower-risk and more healthy subgroups. Through explainability analyses, I demonstrate with three methods that the ECG features most correlated with AI-ECG sex prediction were QRS duration, T wave morphology, QT interval, heart rate and QRS voltage, which are broadly in line with previous descriptions (198).

Sex discordance score is associated with differential phenotypes in males and females in lower-risk populations

Using the continuous sex classification model predictions, I developed the AI-ECG sex discordance score. Siegersma et al described binary 'sex misclassification' as being associated with higher mortality for both males and females (199). Their key finding was derived in a hospital population. Considering the whole cohort, increased sex discordance score in our study was also associated with worse outcomes in both sexes. The likely explanation for this observation is because higher risk individuals with worse outcomes have greater frequencies of subtle or overt ECG abnormalities, which reduce the model's ability to detect sex-specific ECG morphologies, leading to more sex misclassifications in this group.

I therefore subsequently specifically restricted our analyses to ECGs from healthy individuals and those with normal ECG, to investigate the biological significance of increased sex discordance score. I report a novel finding of sex differential effects, with females with higher sex discordance score having a greater CV mortality, without a similar pattern observed in males. This female-specific pattern was also tested in a second large cohort (UK Biobank). Our findings support our hypothesis,

that in lower-risk populations, increased sex discordance score is associated with adverse cardiovascular prognosis in females only, but not in males.

Biological exploration provides insights into mechanisms of adverse CV prognosis conferred by sex misclassification in females but not males

Through biological exploration in both cohorts, I found the AI-ECG sex discordance score to reflect physiological differences between sexes. Even after adjusting for body size, females with greater sex discordance score had a more male phenotype, with greater heart size, reduced fat mass and increased bone density. The opposite was true for males. These findings are consistent with previously described phenotypic differences between males and females (207, 208).

As males in general have an adverse cardiovascular prognosis compared to females (208), these findings support our observation that females with greater sex discordance score (and therefore a more male ECG phenotype) have a comparatively worse CV prognosis. As hypothesised, the opposite was not observed in males.

Interestingly, GWAS suggested that there are genetic determinants of sex discordance score that broadly reflect the phenotypes identified in the PheWAS. I therefore show that genetic modifiers on autosomal chromosomes may contribute to sex discordance score, which may relate to the increased risk seen in females with a more male ECG phenotype. Unsurprisingly, about 90% of the phenotypic variation in sex discordance score is not genetically determined.

Disease analyses highlight heart failure and myocardial infarction as potential contributors to the increased CV mortality seen in females with greater sex

discordance score. The associations with liver disease seen in males may be due to the established bidirectional association of liver disease with sex hormones (209). Overall, our findings support our hypothesis that there is a continuum of sex-related ECG morphologies, and the ECG contains information on sex-specific cardiovascular risk and biological differences beyond the physiological changes due to body shape and size.

AI-ECG may help overcome sex biases in identification and treatment of CV disease

Concerns have been expressed over AI having the potential to exacerbate existing societal, racial and sex biases (210). However, AI models may, instead of exacerbating biases, have the potential to help overcome them. Although cardiovascular disease rates are higher in men (208), cardiovascular disease is also the leading cause of death in females (211). Under the current paradigm, females receive more misdiagnosis and less preventive care and aggressive risk factor modification (193, 194, 212). This may be in part due to the perceived lower cardiovascular risk of females compared to males (195). The current dichotomous paradigm may in part be responsible for the incorrect generalisation that females are at low risk of cardiovascular disease, and in addressing this inequality in cardiovascular care, I propose a shift to an appreciation of sex-specific cardiovascular risk as a continuum rather than as a dichotomous trait when considering cardiovascular risk factors and risk stratification.

I have demonstrated that females with a more male ECG phenotype have adverse CV prognosis. A potential way to address the female disadvantage would be to

integrate sex discordance score into the electronic health records, and flag female subjects for whom an adverse (i.e. more male) ECG phenotype is detected. This personalised approach may help focus preventative treatments to those females at highest risk. The ECG is a ubiquitous and inexpensive investigation, already performed millions of times a day (213). Application of AI to most commonly performed cardiovascular investigation is a natural extension of the current clinical paradigm and may be used to overcome the widespread sex bias that exists in healthcare.

Limitations

These analyses focus on self-identified sex and therefore could not capture any complexities between self-identified sex and biological sex. I was unable to effectively study the associations between sex discordance score and sex hormone levels as the blood results were drawn several years in advance of the ECG recordings in the UK Biobank. I was unable to assess for the impact of significant female life events (menarche, pregnancy, menopause) on the findings.

Conclusion

I describe sex discordance score as a novel AI-ECG biomarker capable of identifying females with disproportionately elevated cardiovascular risk. Biological exploration has identified plausible associations with cardiac structure and body composition. Sex discordance score may help uncover potential underlying mechanisms of cardiovascular disease among females, and future clinical studies are needed to

determine its potential utility in identification of females at higher risk of cardiovascular events.

Chapter 7 - Conclusions

Summary of findings

In this thesis I have explored how AI-ECG can be used for cardiovascular risk prediction. I have used three broad methods: unsupervised machine learning (Chapter 3), supervised machine learning for specific clinical endpoints (Chapters 4 and 5), supervised machine learning for a demographic variable, used as a biomarker (Chapter 6).

These four studies all demonstrate the powerful ability of AI-ECG to predict and therefore risk stratify individuals at risk of cardiovascular diseases. I have also explored important and interesting biological determinants of cardiovascular risk, that can be unravelled with AI-ECG. In this chapter I will review the themes generated from my findings overall. These are:

1. Data-driven research is a powerful mechanism for hypothesis generation
2. AI-ECG has a potentially transformative ability to predict risk of cardiovascular disease
3. AI-ECG predictions are novel digital biomarkers that can be used to explore underlying biological associations and potential identify novel concepts.

Data-driven research

For centuries the scientific method has consisted of observation/question, hypothesis, experiment, conclusion. In contrast, data-driven research is a more recent paradigm (214). The advent of artificial intelligence has the potential to radically change society (for better and worse). However along with advent of AI, the ability to analyse massive amounts of data rapidly also has the potential to change how we approach scientific research. AI is essentially data-driven, it is well established that AI models are only as good as the data they are provided with. In the machine learning community, extensive hyperparameter tuning and architecture changes may be deployed and often return very small improvements that would not be considered meaningful in the clinical world. Instead changes to dataset size and quality labels is the main currency of developing and improving AI models.

The age of data is here and embracing that has the potential to change how discoveries are made. This principle is fundamental to machine learning. As discussed in the introduction, machine learning predicates taking data and (often) labels and using these to identify ‘rules’, these rules may in some cases be novel biological principles, not previously discovered. It is only by removing the constraints of human intuition that is often anchored to existing paradigms that we can make paradigm changing discoveries, like the seminal observation of AI-ECG for asymptomatic left ventricular systolic dysfunction detection (24).

In this thesis I have demonstrated two areas where data driven research has proved beneficial. Chapter 3 involved the application of unsupervised machine learning to identify novel ECG phenogroups. Through this data-driven approach I generated the hypothesis that AI-ECG could be trained, in a supervised way, for mortality and the

five important clinical endpoints evaluated in chapters 4 and 5. Secondly through an exploration of the AI-ECG sex score biomarker, I derived the hypothesis that sex related cardiovascular was non-dichotomous. It is important to note that data-driven research in general should be used to generate hypotheses. It is important to test these hypotheses robustly, as I have done in chapters 4-6. In this age of data and, with this type of approach I suspect that a data-driven approach to clinical research is here to stay.

Clinical relevance

The results of chapters 4 and 5 have the most direct clinical relevance. Risk prediction in clinical medicine is an important goal and AI-ECG has significant promise in being able to improve this. The proposed use of AIRE would be as a platform providing risk scores for mortality (both cardiovascular and non-cardiovascular), ASCVD, HF, VA, CHD and AF for all ECGs performed in healthcare settings. The clinical relevance of each prediction would need incorporation depending on the clinical scenario. Of course, prospective evaluation, ideally in a randomised fashion is needed before anything like this can be adopted. Some AI-ECG clinical trials have already been conducted and have been successful in demonstrating the benefits (215, 216).

The results in chapters 3 and 6 have less direct clinical relevance. Despite the promising observation that sex score discordance associates with adverse cardiovascular outcomes, this is a concept that requires significant further study

before any recommendations can be drawn from the findings. The phenogroups derived in chapter 2 are largely superseded by the findings in chapters 4 and 5.

AI-ECG predictions as a biomarker

Another major finding from this thesis is the potential utility of AI-ECG predictions as biomarkers. AI-ECG models, although often trained for a specific task, may capture important relevant information beyond that they were originally trained for. This was first demonstrated in chapter 3 that shows how NN-derived features could be transferred from a rhythm/conduction disease task, to identify phenogroups relevant to prognosis.

I built on this concept in Chapter 4, showing that a model trained to predict mortality also correlated with measures of biological aging and in fact may represent a biomarker of overall health, including biological age and the presence of clinical and subclinical disease.

Finally, and most interestingly, I found AI-ECG sex predictions were biomarkers of ECG ‘maleness’ or ‘femaleness’ and could be used derive the sex score discordance score that could identify female subjects at higher cardiovascular risk.

These findings together show that AI-ECG is not simply a tool for focussed disease diagnosis and prediction, but could be used to more broadly to act as a biomarker of subclinical disease or specific phenotypes.

Limitations

Although this work has identified several important findings, there are some limitations. Study specific limitations were discussed in the relevant chapters; therefore I address more general limitations here. AI-ECG, although excellent at predicting risk, is limited by the scope of what the ECG can capture. Although it turns out the ECG can accurately represent cardiac, non-cardiac disease, aging, and sex, it is clearly not a full representation of a subject or their risk of disease. In this age of data, many other data modalities exist and can provide important information for risk prediction. I have started to explore this in integrating AI-ECG with other easily accessible datapoints in cox models in chapters 4 and 5, however applying similar deep learning techniques to these other datapoints, in combination with AI-ECG has the potential to significantly improve on the risk prediction described in this thesis.
Throughout this thesis I have used digital ECGs for all analyses. Notably absent are clinical ECGs taken in the NHS. This is because very few NHS trusts have digital ECG capabilities. Herein lies a key challenge in AI-ECG - how to deliver this technology to most of the healthcare settings in the world, where digital ECG systems are lacking.
Although I have explored multiple approaches to explainability, including Grad-CAM, VAE features and median beat morphologies, it is important to note that all methods have limitations. In particular, the VAE features suffer from a degree of feature entanglement, despite efforts to mitigate this issue during model training. When creating a low dimensional feature space from a high dimensional input (such as an ECG), feature entanglement is inevitable, with perfect disentanglement leading to reduced ECG reconstruction performance. The entangled VAE features may not

allow clear interpretation with respect to human-interpretable concepts and may not allow clear ranking of important independent ECG morphological features.

The GWAS findings should be considered hypothesis generating given the absence of a replication dataset. The GWAS and UKB PheWAS results were drawn from a population of predominantly European ancestry and may not apply to other populations. The cohorts studied are not representative of the general population, in particular, the UK Biobank is known to represent and healthier than average subset of the population (90).

Finally, although these retrospective analyses of outcome prediction are compelling, we do not know if these trajectories can be altered without randomised trials of AI-ECG guided care.

Future directions

Expanding AIRE

In this thesis I have described several AIRE submodels based on common and clinically important cardiovascular diseases, particularly those that carry a high morbidity and mortality. Given previous work has shown the remarkable ability of AI-ECG to capture non-cardiac diagnoses (217) I plan to go on to train AIRE models for other important common diseases including hypertension, diabetes mellitus and valvular heart disease.

Applying AI-ECG to images of ECGs

A significant limitation of the work I have shown is the reliance on digital ECGs. In order to allow the application of this work to all healthcare settings I plan to develop techniques to handle images of ECGs. Potential strategies include directly using 2D images of ECGs (218) or digitising the images of ECGs into 1D signals and then applying the existing architectures I have used in this thesis. An additional challenge is applying AI-ECG to poor quality scans or images of ECGs, as are common in the NHS.

Integration of clinical, imaging, biochemical and longitudinal data

Although in this thesis I have shown the ECG has remarkable capabilities in risk prediction, there are clearly important features present in other data modalities that contribute to risk prediction. Additionally, for this thesis I include only a single ECG for risk prediction. Dynamic changes in the ECG have been shown to have additive value in risk prediction (219) and I would hypothesise deep learning could be superior previous approaches. In this thesis I have explored basic approaches to combining AI-ECG predictions with clinical, imaging and biochemical data in cox models. More complex approaches such as traditional machine learning to integrate these tabular variables with AI-ECG predictions may be useful to better take advantage of complex relationships. Additionally, trajectories in clinical diagnoses and ECGs together may yield further additive predictive value and could be integrated with architectures designed to handle temporal or sequence data such as recurrent neural networks and transformers.

Randomised trials of AI-ECG versus usual care

As with any medical intervention, or model of care, the gold standard for demonstrating efficacy is the randomised controlled trial (RCT). Important considerations for AI-ECG RCTs include demonstration of benefit to patients, either through reduction in hard outcomes (death, cardiovascular events), or reduction in investigation burden without an increase in adverse outcomes. Cost-effectiveness is an additional consideration that must be evaluated. AI-ECG may have the potential to help here to target investigations more effectively and reduce overall costs while maintaining safety. We are currently planning trials in several domains to investigate prospectively the benefits of AI-ECG.

Conclusions

The ECG is a ubiquitous, low-cost investigation that has been in use for decades. The advent of AI has the potential to truly revolutionise how we use the ECG. In this thesis I have demonstrated the diverse potential of AI-ECG, in phenotyping, predicting future disease, and as a biomarker. AI-ECG can be used not only in detecting and predicting disease, but also a tool to potentially identify novel biological associations for further investigation. Finally, in developing AIRE, I have created an AI-ECG platform that has the potential for widespread clinical use, for cardiovascular risk prediction.

References

1. AlGhatrif M, Lindsay J. A brief review: history to understand fundamentals of electrocardiography. J Community Hosp Intern Med Perspect. 2012;2(1).

2. HENSON JR. Descartes and the ECG Lettering Series. Journal of the History of Medicine and Allied Sciences. 1971;XXVI(2):181-6.

3. Hollman A. Thomas Lewis--the early years. British Heart Journal. 1981;46(3):233.

4. Silberbauer J. Wilson's Central Terminal, the Keystone to Electrogram Recording—What. Where and Why. 2013.

5. RECOMMENDATIONS for standardization of electrocardiographic and vectorcardiographic leads. Circulation. 1954;10(4):564-73.

6. Schläpfer J, Wellens Hein J. Computer-Interpreted Electrocardiograms. Journal of the American College of Cardiology. 2017;70(9):1183-92.

7. Webster P. Six ways large language models are changing healthcare. Nature Medicine. 2023;29(12):2969-71.

8. Wu T, He S, Liu J, Sun S, Liu K, Han Q-L, et al. A brief overview of ChatGPT: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica. 2023;10(5):1122-36.

9. Goertzel B. Artificial general intelligence: concept, state of the art, and future prospects. Journal of Artificial General Intelligence. 2014;5(1):1-48.

10. Federspiel F, Mitchell R, Asokan A, Umana C, McCoy D. Threats by artificial intelligence to human health and human existence. BMJ global health. 2023;8(5):e010435.

11. Desai AD, Yaw TS, Yamazaki T, Kaykha A, Chun S, Froelicher VF. Prognostic Significance of Quantitative QRS Duration. The American Journal of Medicine. 2006;119(7):600-6.

12. Rosenberg MA. Trusting Magic. Circulation. 2021;143(13):1299-301.

13. Ribeiro MT, Singh S, Guestrin C, editors. " Why should i trust you?" Explaining the predictions of any classifier. Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.

14. DeGrave AJ, Janizek JD, Lee S-I. AI for radiographic COVID-19 detection selects shortcuts over signal. Nature Machine Intelligence. 2021;3(7):610-9.

15. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: Visualising image classification models and saliency maps. arXiv preprint arXiv:13126034. 2013.

16. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. Advances in neural information processing systems. 2018;31.

17. McCoy LG, Brenna CTA, Chen S, Vold K, Das S. Believing in Black Boxes: Machine Learning for Healthcare Does Not Need Explainability to be Evidence-Based. J Clin Epidemiol. 2021.

18. Pérez-Pevida B, Escalada J, Miras AD, Frühbeck G. Mechanisms Underlying Type 2 Diabetes Remission After Metabolic Surgery. Front Endocrinol (Lausanne). 2019;10:641.

19. Toussaint K, Yang XC, Zielinski MA, Reigle KL, Sacavage SD, Nagar S, et al. What do we (not) know about how paracetamol (acetaminophen) works? J Clin Pharm Ther. 2010;35(6):617-38.

20. London AJ. Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report. 2019;49(1):15-21.

21. Smith SW, Walsh B, Grauer K, Wang K, Rapin J, Li J, et al. A deep neural network learning algorithm outperforms a conventional algorithm for emergency department electrocardiogram interpretation. J Electrocardiol. 2019;52:88-95.

22. Hannun AY, Rajpurkar P, Haghpanahi M, Tison GH, Bourn C, Turakhia MP, et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. Nat Med. 2019;25(1):65-9.

23. Ribeiro AH, Ribeiro MH, Paixao GMM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nat Commun. 2020;11(1):1760.

24. Attia ZI, Kapa S, Lopez-Jimenez F, McKie PM, Ladewig DJ, Satam G, et al. Screening for cardiac contractile dysfunction using an artificial intelligence-enabled electrocardiogram. Nat Med. 2019;25(1):70-4.

25. Cohen-Shelly M, Attia ZI, Friedman PA, Ito S, Essayagh BA, Ko WY, et al. Electrocardiogram screening for aortic valve stenosis using artificial intelligence. Eur Heart J. 2021;42(30):2885-96.

26. Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, et al. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. The Lancet. 2019;394(10201):861-7.

27. Attia AI, Lerman G, Friedman PA. Deep neural networks learn by using human-selected ECG features and novel features. Eur Heart J Digit Health. 2021;2:446-55.

28. Ko W-Y, Siontis KC, Attia ZI, Carter RE, Kapa S, Ommen SR, et al. Detection of Hypertrophic Cardiomyopathy Using a Convolutional Neural Network-Enabled

Electrocardiogram. Journal of the American College of Cardiology. 2020;75(7):722-33.

29. Grogan M, Lopez-Jimenez F, Cohen-Shelly M, Dispenzieri A, Attia ZI, Abou Ezzedine OF, et al. Artificial Intelligence–Enhanced Electrocardiogram for the Early Detection of Cardiac Amyloidosis. Mayo Clinic Proceedings. 2021;96(11):2768-78.

30. Adedinsewo DA, Johnson PW, Douglass EJ, Attia IZ, Phillips SD, Goswami RM, et al. Detecting cardiomyopathies in pregnancy and the postpartum period with an electrocardiogram-based deep learning model. European Heart Journal - Digital Health. 2021;2(4):586-96.

31. Feeny AK, Rickard J, Trulock KM, Patel D, Toro S, Moennich LA, et al. Machine Learning of 12-Lead QRS Waveforms to Identify Cardiac Resynchronization Therapy Patients With Differential Outcomes. Circ Arrhythm Electrophysiol. 2020;13(7):e008210.

32. Porumb M, Griffen C, Hattersley J, Pecchia L. Nocturnal low glucose detection in healthy elderly from one-lead ECG using convolutional denoising autoencoders. Biomedical Signal Processing and Control. 2020;62:102054.

33. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nature Medicine. 2021;27(5):815-9.

34. Bachtiger P, Petri CF, Scott FE, Ri Park S, Kelshiker MA, Sahemey HK, 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. The Lancet Digital Health. 2022:S2589750021002569.

35. Gaziano T, Reddy KS, Paccaud F, Horton S, Chaturvedi V. Cardiovascular disease. Disease Control Priorities in Developing Countries 2nd edition. 2006.

36. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990-2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020;76(25):2982-3021.

37. Arnett DK, Blumenthal RS, Albert MA, Buroker AB, Goldberger ZD, Hahn EJ, et al. 2019 ACC/AHA guideline on the primary prevention of cardiovascular disease: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. Circulation. 2019;140(11):e596-e646.

38. Birger M, Kaldjian AS, Roth GA, Moran AE, Dieleman JL, Bellows BK. Spending on cardiovascular disease and cardiovascular risk factors in the United States: 1996 to 2016. Circulation. 2021;144(4):271-82.

39. Ortega FB, Lavie CJ, Blair SN. Obesity and cardiovascular disease. Circulation research. 2016;118(11):1752-70.

40. Lippi G, Sanchis-Gomar F, Cervellin G. Global epidemiology of atrial fibrillation: An increasing epidemic and public health challenge. Int J Stroke. 2021;16(2):217-21.

41. Cowie MR. The heart failure epidemic: a UK perspective. Echo Res Pract. 2017;4(1):R15-r20.

42. Hippisley-Cox J, Coupland C, Vinogradova Y, Robson J, May M, Brindle P. Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. BMJ. 2007;335(7611):136.

43. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. European heart journal. 2021;42(25):2439-54.

44. Goff Jr 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. Circulation. 2014;129(25_#_2):S49-S73.

45. Lip GY, Nieuwlaat R, Pisters R, Lane DA, Crijns HJ. Refining clinical risk stratification for predicting stroke and thromboembolism in atrial fibrillation using a novel risk factor-based approach: the euro heart survey on atrial fibrillation. Chest. 2010;137(2):263-72.

46. Zhu W, He W, Guo L, Wang X, Hong K. The HAS-BLED score for predicting major bleeding risk in anticoagulated patients with atrial fibrillation: a systematic review and meta-analysis. Clinical cardiology. 2015;38(9):555-61.

47. Wilson PW, D'Agostino RB, Levy D, Belanger AM, Silbershatz H, Kannel WB. Prediction of coronary heart disease using risk factor categories. Circulation. 1998;97(18):1837-47.

48. Brindle P, Emberson J, Lampe F, Walker M, Whincup P, Fahey T, et al. Predictive accuracy of the Framingham coronary risk score in British men: prospective cohort study. BMJ. 2003;327(7426):1267.

49. DeFilippis AP, Young R, McEvoy JW, Michos ED, Sandfort V, Kronmal RA, et al. Risk score overestimation: the impact of individual cardiovascular risk factors and preventive therapies on the performance of the American Heart Association-American College of Cardiology-Atherosclerotic Cardiovascular Disease risk score in a modern multi-ethnic cohort. Eur Heart J. 2017;38(8):598-608.

50. Investigators BAATfAF. The effect of low-dose warfarin on the risk of stroke in patients with nonrheumatic atrial fibrillation. New England Journal of Medicine. 1990;323(22):1505-11.

51. Barra S, Providência R. Anticoagulation in atrial fibrillation. Heart. 2021;107(5):419-27.

52. Councils E. 2015 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death. European Heart Journal. 2015;36:2793-867.

53. Stecker EC, Vickers C, Waltz J, Socoteanu C, John BT, Mariani R, et al. Population-based analysis of sudden cardiac death with and without left ventricular systolic dysfunction: two-year findings from the Oregon Sudden Unexpected Death Study. Journal of the American College of Cardiology. 2006;47(6):1161-6.

54. Weng SF, Vaz L, Qureshi N, Kai J. Prediction of premature all-cause mortality: A prospective general population cohort study comparing machine-learning and standard epidemiological approaches. PLoS One. 2019;14(3):e0214365.

55. Hoeper MM, Kramer T, Pan Z, Eichstaedt CA, Spiesshoefer J, Benjamin N, et al. Mortality in pulmonary arterial hypertension: prediction by the 2015 European pulmonary hypertension guidelines risk stratification model. European Respiratory Journal. 2017;50(2).

56. Nchimi A, Dibato JE, Davin L, Schoysman L, Oury C, Lancellotti P. Predicting disease progression and mortality in aortic stenosis: a systematic review of imaging biomarkers and meta-analysis. Front Cardiovasc Med. 2018;5:112.

57. Alkhouli M, Alqahtani F, Aljohani S, Alvi M, Holmes DR. Burden of atrial fibrillation–associated ischemic stroke in the United States. JACC: Clinical Electrophysiology. 2018;4(5):618-25.

58. Patel NJ, Atti V, Mitrani RD, Viles-Gonzalez JF, Goldberger JJ. Global rising trends of atrial fibrillation: a major public health concern. BMJ Publishing Group Ltd and British Cardiovascular Society; 2018. p. 1989-90.

59. Dulli DA, Stanko H, Levine RL. Atrial fibrillation is associated with severe acute ischemic stroke. Neuroepidemiology. 2003;22(2):118-23.

60. Dilaveris PE, Kennedy HL. Silent atrial fibrillation: epidemiology, diagnosis, and clinical impact. Clin Cardiol. 2017;40(6):413-8.

61. Hermans AN, Gawalko M, Dohmen L, van der Velden RM, Betz K, Duncker D, et al. Mobile health solutions for atrial fibrillation detection and management: a systematic review. Clin Res Cardiol. 2022:1-13.

62. McIntyre WF, Benz AP, Becher N, Healey JS, Granger CB, Rivard L, et al. Direct oral anticoagulants for stroke prevention in patients with device-detected atrial fibrillation: a study-level meta-analysis of the NOAH-AFNET 6 and ARTESiA trials. Circulation. 2023.

63. Svennberg E, Friberg L, Frykman V, Al-Khalili F, Engdahl J, Rosenqvist M. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. The Lancet. 2021;398(10310):1498-506.

64. 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. The American Journal of Cardiology. 2006;98(5):644-8.

65. Verdecchia P, Angeli F, Reboldi G, Carluccio E, Benemio G, Gattobigio R, et al. Improved cardiovascular risk stratification by a simple ECG index in hypertension. American Journal of Hypertension. 2003;16(8):646-52.

66. Hughes JW, Tooley J, Torres Soto J, Ostropolets A, Poterucha T, Christensen MK, et al. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease. npj Digit Med. 2023;6(1):169.

67. Khurshid S, Friedman S, Reeder C, Di Achille P, Diamant N, Singh P, et al. ECG-Based Deep Learning and Clinical Risk Factors to Predict Atrial Fibrillation. Circulation. 2022;145(2):122-33.

68. Alonso A, Krijthe BP, Aspelund T, Stepas KA, Pencina MJ, Moser CB, et al. Simple risk model predicts incidence of atrial fibrillation in a racially and geographically diverse population: the CHARGE-AF consortium. J Am Heart Assoc. 2013;2(2):e000102.

69. Kwon JM, Kim KH, Jeon KH, Lee SY, Park J, Oh BH. Artificial intelligence algorithm for predicting cardiac arrest using electrocardiography. Scand J Trauma Resusc Emerg Med. 2020;28(1):98.

70. Sammani A, van de Leur RR, Henkens MTHM, Meine M, Loh P, Hassink RJ, et al. Life-threatening ventricular arrhythmia prediction in patients with dilated cardiomyopathy using explainable electrocardiogram-based deep neural networks EP Europace. 2022;24(10):1645-54.

71. Kolk MZH, Deb B, Ruipérez-Campillo S, Bhatia NK, Clopton P, Wilde AAM, et al. Machine learning of electrophysiological signals for the prediction of ventricular arrhythmias: systematic review and examination of heterogeneity between studies. EBioMedicine. 2023;89:104462.

72. Raghunath S, Ulloa Cerna AE, Jing L, vanMaanen DP, Stough J, Hartzel DN, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nature Medicine. 2020;26(6):886-91.

73. Sun W, Kalmady SV, Sepehrvand N, Salimi A, Nademi Y, Bainey K, et al. Towards artificial intelligence-based learning health system for population-level mortality prediction using electrocardiograms. npj Digit Med. 2023;6(1):21.

74. Ji L, Jazwinski SM, Kim S. Frailty and biological age. Annals of geriatric medicine and research. 2021;25(3):141.

75. Lima EM, Ribeiro AH, Paixão GMM, Ribeiro MH, Pinto-Filho MM, Gomes PR, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. Nature Communications. 2021;12(1):5117.

76. Li X, Patel KHK, Sun L, Peters NS, Ng FS. Neural networks applied to 12-lead electrocardiograms predict body mass index, visceral adiposity and concurrent cardiometabolic ill-health. Cardiovascular Digital Health Journal. 2021.

77. Kumar V, Gu Y, Basu S, Berglund A, Eschrich SA, Schabath MB, et al. Radiomics: the process and the challenges. Magn Reson Imaging. 2012;30(9):1234-48.

78. Vasudevan S, Saha A, Tarver ME, Patel B. Digital biomarkers: Convergence of digital health technologies and biomarkers. NPJ Digit Med. 2022;5(1):36.

79. Raisi-Estabragh Z, Salih A, Gkontra P, Atehortúa A, Radeva P, Boscolo Galazzo I, et al. Estimation of biological heart age using cardiovascular magnetic resonance radiomics. Scientific Reports. 2022;12(1):12805.

80. Bai W, Suzuki H, Huang J, Francis C, Wang S, Tarroni G, et al. A population-based phenome-wide association study of cardiac and aortic structure and function. Nature Medicine. 2020;26(10):1654-62.

81. Bastarache L, Denny JC, Roden DM. Phenome-Wide Association Studies. JAMA. 2022;327(1):75-6.

82. Shah M, de A. Inácio MH, Lu C, Schiratti P-R, Zheng SL, Clement A, et al. Environmental and genetic predictors of human cardiovascular ageing. Nature Communications. 2023;14(1):4941.

83. Wang X, Khurshid S, Choi SH, Friedman S, Weng LC, Reeder C, et al. Genetic Susceptibility to Atrial Fibrillation Identified via Deep Learning of 12-Lead Electrocardiograms. Circ Genom Precis Med. 2023;16(4):340-9.

84. Sau A, Ibrahim S, Ahmed A, Handa B, Kramer DB, Waks JW, et al. Artificial intelligence-enabled electrocardiogram to distinguish cavotricuspid isthmus dependence from other atrial tachycardia mechanisms. European Heart Journal - Digital Health. 2022:ztac042.

85. Sau A, Ibrahim S, Kramer DB, Waks JW, Qureshi N, Koa-Wing M, et al. Artificial intelligence–enabled electrocardiogram to distinguish atrioventricular re-entrant tachycardia from atrioventricular nodal re-entrant tachycardia. Cardiovascular Digital Health Journal. 2023;4(2):60-7.

86. Cardoso CS, Sabino EC, Oliveira CDL, de Oliveira LC, Ferreira AM, Cunha-Neto E, et al. Longitudinal study of patients with chronic Chagas cardiomyopathy in Brazil (SaMi-Trop project): a cohort profile. BMJ Open. 2016;6(5):e011181.

87. Cardoso CS, Ribeiro ALP, Oliveira CDL, Oliveira LC, Ferreira AM, Bierrenbach AL, et al. Beneficial effects of benznidazole in Chagas disease: NIH SaMi-Trop cohort study. PLoS Negl Trop Dis. 2018;12(11):e0006814.

88. Schmidt MI, Duncan BB, Mill JG, Lotufo PA, Chor D, Barreto SM, et al. Cohort Profile: Longitudinal Study of Adult Health (ELSA-Brasil). Int J Epidemiol. 2015;44(1):68-75.

89. 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. 2015;12(3):e1001779.

90. Fry A, Littlejohns TJ, Sudlow C, Doherty N, Adamska L, Sprosen T, et al. Comparison of Sociodemographic and Health-Related Characteristics of UK Biobank

Participants With Those of the General Population. Am J Epidemiol. 2017;186(9):1026-34.

91. Ardissino M, McCracken C, Bard A, Antoniades C, Neubauer S, Harvey NC, et al. Pericardial adiposity is independently linked to adverse cardiovascular phenotypes: a CMR study of 42 598 UK Biobank participants. Eur Heart J Cardiovasc Imaging. 2022;23(11):1471-81.

92. Marmot MG, Smith GD, Stansfeld S, Patel C, North F, Head J, et al. Health inequalities among British civil servants: the Whitehall II study. Lancet. 1991;337(8754):1387-93.

93. Hnatkova K, Andršová I, Novotný T, Britton A, Shipley M, Vandenberk B, et al. QRS micro-fragmentation as a mortality predictor. Eur Heart J. 2022;43(40):4177-91.

94. Ribeiro AH. Predicting age from the electrocardiogram and its usage as a mortality predictor 2021 [

95. Howard JP, Francis DP. Machine learning with convolutional neural networks for clinical cardiologists. Heart. 2021.

96. He K, Zhang X, Ren S, Sun J, editors. Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR); 2016 27-30 June 2016.

97. Xiao Q, Lee K, Mokhtar SA, Ismail I, Pauzi ALbM, Zhang Q, et al. Deep learning-based ECG arrhythmia classification: A systematic review. Applied Sciences. 2023;13(8):4964.

98. Kim J-K, Jung S, Park J, Han SW. Arrhythmia detection model using modified DenseNet for comprehensible Grad-CAM visualization. Biomedical Signal Processing and Control. 2022;73:103408.

99. He K, Zhang X, Ren S, Sun J, editors. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition; 2016.

100. Vrudhula A, Hughes JW, Yuan N, Ouyang D. The impact of task set-up in algorithm design: Regression versus classification. NEJM AI. 2024.

101. van der Net JB, Janssens ACJW, Eijkemans MJC, Kastelein JJP, Sijbrands EJG, Steyerberg EW. Cox proportional hazards models have more statistical power than logistic regression models in cross-sectional genetic association studies. European Journal of Human Genetics. 2008;16(9):1111-6.

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

103. Chollet F. Keras 2015 [Available from: https://keras.io.

104. Martín A, Ashish A, Paul B, Eugene B, Zhifeng C, Craig C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. 2015.

105. Goodfellow I, Bengio Y, Courville A. Deep learning: MIT press; 2016.

106. Chen M, Chen W, Chen W, Cai L, Chai G. Skin cancer classification with deep convolutional neural networks. Journal of Medical Imaging and Health Informatics. 2020;10(7):1707-13.

107. Stensrud MJ, Hernán MA. Why Test for Proportional Hazards? Jama. 2020;323(14):1401-2.

108. Stabenau HF, Waks JW. BRAVEHEART: Open-source software for automated electrocardiographic and vectorcardiographic analysis. Computer Methods and Programs in Biomedicine. 2023;242:107798.

109. Popescu DM, Shade JK, Lai C, Aronis KN, Ouyang D, Moorthy MV, et al. Arrhythmic sudden death survival prediction using deep learning analysis of scarring in the heart. Nature Cardiovascular Research. 2022;1(4):334-43.
110. Meyer HV, Dawes TJW, Serrani M, Bai W, Tokarczuk P, Cai J, et al. Genetic and functional insights into the fractal structure of the heart. Nature. 2020;584(7822):589-94.
111. Jonsson BA, Bjornsdottir G, Thorgeirsson TE, Ellingsen LM, Walters GB, Gudbjartsson DF, et al. Brain age prediction using deep learning uncovers associated sequence variants. Nature Communications. 2019;10(1):5409.
112. Wu P, Gifford A, Meng X, Li X, Campbell H, Varley T, et al. Mapping ICD-10 and ICD-10-CM Codes to Phecodes: Workflow Development and Initial Evaluation. JMIR Med Inform. 2019;7(4):e14325.
113. Yang J, Lee SH, Goddard ME, Visscher PM. GCTA: A Tool for Genome-wide Complex Trait Analysis. The American Journal of Human Genetics. 2011;88(1):76-82.
114. Zvuloni E, Read J, Ribeiro AH, Ribeiro ALP, Behar JA. On Merging Feature Engineering and Deep Learning for Diagnosis, Risk-Prediction and Age Estimation Based on the 12-Lead ECG. 2022.
115. Barish M, Bolourani S, Lau LF, Shah S, Zanos TP. External validation demonstrates limited clinical utility of the interpretable mortality prediction model for patients with COVID-19. Nature Machine Intelligence. 2021;3(1):25-7.
116. Syakur M, Khotimah B, Rochman E, Satoto BD, editors. Integration k-means clustering method and elbow method for identification of the best customer profile cluster. IOP conference series: materials science and engineering; 2018: IOP Publishing.

117. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research. 2011;12(85):2825-30.

118. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision. 2019;128(2):336-59.

119. Chambers JC, Zhao J, Terracciano CM, Bezzina CR, Zhang W, Kaba R, et al. Genetic variation in SCN10A influences cardiac conduction. Nat Genet. 2010;42(2):149-52.

120. Holm H, Gudbjartsson DF, Arnar DO, Thorleifsson G, Thorgeirsson G, Stefansdottir H, et al. Several common variants modulate heart rate, PR interval and QRS duration. Nat Genet. 2010;42(2):117-22.

121. McNair WP, Ku L, Taylor MR, Fain PR, Dao D, Wolfel E, et al. SCN5A mutation associated with dilated cardiomyopathy, conduction disorder, and arrhythmia. Circulation. 2004;110(15):2163-7.

122. Yang KC, Rutledge CA, Mao M, Bakhshi FR, Xie A, Liu H, et al. Caveolin-1 modulates cardiac gap junction homeostasis and arrhythmogenecity by regulating cSrc tyrosine kinase. Circ Arrhythm Electrophysiol. 2014;7(4):701-10.

123. Méndez-Giráldez R, Gogarten SM, Below JE, Yao J, Seyerle AA, Highland HM, et al. GWAS of the electrocardiographic QT interval in Hispanics/Latinos generalizes previously identified loci and identifies population-specific signals. Scientific reports [Internet]. 2017 2017/12//; 7(1):[17075 p.]. Available from: http://europepmc.org/abstract/MED/29213071

https://doi.org/10.1038/s41598-017-17136-0

https://europepmc.org/articles/PMC5719082

https://europepmc.org/articles/PMC5719082?pdf=render.

124. Ntalla I, Weng L-C, Cartwright JH, Hall AW, Sveinbjornsson G, Tucker NR, et al. Multi-ancestry GWAS of the electrocardiographic PR interval identifies 202 loci underlying cardiac conduction. Nature communications [Internet]. 2020 2020/05//; 11(1):[2542 p.]. Available from: http://europepmc.org/abstract/MED/32439900 https://doi.org/10.1038/s41467-020-15706-x https://europepmc.org/articles/PMC7242331 https://europepmc.org/articles/PMC7242331?pdf=render.

125. Young WJ, Lahrouchi N, Isaacs A, Duong T, Foco L, Ahmed F, et al. Genetic analyses of the electrocardiographic QT interval and its components identify additional loci and pathways. Nature communications [Internet]. 2022 2022/09//; 13(1):[5144 p.]. Available from: http://europepmc.org/abstract/MED/36050321 https://doi.org/10.1038/s41467-022-32821-z https://europepmc.org/articles/PMC9436946 https://europepmc.org/articles/PMC9436946?pdf=render.

126. Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. Circulation: Arrhythmia and Electrophysiology. 2019;12(9).

127. Cheng S, Keyes MJ, Larson MG, McCabe EL, Newton-Cheh C, Levy D, et al. Long-term Outcomes in Individuals With Prolonged PR Interval or First-Degree Atrioventricular Block. JAMA. 2009;301(24):2571-7.

128. Haverkamp W, Eckardt L, Mönnig G, Schulze-Bahr E, Wedekind H, Kirchhof P, et al. Clinical aspects of ventricular arrhythmias associated with QT prolongation. European heart journal supplements. 2001;3(suppl_K):K81-K8.

129. Stewart S, Playford D, Scalia GM, Currie P, Celermajer DS, Prior D, et al. Ejection fraction and mortality: a nationwide register-based cohort study of 499 153 women and men. Eur J Heart Fail. 2021;23(3):406-16.

130. Surkova E, Muraru D, Genovese D, Aruta P, Palermo C, Badano LP. Relative Prognostic Importance of Left and Right Ventricular Ejection Fraction in Patients With Cardiac Diseases. J Am Soc Echocardiogr. 2019;32(11):1407-15.e3.

131. Kalam K, Otahal P, Marwick TH. Prognostic implications of global LV dysfunction: a systematic review and meta-analysis of global longitudinal strain and ejection fraction. Heart. 2014;100(21):1673.

132. Kannel WB, Kannel C, Paffenbarger Jr RS, Cupples LA. Heart rate and cardiovascular mortality: the Framingham Study. American heart journal. 1987;113(6):1489-94.

133. Port S, Demer L, Jennrich R, Walter D, Garfinkel A. Systolic blood pressure and mortality. The Lancet. 2000;355(9199):175-80.

134. Shi L, Yu X, Pang Q, Chen X, Wang C. The associations between bone mineral density and long-term risks of cardiovascular disease, cancer, and all-cause mortality. Front Endocrinol (Lausanne). 2022;13:938399.

135. Kalahasti V, Nambi V, Martin DO, Lam CT, Yamada D, Wilkoff BL, et al. QRS duration and prediction of mortality in patients undergoing risk stratification for ventricular arrhythmias. Am J Cardiol. 2003;92(7):798-803.

136. Zhang Y, Post WS, Blasco-Colmenares E, Dalal D, Tomaselli GF, Guallar E. Electrocardiographic QT interval and mortality: a meta-analysis. Epidemiology. 2011;22(5):660-70.

137. Aro AL, Anttonen O, Kerola T, Junttila MJ, Tikkanen JT, Rissanen HA, et al. Prognostic significance of prolonged PR interval in the general population. European Heart Journal. 2013;35(2):123-9.

138. Hu D, Barajas-Martínez H, Pfeiffer R, Dezi F, Pfeiffer J, Buch T, et al. Mutations in SCN10A Are Responsible for a Large Fraction of Cases of Brugada Syndrome. Journal of the American College of Cardiology. 2014;64(1):66-79.

139. Rook MB, Bezzina Alshinawi C, Groenewegen WA, van Gelder IC, van Ginneken ACG, Jongsma HJ, et al. Human SCN5A gene mutations alter cardiac sodium channel kinetics and are associated with the Brugada syndrome*. Cardiovascular Research. 1999;44(3):507-17.

140. Hesse M, Kondo CS, Clark RB, Su L, Allen FL, Geary-Joo CT, et al. Dilated cardiomyopathy is associated with reduced expression of the cardiac sodium channel Scn5a. Cardiovasc Res. 2007;75(3):498-509.

141. Marcsa B, Dénes R, Vörös K, Rácz G, Sasvári-Székely M, Rónai Z, et al. A Common Polymorphism of the Human Cardiac Sodium Channel Alpha Subunit (SCN5A) Gene Is Associated with Sudden Cardiac Death in Chronic Ischemic Heart Disease. PLoS One. 2015;10(7):e0132137.

142. Engelman JA, Zhang XL, Galbiati F, Lisanti MP. Chromosomal localization, genomic organization, and developmental expression of the murine caveolin gene family (Cav-1, -2, and -3). Cav-1 and Cav-2 genes map to a known tumor suppressor locus (6-A2/7q31). FEBS Lett. 1998;429(3):330-6.

143. Ellinor PT, Lunetta KL, Albert CM, Glazer NL, Ritchie MD, Smith AV, et al. Meta-analysis identifies six new susceptibility loci for atrial fibrillation. Nature Genetics. 2012;44(6):670-5.

144. Su ZJ, Hahn CN, Goodall GJ, Reck NM, Leske AF, Davy A, et al. A vascular cell-restricted RhoGAP, p73RhoGAP, is a key regulator of angiogenesis. Proc Natl Acad Sci U S A. 2004;101(33):12212-7.

145. Kim CH, Al-Kindi S, Tarabichi Y, Gohel S, Vyas R, Srinivasan S. Machine Learning to Predict 10-year Cardiovascular Mortality from the Electrocardiogram: Analysis of the Third National Health and Nutrition Examination Survey (NHANES III). medRxiv. 2021:2021.09.09.21263327.

146. van de Leur RR, Bos MN, Taha K, Sammani A, Yeung MW, van Duijvenboden S, et al. Improving explainability of deep neural network-based electrocardiogram interpretation using variational auto-encoders European Heart Journal - Digital Health. 2022;3(3):390-404.

147. Wouters PC, van de Leur RR, Vessies MB, van Stipdonk AMW, Ghossein MA, Hassink RJ, et al. Electrocardiogram-based deep learning improves outcome prediction following cardiac resynchronization therapy. European Heart Journal. 2022.

148. Sau A, Ribeiro AH, McGurk KA, Pastika L, Bajaj N, Ardissino M, et al. Neural network-derived electrocardiographic features have prognostic significance and important phenotypic and genotypic associations. medRxiv. 2023:2023.06.15.23291428.

149. Raghunath S, Ulloa Cerna AE, Jing L, vanMaanen DP, Stough J, Hartzel DN, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. Nat Med. 2020;26(6):886-91.

150. Hughes JW. A deep learning-based electrocardiogram risk score for long term cardiovascular death and disease 2023 [

151. Fine JP. Comparing nonnested Cox models. Biometrika. 2002;89(3):635-48.

152. Suchard MA, Schuemie MJ, Krumholz HM, You SC, Chen R, Pratt N, et al. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. Lancet. 2019;394(10211):1816-26.

153. 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(2):105-21.

154. Verweij N, Benjamins JW, Morley MP, van de Vegte YJ, Teumer A, Trenkwalder T, et al. The Genetic Makeup of the Electrocardiogram. Cell Syst. 2020;11(3):229-38.e5.

155. van der Harst P, van Setten J, Verweij N, Vogler G, Franke L, Maurano MT, et al. 52 Genetic Loci Influencing Myocardial Mass. J Am Coll Cardiol. 2016;68(13):1435-48.

156. Sakaue S, Kanai M, Tanigawa Y, Karjalainen J, Kurki M, Koshiba S, et al. A cross-population atlas of genetic associations for 220 human phenotypes. Nat Genet. 2021;53(10):1415-24.

157. 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(1):10688.

158. Jespersen T, Grunnet M, Olesen S-P. The KCNQ1 potassium channel: from gene to physiological function. Physiology. 2005;20(6):408-16.

159. Graff M, Scott RA, Justice AE, Young KL, Feitosa MF, Barata L, et al. Genome-wide physical activity interactions in adiposity - A meta-analysis of 200,452 adults. PLoS Genet. 2017;13(4):e1006528.

160. Kanai M, Akiyama M, Takahashi A, Matoba N, Momozawa Y, Ikeda M, et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. Nat Genet. 2018;50(3):390-400.

161. Zhao W, Rasheed A, Tikkanen E, Lee JJ, Butterworth AS, Howson JMM, et al. Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease. Nat Genet. 2017;49(10):1450-7.

162. Sillesen H, Muntendam P, Adourian A, Entrekin R, Garcia M, Falk E, et al. Carotid plaque burden as a measure of subclinical atherosclerosis: comparison with other tests for subclinical arterial disease in the High Risk Plaque BioImage study. JACC Cardiovasc Imaging. 2012;5(7):681-9.

163. Otto CM, Nishimura RA, Bonow RO, Carabello BA, Erwin JP, 3rd, Gentile F, et al. 2020 ACC/AHA Guideline for the Management of Patients With Valvular Heart Disease: A Report of the American College of Cardiology/American Heart Association Joint Committee on Clinical Practice Guidelines. Circulation. 2021;143(5):e72-e227.

164. Vahanian A, Beyersdorf F, Praz F, Milojevic M, Baldus S, Bauersachs J, et al. 2021 ESC/EACTS Guidelines for the management of valvular heart disease: Developed by the Task Force for the management of valvular heart disease of the European Society of Cardiology (ESC) and the European Association for Cardio-Thoracic Surgery (EACTS). European Heart Journal. 2022;43(7):561-632.

165. Généreux P. EARLY TAVR: Evaluation of TAVR Compared to Surveillance for Patients With Asymptomatic Severe Aortic Stenosis (EARLY TAVR) [Available from: https://clinicaltrials.gov/study/NCT03042104.

166. McCann G. The Early Valve Replacement in Severe ASYmptomatic Aortic Stenosis Study (EASY-AS).

167. Chang KY, Duval S, Badesch DB, Bull TM, Chakinala MM, De Marco T, et al. Mortality in Pulmonary Arterial Hypertension in the Modern Era: Early Insights From the Pulmonary Hypertension Association Registry. J Am Heart Assoc. 2022;11(9):e024969.

168. Humbert M, Kovacs G, Hoeper MM, Badagliacca R, Berger RMF, Brida M, et al. 2022 ESC/ERS Guidelines for the diagnosis and treatment of pulmonary hypertension. Eur Heart J. 2022;43(38):3618-731.

169. Attia ZI, Harmon DM, Dugan J, Manka L, Lopez-Jimenez F, Lerman A, et al. Prospective evaluation of smartwatch-enabled detection of left ventricular dysfunction. Nat Med. 2022;28(12):2497-503.

170. Istolahti T, Lyytikäinen LP, Huhtala H, Nieminen T, Kähönen M, Lehtimäki T, 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(1):e12799.

171. Schröder LC, Holkeri A, Eranti A, Haukilahti MAE, Kerola T, Kenttä TV, 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(6):952-9.

172. Seyyedi SR, Sharif-Kashani B, Sadr M, Chitsazan M, Malekmohammad M, Abedini A, et al. The Relationship between Electrocardiographic Changes and Prognostic Factors in Severely Symptomatic Pulmonary Hypertension. Tanaffos. 2019;18(1):34-40.

173. Rauseo E, Salih A, Raisi-Estabragh Z, Aung N, Khanderia N, Slabaugh GG, et al. Ischemic Heart Disease and Vascular Risk Factors Are Associated With Accelerated Brain Aging. JACC Cardiovasc Imaging. 2023;16(7):905-15.

174. Ruth EP, Xiaonan L, Jennifer AC, David AC, Benjamin JC, Lei C. Independent external validation of the QRISK3 cardiovascular disease risk prediction model using UK Biobank. Heart. 2023;109(22):1690.

175. Khan SS, Coresh J, Pencina MJ, Ndumele CE, Rangaswami J, Chow SL, et al. Novel prediction equations for absolute risk assessment of total cardiovascular disease incorporating cardiovascular-kidney-metabolic health: a scientific statement from the American Heart Association. Circulation. 2023;148(24):1982-2004.

176. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357.

177. Arbelo E, Protonotarios A, Gimeno JR, Arbustini E, Barriales-Villa R, Basso C, et al. 2023 ESC Guidelines for the management of cardiomyopathies: Developed by the task force on the management of cardiomyopathies of the European Society of Cardiology (ESC). European heart journal. 2023;44(37):3503-626.

178. Zeppenfeld K, Tfelt-Hansen J, de Riva M, Winkel BG, Behr ER, Blom NA, et al. 2022 ESC Guidelines for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death: Developed by the task force for the management of patients with ventricular arrhythmias and the prevention of sudden cardiac death of the European Society of Cardiology (ESC) Endorsed by the Association for European Paediatric and Congenital Cardiology (AEPC). European Heart Journal. 2022;43(40):3997-4126.

179. Glikson M, Nielsen JC, Kronborg MB, Michowitz Y, Auricchio A, Barbash IM, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. Eur Heart J. 2021;42(35):3427-520.

180. Agarwal SK, Chambless LE, Ballantyne CM, Astor B, Bertoni AG, Chang PP, et al. Prediction of incident heart failure in general practice: the Atherosclerosis Risk in Communities (ARIC) Study. Circ Heart Fail. 2012;5(4):422-9.

181. Thompson DJ, Wells D, Selzam S, Peneva I, Moore R, Sharp K, et al. UK Biobank release and systematic evaluation of optimised polygenic risk scores for 53 diseases and quantitative traits. medRxiv. 2022:2022.06.16.22276246.

182. Visseren FLJ, Mach F, Smulders YM, Carballo D, Koskinas KC, Bäck M, et al. 2021 ESC Guidelines on cardiovascular disease prevention in clinical practice. Eur Heart J. 2021;42(34):3227-337.

183. Goff DC, Jr., Lloyd-Jones DM, Bennett G, Coady S, D'Agostino RB, Sr., 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(25 Pt B):2935-59.

184. Hippisley-Cox J, Coupland C, Brindle P. Development and validation of QRISK3 risk prediction algorithms to estimate future risk of cardiovascular disease: prospective cohort study. BMJ. 2017;357:j2099.

185. Abdelaziz Hesham K, Saad M, Pothineni Naga Venkata K, Megaly M, Potluri R, Saleh M, et al. Aspirin for Primary Prevention of Cardiovascular Events. Journal of the American College of Cardiology. 2019;73(23):2915-29.

186. Olgin JE, Pletcher MJ, Vittinghoff E, Wranicz J, Malik R, Morin DP, et al. Wearable Cardioverter–Defibrillator after Myocardial Infarction. New England Journal of Medicine. 2018;379(13):1205-15.

187. Bachtiger P, Kelshiker MA, Petri CF, Gandhi M, Shah M, Kamalati T, 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).

188. Yang H, Negishi K, Otahal P, Marwick TH. Clinical prediction of incident heart failure risk: a systematic review and meta-analysis. Open Heart. 2015;2(1):e000222.

189. Glikson M, Nielsen JC, Kronborg MB, Michowitz Y, Auricchio A, Barbash IM, et al. 2021 ESC Guidelines on cardiac pacing and cardiac resynchronization therapy. Eur Heart J. 2021;42(35):3427-520.

190. Svendsen JH, Diederichsen SZ, Højberg S, Krieger DW, Graff C, Kronborg C, et al. Implantable loop recorder detection of atrial fibrillation to prevent stroke (The LOOP Study): a randomised controlled trial. The Lancet. 2021;398(10310):1507-16.

191. Svennberg E, Friberg L, Frykman V, Al-Khalili F, Engdahl J, Rosenqvist M. Clinical outcomes in systematic screening for atrial fibrillation (STROKESTOP): a multicentre, parallel group, unmasked, randomised controlled trial. Lancet. 2021;398(10310):1498-506.

192. Peters SAE, Muntner P, Woodward M. Sex Differences in the Prevalence of, and Trends in, Cardiovascular Risk Factors, Treatment, and Control in the United States, 2001 to 2016. Circulation. 2019;139(8):1025-35.

193. Pope JH, Aufderheide TP, Ruthazer R, Woolard RH, Feldman JA, Beshansky JR, et al. Missed Diagnoses of Acute Cardiac Ischemia in the Emergency Department. New England Journal of Medicine. 2000;342(16):1163-70.

194. Agarwala A, Michos ED, Samad Z, Ballantyne CM, Virani SS. The Use of Sex-Specific Factors in the Assessment of Women's Cardiovascular Risk. Circulation. 2020;141(7):592-9.

195. Leifheit-Limson EC, D'Onofrio G, Daneshvar M, Geda M, Bueno H, Spertus JA, et al. Sex differences in cardiac risk factors, perceived risk, and health care

provider discussion of risk and risk modification among young patients with acute myocardial infarction: the VIRGO study. Journal of the American College of Cardiology. 2015;66(18):1949-57.

196. Vosberg DE, Syme C, Parker N, Richer L, Pausova Z, Paus T. Sex continuum in the brain and body during adolescence and psychological traits. Nature Human Behaviour. 2021;5(2):265-72.

197. Vosberg DE, Pausova Z, Paus T. The genetics of a "femaleness/maleness" score in cardiometabolic traits in the UK biobank. Sci Rep. 2023;13(1):9109.

198. Moss AJ. Gender differences in ECG parameters and their clinical implications. Ann Noninvasive Electrocardiol. 2010;15(1):1-2.

199. Siegersma KR, van de Leur RR, Onland-Moret NC, Leon DA, Diez-Benavente E, Rozendaal L, et al. Deep neural networks reveal novel sex-specific electrocardiographic features relevant for mortality risk. Eur Heart J Digit Health. 2022;3(2):245-54.

200. Prins BP, Mead TJ, Brody JA, Sveinbjornsson G, Ntalla I, Bihlmeyer NA, et al. Exome-chip meta-analysis identifies novel loci associated with cardiac conduction, including ADAMTS6. Genome Biol. 2018;19(1):87.

201. Ruth KS, Day FR, Tyrrell J, Thompson DJ, Wood AR, Mahajan A, et al. Using human genetics to understand the disease impacts of testosterone in men and women. Nat Med. 2020;26(2):252-8.

202. van Duijvenboden S, Ramírez J, Young WJ, Orini M, Mifsud B, Tinker A, et al. Genomic and pleiotropic analyses of resting QT interval identifies novel loci and overlap with atrial electrical disorders. Hum Mol Genet. 2021;30(24):2513-23.

203. Tadros R, Francis C, Xu X, Vermeer AMC, Harper AR, Huurman R, et al. Shared genetic pathways contribute to risk of hypertrophic and dilated cardiomyopathies with opposite directions of effect. Nat Genet. 2021;53(2):128-34.
204. Graham SE, Clarke SL, Wu KH, Kanoni S, Zajac GJM, Ramdas S, et al. The power of genetic diversity in genome-wide association studies of lipids. Nature. 2021;600(7890):675-9.
205. Kim SK. Identification of 613 new loci associated with heel bone mineral density and a polygenic risk score for bone mineral density, osteoporosis and fracture. PLoS One. 2018;13(7):e0200785.
206. Nielsen JB, Rom O, Surakka I, Graham SE, Zhou W, Roychowdhury T, et al. Loss-of-function genomic variants highlight potential therapeutic targets for cardiovascular disease. Nat Commun. 2020;11(1):6417.
207. Devereux RB, Lutas EM, Casale PN, Kligfield P, Eisenberg RR, Hammond IW, et al. Standardization of M-mode echocardiographic left ventricular anatomic measurements. Journal of the American College of Cardiology. 1984;4(6):1222-30.
208. Zhu K, Briffa K, Smith A, Mountain J, Briggs AM, Lye S, et al. Gender differences in the relationships between lean body mass, fat mass and peak bone mass in young adults. Osteoporosis International. 2014;25(5):1563-70.
209. Bannister P, Losowsky MS. Sex hormones and chronic liver disease. Journal of Hepatology. 1988;6(2):258-62.
210. Howard A, Borenstein J. The Ugly Truth About Ourselves and Our Robot Creations: The Problem of Bias and Social Inequity. Science and Engineering Ethics. 2018;24(5):1521-36.
211. Xhyheri B, Bugiardini R. Diagnosis and treatment of heart disease: are women different from men? Prog Cardiovasc Dis. 2010;53(3):227-36.

212. Mosca L, Linfante AH, Benjamin EJ, Berra K, Hayes SN, Walsh BW, et al. National Study of Physician Awareness and Adherence to Cardiovascular Disease Prevention Guidelines. Circulation. 2005;111(4):499-510.

213. Reichlin T, Abächerli R, Twerenbold R, Kühne M, Schaer B, Müller C, et al. Advanced ECG in 2016: is there more than just a tracing? Swiss Med Wkly. 2016;146:w14303.

214. Shih W, Chai S, editors. Data-driven vs. hypothesis-driven research: making sense of big data. Academy of management proceedings; 2016: Academy of Management Briarcliff Manor, NY 10510.

215. Yao X, Rushlow DR, Inselman JW, McCoy RG, Thacher TD, Behnken EM, et al. Artificial intelligence–enabled electrocardiograms for identification of patients with low ejection fraction: a pragmatic, randomized clinical trial. Nature Medicine. 2021;27(5):815-9.

216. Noseworthy PA, Attia ZI, Behnken EM, Giblon RE, Bews KA, Liu S, et al. Artificial intelligence-guided screening for atrial fibrillation using electrocardiogram during sinus rhythm: a prospective non-randomised interventional trial. The Lancet. 2022;400(10359):1206-12.

217. Ahn JC, Attia ZI, Rattan P, Mullan AF, Buryska S, Allen AM, et al. Development of the AI-cirrhosis-ECG score: an electrocardiogram-based deep learning model in cirrhosis. Official journal of the American College of Gastroenterology| ACG. 2022;117(3):424-32.

218. Sangha V, Mortazavi BJ, Haimovich AD, Ribeiro AH, Brandt CA, Jacoby DL, et al. Automated multilabel diagnosis on electrocardiographic images and signals. Nat Commun. 2022;13(1):1583.

219. Pham HN, Holmstrom L, Chugh H, Uy-Evanado A, Nakamura K, Zhang Z, et al. Dynamic electrocardiogram changes are a novel risk marker for sudden cardiac death. European Heart Journal. 2024;45(10):809-19.