Machine Learning to Predict Cardiovascular Mortality from Electrocardiogram Data

By

Chang H. Kim, M.D.

A Dissertation Submitted to

Rutgers – School of Health Professions

In partial fulfillment of the Requirements for the Degree of

Doctor of Philosophy in Biomedical Informatics

Department of Health Informatics

Rutgers, The State University of New Jersey

School of Health Professions

May 2020

RUTGERS
School of Health Professions

Final Dissertation Defense Approval Form

Machine Learning to Predict Cardiovascular Mortality

from Electrocardiogram Data

By

Chang H. Kim, M.D.

Dissertation Committee:

Shankar Srinivasan, Ph.D.

Suril Gohel, Ph.D.

Riddhi Vyas, Ph.D.

Approved by the Dissertation Committee:

Shankar Srinivasan, Ph.D.	Date signed
Suril Gohel, Ph.D.	Date signed
Riddhi Vyas, Ph.D.	Date signed

ABSTRACT

Atherosclerotic cardiovascular disease (ASCVD) and subsequent adverse cardiovascular events remain highly prevalent in the U.S., making primary prevention an important goal. While the 2013 ACC/AHA Pooled Cohort Equations (PCE) remains the gold standard for cardiovascular event prediction, not represented in the model is cardiac electrophysiology, a major cause of sudden cardiac death. The electrocardiogram (ECG), a routinely available test that reflects one's electrophysiologic health, may thus be useful for cardiovascular risk stratification in addition, and in comparison, to the PCE. Given the automated and highly correlated nature of its measurements, ECG data are well suited for analysis via machine learning. In this work, the value of aggregated ECG measurements for prediction of cardiovascular mortality is assessed in a nationwide cohort (NHANES III), via a comparative analysis of traditional survival analysis and machine learning methods. Overall, machine learning models could predict 10-year cardiovascular mortality with superior accuracy and event detection capacity compared to the PCE. Interestingly, only demographic and ECG data were necessary for such improved performance. Variable comparison between different prediction models provided insight into the relative importance of specific ECG components and the detection of silent myocardial infarctions as a possible underlying mechanism.

ACKNOWLEDGEMENTS

This dissertation would not have been possible without the invaluable input from many who contributed and supported me throughout the process. First, I would like to acknowledge my thesis advisors in the dissertation committee – chair Dr. Shankar Srinivasan for his patient guidance and leadership; Dr. Suril Gohel for his input to training machine learning models; and Dr. Riddhi Vyas, for her insight into the dataset. Second, I would like to thank my advisors from Cleveland, where I currently practice medicine – Dr. Sadeer Al-Kindi and Dr. Yasir Tarabichi from Case Western Reserve University School of Medicine for their valuable clinical insight and advice regarding utilizing healthcare data; and Dr. Jarrod Dalton, Dr. Adam Perzynski, Nikolas Krieger, and many others from the NEOCARE research group (Cleveland Clinic / Lerner College of Medicine, MetroHealth Medical Center / Case Western Reserve University School of Medicine) for their feedback and assistance in coding and analysis.

DEDICATION

This dissertation is dedicated to: my wife Sally Chanmi Bak, for her unwavering love and support throughout the PhD journey; our children Ara and Ian, who tolerated many undeserved hours of inattention; and my father Dr. Dong Joon Kim, for his inspiration as a role model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter I

INTRODUCTION

1. Background

Atherosclerotic cardiovascular disease (ASCVD) is a major cause of morbidity and mortality in the United States, with 580,000 incident heart attacks and 610,000 incident strokes occurring each year[1]. The prevalence and cost burden of ASCVD continue to rise, with over 90 million affected adults and annual healthcare costs of over $400 billion[1]. While cardiovascular mortality has declined in recent decades, the rate of decline appears to be decelerating, thought to be related to the increasing prevalence of cardiovascular risk factors such as hypertension, diabetes, and obesity[2]. Therefore, primary prevention of ASCVD and subsequent major adverse cardiovascular events (MACE), defined as myocardial infarction, coronary heart disease-related death, and fatal and non-fatal stroke[3], remains an important public health goal.

The current gold standard for cardiovascular event risk estimation in the U.S. is the Pooled Cohort Equations (PCE)[3], published by the American College of Cardiology (ACC) and the American Heart Association (AHA) in 2013. Briefly, the PCE is a stratified, multivariable Cox proportional hazards model that estimates 10-year incident ASCVD event risk based on nine clinical variables. While the PCE is in routine clinical use in the U.S., its suboptimal calibration in specific patient populations have been noted, leading to updated clinical practice guidelines in 2019[4]. While cardiac death was originally defined in the PCE as only those related to coronary heart disease, other types of cardiac death, such as those related to fatal

arrhythmias or heart failure, have been included as outcomes of interest in subsequent studies. From this perspective, the PCE may be missing important risk factors related to cardiac electrophysiology. Given such background, it is reasonable to consider whether information captured in the electrocardiogram (ECG), a routinely available and non-invasive test which reflects both coronary health and cardiac electrophysiology, may be useful for prediction of MACE.

2. Electrocardiogram review

The standard 12-lead ECG captures the electrical activity of the heart in twelve separate tracings based on distinct combinations of positive and negative electrodes, known as leads. The twelve leads are named I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, and V6, and are grouped based on their anatomic locations, i.e. lateral (I, aVL, V5, V6), inferior (II, III, aVF), septal (V1, V2), and anterior (V3, V4) leads, which correspond to specific territories of the three coronary arteries. ECG tracings are recorded on standardized grid paper which allows measurement of various amplitudes and intervals, measured in millivolts and microseconds, respectively. The tracing of a normal heartbeat displays the P wave, PR segment, QRS complex, ST segment, and T wave, where diseases of the heart may manifest as pathological waves or aberrations in specific segments or intervals. While modern ECG equipment automatically measures hundreds of wave amplitudes and intervals at the time of ECG capture and produces preliminary interpretations based on proprietary algorithms, the final interpretation is made by a cardiologist, who typically relies on visual pattern recognition rather than quantitative computation. In clinical practice, only a small fraction of information available in the ECG are utilized, e.g. identifying ischemic waveform changes for assessing the risk of an

active heart attack or using the corrected QT interval for medication selection. In medical research, while isolated ECG components have been studied for their capacity in cardiovascular risk stratification, the predictive value of aggregated ECG measurements remain mostly unknown.

3. Healthcare research environment

The widespread adoption of electronic health records (EHR) in the U.S. since the 2009 Health Information Technology for Economic and Clinical Health Act has resulted in large-scale, longitudinal data accumulation in healthcare systems. Availability of such "big data", in addition to recent advances in machine learning (especially deep learning) and availability of affordable computing power, has provided a fertile environment for a new generation of clinical prediction models based on machine learning methods. Compared to traditional statistical analysis, machine learning models provide greater model capacity and flexibility to handle large amounts of correlated data. In this sense, machine learning methods are a natural fit to utilizing aggregated ECG data for cardiovascular event prediction.

4. Problem statement

While the 2013 PCE remains the gold standard for ASCVD event prediction, not represented in the model are risk factors related to cardiac electrophysiology, a major cause of sudden cardiac death. The ECG, a routinely available test that can reflect both coronary and electrophysiologic health, represents an ideal candidate for assessment of risk stratification value in addition, and in comparison, to the PCE. Given the automated, high-dimensional, and highly correlated nature of its

measurements, ECG data are well suited for analysis via machine learning methods that can handle such data complexity.

5. Research objectives

In this work, the primary objective is to assess the value of aggregated ECG measurements for prediction of cardiovascular mortality in a nationwide cohort, via both traditional survival analysis and machine learning methods. Secondary objectives include examination of specific ECG components to assess their relative contribution to cardiovascular risk stratification and to gain insight into the underlying mechanism for prediction.

6. Research hypothesis

Aggregated ECG data are useful for cardiovascular risk prediction, in addition and in comparison to traditional cardiovascular risk factors represented in the PCE.

Chapter II.

REVIEW OF RELATED LITERATURE

1. Standard model for ASCVD event risk estimation

a) The 2013 ACC/AHA Pooled Cohort Equations

The 2013 PCE, developed and endorsed by the ACC and AHA, represents the current gold standard for prediction of ASCVD events in the general U.S. population[4]. Outcome events of interest are adverse events related to ASCVD, or MACE, originally defined as acute coronary syndrome (myocardial infarction), sudden cardiac death (related to coronary heart disease), and fatal- and non-fatal stroke. The PCE are sex- and race- stratified multivariable Cox proportional hazards (Cox PH) models that estimate 10-year ASCVD event risk, derived from five community-based, longitudinal cohorts (Atherosclerosis Risk in Communities, Cardiovascular Health Study, Coronary Artery Risk Development in Young Adults, Framingham Original, Framingham Offspring). These derivation cohorts included 24,626 participants in total, who were 40-79 years of age and enrolled between 1968 and 1990. The PCE utilizes nine routinely available clinical variables with known association with ASCVD risk, namely age, sex, race (white or African American), total cholesterol, high-density lipoprotein (HDL) cholesterol, systolic blood pressure, current treatment for hypertension, diabetes mellitus, and current smoking status, along with selected age-interaction terms. The full model parameters and applied examples are shown in **Figure 1**. The 10-year event risk estimate from the PCE is recommended to be utilized as part of clinician-patient discussion for risk factor modification based on lifestyle changes (e.g. healthy diet, regular exercise, and

smoking cessation) and to guide initiation of pharmacologic therapy, with 10-year risk cutoff of >7.5% for lipid-lowering therapy (i.e. statins)[4].

	White			African American		
	Coefficient	Individual Example Value	Coefficient × Value†	Coefficient	Individual Example Value	Coefficient × Value†
Women (Example: 55 years of age with total cholesterol 213 mg/dL, HDL-C 50 mg/dL, untreated systolic BP 120 mm Hg, nonsmoker, and without diabetes)						
Ln Age (y)	−29.799	4.01	−119.41	17.114	4.01	68.58
Ln Age, Squared	4.884	16.06	78.44	N/A	N/A	N/A
Ln Total Cholesterol (mg/dL)	13.540	5.36	72.59	0.940	5.36	5.04
Ln Age × Ln Total Cholesterol	−3.114	21.48	−66.91	N/A	N/A	N/A
Ln HDL-C (mg/dL)	−13.578	3.91	−53.12	−18.920	3.91	−74.01
Ln Age × Ln HDL-C	3.149	15.68	49.37	4.475	15.68	70.15
Ln Treated Systolic BP (mm Hg)	2.019	—	—	29.291	—	—
Ln Age × Ln Treated Systolic BP	N/A	N/A	N/A	−6.432	—	—
Ln Untreated Systolic BP (mm Hg)	1.957	4.79	9.37	27.820	4.79	133.19
Ln Age × Ln Untreated Systolic BP	N/A	N/A	N/A	−6.087	19.19	−116.79
Current Smoker (1=Yes, 0=No)	7.574	0	0	0.691	0	0
Ln Age × Current Smoker	−1.665	0	0	N/A	N/A	N/A
Diabetes (1=Yes, 0=No)	0.661	0	0	0.874	0	0
Individual Sum			−29.67			86.16
Mean (Coefficient × Value)	N/A	N/A	−29.18	N/A	N/A	86.61
Baseline Survival	N/A	N/A	0.9665	N/A	N/A	0.9533
Estimated 10-y Risk of Hard ASCVD	N/A	N/A	2.1%	N/A	N/A	3.0%
Men (Example: 55 years of age with total cholesterol 213 mg/dL, HDL-C 50 mg/dL, untreated systolic BP 120 mm Hg, nonsmoker, and without diabetes)						
Ln Age (y)	12.344	4.01	49.47	2.469	4.01	9.89
Ln Total Cholesterol (mg/dL)	11.853	5.36	63.55	0.302	5.36	1.62
Ln Age × Ln Total Cholesterol	−2.664	21.48	−57.24	N/A	N/A	N/A
Ln HDL-C (mg/dL)	−7.990	3.91	−31.26	−0.307	3.91	−1.20
Ln Age × Ln HDL-C	1.769	15.68	27.73	N/A	N/A	N/A
Ln Treated Systolic BP (mm Hg)	1.797	—	—	1.916	—	—
Ln Untreated Systolic BP (mm Hg)	1.764	4.79	8.45	1.809	4.79	8.66
Current Smoker (1=Yes, 0=No)	7.837	0	0	0.549	0	0
Ln Age × Current Smoker	−1.795	0	0	N/A	N/A	N/A
Diabetes (1=Yes, 0=No)	0.658	0	0	0.645	0	0
Individual Sum			60.69			18.97
Mean (Coefficient × Value)	N/A	N/A	61.18	N/A	N/A	19.54
Baseline Survival	N/A	N/A	0.9144	N/A	N/A	0.8954
Estimated 10-y Risk of Hard ASCVD	N/A	N/A	5.3%	N/A	N/A	6.1%

*Defined as first occurrence of nonfatal myocardial infarction or CHD death, or fatal or nonfatal stroke.

†Coefficient × Value: For age, lipids, and BP, defined as the natural log of the value multiplied by the parameter estimate. When an age interaction is present with lipids or BP, the natural log of age is multiplied by the natural log of the lipid or BP, and the result is multiplied by the parameter estimate. N/A indicates that that specific covariate was not included in the model for that sex–race group; — indicates that this value was not included in the example (eg, this example used untreated systolic BP, not treated systolic BP).

ASCVD indicates atherosclerotic cardiovascular disease; BP indicates blood pressure; CHD, coronary heart disease; HDL-C, high-density lipoprotein cholesterol; Ln, natural logarithm; and N/A, not included.

Figure 1. The 2013 ACC/AHA Pooled Cohort Equations. Adapted from: [3]

b) Assessing the PCE and opportunities for improvement

While the PCE reported good discrimination (c-index 0.718-0.818) and calibration (chi square 4.86-7.25)[3] at the time of its publication, follow-up studies

have since revealed suboptimal calibration in various population subgroups. In general, the PCE were shown to overestimate ASCVD event risk in more contemporary cohorts (including Physicians' Health Study, Women's Health Study, Women's Health Initiative-Observational Cohort, Multi-Ethnic Study of Atherosclerosis, Reasons for Geographic and Racial Differences in Stroke)[5], with particularly variable risk estimates among certain demographic groups (e.g. black men)[6]. Meanwhile, other studies have found significant underestimation of risk among specific patient subgroups, such as those with autoimmune diseases[7], human immunodeficiency virus infection[8], or disadvantaged socioeconomic background[9,10]. Such miscalibration may be due to the outdated nature of the PCE derivation cohorts which do not reflect secular changes in cardiovascular risk profile and disease management that have occurred in recent decades[5], as well as additional cardiovascular risk factors that are not captured by the PCE[4].

In addition to suboptimal calibration, statistical design concerns have been raised regarding the PCE. These include possible violation of the underlying proportional hazards assumption, whose validity is necessary for a Cox PH model for proper estimation of its coefficients, and selection of interaction terms based on statistical significance, which can lead to overfitting of the model in the absence of appropriate regularization techniques[6]. Furthermore, there is mounting evidence that in addition to presence or magnitude of a risk factor, duration and variability over time are also important determinants of risk[11]. Among known risk factors for ASCVD, high variability in blood pressure, LDL cholesterol, blood glucose, and body weight have been shown to be associated with increased cardiovascular risk, where in contrast, low variability in heart rate appears to be associated with increased cardiovascular risk[11]. Such temporal variation, while containing important risk-

stratifying information, cannot be captured by models based on the standard Cox PH framework such as the PCE, signaling the need for a more advanced model framework.

Despite these concerns, the PCE remains the best validated ASCVD risk assessment tool in the U.S., compared to older prediction models such as the Framingham risk score[4,12]. There have been efforts to recalibrate and revise the PCE based on more modern cohorts[5,6], and to expand its scope to address change in ASCVD risk factors in response to therapy[13]. The recently updated ACC/AHA clinical practice guidelines from 2019 recommend active consideration of other ASCVD risk-enhancing factors that are not part of the PCE, including family history, high-risk race/ethnicity (e.g. South Asian ancestry), primary hypercholesterolemia, metabolic syndrome, chronic kidney disease, chronic inflammatory conditions, premature menopause, pre-eclampsia, as well as additional biomarkers (C-reactive protein, Lp(A), apoB), ankle-brachial index, and coronary artery calcium score (CAC score)[4]. The PCE is recommended to be used as a baseline risk assessment tool, from which additional risk stratification and management decisions can be made based on consideration of patient-specific factors[4].

However, aside from the recommendation to inquire about family history, there are still no specific recommendations related to electrophysiologic risk factors or sudden cardiac death. Several studies have identified specific ECG components that can provide useful cardiovascular risk stratification beyond a standardized risk calculator[14,15], suggesting the value of the ECG for further improving upon the PCE.

2. Challenges in modern healthcare research

The specific challenges related to leveraging modern healthcare data for clinical prediction models are twofold: issues related to data complexity and model complexity[16,17]. These issues and relevant examples are examined in the following sections.

a) Issues related to data complexity

i) Heterogeneity and scale

Healthcare data consist of large collections of different data types, both clinical and non-clinical, in structured and unstructured forms. They may be continuous/numeric (vital signs, laboratory values, sensor data), categorical (medical diagnosis and procedure codes), free text (clinical notes), image (radiology studies, pathology slides), administrative (encounter and claims data), among others. Utilizing such heterogeneous data in traditional statistical models is challenging as model input must typically be structured in a limited number of relevant variables.

Potential solutions to address such large-scale heterogeneity in clinical data include computational phenotyping and representation learning[16,17]. In computational phenotyping, clustering methods or deep learning methods can be used to uncover underlying patterns and natural groupings in complex diseases. In representation learning, multimodal data are first embedded into a structured vector that can then be used for future predictive modeling. Pertinent examples are discussed below.

In a prospective cohort study of 397 patients with heart failure with preserved ejection fraction, a heterogeneous disease, Shah *et al.*[18] used hierarchical, penalized model-based clustering methods to discover three distinct phenotypes

based on structured clinical, laboratory, ECG, and echocardiography data. Significantly, these three phenotypes could be recognized as distinct clinical entities based on their parameter distribution, and were associated with important clinical outcomes (hospitalizations or death) in a separate validation cohort of 107 patients. This study demonstrated that a data-driven unsupervised learning approach could be utilized to discover clinically meaningful phenotypes in a heterogeneous cardiovascular disease.

In another study of novel phenotype discovery, Seymour *et al.*[19] applied *k*-means clustering to 29 clinical variables derived from EHR data of 20,189 patients to identify four novel phenotypes of sepsis, another heterogeneous clinical syndrome. Results were validated in a separate cohort of 43,086 patients, who displayed similar phenotype distributions and consistent biomarker patterns that were recognizable by clinicians. These novel sepsis phenotypes were effective at predicting clinical outcomes, such as 28-day and 365-day mortality. In addition, these phenotypes could be applied to randomized controlled trial data to assess treatment effects, where simulation studies showed that treatment recommendations would have changed significantly had these phenotypic groups been used to stratify patients at the time of the clinical trial. This study suggested another utility of identification of data-driven phenotypes, that of potential use in reassessing trial outcomes and aiding future study design.

ii) Lack of data labels

A related issue to data heterogeneity is the frequent lack of data labels, which is a significant hindrance to training accurate clinical prediction models. In traditional clinical studies, data labels are typically provided by an expert panel,

who identify and validate outcomes through an adjudication process. However, aggregated healthcare data such as that from an EHR often come unlabeled in relation to the outcome of interest. For example, a disease may be present based on diagnostic criteria but may not be labeled with an appropriate diagnosis code, or continuous sensor output may have been recorded but without formal documentation of its interpretation. While outcome adjudication by an expert panel is considered the gold standard, this is frequently unavailable nor practical for large amounts of aggregated data. A practical solution may involve constructing "silver standard" labels based on specific criteria recorded in EHR data[16], for example diagnosing a disease via co-occurrence of specific diagnosis codes[20]. Other approaches include implicit labeling via transfer learning, where a deep learning model transfers learned knowledge about an outcome label to another model, without the need to explicitly create those labels[21].

In a study of healthcare data heterogeneity and disease labels, Wei *et al.*[20] investigated the relative importance of individual EHR components (disease diagnosis codes, clinician notes, and medications) in identifying common diseases. With the goal of properly identifying the ten most common diseases, it was discovered that each EHR component on its own was unreliable in both consistency and accuracy, while utilizing at least two components significantly improved and stabilized the positive predictive value. Interestingly, as a single component, primary clinician notes (from which text diagnoses were extracted) had better sensitivity than diagnosis codes, highlighting the value of incorporating unstructured data elements when feasible. Once multiple EHR components were combined, good empiric accuracy could be achieved in creating proper disease labels.

iii) Temporality and irregularity

While longitudinal healthcare data capture a patient's health trajectory over time, data sampling occurs at irregular intervals during healthcare encounters, resulting in data sparsity, irregularity, and censoring. Prediction models based on such data must actively address how to capture temporal information without being overly biased. Traditional survival analysis is the usual go-to for building clinical prediction models based on such temporal data, although it remains susceptible to aforementioned biases. In terms of machine learning, gated deep learning architectures such as recurrent neural network variants (e.g. Long short-term memory, gated recurrent unit, attention mechanism)[22-25] and deep learning survival models[26-30] enable capture of such temporal relationships, where data normalization methods and architectural modifications have been proposed to overcome data irregularity[25,31-33].

iv) High dimensionality

The promise of modern healthcare data – that among a great number of variables certain combinations may prove to be useful for clinical event prediction or uncover previously unknown disease mechanisms – comes with the challenge of high dimensionality. Having too many variables to consider is problematic for two main reasons. First, variable selection becomes paramount, especially for traditional statistical prediction models that have limited capacity to handle high correlations or interactions. However, variable selection is a complicated problem, requiring expert knowledge for *a priori* variable selection that could potentially negate the benefits of utilizing large scale data, or having to rely on a statistical approach that may exclude clinically meaningful variables[34]. Furthermore, high dimensionality

necessarily leads to data sparsity (known as curse of dimensionality), leading to instances of model non-convergence and inaccurate parameter estimates. Approaches to deal with high dimensionality include knowledge-driven variable selection or data-driven dimension reduction techniques such as principal components analysis or multidimensional scaling. Notably, deep learning models are naturally adept at overcoming high dimensionality through its representation learning capacity, obviating the need to explicitly specify a dimension reduction approach[35].

One interesting approach for addressing high dimensionality using deep learning is to represent the entire medical history of a patient in an embedding vector, in a process called general purpose patient representation. For a given patient, all the data present in the EHR can be organized as a time series of healthcare encounters, with each encounter storing clinical data elements related to that visit. Specific deep learning methods that have been used include: stacked denoising autoencoders[36], ensemble of neural network models[33], and recurrent neural network variants[23,25]. Once built, these general-purpose patient representations have been shown to be empirically effective at predicting various clinical outcomes such as new onset heart failure, in-hospital mortality, etc. The main advantage of this approach is that these general representations can be utilized for many different prediction tasks, while the main disadvantages arise from difficulties related to model training due to the extreme size of data and computational power that is required, as well as poor interpretability and questions of external validity that arise from its overarching scope.

a) Issues related to model complexity

Prediction models must appropriately reflect the type of input data and nature of the prediction task. In the setting of cardiovascular risk prediction, the task may be formulated as a survival analysis problem, given longitudinal data with intermittent observation times and censoring. If there is a specific time point of interest, such as event risk at 10-years, the prediction task can also be formulated as a binary classification problem, though with information loss and biased estimates when censoring is not considered. A review of traditional survival analysis methods and relevant machine learning adaptations are discussed below.

i) Traditional survival analysis

The goal of survival analysis is to utilize longitudinal data to predict the occurrence and if possible, timing, of the outcome event(s) of interest. In such time-to-event data, survival time remains unknown for some subjects, a phenomenon known as censoring. Data may be left censored (event occurs prior to observation) or right censored (event has not yet occurred at end of follow-up). The term survival refers to a state where the outcome event of interest has not (yet) occurred, and does not necessarily refer to mortality. The main quantities of interest include the survival function $S(t) = P(T \geq t)$, representing the probability of survival to point t; the hazard function $h(t) = -d[\log S(t)]/dt$, representing the probability of outcome event occurring at t given survival up to t; and the cumulative hazard function $H(t) = \int h(u)du = -\log S(t)$, approximating the cumulative event risk up to t [37]. Other related functions include the cumulative density function $F(t) = 1 - S(t)$, and the death density function $f(t) = dF(t)/dt = -dS(t)/dt$ [38]. These functions are related to

each other in a fixed mathematical relationship, allowing different survival models to focus on estimating specific quantities to form the basis of outcome prediction[37].

Among traditional survival analysis methods, the Kaplan-Meier method and Cox PH regression are commonly used. The Kaplan-Meier method along with the log-rank test are nonparametric methods and are useful for comparing survival between groups, though they are unable to account for covariates or make specific predictions for individuals[39]. The Cox PH model, in contrast, is a semi-parametric method that allows for multivariable analysis, and is a common choice for clinical event prediction models such as the PCE.

In the Cox PH model, the hazard function is expressed as a combination of baseline hazard and an exponentiated, linear combination of covariates: $h(t) = h_0(t) * exp(\beta_1 x_1 + \beta_2 x_3 + \ldots + \beta_p x_p)$. The model is semi-parametric as the baseline hazard function remains unspecified while the coefficients $\beta_1 \ldots \beta_p$ are estimated through regression[39]. Advantages of the Cox PH model include the lack of need to make survival distribution assumptions, reasonable interpretability of the estimated coefficients (a constant multiplicative effect on the hazard), robustness compared to fully parametric models, and greater precision compared to non-parametric models[38,39]. In contrast, disadvantages of the standard Cox PH model include its assumption of proportional hazards (hazard of one group must remain a constant multiple of the hazard of the other group over the entire follow up period), fixed covariate effects over time (unable to account for nonlinear relationships or temporal changes occurring over the follow up period), and lack of direct estimation of the survival function (does not allow straightforward prediction of survival times)[39]. Variants of Cox PH models such as the regularized (e.g. lasso, ridge, elastic net) and

time-dependent Cox models enable handling of issues such as variable selection, collinearity, and time-varying covariates[38].

Beyond specific issues related to the Cox PH model, there are important general assumptions that must be considered when dealing with time-to-event data[34]. First, censoring must be noninformative; when censoring carries prognostic information (e.g. when patients withdraw from a clinical trial due to drug toxicity), survival model estimates will become biased. Second, when utilizing a regression framework, the effect of covariates are assumed to be constant during follow-up; otherwise scalar coefficient estimation would be inadequate. Finally, when there is a possibility of multiple outcome events interfering with one another, a competing risks framework must be adopted to avoid biased estimates.

ii) Machine learning for survival analysis

For survival analysis, machine learning methods must be adapted to handle issues related to time-to-event data and censoring, as discussed above. In comparison to traditional survival analysis, advantages of machine learning methods include: greater model flexibility due to less stringent assumptions, higher model capacity with ability to handle high dimensional data with nonlinear relationships and interactions, and improved prediction performance[38]. Many different types of machine learning methods that were originally developed for classification tasks have been adapted for survival analysis, including survival trees and related ensemble methods[40-43], generalized additive models[44,45], support vector machines[46], and deep learning-based methods [29,30]. Selected examples are discussed below.

Yu *et al.*[47] proposed multi-task logistic regression (MTLR), where the survival function is modeled as a dependent sequence of logistic regression submodels. The name MTLR refers to its multi-task learning framework, where survival status at each time point modeled by the submodels are jointly learned over the entire time sequence. In effect a discretized linear version of the survival function, MTLR retains the linear combination framework of the Cox PH model, but is more flexible in the sense that it naturally allows for time-varying covariates and coefficients. Tested in a cohort of more than 2,000 cancer patients, the MTLR model outperformed traditional survival analysis models (Cox PH and Aalen additive hazards models) by as much as 20% in classification accuracy.

In neural multi-task logistic regression (N-MTLR), Fotso el al[26] expanded on the MTLR framework by introducing an overarching neural network architecture over the sequence of dependent logistic regression submodels. The aim was to expand on the MTLR to capture nonlinear dependencies between covariates and survival, which linear models are inherently unable to capture. To achieve this, a non-linear transformation was performed on the input feature vector using a multi-layer perceptron neural network, whose output vectors were used to represent the covariates at each subdivision of the time axis. In simulated and real-life datasets (Worcester Heart Attack Study, Veterans' Administration Lung Cancer), the N-MTLR performed similarly or better than both Cox PH and baseline MTLR models, particularly when nonlinear dependencies were present in simulated data.

Katzman *et al.* proposed DeepSurv[28], a deep learning model based on a multi-layer perceptron architecture with its output layer modeled as a Cox PH model. The primary aim was to explicitly model treatment assignment and its interaction with other covariates in data-driven fashion, in contrast to feature engineering based on

expert knowledge that would be required in a standard Cox PH model. DeepSurv was shown to perform as well or better than other survival methods in both simulated datasets and real clinical studies (including the Worcestor Heart Attack Study), while also acting as an individualized treatment recommender system by predicting which treatment assignment would lead to improved survival given an individual patient's baseline covariates.

Survival trees refer to decision tree-based machine learning methods that have been adopted for analysis of survival data[34,40]. Briefly, decision trees are partitioning algorithms based on recursive binary splits, where the selection of covariate and cutoff value at each decision point is determined by maximizing the in-group homogeneity (i.e. survival time) of the resulting subgroups. Advantages of survival trees include lack of baseline survival distribution assumptions, natural clustering of subjects, and clear interpretability in terms of important variables and cutoff points. Disadvantages include the lack of effect size estimate for each covariate, and for individual trees, high sensitivity to small changes in input data.

The disadvantages of a single survival tree can be overcome by combining many trees, an instance of a general approach called ensemble learning[40]. The main principle behind ensemble learning is to take multiple base learners (e.g. individual survival trees), each of which may be good at making predictions based on certain data patterns, and combine their outputs into a single prediction. Ensemble learning models have an allocation function, which decides how much training data each base learner receives, and a combination function, which decides how to combine the prediction outcomes of each base learner (e.g. equal vs. weighted voting). Advantages of ensemble learning include improved robustness and performance, especially in extremes of data size or imbalance, and its ability to incorporate

distinct data domains for each base learner. Disadvantages include greater model complexity and increased computation needs. In the context of survival trees, ensemble learning methods include bootstrap aggregating (bagging)[41], random forests[42], and boosting[43].

Generalized additive models are generalized linear models in which the output variable depends on a smooth function of predictor variables, with the goal of estimating the shape of such functions. In its basic form a generalized additive model is formulated as $y = f_1(x_1) + f_2(x_2) + \ldots + f_p(x_p)$, where the functions f_i replace the coefficients β_i in ordinary linear regression. Interaction terms may also be modeled with functions $f_{ij}(x_i, x_j)$, which can be expanded to arbitrarily higher dimensions up to a full complexity model $y = f(x_1, \ldots, x_p)$. Estimation of the predictor functions can be done via smoothing splines or local linear regression using the backfitting algorithm, although a boosting trees approach is also feasible. Generalized additive models have been adapted for survival analysis[45] and studied in the healthcare setting, where modeling up to pairwise interactions showed good empiric performance in predicting clinical outcomes (pneumonia risk and hospital readmissions), comparable to other machine learning methods (random forest, LogitBoost)[44]. A major advantage of generalized additive models over other machine learning methods is interpretability, enabled by its modular construction and intuitive visualization through bivariate shape plots[44].

Support vector machines refer to a class of supervised machine learning methods based on finding an optimal decision boundary in high-dimensional covariate space, combining aspects of nearest neighbor classification and linear regression modeling. It is a powerful technique that has been adapted for survival analysis[46], with advantages such as the ability to find a globally optimal solution by

capturing complex, nonlinear relationships occurring in high dimensional space, compact solution representation using only a subset of training data, and strong empiric performance with good generalization capacity. Its disadvantages include the arbitrary nature of kernel function selection and lack of interpretability.

The recurrent neural network family of deep learning methods are designed to handle sequential data, and can naturally represent time-varying effects of covariates in survival analysis. Giunchiglia *et al.* proposed RNN-SURV[29], a deep learning survival model based on a composite framework of a 2-layer feed forward neural network and a 2-layer recurrent neural network with LSTM cells, used to predict both the survival function and individualized risk score for each patient. RNN-SURV demonstrated improved discriminative power of up to 28.4% over traditional survival models (Cox PH and Aalen's additive hazards models) and machine learning based methods (including DeepSurv[28]) in several healthcare data sets including the United Network for Organ Sharing transplant and waitlist registries. While directly interpreting the parameters of a deep learning model remain difficult, RNN-SURV allows for plotting of the unique survival function for each patient, a useful visualization tool for discussing individual risk.

In another deep learning application, Lee *et al.* proposed DeepHit [30], a multitask deep neural network to address the competing risks scenario. DeepHit architecture consists of a single shared multilayer perceptron network, upon which several cause-specific sub-networks are built. To enforce joint learning, the output of each sub-network is combined in a final softmax output layer, which predicts probability of each event at each time point. DeepHit learns the distribution of survival times without making any survival distribution assumptions, and is able to model time-varying effects of covariates. Significant improvements in prediction

accuracy in terms of the time-dependent concordance index (C^{td}) were demonstrated over previous state-of-the-art methods in multiple large health datasets.

iii) Model performance and evaluation

To compare the performance of prediction models, standardized measures of comparison are needed. Traditionally, there are three main areas of assessment: overall model performance, discrimination, and calibration. More recently, measures such as reclassification tables and net reclassification improvement have been developed[48]. Relevant details are discussed below.

Assessing overall model performance can be thought of as measuring the distance between predicted outcomes and actual outcomes. For continuous outcomes, mean absolute error or explained variation (R^2 statistic) can be used, whereas for binary outcomes, measures such as accuracy, balanced accuracy, and Brier score are commonly used[48]. In survival analysis, prediction models can be evaluated for their accuracy at a given time point, where performance measures for a binary outcome would apply, or evaluated over the entire duration of follow up, where measures such as the integrated Brier score can be computed using time-dependent weights based on censoring information[38].

Beyond assessing overall model performance, its component characteristics, i.e. discrimination and calibration, are important to consider in the context of clinical decision making[48,49]. Discrimination refers to whether a model can accurately distinguish between those who will have an event versus those who will not. A common measure of discrimination is the concordance statistic (c-index), which in a binary setting is identical to the area under the receiver operating characteristic (ROC-AUC) curve, which plots the true positive rate (sensitivity)

against the false positive rate (1- specificity). An extension of the c-index is available (Harrell's c-index) for survival analysis. In contrast, calibration, which assesses a model's goodness of fit, refers to the agreement between predicted and observed outcomes. Calibration may be measured via graphical assessment of a plot of predicted vs. observed outcomes where the 45 degree line would indicate perfect calibration, or by using statistical tests such as the Hosmer-Lemeshow test.

It is important to consider context when utilizing a prediction model. In a clinical scenario where the focus is to accurately identify highest risk candidates for targeted therapy, discrimination may be the most important consideration, in order to maximize therapeutic effect and minimize side effects. In other clinical scenarios where the focus is in discussing prognosis or determining thresholds for initiation of therapy, calibration may be more important. Thus the importance and utility of a given clinical prediction model will vary depending on the specific clinical scenario[49].

For prediction models based on machine learning, appropriate measures of assessment depend on the model type[38]. For supervised learning methods, standard model assessment metrics described above can be used, as the outcome event is clearly defined. In the binary outcome setting, especially in cases of class imbalance (e.g. rare outcome events), additional measures of interest may include precision, recall, F1 score, and precision-recall (PR) curve, which are related to positive event detection rate. More generally, measures such as Cohen's kappa statistic, which measures improvement in classifier model performance over random guesses based on observed frequency of events, can be useful for comparing multiple prediction models. In contrast to supervised learning, for unsupervised learning methods, directly assessing model performance is not possible as the model is not trained with regards to a specific task; in such cases, a common approach is to apply the trained

model to different prediction tasks and then assess comparative performance metrics across those tasks as an indirect measure of overall model performance.

Beyond assessing model performance in the derivation dataset (internal validity), a prediction model must be validated in external dataset(s) before it can be considered for deployment in actual clinical practice (external validity). Preferably, the model should be developed in a similar setting as the intended use environment. In the context of cardiovascular risk prediction, the PCE is meant to be used for the general undifferentiated patient in the U.S. without pre-existing ASCVD; as such, novel prediction models should ideally be developed and validated in similar patient cohorts representative of the intended target population.

iv) Model interpretability

While complex prediction models may have better performance compared to simpler counterparts, interpretability is an important factor when considering real-world usage of a prediction model[50]. For regression-based models such as the PCE, adding additional covariates and interaction terms may increase model performance, but will make it more cumbersome to use and interpret. When considering more complex models based on machine learning, different model types will have varying degrees of interpretability depending on factors such as identification of important variables, estimation of variable effect, assessment of interaction between variables, ability to incorporate existing knowledge, etc. The "black box" nature of certain machine learning methods are considered major barriers to clinical adoption, and such models will likely need extensive external validation and further work regarding interpretability before they can be incorporated into clinical practice.

iv) Causal inference

Beyond interpretability, causality is an important consideration in the clinical setting. Ideally, a clinician would like to know which factors cause adverse outcomes, so that modifiable risk factors can be addressed via therapy. However, prediction models based on observational data can only capture associations that are present in the data, and do not represent causal knowledge. While theoretical advances in causal inference have demonstrated that it is possible to extract causal information from observational data given causal hypotheses and assumptions represented as directed acyclic graphs[51,52], in many clinical settings the underlying mechanisms are multifactorial, complex, or unknown, making it difficult to apply such methods. Still, machine learning methods that can incorporate existing medical knowledge allow for richer disease representations based on EHR data[53], which may eventually lead to elucidation of complex pathophysiology and causal discovery.

3. New models for cardiovascular event risk estimation

a) Augmented PCE models

A common approach to improve predictive power of a regression model, such as the PCE, is to extend the model with additional factor(s) and assess their incremental and independent prediction value. For ASCVD event prediction, parameters from cardiovascular imaging studies and serum biomarkers are popular candidates. Several examples are discussed below.

Yeboah *et al.*[54] assessed whether the addition of coronary artery calcium score (CAC score; degree of coronary artery calcification seen on computed tomography), the ankle-brachial index (ABI; a measure of peripheral vascular disease), high sensitivity C-reactive protein (a serum marker of inflammation), and

family history of ASCVD to the PCE improve prediction of ASCVD events. To address the known overestimation of risk by the PCE in the study cohort (Multi-Ethnic Study of Atherosclerosis), the authors recalibrated the PCE prior to evaluation. Among the studied factors, CAC score, ABI, and family history were independent predictors of ASCVD events, although only the CAC score resulted in improvement in the overall c-index (0.74 to 0.76) when added to the recalibrated PCE. These results are consistent with the 2019 ACC/AHA recommendation to consider the CAC score as the test of choice for further risk stratification in intermediate risk patients[4].

Data from the ECG represent a potentially attractive addition to the set of risk factors considered in the PCE, as the ECG represents an unrepresented dimension of cardiac health (electrophysiology), is non-invasive and readily performed, and is easily interpreted and quantified. Prior studies have examined individual ECG components as a predictor of adverse cardiovascular outcomes, e.g. P wave duration[55], deep terminal negativity of P wave in V1 (DTNPV1) [56,57], QRS duration[58], QT interval[59-62], JT interval[62], and isolated ST-segment and T-wave abnormalities[63]. Individual ECG components have also been evaluated for their additive and independent predictive value to standard cardiovascular risk calculators such as the Framingham Risk Score and the PCE[14,15]. Beyond individual ECG components, groups of components have been evaluated in the framework of global electrical heterogeneity[64,65]. Despite their promise as additional predictors of cardiovascular risk, ECG data have not yet been incorporated into standard risk calculators such as the PCE or guideline recommendations.

b) Machine learning models

In contrast to augmenting the standard PCE, machine learning based models provide greater flexibility and higher capacity that are better suited for high-dimensional healthcare data. Applications to cardiovascular risk prediction are detailed below.

Ambale-Venkatesh *et al.*[66] applied the random survival forest technique to a wide range of clinical variables (735 total variables from imaging, noninvasive tests, questionnaires, and biomarker panels) collected in 6,814 participants from the Multi-Ethnic Study of Atherosclerosis cohort to develop separate prediction models for ASCVD and related outcomes (all-cause death, stroke, coronary artery disease, all cardiovascular disease, atrial fibrillation, heart failure). In this study, the random survival forest model outperformed both the PCE and variants of the Cox PH model (lasso regularization, forward and backward variable selection) in terms of c-index and Brier score over all outcomes. Interestingly, the most of the top 20 important predictors identified by the random survival forest model were not part of the PCE, and varied for each outcome. Furthermore, these variable importance ranking results could then be used to improve the performance of the Cox PH model. Overall, this study demonstrated the superior performance of the random survival forest method over the PCE and related Cox PH variants in predicting distinct cardiovascular outcomes and identifying important risk factors for each outcome, based on richly phenotyped data.

In another study involving the Multi-Ethnic Study of Atherosclerosis cohort, Kakadiaris *et al.*[67] developed a novel ASCVD event calculator based on support vector machines. The model was trained based on baseline clinical variables identical to that of the PCE, enhanced by the NEATER data augmentation

algorithm to address class imbalance. Based on 13-year follow up data of 6,459 patients, the support vector machine model significantly outperformed the PCE in ASCVD event prediction, in terms of sensitivity (0.86 vs. 0.76), specificity (0.95 vs. 0.56), accuracy (0.94 vs. 0.58), ROC-AUC (0.92 vs. 0.71), and net reclassification index of 0.49. Results were further validated in an external cohort (Flemish Study of Environment, Genes and Health Outcome cohort). In addition to ASCVD-related events, clinically relevant outcomes such as other cardiovascular events and statin therapy were also evaluated, with similar results. While the performance gain over the PCE was significant and showcased the predictive power of the support vector machine model over the Cox PH model, limitations included lack of adjustment for survival data and unclear interpretability.

In a retrospective cohort study of ~114,000 Veterans Health Administration patients in the U.S., Kennedy *et al.*[68] compared the Framingham Risk Score to various machine learning methods (parametric logistic regression, nonparametric generalized additive model, and gradient tree boosting), with cerebrovascular- and cardiovascular- death as the outcome events of interest. For machine learning models, data were augmented with additional details such as medication, laboratory, vital signs, disease diagnoses and other data elements drawn from the EHR. Machine learning models performed significantly better compared to the Framingham Risk Score when using the same variables (ROC-AUC 71% vs. 73%), and even better when using augmented variables (ROC-AUC 78%, net reclassification improvement 0.29). This study demonstrated that internally developed prediction models specific to a healthcare system could outperform general prediction models within a specific target population.

In a study of the Framingham Offspring cohort, Dogan *et al.*[69] constructed a new risk calculator for incident ASCVD events based on an ensemble of random forest models trained on four genetic (single nucleotide polymorphisms) and four epigenetic (DNA methylation) markers and their interactions. The integrated genetic-epigenetic model, in comparison to the PCE, predicted 5-year ASCVD risk with good accuracy (0.82), ROC-AUC (0.82), sensitivity (0.75 vs. 0.38) and specificity (0.73 vs. 0.85) in the internal validation sample. An interesting finding was that when regression analysis was performed between 8 traditional cardiovascular risk factors versus 8 genetic and epigenetic markers and their interaction terms, about half of the statistically significant relationships occurred between single loci and a cardiovascular risk factor, while the other half were between the genetic/epigenetic interaction term and a cardiovascular risk factor. These findings are illustrative of the complex interplay underlying genetic, environmental, and clinically manifested risk factors of ASCVD. Overall, this study demonstrated the potential of using genetic and epigenetic data in a machine learning framework for cardiovascular risk prediction, while limited by lack of external validation.

In a prospective cohort study of ~380,000 outpatients based in the U.K, Weng *et al.*[70] utilized 30 routinely available clinical data elements (demographic, social, laboratory, diagnoses, and treatment categories) to build four machine learning algorithms (logistic regression, random forest, gradient boosting machines, and neural network with multilayer perceptron architecture) and compared them to the PCE. In a 75/25 split internal validation sample, the PCE performed reasonably well (ROC-AUC 0.728), while machine learning methods showed slight gains in prediction performance (+1.7% to +3.6% gain in ROC-AUC, highest for the neural network model). Variable importance was determined based on coefficient size for

the PCE and logistic regression, variable importance ranking based on selection frequency in decision-tree based models (random forest and gradient boosting machine), and with assessment of overall variable weighting in the neural network. Among the PCE variables, age, sex, race, and smoking status were featured prominently in the machine learning models. This study was significant in terms of demonstrating the generalizability of the PCE to a population outside the U.S., highlighting the potential performance gain from machine learning, and identifying important variables based on a comparison between different machine learning methods.

Chapter III.

METHODS

1. Description of dataset (NHANES III)

The third iteration of the National Health and Nutrition Examination Survey (NHANES III)[71] consists of healthcare survey data compiled from a nationally representative sample of 39,695 persons from 1988 to 1994. In addition to survey features including demographic, historical, and physical elements, biochemical laboratory studies and ECG data are available for a subset of the surveyed population. Mortality outcome data, including the cause of death, are available via linked National Death Index files. As a nationwide probability-weighted sample consisting of mostly healthy persons with defined outcome event data over long-term follow up, the NHANES III dataset is well suited for development of population-based event prediction models. As a publicly available data set, Institutional Review Board approval was not required for this study.

As not all data were available for all subjects, subsets of the NHANES III dataset were used for this study. Those with complete demographic information (N=17,860) formed the base group, while those with complete PCE data components (i.e. additional medical history, exam, social, laboratory, and medication data) and ECG measurements (N=7,067) formed the main study group.

2. Data analysis

All data were imported and analyzed using R 3.5.1[72] and R Studio[73] statistical software. For statistical analysis and general machine learning, publicly

available R packages (tidyverse, data.table, survival, survminer, pROC, PRROC, mice, ROSE, caret, SuperLearner) were utilized where applicable. For deep learning, keras and CUDA software packages were adopted for R implementation and trained with NVIDIA GPU. R code for the PCE was provided by Dr. Jarrod Dalton of the NEOCARE research group at Cleveland Clinic Foundation via personal communication.

3. Data preparation: Clinical data

All adult (age 18 and above) participants from NHANES III were included in this study, except those with pre-existing cardiovascular disease (history of heart attack (HAF10), congestive heart failure (HAC1C), or stroke (HAC1D)). Data were recorded and averaged from relevant sections of NHANES III (questionnaire, examination, laboratory, medications, linked National Mortality Index files), and secondary measures (e.g. body mass index (BMI), pulse pressure, adjusted total cholesterol, adjusted HDL cholesterol) were computed according to descriptions below.

Baseline age and sex were recorded as noted in multiple parts of the survey. Race categories were simplified to White, Black, or Other based on the DMARACER variable. Vital signs were combined as the median value among up to seven recorded measurements between the questionnaire (Systolic blood pressure: HAZA8AK1, HAZA8BK1, HAZA8CK1, HAZA8DK1; Diastolic blood pressure: HAZA8AK5, HAZA8BK5, HAZA8CK5, HAZA8DK5; Heart rate: HAZA5R) and exam (Systolic blood pressure: PEP6G1, PEP6H1, PEP6I1; Diastolic blood pressure: PEP6G3, PEP6H3, PEP6I3; Heart rate: PEP6DR) data. Median pulse pressure was computed for each participant based on above data. For body measurements, corresponding

Exam elements were recorded (weight: BMPWT, height: BMPHT, body mass index: BMPBMI), although secondarily computed BMI was used for subsequent analysis.

For history of hypertension, qualifying criteria included answering 'yes' to questionnaire items (HAE4A, HAE5A), while blood pressure measurements were recorded separately as above. For history of hyperlipidemia, qualifying criteria included answering 'yes' to questionnaire item (HAE7), while lipid panel laboratory measurements (total cholesterol, HDL cholesterol, low density lipoprotein cholesterol, triglycerides) were separately recorded. For history of diabetes, qualifying criteria included answering 'yes' to questionnaire item (HAD1) or having laboratory values of any fasting glucose >=126 or HgbA1c >=6.5. For current tobacco use, qualifying criteria included answering 'yes' to questionnaire items (HAR3, HAR24, HAR27) or exam item (MYPB5).

Data regarding medication use were obtained from a separate medications file. Medication codes used to identify treatment for high blood pressure included: 0506: antihypertensives, 0507: diuretics, 0510: calcium channel blockers, 0512: beta blockers, 0513: alpha agonist/alpha blocker, 0514: angiotensin converting enzyme inhibitors. In addition, answering 'yes' to questionnaire items HAE4A, HAE5A also qualified for history of antihypertensive treatment. Medication codes used to identify treatment for high cholesterol included: 0912: hyperlipidemia.

For those participants taking medications for high cholesterol, laboratory values were adjusted to adjust for average statin effect on cardiovascular outcomes (total cholesterol: 21% reduction, HDL cholesterol: 3.5% increase). Other laboratory measurements relevant to cardiovascular disease, including HgbA1c and C-reactive protein, were also recorded.

For outcomes, the main outcome of interest was cardiac death occurring during 10 years of follow up. Secondary outcomes of interest were all-cause death and cerebrovascular death. Death status and cause of death were determined using codes International Statistical Classification of Diseases and Related Health Problems - Tenth Revision (ICD-10). Cardiac death was identified by ICD-10 codes: I00-I09, I11, I13, I20-I51. Cerebrovascular death was identified by ICD-10 codes: I60-I69. All other deaths with or without recorded underlying cause of death were included in all-cause death. Other outcomes related to ASCVD, including nonfatal myocardial infarction and stroke, were not available in the NHANES III dataset. Outcome events were right-censored at maximum follow-up of 10 years for analysis.

4. Data preparation: ECG data

Combined ECG data in NHANES III were available in 166 column entries. Among these, 133 columns consisted of direct ECG measurements while 33 columns consisted of other ancillary data and interpretations based on the Minnesota code. Columns not based on direct measurements were excluded, due to lack of clinical relevance or requirement of human interpretation and labeling. Due to high proportion of missing values, preprocessing steps for ECG data included removing rows (participants) and columns (ECG measure) which had >50% missing data. For the remaining missing values, multiple imputation was performed using the Multivariate Imputation by Chained Equations (mice) package[74], based on demographic and other ECG measurements. Further preprocessing steps included converting the rhythm code (ECPBEAT) to binary (Sinus vs. non-sinus rhythm) to avoid data sparsity and replacing QT interval (ECPQT) with corrected QT interval

(QTc) based on Bazett's formula. For machine learning, all ECG data columns were standardized to have a mean of 0 and standard deviation of 1.

5. Data splitting and augmentation

All available data were split into train:test partitions in 80:20 ratio, based on random sampling. Given the low frequency of outcome events and resulting class imbalance, the training set was augmented by 1) oversampling of positive events and 2) synthetic data generation using the Random Over-Sampling Examples (ROSE) package[75].

6. Model training

All models were trained on base, oversampled, and synthetic train datasets and with different data combinations (PCE variables, PCE + ECG variables, demographic + ECG variables, etc.) using 10-fold cross validation. Model performance was assessed in the test set. For traditional survival analysis, the PCE was implemented based on published parameters, while various survival models (Cox PH models, Cox PH models with L1 (lasso) regularization) were trained using survival[76], survminer[77], glmnet[78] packages. For machine learning for survival analysis, the random survival forest method was implemented using the randomForestSRC[79] package. For machine learning for classification, where probability of cardiovascular mortality outcome at 10 years was assessed as a binary outcome, models based on logistic regression, random forest, gradient boosting machine, support vector machine, and neural network were implemented using the caret[80] package. For neural networks, the R implementation for keras[81] package was used to design a multilayer perceptron with 3 hidden layers of 16 units each with

RELU activation for the hidden layers and sigmoid activation for the final layer, with he_uniform initialization and 25% dropout regularization. Finally, ensemble models were trained using the SuperLearner[82] package, where optimal model weighting was determined based on 5-fold cross validation and maximizing overall ROC-AUC using the Nelder-Mead method.

7. Model assessment and comparison

Survival models were compared for discrimination using Harrell's c-index and assessed for calibration by plotting their calibration curves. Classification models and their ensembles were compared using accuracy, Cohen's kappa, sensitivity, specificity, ROC-AUC, and PR-AUC. In the final combined comparison, classification performance metrics were computed for survival models by assessing their performance at 10 years, with varying probability cutoffs applied to maximize PR-AUC.

For variable importance comparison, hazard ratios were compared in survival and regression models, while variable importance rank metric from the caret package was used for classification models. For aggregate assessment of classification models, the cumulative count of top ten important predictors for each model was computed and plotted on an electrocardiogram for visual assessment and clinical interpretation.

Chapter IV.

RESULTS

1. Study population

Among 20,050 adult persons who participated in the NHANES III survey, 1,742 participants who reported pre-existing cardiovascular conditions (heart attack, congestive heart failure, or stroke) were excluded from analysis, and outcome data from the linked National Death Index were available for 17,860 participants (base group). Among them, 7,067 participants had PCE and ECG data available after preprocessing (study group). Baseline characteristics of the study population are reported in **Table 1**.

There were notable differences between the base (N= 17,860) and study (N=7,067) groups. Most significantly, there were differences in age (mean ± standard deviation: 46.1 ± 19.8 vs. 59.2 ± 13.2), underlying medical comorbidities (hyperlipidemia 15.1% vs. 23.3%, hypertension 24.7% vs. 33.8%, diabetes 23.7% vs. 46.7%), medications (for blood pressure: 21.0% vs. 31.6%, for cholesterol 1.5% vs 2.9%), and mortality (all-cause: 15.0% vs. 19.8%, cardiac: 3.7% vs. 4.8%, cerebrovascular: 1.2% vs. 1.6%). Overall, the study group, who had additional laboratory and ECG measurements, were older and with more medical comorbidities compared to the base group.

	NHANES 3 with demographics (N=17,860)		NHANES 3 with PCE + ECG data (N=7,067)	
	Mean±SD [Range] or N (%)	Missing (%)	Mean [Range] or N (%)	Missing (%)
Age (years)	46.1±19.8 [18.0-90.0]		59.2±13.4 [40.0-90.0]	
Sex				
Male	8,260 (46.2%)		3,355 (47.5%)	
Female	9,600 (53.8%)		3,712 (52.5%)	
Race				
White	12,138 (68.0%)		5,223 (73.9%)	
Black	5,116 (28.6%)		1,645 (23.3%)	
Other	606 (3.4%)		199 (2.8%)	
Vital signs				
Heart rate (beats/min)	73±10 [40-164]	256 (1.4%)	74±10 [43-164]	0 (0.0%)
Systolic BP (mmHg)	125±20 [70-248]	279 (1.6%)	132±20 [78-248]	0 (0.0%)
Diastolic BP (mmHg)	74±11 [2-144]	292 (1.6%)	77±10 [16-136]	1 (0.01%)
Pulse pressure (mmHg)	51±17 [12-139]	292 (1.6%)	56±18 [20-189]	1 (0.01%)
Body measurements				
Weight (kg)	74.7±17.9 [21.8-218.9]	1,696 (9.5%)	76.2±17.1 [33.4-182.3]	9 (0.1%)
Height (cm)	166.4±9.9 [118.5-206.5]	1,678 (9.4%)	165.9±9.9 [126.9-200.0]	4 (0.06%)
Body mass index (kg/m2)	26.9±5.8 [11.7-79.6]	1,699 (9.5%)	27.6±5.5 [13.3-64.5]	9 (0.1%)
Medical history				
Hyperlipidemia	2,699 (15.1%)		1,645 (23.3%)	
Hypertension	4,402 (24.7%)		2,385 (33.8%)	
Diabetes	4,226 (23.7%)		3,297 (46.7%)	
Tobacco use (current)	4,854 (27.2%)		1,732 (24.5%)	
Medication use				
For blood pressure	3,749 (21.0%)		2,238 (31.7%)	
For cholesterol	273 (1.5%)		203 (2.9%)	
Lab measures				
Total cholesterol	203±44 [59-676]	2,549 (14.3%)	217±43 [59-501]	0 (0.0%)
HDL cholesterol	52±16 [8-196]	2,657 (14.9%)	51±16 [12-196]	0 (0.0%)
LDL cholesterol	126±38 [20-380]	11,336 (63.5%)	136±38 [20-361]	3,935 (55.7%)
Triglyceride	140±111 [23-3616]	2,585 (14.5%)	157±116 [27-3616]	6 (0.1%)
HgbA1c	5.5±1.1 [2.8-16.1]	2,382 (13.3%)	5.8±1.2 [2.7-16.1]	32 (0.5%)
C-reactive protein	0.5±0.8 [0.2-25.2]	2,659 (14.9%)	0.5±0.8 [0.2-18.3]	62 (0.9%)
Follow-up (years)	19.8±7.2 [0.0-27.2]		18.1±7.6 [0.0-27.2]	
Outcome (at 10 years)				
Death (all cause)	2,681 (15.0%)		1,399 (19.8%)	
Death (cardiac)	652 (3.7%)		338 (4.8%)	
Death (cerebrovascular)	209 (1.2%)		111 (1.6%)	

Table 1. Characteristics of the study population

Combined ECG data in NHANES III were available in 166 ECG variables, where 133 columns consisted of direct physical measurements, which were utilized for analysis, while 33 columns consisted of interpretations based on the Minnesota code and other miscellaneous data, which were excluded. There was a high frequency of missing data, with 44.7% of overall cells missing, and 34.8% of direct

measurement cells missing. Following removal of 68 rows (participants) and 41 columns (ECG measures) which had >50% missing per row or per column, 7,413 participants remained with mostly complete ECG data in 92 separate columns. The remaining 12.6% of cells with missing data were imputed with multiple imputation based on demographic and other ECG measurements. The 92 ECG variables used for analysis are highlighted in **Appendix A**.

Following 80:20 train:test data splitting, the base training set (n=5,654, event rate: 4.6%) and test set (n=1,413, event rate: 5.6%) were created. Given the low frequency of outcome event and resulting class imbalance, oversampled (n=11,308, event rate: 52.2%) and synthetic datasets (n=5,654, event rate: 49.4%) were created for training of machine learning models.

2. Assessment of the PCE in NHANES III

Because the PCE is only defined for white or black persons, NHANES III participants with other race categories were excluded, leaving 6,868 participants for assessment of the performance of the PCE. The c-index and calibration plot are shown in **Figure 2**. With a c-index of 0.82 (95% CI: 0.78-0.84), discrimination performance was good, however the calibration curve clearly showed overestimation of event risk by the PCE. Some degree of risk overestimation was expected, however, given that the PCE was originally designed to predict ASCVD events, which include, but are not exclusive to, cardiovascular mortality.

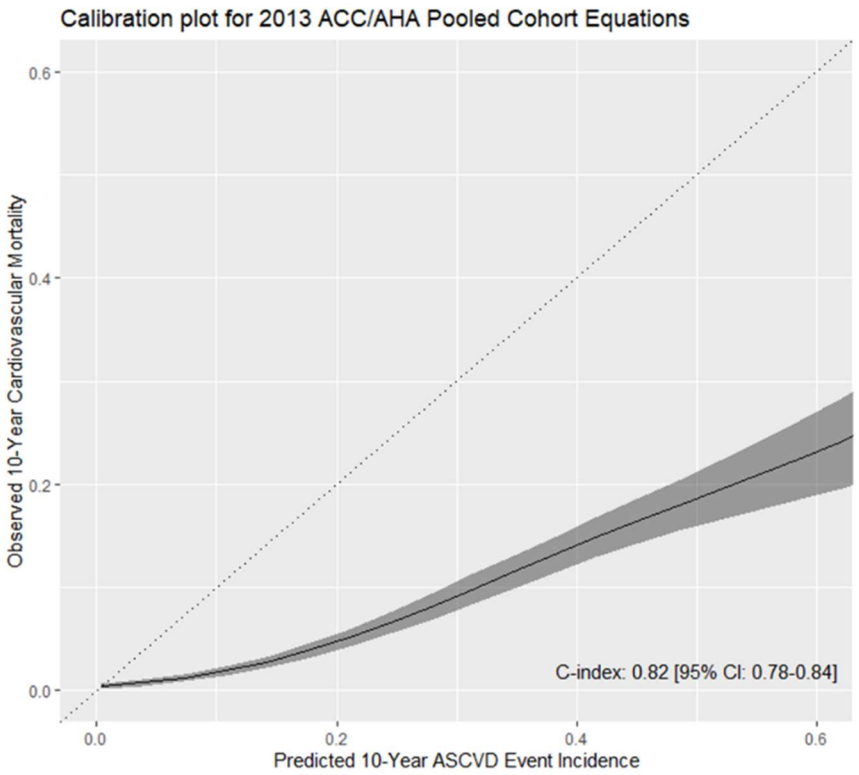

Figure 2. Assessment of the PCE in NHANES III

3. Survival models in NHANES III

a) Cox proportional hazards models

Seven Cox PH models were trained, based on various portions of clinical data available in NHANES III. Data subsets included: PCE variables only, PCE + ECG variables, demographic variables only, demographic + body measurements (BMI/vitals), demographic + ECG variables, demographic + body measurements + ECG variables, and ECG variables only. Comparison of model c-index and calibration assessment are shown in **Table 2**.

Model description	C-index	Calibration
Pooled cohort equations	0.82 [0.78-0.84]	Poor
Cox PH model (PCE variables only)	0.84 [0.83-0.85]	Fair
Cox PH model (PCE+ECG vars)	0.87 [0.86-0.89]	Good
Cox PH model (Dem variables only)	0.83 [0.81-0.85]	Poor
Cox PH model (BMI/Vitals+Dem vars)	0.83 [0.81-0.85]	Poor
Cox PH model (Dem+ECG vars)	0.87 [0.86-0.88]	Good
Cox PH model (BMI/Vitals+Dem+ECG)	0.84 [0.82-0.86]	Good
Cox PH model (ECG variables only)	0.82 [0.80-0.84]	Fair

Table 2. Cox PH model comparison

All of the fitted Cox PH models performed as well as the PCE in terms of discrimination, while only some had better calibration than the PCE. The two best performers were Cox PH models based on PCE + ECG variables and demographic + ECG variables, whose calibration plots are shown in **Figures 3A** and **3B**, respectively, which showed clear improvement in calibration compared to the PCE. Interestingly, when ECG data were available, the PCE variables did not materially add to model performance compared to demographic (age, sex, race) information.

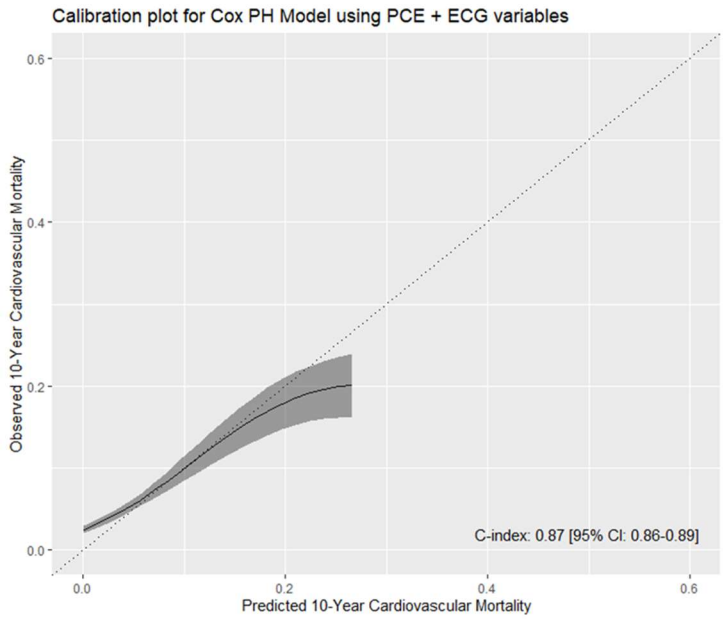

Figure 3A. Calibration plot for Cox PH model (PCE + ECG variables)

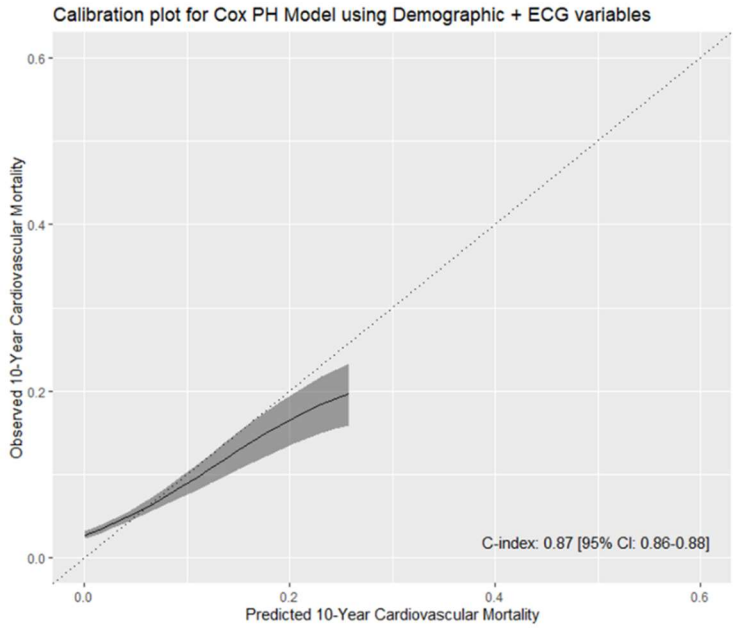

Figure 3B. Calibration plot for Cox PH model (Demographic + ECG variables)

b) Regularized Cox proportional hazards models

Given the numerous and highly correlated nature of ECG measurements, L1 (lasso) regularized Cox PH models were trained for automated variable selection. In terms of performance, L1-regularized Cox PH models had better discrimination compared to the PCE (**Table 3**), though with less outperformance compared to the full Cox PH models.

Model description	C-index
Pooled cohort equations	0.820
Cox PH with L1 regularization (PCE+ECG vars)	0.857
Cox PH with L1 regularization (Dem+ECG vars)	0.846
Cox PH with L1 regularization (BMI/Vitals+Dem+ECG)	0.852

Table 3. L1-regularized Cox PH model comparison

When hazard ratios were examined for variable selection, age, sex, systolic blood pressure, and current smoking status among PCE variables were retained in the L1-regularized Cox PH model, while race, total cholesterol, HDL cholesterol, and diabetes were excluded, in favor of ECG measurements (**Table 4**). Variable selection and hazard ratios were similar between PCE + ECG and demographic + ECG -based models. Among ECG variables, aside from sinus/non-sinus rhythm, the magnitude of hazard ratios for individual ECG components were very small (i.e. near 1), suggesting risk stratification value in aggregate rather than in specific ECG components.

Variable	PCE+ECG	Dem+ECG	Variable	PCE+ECG	Dem+ECG	Variable	PCE+ECG	Dem+ECG	Variable	PCE+ECG	Dem+ECG	Variable	PCE+ECG	Dem+ECG
HSSEX1	1.1607	1.1549	ECPLEADS			ECPRA3			ECPSA2			ECPJ3		
HSSEX2			ECPWIDTH			ECPRA4			ECPSA3			ECPJ4		
HSAGEIR	1.0927	1.0910	ECPDEPTH	1.0005	1.0002	ECPRA5			ECPSA5			ECPJ5	0.9970	0.9970
DMARACER2			ECPRATE	1.0015		ECPRA6			ECPSA6			ECPJ6		
DMARACER3			ECPPR	0.9994		ECPRA7			ECPSA7			ECPJ7		
Totalchol_adj			ECPQRS			ECPRA8			ECPSA8			ECPJ8		
HDLchol_adj			ECPAXIS1			ECPRA9			ECPSA9			ECPJ9		
medianSBP	1.0049		ECPAXIS2			ECPRA10	1.0001	1.0001	ECPSA10			ECPJ10		
HTN_tx1	1.2247		ECPAXIS3	0.9993		ECPRA11			ECPSA11	0.9999		ECPJ11	0.9973	0.9975
Diabetes1			ECPP1	0.9993	1.0000	ECPRA12			ECPSA12	0.9998		ECPJ12	0.9970	0.9964
Smoker_cur1	1.5297		ECPP2			ECPRD1			ECPSD1	1.0060	1.0039	ECPNTA4	1.0014	1.0013
			ECPP3			ECPRD2			ECPSD2			ECPPTA1	1.0000	
			ECPP4	0.9995	0.9999	ECPRD3			ECPSD3			ECPPTA2		
			ECPQA1			ECPRD4	1.0006	1.0001	ECPSD5			ECPPTA3		
			ECPQA4	1.0004		ECPRD5			ECPSD6			ECPPTA5		
			ECPQA10			ECPRD6			ECPSD7			ECPPTA6		
			ECPQA11			ECPRD7			ECPSD8			ECPPTA8	1.0002	1.0001
			ECPQD1	1.0126	1.0117	ECPRD8			ECPSD9			ECPPTA9	1.0001	
			ECPQD4			ECPRD9			ECPSD10			ECPPTA10		
			ECPQD10			ECPRD10			ECPSD11			ECPPTA11		
			ECPQD11			ECPRD11			ECPSD12	0.9982		ECPPTA12		
			ECPRA1			ECPRD12			ECPJ1			ECPSINUSNS	1.1471	1.1499
			ECPRA2	0.9999		ECPSA1			ECPJ2			ECPQTC	1.0011	1.0011

Table 4. Variable selection in L1-regularized Cox PH models

c) Random survival forest models

To assess whether non-linear, machine learning-based survival analysis would improve model performance, random survival forest models were trained, with various data combinations among PCE, demographic, and ECG variables, in base, oversampled, and synthetic training sets. Model comparison results are shown

in **Table 5**. Overall, the performance of random survival forest models was not significantly better compared to the PCE, both in terms of discrimination and calibration. Interestingly, the poor calibration in random survival forest models were due to underestimation of event risk, in contrast to overestimation of event risk in the PCE. An example calibration plot for the random survival forest model trained on PCE variables in the base training set is shown in **Figure 4**.

Model description	C-index	Calibration
Pooled cohort equations	0.82 [0.78-0.84]	Poor
Random survival forest (base) (PCE vars only)	0.817	Poor
Random survival forest (base) (PCE+ECG vars)	0.816	Poor
Random survival forest (base) (Dem+ECG vars)	0.809	Poor
Random survival forest (over) (PCE vars only)	0.799	Poor
Random survival forest (over) (PCE+ECG vars)	0.808	Poor
Random survival forest (over) (Dem+ECG vars)	0.798	Poor
Random survival forest (syn) (PCE vars only)	0.821	Poor
Random survival forest (syn) (PCE+ECG vars)	0.813	Poor
Random survival forest (syn) (Dem+ECG vars)	0.802	Poor

Table 5. Random survival forest model comparison

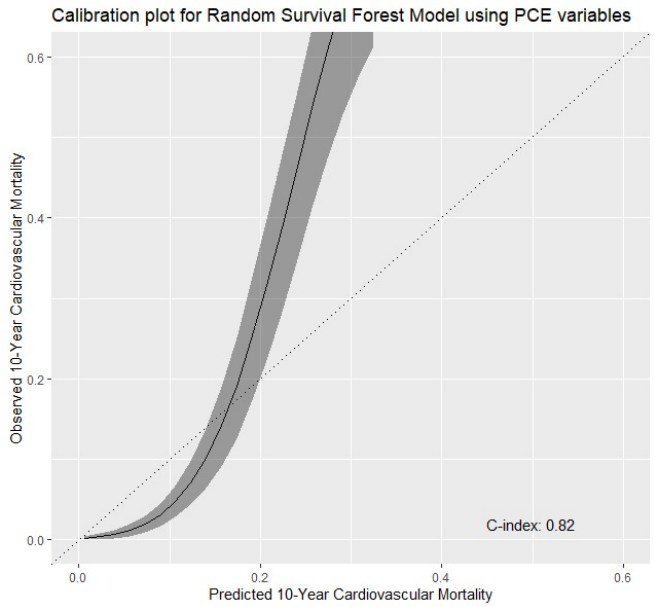

Figure 4. Calibration plot for random survival forest model (PCE variables)

4. Classification models in NHANES III

Machine learning models for classification were trained to predict cardiovascular mortality at 10 years, ignoring censoring. Models fitted include: logistic regression, random forest, gradient boosting machine, support vector machine, and neural networks, based on various data combinations. Ensemble models were created to combine multiple models and assess improvement in prediction performance.

a) Logistic regression models

In total, six logistic regression models were trained, with two data combinations ((1) PCE + ECG variables, (2) Demographic + ECG variables), for each of the three training sets (base, oversampling, synthetic). ROC curve, PR curve, and calibration plots are shown in **Figures 5A** , **5B**, and **5C**. For both PCE + ECG and demographic + ECG data, utilizing oversampling or synthetic training sets significantly improved AUC-ROC and PR-AUC. Calibration curves showed overestimation of risk by all models, though more smoothed for models derived from oversampling and synthetic training sets.

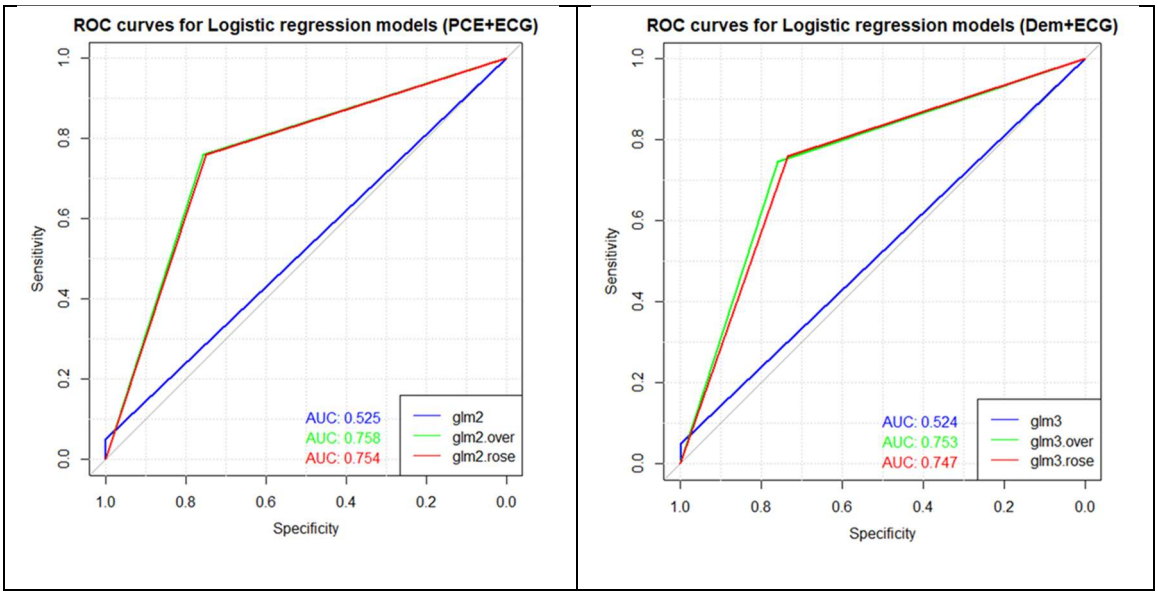

Figure 5A. Logistic regression models – ROC plot

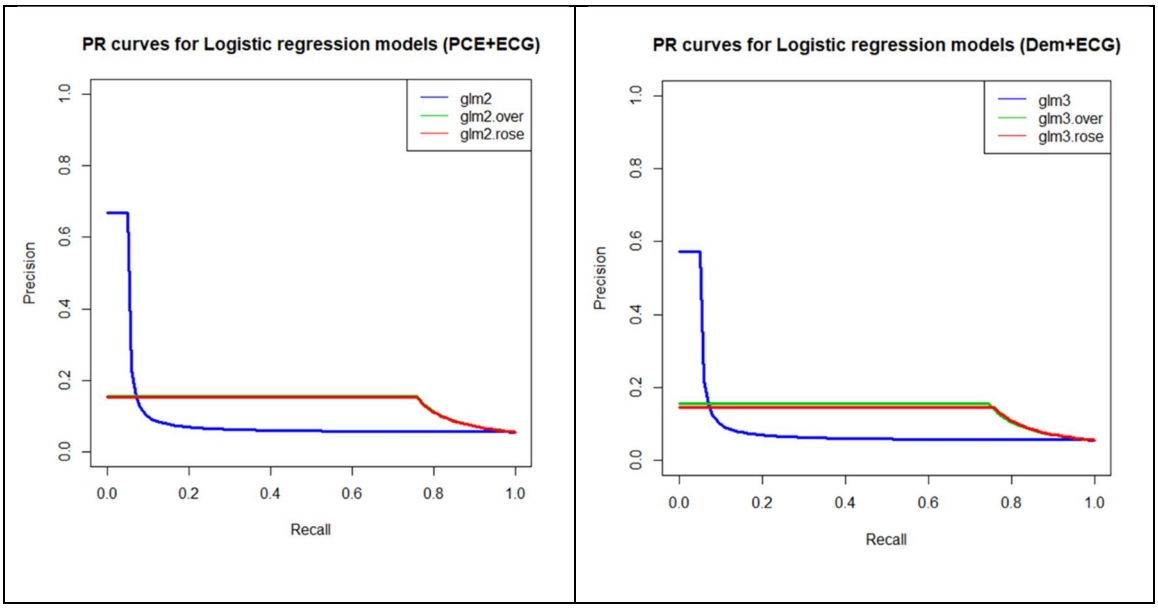

Figure 5B. Logistic regression models – PR plot

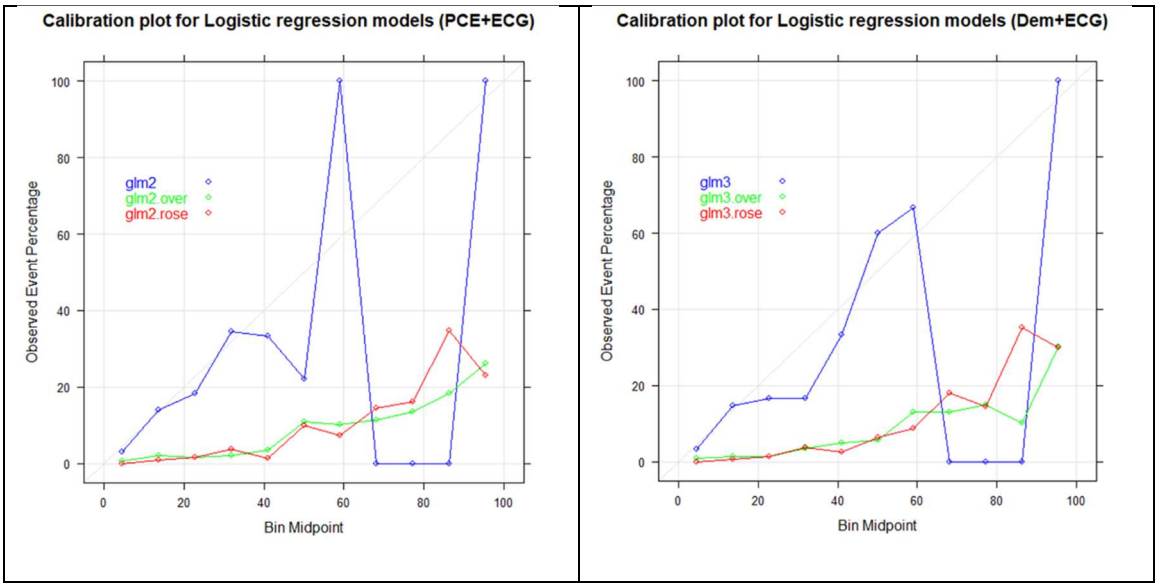

Figure 5C. Logistic regression models – Calibration plot

b) Random forest models

In total, twelve random forest models were trained, with two data combinations ((1) PCE + ECG variables, (2) Demographic + ECG variables) and two different optimization goals (accuracy vs. kappa) for each of the three training sets (base, oversampling, synthetic). Tuning hyperparameter mtry was optimized via automated grid search. A separate hyperparameter ntree was fixed at 500. ROC curve, PR curve, and calibration plots are shown in **Figures 6A** , **6B**, and **6C**.

Overall, random forest models performed very poorly, with negligible performance gain over random chance and without significant difference between training to maximize accuracy versus kappa. It is likely that these models were overfitted to training data, where specific cutoffs used for decision splits did not at all reflect generalizable patterns also present in the test data.

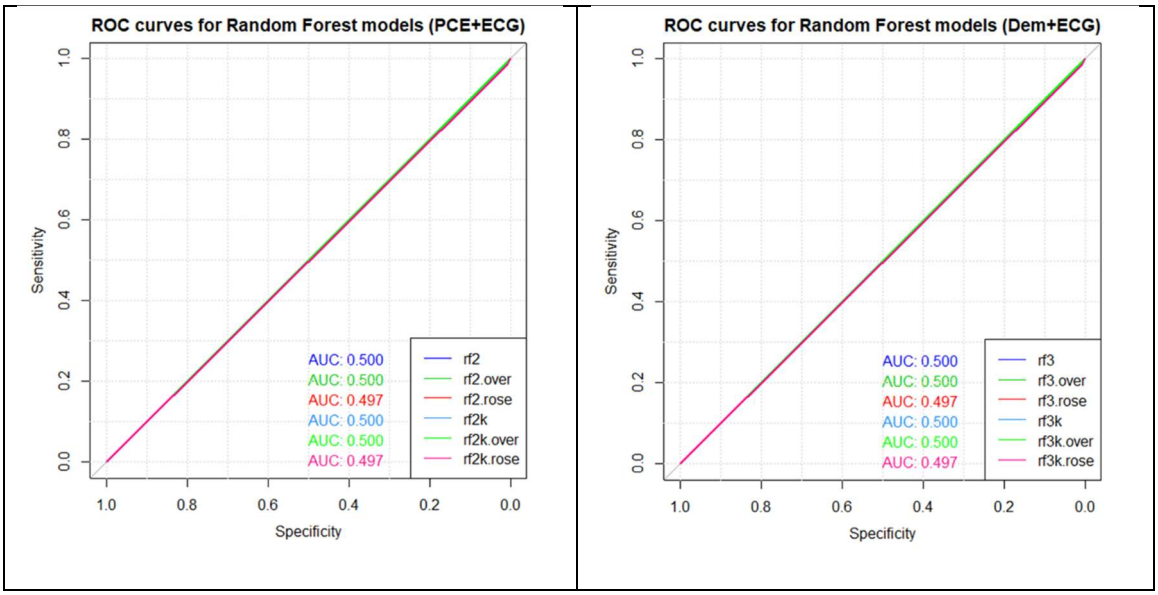

Figure 6A. Random forest models – ROC plot

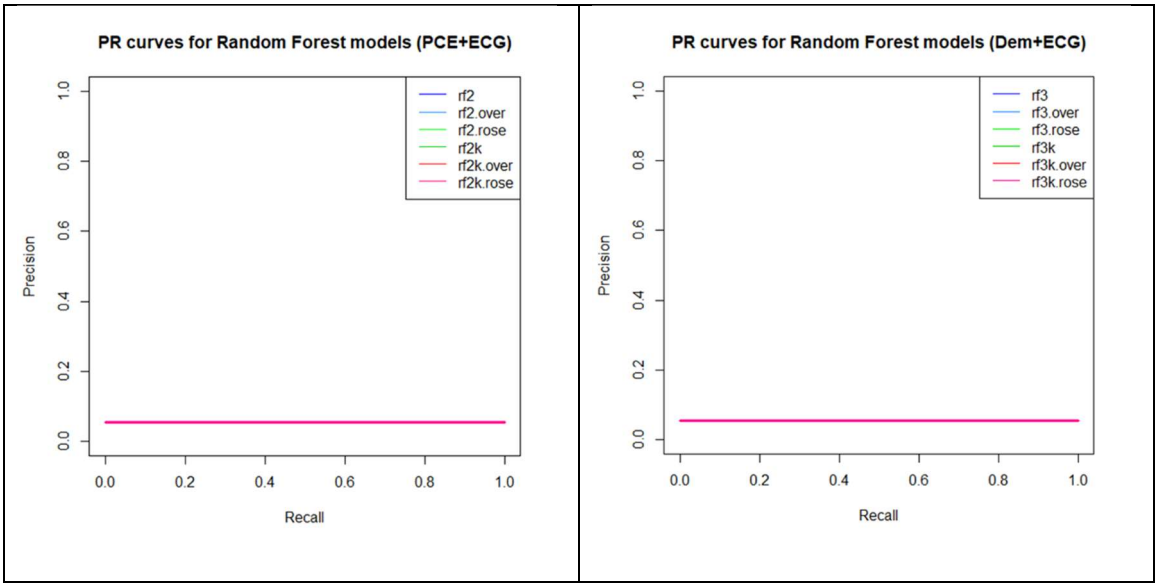

Figure 6B. Random forest models – PR plot

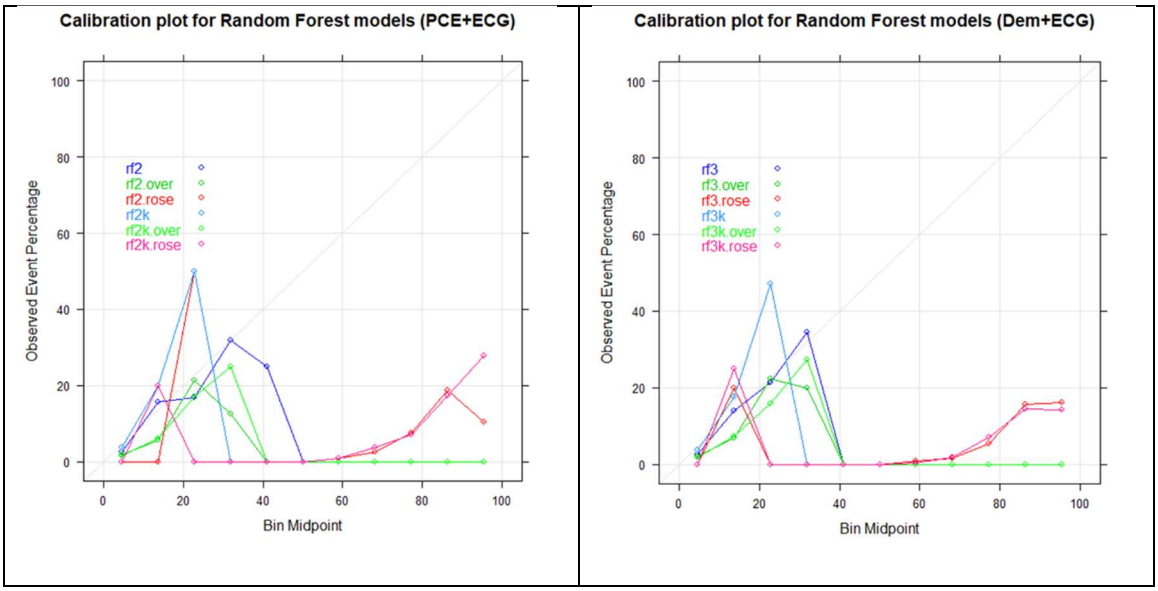

Figure 6C. Random forest models – Calibration plot

c) Gradient boosting machine models

In total, twelve gradient boosting machine models were trained, with two data combinations ((1) PCE + ECG variables, (2) Demographic + ECG variables) and two different optimization goals (accuracy vs. kappa) for each of the three training sets (base, oversampling, synthetic). Tuning hyperparameters included: nround, max_depth, eta, gamma, colsample_bytree, min_child_weight, and subsample, which were optimized for each model using automated grid search. The ROC curve, PR curve, and calibration plots for gradient boosting machine models are shown in **Figures 7A** , **7B**, and **7C**.

Overall, gradient boosting models performed very poorly, with negligible performance gain over random chance and without significant difference between training to maximize accuracy versus kappa. It is likely that these models were overfitted to training data, especially given the version of the gradient boosting

model used (extreme gradient boosting), which has more hyperparameters and higher model capacity, but also prone to overfit more easily.

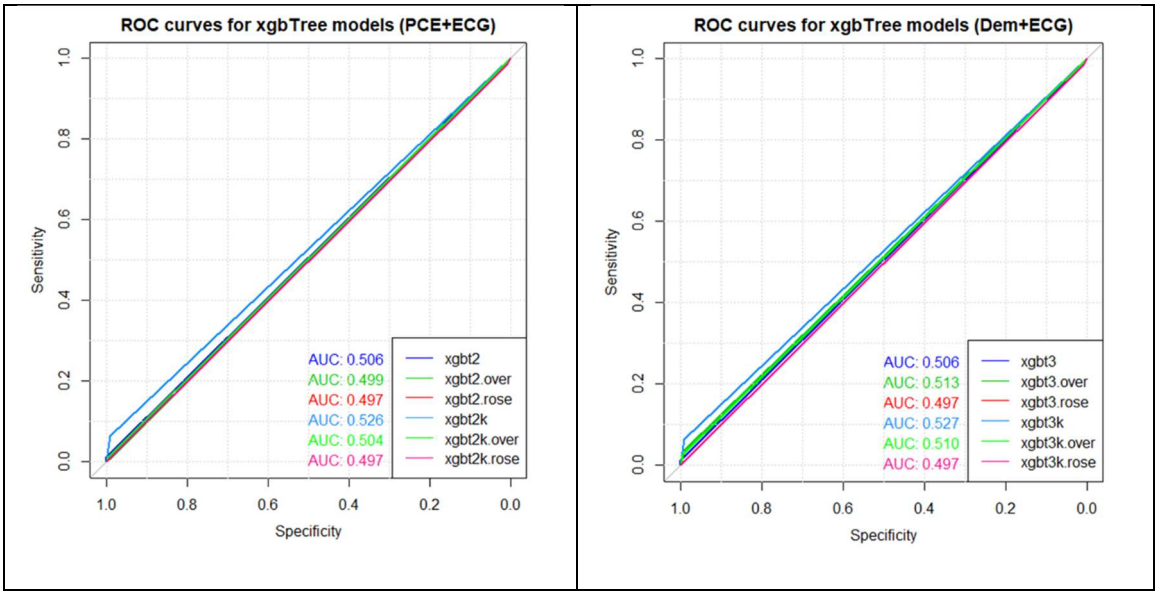

Figure 7A. Gradient boosting machine models – ROC plot

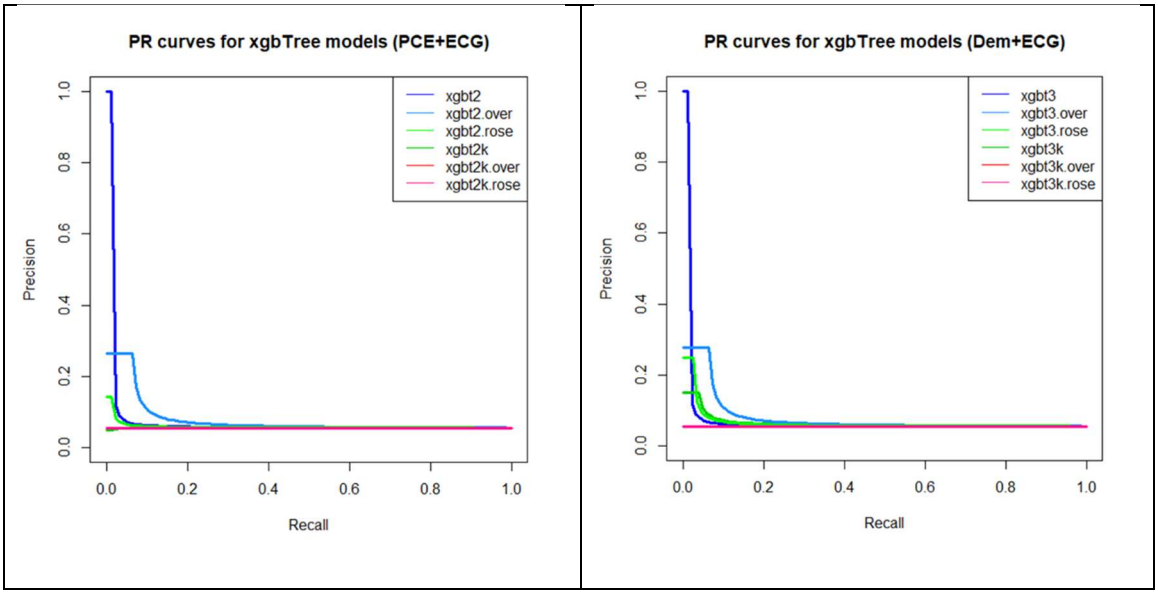

Figure 7B. Gradient boosting machine models – PR plot

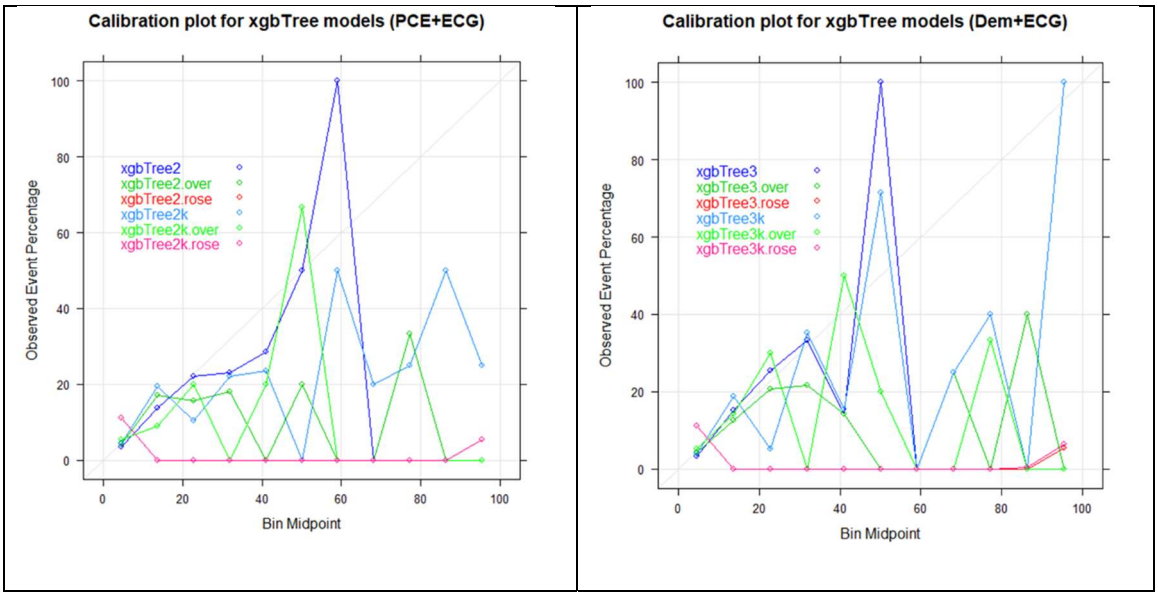

Figure 7C. Gradient boosting machine models – Calibration plot

d) Support vector machine models

In total, twelve support vector machine models were trained, with two data combinations ((1) PCE + ECG variables, (2) Demographic + ECG variables) and two different optimization goals (accuracy vs. kappa) for each of the three training sets (base, oversampling, synthetic). Tuning hyperparameters included: sigma and C, which were optimized for each model using automated grid search. The ROC curve, PR curve, and calibration plots for support vector machine models are shown in **Figures 8A** , **8B**, and **8C**.

For both data combinations, support vector machine models based on the synthetic training set showed superior performance in terms of ROC-AUC, PR-AUC, and smoother calibration within this model family. However, performance gains were marginal and calibration remained poor. There were no discernable differences

in performance between models based on PCE + ECG vs. demographic + ECG data groups.

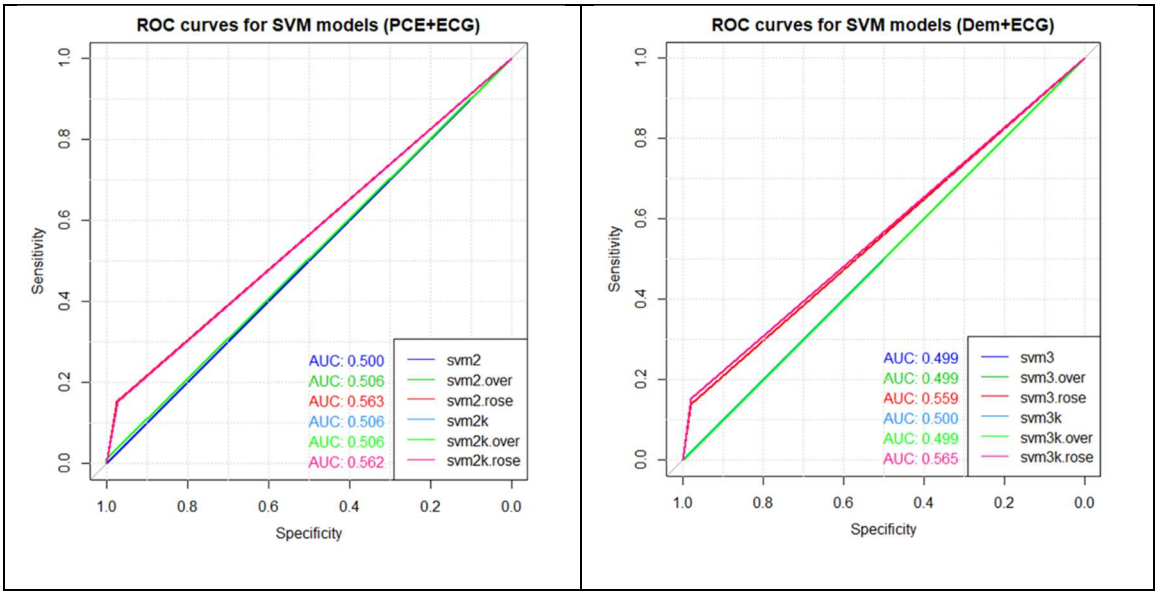

Figure 8A. Support vector machine models – ROC plot

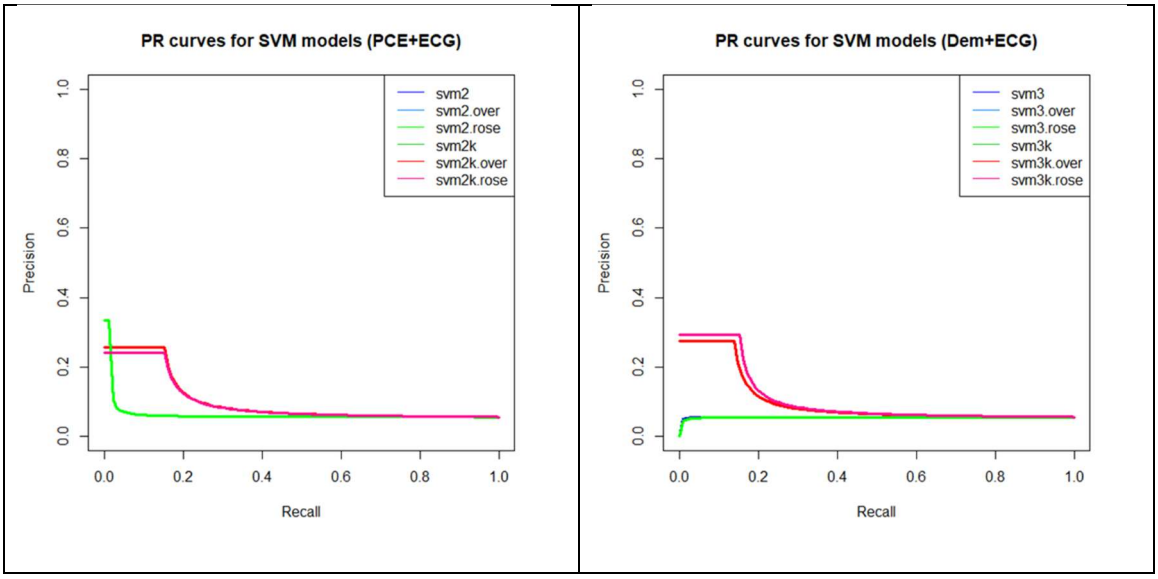

Figure 8B. Support vector machine models – PR plot

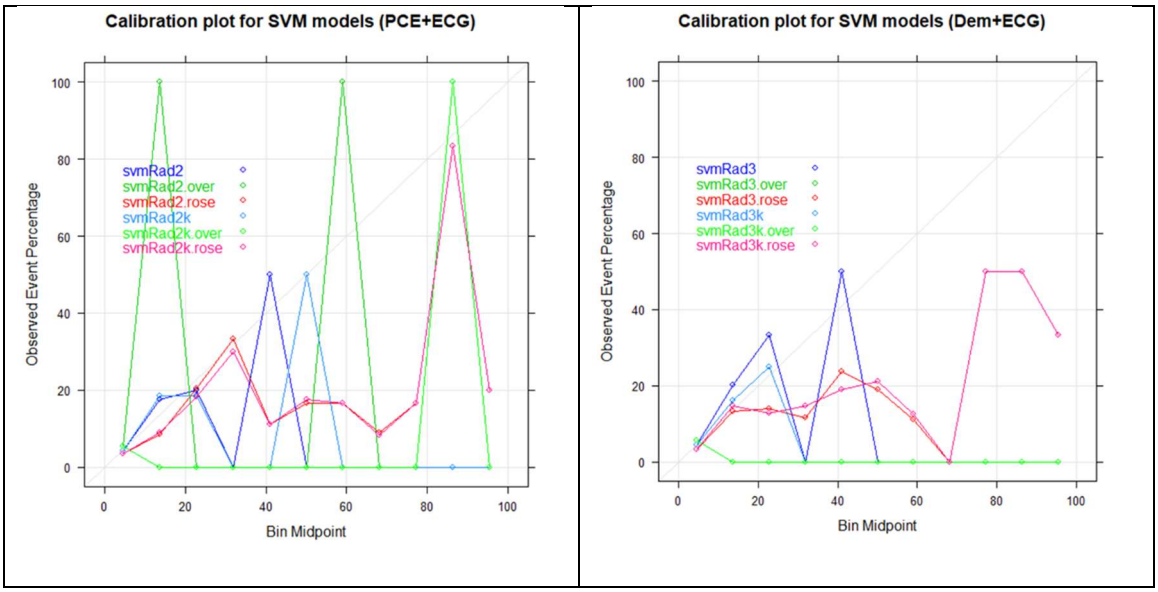

Figure 8C. Support vector machine models – Calibration plot

e) Neural network models

In total, twelve neural network models were trained, with two data combinations ((1) PCE + ECG variables, (2) Demographic + ECG variables) and two different optimization goals (accuracy vs. cost-sensitive learning with 20:1 weighting for positive cases) for each of the three training sets (base, oversampling, synthetic). Model design and tuning hyperparameters included choice of network architecture, number and size of hidden layers, activation function for each layer, batch size, cost function, initialization scheme, and regularization, which were determined via manual grid search. Final model specification was as follows: multilayer perceptron with five layers (one input layer, three densely connected hidden layers with 16 units each with ReLu activation, and one output layer with sigmoid activation), resulting in 2,305 trainable parameters. Models were trained with Rmsprop

optimizer, using accuracy vs. cost-sensitive accuracy metrics for the binary_crossentropy loss function. Further optimization steps included he_uniform initialization and 25% dropout regularization. Final epoch size was determined following examination of the training history for each neural network. The ROC curve, PR curve, and calibration plots for the neural network models are shown in **Figures 9A**, **9B**, and **9C**. Note for the calibration plot, sigmoid likelihoods were substituted for probabilities, which may result in graphical skew.

Overall, all neural network models performed well above baseline. In the PCE + ECG group, both oversampled and synthetic datasets resulted in strong model performance, while in the demographic + ECG group, only the synthetic dataset resulted in models with good ROC-AUC and PR-AUC. Calibration plots demonstrated a tendency for overestimation of risk for all neural network models.

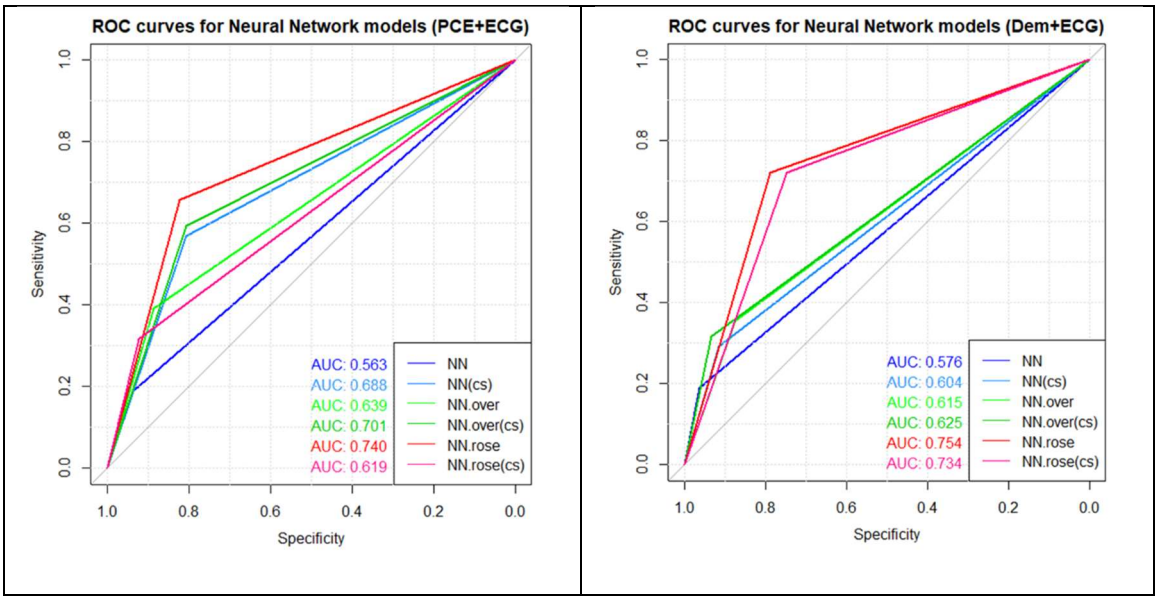

Figure 9A. Neural network models – ROC plot

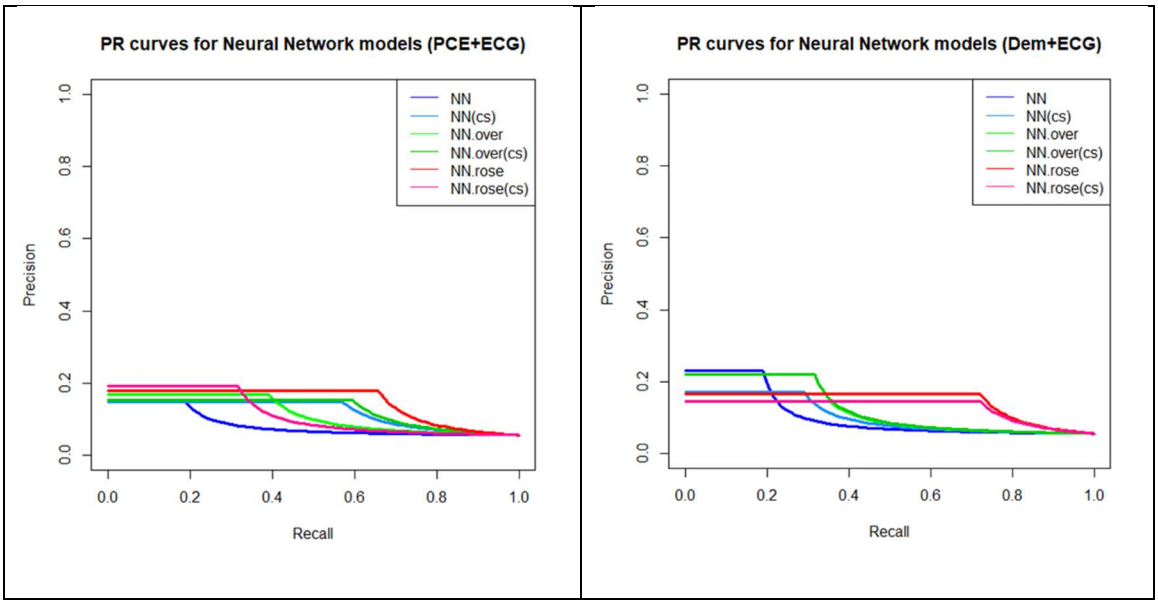

Figure 9B. Neural network models – PR plot

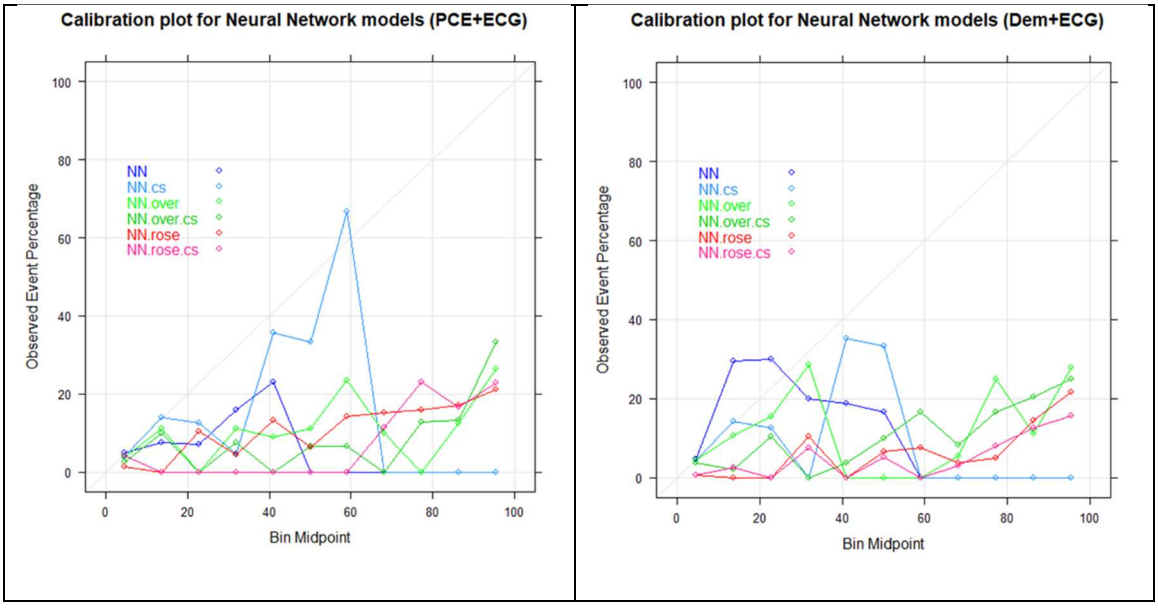

Figure 9C. Neural network models – Calibration plot

5. Ensemble model

In total, six ensemble models were created, for the two data combinations (PCE + ECG vs. Demographic + ECG) and three training sets (base, oversampling, synthetic). Of note, the neural network for the ensemble model differed from the individually trained neural network models, and consisted of a simplified feed-forward neural network with a single hidden layer and without optimizations such as nonrandom initialization or dropout regularization, due to limitations in implementation. Example ensemble model characteristics, for the ensemble model based on PCE + ECG variables in the base training set, are shown in **Table 6**. As expected, ensemble weights were skewed towards the better performing individual base learners (e.g. logistic regression), while the best average ROC-AUC was achieved by the ensemble model. The classification performance characteristics of the six ensemble models are shown in **Table 7**. Interestingly, the ensemble models based on the base training set appeared to have the best kappa, ROC-AUC, and PR-AUC in the PCE + ECG data group, while for the demographic + ECG data group, different ensemble models performed best in different categories.

Model name	Ensemble weight [range]		Average ROC-AUC [range]	
Mean (syn)	0.132	[0.093-0.159]	0.500	[0.500-0.500]
Logistic regression (syn)	0.308	[0.222-0.432]	0.801	[0.763-0.835]
Random Forest (syn)	0.152	[0.105-0.205]	0.774	[0.707-0.797]
Gradient boosting machine (syn)	0.174	[0.063-0.249]	0.781	[0.726-0.822]
Support vector machine (syn)	0.064	[0.000-0.167]	0.753	[0.691-0.790]
Neural network (syn)*	0.170	[0.159-0.205]	0.500	[0.500-0.500]
Ensemble model			0.812	[0.750-0.845]

*Single hidden layer FFNN vs. 3-hidden layer Deep FFNN with dropout in individual mode

Table 6. Ensemble model characteristics

Classification model	Accuracy	Kappa	Sensitivity	Specificity	ROC-AUC	PR-AUC
Ensemble model (base) (PCE+ECG)	0.834	0.234	0.633	0.846	0.740	0.157
Ensemble model (over) (PCE+ECG)	0.871	0.198	0.405	0.898	0.652	0.126
Ensemble model (syn) (PCE+ECG)	0.862	0.224	0.494	0.884	0.689	0.143
Ensemble model (base) (Dem+ECG)	0.857	0.237	0.544	0.876	0.710	0.152
Ensemble model (over) (Dem+ECG)	0.779	0.183	0.684	0.785	0.734	0.135
Ensemble model (syn) (Dem+ECG)	0.887	0.262	0.468	0.912	0.690	0.159

Table 7. Ensemble model performance

6. Model performance comparison

Performance metrics of all trained classification models, for the PCE + ECG data group and demographic + ECG data group, are shown in **Table 8** and **Table 9**, respectively. Highlighted in orange and light orange are baseline models for comparison, i.e. sample mean or base logistic regression. Highlighted in yellow are where the performance of a given model exceeded that of the baseline model for each performance category (column). Highlighted in blue are where an individual model performed best in each performance category. Highlighted in pink are where an ensemble model performed best in each performance category.

Overall, model performance was heavily affected by the low event rate in the test set, resulting in no information rate of 94%. Given this limitation, kappa and PR-AUC were considered important metrics for model evaluation. Within each model class, there was a clear trend towards improved performance when utilizing augmented training sets, either oversampled or synthetic, although this was not evident in the ensemble models.

Classification model (PCE+ECG vars)	Accuracy	Kappa	Sensitivity	Specificity	ROC-AUC	PR-AUC
Mean (No information rate)	0.944	0.000	0.000	1.000	0.500	NA
Logistic regression (base)	0.946	0.087	0.051	0.999	0.525	0.098
Logistic regression (over)	0.756	0.182	0.759	0.756	0.758	0.139
Logistic regression (syn)	0.750	0.176	0.759	0.749	0.754	0.136
Random forest (base)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (over)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (syn)	0.061	0.000	0.987	0.006	0.497	0.056
Random forest (base) (K)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (over) (K)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (syn) (K)	0.062	0.000	0.987	0.007	0.497	0.056
Gradient boosting machine (base)	0.945	0.024	0.013	1.000	0.506	0.072
Gradient boosting machine (over)	0.931	-0.002	0.013	0.986	0.499	0.056
Gradient boosting machine (syn)	0.061	0.000	0.987	0.006	0.497	0.056
Gradient boosting machine (base) (K)	0.938	0.082	0.063	0.990	0.526	0.078
Gradient boosting machine (over) (K)	0.941	0.014	0.013	0.996	0.504	0.059
Gradient boosting machine (syn) (K)	0.061	0.000	0.987	0.006	0.497	0.056
Support vector machine (base)	0.944	0.000	0.000	1.000	0.500	NA
Support vector machine (over)	0.943	0.020	0.013	0.999	0.506	0.062
Support vector machine (syn)	0.928	0.155	0.152	0.974	0.563	0.101
Support vector machine (base) (K)	0.943	0.020	0.013	0.999	0.506	0.062
Support vector machine (over) (K)	0.943	0.020	0.013	0.999	0.506	0.062
Support vector machine (syn) (K)	0.926	0.149	0.152	0.972	0.562	0.098
Neural network (base)	0.894	0.111	0.190	0.936	0.563	0.084
Neural network (over)	0.859	0.178	0.405	0.885	0.639	0.114
Neural network (syn)	0.813	0.214	0.658	0.822	0.740	0.148
Neural network (base) (cs)	0.793	0.161	0.570	0.806	0.688	0.119
Neural network (over) (cs)	0.795	0.171	0.595	0.807	0.701	0.125
Neural network (syn) (cs)	0.888	0.182	0.316	0.921	0.619	0.114
Ensemble model (base) (PCE+ECG)	0.834	0.234	0.633	0.846	0.740	0.157
Ensemble model (over) (PCE+ECG)	0.871	0.198	0.405	0.898	0.652	0.126
Ensemble model (syn) (PCE+ECG)	0.862	0.224	0.494	0.884	0.689	0.143

Table 8. Classification models comparison (PCE + ECG variables)
Base: original training set, Over: oversampled training set, Syn: synthetic training set. All models were trained to maximize accuracy, except for: (K) maximized kappa, or (cs): cost-sensitive learning with 20:1 weight for positive events.

For classification models based on PCE + ECG variables (**Table 8**), best kappa and PR-AUC were achieved by the neural network model based on the synthetic training set (kappa: 0.214, ROC-AUC: 0.740, PR-AUC: 0.148), while the logistic regression model based on the oversampled training set had the best ROC-AUC (kappa: 0.182, ROC-AUC: 0.758, PR-AUC: 0.139). Overall best performance was achieved by the ensemble model based on the base training set (kappa: 0.234, ROC-AUC: 0.740, PR-AUC: 0.157).

Classification model (Dem+ECG vars)	Accuracy	Kappa	Sensitivity	Specificity	ROC-AUC	PR-AUC
Mean (No information rate)	0.944	0.000	0.000	1.000	0.500	NA
Logistic regression (base)	0.945	0.085	0.051	0.998	0.524	0.093
Logistic regression (over)	0.758	0.181	0.747	0.759	0.753	0.137
Logistic regression (syn)	0.736	0.165	0.759	0.735	0.747	0.130
Random forest (base)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (over)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (syn)	0.061	0.000	0.987	0.006	0.497	0.056
Random forest (base) (K)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (over) (K)	0.944	0.000	0.000	1.000	0.500	NA
Random forest (syn) (K)	0.061	0.000	0.987	0.006	0.497	0.056
Gradient boosting machine (base)	0.945	0.024	0.013	1.000	0.506	0.072
Gradient boosting machine (over)	0.934	0.039	0.038	0.987	0.513	0.064
Gradient boosting machine (syn)	0.061	0.000	0.987	0.006	0.497	0.056
Gradient boosting machine (base) (K)	0.938	0.084	0.063	0.990	0.527	0.079
Gradient boosting machine (over) (K)	0.941	0.036	0.025	0.996	0.510	0.065
Gradient boosting machine (syn) (K)	0.061	0.000	0.987	0.006	0.497	0.056
Support vector machine (base)	0.943	-0.003	0.000	0.999	0.499	0.055
Support vector machine (over)	0.941	-0.005	0.000	0.997	0.499	0.055
Support vector machine (syn)	0.931	0.153	0.139	0.978	0.559	0.101
Support vector machine (base) (K)	0.944	0.000	0.000	1.000	0.500	NA
Support vector machine (over) (K)	0.941	-0.005	0.000	0.997	0.499	0.055
Support vector machine (syn) (K)	0.932	0.168	0.152	0.978	0.565	0.108
Neural network (base)	0.919	0.166	0.190	0.963	0.576	0.103
Neural network (over)	0.846	0.137	0.544	0.876	0.615	0.099
Neural network (syn)	0.841	0.244	0.633	0.853	0.754	0.145
Neural network (base) (cs)	0.881	0.155	0.291	0.916	0.604	0.103
Neural network (over) (cs)	0.900	0.208	0.316	0.934	0.625	0.125
Neural network (syn) (cs)	0.745	0.163	0.722	0.747	0.734	0.127
Ensemble model (base) (Dem+ECG)	0.857	0.237	0.544	0.876	0.710	0.152
Ensemble model (over) (Dem+ECG)	0.779	0.183	0.684	0.785	0.734	0.135
Ensemble model (syn) (Dem+ECG)	0.887	0.262	0.468	0.912	0.690	0.159

Table 9. Classification models comparison (Demographic + ECG variables)
Base: original training set, Over: oversampled training set, Syn: synthetic training set. All models were trained to maximize accuracy, except for: (K) maximized kappa, or (cs): cost-sensitive learning with 20:1 weight for positive events.

For classification models based on demographic + ECG variables (**Table 9**), best kappa, ROC-AUC, and PR-AUC were all achieved by a single individual model, the neural network model based on the synthetic training set (kappa: 0.244, ROC-AUC: 0.754, PR-AUC: 0.145). The ensemble model based on the synthetic dataset (kappa: 0.262, ROC-AUC: 0.690, PR-AUC: 0.159) performed better than the best neural network model in kappa and PR-AUC metrics, but not in ROC-AUC.

For individual classification models that performed best within each model family, a combined tabular comparison versus the PCE, Cox PH models, and ensemble models is shown in **Table 10**, and a summary visualization of their ROC curves, PR curves, and calibration plots are shown in **Figure 10**. Compared to the PCE, whose classification performance metrics were computed at 10 years (kappa: 0.170, ROC-AUC: 0.676, PR-AUC: 0.111), logistic regression and neural network models trained on augmented data sets and their ensemble models clearly outperformed in terms of all three metrics. As noted previously, there were no significant differences between utilizing PCE + ECG versus demographic + ECG data in terms of model performance.

Survival model	Accuracy	Kappa	Sensitivity	Specificity	ROC-AUC	PR-AUC
Pooled cohort equations	0.841	0.170	0.492	0.859	0.676	0.111
Cox PH model (PCE+ECG)	0.825	0.132	0.447	0.844	0.646	0.093
Cox PH model (Dem+ECG)	0.772	0.105	0.509	0.785	0.647	0.085
Classification model	**Accuracy**	**Kappa**	**Sensitivity**	**Specificity**	**ROC-AUC**	**PR-AUC**
Logistic regression (over) (PCE+ECG)	0.756	0.182	0.759	0.756	0.758	0.139
Logistic regression (over) (Dem+ECG)	0.758	0.181	0.747	0.759	0.753	0.137
Logistic regression (syn) (PCE+ECG)	0.750	0.176	0.759	0.749	0.754	0.136
Logistic regression (syn) (Dem+ECG)	0.736	0.165	0.759	0.735	0.747	0.130
Support vector machine (syn) (PCE+ECG)	0.928	0.155	0.152	0.974	0.563	0.101
Support vector machine (syn) (Dem+ECG)	0.931	0.153	0.139	0.978	0.559	0.101
Support vector machine (syn) (K) (PCE+ECG)	0.926	0.149	0.152	0.972	0.562	0.098
Support vector machine (syn) (K) (Dem+ECG)	0.932	0.168	0.152	0.978	0.565	0.108
Neural network (syn) (PCE+ECG)	0.813	0.214	0.658	0.822	0.740	0.148
Neural network (syn) (Dem+ECG)	0.841	0.244	0.633	0.853	0.754	0.145
Neural network (over) (cs) (PCE+ECG)	0.795	0.171	0.595	0.807	0.701	0.125
Neural network (syn) (cs) (Dem+ECG)	0.745	0.163	0.722	0.747	0.734	0.127
Ensemble model	**Accuracy**	**Kappa**	**Sensitivity**	**Specificity**	**ROC-AUC**	**PR-AUC**
Ensemble model (base) (PCE+ECG)	0.834	0.234	0.633	0.846	0.740	0.157
Ensemble model (over) (PCE+ECG)	0.871	0.198	0.405	0.898	0.652	0.126
Ensemble model (syn) (PCE+ECG)	0.862	0.224	0.494	0.884	0.689	0.143
Ensemble model (base) (Dem+ECG)	0.857	0.237	0.544	0.876	0.710	0.152
Ensemble model (over) (Dem+ECG)	0.779	0.183	0.684	0.785	0.734	0.135
Ensemble model (syn) (Dem+ECG)	0.887	0.262	0.468	0.912	0.690	0.159

Table 10. Combined models comparison
Base: original training set, Over: oversampled training set, Syn: synthetic training set. All models were trained to maximize accuracy, except for: (K) maximized kappa, or (cs): cost-sensitive learning with 20:1 weight for positive events. For survival models, classification performance was assessed at 10 years, with varying probability cutoffs to maximize PR-AUC.

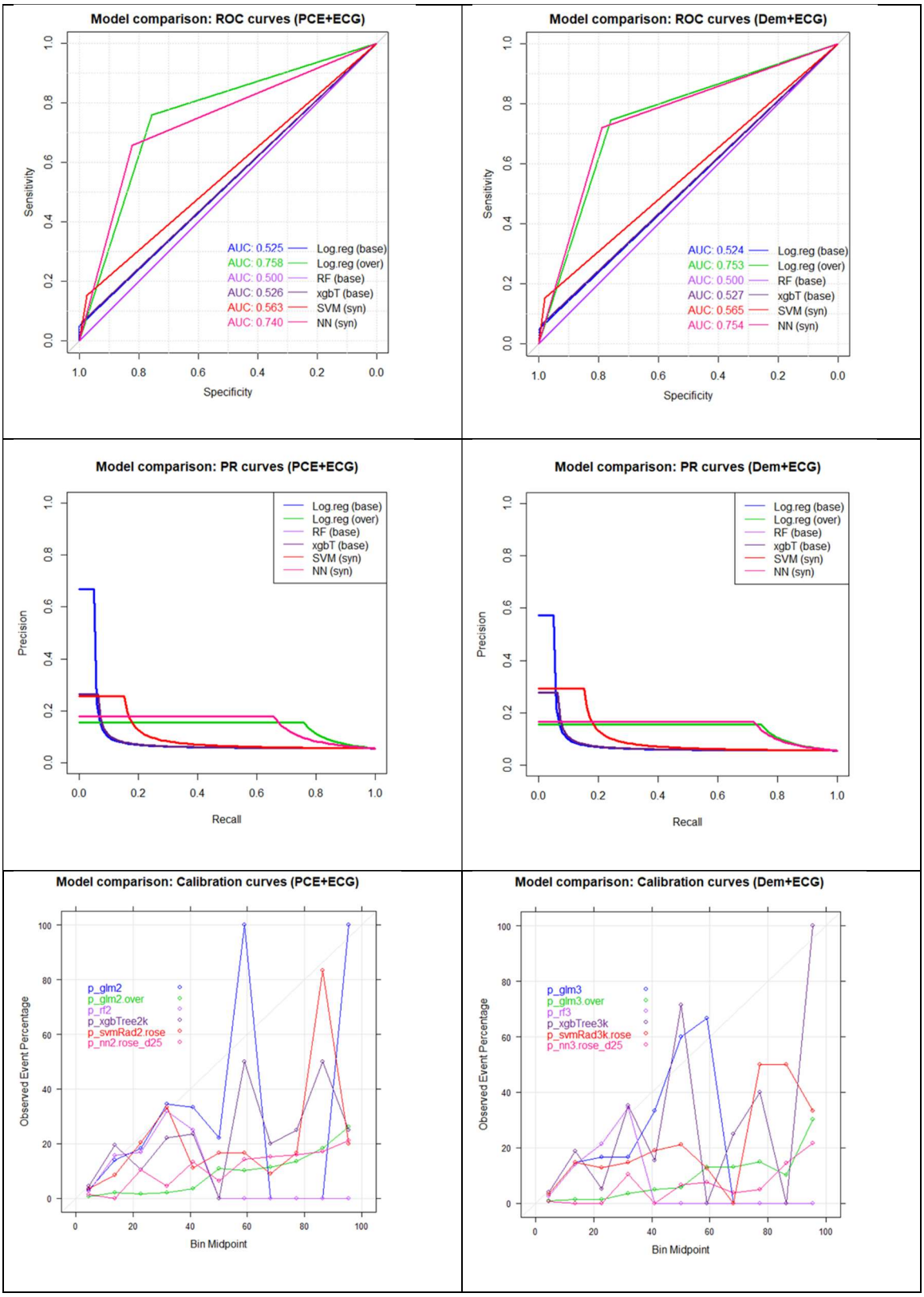

Figure 10. Classification models comparison plots

7. Variable importance comparison

Ranking of important predictor variables for individual models and the aggregate counts for the top ten predictors for each model are shown in **Table 11**. Models were selected for inclusion based on individual model performance as assessed in the preceding sections, and included the Cox PH model, L1-regularized Cox PH model, logistic regression models based on each of the training sets (base, oversampled, synthetic), a gradient boosting machine model (base), and a support vector machine model (synthetic). Aside from the Cox PH model which was based on only PCE variables, the rest of the models were based on both PCE and ECG data. Neural network models were not included for this analysis due to limitations in assessing the relative ranking of predictor variables.

Age was clearly the most important predictor of 10-year cardiovascular mortality, occurring in the top 10 list in all models. Next most important were systolic blood pressure and treatment for hypertension, which occurred in top 10 predictors in 6 out of 7 examined models. Other variables of the PCE, including race, total cholesterol, HDL cholesterol, and diabetes, were deemed less important, occurring only in the Cox PH model which was forced to rank all nine variables (with two levels for race).

Variable	Description	Count	PCE-Cox	L1-Cox	LR (base)	LR (over)	LR (syn)	GBM (base)(K)	SVM (syn)
HSSEX	Sex	2	4	3					
HSAGEIR	Age	7	7	5	1	1	1	1	1
DMARACER2	Race-Black	1	5						
DMARACER3	Race-Other	1	3						
Totalchol_adj	Total cholesterol	1	10						
HDLchol_adj	HDL cholesterol	1	9						
medianSBP	Systolic BP	6	8	7	10		2	3	2
HTN_tx1	Tx for Hypertension	6	2	2		8	4		7
Diabetes1	Hx of Diabetes	1	6						
Smoker_cur1	Current smoker	4	1	1	3	2			
ECPWIDTH	Chest half-width (mm)	1					9		
ECPDEPTH	Chest half-depth (mm)	2				10	10		
ECPRATE	Heart rate on ECG	1		8					
ECPP4	P amplitude, negative phase, lead V1	1					3		
ECPQA4	Q/QS amplitude, lead aVL	2			6	5			
ECPQA11	Q/QS amplitude, lead V6	1						4	
ECPQD1	Q/QS duration, lead I	1		6					
ECPRA2	R amplitude, lead II	2					5		5
ECPRA4	R amplitude, lead aVR	1						5	
ECPRA5	R amplitude, lead aVL	3			4	7		2	
ECPRD2	R duration, lead II	1			8				
ECPSA1	S amplitude, lead I	2			9	9			
ECPSA3	S amplitude, lead III	3			2	3		8	
ECPSA6	S amplitude, lead aVF	3			7	6		10	
ECPSA12	S amplitude, lead V6	1						9	
ECPSD1	S duration, lead I	1		9					
ECPSD12	S duration, lead V6	3			5	4	6		
ECPJ1	J amplitude, lead I	1							6
ECPJ4	J amplitude, lead aVR	1							10
ECPJ5	J amplitude, lead aVL	1		10					
ECPJ7	J amplitude, lead V1	1						6	
ECPJ10	J amplitude, lead V4	1							9
ECPJ11	J amplitude, lead V5	2					8		3
ECPJ12	J amplitude, lead V6	1							4
ECPPTA1	Positive T amplitude, lead I	1							8
ECPSINUSNS	Rhythm code	2		4			7		
ECPQTC	QT interval, corrected	1						7	

Table 11. Variable importance comparison
Variables are color-coded by the frequency of occurrence in the top ten important predictor variables as identified by each model. Blue: 7, green: 6, light green: 4, yellow: 3, orange: 2.

Among ECG variables, several occurred in the top 10 predictor list three or two times, and are highlighted in yellow and orange, respectively. Interestingly, when plotted on an ECG, the important ECG variables were seen to be clustering around inferior (II, III, aVF) and lateral (I, aVL, V5, V6) ECG leads (**Figure 11**).

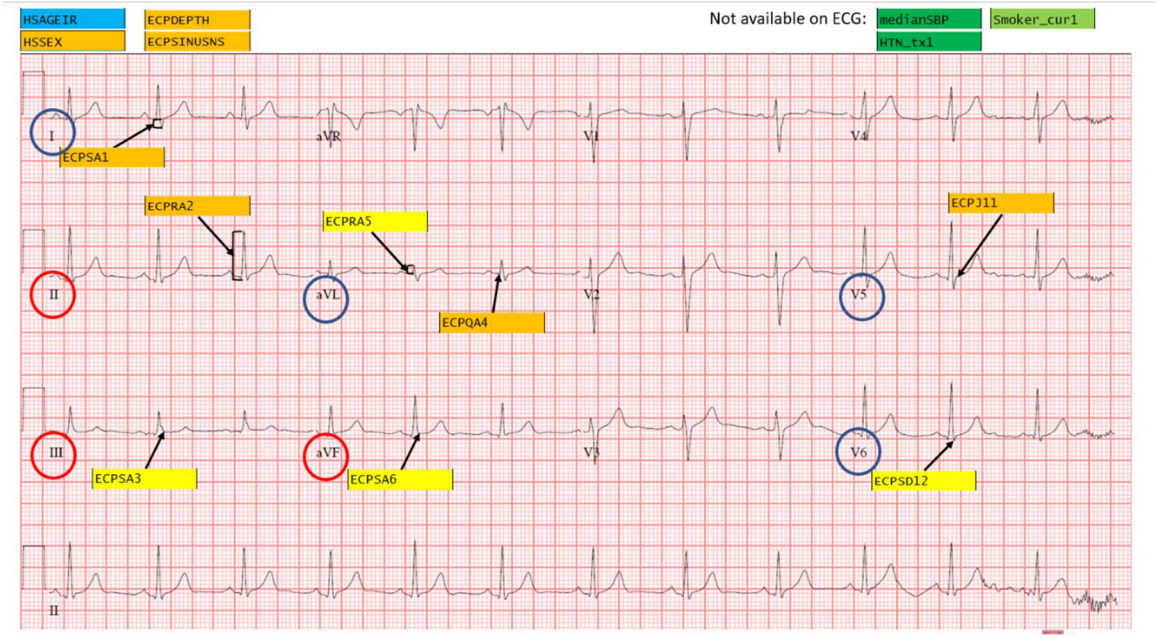

Figure 11. Important variables plotted on ECG
Variables are color-coded by the frequency of occurrence in the top ten important predictor variables as identified by each model. Blue: 7, green: 6, light green: 4, yellow: 3, orange: 2. Important ECG variables appear to be clustering around inferior (II, III, aVF) and lateral (I, aVL, V5, V6) ECG leads.

Chapter V.

DISCUSSION

1. Summary of important findings

The primary objective of this study was to assess whether ECG data could be used to predict cardiovascular mortality in the general population. This was motivated by the limitations of the current clinical standard, the 2013 PCE, which is known to be poorly calibrated in modern populations and does not reflect electrophysiologic risk factors that can lead to sudden cardiac death. Through a comparative analysis of traditional survival analysis and machine learning methods, this study demonstrated that 10-year cardiovascular mortality could be predicted from clinical and ECG data by machine learning models, with superior performance characteristics compared to the PCE.

Interestingly, the superior performance of machine learning models, particularly that of neural networks and ensemble models, could be achieved with just demographic and ECG data, without the need for other traditional cardiovascular risk factors represented in the PCE. While the PCE is intended to be used as a screening tool for primary prevention, it is rather cumbersome to implement in practice due to the number of data categories required for its computation. The PCE requires demographic (age, sex, race), historical (diabetes), physical (systolic blood pressure), social (smoking status), laboratory (total cholesterol, HDL cholesterol), and medication (treatment for hypertension) data elements, which usually require multiple visits to determine. In contrast, the ECG is a single-instance, non-invasive, and routinely available test in the outpatient

setting, and in the modern era with built-in measurements that allows for automated computation of cardiovascular risk based on prediction models such as those developed in this study. Thus, risk prediction models based on the ECG represent an attractive alternative to the PCE.

A comparative analysis of important variables between different prediction models shed insight into a possible underlying mechanism for cardiovascular risk determination in an ECG. The most frequently utilized ECG variables all localized to inferior (II, III, aVF) and lateral (I, aVL, V5, V6) leads, which are clinically meaningful in that they represent locations of "silent" myocardial infarctions, or minor myocardial infarctions that occur without any noticeable symptoms. Despite their lack of symptoms, the presence or history of these silent myocardial infarctions, as manifested in the ECG, would imply that the underlying risk factors for adverse cardiovascular events are already present, which would then lead to increased risk over the following 10 years. While it was initially hypothesized that risk for sudden cardiac death related to arrhythmia may be underrepresented in the PCE, it may in fact be the case that aggregated ECG measurements effectively capture underlying risk for ASCVD-related events instead.

In terms of performance improvement, training data augmentation via oversampling and synthetic data generation were clearly beneficial in improving positive event detection rate in most machine learning models. Among individual models, the neural network model based on demographic and ECG data with synthetic data augmentation had the best ROC-AUC and PR-AUC, though it was expectedly outperformed by the ensemble model based on the same training data. However, the suite of prediction models trained clearly displayed a wide spectrum of model characteristics, where different models may be suited for different clinical

scenarios. In the context of cardiovascular risk prediction for primary prevention, being able to detect future events (high sensitivity, high PR-AUC) would arguably be the most important, and this was achieved by the neural network and ensemble models described above.

The aggregated ECG components analysis for cardiovascular risk stratification is a unique approach taken for this study, which is in distinction compared to prior studies in medical literature. Most prior studies have examined individual ECG components (e.g. P wave duration[55], deep terminal negativity of P wave in V1 (DTNPV1) [56,57], QRS duration[58], QT interval[59-62], JT interval[62], and isolated ST-segment and T-wave abnormalities[63]) as independent and/or additive predictor(s) in the setting of standardized cardiovascular risk calculators such as the Framingham Risk Score and the PCE[14,15]. Interestingly, these previously identified individual ECG components did not feature prominently in the comparative analysis in this study, suggesting that ECG components in aggregate may capture additional cardiovascular risk beyond that available in individual ECG components. Other studies have explored the concept of global electrical heterogeneity, or abnormalities of the spatial ventricular gradient captured in 3-dimensions through an X- Y- Z- ECG lead system or by matrix transformation of the standard 12-lead ECG[64]. While global electrical heterogeneity has been shown to be a risk factor for adverse cardiovascular events such as sudden cardiac death[65], how it compares to an aggregated examination of scalar ECG measurements such as in this study requires further research.

2. Limitations

There are important limitations to this study. First, this was a retrospective study based on a single data source, with the usual limitations associated with such study design. Not all data components were available for all participants of NHANES III; in particular, ECG data were available in less than half of survey participants, who had greater comorbidity burden at baseline and thus may represent a different population than the target population of healthy, undifferentiated U.S. persons. Even where ECG data were available, there was a significant proportion of missing values, requiring removal of specific components and data imputation. This may be problematic if the data are not missing at random, which could not be ascertained and thus may have led to biased results. Next, available outcome events were limited to cardiovascular mortality, which is only part of the ultimate outcome of interest, that of both fatal and nonfatal adverse cardiovascular and cerebrovascular events. Methodologically, while some models were trained as survival models, other models were trained as classification models, which results in information loss and biased parameter estimates. Finally, many of the machine learning models overfitted training data and were not effective for event prediction in test data, only some of which may have improved with further optimization.

Despite these limitations, the methodological framework pursued in this study, that of comparative analysis among many different types of prediction models with less focus on individual model parameters but with greater emphasis on aggregate findings and empiric performance, led to superior prediction performance compared to the current standard and also allowed for important insight into the underlying mechanisms of risk stratification using aggregated ECG data. There

were many novel insights which are hypothesis generating for future work, as outlined below.

3. Future research

Findings of this study would greatly benefit from a validation study in a separate dataset, ideally in a nationwide sample with ECG data and all MACE outcomes over a long follow-up period. Specific patient subgroups (e.g. those with diabetes, obesity, or other risk-enhancing features as outlined in the 2019 ASCVD guidelines[4]) may be of particular interest to assess whether ECG data can improve cardiovascular risk prediction in these populations. Regarding ECG data, while measurements from a single ECG were used for this analysis, a series of ECGs obtained over time may provide additional insight into the nature of the predicted risk and how it changes over time. Beyond numeric ECG measurements, a direct visual pattern assessment based on advanced machine learning methods for image processing (e.g. convolutional neural networks) or computational phenotyping approaches may allow for extraction of additional risk stratifying information from a given ECG. Finally, the wealth of longitudinal and real-world data stored in EHRs across the U.S. could be leveraged in a similar methodological framework to complement the findings of this study.

Chapter VI.

CONCLUSIONS

Machine learning models trained on demographic and aggregated ECG data were superior to the PCE for prediction of 10-year cardiovascular mortality in a nationwide sample. Prediction models based on automated ECG measurements could be useful for routine screening for primary prevention due to their superior performance and ease of testing and use. Comparison between multiple models provided insight into important predictors and possible underlying mechanisms for cardiovascular risk estimation.

REFERENCES

1. Benjamin EJ, Blaha MJ, Chiuve SE, et al. Heart Disease and Stroke Statistics-2017 Update: A Report From the American Heart Association. *Circulation.* 2017;135(10):e146-e603.

2. Cardiovascular Disease: A Costly Burden For America: Projections Through 2035. *American Heart Association CVD Burden Report.* 2017.

3. Goff DC, Jr., Lloyd-Jones DM, Bennett G, 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):2935-2959.

4. Arnett DK, Blumenthal RS, Albert MA, 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. *J Am Coll Cardiol.* 2019.

5. Ridker PM, Cook NR. The Pooled Cohort Equations 3 Years On: Building a Stronger Foundation. *Circulation.* 2016;134(23):1789-1791.

6. Yadlowsky S, Hayward RA, Sussman JB, McClelland RL, Min YI, Basu S. Clinical Implications of Revised Pooled Cohort Equations for Estimating Atherosclerotic Cardiovascular Disease Risk. *Ann Intern Med.* 2018;169(1):20-29.

7. Crowson CS, Gabriel SE, Semb AG, et al. Rheumatoid arthritis-specific cardiovascular risk scores are not superior to general risk scores: a validation analysis of patients from seven countries. *Rheumatology (Oxford).* 2017;56(7):1102-1110.

8. Triant VA, Perez J, Regan S, et al. Cardiovascular Risk Prediction Functions Underestimate Risk in HIV Infection. *Circulation.* 2018;137(21):2203-2214.

9. Dalton JE, Perzynski AT, Zidar DA, et al. Accuracy of Cardiovascular Risk Prediction Varies by Neighborhood Socioeconomic Position: A Retrospective Cohort Study. *Ann Intern Med.* 2017;167(7):456-464.

10. Colantonio LD, Richman JS, Carson AP, et al. Performance of the Atherosclerotic Cardiovascular Disease Pooled Cohort Risk Equations by Social Deprivation Status. *J Am Heart Assoc.* 2017;6(3).

11. Messerli FH, Hofstetter L, Rimoldi SF, Rexhaj E, Bangalore S. Risk Factor Variability and Cardiovascular Outcome: JACC Review Topic of the Week. *J Am Coll Cardiol.* 2019;73(20):2596-2603.

12. Muntner P, Colantonio LD, Cushman M, et al. Validation of the atherosclerotic cardiovascular disease Pooled Cohort risk equations. *Jama.* 2014;311(14):1406-1415.

13. Lloyd-Jones DM, Huffman MD, Karmali KN, et al. Estimating Longitudinal Risks and Benefits From Cardiovascular Preventive Therapies Among Medicare Patients: The Million Hearts Longitudinal ASCVD Risk Assessment Tool: A Special Report From the American Heart Association and American College of Cardiology. *J Am Coll Cardiol.* 2017;69(12):1617-1636.

14. Badheka AO, Patel N, Tuliani TA, et al. Electrocardiographic abnormalities and reclassification of cardiovascular risk: insights from NHANES-III. *Am J Med.* 2013;126(4):319-326.e312.

15. Shah AJ, Vaccarino V, Janssens AC, et al. An Electrocardiogram-Based Risk Equation for Incident Cardiovascular Disease From the National Health and Nutrition Examination Survey. *JAMA Cardiol.* 2016;1(7):779-786.

16. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *J Am Med Inform Assoc.* 2018;25(10):1419-1428.

17. Shickel B, Tighe P, Bihorac A, Rashidi P. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *arXiv e-prints.* 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170603446S. Accessed June 01, 2017.

18. Shah SJ, Katz DH, Selvaraj S, et al. Phenomapping for novel classification of heart failure with preserved ejection fraction. *Circulation.* 2015;131(3):269-279.

19. Seymour CW, Kennedy JN, Wang S, et al. Derivation, Validation, and Potential Treatment Implications of Novel Clinical Phenotypes for Sepsis. *Jama.* 2019;321(20):2003-2017.

20. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *J Am Med Inform Assoc.* 2016;23(e1):e20-27.

21. Dubois SR, N.; Jung, K.; Shah N.; Kale, D.C. The Effectiveness of Transfer Learning in Electronic Health Records Data. Paper presented at: ICLR 20172017.

22. Pham T, Tran T, Phung D, Venkatesh S. DeepCare: A Deep Dynamic Memory Model for Predictive Medicine. *arXiv e-prints.* 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv160200357P. Accessed January 01, 2016.

23. Choi E, Taha Bahadori M, Schuetz A, Stewart WF, Sun J. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. *arXiv e-prints.*

2015. https://ui.adsabs.harvard.edu/abs/2015arXiv151105942C. Accessed November 01, 2015.

24. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc.* 2017;24(2):361-370.

25. Bai TZ, S.; Egleston, B.L.; Vucetic, S. Interpretable Representation Learning for Healthcare via Capturing Disease Progression through Time. Paper presented at: KDD 20182018.

26. Fotso S. Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework. *arXiv e-prints.* 2018. https://ui.adsabs.harvard.edu/abs/2018arXiv180105512F. Accessed January 01, 2018.

27. Luck M, Sylvain T, Cardinal H, Lodi A, Bengio Y. Deep Learning for Patient-Specific Kidney Graft Survival Analysis. *arXiv e-prints.* 2017. https://ui.adsabs.harvard.edu/abs/2017arXiv170510245L. Accessed May 01, 2017.

28. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology.* 2018;18(1):24.

29. Giunchiglia EN, A.; van der Schaar, M. RNN-SURV: a Deep Recurrent Model for Survival Analysis. Paper presented at: ICANN 20182018.

30. Lee CZ, W.R.; Yoon, J.; van der Schaar, M. DeepHit: A Deep Learning Approach to Survival Analysis with Competing Risks. Paper presented at: AAAI 20182018.

31. Baytas IM, Xiao C, Zhang X, Wang F, Jain AK, Zhou J. Patient Subtyping via Time-Aware LSTM Networks. Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; 2017; Halifax, NS, Canada.

32. Che C, Xiao C, Liang J, Jin B, Zho J, Wang F. An RNN Architecture with Dynamic Temporal Matching for Personalized Predictions of Parkinson's Disease. Paper presented at: Proceedings of the 2017 SIAM International Conference on Data Mining2017.

33. Rajkomar A, Oren E, Chen K, et al. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine.* 2018;1(1):18.

34. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part IV: further concepts and methods in survival analysis. *Br J Cancer.* 2003;89(5):781-786.

35. Goodfellow IB, Y.; Courville, A. *Deep Learning*. MIT Press; 2016.

36. Miotto R, Li L, Kidd BA, Dudley JT. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports.* 2016;6:26094.

37. Clark TG, Bradburn MJ, Love SB, Altman DG. Survival analysis part I: basic concepts and first analyses. *Br J Cancer.* 2003;89(2):232-238.

38. Wang PL, Y.; Reddy, C.K. Machine Learning for Survival Analysis: A Survey. *ACM Computing Surveys.* 2019;51(6).

39. Bradburn MJ, Clark TG, Love SB, Altman DG. Survival analysis part II: multivariate data analysis--an introduction to concepts and methods. *Br J Cancer.* 2003;89(3):431-436.

40. Bou-Hamad IL, D.; Ben-Ameur, H. A review of survival trees. *Statistics Surveys.* 2011;5:44-71.

41. Hothorn T, Lausen B, Benner A, Radespiel-Troger M. Bagging survival trees. *Stat Med.* 2004;23(1):77-91.

42. Ishwaran HK, U.B.; Blackstone, E.H.; Lauer, M.S. Random survival forests. *The Annals of Applied Statistics.* 2008;2(3):841-860.

43. Hothorn T, Buhlmann P, Dudoit S, Molinaro A, van der Laan MJ. Survival ensembles. *Biostatistics.* 2006;7(3):355-373.

44. Caruana RL, Y.; Gehrke, J.; Koch, P.; Sturm, M.; Elhadad, N. Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission. Paper presented at: KDD 2015 Proceedings of the 21th ACM SIGKDD International Conference on Knolwedge Discovery and Data Mining2015.

45. Tsujitani M, Tanaka Y, Sakon M. Survival data analysis with time-dependent covariates using generalized additive models. *Comput Math Methods Med.* 2012;2012:986176.

46. Shivaswamy PKC, W.; Jansche, M. A Support Vector Approach to Censored Targets. Paper presented at: IDCM 2008 Proceedings of the 2007 Seventh IEEE International Conference on Data Mining2007.

47. Yu CG, R.; Lin, H.; Baracos, V. Learning Patient-Specific Cancer Survival Distributions as a Sequence of Dependent Regressors. Paper presented at: NIPS 20112011.

48. Steyerberg EW, Vickers AJ, Cook NR, et al. Assessing the performance of prediction models: a framework for traditional and novel measures. *Epidemiology.* 2010;21(1):128-138.

49. Alba AC, Agoritsas T, Walsh M, et al. Discrimination and Calibration of Clinical Prediction Models: Users' Guides to the Medical Literature. *Jama.* 2017;318(14):1377-1384.

50. Rudin C. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *ArXiv e-prints.* 2019.

51. Pearl J. *Causality: Models, Reasoning, and Inference.* 2nd ed: Cambridge University Press; 2009.

52. Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM.* 2019;62(3):54-60.

53. Choi E, Taha Bahadori M, Song L, Stewart WF, Sun J. GRAM: Graph-based Attention Model for Healthcare Representation Learning. *arXiv e-prints.* 2016. https://ui.adsabs.harvard.edu/abs/2016arXiv161107012C. Accessed November 01, 2016.

54. Yeboah J, Young R, McClelland RL, et al. Utility of Nontraditional Risk Markers in Atherosclerotic Cardiovascular Disease Risk Assessment. *J Am Coll Cardiol.* 2016;67(2):139-147.

55. Magnani JW, Gorodeski EZ, Johnson VM, et al. P wave duration is associated with cardiovascular and all-cause mortality outcomes: the National Health and Nutrition Examination Survey. *Heart Rhythm.* 2011;8(1):93-100.

56. Tereshchenko LG, Shah AJ, Li Y, Soliman EZ. Electrocardiographic deep terminal negativity of the P wave in V1 and risk of mortality: the National Health and Nutrition Examination Survey III. *J Cardiovasc Electrophysiol.* 2014;25(11):1242-1248.

57. Tereshchenko LG, Henrikson CA, Sotoodehnia N, et al. Electrocardiographic deep terminal negativity of the P wave in V(1) and risk of sudden cardiac death: the Atherosclerosis Risk in Communities (ARIC) study. *J Am Heart Assoc.* 2014;3(6):e001387.

58. Badheka AO, Singh V, Patel NJ, et al. QRS duration on electrocardiography and cardiovascular mortality (from the National Health and Nutrition Examination Survey-III). *Am J Cardiol.* 2013;112(5):671-677.

59. Zhang Y, Post WS, Dalal D, Blasco-Colmenares E, Tomaselli GF, Guallar E. QT-interval duration and mortality rate: results from the Third National Health and Nutrition Examination Survey. *Arch Intern Med.* 2011;171(19):1727-1733.

60. Malik R, Waheed S, Parashara D, Perez J, Waheed S. Association of QT interval with mortality by kidney function: results from the National Health and Nutrition Examination Survey (NHANES). *Open Heart.* 2017;4(2):e000683.

61. Waheed S, Dawn B, Gupta K. Association of corrected QT interval with body mass index, and the impact of this association on mortality: Results from the Third National Health and Nutrition Examination Survey. *Obes Res Clin Pract.* 2017;11(4):426-434.

62. Zulqarnain MA, Qureshi WT, O'Neal WT, Shah AJ, Soliman EZ. Risk of Mortality Associated With QT and JT Intervals at Different Levels of QRS Duration (from the Third National Health and Nutrition Examination Survey). *Am J Cardiol.* 2015;116(1):74-78.

63. Badheka AO, Rathod A, Marzouka GR, et al. Isolated nonspecific ST-segment and T-wave abnormalities in a cross-sectional United States population and Mortality (from NHANES III). *Am J Cardiol.* 2012;110(4):521-525.

64. Waks JW, Tereshchenko LG. Global electrical heterogeneity: A review of the spatial ventricular gradient. *Journal of electrocardiology.* 2016;49(6):824-830.

65. Waks JW, Sitlani CM, Soliman EZ, et al. Global Electric Heterogeneity Risk Score for Prediction of Sudden Cardiac Death in the General Population: The Atherosclerosis Risk in Communities (ARIC) and Cardiovascular Health (CHS) Studies. *Circulation.* 2016;133(23):2222-2234.

66. Ambale-Venkatesh B, Yang X, Wu CO, et al. Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis. *Circ Res.* 2017;121(9):1092-1101.

67. Kakadiaris IA, Vrigkas M, Yen AA, Kuznetsova T, Budoff M, Naghavi M. Machine Learning Outperforms ACC / AHA CVD Risk Calculator in MESA. *J Am Heart Assoc.* 2018;7(22):e009476.

68. Kennedy EH, Wiitala WL, Hayward RA, Sussman JB. Improved cardiovascular risk prediction using nonparametric regression and electronic health record data. *Med Care.* 2013;51(3):251-258.

69. Dogan MV, Beach SRH, Simons RL, Lendasse A, Penaluna B, Philibert RA. Blood-Based Biomarkers for Predicting the Risk for Five-Year Incident Coronary Heart Disease in the Framingham Heart Study via Machine Learning. *Genes (Basel).* 2018;9(12).

70. Weng SF, Reps J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLoS One.* 2017;12(4):e0174944.

71. NHANES III (1988-1994). Centers for Disease Control and Prevention. https://wwwn.cdc.gov/nchs/nhanes/nhanes3/Default.aspx. Accessed 08/29/2019.

72. R: A language and environment for statistical computing. In. https://www.R-project.org/: R Foundation for Statistical Computing, Vienna, Austria. ; 2019.

73. RStudio: Integrated Development for R. In. RStudio, Inc. Boston, MA http://www.rstudio.com2015.

74. van Buuren S, Groothuis-Oudshoorn K, Vink G, al. e. mice: Multivariate Imputation by Chained Equations. In: https://cran.r-project.org/web/packages/mice/index.html; 2020.

75. Lunardon N, Menardi G, Torelli N. ROSE: Random Over-Sampling Examples. In. https://cran.r-project.org/web/packages/ROSE/ROSE.pdf2014.

76. Therneau TML, T.; Atkinson, E.; Crowson, C. survival: Survival Analysis. In. https://cran.r-project.org/web/packages/survival/index.html2020.

77. Kassambara AK, M.; Biecek, P.; Fabian, S. survminer: Drawing Survival Curves using 'ggplot2'. In. https://cran.r-project.org/web/packages/survminer/index.html2020.

78. Friedman JH, T.; Tibshirani, R.; Narasimhan, B.; Simon, N.; Qian, J. glmnet: Lasso and Elastic-Net Regularized Generalized Linear Models. In. https://cran.r-project.org/web/packages/glmnet/index.html2020.

79. Ishwaran HK, U.B. randomForestSRC: Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC). In. https://cran.r-project.org/web/packages/randomForestSRC/index.html2020.

80. Kuhn MW, J.; Weston, S.; Williams, A.; Keefer, C.; Engelhardt, A.; Cooper, T.; Mayer, Z.; Kenkel, B.; Benesty, M.; Lescarbeau, R.; Ziem, A.; Scrucca, L.; Tang, Y.; Candan, C.; Hunt, T. caret: Classification and Regression Training. In: https://cran.r-project.org/web/packages/caret/caret.pdf; 2020.

81. Falbel DA, J.J.; Chollet, F.; Tang, Y.; van der Bifl, W.; Struder, M.; Keydana, S. keras: R Interface to 'Keras'. In. https://cran.r-project.org/web/packages/keras/index.html2020.

82. Polley E. SuperLearner: Super Learner Prediction. In. https://cran.r-project.org/web/packages/SuperLearner/SuperLearner.pdf2019.

APPENDIX A.

ECG variables used for analysis (adapted from NHANES III Electrocardiography Data File Index)

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
DEMOGRAPHIC DATA		
Sample person identification number	SEQN	1-5
NHANES III Survey (1988-94)	ECPSNUM	6
Sex	HSSEX	7
Race	DMARACER	8
Age at interview (Screener) in years	HSAGEIR	9-10
Pseudo-PSU	SDPPSU	11
Pseudo-stratum	SDPSTRA	12-13
MEC-examined sample final weight	WTPFEX	14-22
INTRODUCTORY INFORMATION		
Technician number	ECPTECH1	23-27
Number of leads	ECPLEADS	28-29
Chest half-width (mm)	ECPWIDTH	30-32
Chest half-depth (mm)	ECPDEPTH	33-35
Major ECG abnormalities	ECPG1	36
Minor ECG abnormalities	ECPG2	37
Probable myocardial infarction (MI)	ECPG3	38
Possible MI	ECPG4	39
Probable left ventricular hypertrophy	ECPG5	40
Possible LVH by MC	ECPG6	41
MINNESOTA CODES		
MC 1 Leadgroup L(I, aVL, V6)	ECPL1	42-43
MC 1 Leadgroup F(II, III, aVF)	ECPF1	44-45
MC 1 Leadgroup V(V1-V5)	ECPV1	46-47
MC 4 Leadgroup L	ECPL4	48-49
MC 4 Leadgroup F	ECPF4	50-51
MC 4 Leadgroup V	ECPV4	52-53
MC 5 Leadgroup L	ECPL5	54
MC 5 Leadgroup F	ECPF5	55
MC 5 Leadgroup V	ECPV5	56
MC 9.2 Leadgroup L	ECPL9	57
MC 9.2 Leadgroup F	ECPF9	58
MC 9.2 Leadgroup V	ECPV9	59

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
MC 2 (QRS axis code)	ECPMC2	60-61
MC 3 (High-amplitude R waves)	ECPMC3	62-63
MC 6 (A-V conduction)	ECPMC6	64-65
MC 7 (Ventricular conduction)	ECPMC7	66
MC 9.1 (Low-amplitude QRS)	ECPMC91	67
MC 9.3 (High-amplitude P)	ECPMC93	68
MC 9.4 (QRS transition zone)	ECPMC94	69
MC 9.5 (High-amplitude T)	ECPMC95	70
CARDIAC/INFARCTION INJURY SCORE		
Cardiac infarction score (12-lead by 10)	ECPCIIS	71-73
Probable infarction/injury	ECPCIIS2	74
Possible infarction/injury	ECPCIIS3	75
Consider infarction/injury	ECPCIIS4	76
LEFT VENTRICULAR MASS		
ECG estimate of LV mass	ECPLVM	77-79
ECG estimate of LV mass index	ECPLVMI	80-82
Probable LVH	ECPLVM3	83
HEART RATE, BASIC ECG INTERVALS, AND MEAN AXIS DATA		
Heart rate (beats per minute)	ECPRATE	84-86
PR interval (msec)	ECPPR	87-89
QRS interval (msec)	ECPQRS	90-92
QT interval (msec)	ECPQT	93-95
P axis, frontal plane (degrees)	ECPAXIS1	96-99
QRS axis, frontal plane (degrees)	ECPAXIS2	100-103
T axis, frontal plane (degrees)	ECPAXIS3	104-107
Rhythm code	ECPBEAT	108
ECG WAVE MEASUREMENTS		
P amplitude, positive phase, lead II(uV)	ECPP1	109-111

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
P duration, lead II (msec)	ECPP2	112-114
P amplitude, positive phase, lead V1(uV)	ECPP3	115-117
P amplitude, negative phase, lead V1(uV)	ECPP4	118-121
Q or QS amplitude, lead I (uV)	ECPQA1	122-125
Q or QS amplitude, lead II (uV)	ECPQA2	126-129
Q or QS amplitude, lead III (uV)	ECPQA3	130-133
Q or QS amplitude, lead aVL (uV)	ECPQA4	134-137
Q or QS amplitude, lead AVF (uV)	ECPQA5	138-141
Q or QS amplitude, lead V1 (uV)	ECPQA6	142-145
Q or QS amplitude, lead V2 (uV)	ECPQA7	146-149
Q or QS amplitude, lead V3 (uV)	ECPQA8	150-153
Q or QS amplitude, lead V4 (uV)	ECPQA9	154-157
Q or QS amplitude, lead V5 (uV)	ECPQA10	158-161
Q or QS amplitude, lead V6 (uV)	ECPQA11	162-165
Q or QS duration, lead I (msec)	ECPQD1	166-168
Q or QS duration, lead II (msec)	ECPQD2	169-171
Q or QS duration, lead III (msec)	ECPQD3	172-174
Q or QS duration, lead aVL (msec)	ECPQD4	175-177
Q or QS duration, lead aVF (msec)	ECPQD5	178-180
Q or QS duration, lead V1 (msec)	ECPQD6	181-183
Q or QS duration, lead V2 (msec)	ECPQD7	184-186
Q or QS duration, lead V3 (msec)	ECPQD8	187-189
Q or QS duration, lead V4 (msec)	ECPQD9	190-192
Q or QS duration, lead V5 (msec)	ECPQD10	193-195
Q or QS duration, lead V6 (msec)	ECPQD11	196-198
R amplitude, lead I (uV)	ECPRA1	199-202
R amplitude, lead II (uV)	ECPRA2	203-206
R amplitude, lead III (uV)	ECPRA3	207-210
R amplitude, lead aVR (uV)	ECPRA4	211-214
R amplitude, lead aVL (uV)	ECPRA5	215-218
R amplitude, lead aVF (uV)	ECPRA6	219-222
R amplitude, lead V1 (uV)	ECPRA7	223-226
R amplitude, lead V2 (uV)	ECPRA8	227-230
R amplitude, lead V3 (uV)	ECPRA9	231-234
R amplitude, lead V4 (uV)	ECPRA10	235-238
R amplitude, lead V5 (uV)	ECPRA11	239-242
R amplitude, lead V6 (uV)	ECPRA12	243-246
R duration, lead I (msec)	ECPRD1	247-249
R duration, lead II (msec)	ECPRD2	250-252
R duration, lead III (msec)	ECPRD3	253-255

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
R duration, lead aVR (msec)	ECPRD4	256-258
R duration, lead aVL (msec)	ECPRD5	259-261
R duration, lead aVF (msec)	ECPRD6	262-264
R duration, lead V1 (msec)	ECPRD7	265-267
R duration, lead V2 (msec)	ECPRD8	268-270
R duration, lead V3 (msec)	ECPRD9	271-273
R duration, lead V4 (msec)	ECPRD10	274-276
R duration, lead V5 (msec)	ECPRD11	277-279
R duration, lead V6 (msec)	ECPRD12	280-282
S amplitude, lead I (uV)	ECPSA1	283-286
S amplitude, lead II (uV)	ECPSA2	287-290
S amplitude, lead III (uV)	ECPSA3	291-294
S amplitude, lead aVR (uV)	ECPSA4	295-298
S amplitude, lead aVL (uV)	ECPSA5	299-302
S amplitude, lead aVF (uV)	ECPSA6	303-306
S amplitude, lead V1 (uV)	ECPSA7	307-310
S amplitude, lead V2 (uV)	ECPSA8	311-314
S amplitude, lead V3 (uV)	ECPSA9	315-318
S amplitude, lead V4 (uV)	ECPSA10	319-322
S amplitude, lead V5 (uV)	ECPSA11	323-326
S amplitude, lead V6 (uV)	ECPSA12	327-330
S duration, lead I (msec)	ECPSD1	331-333
S duration, lead II (msec)	ECPSD2	334-336
S duration, lead III (msec)	ECPSD3	337-339
S duration, lead aVR (msec)	ECPSD4	340-342
S duration, lead aVL (msec)	ECPSD5	343-345
S duration, lead aVF (msec)	ECPSD6	346-348
S duration, lead V1 (msec)	ECPSD7	349-351
S duration, lead V2 (msec)	ECPSD8	352-354
S duration, lead V3 (msec)	ECPSD9	355-357
S duration, lead V4 (msec)	ECPSD10	358-360
S duration, lead V5 (msec)	ECPSD11	361-363
S duration, lead V6 (msec)	ECPSD12	364-366
R' amplitude, lead I (uV)	ECPRPA1	367-370
R' amplitude, lead II (uV)	ECPRPA2	371-374
R' amplitude, lead III (uV)	ECPRPA3	375-378
R' amplitude, lead aVR (uV)	ECPRPA4	379-382
R' amplitude, lead aVL (uV)	ECPRPA5	383-386
R' amplitude, lead aVF (uV)	ECPRPA6	387-390
R' amplitude, lead V1 (uV)	ECPRPA7	391-394

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
R' amplitude, lead V2 (uV)	ECPRPA8	395-398
R' amplitude, lead V3 (uV)	ECPRPA9	399-402
R' amplitude, lead V4 (uV)	ECPRPA10	403-406
R' amplitude, lead V5 (uV)	ECPRPA11	407-410
R' amplitude, lead V6 (uV)	ECPRPA12	411-414
J amplitude, lead I (uV)	ECPJ1	415-418
J amplitude, lead II (uV)	ECPJ2	419-422
J amplitude, lead III (uV)	ECPJ3	423-426
J amplitude, lead aVR (uV)	ECPJ4	427-430
J amplitude, lead aVL (uV)	ECPJ5	431-434
J amplitude, lead aVF (uV)	ECPJ6	435-438
J amplitude, lead V1 (uV)	ECPJ7	439-442
J amplitude, lead V2 (uV)	ECPJ8	443-446
J amplitude, lead V3 (uV)	ECPJ9	447-450
J amplitude, lead V4 (uV)	ECPJ10	451-454
J amplitude, lead V5 (uV)	ECPJ11	455-458
J amplitude, lead V6 (uV)	ECPJ12	459-462
Negative T amplitude, lead I (uV)	ECPNTA1	463-467
Negative T amplitude, lead II (uV)	ECPNTA2	468-471
Negative T amplitude, lead III (uV)	ECPNTA3	472-475
Negative T amplitude, lead aVR (uV)	ECPNTA4	476-479
Negative T amplitude, lead aVL (uV)	ECPNTA5	480-483
Negative T amplitude, lead aVF (uV)	ECPNTA6	484-487
Negative T amplitude, lead V1 (uV)	ECPNTA7	488-491
Negative T amplitude, lead V2 (uV)	ECPNTA8	492-496
Negative T amplitude, lead V3 (uV)	ECPNTA9	497-501
Negative T amplitude, lead V4 (uV)	ECPNTA10	502-506
Negative T amplitude, lead V5 (uV)	ECPNTA11	507-511
Negative T amplitude, lead V6 (uV)	ECPNTA12	512-516
Positive T amplitude, lead I (uV)	ECPPTA1	517-520
Positive T amplitude, lead II (uV)	ECPPTA2	521-524
Positive T amplitude, lead III (uV)	ECPPTA3	525-528
Positive T amplitude, lead aVR (uV)	ECPPTA4	529-532
Positive T amplitude, lead aVL (uV)	ECPPTA5	533-536
Positive T amplitude, lead aVF (uV)	ECPPTA6	537-540
Positive T amplitude, lead V1 (uV)	ECPPTA7	541-544
Positive T amplitude, lead V2 (uV)	ECPPTA8	545-548
Positive T amplitude, lead V3 (uV)	ECPPTA9	549-552
Positive T amplitude, lead V4 (uV)	ECPPTA10	553-556
Positive T amplitude, lead V5 (uV)	ECPPTA11	557-560

NHANES III Electrocardiography Data File Index

Description	Variable Name	Positions
Positive T amplitude, lead V6 (uV)	ECPPTA12	561-564

NHANES III Electrocardiography Data File

FILENAME=NH3ECG VERSION 1.0 N=8,561

DEMOGRAPHIC DATA

Positions SAS name	Counts	Item description and code		Notes
1-5		Sample person identification number		
SEQN	8561	00009-53616		
6		NHANES III Survey (1988-94)		
ECPSNUM	8561	3	NHANES III	
7		Sex		
HSSEX	4155	1	Male	
	4406	2	Female	
8		Race		
DMARACER	6286	1	White	
	2041	2	Black	
	234	3	Other	
9-10		Age at interview (Screener) in years		
HSAGEIR	8494	40-89		
	67	90	90+	
11		Pseudo-PSU		
SDPPSU	8561	1-2		
12-13		Pseudo-stratum		
SDPSTRA	8561	01-49		
14-22		MEC-examined sample final weight		
WTPFEX	8561	000227.87-129993.17		

APPENDIX B.

Abbreviations

ACC: American College of Cardiology

AHA: American Heart Association

ASCVD: Atherosclerotic cardiovascular disease

BMI: Body mass index

Cox PH model: Cox proportional hazards model

ECG: Electrocardiogram

EHR: Electronic health records

HDL: High-density lipoprotein

MACE: Major adverse cardiovascular events

NHANES: National Health and Nutrition Examination Survey

PCE: Pooled Cohort Equations

ROC curve: Receiver operating characteristic curve

ROC-AUC: Receiver operating characteristic - area under curve

PR curve: Precision-recall curve

PR-AUC: Precision-recall - area under curve