

Machine Learning for Sudden Cardiac Death Prediction in the Atherosclerosis Risk in Communities Study

Zhi Yu^{1,2,3}, Shannon Wongvibulsin⁴, Natalie R. Daya³, Linda Zhou³, Kunihiro Matsushita^{3,4}, Pradeep Natarajan^{1,2,5}, Josef Coresh^{3,4}, Scott L. Zeger^{3,4,6}

1. Program in Medical and Population Genetics and Cardiovascular Disease Initiative, Broad Institute of Harvard and MIT, Cambridge, MA

2. Cardiovascular Research Center, Massachusetts General Hospital, Boston, MA

3. Johns Hopkins Bloomberg School of Public Health, Baltimore, MD

4. Johns Hopkins School of Medicine, Baltimore, MD

5. Department of Medicine, Harvard Medical School, Boston, MA

6. Johns Hopkins Krieger School of Arts and Sciences, Baltimore, MD

Abstract word count: 271

Main text work count: 3,491

Figures: 2

Tables: 2

References: 44

Corresponding Authors:

Scott L. Zeger, PhD

Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health

615 N. Wolfe Street, Room E3650

Baltimore MD 21205

Phone: 410-502-9054

e-mail: sz@jhu.edu

Zhi Yu, MB, PhD

Broad Institute of MIT and Harvard

75 Ames Steet

Cambridge, MA 02142

Phone: 617-714-7398

Email: zyu@broadinstitute.org

Abstract

Introduction

Sudden cardiac death (SCD) is a devastating consequence often without antecedent expectation. Current risk stratification methods derived from baseline independently modeled risk factors are insufficient. Novel random forest machine learning (ML) approach incorporating time-dependent variables and complex interactions may improve SCD risk prediction.

Methods

Atherosclerosis Risk in Communities (ARIC) study participants were followed for adjudicated SCD. ML models were compared to standard Poisson regression models for interval data, an approximation to Cox regression, with stepwise variable selection. Eighty-two time-varying variables (demographics, lifestyle factors, clinical characteristics, biomarkers, etc.) collected at four visits over 12 years (1987-98) were used as candidate predictors. Predictive accuracy was assessed by area under the receiver operating characteristic curve (AUC) through out-of-bag prediction for ML models and 5-fold cross validation for the Poisson regression models.

Results

Over a median follow-up time of 23.5 years, 583 SCD events occurred among 15,661 ARIC participants (mean age 54 years and 55% women). Compared to different Poisson regression models (AUC at 6-year ranges from 0.77-0.83), the ML model improved prediction (AUC at 6-year 0.89). Top predictors identified by ML model included prior coronary heart disease, which explained 47.9% of the total phenotypic variance, diabetes mellitus, hypertension, and T wave abnormality in any of leads I, aVL, or V6. Using the top ML predictors to select variables,

the Poisson regression model AUC at 6-year was 0.77 suggesting that the non-linear dependencies and interactions captured by ML, are the main reasons for its improved prediction performance.

Conclusions

Applying novel ML approach with time-varying predictors improves the prediction of SCD. Interactions of dynamic clinical characteristics are important for risk-stratifying SCD in the general population.

Introduction

Sudden cardiac death (SCD) accounts for approximately 400,000 adult deaths in the United States each year.¹⁻³ Risk stratification for SCD continues to lag contributing to the significant public health burden.⁴ Current guidelines directing the use of primary prevention, implanted cardioverter defibrillator (ICD), largely rely on a single parameter: reduced left ventricular ejection fraction (LVEF). Using LVEF alone is inadequate for two main reasons: (1) it does not account for dynamic factors including interim clinical events;⁵ and (2) it is seldom measured for approximately 50% of SCD victims who did not have a prior diagnosis of heart disease.^{1,4} Previous efforts to expand the list of SCD predictors examined single or small numbers of static predictors, or were conducted among patients with known existing cardiovascular disease.⁶⁻¹¹

By combining data from a general population cohort with dense phenotyping and a novel machine learning (ML) approach capable of handling large number of time-dependent variables and incorporating non-linear relationships as well as complex interactions between risk factors, we aimed to develop a population-based approach to identify individuals at high SCD risk.

Methods

Study Cohort

The Atherosclerosis Risk in Communities (ARIC) study is an ongoing longitudinal cohort of 15,792 middle-aged men and women recruited from four communities in the U.S.: Forsyth County, North Carolina; Jackson, Mississippi; suburbs of Minneapolis, Minnesota; and Washington County, Maryland at 1987-1989 (visit 1). The 3 short-term follow-up visits occurred approximately three years apart: 1990-1992 (visit 2), 1993-1995 (visit 3), 1996-1998 (visit 4).

Each study visit consisted of a comprehensive examination, which included physical and clinical examination, blood and urine specimens' collection for laboratory testing, administration of extensive questionnaires, and a 12-lead electrocardiogram (ECG). In addition, participants were contacted by phone annually for hospitalizations and death information during the prior year. If any clinical events happened, hospital records and death certificates would be obtained for ascertainment by physicians.¹² In the current study, participants with ICD (N=131) were excluded from the analysis. Study protocols were approved by the Institutional Review Boards and all study participants provided informed consent.

Assessment of Candidate Predictors

Clinical predictors were measured and updated during the first four ARIC visits from 1987-89 to 1996-98. At visit 1, demographics variables including age at the time of the visit, sex, race, center, education level, income, and family history (mother and father) of diseases were collected using an interviewer-administered questionnaire. Anthropometric variables including height, weight, and waist circumferences were measured by standard protocol at all visits, and body mass index (BMI) was calculated as weight (in kilograms) divided by the square of height (in meters). Lifestyle factors including smoking status, alcohol intake, physical activity [intensity and meeting American Heart Association (AHA) recommendations or not], and dietary quality were evaluated using questionnaires that assess each participant's self-reported information at all visits (alcohol intake and smoking status) or visit 1 and 3 (physical activity and dietary quality).¹³⁻¹⁶ Clinical factors included systolic and diastolic blood pressure, hypertension, diabetes mellitus, coronary heart disease (CHD), stroke, atrial fibrillation, heart failure, and hospitalizations (yes/no and number of hospitalizations). These variables were evaluated at all visits.

The definitions and ascertainment of the clinical factors are described as follows. Blood pressure was measured in seated participants after a 5-minute rest. Hypertension was defined as systolic blood pressure ≥ 140 mm Hg, diastolic blood pressure ≥ 90 mm Hg, or use of antihypertensive medication in the 2 weeks prior to visits.¹⁷ Diabetes mellitus was defined as fasting blood glucose ≥ 126 mg/dL, non-fasting glucose ≥ 200 mg/dL, self-reported doctor-diagnosed diabetes, or use of diabetes medication in the 2 weeks prior to visits. CHD, at visit 1, was defined as myocardial infarction (MI) observed on ECG, self-reported history of MI, self-reported heart or arterial surgery, coronary bypass, balloon angioplasty, and coronary angioplasty and, at visit 2-4, additionally included CHD cases occurring after visit 1 but before the relevant visit which were identified based on hospitalization records and death records and adjudicated by physicians. Stroke, at visit 1, was defined self-reported stroke and, at visit 2-4, additionally included stroke cases occurring after visit 1 but before the relevant visit which were identified by a computer diagnostic algorithm detailed elsewhere and adjudicated by physicians.¹⁸ Atrial fibrillation was identified through ECGs from follow-up exams, hospital discharge records, and death certificates. Heart failure, at visit 1, was defined as use of heart failure medication or evidence of symptoms defined by stage 3 of the Gothenburg criteria and, at visit 2-4, additionally included heart failure cases occurring between visit 1 and the relevant visit which were identified through ICD codes in hospitalization and death records.¹⁹ For clinical events including diabetes, CHD, stroke, atrial fibrillation, and heart failure, incidences were only considered until visit 4. All-cause hospitalizations, cardiac-related hospitalization, and pulmonary-related hospitalization were ascertained through hospitalization records obtained from annual telephone contact with participants and active surveillance in the study community hospitals. Use of medication, such as

anti-hypertensive, anti-arrhythmic, lipid-lowering, anti-diabetic medication, were obtained by extracting medication names or codes from the lists transcribed from the medication containers study participants brought to the study visits.

Laboratory values or biomarkers included C-reactive protein (CRP), white blood cells (WBC) count, hematocrit, hemoglobin, N-terminal pro-brain natriuretic peptide (NT-proBNP), troponin I, troponin T, fibrinogen, total cholesterol, high-density lipoprotein cholesterol (HDL-C), low-density lipoprotein cholesterol (LDL-C), triglycerides, serum creatinine, estimated glomerular filtration rate (eGFR), serum albumin, urine albumin, blood glucose (fasting and non-fasting), serum sodium, serum potassium, blood urea nitrogen, and serum magnesium. Definition of each marker and at which visit(s) each marker was measured are described in detail in the

Supplemental Table 1.

ECG variables were derived based on the 12-lead ECGs results measured at all visits. Details of the ECG measuring protocol, data processing, monitoring, and quality control have been described elsewhere.²⁰ Twenty-one variables of ECG patterns classified by Minnesota Code, which utilized a series of measurement rules to assign numerical codes based on the location and severity of ECG findings, were included.^{21,22} The presence of diagnostic Q-wave, QRS duration, QT duration, and heart rate observed from ECG were also used as candidate predictors. Cornell voltage (S in $V3+R$ in aVL) calculated from ECG results for evaluating left ventricular thickness, as well as left ventricular hypertrophy (LVH) defined by Cornell voltage criteria.²³

Assessment of Outcome

SCD was defined as a sudden pulseless condition presumed due to a ventricular tachyarrhythmia in a previously stable individual without evidence of a non-cardiac cause of cardiac arrest occurring out of the hospital or in the emergency department. Details regarding the ascertainment of SCD events in ARIC have been described elsewhere.²⁴ Briefly, all fatal CHD events documented in death certificates, informant interviews, physician questionnaires, coroner reports, prior medical history, or hospital discharge summaries that occurred through 2012 were reviewed and independently adjudicated by 2 physicians. Cases were classified as definite sudden arrhythmic death, possible sudden arrhythmic death, not sudden arrhythmic death, or unclassifiable. For the current analysis, we defined SCD as adjudicated definite or possible sudden arrhythmic death that occurred by December 31, 2012.

Statistical Analysis

We used the Random Forest for Survival, Longitudinal, and Multivariate data analysis (RF-SLAM) methods for SCD risk prediction. Details of this method have been described elsewhere.²⁵ Briefly, random forest is an ensemble learning method that combines multiple decision trees trained on uncorrelated bootstrap replications of the original training data and outputs the averaged prediction across all trees.²⁶ RF-SLAM uses data partitioned into discrete intervals (also known as person-time intervals) and uses a Poisson regression log-likelihood as the split statistic thus allowing for modeling time-varying predictor variables. In this analysis, we followed the recommended settings of 1,000 as the number of trees, 10% of the total number of intervals as the minimum terminal node size, and the square root of the number of input variables as the number of predictors for each tree.²⁵ To simplify the interpretation of the RF-SLAM predictions, we grew a single regression tree that best approximates the RF-SLAM predictions

among those that explained at least 90% of the variation in the RF-SLAM predictions.²⁷ We grew a second tree that explained at least 80% of the variation for facilitating visualization and discussion of key predictors and their interactions.

We compared the performance of RF-SLAM to a series of standard Poisson regression models that share the interval event indicator as the outcome and differ by their specific variable selection approach. Missing data was handled through imputation by randomly drawing from non-missing data by RF-SLAM and by multivariate imputation by chained equations for the Poisson regression models.²⁸⁻³⁰ The RF-SLAM and the Poisson regression models shared the same pool of 82 potential (dynamic) predictor variables collected at the four visits including demographics, anthropometric variables, lifestyle factors, clinical characteristics, medication, laboratory values and biomarkers, ECG variables, and other cardiac functional indices. To evaluate the importance of time-varying covariates, we additionally include a random forest survival (RFS) model incorporating only baseline covariates.³¹

We considered the following pools of predictor variables for the Poisson regressions: (1) using the top predictors as selected in the summary regression tree that explained at least 80% of the variation of RF-SLAM (referred to as top predictors of RF-SLAM); (2) using top predictors of RF-SLAM as candidate predictors with selection by stepwise regression, (3) using top predictors of RF-SLAM plus interactions formed by parent-child node pairs identified from the summary regression tree of RF-SLAM and modeled based on the tree structure (referred to as top interactions of RF-SLAM) as predictors, and (4) using top predictors of RF-SLAM plus top interactions of RF-SLAM as candidate predictors with selection by stepwise regression.

Predictive accuracy was assessed at 3, 6, 9, and 25 years by time-dependent area under the receiver operating characteristic curve (AUC) through out-of-bag prediction for RF-SLAM and RFS models and 5-fold cross validation for Poisson regression models.³² The comparison between the RF-SLAM model, RFS model, and Poisson regression models is also summarized in **Supplemental Table 2**. Since multiple variables were measured in only one or two out of the total four visits resulting in substantial missingness we excluded all variables with greater than 50% missingness in sensitivity analysis to test the robustness of our results. Also, as CHD appeared to be the key predictor of SCD in our main model, as shown below, we excluded all participants with existing CHD at baseline as a sensitivity analysis. Analyses were conducted with R version 3.6.3 (<https://www.r-project.org>). Statistical significance was defined as P value < 0.05.

Results

Baseline Characteristics

Our primary study population included 15,661 participants (mean age 54.2 years; 55% female) in ARIC. Over a median follow-up of 23.5 years, there were a total of 583 adjudicated SCD cases, with 97, 93, 116, and 277 cases having occurred during the intervals of visit 1 to 2, visit 2 to 3, visit 3 to 4, and visit 4 to end of follow-up respectively. At baseline, participants who subsequently developed SCD were frequently male, self-identified as Black, had less than a high school degree, were current smokers and non-drinkers, and less physically active compared with individuals who did not develop SCD. They also had higher mean BMI and blood pressure, less favorable lipid profile, and lower kidney function when comparing with those who did not

develop SCD. CHD, heart failure, stroke, and diabetes mellitus, were more prevalent among participants who developed SCD, which was also reflected by the higher percentage of participants taking medications for managing those conditions. ECG parameters among participants with SCD included higher prevalence of diagnostic Q wave, atrial fibrillation and LVH, and higher Cornell voltage (**Table 1**). The prevalence of clinical conditions at visit 4 among participants who attended visit 4 is presented at **Supplemental Table 3**. Several variables were not measured in all four visits and therefore had a high prevalence of missingness (**Supplemental Table 4**).

RF-SLAM prediction

Using time-varying covariates from visits 1-4 of the ARIC study, the AUCs of RF-SLAM for SCD prediction at 3, 6, 9, and 25 years of follow-up were 0.83, 0.89, 0.83, and 0.77 respectively. In comparison, the AUCs for RSF were 0.82, 0.86, 0.80, and 0.75 at 3, 6, 9, and 25 years, respectively, when limiting to baseline covariates (**Figure 1A**). The predicted SCD risk by RF-SLAM with time-varying covariates ranged from 5.6×10^{-2} to 11.9 per 1000 person-years and was significantly different between SCD cases and controls, with the median of predicted risk being 3.4 per 1000 person-year ((25th percentile - 75th percentile, 1.7 - 5.4 per 1000 person-year) for cases and 1.1 per 1000 person-year (25th percentile - 75th percentile, 0.7 - 1.8 per 1000 person-year) for controls ($p=1.6 \times 10^{-84}$) (**Figure 1B**).

Approximately 80% of input candidate covariates were included in the summary tree that explain 90.0 % of the RF-SLAM predictions. Top predictors that explained most of the variation

included (1) clinical characteristics: CHD (variation explained: 47.9%), diabetes mellitus (12.3%), hypertension (4.9%), and number of cardiac-related hospitalization (2.2%) (2) ECG variables: Minnesota Code for T wave abnormality in any of leads I, aVL, or V6 (3.2%) and ST junction & segment depression in any of leads I, aVL, or V6 (1.2%), (3) use of medication: anti-diabetic medications (2.0%), and (4) biomarkers: NT-proBNP (3.3%), as well as sex (1.3%) (**Table 2**). Figure 2 shows the summary tree that explains 80.7% of the variation of the variation of the RF-SLAM predictions in **Figure 2** for visualizing the dependencies.

Poisson models comparison

Among all time-varying covariates, the following predictors of SCD were selected in all 5 cross-validations by stepwise regression, more specifically, a forward-backward search starting from the full model with the Akaike information criterion (AIC) for evaluating model fit: CHD, heart failure, stroke, number of all-cause hospitalization, number of cardiac-related hospitalization, three variables for ECG patterns classified by Minnesota Code (T wave in II, III, aVF, A-V conduction defect codes, and QRS transition zone), heart rate, glucose level, systolic blood pressure, diastolic blood pressure, estimated glomerular filtration rate, urinary albumin, troponin T, troponin I, fibrinogen, visit, and smoking status. Five-fold cross validation yielded AUC of 0.78, 0.83, 0.74, and 0.70 at 3, 6, 9, and 25 years of follow-up, respectively. Instead of using the stepwise variable selection, we also use the predictors selected by RF-SLAM model with and without an additional stepwise selection in the Poisson regression mode, which yielded AUC of around 0.74, 0.78, 0.73, and 0.73 at 3, 6, 9, and 25 years of follow-up, respectively, for the models with or without stepwise selection. Finally, manually adding in the top interactions

identified by RF-SLAM improved the AUC at 3 and 6 years of follow-up but not for AUC at 9 and 25 years (**Figure 1A**).

For sensitivity analysis, building summary trees with bootstrapping predicted values of the RF-SLAM model yield consistent selections of predictors and percentages of total variance explained by those predictors (**Supplemental Table 5**); limiting to variables with less than 50% missingness did not change the performance of RF-SLAM prediction (AUC at 3 years: 0.83, 6 years: 0.88, 9 years: 0.83, 25 years: 0.77) nor the selection of predictors (**Supplemental Figure 1 and Supplemental Table 6**); excluding participants with prevalent CHD at baseline resulted in slight drop in the performance (AUC at 3 years: 0.81, 6 years: 0.87, 9 years: 0.78, 25 years: 0.75) and the predictors other than CHD remained similar (**Supplemental Figure 1 and Supplemental Table 7**).

Discussion

In this community-based cohort of 15,663 middle-aged adults, we derived a prediction model for SCD with a novel ML approach, RF-SLAM using a large number of time-varying covariates. Participants who developed SCD had statistically significant higher predicted risk of SCD than those who did not. This model substantially outperformed a random forest survival model with only baseline covariates as well as Poisson regression models with time-varying predictors selected through stepwise regression. Prior CHD was identified as the top predictor for SCD, explaining 47.9% of the total phenotypic variance. The combination of these time-varying data and ML methods that accommodate dynamic data can contribute to improved risk stratification

for SCD in the generally healthy population and thus aid in targeted primary prevention strategies for high-risk individuals.

Our study highlighted the importance of the dynamic dependency of time-varying risk factors on SCD prediction. Leveraging novel ML approach that incorporates large number of time-varying predictor variables, we generated models that providing continuous a gradient of predicted SCD risk with significant differentiation between those who ultimately have versus do not have an SCD. This model demonstrated better performance than both using same ML approach but only inputting baseline variables and using traditional statistical regression analysis with stepwise selection of time-varying variables.

A single summary decision tree facilitates the portability and interpretability of complex clinical ML algorithms. The “black box” algorithms of ML methods can limit their clinical or epidemiologic utility. To surmount this limitation, we approximated the predictions from the complex algorithm into a single summary decision tree with represents the sequential subsetting that produces groups of differential risk and mimics clinical reasoning. In this way, we (1) identified the key predictors as well as calculated the variation in the risks among people explained by each predictor, which can be easily used as predictors by other comparable general population cohorts; and (2) visualize the interactions among predictors as reflected in the sequences of splitting variables, which facilitate understanding the dependency among risk factors. Top time-varying predictors identified using RF-SLAM including CHD, diabetes mellitus, hypertension, NT-proBNP, and T wave abnormality in any of leads I, aVL, and V6, , which are consistent with existing epidemiological findings.^{9,33,34} Among them, prior CHD alone

explained 47.9% of total variation of the RF-SLAM prediction, which corroborate with the fact that approximately 50% of SCD victims have prior diagnosis of heart disease.¹

Our identification of key predictors of SCD in a community-based cohort can better inform the prevention and management strategy for the general population. In our study, in addition to CHD, diabetes, hypertension along with their corresponding biomarkers and medications, explained a large proportion of the total variation of SCD prediction. Diabetes and hypertension are widely recognized modifiable cardiovascular risk factors. Furthermore, our study indicates that clinical surveillance may be appropriate for patients with CHD, diabetes, and hypertension for risk of SCD. For instance, monitoring could include serial electrocardiograms to assess features such as T wave abnormalities and ST junction & segment depression, as well as cardiac biomarkers, such as NT-proBNP, which were selected as top predictors for SCD in our study. A multitude of high-risk features may also prompt further risk stratification with echocardiography among asymptomatic patients.

Other features of the ML approach favor its use over classical regression methods. In random forest, each tree is built from a bootstrap sample of the observations in the training data. As a result of the bootstrapping, roughly one in three observations is randomly left out in a particular bootstrap. We can predict these "out-of-bag" observations from the model fit to the "in-bag" values and obtain a cross-validated estimate of prediction error along the way. Another key feature is that missing data imputation can be integrated within random forests but must be handled as separate steps for regression analysis.²⁵ Particularly in our study, through randomly drawing from the non-missing "in-bag" data within the current node during the tree growing

process, RF-SLAM demonstrated its robustness to missingness, a common and important issue in real-world longitudinal data, compared to conventional methods.

Several limitations should be considered when interpreting our findings. First, ECG has long offered valuable insights into cardiac health and disease.^{35,36} Several ECG abnormalities have been associated with SCD risk³⁷ and our models also selected some of them as predictors. We used the Minnesota code for ECG variables while recent advances in deep-learning convolutional neural networks (CNNs) have been used to extract information beyond that being captured by Minnesota code.³⁸ We will incorporate these additional ECG features generated through CNNs when they become available in the study population. Second, there is a time gap between visit 4 and the end of follow-up of our study, limiting the incorporation of features during this interval and likely contributing to the modest AUC decline during this interval. Third, for important clinical events predictors such as CHD and heart failure, we updated them just at each ARIC visit but not for events during each interval. Fourth, some candidate covariates in our study were not measured in all four visits resulting in substantial missingness. These include important cardiac biomarkers NT-proBNP and troponin T that have been associated with risk of multiple cardiovascular events including SCD.³⁹⁻⁴⁴ To address this, we conducted sensitivity analysis with excluding those covariates, and observed results consistent with our main analysis. Finally, cross-validation was performed in the diverse ARIC cohort and future research is needed to confirm generalizability in external cohorts.

In conclusion, our study highlighted the improved prediction for SCD of using a novel machine-learning approach with time-varying predictors, as well as the feasibility of applying this

approach in large cohorts and biobanks. Our findings allow identification of higher-risk individuals appropriate for targeted interventions designed to reduce the burden of SCD in the general population.

Acknowledgements

The authors thank the staff and participants of the ARIC study for their important contributions.

Fundings

N.D. was supported by NIH/NHLBI Cardiovascular Epidemiology training grant T32HL007024. S.W. was funded by the National Institutes of Health (NIH) F30HL142131 and 5T32GM007309 grants. S.Z. received partial support for this research from NIH grants 5U01HL096812-10, P30AR070254, 1R01AR073208-04, and 5UL1TR003098-03. The Atherosclerosis Risk in Communities study has been funded in whole or in part with Federal funds from the National Heart, Lung, and Blood Institute, National Institutes of Health, Department of Health and Human Services, under Contract nos. (HHSN268201700001I, HHSN268201700002I, HHSN268201700003I, HHSN268201700004I, HHSN268201700005I). The funding sources had no role in: a. the design or conduct of the study, b. the collection, management, analysis, and interpretation of the data, or c. preparation, review, or approval of the manuscript. The content is solely the responsibility of the authors and does not necessarily represent the official views of the funding sources.

Disclosure

P.N. reports grants from Amgen, Apple, AstraZeneca, Boston Scientific, and Novartis, personal fees from Apple, AstraZeneca, Blackstone Life Sciences, Foresite Labs, Genentech / Roche, Novartis, and TenSixteen Bio, equity in geneXwell, and TenSixteen Bio, co-founder of TenSixteen Bio, and spousal employment at Vertex, all unrelated to the present work. Other authors have no conflict of interest to declare.

References

1. Myerburg RJ, Goldberger JJ. Sudden Cardiac Arrest Risk Assessment: Population Science and the Individual Risk Mandate. *JAMA Cardiol* 2017;**2**:689-694. doi: 10.1001/jamacardio.2017.0266
2. Fishman GI, Chugh SS, Dimarco JP, *et al.* Sudden cardiac death prediction and prevention: report from a National Heart, Lung, and Blood Institute and Heart Rhythm Society Workshop. *Circulation* 2010;**122**:2335-2348. doi: 10.1161/circulationaha.110.976092
3. Hayashi M, Shimizu W, Albert CM. The spectrum of epidemiology underlying sudden cardiac death. *Circ Res* 2015;**116**:1887-1906. doi: 10.1161/circresaha.116.304521
4. Wellens HJJ, Schwartz PJ, Lindemans FW, *et al.* Risk stratification for sudden cardiac death: current status and challenges for the future†. *European Heart Journal* 2014;**35**:1642-1651. doi: 10.1093/eurheartj/ehu176
5. Buxton AE, Lee KL, Hafley GE, *et al.* Limitations of ejection fraction for prediction of sudden death risk in patients with coronary artery disease: lessons from the MUSTT study. *J Am Coll Cardiol* 2007;**50**:1150-1157. doi: 10.1016/j.jacc.2007.04.095
6. Wu KC, Wongvibulsin S, Tao S, *et al.* Baseline and Dynamic Risk Predictors of Appropriate Implantable Cardioverter Defibrillator Therapy. *J Am Heart Assoc* 2020;**9**:e017002. doi: 10.1161/jaha.120.017002
7. Suzuki T, Wang W, Wilsdon A, *et al.* Carotid Intima-Media Thickness and the Risk of Sudden Cardiac Death: The ARIC Study and the CHS. *J Am Heart Assoc* 2020;**9**:e016981. doi: 10.1161/jaha.120.016981
8. Perez-Alday EA, Bender A, German D, *et al.* Dynamic predictive accuracy of electrocardiographic biomarkers of sudden cardiac death within a survival framework: the

Atherosclerosis Risk in Communities (ARIC) study. *BMC cardiovascular disorders*

2019;**19**:255-255. doi: 10.1186/s12872-019-1234-9

9. Deo R, Norby FL, Katz R, *et al.* Development and Validation of a Sudden Cardiac Death Prediction Model for the General Population. *Circulation* 2016;**134**:806-816. doi:

doi:10.1161/CIRCULATIONAHA.116.023042

10. Koene RJ, Norby FL, Maheshwari A, *et al.* Predictors of sudden cardiac death in atrial fibrillation: The Atherosclerosis Risk in Communities (ARIC) study. *PloS one*

2017;**12**:e0187659-e0187659. doi: 10.1371/journal.pone.0187659

11. Goldenberg I, Vyas AK, Hall WJ, *et al.* Risk Stratification for Primary Implantation of a Cardioverter-Defibrillator in Patients With Ischemic Left Ventricular Dysfunction. *Journal of the American College of Cardiology* 2008;**51**:288-296. doi:

<https://doi.org/10.1016/j.jacc.2007.08.058>

12. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. *Am J Epidemiol* 1989;**129**:687-702. doi:

13. Ainsworth BE, Haskell WL, Herrmann SD, *et al.* 2011 Compendium of Physical Activities: a second update of codes and MET values. *Med Sci Sports Exerc* 2011;**43**:1575-1581. doi: 10.1249/MSS.0b013e31821ece12

14. Garber CE, Blissmer B, Deschenes MR, *et al.* American College of Sports Medicine position stand. Quantity and quality of exercise for developing and maintaining cardiorespiratory, musculoskeletal, and neuromotor fitness in apparently healthy adults: guidance for prescribing exercise. *Med Sci Sports Exerc* 2011;**43**:1334-1359. doi:

10.1249/MSS.0b013e318213febf

15. Piercy KL, Troiano RP, Ballard RM, *et al.* The Physical Activity Guidelines for Americans. *Jama* 2018;**320**:2020-2028. doi: 10.1001/jama.2018.14854
16. Hu EA, Steffen LM, Grams ME, *et al.* Dietary patterns and risk of incident chronic kidney disease: the Atherosclerosis Risk in Communities study. *Am J Clin Nutr* 2019;**110**:713-721. doi: 10.1093/ajcn/nqz146
17. Whelton PK, Carey RM, Aronow WS, *et al.* 2017 ACC/AHA/AAPA/ABC/ACPM/AGS/APhA/ASH/ASPC/NMA/PCNA Guideline for the Prevention, Detection, Evaluation, and Management of High Blood Pressure in Adults: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines. *Hypertension* 2018;**71**:1269-1324. doi: 10.1161/hyp.000000000000066
18. Rosamond WD, Folsom AR, Chambless LE, *et al.* Stroke incidence and survival among middle-aged adults: 9-year follow-up of the Atherosclerosis Risk in Communities (ARIC) cohort. *Stroke* 1999;**30**:736-743. doi: 10.1161/01.str.30.4.736
19. Eriksson H, Caidahl K, Larsson B, *et al.* Cardiac and pulmonary causes of dyspnoea--validation of a scoring test for clinical-epidemiological use: the Study of Men Born in 1913. *Eur Heart J* 1987;**8**:1007-1014. doi: 10.1093/oxfordjournals.eurheartj.a062365
20. Vitelli LL, Crow RS, Shahar E, *et al.* Electrocardiographic findings in a healthy biracial population. Atherosclerosis Risk in Communities (ARIC) Study Investigators. *Am J Cardiol* 1998;**81**:453-459. doi: 10.1016/s0002-9149(97)00937-5
21. Crow RS, Prineas RJ, Hannan PJ, Grandits G, Blackburn H. Prognostic associations of Minnesota Code serial electrocardiographic change classification with coronary heart disease

mortality in the Multiple Risk Factor Intervention Trial. *Am J Cardiol* 1997;**80**:138-144. doi: 10.1016/s0002-9149(97)00307-x

22. Prineas RJ, Crow RS, Zhang Z. The Minnesota Code Manual of Electrocardiographic Findings (including measurement and comparison with the Novacode) 2nd Edition new and enlarged ed ed. London: Springer; 2010.

23. Jain A, Tandri H, Dalal D, *et al.* Diagnostic and prognostic utility of electrocardiography for left ventricular hypertrophy defined by magnetic resonance imaging in relationship to ethnicity: the Multi-Ethnic Study of Atherosclerosis (MESA). *Am Heart J* 2010;**159**:652-658. doi: 10.1016/j.ahj.2009.12.035

24. 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**:2222-2234. doi: 10.1161/CIRCULATIONAHA.116.021306

25. Wongvibulsin S, Wu KC, Zeger SL. Clinical risk prediction with random forests for survival, longitudinal, and multivariate (RF-SLAM) data analysis. *BMC Med Res Methodol* 2019;**20**:1. doi: 10.1186/s12874-019-0863-0

26. Breiman L. Random Forests. *Machine Learning* 2001;**45**:5-32. doi: 10.1023/A:1010933404324

27. Wongvibulsin S, Wu KC, Zeger SL. Improving Clinical Translation of Machine Learning Approaches Through Clinician-Tailored Visual Displays of Black Box Algorithms: Development and Validation. *JMIR medical informatics* 2020;**8**:e15791-e15791. doi: 10.2196/15791

28. Graham JW. Missing data analysis: making it work in the real world. *Annu Rev Psychol* 2009;**60**:549-576. doi: 10.1146/annurev.psych.58.110405.085530
29. Schafer JL. Multiple imputation: a primer. *Stat Methods Med Res* 1999;**8**:3-15. doi: 10.1177/096228029900800102
30. Lee KJ, Carlin JB. Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation. *Am J Epidemiol* 2010;**171**:624-632. doi: 10.1093/aje/kwp425
31. Ishwaran H. and Kogalur U.B. (2021). Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), R package version 2.12.1. In.
32. Bansal A, Heagerty PJ. A Tutorial on Evaluating the Time-Varying Discrimination Accuracy of Survival Models Used in Dynamic Decision Making. *Med Decis Making* 2018;**38**:904-916. doi: 10.1177/0272989x18801312
33. Pham Q, Quan KJ, Rosenbaum DS. T-wave alternans: marker, mechanism, and methodology for predicting sudden cardiac death. *Journal of Electrocardiology* 2003;**36**:75-81. doi: <https://doi.org/10.1016/j.jelectrocard.2003.09.018>
34. Myerburg RJ, Junttila MJ. Sudden Cardiac Death Caused by Coronary Heart Disease. *Circulation* 2012;**125**:1043-1052. doi: doi:10.1161/CIRCULATIONAHA.111.023846
35. Mehta NJ, Khan IA. Cardiology's 10 greatest discoveries of the 20th century. *Tex Heart Inst J* 2002;**29**:164-171. doi:
36. Drew BJ, Califf RM, Funk M, *et al.* Practice standards for electrocardiographic monitoring in hospital settings: an American Heart Association scientific statement from the Councils on Cardiovascular Nursing, Clinical Cardiology, and Cardiovascular Disease in the Young: endorsed by the International Society of Computerized Electrocardiology and the

American Association of Critical-Care Nurses. *Circulation* 2004;**110**:2721-2746. doi:

10.1161/01.Cir.0000145144.56673.59

37. Terho HK, Tikkanen JT, Kenttä TV, *et al.* Electrocardiogram as a predictor of sudden cardiac death in middle-aged subjects without a known cardiac disease. *Int J Cardiol Heart Vasc* 2018;**20**:50-55. doi: 10.1016/j.ijcha.2018.08.002

38. Siontis KC, Noseworthy PA, Attia ZI, Friedman PA. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology* 2021;**18**:465-478. doi: 10.1038/s41569-020-00503-2

39. Levin ER, Gardner DG, Samson WK. Natriuretic Peptides. *New England Journal of Medicine* 1998;**339**:321-328. doi: 10.1056/nejm199807303390507

40. Reichlin T, Hochholzer W, Bassetti S, *et al.* Early Diagnosis of Myocardial Infarction with Sensitive Cardiac Troponin Assays. *New England Journal of Medicine* 2009;**361**:858-867. doi: 10.1056/NEJMoa0900428

41. Patton KK, Sotoodehnia N, DeFilippi C, *et al.* N-terminal pro-B-type natriuretic peptide is associated with sudden cardiac death risk: the Cardiovascular Health Study. *Heart Rhythm* 2011;**8**:228-233. doi: 10.1016/j.hrthm.2010.10.038

42. Wang TJ, Larson MG, Levy D, *et al.* Plasma Natriuretic Peptide Levels and the Risk of Cardiovascular Events and Death. *New England Journal of Medicine* 2004;**350**:655-663. doi: 10.1056/NEJMoa031994

43. Nakamura H, Niwano S, Fukaya H, *et al.* Cardiac troponin T as a predictor of cardiac death in patients with left ventricular dysfunction. *J Arrhythm* 2017;**33**:463-468. doi: 10.1016/j.joa.2017.07.004

44. Willeit P, Welsh P, Evans JDW, *et al.* High-Sensitivity Cardiac Troponin Concentration and Risk of First-Ever Cardiovascular Outcomes in 154,052 Participants. *Journal of the American College of Cardiology* 2017;**70**:558-568. doi: doi:10.1016/j.jacc.2017.05.062

Table 1. Baseline characteristics of Atherosclerosis Risk in Communities (ARIC) study participants.

Metric*	No SCD (N = 15,078)	SCD (N = 583)	P-value
Demographics & anthropometric variables			
Age (years)	54.1 (5.8)	56.2 (5.6)	<0.001
Male	6623 (43.9)	365 (62.6)	<0.001
White	11048 (73.5)	338 (58.0)	<0.001
At least high school degree	11546 (76.7%)	358 (61.5%)	<0.001
Body mass index (kg/m ²)	27.7 (5.3)	29.3 (6.0)	<0.001
Lifestyle factors			
Current smoker	3886 (25.8)	216 (37.0)	<0.001
Current alcohol drinker	8422 (56.1)	273 (47.2)	<0.001
Physical activity (MET-min/week)	611.8 (764.7)	502.2 (684.1)	0.001
Clinical factors			
Systolic blood pressure (mmHg)	120.9 (18.7)	131.0 (22.6)	<0.001
Diastolic blood pressure (mmHg)	73.6 (11.2)	76.9 (13.3)	<0.001
Hypertension	5086 (33.9)	351 (60.6)	<0.001
Coronary heart disease	600 (4.1)	133 (23.5)	<0.001
Heart failure	683 (4.6)	61 (10.7)	<0.001
Stroke	239 (1.6)	43 (7.4)	<0.001
Diabetes mellitus	1089 (7.2)	131 (22.5)	<0.001
Medication classes			
Anti-hypertensive	4455 (29.6)	319 (54.7)	<0.001
Lipid-lowering	423 (2.8)	23 (4.0)	0.129
Anti-diabetic	774 (6.1)	100 (19.6)	<0.001
Anti-arrhythmic	104 (0.7)	14 (2.4)	<0.001

Laboratory values or biomarkers

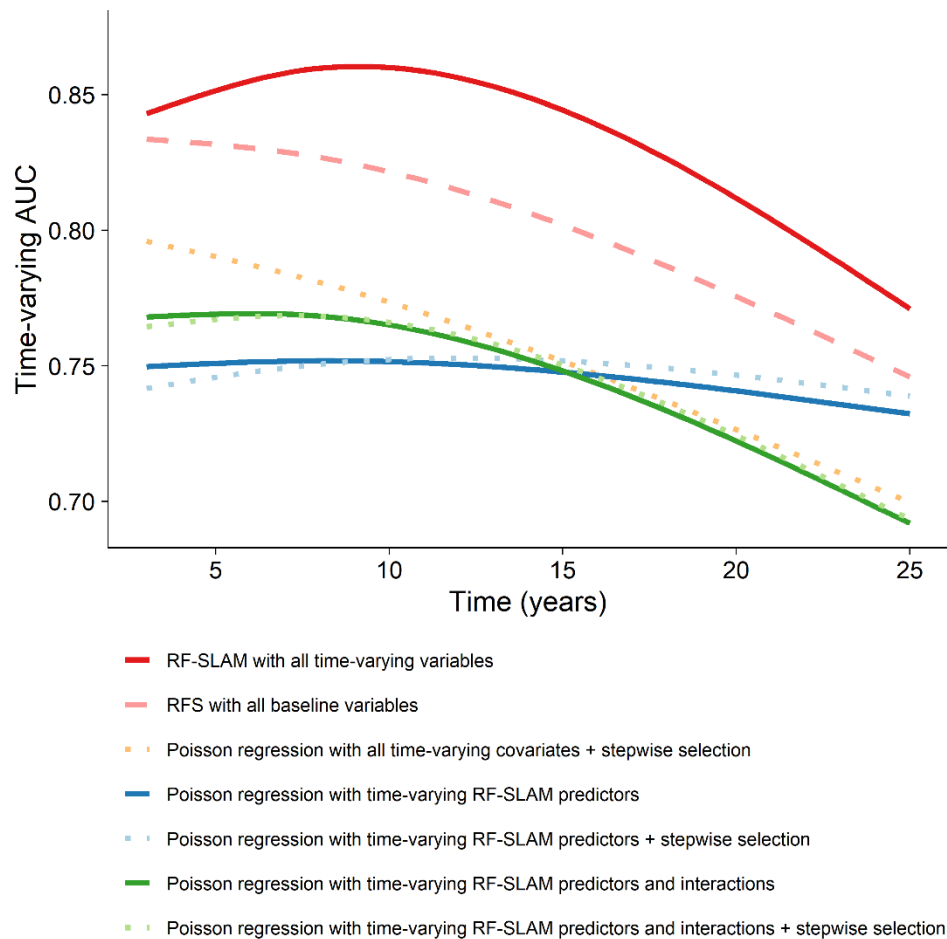
HDL-C, md/dL	49.1 [39.5, 61.6]	43.3 [35.6, 53.7]	<0.001
LDL-C, md/dL	134.9 [110.8, 160.8]	146.3 [119.3, 174.9]	<0.001
Triglyceride, md/dL	109.0 [78.0, 156.0]	124.0 [88.0, 188.8]	<0.001
Total cholesterol, md/dL	212.0 [186.0, 239.0]	219.0 [193.0, 251.0]	<0.001
eGFR, mL/min/1.73 m ²	102.6 (15.7)	99.6 (19.2)	<0.001

Electrocardiogram variables

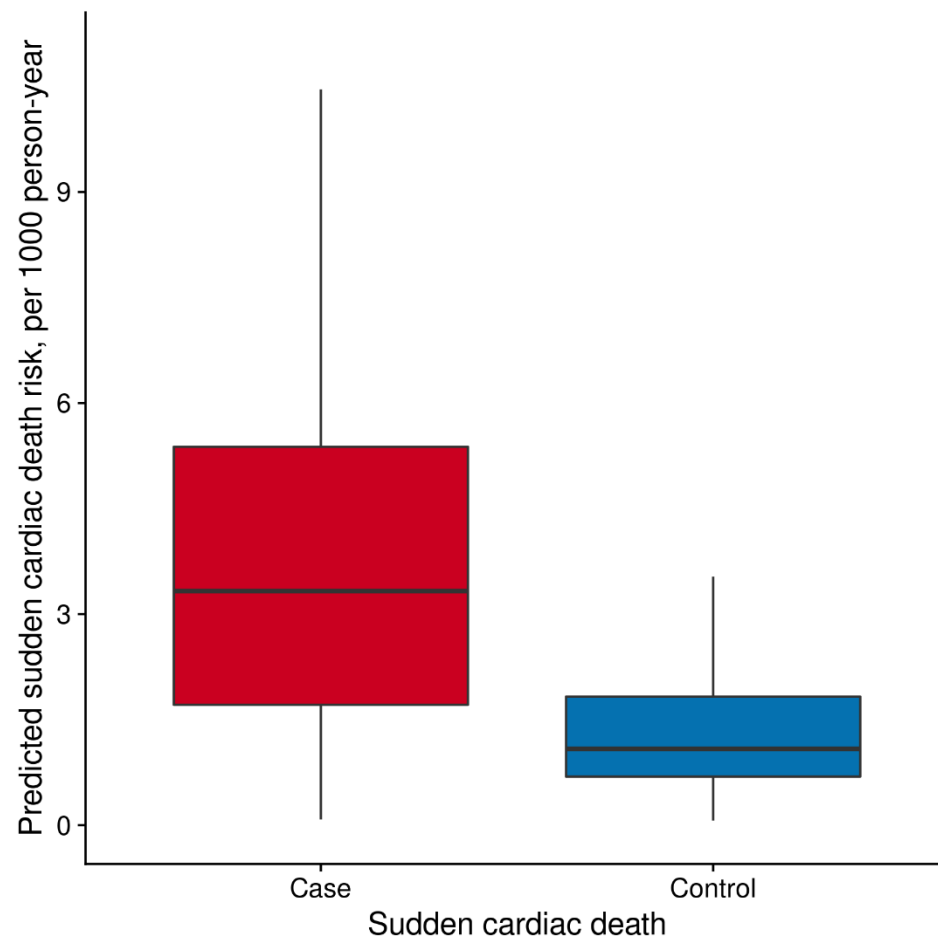
Atrial fibrillation	34 (0.2)	3 (0.5)	0.324
Heart rate, beats per minute	66.7 (10.3)	67.8 (11.5)	0.015
Presence of diagnostic Q wave	131 (0.9)	44 (7.7)	<0.001
Cornell voltage (S in V3+R in aVL), uV	1245.8 (556.6)	1556.1 (650.1)	<0.001
Left ventricular hypertrophy	308 (2.1)	33 (5.9)	<0.001

*Values are expressed as mean (SD) or median [IQR] for continuous variables and N (%) for categorical variables.

eGFR: Estimated glomerular filtration rate; HDL-C: high-density lipoprotein cholesterol; LDL-C: low-density lipoprotein cholesterol.



(A)



(B)

Figure 1. Performance of the RF-SLAM model. (A) AUC performances for predicting sudden cardiac death by RF-SLAM model incorporating time-varying covariates comparing with RFS model incorporating baseline covariates and five Poisson regression models incorporating time-varying covariates. Random forest models are in red; the model using time-varying covariates is shown as a solid line and the one using baseline covariates is shown as dashed lines with lighter color. For the five Poisson regression models, models with all candidate covariates, with predictors that accounted for > 80% of the RF-SLAM prediction, and with predictors that accounted for > 80% of the RF-SLAM prediction and their interactions are in yellow, blue, and green respectively. Models with stepwise selection

are shown as solid lines and those without are shown as dashed lines with lighter colors. (B) Predicted sudden cardiac death risk per 1,000 person-year by RF-SLAM among cases and controls. Predicted risk was calculated as the mean of predicted annual risks (unit: 1,000 person-year) of all follow-up time units. Red box indicates cases and blue box indicates controls. AUC: area under the curve; RF-SLAM: random forest statistical method for survival, longitudinal, and multivariable outcomes. RFS: random forest survival.

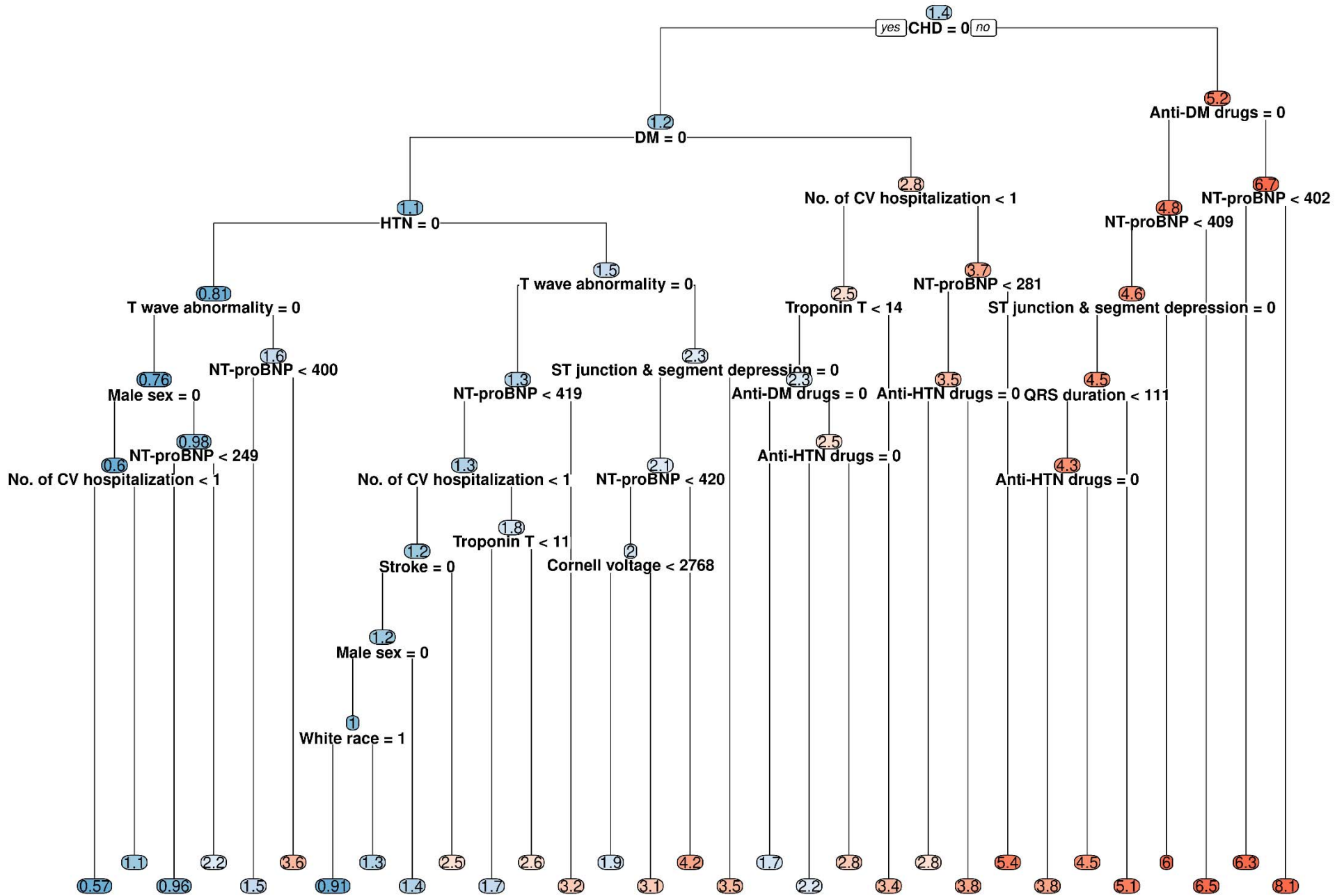


Figure 2. Summary tree of RF-SLAM depicting the time-varying predictors for sudden cardiac death that accounted for > 80% of the prediction. The number in each node indicate the mean predicted sudden cardiac death risk (unit: 1,000 person-year) of that

node. Red nodes indicate high risk and blue nodes indicate low risk. The darker the color, the higher/lower the risk. RF-SLAM: random forest statistical method for survival, longitudinal, and multivariable outcomes. CHD: coronary heart disease, CV: cardiovascular, DM: diabetes mellitus, HTN: hypertension, NT-proBNP: N-terminal pro-brain natriuretic peptide.

Table 2. Predictors that together explained > 90% of the RF-SLAM prediction for sudden cardiac death.

	Variable	Category	% of total variation explained
1	Prior coronary heart disease*	Clinical factors	47.95%
2	Diabetes mellitus*	Clinical factors	12.29%
3	Hypertension*	Clinical factors	4.97%
4	N-terminal pro-brain natriuretic peptide*	Laboratory values or biomarkers	3.78%
5	T wave abnormality in any of leads I, aVL, and V6*	Electrophysiologic variables	3.61%
6	Number of cardiac-related hospitalization*	Clinical factors	2.39%
7	Use of anti-diabetic medications*	Medications	2.13%
8	ST junction & segment depression in any of leads I, aVL, or V6*	Electrophysiologic variables	1.47%
9	Sex*	Demographics & anthropometric variables	1.35%
10	Troponin T*	Laboratory values or biomarkers	1.33%
11	QRS duration*	Electrophysiologic variables	1.15%
12	Cornell voltage *	Other cardiac indices	0.96%
13	Use of anti-hypertensive medications*	Medications	0.79%
14	Prior stroke*	Clinical factors	0.78%
15	Race*	Demographics & anthropometric variables	0.78%
16	Visit	Demographics & anthropometric variables	0.61%
17	Prior heart failure	Clinical factors	0.55%
18	Troponin I	Laboratory values or biomarkers	0.44%
19	Blood glucose	Laboratory values or biomarkers	0.33%
20	Systolic blood pressure	Clinical factors	0.28%
21	Q-Q.S. pattern II, III, aVF	Electrophysiologic variables	0.24%
22	Smoking status	Lifestyle factors	0.21%
23	Serum creatinine	Laboratory values or biomarkers	0.19%
24	Heart rate	Electrophysiologic variables	0.17%
25	Education	Demographics & anthropometric variables	0.16%

26	High-density lipoprotein cholesterol	Laboratory values or biomarkers	0.13%
27	Age	Demographics & anthropometric variables	0.12%
28	Hematocrit	Laboratory values or biomarkers	0.12%
29	Urine albumin	Laboratory values or biomarkers	0.08%
30	Body mass index	Demographics & anthropometric variables	0.08%
31	Estimated glomerular filtration rate	Laboratory values or biomarkers	0.08%
32	Waist circumferences	Demographics & anthropometric variables	0.08%
33	Number of hospitalizations	Clinical factors	0.08%
34	Diastolic blood pressure	Clinical factors	0.06%
35	Triglycerides	Laboratory values or biomarkers	0.04%
36	ST junction & segment depression V1-V5	Electrophysiologic variables	0.04%
37	Jackson field center	Demographics & anthropometric variables	0.03%
38	Dietary quality score	Lifestyle factors	0.03%
39	Low-density lipoprotein cholesterol	Laboratory values or biomarkers	0.02%
40	Serum albumin	Laboratory values or biomarkers	0.02%
41	C-reactive protein	Laboratory values or biomarkers	0.02%
42	Atrial fibrillation	Clinical factors	0.02%
43	Q-Q.S. pattern V1-V5	Electrophysiologic variables	0.02%
44	Income	Demographics & anthropometric variables	0.02%
45	White blood cells count	Laboratory values or biomarkers	0.02%
46	Father had coronary heart disease	Demographics & anthropometric variables	0.01%
47	Total cholesterol	Laboratory values or biomarkers	0.01%
48	Physical activity	Lifestyle factors	0.01%
49	Hemoglobin	Laboratory values or biomarkers	0.01%
50	Washington field center	Demographics & anthropometric variables	0.01%
51	Mother had coronary heart disease	Demographics & anthropometric variables	0.01%
52	High Amplitude R Wave Codes	Electrophysiologic variables	0.01%
53	T Wave V1-V5	Electrophysiologic variables	0.01%
54	QRS axis deviation codes	Electrophysiologic variables	0.01%
55	Minneapolis field center	Demographics & anthropometric variables	0.005%

56	QRS transition zone	Electrophysiologic variables	0.004%
57	Drinking status	Lifestyle factors	0.004%
58	T Wave II, III, aVF	Electrophysiologic variables	0.004%
59	Serum nitrogen	Laboratory values or biomarkers	0.003%

*Indicate variables that together explained > 80% of the RF-SLAM prediction.

RF-SLAM: random forest statistical method for survival, longitudinal, and multivariable outcomes.