# Cardiovascular Event Prediction by Machine Learning

## The Multi-Ethnic Study of Atherosclerosis

Bharath Ambale-Venkatesh, Xiaoying Yang, Colin O. Wu, Kiang Liu, W. Gregory Hundley, Robyn McClelland, Antoinette S. Gomes, Aaron R. Folsom, Steven Shea, Eliseo Guallar, David A. Bluemke, João A.C. Lima

***Rationale:*** **Machine learning may be useful to characterize cardiovascular risk, predict outcomes, and identify biomarkers in population studies.**

***Objective:*** **To test the ability of random survival forests, a machine learning technique, to predict 6 cardiovascular outcomes in comparison to standard cardiovascular risk scores.**

***Methods and Results:*** **We included participants from the MESA (Multi-Ethnic Study of Atherosclerosis). Baseline measurements were used to predict cardiovascular outcomes over 12 years of follow-up. MESA was designed to study progression of subclinical disease to cardiovascular events where participants were initially free of cardiovascular disease. All 6814 participants from MESA, aged 45 to 84 years, from 4 ethnicities, and 6 centers across the United States were included. Seven-hundred thirty-five variables from imaging and noninvasive tests, questionnaires, and biomarker panels were obtained. We used the random survival forests technique to identify the top-20 predictors of each outcome. Imaging, electrocardiography, and serum biomarkers featured heavily on the top-20 lists as opposed to traditional cardiovascular risk factors. Age was the most important predictor for all-cause mortality. Fasting glucose levels and carotid ultrasonography measures were important predictors of stroke. Coronary Artery Calcium score was the most important predictor of coronary heart disease and all atherosclerotic cardiovascular disease combined outcomes. Left ventricular structure and function and cardiac troponin-T were among the top predictors for incident heart failure. Creatinine, age, and ankle-brachial index were among the top predictors of atrial fibrillation. TNF-α (tissue necrosis factor-α) and IL (interleukin)-2 soluble receptors and NT-proBNP (N-Terminal Pro-B-Type Natriuretic Peptide) levels were important across all outcomes. The random survival forests technique performed better than established risk scores with increased prediction accuracy (decreased Brier score by 10%–25%).**

***Conclusions:*** **Machine learning in conjunction with deep phenotyping improves prediction accuracy in cardiovascular event prediction in an initially asymptomatic population. These methods may lead to greater insights on subclinical disease markers without apriori assumptions of causality.**

***Clinical Trial Registration:*** **URL: http://www.clinicaltrials.gov. Unique identifier: NCT00005487.**

**(*Circ Res*. 2017;121:1092-1101. DOI: 10.1161/CIRCRESAHA.117.311312.)**

**Key Words:** atrial fibrillation ■ cardiovascular disease ■ coronary heart disease ■ heart failure ■ machine learning ■ mortality ■ stroke



Event prediction has been the cornerstone of cardiovascular epidemiology as exemplified by the Framingham study and other prospective studies that function as pillars for much of what comprises current cardiovascular medicine.[1] A fundamental goal of such efforts has been event prediction over relatively long periods of time such as 10 years or a lifetime. These efforts have allowed us to characterize subclinical disease processes and target key risk factors for modification (eg, smoking cessation, statin therapy, blood pressure control).[2] Epidemiological studies used to derive such predictive models frequently contain hundreds or thousands of variables. It is in this context that machine learning methods might be useful as a means to identify the best predictors of outcomes from among millions of phenotypic data points.

Original received May 9, 2017; revision received July 28, 2017; accepted August 9, 2017. In July 2016, the average time from submission to first decision for all original research papers submitted to *Circulation Research* was 12.80 days.

From the Department of Radiology (B.A.-V.), Bloomberg School of Public Health (E.G.), and Department of Medicine, Cardiology and Radiology (J.A.C.L.), Johns Hopkins University, Baltimore, MD; George Washington University, DC (X.Y.); Office of Biostatistics, NHLBI, NIH, Bethesda, MD (C.O.W.); Department of Preventive Medicine, Northwestern University Medical School, Chicago, IL (K.L.); Department of Cardiology, Wake Forest University Health Sciences, Winston-Salem, NC (W.G.H.); Department of Biostatistics, University of Washington, Seattle (R.M.); Department of Radiology, UCLA School of Medicine, Los Angeles, CA (A.S.G.); Division of Epidemiology and Community Health, University of Minnesota, Minneapolis (A.R.F.); Departments of Medicine and Epidemiology, Columbia University, New York, NY (S.S.); and Radiology and Imaging Sciences, NIH Clinical Center, Bethesda, MD (D.A.B.).

**The online-only Data Supplement is available with this article at http://circres.ahajournals.org/lookup/suppl/doi:10.1161/CIRCRESAHA.117.311312/-/DC1.**

Correspondence to João A.C. Lima, MD, Johns Hopkins Hospital, 600 N Wolfe St, Blalock 524, Baltimore, MD 21287. E-mail jlima@jhmi.edu

## Novelty and Significance

### What Is Known?

- Machine learning techniques, such as the random survival techniques, may be an effective statistical methodology for handling biomedical data of increased volume, velocity, and variety, under the curse of dimensionality.
- These methods do not require a priori assumptions on causality and may thus be suitable for defining the role of novel biomarkers in cardiovascular disease prediction.

### What New Information Does This Article Contribute?

- Machine learning methods are better suited for meaningful risk prediction in extensively phenotyped large-scale epidemiological studies than regular Cox proportional Hazards models or risk scores.
- Random survival forests may be an effective machine learning strategy for incident cardiovascular event prediction and risk stratification in large populations with large phenotypic data sets.

There is a lack of studies using machine learning techniques with deep phenotyping (multiple evaluations of different aspects of a specific disease process) for cardiovascular event prediction. We examined the ability of combining deep phenotyping with machine learning for cardiovascular event prediction in the MESA (Multi-Ethnic Study of Atherosclerosis). The random survival forests–based method of risk prediction yielded an entirely unexpected perspective on event prediction of specific outcomes such as death, stroke, cardiovascular events, incident heart failure, and atrial fibrillation, with superior predictive power and improved accuracy than established risk scores. The results also suggest the importance of subclinical disease markers determined by imaging, electrocardiography, and blood tests, as revealed by their prominent presence on the lists of the top-20 phenotyping predictors for the selected outcomes. This strategy could yield insights about specific use of variables for specific event prediction and guiding strategies to prevent cardiovascular disease outcomes. Potentially, these techniques could be applied retrospectively to analyze large phenotyping data sets for identifying disease mechanisms, and as a means of hypothesis generation, without prior assumptions.

### Nonstandard Abbreviations and Acronyms

| | |
|---|---|
| **ABI** | ankle-brachial index |
| **AF** | atrial fibrillation |
| **AIC-Cox** | Akaike Information Criterion for Cox regression |
| **BS** | Brier score |
| **C-index** | concordance index |
| **CAC** | Coronary Artery Calcium score |
| **CHD** | coronary heart disease |
| **Cox-PHM** | Cox Proportional Hazard regression model |
| **CVD** | cardiovascular disease |
| **HF** | heart failure |
| **IL-2 SR** | interleukin-2 soluble receptor |
| **LASSO-Cox** | least absolute shrinkage and selection operator for Cox |
| **LV** | left ventricle |
| **MESA** | Multi-Ethnic Study of Atherosclerosis |
| **NT-proBNP** | N-Terminal Pro-B-Type Natriuretic Peptide |
| **RSF** | random survival forest |
| **TNF-α SR** | tissue necrosis factor-α soluble receptor |

Editorial, see p 1032
In This Issue, see p 1021
Meet the First Author, see p 1022

The Cox Proportional Hazard regression model (Cox-PHM) is often limited for data mining purposes because of correlation between variables, nonlinearity of variables (including potential complex interactions among them), and the possibility of overfitting. On the other hand, machine learning methods, such as random survival forests (RSFs), use a nonparametric decision tree approach to overcome these issues.[3] The purpose of this study was to (1) compare machine learning approaches to Cox-PHM and traditional risk scores for cardiovascular event prediction and (2) identify the important predictors for each of 6 cardiovascular clinical outcomes in a large epidemiological study.

## Methods

The design of the MESA study (Multi-Ethnic Study of Atherosclerosis) has been described previously.[4] In brief, MESA is a prospective, population-based observational cohort study of 6814 men and women representing 4 racial/ethnic groups, aged 45 to 84 years, and free of clinical cardiovascular disease (CVD) at enrollment. As part of the baseline examination (2000–2002), study participants were recruited at 6 field centers in the United States. Institutional review boards of all field centers approved the study protocol, and all participants gave informed consent. Information on assessment of markers within MESA has been described previously,[4] a detailed description is provided in the Online Data Supplement, and a list of markers is shown in Table 1. We included markers from questionnaires, demographics, traditional risk factors, anthropometry, medication use, biochemistry, magnetic resonance imaging of the heart and aorta, coronary computed tomography, carotid ultrasound, ECG exams, and ankle-brachial index (ABI).

### Outcomes

All-cause death, stroke, all CVD, coronary heart disease (CHD), atrial fibrillation (AF), and heart failure (HF) events adjudicated as part of MESA were used as end points (details in the Online Data Supplement). A telephone interviewer contacted each participant (or representative) every 6 to 9 months to inquire about all interim hospital admissions, outpatient diagnoses, and deaths. Two physicians reviewed all medical records for independent end point classification and assignment of event dates. Stroke was defined as rapid onset of a documented focal neurological deficit (vascular causes) lasting 24 hours or until death, or if <24 hours, when there was a clinically relevant brain lesion. Criteria for CHD included any of myocardial infarction, resuscitated cardiac arrest, definite angina, probable angina followed by revascularization, and CHD death. CVD outcomes represented a composite of CVD death, stroke, and CHD. Criteria for incident HF as an end point included symptomatic HF diagnosed by a physician and patient receiving medical treatment for HF, in addition to (1) pulmonary edema/congestion, and (2) dilated ventricle or poor left ventricular (LV) function, or evidence of LV diastolic dysfunction. Criteria for incident AF as an end point required in-hospital AF diagnosis according to ICD-9 (*International Classification of Disease-9*) codes.

### Statistical Analysis

Figure 1 shows the statistical analysis procedures followed in this study. Data transformation, indexing, and imputation (details in the

**Table 1. A List of the Markers That Were Used for Prediction in This Study**

| |
|---|
| Traditional risk factors, demographics, anthropometry, site |
| Age, sex, race, body mass index, body surface area, waist/hip ratio, systolic blood pressure, diastolic blood pressure, pulse pressure, diabetes mellitus, smoking status, pack-years, high-density lipoprotein cholesterol, low-density lipoprotein cholesterol, total cholesterol, triglycerides, heart rate, creatinine, site, waist circumference, hip circumference, fasting glucose |
| Medication use |
| All hypertension, angiotensin-converting enzyme, angiotensin-II receptor blockers, lipid control, statins, β-blockers, calcium channel blockers |
| Atherosclerotic markers—computed tomography, carotid ultrasonography |
| Coronary Artery Calcium score, ankle-brachial index, common and internal carotid artery intima media thickness, maximum carotid stenosis |
| Questionnaire |
| Family history of heart attacks, alcohol use, no. of drinks per week, emphysema, asthma, arthritis, cancer, liver disease, education level, economic status/income, exercise metabolic equivalents |
| Magnetic resonance imaging markers |
| Left ventricular (LV) mass, LV end-diastolic volume, LV end-systolic volume, LV ejection fraction, LV mass/volume ratio, LV stroke volume, LV sphericity index at end diastole and end systole, LV cardiac output, LV end-diastolic wall thickness, LV end-systolic wall thickness, ascending aortic distensibility, descending aortic distensibility, pulse wave velocity, maximum ascending aortic area, maximum descending aortic area, aortic arch distance, maximum left atrial (LA) volume, minimum LA volume, maximum LA strain, total LA ejection fraction, passive LA ejection fraction, active LA ejection fraction, right ventricular (RV) mass, RV end-diastolic volume, RV end-systolic volume, RV ejection fraction, RV stroke volume |
| Laboratory Biomarkers |
| Interleukin-2 soluble receptor, plasmin-antiplasmin complex, D-dimer, Factor VIII, NT-proBNP (N-Terminal Pro-B-Type Natriuretic Peptide), cardiac troponin-T, C-reactive protein, interleukin-6, fibrinogen, homocysteine, tissue necrosis factor-α soluble receptor |
| Electrocardiographic main |
| PR duration, QRS duration, QT duration, P axis, QRS axis, T axis, Minnesota codes, ECG LV hypertrophy by cornell voltage and novacode, heart rate variability short-term and overall components, Cornell voltage |
| ECG all |
| P-, P′-, Q-, R-, R′-, S-, S′-, T-, and T′-wave duration, amplitude, area, and intrinsicoid; middle and end of ST-segment amplitudes; amplitude at the point of 60 ms from J point; STJ amplitude; total QRS area, balance, deflection balance, intrinsicoid; for each of the leads (aVL, aVr, aVF, I, II, III, $V_1$, $V_2$, $V_3$, $V_4$, $V_5$, $V_6$) |

Online Data Supplement) were performed as necessary to generate data points to predict 6 outcomes over the follow-up period. A subset of the data set (66%) was randomly selected from the overall group of participants for training; the remaining 33.3% was used for testing/validation. The training data set was used for model construction using different approaches and optimized to reduce prediction error. These models were then tried on the test data set to examine model performance and identify the best predictors.

## Models Tested

We tested 9 different models in our analysis. The first model used the RSF algorithm on all available variables.[3] RSF is an ensemble tree method for analysis of right-censored data. Details of the RSF implementation are provided in the Online Data Supplement. In short, trees are grown by binary recursive splitting of data. At each split, a candidate variable that maximizes the difference in cumulative hazard between the daughter nodes (and the cutoff that identifies this maximum difference) is chosen. The splitting stops at the terminal nodes when the data at hand can no longer be split such that each terminal node has at least 1 unique outcome. For each tree, the cumulative hazard rate of a case is determined based on the terminal node that contains it. An ensemble hazard function (and survival probability) is estimated by averaging over all trees in a forest.

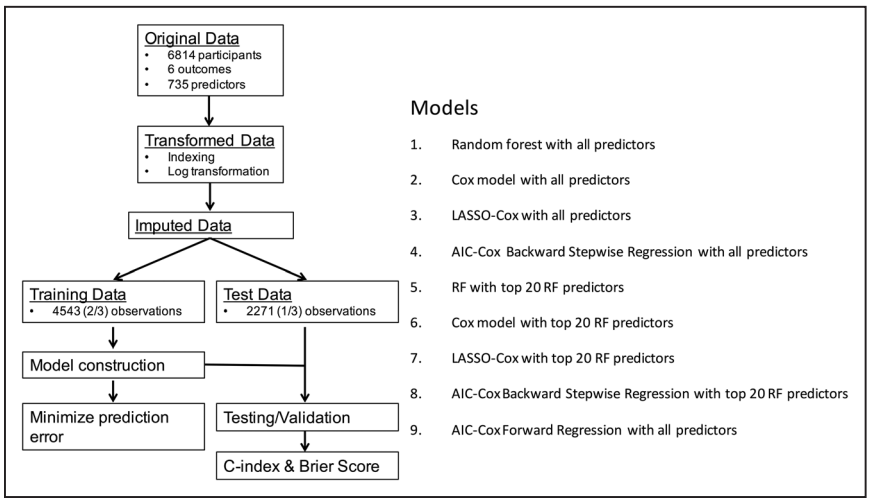


**Figure 1. A flowchart describing the general framework of the study.** Models were built using the training data set, and the test data set was used for computing the C-index and the Brier score shown in Table 4. AIC-Cox indicates Akaike Information Criterion for Cox regression; LASSO-Cox, least absolute shrinkage and selection operator for Cox regression; and RSF, random survival forest.

The Akaike Information Criterion[5] for Cox regression (AIC-Cox) with backward stepwise elimination and the AIC-Cox with forward stepwise regression, as well as the least absolute shrinkage and selection operator for Cox regression (LASSO-Cox), were tested in addition to the Cox-PHM. LASSO-Cox minimizes the log partial likelihood subject to the sum of the absolute values of the parameters being bounded by a constant. It shrinks coefficients and produces some coefficients that are zero, allowing efficient variable selection.[6]

Although RSF can be used instead of Cox regression analysis for risk prediction, it can also be used as an efficient variable selection technique. For variable selection using RSF, the variables were ranked by the mean of the minimal depth of the maximal subtree (highest point in the tree of a variable) over the entire forest (averaged over 1000 trees). Variables appearing higher on the tree have a higher rank. The top-20 ranked variables were then used again with RSF, AIC-Cox with forward stepwise, AIC-Cox with backward stepwise, LASSO-Cox, and Cox-PHM models.

## Performance Evaluation

We assessed the performance of each prediction model to discriminate outcomes on the test data set using Harrell concordance index (C-index),[7] and the accuracy of prediction (mean squared distance between the predicted probabilities and actual outcomes) using the

**Table 2. General Characteristics of the MESA Sample at Baseline, 2000 to 2002**

| Variable | Value |
|---|---|
| Age (in y) | 62.15 (10.23) |
| Sex (% female) | 52.85 |
| Race | |
| % Black | 27.78 |
| % White | 38.48 |
| % Chinese American | 11.78 |
| % Hispanic | 21.95 |
| Body mass index, kg/m$^2$ | 28.34 (5.48) |
| Diabetes mellitus | |
| % Impaired fasting glucose | 13.83 |
| % Treated | 10.01 |
| % Untreated | 2.64 |
| Systolic blood pressure, mm Hg | 126.59 (21.48) |
| Use of hypertension medication (%) | 37.23 |
| Heart rate, bpm | 63.13 (9.66) |
| Smoking status | |
| % current | 13.06 |
| % former | 36.62 |
| Total cholesterol, mg/dL | 194.16 (3573) |
| High-density lipoprotein cholesterol, mg/dL | 50.96 (14.83) |
| Lipid medication use, % | 16.15 |
| Heart failure, n (%) | 259 (3.8) |
| All cardiovascular disease, n (%) | 710 (10.4) |
| Coronary heart disease, n (%) | 498 (7.3) |
| Atrial fibrillation, n (%) | 317 (4.7) |
| All-cause death, n (%) | 831 (12.2) |
| Stroke, n (%) | 200 (2.9) |

**Table 3. The Top-20 Ranked Variables by the Variable Importance From the Random Survival Forest Method for Each of the Outcomes of Interest**

| Rank | Death | RVI | Stroke | RVI |
|---|---|---|---|---|
| 1 | Age | 0.00 | Fasting glucose | 0.00 |
| 2 | Tissue necrosis factor-α soluble receptor | 0.07 | Interleukin-2 soluble receptor | 0.09 |
| 3 | Interleukin-2 soluble receptor | 0.09 | Maximum carotid stenosis | 0.11 |
| 4 | NT-proBNP | 0.16 | Tissue necrosis factor-α soluble receptor | 0.13 |
| 5 | Ankle-brachial index | 0.21 | NT-proBNP | 0.16 |
| 6 | Coronary Artery Calcium score | 0.25 | Internal carotid intima media thickness | 0.18 |
| 7 | Common carotid intima media thickness | 0.26 | Systolic blood pressure | 0.24 |
| 8 | Internal carotid intima media thickness | 0.32 | Pulse pressure | 0.28 |
| 9 | Descending aortic distensibility | 0.33 | Descending aortic distensibility | 0.32 |
| 10 | Plasmin-antiplasmin complex | 0.35 | Ankle-brachial index | 0.32 |
| 11 | Cardiac troponin-T | 0.37 | Coronary Artery Calcium score | 0.32 |
| 12 | D-dimer | 0.37 | R amplitude in lead $V_2$ | 0.32 |
| 13 | Maximum ascending aortic area | 0.38 | R amplitude in lead $V_6$ | 0.35 |
| 14 | Ascending aortic distensibility | 0.39 | Minnesota code 1 score: F lead group | 0.35 |
| 15 | Homocysteine | 0.39 | Ascending aortic distensibility | 0.37 |
| 16 | Thoracic aorta arch length | 0.41 | Age | 0.38 |
| 17 | R amplitude in lead V | 0.41 | Cardiac output | 0.39 |
| 18 | Interleukin-6 | 0.41 | JT duration | 0.40 |
| 19 | Economic status/ income | 0.42 | LV mass/volume ratio | 0.40 |
| 20 | Maximum descending aortic area | 0.42 | End-diastolic septal anterior wall thickness | 0.41 |
| Rank | Coronary heart disease | RVI | All CVD | RVI |
| 1 | Coronary Artery Calcium score | 0.00 | Coronary Artery Calcium score | 0.00 |
| 2 | Tissue necrosis factor-α soluble receptor | 0.28 | Tissue necrosis factor-α soluble receptor | 0.24 |
| 3 | Cardiac troponin-T | 0.31 | NT-proBNP | 0.25 |
| 4 | NT-proBNP | 0.35 | Interleukin-2 soluble receptor | 0.28 |
| 5 | Minnesota code 1 score: F lead group | 0.36 | Cardiac troponin-T | 0.35 |

(*Continued*)

**Table 3. Continued**

| | | | | |
|---|---|---|---|---|
| 6 | Ankle-brachial index | 0.37 | Ankle-brachial index | 0.40 |
| 7 | Common carotid intima media thickness | 0.44 | Common carotid intima media thickness | 0.41 |
| 8 | Interleukin-2 soluble receptor | 0.48 | Pulse pressure | 0.41 |
| 9 | Pack-years of smoking | 0.50 | Maximum ascending aortic area | 0.42 |
| 10 | Internal carotid intima media thickness | 0.50 | Internal carotid intima media thickness | 0.42 |
| 11 | Factor VIII | 0.50 | Age | 0.42 |
| 12 | End-systolic midventricular septal wall thickness | 0.52 | R amplitude in lead $V_4$ | 0.44 |
| 13 | Maximum descending aortic area | 0.53 | Systolic blood pressure | 0.44 |
| 14 | End-systolic midventricular anteroseptal wall thickness | 0.54 | Factor VIII | 0.46 |
| 15 | Maximum ascending aortic area | 0.54 | Ascending aortic distensibility | 0.47 |
| 16 | S amplitude in lead aVr | 0.55 | Waist/hip ratio | 0.47 |
| 17 | End-diastolic basal septal wall thickness | 0.55 | Minnesota code 1 score: F lead group | 0.48 |
| 18 | Left ventricular ejection fraction | 0.56 | End-diastolic basal septal wall thickness | 0.48 |
| 19 | Pulse pressure | 0.56 | Plasmin-antiplasmin complex | 0.49 |
| 20 | Descending aortic distensibility | 0.56 | End-diastolic basal inferior wall thickness | 0.49 |
| Rank | Heart failure | RVI | Atrial fibrillation | RVI |
| 1 | NT-proBNP | 0.00 | NT-proBNP | 0.00 |
| 2 | Tissue necrosis factor-$\alpha$ soluble receptor | 0.07 | Coronary Artery Calcium score | 0.23 |
| 3 | Coronary Artery Calcium score | 0.07 | Age | 0.27 |
| 4 | End-systolic left ventricular volume | 0.10 | Creatinine | 0.31 |
| 5 | Cardiac troponin-T | 0.20 | Ankle-brachial index | 0.36 |
| 6 | End-diastolic left ventricular volume | 0.21 | Interleukin-2 soluble receptor | 0.39 |
| 7 | Left ventricular ejection fraction | 0.24 | Tissue Necrosis Factor-$\alpha$ soluble receptor | 0.41 |
| 8 | QTc interval | 0.32 | Common carotid intima media thickness | 0.45 |
| 9 | QT index | 0.34 | R amplitude in lead $V_4$ | 0.53 |
| 10 | Interleukin-2 soluble receptor | 0.36 | STJ amplitude in lead $V_5$ | 0.53 |
| 11 | Waist/hip ratio | 0.38 | Internal carotid intima media thickness | 0.55 |
| 12 | Ankle-brachial index | 0.42 | Pulse pressure | 0.55 |
| 13 | PR interval | 0.45 | Estimate of overall heart rate variability | 0.56 |
| 14 | Creatinine | 0.45 | End-systolic basal lateral wall thickness | 0.56 |
| 15 | Pulse pressure | 0.46 | End-systolic midventricular anterior wall thickness | 0.56 |
| 16 | End-diastolic left ventricular mass | 0.47 | Heart rate | 0.57 |
| 17 | Estimate of overall heart rate variability | 0.51 | QRS Axis (degrees) | 0.57 |
| 18 | T amplitude in lead $V_1$ | 0.51 | Cardiac troponin-T | 0.57 |
| 19 | Minnesota code 1 score: V lead group | 0.51 | Total left atrial ejection fraction | 0.57 |
| 20 | Minnesota code 1 score: F lead group | 0.52 | Pack-years of smoking | 0.58 |

The relative variable importance (RVI) of each variable can be assessed using the normalized minimal depth of the maximal subtree (Figure 3). The normalized RVI values vary from 0 (most important) to 1 (least important). CVD indicates cardiovascular disease; LV, left ventricular; MESA, Multi-Ethnic Study of Atherosclerosis; and NT-proBNP, N-Terminal Pro-B-Type Natriuretic Peptide.

Brier score (BS).[8] Higher C-index and lower BS indicate better prediction performance. C-index and BS for nested models generated using subsets of predictors (chosen based on increasing variable importance) were calculated to assess problems of overfitting. We also compared the results of RSF techniques to published risk scores.[9–11] When the models failed to converge, no results were reported and this was considered the worst possible outcome.

Data analysis was performed using R software, using publically available libraries for Cox-PHM,[12] LASSO-Cox,[6] AIC-Cox,[5] and RSF methods.[13]

## Results

A total of 6814 participants were included. The average age was 62 years with 53% women, 28% black, 38% white, 12% Chinese American, and 22% Hispanic. Thirteen percent of the participants were diabetic, 45% classified as hypertensive based on Joint National Committee VI criteria, and 50% were current or former smokers. Over a median of 11.2 years (interquartile range, 10.6–11.7 years), MESA identified 831 all-cause deaths, 710 cardiovascular events (CVD) including 498 CHD events among which 229 were nonfatal myocardial infarctions, 200 strokes, 259 cases of incident HF, and 317 incident AF events (Table 2).

### Predictors by Outcomes

Table 3 shows the top-20 predictors using RSF for each of the outcomes ordered by the minimal depth of maximal subtree. These were the predictors used for the RSF-20, Cox-20, LASSO-20, and AIC-20 models. Figure 2 shows Lowess plots of the 12-year survival probability calculated from the RSF method over the range of values for the top-5 variables.

Increasing age, perhaps reflecting duration of risk exposure, was the most important predictor of all-cause death.

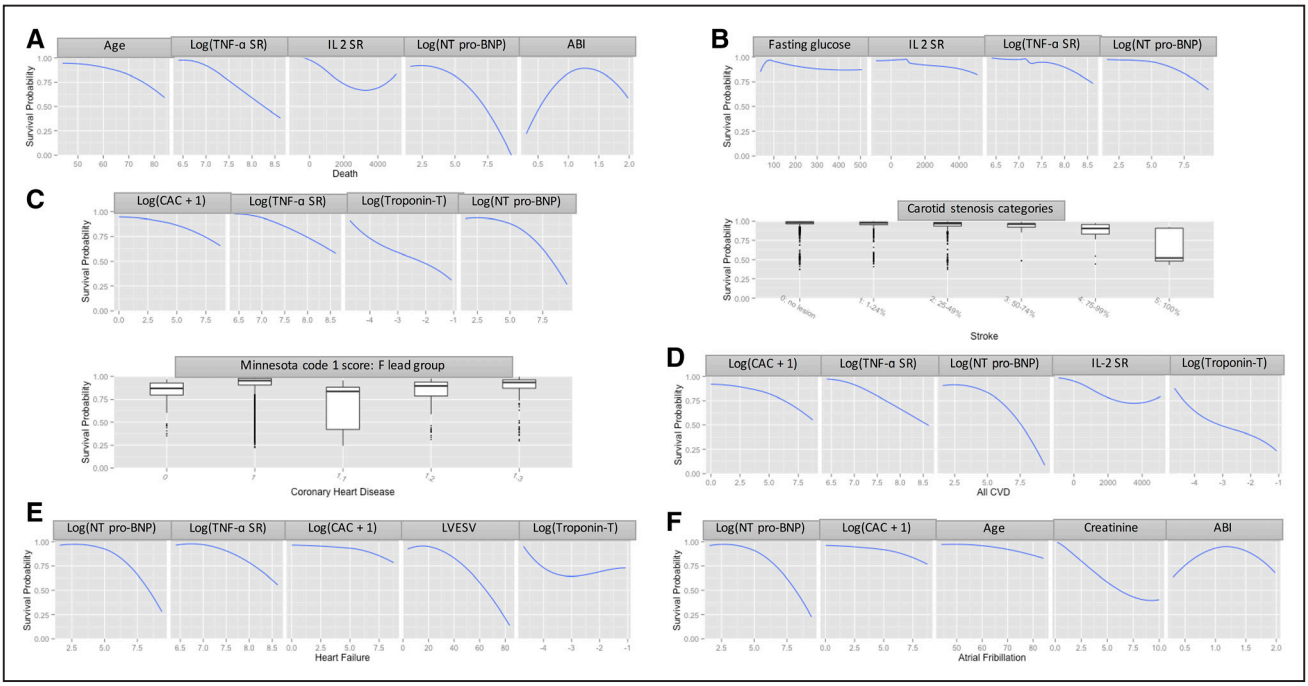


**Figure 2. Plots showing Lowess curves (for continuous variables) and box plots (for categorical variables) of the survival probability vs variable values for the top-5 predictors for each of the outcomes at 12 y.** The *y* axis represents survival probability calculated from the RSF-20 algorithm (range: 0–1). The *x* axis spans the range (or categories) of the variable of interest. Units for each variable: NT-proBNP, pg/mL; TNF-α SR, pg/mL; IL-2 SR, pg/mL; CAC, Agatston units, cardiac troponin-T, ng/mL; ABI, ratio; age, y; and fasting glucose, mg/dL. ABI indicates ankle-brachial index; CAC, Coronary Artery Calcium score; CVD, cardiovascular disease; IL-2 SR, interleukin-2 soluble receptor; NT-proBNP, N-Terminal Pro-B-Type Natriuretic Peptide; RSF, random survival forest; TNF-α SR, tissue necrosis factor-α soluble receptor; and LVESV, left ventricle end-systolic volume.



Inflammation and immune response measured by increased IL (interleukin)-6, fibrinogen, homocysteine, TNF-α SR (tissue necrosis factor-α soluble receptor), and IL-2 SR levels and abnormal hemostasis measured by increased D-Dimer, plasmin-antiplasmin complex, and factor VIII levels were among the top-20 markers of all-cause death underlining the role of inflammation and thrombosis as common pathways for chronic diseases leading to death. Similarly, cardiac stress measured by increased NT-proBNP (N-Terminal Pro-B-Type Natriuretic Peptide) levels and myocyte damage by increased cardiac troponin-T levels were among the top predictive markers of death reflecting the role of cardiac failure on mortality. In addition to biomarkers that featured so prominently, low and high values of ABI, increased carotid intima media thickness, increased Coronary Artery Calcium (CAC) score, aortic dimension, and distensibility also showed lower survival probability, indicating the importance of atherosclerosis and vascular abnormalities to mortality in the general population. Importantly, also, economic status/income was among the top-20 markers of all-cause death in MESA highlighting the potential power of inequality as a mortality risk factor even when one considers the theoretical distance between such a risk factor and death in the accepted causation chain for most chronic degenerative diseases.

Increased fasting glucose levels were the most important risk factor for stroke, whereas high blood pressure, a known stroke risk, and age also featured in the top-20 list. Measures of atherosclerosis (with carotid stenosis being most important) and inflammation were also top-5 predictors of stroke.

Expectedly, a composite of atherosclerosis measures (low and high ABI, increased carotid intima media thickness, decreased aortic distensibility) was among the most important predictors of CHD which represents a subset of CVD events, with CAC being by far the most important, reflecting the specific influence of coronary atherosclerosis. Increased LV regional wall thickness (myocardial hypertrophy), decreased ejection fraction (myocardial function), and increased aortic cross-sectional area (aortic dilatation), as well as biomarkers of abnormal hemostasis, inflammation, myocardial stress, and damage, featured among the other top predictors of CHD. Importantly, ECG LV hypertrophy and major Q-wave and repolarization abnormalities were markers of CHD and CVD events in MESA. In addition, among traditional risk factors, pack-years of smoking and pulse pressure were among the top predictors for CHD, whereas systolic blood pressure and age were among the top predictors for CVD.

For incident HF as the end point, cardiac chamber stress (increased LV volume and increased NT-proBNP levels) and decreased LV function from magnetic resonance imaging were the most important markers. LV hypertrophy on ECG, a lengthened QT interval indicating increased risk for tachyarrhythmias, increased creatinine levels, increased vascular stiffness, atherosclerosis as measured by CAC and ABI, and inflammation were also among the top predictors for HF. Increased pulse pressure and increased waist/hip ratio were also among the top risk factors for incident HF reflecting the role of obesity and hypertension on incident HF development.

**Table 4. The Number of Variables and the Performance (Concordance Index and Brier Score) for Each of the Models Tested and for the Risk Scores at the End of Follow-Up**

| | DTH | STRK | CHD | CVD | HF | AF |
|---|---|---|---|---|---|---|
| No. of variables | | | | | | |
| RSF with all covariates | 735 | 735 | 735 | 735 | 735 | 735 |
| RSF with top-20 covariates | 20 | 20 | 20 | 20 | 20 | 20 |
| AIC-Cox with forward selection | 13 | 9 | 5 | 6 | 5 | 6 |
| Cox with top-20 RSF covariates | 20 | 20 | 20 | 20 | 20 | 20 |
| LASSO-Cox with top-20 RSF covariates | 19 | 17 | 19 | 19 | 10 | 15 |
| AIC-Cox backward selection with top-20 RSF covariates | 16 | 12 | 13 | 13 | 11 | 12 |
| Concordance index at 12 y | | | | | | |
| RSF with all covariates | 0.86 | 0.77 | 0.81 | 0.81 | 0.84 | 0.82 |
| RSF with top-20 covariates | 0.84 | 0.75 | 0.80 | 0.80 | 0.84 | 0.75 |
| AIC-Cox with forward selection | 0.78 | 0.70 | 0.74 | 0.74 | 0.78 | 0.79 |
| Cox with top-20 RSF covariates | 0.80 | 0.66 | 0.75 | 0.76 | 0.81 | 0.78 |
| LASSO-Cox with top-20 RSF covariates | 0.80 | 0.67 | 0.75 | 0.76 | 0.82 | 0.78 |
| AIC-Cox Backward Selection with top-20 RSF covariates | 0.80 | 0.68 | 0.75 | 0.76 | 0.80 | 0.78 |
| Brier score at 12 y | | | | | | |
| RSF with all covariates | 0.083 | 0.031 | 0.067 | 0.083 | 0.035 | 0.039 |
| RSF with top-20 covariates | 0.076 | 0.030 | 0.065 | 0.079 | 0.033 | 0.038 |
| AIC-Cox with forward selection | 0.088 | 0.032 | 0.069 | 0.087 | 0.035 | 0.045 |
| Cox with top-20 RSF covariates | 0.086 | 0.031 | 0.070 | 0.087 | 0.035 | 0.037 |
| LASSO-Cox with top-20 RSF covariates | 0.086 | 0.031 | 0.070 | 0.087 | 0.033 | 0.038 |
| AIC-Cox Backward Selection with top-20 RSF covariates | 0.086 | 0.031 | 0.069 | 0.087 | 0.035 | 0.038 |

AF indicates atrial fibrillation; AIC-Cox, Akaike Information Criteria for Cox regression; CHD, coronary heart disease; CVD, cardiovascular disease; DTH, death; HF, heart failure; LASSO-Cox, least absolute shrinkage and selection operator for Cox regression; RSF, random survival forest; and STRK, stroke.

For incident AF as the end point, inflammation, higher levels of creatinine, atherosclerosis (CAC and ABI), and repolarization abnormalities were the most important markers. Decreased left atrial function and increased age and pulse pressure were also among the top risk factors for AF development.

## Predictors Across Outcomes

In general, variables from imaging markers, ABI, and serum biomarkers were of intermediate-to-high prediction importance, whereas questionnaires and medication exposures were of lower importance. Components of ECG related to ST segments were of intermediate importance, whereas other ECG indices ranged from low-to-intermediate prediction importance. As illustrated in Online Figure III, just the first 5 to 6 variables listed by the RSF algorithm produced C-indices >0.75 for CVD, CHD, incident HF, and AF prediction reflecting the importance of NT-proBNP (HF and AF) and CAC (CHD and CVD). Prediction of all-cause mortality and stroke with a C-index >0.75 required a larger group of variables.

## Comparison of Prediction Models

Table 4 shows the C-index and the BS for the 8 tested models using the test data sets. The standard Cox, LASSO-Cox, and AIC-Cox methods failed to converge when all the 735 variables were included, and hence BS and C-index could not be calculated. As shown in Online Figure III, using the nested models, <20 variables were necessary to obtain a stable and high C-index for the RSF method. Addition of more variables into the model beyond 30 resulted in minimal improvement of the C-index, if any. Figure 3 shows variable importance measured using the minimum depth of the maximal subtree, for each of the 735 variables used in analysis. Lower values correspond to greater prediction importance.

For all outcomes of interest, the RSF model with all 735 covariates showed a very high C-index and low BS. The RSF-20 model was comparable and even outperformed the RSF model with all covariates in some cases. Both the RSF models outperformed the AIC-Cox with forward stepwise across all end points with higher C-index and lower BS. The use of RSF for variable selection with top-20 (RSF-20) covariates and subsequent application of LASSO-Cox and AIC-Cox with backward stepwise resulted in fewer variables selected into the final models for most of the outcomes. The C-indices from these standard Cox, LASSO-Cox, and AIC-Cox with backward stepwise models were high and the BS low in general, and very similar to that of the RSF-20 model. Figure 4A through 4F show the C-index values over time for all the models (models that did not converge are not shown). In general, the C-index values were higher for prediction of short-term when compared with long-term events.

The models from machine learning that included biomarkers and measures of subclinical disease were, as expected, better than the ACC/AHA ASCVD (American College of Cardiology/American Heart Association Atherosclerotic Cardiovascular Disease; C-index: 0.73, BS: 0.11) and the Framingham (C-index: 0.73, BS: 0.089) risk scores for incident CVD prediction (Table 4). The performance of the RSF-20 model for incident CHD prediction was better than the Framingham CHD risk score (C-index: 0.69, BS: 0.072) with a higher C-index and lower BS.

When comparing the population-specific risk scores, the RSF-20 model from machine learning that included biomarkers and measures of subclinical disease was better than the MESA CHD Risk score (C-index: 0.79, BS: 0.074) for incident CHD prediction[14] (Table 4). The performance of the RSF-20 model for incident CHF prediction was better than


**Figure 3. Plots showing the variable importance for each of the 735 variables used in analysis.** The color of the dots represents the category or type of measurement. The legend on the right provides the phenotype category ordered from left-to-right on the individual plots. The variable importance is measured using the minimum depth of the maximal subtree, with lower values representing greater importance of corresponding variable. ABI indicates ankle-brachial index; CAC, Coronary Artery Calcium score; IL-2 SR, interleukin-2 soluble receptor; IMT, intima media thickness; NT-proBNP, N-Terminal Pro-B-Type Natriuretic Peptide; SBP, systolic blood pressure; and TNF-α SR, tissue necrosis factor-α soluble receptor.

the MESA HF[15] risk score (C-index: 0.80, BS: 0.038) with a higher C-index and lower BS.

## Discussion

The results of this study suggest that machine learning methods are well suited for meaningful risk prediction in extensively phenotyped large-scale epidemiological studies. The RSF-based method of risk prediction provided better event prediction over standard risk scores. RSF-based methods of variable selection followed by Cox regression methods also allowed for improved prediction of outcomes, without problems of overfitting and nonconvergence while accounting for nonlinearities. The results also suggest the importance of deep phenotyping using subclinical markers defined by imaging, electrocardiographic, and blood biochemistry, as revealed by their prominent presence on the lists of top-20 predictors, for CVD event prediction.

This work is unique by demonstrating patterns of predictors that vary for specific disease outcomes. While age, inflammation, cardiac stress, and vascular disease dominate the prediction of death in the MESA study, impaired glucose metabolism and hypertension lead in the prediction of stroke, and subclinical atherosclerosis markers occupy center stage in forecasting overall cardiovascular events—be they limited to the heart (CHD) or involving the systemic circulation. For incident HF and AF, a combination of markers reflecting increased cardiac chamber stress coupled with electric dysregulation is at the forefront of potential outcome determinants.

Another important pattern of findings from this investigation was the underrepresentation of certain traditional cardiovascular risk factors such as sex, race/ethnicity, and therapy exposure (medication use) among the top predictors of disease outcome. Important exceptions to such trend were the presence of socioeconomic status as one of the top predictors of death and the role of hypertension as a top predictor of stroke, CVD, CHD, and incident HF and AF. The lower than expected representation of traditional risk factors may stem from the fact that they are fundamental to the genesis, maintenance, and progression of CVDs; therefore, they are intrinsic components of other phenotypes, particularly subclinical phenotypes that are more distal to disease initiation but closer to adverse outcomes. Even though some of these risk factors did not feature in the top 20, they remain crucial to medical practice, particularly, disease prevention.

### Machine Learning and Deep Phenotyping

The application and use of machine learning tools for CVD are still controversial.[10] This is so, even as there has been an increase in use of imaging tests, ECG exams, and laboratory tests in recent years.[16] In most cases, even when multiple markers are acquired, not all are used for diagnosis.[17] For example, regional function measures from imaging, large portions of biomarker panels, or ECG signals are frequently ignored by many clinicians. As we move into the age of precision medicine, understanding the use of phenotypic data and methods to analyze already acquired information is of paramount importance. Machine learning methods, and RSF particularly, have been used before for CVD risk prediction.[18–27] In this study, we were able to use deeply phenotyped data to predict outcomes in a population study that accounts for time to event.[28] The

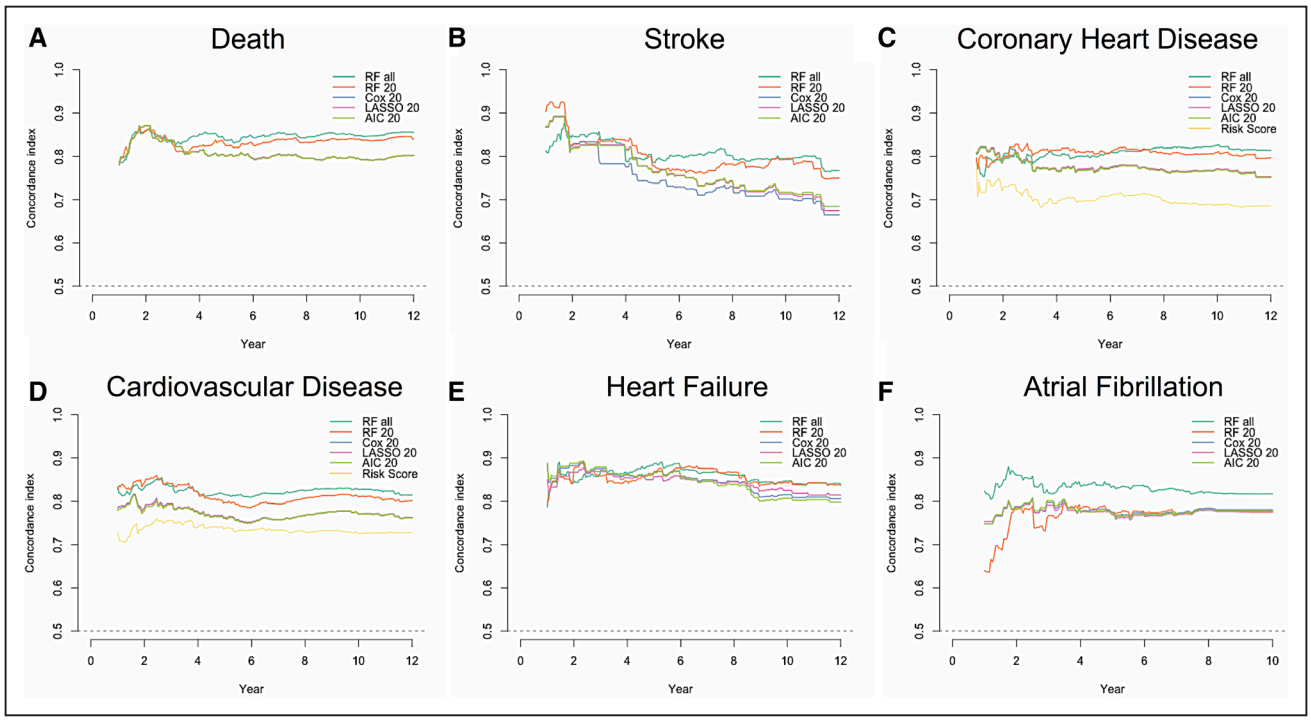


**Figure 4. The concordance index for each of the models tested over time.** The full models (models with all 735 variables) did not converge for the LASSO-Cox, AIC-Cox and the Cox Proportional Hazard regression model (Cox-PHM) models and hence are not shown here. The prediction ability of conventional risk scores for heart failure (MESA HF risk score), cardiovascular disease (AHA/ASCVD risk score), and coronary heart disease (Framingham CHD risk score) are also shown (yellow curve). In general, the C-index for all variables decreased over time. AIC-Cox, Akaike Information Criterion for Cox regression; CHD, coronary heart disease; LASSO-Cox, least absolute shrinkage and selection operator for Cox regression; MESA HF, Multi-Ethnic Study of Atherosclerosis heart failure; and RSF, random survival forest.



added advantage is that these methods can be extended and refined, regularly, with new data. They also account for nonlinearity in relationships (Figure 2), for example, both high and low values of ABI (as previously shown) were predictive of incidence of CVD.

This work also confirms the influence of certain markers and risk factors on cardiovascular events. From this analysis, the importance of markers, heretofore underestimated, such as TNF-α SR and IL-2 SR, come to fore with machine learning. In this regard, machine learning opens the possibility of discovering new relationships that are not hypothesis driven and without prior assumptions. Identifying effective disease markers and discovering unknown mechanisms may be of benefit for effective screening strategies and suggest specific targets for risk reduction. Yet another advantage of this technique is the ability to recognize the best predictors within a domain (questionnaires, imaging, etc) and their importance with respect to predictors from other domains. This approach to biomarker identification may be of particular benefit in intermediate-risk groups where underlying subclinical risk is not apparent in traditional cardiovascular risk factors.

## Methodological Considerations

In this study, We used the minimum depth of the maximal subtree as the main measure of variable importance because of prior research that showed inherent advantages to using this over permutation testing.[29] Although there are other methods to do the same, the change in Gini index, for example, they may not use survival data and hence may not be the best method in the case of survival analysis as ours. However, we provide the top-10 variables from both the Gini index (12-year cutoff) and permutation testing in the Online Data Supplement.

While deep phenotyping might help in biomarker discovery, it is seen from Online Figure III that far fewer than the measured variables are necessary for obtaining a high C-index. To this end, RSF methods may help in identifying important variables. We have shown the top-20 variables, as well as the C-index and BS using just the top-20 models. It is plausible that the optimal number of variables is <20. However, formal methods need to be developed with consideration to cost, appropriateness, ease of access, and reproducibility of measurements for a more judicious approach to variable selection for event prediction.

MESA, designed to study progression of subclinical disease to manifest symptoms and outcomes, was performed in a middle-aged population free of CVD at baseline. Therefore, results may not generalize to other study populations. We did not include genetic data; the identification of the phenome–genome interaction and assessment of their combined prediction ability may potentially improve our findings.[30] Although phenomorphing (longitudinal covariate data) and risk prediction is of interest, it is out of the scope of this study. Although this study identifies top predictors as a method of biomarker discovery, further work is required including validation in other populations, as the training and test data sets were drawn from within the MESA study population.

## Conclusions

In an extensively phenotyped population free of CVD at baseline, using random forests, we show efficient cardiovascular risk prediction for specific outcomes including death, stroke, cardiovascular events, incident HF, and AF. Inflammation, subclinical atherosclerosis, myocardial damage, and cardiac chamber stress were among the most important predictors

across all outcomes. We provide a framework for big data applications to obtain meaningful risk prediction, biomarker identification, and generate data-driven hypotheses.

## Acknowledgments

The information contained herein (for the MESA [Multi-Ethnic Study of Atherosclerosis] Columbia Field Center) was derived in part from data provided by the Bureau of Vital Statistics, New York City Department of Health and Mental Hygiene. We thank the other investigators, the staff, and the participants of the MESA study for their valuable contributions. The MESA protocol, including information about the populations from which recruitment occurred, detailed exclusion criteria, investigators, and other information, is available at www.mesa-nhlbi.org. A full list of participating MESA investigators and institutions can also be found.

## Sources of Funding

This research was supported by contracts N01-HC-95159, N01-HC-95160, N01-HC-95161, N01-HC-95162, N01-HC-95163, N01-HC-95164, N01-HC-95165, N01-HC-95166, N01-HC-95167, N01-HC-95168 and N01-HC-95169 from the National Heart, Lung, and Blood Institute and by grants UL1-TR-000040 and UL1-TR-001079 from the National Center for Research Resources.

## Disclosures

None.

## References

1. Lloyd-Jones DM. Cardiovascular risk prediction: basic concepts, current status, and future directions. *Circulation*. 2010;121:1768–1777. doi: 10.1161/CIRCULATIONAHA.109.849166.
2. Wong ND. Epidemiological studies of CHD and the evolution of preventive cardiology. *Nat Rev Cardiol*. 2014;11:276–289. doi: 10.1038/nrcardio.2014.26.
3. Gorodeski EZ, Ishwaran H, Kogalur UB, Blackstone EH, Hsich E, Zhang Z-M, Vitolins MZ, Manson JE, Curb JD, Martin LW. Use of hundreds of electrocardiographic biomarkers for prediction of mortality in postmenopausal women: the women's health initiative. *Circ Cardiovasc Qual Outcomes*. 2011;4:521–532.
4. Bild DE, Bluemke DA, Burke GL, et al. Multi-Ethnic Study of Atherosclerosis: objectives and design. *Am J Epidemiol*. 2002;156:871–881.
5. Akaike H. Likelihood of a model and information criteria. *J Econom*. 1981;16:3–14.
6. Tibshirani R. The lasso method for variable selection in the Cox model. *Stat Med*. 1997;16:385–395.
7. Harrell FE, Jr, Califf RM, Pryor DB, Lee KL, Rosati RA. Evaluating the yield of medical tests. *JAMA*. 1982;247:2543–2546.
8. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78:1–3.
9. D'Agostino RB, Sr, Vasan RS, Pencina MJ, Wolf PA, Cobain M, Massaro JM, Kannel WB. General cardiovascular risk profile for use in primary care: the Framingham Heart Study. *Circulation*. 2008;117:743–753. doi: 10.1161/CIRCULATIONAHA.107.699579.
10. Goff DC, Jr, Lloyd-Jones DM, Bennett G, et al; American College of Cardiology/American Heart Association Task Force on Practice Guidelines. 2013 ACC/AHA guideline on the assessment of cardiovascular risk: a report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines. *J Am Coll Cardiol*. 2014;63:2935–2959. doi: 10.1016/j.jacc.2013.11.005.
11. Lloyd-Jones DM, Wilson PW, Larson MG, Beiser A, Leip EP, D'Agostino RB, Levy D. Framingham risk score and prediction of lifetime risk for coronary heart disease. *Am J Cardiol*. 2004;94:20–24. doi: 10.1016/j.amjcard.2004.03.023.
12. Cox DR. Regression models and life-tables. *J R Stat Soc Series B*. 1972;34:187–220.
13. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS. Random survival forests. *Annals Appl Stat*. 2008;2:841–860.
14. McClelland RL, Jorgensen NW, Budoff M, et al. 10-year coronary heart disease risk prediction using coronary artery calcium and traditional risk factors: derivation in the MESA (Multi-Ethnic Study of Atherosclerosis) with validation in the HNR (Heinz Nixdorf Recall) Study and the DHS (Dallas Heart Study). *J Am Coll Cardiol*. 2015;66:1643–1653. doi: 10.1016/j.jacc.2015.08.035.
15. Chahal H, Bluemke DA, Wu CO, McClelland R, Liu K, Shea SJ, Burke G, Balfour P, Herrington D, Shi P, Post W, Olson J, Watson KE, Folsom AR, Lima JA. Heart failure risk prediction in the Multi-Ethnic Study of Atherosclerosis. *Heart*. 2015;101:58–64. doi: 10.1136/heartjnl-2014-305697.
16. Andrus BW, Welch HG. Medicare services provided by cardiologists in the United States: 1999-2008. *Circ Cardiovasc Qual Outcomes*. 2012;5:31–36. doi: 10.1161/CIRCOUTCOMES.111.961813.
17. Lanktree MB, Hassell RG, Lahiry P, Hegele RA. Phenomics: expanding the role of clinical evaluation in genomic studies. *J Investig Med*. 2010;58:700–706. doi: 10.231/JIM.0b013e3181d844f7.
18. Deo RC. Machine learning in medicine. *Circulation*. 2015;132:1920–1930. doi: 10.1161/CIRCULATIONAHA.115.001593.
19. Ishwaran H, Blackstone EH, Pothier CE, Lauer MS. Relative risk forests for exercise heart rate recovery as a predictor of mortality. *J Am Stat Assoc*. 2004;99:591–600.
20. Inuzuka R, Diller GP, Borgia F, Benson L, Tay EL, Alonso-Gonzalez R, Silva M, Charalambides M, Swan L, Dimopoulos K, Gatzoulis MA. Comprehensive use of cardiopulmonary exercise testing identifies adults with congenital heart disease at increased mortality risk in the medium term. *Circulation*. 2012;125:250–259. doi: 10.1161/CIRCULATIONAHA.111.058719.
21. Hsich E, Gorodeski EZ, Blackstone EH, Ishwaran H, Lauer MS. Identifying important risk factors for survival in patient with systolic heart failure using random survival forests. *Circ Cardiovasc Qual Outcomes*. 2011;4:39–45. doi: 10.1161/CIRCOUTCOMES.110.939371.
22. Sitar-tăut A, Zdrenghea D, Pop D, Sitar-tăut D. Using machine learning algorithms in cardiovascular disease risk evaluation. *Age*. 2009;1:4.
23. Colombet I, Ruelland A, Chatellier G, Gueyffier F, Degoulet P, Jaulent M. Models to predict cardiovascular risk: comparison of CART, multilayer perceptron and logistic regression. *Proc AMIA Symp*. 2000:156.
24. Park GM, Han S, Kim SH, et al. Model for assessing cardiovascular risk in a Korean population. *Circ Cardiovasc Qual Outcomes*. 2014;7:944–951. doi: 10.1161/CIRCOUTCOMES.114.001305.
25. Shardell MD, Alley DE, Hicks GE, El-Kamary SS, Miller RR, Semba RD, Ferrucci L. Low-serum carotenoid concentrations and carotenoid interactions predict mortality in US adults: the Third National Health and Nutrition Examination Survey. *Nutr Res*. 2011;31:178–189. doi: 10.1016/j.nutres.2011.03.003.
26. Rizza S, Copetti M, Rossi C, Cianfarani MA, Zucchelli M, Luzi A, Pecchioli C, Porzio O, Di Cola G, Urbani A, Pellegrini F, Federici M. Metabolomics signature improves the prediction of cardiovascular events in elderly subjects. *Atherosclerosis*. 2014;232:260–264. doi: 10.1016/j.atherosclerosis.2013.10.029.
27. Rebholz CM, Grams ME, Matsushita K, Inker LA, Foster MC, Levey AS, Selvin E, Coresh J. Change in multiple filtration markers and subsequent risk of cardiovascular disease and mortality. *Clin J Am Soc Nephrol*. 2015;10:941–948.
28. Yoo I, Alafaireet P, Marinov M, Pena-Hernandez K, Gopidi R, Chang JF, Hua L. Data mining in healthcare and biomedicine: a survey of the literature. *J Med Syst*. 2012;36:2431–2448. doi: 10.1007/s10916-011-9710-5.
29. Ishwaran H, Kogalur UB, Gorodeski EZ, Minn AJ, Lauer MS. High-dimensional variable selection for survival data. *J Am Stat Assoc*. 2010;105:205–217.
30. Benjamin I, Brown N, Burke G, Correa A, Houser SR, Jones DW, Loscalzo J, Vasan RS, Whitman GR. American Heart Association Cardiovascular Genome-Phenome Study: foundational basis and program. *Circulation*. 2015;131:100–112. doi: 10.1161/CIRCULATIONAHA.114.014190.