Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score

Matthew W. Segar,[1] Muthiah Vaduganathan,[2] Kershaw V. Patel,[1] Darren K. McGuire,[1] Javed Butler,[3] Gregg C. Fonarow,[4] Mujeeb Basit,[1] Vaishnavi Kannan,[5] Justin L. Grodin,[1] Brendan Everett,[2] Duwayne Willett,[1] Jarett Berry,[1] and Ambarish Pandey[1]

Diabetes Care 2019;42:2298–2306 | https://doi.org/10.2337/dc19-0587

OBJECTIVE

To develop and validate a novel, machine learning–derived model to predict the risk of heart failure (HF) among patients with type 2 diabetes mellitus (T2DM).

RESEARCH DESIGN AND METHODS

Using data from 8,756 patients free at baseline of HF, with <10% missing data, and enrolled in the Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial, we used random survival forest (RSF) methods, a nonparametric decision tree machine learning approach, to identify predictors of incident HF. The RSF model was externally validated in a cohort of individuals with T2DM using the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT).

RESULTS

Over a median follow-up of 4.9 years, 319 patients (3.6%) developed incident HF. The RSF models demonstrated better discrimination than the best performing Cox-based method (C-index 0.77 [95% CI 0.75–0.80] vs. 0.73 [0.70–0.76] respectively) and had acceptable calibration (Hosmer-Lemeshow statistic $\chi^2 = 9.63$, $P = 0.29$) in the internal validation data set. From the identified predictors, an integer-based risk score for 5-year HF incidence was created: the WATCH-DM (Weight [BMI], Age, hyperTension, Creatinine, HDL-C, Diabetes control [fasting plasma glucose], QRS Duration, MI, and CABG) risk score. Each 1-unit increment in the risk score was associated with a 24% higher relative risk of HF within 5 years. The cumulative 5-year incidence of HF increased in a graded fashion from 1.1% in quintile 1 (WATCH-DM score ≤7) to 17.4% in quintile 5 (WATCH-DM score ≥14). In the external validation cohort, the RSF-based risk prediction model and the WATCH-DM risk score performed well with good discrimination (C-index = 0.74 and 0.70, respectively), acceptable calibration (P ≥0.20 for both), and broad risk stratification (5-year HF risk range from 2.5 to 18.7% across quintiles 1–5).

CONCLUSIONS

We developed and validated a novel, machine learning–derived risk score that integrates readily available clinical, laboratory, and electrocardiographic variables to predict the risk of HF among outpatients with T2DM.

[1]Division of Cardiology, Department of Internal Medicine, University of Texas Southwestern Medical Center, Dallas, TX
[2]Brigham and Women's Hospital Heart and Vascular Center, Department of Medicine, Harvard Medical School, Boston, MA
[3]Department of Medicine, University of Mississippi Medical Center, Jackson, MS
[4]Division of Cardiology, Ahmanson-UCLA Cardiomyopathy Center, Ronald Reagan UCLA Medical Center, Los Angeles, CA
[5]Department of Health System Information Resources (Clinical Informatics), University of Texas Southwestern Medical Center, Dallas, TX

Corresponding author: Ambarish Pandey, ambarish.pandey@utsouthwestern.edu

Received 22 March 2019 and accepted 5 August 2019

Clinical trial reg. no. NCT00000620, clinicaltrials.gov

This article contains Supplementary Data online at http://care.diabetesjournals.org/lookup/suppl/doi:10.2337/dc19-0587/-/DC1.

M.W.S. and M.V. contributed equally as co–first authors of this manuscript.

Prevention of atherosclerotic cardiovascular events has been a major goal of therapeutic approaches in type 2 diabetes mellitus (T2DM) clinical practice guidelines (1). Similarly, in response to the 2008 U.S. Food and Drug Administration and the European Medicines Agency Committee for Medicinal Products for Human Use guidance to industry for the development of antihyerglycemic therapies for the treatment of T2DM, exclusion of significant risk of composite major adverse cardiovascular events has been the focus of cardiovascular outcome trial programs of novel and established antihyperglycemic therapies over the last decade (2). There has been comparatively less attention toward preventing heart failure (HF), despite its frequency as an initial presentation of cardiovascular disease in T2DM (3,4). Patients with T2DM with adequate control of major risk factors within target ranges appear to have risk of atherosclerotic cardiovascular disease (ASCVD) comparable with that of the general population; however, even patients with T2DM with no additional risk factors face a substantial residual risk of hospitalization for HF (5). Unfortunately, these patients with T2DM complicated by HF experience particularly high rates of mortality (5). As such, the prevention of HF in T2DM is of upmost importance.

The sodium–glucose cotransporter 2 inhibitors (SGLT2i), a class of antihyperglycemic therapies, have been shown to reduce risk of hospitalization for HF in at-risk patients with T2DM (6–9) and are now supported as second-line therapies (after metformin) in patients with T2DM and prevalent ASCVD or CKD (7,10,11). However, limited guidance is available regarding targeted introduction of these therapies in patients with T2DM at heightened risk of HF, independent of ASCVD considerations. Importantly, current risk prediction models with traditional risk factors incompletely capture HF risk in T2DM (5). As such, we hypothesized that a novel approach leveraging machine learning methods that can handle multidimensional data may offer superior risk prediction abilities. We aimed to develop a novel risk prediction model and integer-based score for incident HF in patients with T2DM at high cardiovascular risk enrolled in the ACCORD (Action to Control Cardiovascular Risk in Diabetes) trial and externally validate the findings in the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) database.

RESEARCH DESIGN AND METHODS

The ACCORD Trial

The detailed protocol, study design (12), and primary results (13) of ACCORD have previously been reported. Briefly, ACCORD was conducted in 77 centers across the U.S. and Canada. The trial included a total of 10,251 men and women, aged 40–79 years, with T2DM and inadequate glycemic control (glycated hemoglobin [HbA_{1c}] $\geq$7.5% [58 mmol/mol]). All participants had established ASCVD or were 55–79 years of age with anatomic evidence of atherosclerosis, albuminuria, left ventricular hypertrophy, or two or more other cardiovascular risk factors (current smoking, hyperlipidemia, hypertension, or obesity). In ACCORD, randomization to a more intensive glycemic control group (target HbA_{1c} $<$6% [42 mmol/mol]) versus usual care did not affect risk of major adverse cardiovascular events compared with standard glycemic control (target HbA_{1c} 7–7.9% [53–63 mmol/mol]) (13). As more intensive glycemic control did not influence HF risk and as ACCORD was conducted prior to the availability of SGTL2i, patients randomized into both trial arms were included in the present risk modeling for incident HF, and randomized treatment assignment was included as a covariate in the risk prediction models. Participants with prevalent HF at baseline and those with $>$10% missing data were excluded.

Incident HF Events

The primary end point of interest for the current study was incident hospitalization or death due to HF, which was a prespecified secondary outcome that was prospectively captured and centrally adjudicated by two independent multidisciplinary reviewer physicians (general internists, cardiologists, and endocrinologists). Specific event definitions have been detailed previously (14). Hospitalization for HF required documented clinical and radiologic evidence of clinical HF and congestion. Death due to HF or cardiogenic shock was defined as a death with clinical, radiologic, or postmortem evidence of HF, in the absence of acute ischemic event. We compared descriptive statistics in patients who did and did not experience an incident HF event during follow-up.

Candidate Variables

In total, 147 covariates collected at baseline were considered as candidates for analysis. Covariates encompassed a range of domains including demographics, clinical variables, laboratory data, electrocardiographic parameters, baseline antihyperglycemic therapies, and treatment randomization (to intensive vs. standard glycemic control). Covariates with $>$10% missing data were excluded. We further removed variables with a correlation coefficient $>$0.7; however, no highly correlated covariates were identified. The present analysis included 109 predictor variables (Supplementary Table 1). Continuous variables that did not follow a normal distribution were log transformed. Missing values were imputed using a random forest imputation (15).

Model Development

Prediction model development consisted of two main stages: variable selection and relationship modeling (Supplementary Fig. 1). The different methods used in each of the stages and the resulting prediction models are described below.

Variable Selection

The variable selection stage reduces the number of variables used in the prediction model by evaluating changes in performance resulting from addition or removal of variables (16). We considered three variable selection methods, including stepwise backward selection, stepwise forward selection, and permutation-based random survival forest (RSF) selection. Stepwise backward selection removes variables sequentially according to their strength of association with the outcomes until the Akaike information criterion stops improving (i.e., is minimized) (16). Similarly, stepwise forward selection adds variables sequentially until the Akaike information criterion reaches a minimum value. The permutation-based selection was conducted using the variable importance (VIMP) metric of the RSF. For VIMP, a random subset of predictor variable values was permuted and the difference in prediction error between the observed and randomly permutated variables was calculated. A high VIMP suggests that misspecification worsens the predictive accuracy in the forest, while a low VIMP

suggests noise is more informative than the observed variable.

Relationship Modeling

For each of the variable selection methods, the relationship between the selected variables and the outcome of interest was assessed using traditional Cox proportional hazards (PH) models. The best performing variable selection method was further assessed using a machine learning–based method, RSFs. Briefly, an RSF is an ensemble classification method that determines a consensus prediction by averaging the results of many individual decision trees (17). Each individual tree is fitted using randomly selected data using a subset of the observations (18). Competing risks were modeled using a log-rank split rule (19).

Model Evaluation

For elucidation of the contribution of various clinical variables and model development strategies, the performance of the three models was compared as illustrated in Supplementary Fig. 1. For development and comparison of the models, the following procedure was repeated 1,000 times. First, the study cohort was randomly split into a development set (50%) and a validation set (50%). The three models were built using data from the development data set only. Imputation was performed separately on the development and internal validation data sets. Finally, the relative discrimination performance of the models against the validation data set was calculated using the Harrell concordance index (C-index). A C-index ranges from 0.5 (no better than chance) to 1.0 (perfect discrimination) and is analogous to the area under the receiver operating characteristic (ROC) curve for survival data. Performance was reported as mean and 95% CI from the 1,000 bootstrapped replicates. Improvement in discrimination between models was assessed by the DeLong test (20).

Integer-Based Risk Score Development

For improvement of the clinical utility of our risk prediction model, an integer-based score was developed to estimate the 5-year risk of incident HF using regression coefficients from the Cox PH model and an age-standardized points scoring system similar to the Framingham framework (21). Calibration of the model was evaluated by the Hosmer-Lemeshow statistic χ^2. Similarly, the final model was visualized graphically by comparing the observed probability with the predicted probability across 10 deciles of predicted risk (22). Finally, participants were further divided into five equally sized risk strata using the quintiles of the calculated risk score.

External Validation of the Risk Scores

The RSF risk prediction model and the integer-based WATCH-DM risk score were externally validated in a separate cohort of patients with baseline T2DM and free of HF from ALLHAT (23). ALLHAT was a randomized, double-blind, multicenter clinical trial designed to investigate whether treatment with a calcium channel blocker (amlodipine), an ACE inhibitor (lisinopril), or an α-adrenergic blocker (doxazosin) would reduce the incidence of fatal coronary heart disease or nonfatal myocardial infarction compared with treatment with a thiazide-type diuretic (chlorthalidone) (24). The study enrolled 42,418 participants aged ≥55 years with baseline hypertension and at least one additional coronary artery disease risk factor (including T2DM, hyperlipidemia, current cigarette smoking, left ventricular hypertrophy, previous myocardial infarction or stroke, or ASCVD). The participants of ALLHAT in the chlorthalidone, lisinopril, or amlodipine arms of the study with prevalent T2DM at baseline were considered for inclusion in the external validation analysis (N = 12,063). Given that the α-adrenergic blocker arm was terminated early due to increased incidence of major cardiovascular events, patients enrolled in that arm were not considered for the analysis.

The final external validation cohort included 10,819 participants (25.5% of ALLHAT) after further exclusion of participants with prevalent HF (N = 791) and missing data for the risk prediction model (N = 453). The outcome of interest for our analysis, as defined in ALLHAT, was new-onset HF. A subset of participants in ALLHAT also had available data on HF subtype, HF with preserved ejection fraction (HFpEF), and HF with reduced ejection fraction (HFrEF) at the time of HF diagnosis. All HF events were adjudicated using the hospitalization data in a centrally blinded manner (25).

Data Sharing Statement

All patients provided written informed consent to participate in ACCORD and ALLHAT, and study protocols were approved by local institutional review boards. Both ACCORD and ALLHAT were supported by the National Heart, Lung, and Blood Institute, and limited anonymized data are available by request to the National Institutes of Health Biologic Specimen and Data Repository Information Coordinating Center.

Analyses were computed using R 3.5.1 (R Foundation for Statistical Computing, Vienna, Austria). The RSF and stepwise-selection models were implemented using the randomForestSRC and MASS packages, respectively (26,27).

RESULTS

In ACCORD (n = 10,251), 492 (4.8%) had history of diagnosed HF at baseline and 1,003 (9.8%) had >10% missing data; our final study sample included 8,756 participants (85.4%). Overall, 319 (cumulative rate 3.6%) developed incident HF during a median follow-up of 4.9 years. Participants who developed HF were older, more commonly men, and had a higher BMI (mean 33.0 vs. 32.1 kg/m^2, $P < 0.01$). Patients who experienced incident HF events also had higher frequencies of prevalent ASCVD and had longer average durations of T2DM, hypertension, and hyperlipidemia (all $P < 0.05$) (Table 1).

Performance of Models With Different Variable Selection and Risk Prediction Methods

Of the three different methods for variable selection, each method selected a unique subset of the available candidate variables. The stepwise backward selection method identified 11 significant variables, the stepwise forward selection method identified 8 significant variables, and the RSF method identified 10 significant variables. The variables identified by each method are shown in Supplementary Table 2. A total of eight variables were common between all three methods (age, diastolic blood pressure, glycated hemoglobin, serum creatinine, HDL-C, T-wave axis, QRS duration, and history of myocardial infarction). T-wave axis, PR duration, and QTc duration were not included in the relationship modeling stage, as they were not routine clinical markers (T-wave axis), were clinically similar to other variables, or were not available in the validation cohort. Both fasting plasma glucose and HbA_{1c}

Table 1—Baseline characteristics of the study participants with versus without incident HF during the study period

	ACCORD patients (free from baseline HF)	No HF event during follow-up	Incident HF event during follow-up	*P*
n	8,756	8,437	319	
Demographics				
Female, *n* (%)	3,370 (38.5)	3,275 (38.8)	95 (29.8)	<0.01
Age (years), *n* (%)	62.7 (6.6)	62.6 (6.5)	65.3 (6.9)	<0.01
Race, *n* (%)				<0.01
Black	1,622 (18.5)	1,558 (18.5)	64 (20.1)	
Hispanic	658 (7.5)	643 (7.6)	15 (4.7)	
Other	987 (11.3)	967 (11.5)	20 (6.3)	
White	5,489 (62.7)	5,269 (62.5)	220 (69.0)	
BMI (kg/m^2), mean (SD)	32.1 (5.4)	32.1 (5.4)	33.0 (5.6)	<0.01
Waist circumference (cm), mean (SD)	106.6 (13.6)	106.4 (13.5)	110.8 (14.6)	<0.01
Current cigarette smoker, *n* (%)	1,078 (12.3)	1,031 (12.2)	47 (14.7)	0.21
Lives with one or more adults, *n* (%)	371 (4.2)	338 (4.0)	33 (10.3)	<0.01
Highest level of education, *n* (%)				
Less than high school graduate	1,260 (14.4)	1,191 (14.1)	69 (21.6)	<0.01
High school graduate or GED	2,300 (26.3)	2,223 (26.4)	77 (24.1)	
Some college or technical school	2,891 (33.0)	2,770 (32.9)	121 (37.9)	
College graduate	2,300 (26.3)	2,248 (26.7)	52 (16.3)	
Alcoholic drinks consumed weekly, mean (SD)	1.0 (2.7)	1.0 (2.7)	0.9 (2.7)	0.70
Vital signs at baseline, mean (SD)				
Systolic blood pressure (mmHg)	136.4 (16.9)	136.3 (16.9)	139.4 (18.4)	<0.01
Diastolic blood pressure (mmHg)	75.0 (10.5)	75.1 (10.5)	72.5 (11.6)	<0.01
Heart rate (bpm)	72.6 (11.7)	72.6 (11.7)	73.0 (12.2)	0.56
Medical history				
History of myocardial infarction, *n* (%)	1,237 (14.1)	1,138 (13.5)	99 (31.0)	<0.01
History of stroke, *n* (%)	502 (5.7)	472 (5.6)	30 (9.4)	0.01
History of angina, *n* (%)	928 (10.6)	880 (10.4)	48 (15.0)	0.01
History of coronary artery bypass graft surgery, *n* (%)	918 (10.5)	830 (9.8)	88 (27.6)	<0.01
History of percutaneous coronary intervention, *n* (%)	870 (10.1)	818 (9.8)	52 (18.1)	<0.01
History of other revascularization procedure, *n* (%)	325 (3.7)	297 (3.5)	28 (8.8)	<0.01
History of foot ulcer requiring antibiotics, *n* (%)	371 (4.2)	338 (4.0)	33 (10.3)	<0.01
Years of diabetes diagnosis, median (IQR)	9.0 (10.0)	9.0 (10.0)	11.0 (13.0)	<0.01
Years of hyperlipidemia diagnosis, median (IQR)	7.0 (12.0)	7.0 (11.0)	10.0 (15.0)	<0.01
Years of hypertension diagnosis, median (IQR)	4.0 (6.0)	4.0 (6.0)	5.0 (7.3)	0.05
Health Utilities Index Mark 3 score, mean (SD)*	0.5 (0.3)	0.5 (0.3)	0.4 (0.3)	<0.01
Health Utilities Index Mark 2 score, mean (SD)†	0.6 (0.2)	0.6 (0.2)	0.51 (0.2)	<0.01
Laboratory values, mean (SD)				
Glycated hemoglobin (mg/dL)	8.3 (1.1)	8.3 (1.1)	8.6 (1.1)	<0.01
Total cholesterol (mg/dL)	183.6 (41.7)	183.7 (41.6)	180.7 (44.0)	0.20
Triglycerides (mg/dL)	190.6 (149.8)	190.3 (150.8)	197.7 (118.9)	0.39
VLDL cholesterol (mg/dL)	36.6 (24.5)	36.6 (24.6)	38.1 (20.1)	0.27
LDL cholesterol (mg/dL)	105.1 (33.9)	105.2 (33.8)	104.3 (35.4)	0.63
HDL cholesterol (mg/dL)	41.9 (11.5)	42.0 (11.5)	38.3 (10.4)	<0.01
Fasting plasma glucose (mg/dL)	171.2 (55.9)	171.0 (55.6)	180.6 (63.1)	0.03
Alanine aminotransferase (mg/dL)	27.8 (16.7)	27.9 (16.8)	24.3 (10.9)	<0.01
Creatine kinase (mg/dL)	139.5 (127.4)	139.6 (128.0)	136.5 (109.6)	0.67
Potassium (mg/dL)	4.5 (0.5)	4.5 (0.5)	4.5 (0.5)	0.31
Serum creatinine (mg/dL)	0.9 (0.2)	0.9 (0.2)	1.0 (0.3)	<0.01
Estimated glomerular filtration rate (mL/min)	91.5 (27.3)	91.8 (27.4)	84.3 (23.5)	<0.01
Urinary albumin (mg/dL)	10.0 (36.2)	9.4 (34.8)	26.2 (61.2)	<0.01
Urinary creatinine (mg/dL)	124.6 (66.4)	124.7 (66.4)	121.2 (65.7)	0.36
Urinary albumin-to-creatinine ratio (mg/g)‡	95.1 (351.6)	88.9 (333.1)	254.7 (653.7)	<0.01
Electrocardiogram, mean (SD)				
PR duration (ms)	164.4 (31.5)	164.3 (30.9)	168.2 (44.9)	0.03

Continued on p. 2302

Table 1—Continued

	ACCORD patients (free from baseline HF)	No HF event during follow-up	Incident HF event during follow-up	*P*
P-axis	48.1 (21.6)	48.0 (21.5)	51.7 (23.9)	<0.01
QRS-axis	12.9 (33.4)	13.2 (33.1)	6.9 (38.1)	<0.01
T-axis	44.2 (39.1)	43.3 (38.0)	69.1 (55.7)	<0.01
QRS duration (ms)	94.9 (16.6)	94.6 (16.2)	102.8 (22.2)	<0.01
Bazzett QTc calculated (ms)	420.7 (20.5)	420.4 (20.3)	429.4 (25.1)	<0.01
R-wave amplitude in aVL	622.3 (312.8)	621.1 (311.8)	652.5 (336.9)	0.08
S-wave amplitude in V3	836.8 (477.7)	830.6 (471.6)	1,000.3 (595.0)	<0.01
Heart rate variability SD of NN intervals	16.6 (14.0)	16.7 (14.0)	14.4 (14.2)	0.01
Cornell voltage	1,461.6 (578.2)	1,454.2 (572.1)	1,657.1 (694.3)	<0.01

GED, General Educational Development; IQR, interquartile range. *Health Utilities Index Mark 3 (HUI3): aggregate score of vision, hearing, speech, ambulation, dexterity, emotion, cognition, and pain. †Health Utilities Index Mark 2 (HUI2): aggregate score of sensation, mobility, cognition, self-care, emotion, pain, and fertility. ‡Estimated by the MDRD four-variable equation.

were identified as significant predictors of incident HF in the RSF and backward selection models. Fasting plasma glucose was preferentially included over HbA_{1c} in the relationship modeling stage owing to lack of HbA_{1c} data in the validation cohort. Presence of atrial fibrillation was noted in the baseline electrocardiogram assessment in 1.1% of the participants. While presence of atrial fibrillation was significantly associated with risk of HF in univariate Cox regression (hazard ratio 3.45 [95% CI 1.94–6.47]), it was not identified as a top predictor of incident HF in any of the variable selection models (ranked 53, 59, and 55 in the RSF, forward selection, and backward selection models, respectively).

The RSF-selected variables had a higher overall C-index, 0.74 (95% CI 0.71–0.74), compared with the stepwise forward, 0.71 (95% CI 0.67–0.74) and backward 0.73 (95% CI 0.70–0.76) selection methods when used with Cox PH relationship modeling ($P < 0.01$ and $P = 0.01$, respectively). With use of the same RSF-selected variables with RSF relationship modeling, the performance improved to an overall C-index of 0.77 (95% CI 0.75–0.80, $P < 0.001$) (Fig. 1*A*). Calibration of the RSF-based model was acceptable (Hosmer-Lemeshow statistic $\chi^2 = 9.63$, $P = 0.29$) (Fig. 1*B*). An online tool to calculate the RSF-based risk models has been made publicly available at www.cvriskscores.com to allow for its use with other data sets.

Development and Internal Validation of the WATCH-DM Score

From the 10 identified top-performing RSF-selected predictors, a risk score for 5-year HF incidence was created (Fig. 2): the WATCH-DM risk score (Weight [BMI], Age, hyperTension, Creatinine, HDL-C, Diabetes control [fasting plasma glucose], QRS Duration, MI, and CABG). The Cox PH β-coefficients, hazard ratios, and 95% CIs for each of the RSF-selected variables are displayed in Supplementary Table 3. The WATCH-DM risk score model demonstrated good discrimination with an overall C-index of 0.72 (95% CI 0.69–0.75) and acceptable calibration (Hosmer-Lemeshow $\chi^2 = 10.58$, $P = 0.23$) (Supplementary Fig. 2) for predicting HF risk in the internal validation subset of the ACCORD trial cohort.

The median WATCH-DM risk score was 10 with a theoretical range of 0–36. The observed scores ranged from 3 to 27. A 1-unit increment in the risk score was associated with a 24% higher risk of HF at 5 years. The cumulative 5-year incidence of HF increased in a graded fashion across data-derived quintiles of the risk score (Fig. 3), from 1.1% in quintile 1 (WATCH-DM score ≤7) to 17.4% in quintile 5 (WATCH-DM score ≥14). No significant interaction was observed between the intensive glucose control arm and the WATCH-DM risk score for the risk of incident HF.

External Validation of the RSF-Based Risk Prediction Model and the WATCH-DM Risk Score

The RSF-based model for predicting HF risk among individuals with T2DM was externally validated in the subgroup of ALLHAT participants with prevalent T2DM at baseline. The external validation cohort included 10,819 participants with 942 incident HF events (cumulative rate 8.7%) over a median follow-up of 4.8 years. The differences in baseline characteristics between the ACCORD and ALLHAT participants are shown in Supplementary Table 4. The RSF-based risk prediction model had good discrimination (C-index 0.74 [95% CI 0.72–0.76]) and acceptable calibration (Hosmer-Lemeshow statistic $\chi^2 = 11.05$, $P = 0.20$) (Fig. 1*D*) in the ALLHAT T2DM cohort.

The integer-based WATCH-DM risk score also demonstrated good discrimination (C-index 0.70 [95% CI 0.67–0.72]) and acceptable calibration (Hosmer-Lemeshow statistic $\chi^2 = 10.11$, $P = 0.29$) (Supplementary Fig. 3) for predicting HF risk in the ALLHAT T2DM cohort. The cumulative incidence of HF increased from 2.5% in the lowest quintile (score ≤7) based on the WATCH-DM risk score to 18.7% in the highest quintile (total score ≥14), indicating good risk stratification (Supplementary Fig. 4).

Information on HF subtype was available in 37% of the incident HF cases, with 44.3% identified as HFpEF ($N = 154$) and 55.7% as HFrEF ($N = 194$) (Supplementary Fig. 5). The median integer-based WATCH-DM risk score was higher in participants with incident HFrEF versus HFpEF events (median 14 [25–75% percentile 11–16] vs. 12 [9–15], $P < 0.01$). The WATCH-DM risk score also demonstrated better discrimination of risk in HFrEF versus HFpEF (C-index 0.72 [95% CI 0.67–0.75] vs. 0.64 [0.59–0.68], respectively, $P < 0.001$]. The cumulative incidence of HFrEF and HFpEF in the lowest quintile of WATCH-DM risk score were 0.4% and 0.8%, respectively, and in the highest quintile was 7.1% and 4%, respectively (Supplementary Fig. 6). Calibration was acceptable for each phenotype (Hosmer-Lemeshow statistic $\chi^2 = 8.11$ and 10.49 and $P = 0.42$ and 0.23 for HFrEF and HFpEF, respectively).

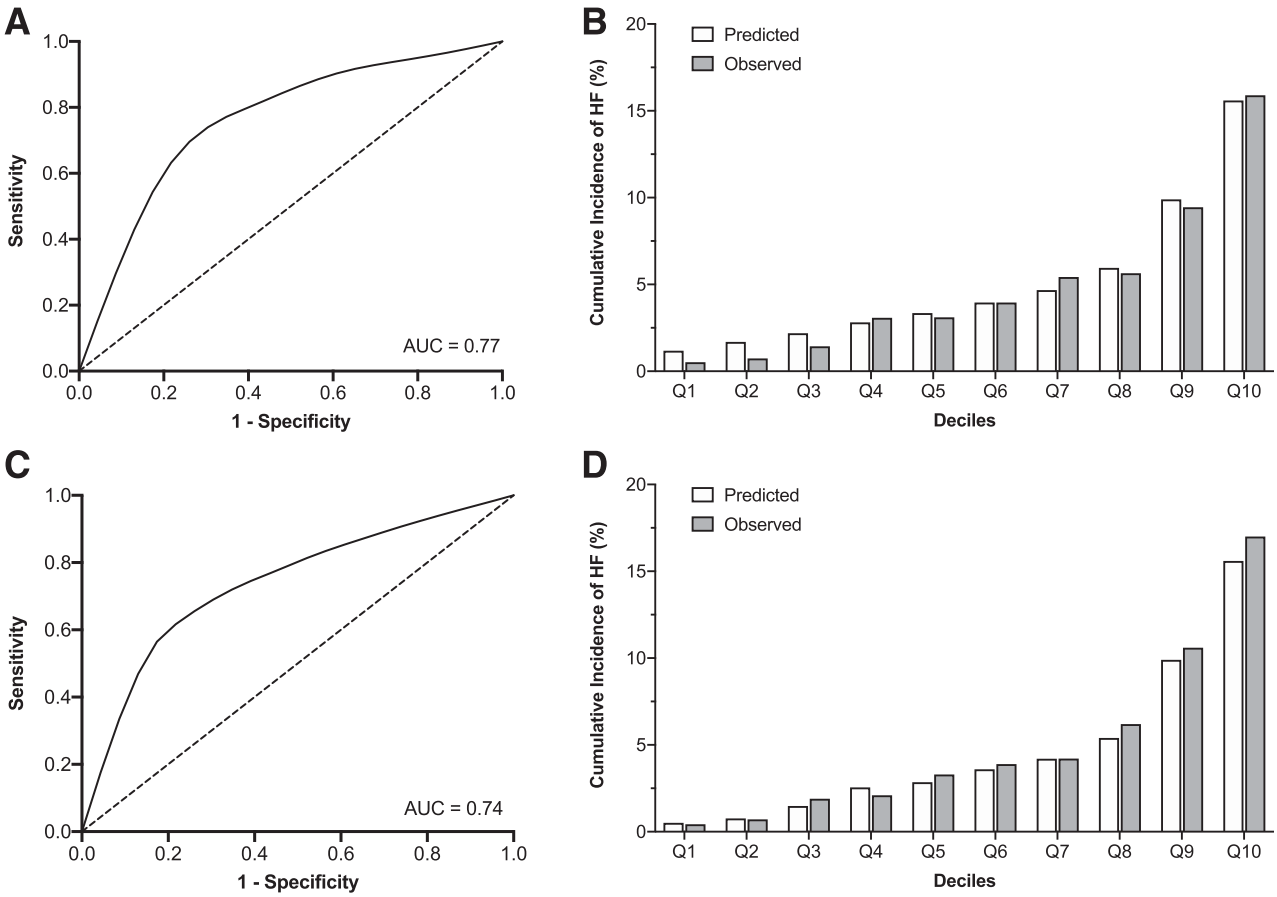

Figure 1—*A*: The ROC curve for the RSF-based model for predicting incident HF at year 5 in the derivation data set (ACCORD). AUC, area under the curve. *B*: Calibration of the RSF-based model in the derivation data set. Predicted vs. observed 5-year incidence of HF based on deciles of predicted risk. Calibration was acceptable (Hosmer-Lemeshow statistic $\chi^2 = 9.63$, $P = 0.29$). *C*: The ROC curve for the RSF-based model for predicting incident HF at year 5 in the external validation data set (ALLHAT). *D*: Calibration of the RSF-based model in the externally validated data set. Predicted vs. observed 5-year incidence of HF based on deciles of predicted risk. Calibration was acceptable (Hosmer-Lemeshow statistic $\chi^2 = 11.05$, $P = 0.20$). Q1 to Q10 refer to deciles of WATCH-DM risk score.

CONCLUSIONS

We present a novel, machine learning–derived risk score (WATCH-DM) that integrates readily available clinical, laboratory, and electrocardiographic variables to efficiently predict incident HF risk among high-risk patients with T2DM. The machine learning–based risk prediction model yielded favorable discrimination and greater accuracy compared with traditional risk scores. We derived this streamlined integer-based risk score after assessment of >60 candidate variables in a well-characterized population free from HF at baseline. Patients in the highest WATCH-DM risk category faced a 5-year risk of incident HF approaching 20%. The machine learning–based risk prediction model and the WATCH-DM risk score for HF performed well in an external cohort of patients with T2DM. Taken together, our study findings may inform risk-based monitoring strategies and targeted introduction of therapies known to influence HF risk.

A Machine Learning–Derived Risk Score

Although a number of risk prediction models have become available in T2DM (28), none to our knowledge have been specific to HF risk. This machine learning–based approach has unique advantages over traditional risk prediction, as it can handle large multidimensional sets of time-to-event data without need for assumptions of normality of distributions, linearity of risk prediction, and overfitting of models. Approaches adequately handling these issues may be especially important for complex phenotypes such as HF. Indeed, in the ACCORD and ALLHAT T2DM cohorts, the RSF-based methods offered better risk prediction than standard Cox-based methods. A web-based version of the RSF-based risk prediction model has been made available on www.cvriskscores.com to allow for widespread use of the risk prediction tool. This tool will calculate the 5-year incident HF risk for patients with T2DM using patient data on risk factors. Furthermore, we have also developed an integer-based risk score, the WATCH-DM risk score, to facilitate the ease of use in clinical settings without the need for a web-based platform or programming into the electronic health record system.

Transition From Cardiometabolic Disease to HF

Our robust risk prediction model also contributes to further understanding of broad mechanistic contributions to development of HF among at-risk patients with T2DM. Although in clinical practice, risk categories (such as obese BMI category) are commonly used, we

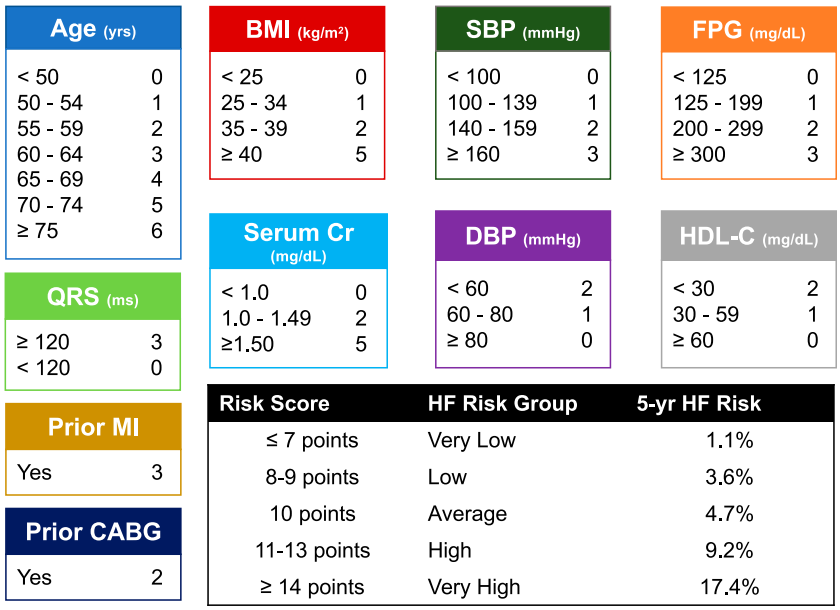

Figure 2—The WATCH-DM score for 5-year HF incidence and the five risk groups by quintiles of WATCH-DM (very low ≤7, low 8–9, average 10, high 11–13, and very high ≥14). CABG, coronary artery bypass grafting; CR, creatinine; DBP, diastolic blood pressure; FPG, fasting plasma glucose; HDL-C, HDL cholesterol; MI, myocardial infarction; SBP, systolic blood pressure; yrs, years.

demonstrate that risk associated with many of these clinical parameters operates on a continuum. In fact, most of these measures could be considered clinically silent and may not be routinely flagged as abnormal in many patients. Furthermore, these data serve to validate the multisystem inputs (spanning cardiovascular, kidney, and general health domains) that may inform the transition from cardiometabolic disease to HF. Beyond established risk factors of incident HF (such as age, adiposity, and blood pressure), these data also help to identify potential markers that were not hypothesis driven (such as select electrocardiographic parameters and HDL cholesterol) in forecasting future HF risk. It is reassuring that nearly identical predictors of incident HF risk were identified as the Prevent HF tool (29), which was derived from <33,000 patients across seven community-based cohorts. Importantly, Prevent HF was derived from a general sample of patients free from cardiovascular disease and was not specific for T2DM. In this higher-risk cohort with T2DM with or at risk for ASCVD, we additionally found prior ASCVD and serum creatinine to be important predictors of incident HF.

In a subset of patients in the external validation cohort (ALLHAT T2DM), we uniquely assessed the performance of our risk models in predicting future incident HFrEF or HFpEF. Although the WATCH-DM carried modest ability in predicting incident HFrEF, discrimination was lower for incident HFpEF events. These data highlight the need that risk predictors, and potentially strategies to attenuate risk, may differ between HFrEF and HFpEF and should be evaluated independently.

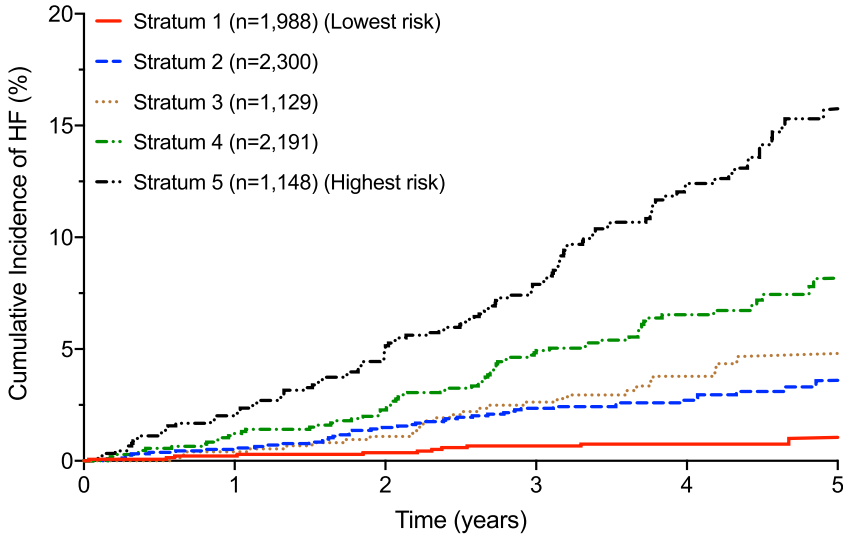

Figure 3—Cumulative incidence of HF across quintiles of WATCH-DM: quintile 1 (≤7), 2 (8–9), 3 (10), 4 (11–13), and 5 (≥14).

SGLT2i and HF Risk Reduction Strategies

The SGLT2i class has been shown to reduce risk of HF and kidney events in at-risk patients with T2DM (6–9). In the BI 10773 (Empagliflozin) Cardiovascular Outcome Event Trial in Type 2 Diabetes Mellitus Patients (EMPA-REG OUTCOME) trial, the benefits of empagliflozin were consistent across risk strata identified by the nine-variable Health ABC HF Risk Score (30). However, unlike our study, the trial enrolled only patients with prevalent cardiovascular disease, and the Health ABC HF Risk Score was not derived in a T2DM cohort. Given the attendant costs of global introduction of SGLT2i especially in limited resource settings (2) and since limited guidance is available regarding use of these therapies among T2DM without prevalent ASCVD, targeted integration of this therapeutic class in patients with T2DM at highest risk for HF based on the WATCH-DM score may be important and requires further study. Beyond the SGLT2i class, guideline-recommended strategies (31), including greater engagement with team-based care and control of target risk factors, may be particularly relevant in patients identified as at higher risk by WATCH-DM. Conversely, this risk score may also be used to select patients to avoid or limit use of therapies that may increase the risk of HF. For instance, cautious use of certain antihyperglycemic therapies, including the thiazolidinediones and select dipeptidyl peptidase-4 inhibitors (namely, saxagliptin and possibly alogliptin) that have been demonstrated to increase risk of HF in

randomized clinical trials (2,32,33), may be considered.

Study Strengths and Limitations

Our study has several strengths including the large sample size and event rates for HF in the derivation cohort; use of a machine learning–based statistical tool for variable selection and risk modeling that could handle large, multidimensional sets of time-to-event data and was not limited by the statistical assumptions of traditional risk prediction techniques; independent validation in an external cohort of patients with T2DM; and availability of HF subtype data in the validation cohort that allowed us to contrast the performance of the risk prediction model for HFpEF versus HFrEF events.

Our study is also subject to certain limitations. ACCORD was conducted between 1999 and 2009, and the relative importance of predictors of HF may have evolved over the last decade. Certain biomarkers, including circulating natriuretic peptides (which is a class IIA recommendation for HF risk screening in the American College of Cardiology/American Heart Association/Heart Failure Society of America heart failure guidelines) (31) and high-sensitivity troponin (34,35), were not available to assess incremental risk predictive value; however, these assays are not routinely collected in stable patients with T2DM without HF. Similarly, data on other potential predictors of HF such as atrial fibrillation, anemia (hemoglobin levels), and cardiorespiratory fitness levels were not available or inadequately captured (atrial fibrillation status only reported based on baseline electrocardiogram assessment) in the ACCORD trial cohort. More details regarding incident HF events (such as HF severity, ischemic vs. nonischemic etiology, and ejection fraction) were also not available in the derivation cohort, and risk predictors may be different in patients with HF with reduced, midrange, and preserved ejection fraction. However, considering the high predictive value of prior history of myocardial infarction and coronary artery bypass grafting in the WATCH-DM risk score, it is expected that the risk score identifies participants at a higher risk of developing ischemic cardiomyopathy. Furthermore, we were able to assess the performance of our risk prediction model for HFpEF versus HFrEF events in the validation cohort and observed better performance for HFrEF than HFpEF. Future studies are needed to develop specific risk scores for predicting HFpEF risk in the general population as well as among those with T2DM. As incident HF risk trajectories are known to vary widely by sex, race, and socioeconomic status (36), WATCH-DM will need to be tested in broader and more diverse cohorts beyond clinical trial settings. Furthermore, both ACCORD and ALLHAT included patients with higher cardiovascular risk (>10% with prior myocardial infarction in both cohorts), and future studies are needed to validate the WATCH-DM risk score in lower-risk cohorts of individuals with T2DM. Finally, studies are also needed to assess the efficacy of SGLT2i and other HF risk reduction strategies across a spectrum of HF risk based on this integer-based score.

Conclusion

In a well-phenotyped clinical trial population of patients with T2DM and cardiovascular disease or risk factors, but free from baseline HF, our novel risk prediction tool, WATCH-DM, identifies patients who face a HF risk of up to 20% in the next 5 years. As data elements needed to calculate the WATCH-DM risk score are collected in the routine clinical care of patients with T2DM, its integration in electronic health record systems or mobile health applications may facilitate its practical use. This risk score benefits from not requiring specific cardiovascular biomarker or adjunctive imaging assessment. Future investigations are needed to understand whether this identified risk is modifiable with currently available therapeutic strategies, including with SGLT2i. Machine learning–based approaches, which appear to outperform traditional risk prediction modeling in this setting, may efficiently validate known and uncover novel subclinical markers that inform the dynamic transition between cardiometabolic disease and manifest HF.

Funding. The Texas Health Resources Clinical Scholars Program funded this study. M.V. is supported by the KL2/Catalyst Medical Research Investigator Training award from Harvard Catalyst (National Institutes of Health [NIH]/National Center for Advancing Translational Sciences award UL 1TR002541). M.V. participates on clinical end point committees for studies sponsored by the NIH. J.B. has received research support from the NIH, Patient-Centered Outcomes Research Institute, and the European Union. J.L.G. and A.P. are supported by the Texas Health Resources Clinical Scholars Program.

Duality of Interest. M.V. serves on advisory boards for Amgen, AstraZeneca, Bayer AG, Baxter Healthcare, and Boehringer Ingelheim and participates on clinical end point committees for studies sponsored by Novartis. D.K.M. reports honoraria for trial leadership from AstraZeneca, Sanofi, Lilly USA, Janssen, Boehringer Ingelheim, Merck & Co, Pfizer, Novo Nordisk, Lexicon, Eisai, GlaxoSmithKline, and Esperion and honoraria for consulting for AstraZeneca, Sanofi, Lilly USA, Boehringer Ingelheim, Merck & Co, Pfizer, Novo Nordisk, Metavant, and Afimmune. J.B. has served as a consultant for Amgen, Array, AstraZeneca, Bayer, Boehringer Ingelheim, Bristol-Myers Squibb, CVRx, G3 Pharmaceuticals, Innolife, Janssen, Luitpold, Medtronic, Merck, Novartis, Novo Nordisk, Relypsa, and Vifor. G.C.F. reports consulting for Abbott, Amgen, Bayer, Janssen, Medtronic, and Novartis. J.L.G. serves as a consultant for Pfizer. No other potential conflicts of interest relevant to this article were reported.

Author Contributions. M.W.S., M.V., and A.P. developed the study concept and design, interpreted data, and critically revised and drafted the manuscript. K.V.P., D.K.M., J.Bu., G.C.F., J.L.G., B.E., and J.Be. contributed to discussion and critically revised and reviewed the manuscript. M.B., V.K., and D.W. performed statistical analyses and reviewed the manuscript. M.W.S. and A.P. are the guarantors of this work and, as such, had full access to all the data in the study and take responsibility for the integrity of the data and the accuracy of the data analysis.

Prior Presentation. Parts of this study were presented in abstract form at the EPI|Lifestyle Scientific Sessions of the American Heart Association, Houston, TX, 6–8 March 2019, and at the Heart Failure Society of America 23rd Annual Scientific Meeting, Philadelphia, PA, 13–16 September 2019.

References

1. American Diabetes Association. 8. Pharmacologic approaches to glycemic treatment: *Standards of Medical Care in Diabetes—2018*. Diabetes Care 2018;41(Suppl. 1):S73–S85

2. Greene SJ, Vaduganathan M, Khan MS, et al. Prevalent and incident heart failure in cardiovascular outcome trials of patients with type 2 diabetes. J Am Coll Cardiol 2018;71:1379–1390

3. McAllister DA, Read SH, Kerssens J, et al. Incidence of hospitalization for heart failure and case-fatality among 3.25 million people with and without diabetes mellitus. Circulation 2018;138:2774–2786

4. Standl E, Schnell O, McGuire DK. Heart failure considerations of antihyperglycemic medications for type 2 diabetes. Circ Res 2016;118:1830–1843

5. Rawshani A, Rawshani A, Franzén S, et al. Risk factors, mortality, and cardiovascular outcomes in patients with type 2 diabetes. N Engl J Med 2018;379:633–644

6. Zinman B, Wanner C, Lachin JM, et al.; EMPA-REG OUTCOME Investigators. Empagliflozin,

cardiovascular outcomes, and mortality in type 2 diabetes. N Engl J Med 2015;373:2117–2128
7. Neal B, Perkovic V, Mahaffey KW, et al.; CANVAS Program Collaborative Group. Canagliflozin and cardiovascular and renal events in type 2 diabetes. N Engl J Med 2017;377:644–657
8. Wiviott SD, Raz I, Bonaca MP, et al.; DECLARE–TIMI 58 Investigators. Dapagliflozin and cardiovascular outcomes in type 2 diabetes. N Engl J Med 2019;380:347–357
9. Perkovic V, Jardine MJ, Neal B, et al.; CREDENCE Trial Investigators. Canagliflozin and renal outcomes in type 2 diabetes and nephropathy. N Engl J Med 2019;380:2295–2306
10. Davies MJ, D'Alessio DA, Fradkin J, et al. Management of hyperglycemia in type 2 diabetes, 2018. A consensus report by the American Diabetes Association (ADA) and the European Association for the Study of Diabetes (EASD). Diabetes Care 2018;41:2669–2701
11. Das SR, Everett BM, Birtcher KK, et al. 2018 ACC expert consensus decision pathway on novel therapies for cardiovascular risk reduction in patients with type 2 diabetes and atherosclerotic cardiovascular disease: a report of the American College of Cardiology Task Force on Expert Consensus Decision Pathways. J Am Coll Cardiol 2018;72:3200–3223
12. Buse JB, Bigger JT, Byington RP, et al.; ACCORD Study Group. Action to Control Cardiovascular Risk in Diabetes (ACCORD) trial: design and methods. Am J Cardiol 2007;99(12A):21i–33i
13. Action to Control Cardiovascular Risk in Diabetes Study Group; Gerstein HC, Miller ME, Byington RP, et al. Effects of intensive glucose lowering in type 2 diabetes. N Engl J Med 2008;358:2545–2559
14. Cushman WC, Evans GW, Byington RP, et al.; ACCORD Study Group. Effects of intensive blood-pressure control in type 2 diabetes mellitus. N Engl J Med 2010;362:1575–1585
15. Stekhoven DJ, Bühlmann P. MissForest–non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112–118
16. Huang C, Murugiah K, Mahajan S, et al. Enhancing the prediction of acute kidney injury risk after percutaneous coronary intervention using machine learning techniques: a retrospective cohort study. PLoS Med 2018;15:e1002703
17. Ishwaran H, Kogalur U, Blackstone E, Lauer MS. Random survival forests. Ann Appl Stat 2008;2:841–860
18. Gurm HS, Kooiman J, LaLonde T, Grines C, Share D, Seth M. A random forest based risk model for reliable and accurate prediction of receipt of transfusion in patients undergoing percutaneous coronary intervention. PLoS One 2014;9:e96385
19. Ishwaran H, Gerds TA, Kogalur UB, Moore RD, Gange SJ, Lau BM. Random survival forests for competing risks. Biostatistics 2014;15:757–773
20. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. Biometrics 1988;44:837–845
21. Austin PC, Lee DS, D'Agostino RB, Fine JP. Developing points-based risk-scoring systems in the presence of competing risks. Stat Med 2016;35:4056–4072
22. Wolbers M, Koller MT, Witteman JC, Steyerberg EW. Prognostic models with competing risks: methods and application to coronary risk prediction. Epidemiology 2009;20:555–561
23. ALLHAT Officers and Coordinators for the ALLHAT Collaborative Research Group. The Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial. Major outcomes in high-risk hypertensive patients randomized to angiotensin-converting enzyme inhibitor or calcium channel blocker vs diuretic: the Antihypertensive and Lipid-Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) [published correction appears in JAMA 2003;289:178; published correction appears in JAMA 2004;291:2196]. JAMA 2002;288:2981–2997
24. Davis BR, Kostis JB, Simpson LM, et al.; ALLHAT Collaborative Research Group. Heart failure with preserved and reduced left ventricular ejection fraction in the antihypertensive and lipid-lowering treatment to prevent heart attack trial. Circulation 2008;118:2259–2267
25. Einhorn PT, Davis BR, Massie BM, et al.; ALLHAT Collaborative Research Group. The Antihypertensive and Lipid Lowering Treatment to Prevent Heart Attack Trial (ALLHAT) Heart Failure Validation Study: diagnosis and prognosis. Am Heart J 2007;153:42–53
26. Kogalur U, Ishwaran H. randomForestSRC: random forests for survival, regression, and classification (RF-SRC) [Internet], 2018. Available from https://cran.r-project.org/web/packages/randomForestSRC/index.html. Accessed 11 January 2019
27. Ripley B, Venables B, Bates D, Hornik K, Gebhardt A, Firth D. MASS: support functions and datasets for Venables and Ripley's MASS [Internet], 2018. Available from https://cran.r-project.org/web/packages/MASS/index.html. Accessed 11 January 2019
28. Noble D, Mathur R, Dent T, Meads C, Greenhalgh T. Risk models and scores for type 2 diabetes: systematic review. BMJ 2011;343:d7163
29. Khan SS, Ning H, Shah SJ, et al. 10-year risk equations for incident heart failure in the general population. J Am Coll Cardiol 2019;73:2388–2397
30. Fitchett D, Butler J, van de Borne P, et al.; EMPA-REG OUTCOME® Trial Investigators. Effects of empagliflozin on risk for cardiovascular death and heart failure hospitalization across the spectrum of heart failure risk in the EMPA-REG OUTCOME® trial. Eur Heart J 2018;39:363–370
31. Yancy CW, Jessup M, Bozkurt B, et al. 2017 ACC/AHA/HFSA focused update of the 2013 ACCF/AHA guideline for the management of heart failure: a report of the American College of Cardiology/American Heart Association Task Force on Clinical Practice Guidelines and the Heart Failure Society of America. Circulation 2017;136:e137–e161
32. Vijayakumar S, Vaduganathan M, Butler J. Glucose-lowering therapies and heart failure in type 2 diabetes mellitus: mechanistic links, clinical data, and future directions. Circulation 2018;137:1060–1073
33. McGuire DK, Alexander JH, Johansen OE, et al.; CARMELINA Investigators. Linagliptin effects on heart failure and related outcomes in individuals with type 2 diabetes mellitus at high cardiovascular and renal risk in CARMELINA. Circulation 2019;139:351–361
34. Everett BM, Brooks MM, Vlachos HE, Chaitman BR, Frye RL, Bhatt DL; BARI 2D Study Group. Troponin and cardiac events in stable ischemic heart disease and diabetes. N Engl J Med 2015;373:610–620
35. Scirica BM, Bhatt DL, Braunwald E, et al. Prognostic implications of biomarker assessments in patients with type 2 diabetes at high cardiovascular risk: a secondary analysis of a randomized clinical trial. JAMA Cardiol 2016;1:989–998
36. Glynn P, Lloyd-Jones DM, Feinstein MJ, Carnethon M, Khan SS. Disparities in cardiovascular mortality related to heart failure in the United States. J Am Coll Cardiol 2019;73:2354–2355