Supplemental Online Content

Segar MW, Hall JL, Jhund PS, et al. Machine learning–based models incorporating social determinants of health vs traditional models for predicting in-hospital mortality in patients with heart failure. *JAMA Cardiol*. Published online July 6, 2022. doi:10.1001/jamacardio.2022.1900

This supplemental material has been provided by the authors to give readers additional information about their work.

eMethods. GWTG-HF cohort description

Briefly, GWTG-HF is a hospital-based program that began in 2005 to assess and improve adherence to guideline-directed medical therapy in patients with HF hospitalization. Patients of participating centers admitted with new or worsening HF or who develop significant HF symptoms during their hospital stay are included in the registry. Participating centers include rural and urban, teaching and nonteaching, and small and large hospitals from all United States regions. Personnel trained in data abstraction reported patient data using the AHA's web-based Patient Management tool (Quintiles Real World and Late Phase Research, Cambridge, MA) in compliance with the Joint Commission and CMS standards.

eMethods: Candidate variables

Recorded data in the GWTG-HF registry encompass a range of domains, including patient demographics, vital signs, socioeconomic status, medical history, laboratory values, cardiac biomarkers, and electrocardiography and ejection fraction. A list of 38 candidate variables is shown in **eTable 1**. Race was self-reported as Black, White, Asian, or other. Ethnicity was reported as Hispanic vs. non-Hispanic. B-type natriuretic peptide (BNP) and N-terminal pro-B-type natriuretic peptide (NT-proBNP) levels were combined into a single log-transformed Z-score as previously described.(1) Abnormal troponin was defined as a value greater than the upper limit of normal specific to each hospital center. Only the cohorts with <15% missingness were imputed using random forest imputation.(2)

eMethods: Zip-code level social determinants of health (SDOH) parameters

All ZIP code data was that of the patient and not the hospital. Data from the 2019 Internal Revenue Service Statistics of Income were used to determine median household income.

Publicly available from the 2018 CDC Social Vulnerability Index,(3) the Graham Center 2011-2015 Social Deprivation Index,(4) and 2014-2018 Distressed Communities Index(5) were obtained and assigned to GWTG-HF participants by linking the participant's residential ZIP codes (**eTable 2**). The DCI compiles an aggregate score, termed the distress score, for each ZIP-code accounting for performance across seven individual metrics (< high-school education, housing vacancy rate, adults not in work, poverty rate, median income ratio, and change in employment and establishment proportions from 2014-2018). Scores range from 0-100, reflecting how prosperous (low score) or distressed (high score) a ZIP-code is. These indices provide neighborhood- and ZIP-code level measure varying aspects of social determinants of health and derive their scores from the American Communities Survey.(6) The social vulnerability index and the social deprivation index are presented as neighborhood-level data, while the Distressed Communities Index is presented as ZIP-code level data. Because these indices span across different years and our analysis was performed only on the ZIP-code level, we opted to extract matching ZIP-code variables from the 5-year estimates of the 2015-2019 American Communities Survey. The detailed methodology for measuring these social determinants of health have been reported previously.(3-5)

Hospital-level covariates included geographic classification (either rural or urban), sole community hospital, essential hospital, and disproportionate share hospital. Essential hospital was defined by the America's Essential Hospital association. In order to be defined as an Essential Hospital member, applicants must be publicly owned and operated by state/local governments, private nonprofits and/or via hybrid structures.(7) Member hospitals must also have over 75% of inpatient admissions and 70% outpatient visits be for patients with no insurance, Medicare, or Medicaid. Disproportionate share hospital was defined using the Medicare Disproportionate Share Hospital Payment report and defined as hospital serving a disproportionately large percentage (≥ 15%) of low income patients.(8)

eMethods: External validation cohort (ARIC Study) description

Briefly, the Atherosclerosis Risk in Communities (ARIC) study is an observational, community-based cohort study of 15,792 participants that began enrolling individuals in 1987 from 4 US communities (Washington County, MD; Minneapolis, MN; Jackson, MS; Forsyth County, NC). Beginning in 2005, physician reviewers began adjudicating HF hospitalizations into definite, probable, stable chronic, not HF, or unclassifiable events using standardized

criteria.(9) For the present analysis, hospitalizations categorized as not HF or unclassifiable were excluded. Cohort participants hospitalized with HF had demographic, medical history, and laboratory values abstracted from their hospital records by trained personnel.

e-Methods: Description of random forest-based variable importance selection

This technique is an ensemble classification method that aggregates the results of multiple decision trees – each a random subset of the derivation cohort.(10) Permutation-based variable selection was performed to optimize the number of covariates included in the final prediction model by quantifying a covariate's variable importance (VIMP).(11) A high VIMP score indicates higher variable importance as randomly permutating the variable degrades model performance. For both groups (Black and non-Black race), a ML model was generated for all data in the derivation cohort and variables were ranked by VIMP score. The process was repeated 20 times with the final covariate VIMP rank being an average of all iterations. The final number of variables included in the model was determined by visually assessing the area under the receiver operating characteristics (C-index) curve across an increasing number of VIMP-ordered covariates.

e-Methods: Development of the logistic regression model

A logistic regression model was developed using the same training data as the machine learning model. Using the same 40 candidate variables (Supplemental Table 1) as the machine learning model, variables were selected using a backwards selection method with minimization of the Akaike Information Criterion. Race was then forced into the model. Multicollinearity of the resultant model was assessed using the variable inflation factor (VIF). The resultant model did not show evidence of multicollinearity with a VIF range of 1.04-2.57. Since 3 of the ML model variables were missing in the ARIC dataset, the logistic regression model was rederived using available data in the derivation cohort. Conversely, the ML model can overcome missing data using the decision trees with the available data and built-in imputation.(12)

e-Methods: Dichotomization of continuous variables and PARP analysis methodology

A logistic regression model was developed for both Black and non-Black race groups using the variables identified in the clinical + socioeconomic model. The exposure of interest was removed by reclassifying all individuals as unexposed. The sum of the new model predicted probabilities for each individual represents the expected number of cases if the exposure were removed from the population. The PARP was subsequently calculated as the difference in the

observed minus expected cases and expressed as a percentage. Continuous variables were dichotomized based on previously established clinical cutoffs or median values: Age ≥ 70 years, BNP ≥ 1000 or NT-proBNP ≥ 4000 pg/mL, BUN ≥ 22 mg/dL, BMI ≥ 30 kg/m^2, creatinine ≥ 1.3 mg/dL, SBP ≥ 125 mmHg, DBP ≥ 80 mmHg, EF < 40%, hemoglobin ≤ 11 g/dL, heart rate ≥ 80 bpm, income < $54,471, potassium ≤ 4 mg/dL, sodium ≤ 134 mg/dL, and QRS duration ≥ 120 ms. Socioeconomic covariates were dichotomized by the highest quartile and included poverty rate ≥ 28%, disability rate ≥ 17.8%, total population ≥ 41,275, percentage of households without vehicle ≥ 22.3%, unemployment rate ≥ 6.2%, and percentage of housing units with 10+ units ≥ 22.3%.

eReferences

1. Anand IS, Claggett B, Liu J et al. Interaction Between Spironolactone and Natriuretic Peptides in Patients With Heart Failure and Preserved Ejection Fraction: From the TOPCAT Trial. JACC Heart Fail 2017;5:241-252.

2. Stekhoven DJ, Buhlmann P. MissForest--non-parametric missing value imputation for mixed-type data. Bioinformatics 2012;28:112-8.

3. https://www.atsdr.cdc.gov/placeandhealth/svi/data_documentation_download.html. Date of Access: 8/1/2021.

4. (https://www.graham-center.org/rgc/maps-data-tools/sdi/social-deprivation-index.html. Date of Access: 7/25/2021.

5. https://eig.org/dci. Date of Access: 6/24/2021.

6. https://www.census.gov/programs-surveys/acs). Date of Access: 7/25/2021.

7. Americas Essential Hospitals. http://essentialhospitals.org, Date of access: 11/2/2021.

8. Disproportionate Share Hospital. https://www.cms.gov/Medicare/Medicare-Fee-for-Service-Payment/AcuteInpatientPPS/dsh, Date of access: 11/2/2021.

9. Rosamond WD, Chang PP, Baggett C et al. Classification of heart failure in the atherosclerosis risk in communities (ARIC) study: a comparison of diagnostic criteria. Circ Heart Fail 2012;5:152-9.

10. Breiman L. Random Forests. Machine Learning 2001;45:5-32.

11. Segar MW, Vaduganathan M, Patel KV et al. Machine Learning to Predict the Risk of Incident Heart Failure Hospitalization Among Patients With Diabetes: The WATCH-DM Risk Score. Diabetes Care 2019;42:2298-2306.

12. Tang F, Ishwaran H. Random Forest Missing Data Algorithms. Stat Anal Data Min 2017;10:363-377.

13. Fonarow GC, Adams KF, Jr., Abraham WT et al. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. JAMA 2005;293:572-80.

14. Auble TE, Hsieh M, Gardner W et al. A prediction rule to identify low-risk patients with heart failure. Acad Emerg Med 2005;12:514-21.

15. Peterson PN, Rumsfeld JS, Liang L et al. A validated risk score for in-hospital mortality in patients with heart failure from the American Heart Association get with the guidelines program. Circ Cardiovasc Qual Outcomes 2010;3:25-32.

16. Gok G, Karadag M, Sinan UY, Zoghi M. A New Risk Score to Predict In-Hospital Mortality in Elderly Patients With Acute Heart Failure: On Behalf of the Journey HF-TR Study Investigators. Angiology 2020;71:948-954.

17. Abraham WT, Fonarow GC, Albert NM et al. Predictors of in-hospital mortality in patients hospitalized for heart failure: insights from the Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients with Heart Failure (OPTIMIZE-HF). J Am Coll Cardiol 2008;52:347-56.

18. Kwon JM, Kim KH, Jeon KH et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. PLoS One 2019;14:e0219302.

eTable 1. Candidate covariates and their respective domain.

Demographic (3): age, gender, ethnicity

Medical history (15): smoking, atrial fibrillation, COPD/asthma, hyperlipidemia, hypertension, coronary artery disease, prior myocardial infarction, cerebrovascular accident or transient ischemic attack, heart failure, dialysis, percutaneous coronary intervention, coronary artery bypass graft, valvular heart disease, diabetes, ischemic heart failure etiology

Echocardiogram (1): Ejection fraction

Electrocardiogram (2): QRS duration, ECG morphology

Vital signs (5): heart rate, systolic blood pressure, diastolic blood pressure, body mass index, respiratory rate

Laboratory (10): sodium, creatinine, blood urea nitrogen, potassium, hemoglobin, low-density lipoprotein-cholesterol, high-density lipoprotein-cholesterol, triglycerides, total cholesterol, plasma glucose

Biomarkers (2): abnormal troponin, natriuretic peptide level

Socioeconomic parameters (27):

Participant level (1): insurance status (Medicaid, none, vs. other)

Zip-code level (21): median household income, total population, population aged ≥ 65 years, percent with bachelor's degree, percent minority, percent renter-occupied housing, housing vacancy rate, percent with housing units ≥ 10, percent mobile homes, percent of households without vehicle, average number of housing units occupants per room, percent unemployed, percent single-parent households, percent non-institutionalized with a disability, percent foreign born, percent who speak English "less than well", percent below poverty level, percent in group home, percent in urban designation, percent change in employment, percent change in establishments

Hospital level (5): Proportion of Non-White participants, geographic classification (rural vs. urban), sole community hospital, essential hospital, disproportionate share hospital

eTable 2. ZIP-code level social determinants of health parameters that were considered for predicting in-hospital mortality following HF hospitalization. The table here shows the variables included and their source.

Internal Revenue Service Statistics of Income 2018	**Social Vulnerability Index 2016**	**Social Deprivation Index 2011-2015**	**Distressed Communities Index 2014-2018**
Median household income	Percent of population aged 65 and older	Percent of Households without a Vehicle	Total Population
	Percent renter-occupied housing	Percent Single Parent Household	Percent of Adults with bachelor's degree
	Percent of housing in structures w/ 10+ units		Percent Minority
	Percent in Mobile Homes		Housing Vacancy Rate
	Number of Housing Units Occupants per Room		Percent Unemployed
	Percent Non-Institutionalized with a Disability		Percent Foreign Born
	Percent of Adults who speak English "less than well"		Percent in Urban Designation
	Percent Below Poverty Level		Change in Employment from 2014 to 2018
	Percent in Group Homes		Change in Establishments from 2014 to 2018

eTable 3. Baseline characteristics of participants in the internal and external validation cohorts by race.

	GWTG Internal Validation		ARIC External Validation	
	Black	**Non-Black**	**Black**	**Non-Black**
N	15634	66786	1205	2264
Age, years	62.8 (14.7)	73.9 (13.7)	76.3 (6.4)	79.0 (6.1)
Women, %	7490 (47.9)	31547 (47.2)	720 (59.8)	1174 (51.9)
Systolic blood pressure, mmHg	148.0 (31.7)	141.1 (29.0)	146.3 (35.5)	138.3 (30.4)
Body mass index, kg/m2	32.5 (10.4)	30.1 (8.9)	39.1 (23.4)	34.8 (17.2)
No insurance	1638 (10.7)	1902 (2.9)	6 (0.5)	11 (0.5)
Median Zip code household income, $	4416 (28.2)	10080 (15.1)	-	-
Current smoker, %	13805 (88.3)	55237 (82.7)	121 (10.0)	140 (6.2)
Hypertension, %	5471 (35.0)	33434 (50.1)	1135 (94.2)	1997 (88.2)
CAD, %	7652 (48.9)	30016 (44.9)	236 (19.6)	237 (10.5)
Diabetes, %	138.8 (4.0)	137.8 (4.6)	767 (63.7)	1085 (47.9)
Sodium, mg/dL	2.0 (1.9)	1.6 (1.3)	135.9 (4.5)	135.7 (4.4)
Creatinine, mg/dL	1.9 (1.6)	1.6 (1.2)	1.9 (1.5)	1.5 (0.9)
Hemoglobin, g/dL	144.3 (32.0)	137.9 (27.2)	10.3 (2.3)	10.6 (2.2)
BNP, pg/mL	915 [390, 1898]	794 [395, 1569]	1048 [498, 2040]	787 [412, 1534]
NT-proBNP, pg/mL	4587[1861, 11444]	4821 [2126, 10865]	7753 [3096, 18275]	5212 [2736, 11392]
Abnormal troponin, %	5612 (35.9)	19724 (29.5)	507 (42.1)	846 (37.4)
QRS duration, ms	106.4 (28.4)	115.7 (33.1)	-	-

Abbreviations:
BNP, B-type natriuretic peptide; BUN, blood urea nitrogen; CAD, coronary artery disease; NT-proBNP, N-terminal pro-B-type natriuretic peptide

eTable 4. Discrimination and calibration performance of the race-specific models for predicting in-hospital mortality among patients with heart failure in the internal GWTG testing cohort with complete data available and with up to 50% missingness in the covariate data. A higher AUROC and lower Brier score indicate better performance. Among calibration slope measures, an intercept closer to 0 and slope closer to 1 indicates better calibration.

	Discrimination	**Calibration**		
	C-index	**Brier Score ($\times 10^{-3}$)**	**Intercept**	**Slope**
Testing dataset with < 15% missingness *(N = 15,634 Black patients, n = 66,786 non-Black patients)*				
Race-specific model in Black patients	0.81 (0.79, 0.83)	17 (15, 19)	-0.08	0.92
Race-specific model in non-Black patients	0.82 (0.81, 0.83)	29 (28, 30)	-0.12	0.99
Race-agnostic model in Black patients	0.81 (0.79, 0.83)	17 (15, 18)	-0.10	0.98
Race-agnostic model in non-Black patients	0.82 (0.80, 0.83)	29 (28, 30)	-0.14	0.97
Testing dataset with < 50% missingness *(n=107,508 Black patients, n = 445,998 non-Black patients)*				
Race-specific model in Black patients	0.74 (0.71, 0.77)	16 (11, 21)	-0.17	0.92
Race-specific model in non-Black patients	0.75 (0.73, 0.78)	31 (30, 31)	0.11	1.10
Race-agnostic model in Black patients	0.74 (0.71, 0.76)	16 (12, 21)	-0.20	0.93
Race-agnostic model in non-Black patients	0.75 (0.73, 0.78)	31 (30, 31)	0.12	1.09

eTable 5. Discrimination and calibration performance of the models for predicting in-hospital mortality among patients with heart failure in the internal GWTG validation cohort across age, sex, and socioeconomic status-based subgroups. A higher AUROC and lower Brier score indicate better performance. Among calibration slope measures, an intercept closer to 0 and slope closer to 1 indicates better calibration.

	C-index	Brier Score (x 10^{-5})
Age-based subgroup (<70 years) *(n=10,952 Black race patients, n=24,281 non-Black patients)*		
Race-specific model in Black patients	0.77 (0.73, 0.82)	114 (98, 130)
Race-specific model in non-Black patients	0.80 (0.78, 0.81)	201 (191, 216)
Race-agnostic model in Black patients	0.78 (0.76, 0.82)	172 (150, 197)
Race-agnostic model in non-Black patients	0.78 (0.77, 0.80)	277 (264, 288)
Age-based subgroup (>=70 years) (*n=5,229 Black race patients, n=46,790 non*-Black patients)		
Race-specific model in Black patients	0.79 (0.74, 0.83)	263 (225, 305)
Race-specific model in non-Black patients	0.80 (0.78, 0.82)	338 (325, 356)
Race-agnostic model in Black patients	0.80 (0.76, 0.84)	254 (232, 288)
Race-agnostic model in non-Black patients	0.78 (0.77, 0.79)	297 (282, 311)
Sex-based subgroups (Men) (*n=8,446 Black race patients, n=37,540 non*-Black patients)		
Race-specific model in Black patients	0.78 (0.74, 0.82)	158 (135, 184)
Race-specific model in non-Black patients	0.81 (0.79, 0.83)	296 (280, 309)
Race-agnostic model in Black patients	0.78 (0.74, 0.82)	172 (150, 200)
Race-agnostic model in non-Black patients	0.78 (0.77, 0.80)	278 (263, 292)
Sex-based subgroups (Women) (*n=7,735 Black race patients, n=33,531 non*-Black patients)		
Race-specific model in Black patients	0.80 (0.76, 0.84)	164 (141, 185)
Race-specific model in non-Black patients	0.80 (0.79, 0.82)	282 (269, 299)
Race-agnostic model in Black patients	0.80 (0.76, 0.84)	155 (135, 180)
Race-agnostic model in non-Black patients	0.78 (0.77, 0.79)	295 (283, 309)
HF with reduced ejection fraction		
Race-specific model in Black patients	0.86 (0.80, 0.89)	170 (96, 255)

Race-specific model in non-Black patients	0.80 (0.78, 0.81)	202 (138, 269)
Race-agnostic model in Black patients	0.80 (0.76, 0.84)	172 (146, 195)
Race-agnostic model in non-Black patients	0.79 (0.78, 0.81)	340 (322, 356)
HF with preserved ejection fraction		
Race-specific model in Black patients	0.77 (0.72, 0.83)	75 (33, 152)
Race-specific model in non-Black patients	0.82 (0.80, 0.83)	211 (159, 271)
Race-agnostic model in Black patients	0.78 (0.74, 0.83)	152 (127, 178)
Race-agnostic model in non-Black patients	0.78 (0.77, 0.80)	253 (241, 266)
SES based subgroups (below median income [household income <$54,471per annum]) (*n=8,397 Black race patients, n=24,208 non*-Black patients)		
Race-specific model in Black patients	0.78 (0.75, 0.82)	164 (143, 186)
Race-specific model in non-Black patients	0.77 (0.75, 0.79)	274 (257, 291)
Race-agnostic model in Black patients	0.78 (0.75, 0.81)	170 (148, 189)
Race-agnostic model in non-Black patients	0.76 (0.75, 0.77)	270 (251, 285)
SES based subgroups (above median income [household income >=$54,471per annum]) (*n=7,784 Black race patients, n=46,863 non*-Black patients)		
Race-specific model in Black patients	0.79 (0.74, 0.83)	157 (133, 185)
Race-specific model in non-Black patients	0.81 (0.79, 0.83)	296 (283, 309)
Race-agnostic model in Black patients	0.79 (0.74, 0.83)	160 (134, 188)
Race-agnostic model in non-Black patients	0.80 (0.79, 0.83)	288 (279, 303)

eTable 6. Discrimination and calibration performance of the non-Black race-specific and race-agnostic models for predicting in-hospital mortality among patients with heart failure with different self-identified race/ethnicities.

	C-index	**Brier Score ($\times 10^{-3}$)**
Race-specific ML Model		
Race-based subgroups		
Asian (n=1,372)	0.79 (0.72, 0.87)	28 (22, 36)
Other (n=4,648)	0.80 (0.74, 0.85)	32 (27, 39)
White (n=60,766)	0.81 (0.80, 0.82)	29 (28, 30)
Ethnicity-based subgroups		
Hispanic, non-Black race (n=59,662)	0.78 (0.77, 0.79)	29 (28, 30)
Non-Hispanic, non-Black race (n=7,124)	0.79 (0.76, 0.83)	24 (21, 27)
Race-agnostic ML Model		
Race-based subgroups		
Asian (n=1,372)	0.78 (0.71, 0.85)	38 (29, 44)
Other (n=4,648)	0.78 (0.72, 0.83)	40 (30, 49)
White (n=60,766)	0.81 (0.80, 0.82)	28 (27, 30)
Ethnicity-based subgroups		
Hispanic, non-Black race (n=59,662)	0.77 (0.76, 0.79)	29 (28, 30)
Non-Hispanic, non-Black race (n=7,124)	0.79 (0.76, 0.82)	24 (21, 28)

eTable 7. Reclassification metrics in the ARIC external validation between the ML model and the original GWTG risk score. Categorical NRI was calculated using race-specific event rate risk thresholds.

	Categorical NRI (95% CI)	**P-value**	**IDI (95% CI)**	**P-value**
Black patients (n = 1,205)				
Race-specific ML model (vs. GWTG risk score)	0.34 (0.12, 0.55)	0.007	0.01 (0.001, 0.02)	0.02
Race-agnostic ML model (vs. GWTG risk score)	0.31 (0.10, 0.51)	0.01	0.01 (0.001, 0.014)	0.03
LR race-specific (vs. LR with race as covariate)	0.18 (0.02, 0.29)	0.03	0.01 (0.006, 0.015)	0.003
Non-Black patients (n = 2,264)				
Race-specific ML model (vs. GWTG risk score)	0.39 (0.28, 0.57)	<0.001	0.004 (0.001, 0.008)	0.04
Race-agnostic ML model (vs. GWTG risk score)	0.33 (0.25, 0.50)	<0.001	0.003 (0.001, 0.006)	0.04
LR race-specific (vs. LR with race as covariate)	0.22 (0.04, 0.42)	0.02	0.03 (0.003, 0.05)	0.03
Abbreviations: *IDI, integrated discrimination improvement; ML, machine learning; NRI, net reclassification improvement*				

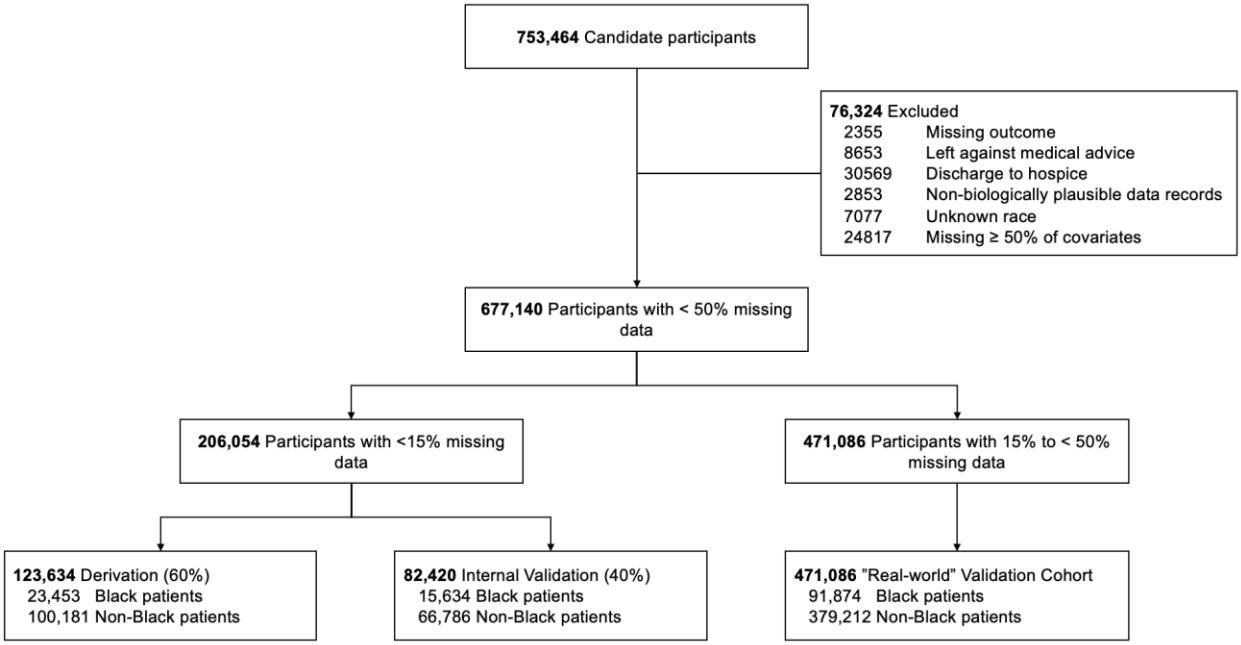
eTable 8. Discrimination and calibration performance of the models for predicting in-hospital mortality among patients with heart failure in the internal GWTG validation cohort across disproportional share hospital-based subgroups. A higher AUROC and lower Brier score indicate better performance.

	C-index	Brier Score (x 10^{-3})
Disproportionate Share Hospital *(n=24,752 Black patients, n=70,642 non-Black patients)*		
Race-specific model in Black patients	0.75 (0.72, 0.78)	16 (14, 17)
Race-specific model in non-Black patients	0.74 (0.73, 0.75)	30 (29, 31)
Race-agnostic model in Black patients	0.75 (0.72, 0.77)	16 (14, 18)
Race-agnostic model in non-Black patients	0.74 (0.73, 0.75)	30 (29, 31)
No Disproportionate Share Hospital *(n=12,925 Black patients, n=75,130 non-Black patients)*		
Race-specific model in Black patients	0.77 (0.73, 0.80)	14 (12, 15)
Race-specific model in non-Black patients	0.74 (0.73, 0.75)	29 (28, 30)
Race-agnostic model in Black patients	0.76 (0.73, 0.79)	15 (13, 16)
Race-agnostic model in non-Black patients	0.74 (0.73, 0.75)	29 (28, 30)

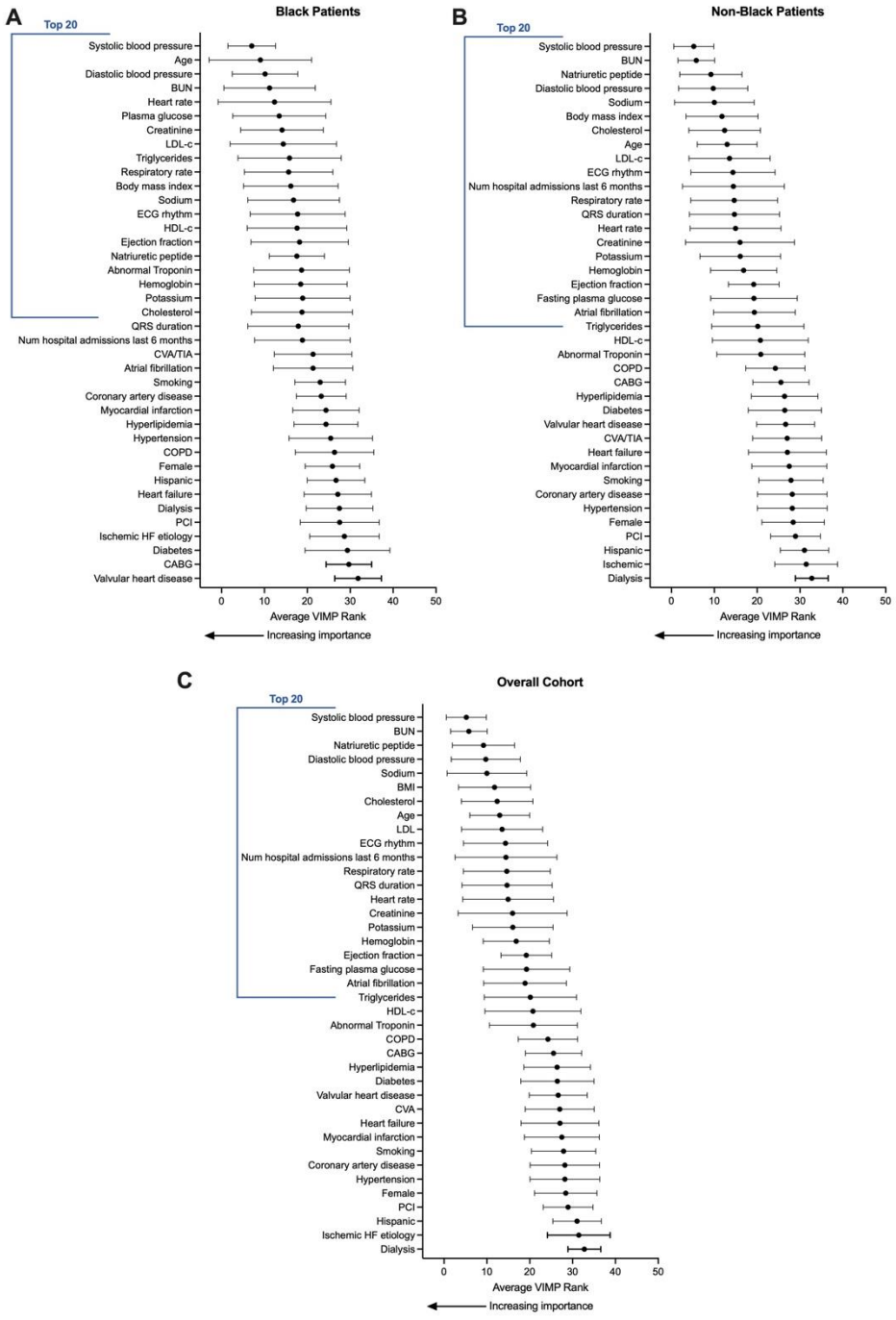
eTable 9. Comparison of models to predict risk of in-hospital mortality among patients with hospitalization for heart failure.

Study	Model Development	Sample Size (Derivation/ Validation)	In-hospital mortality rate	% Black race	No. of features	cardiac biomarkers	AUC
ADHERE (13)	Decision tree and logistic regression	33,046/32,229	4.0-4.2%	Unknown	3	No	Decision tree: 0.67 LR: 0.76
AHFI (14)	Decision tree	33,533/8,384	4.5%	19.8	21	No	Risk stratification only
GWTG-HF (15)	Logistic regression	27,850/11,933	2.86%	17.6	7	No	0.75
Journey HF-TR (16)	Logistic regression	702/346	7.4%	0	6	No	Risk stratification only
OPTIMIZE-HF (17)	Logistic regression	37,548/181,830	3.8%	18	7	No	0.75
KorAHF (18)	Deep neural network	2,165/4,759	2.9%	0	22	No	0.88

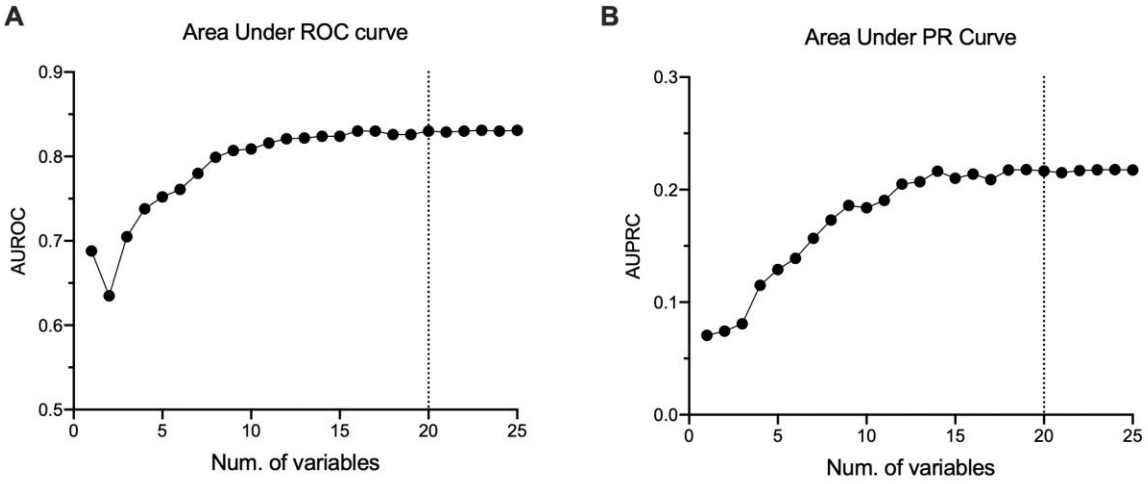
eFigure 1. CONSORT diagram.



eFigure 2. Variable importance of **A)** Black and **B)** non-Black patients determined by the VIMP metric of a race-specific random forest model with 20 bootstrap replicates. **Figure C** shows the variable importance determined by the VIMP metric of a race-agnostic model in the overall cohort. A lower VIMP rank indicates higher variable importance. The blue box indicates the top 20 variables.

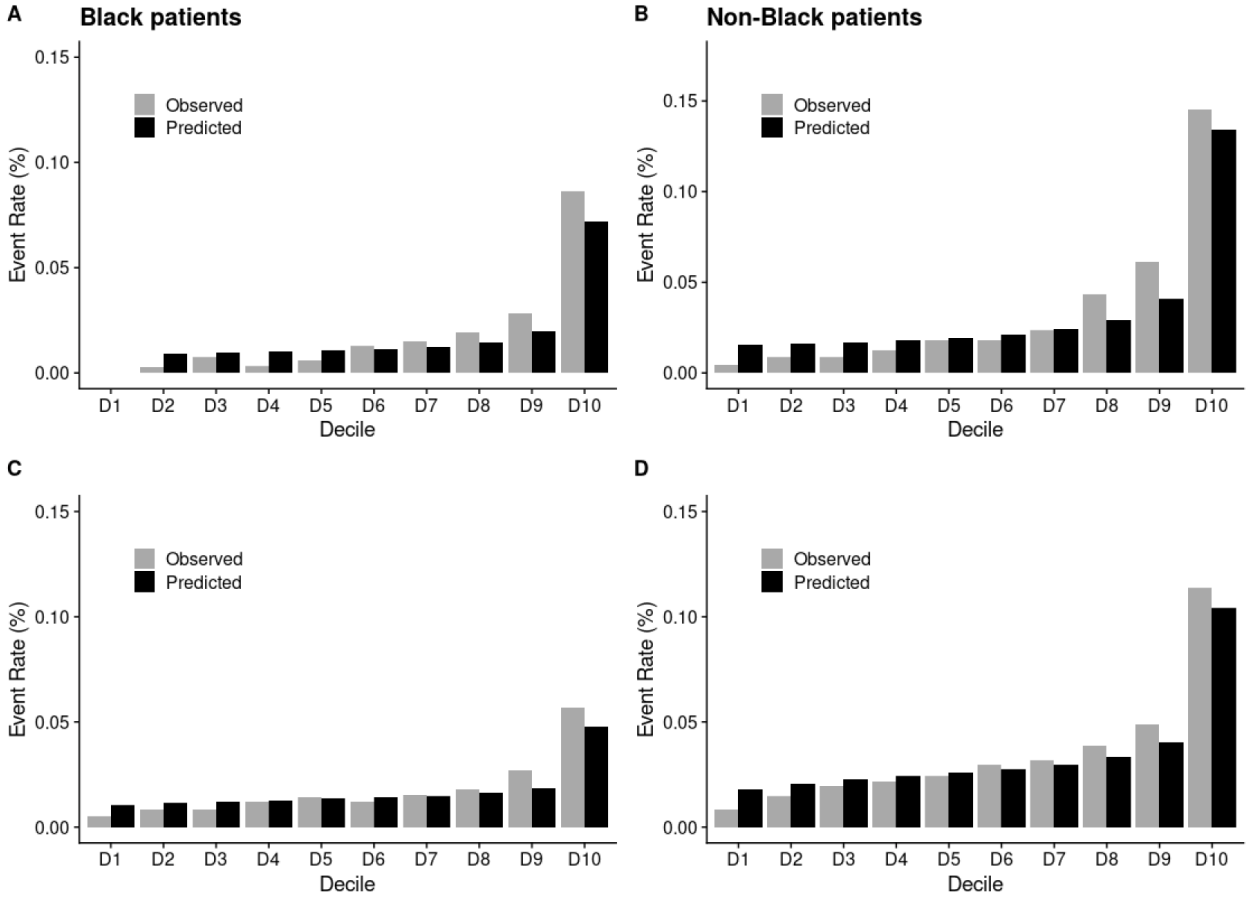


eFigure 3. Area under the **A)** receiver operating characteristics and **B)** precision-recall curve for increasing number of variables in a random forest model to predict in-hospital mortality in the overall cohort.

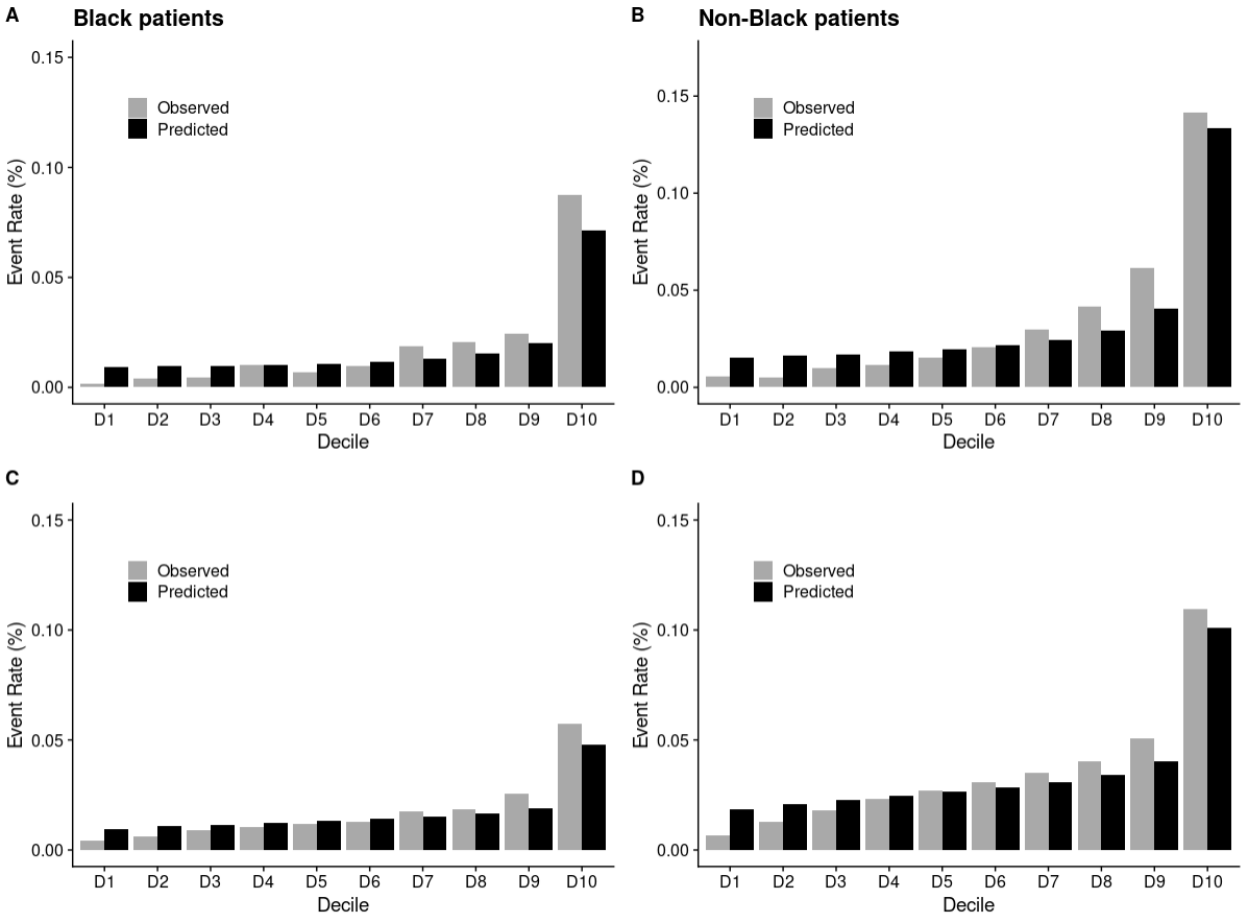


Abbreviations:
AUPRC, area under the precision-recall curve; AUROC, area under the receiver operating characteristics curve

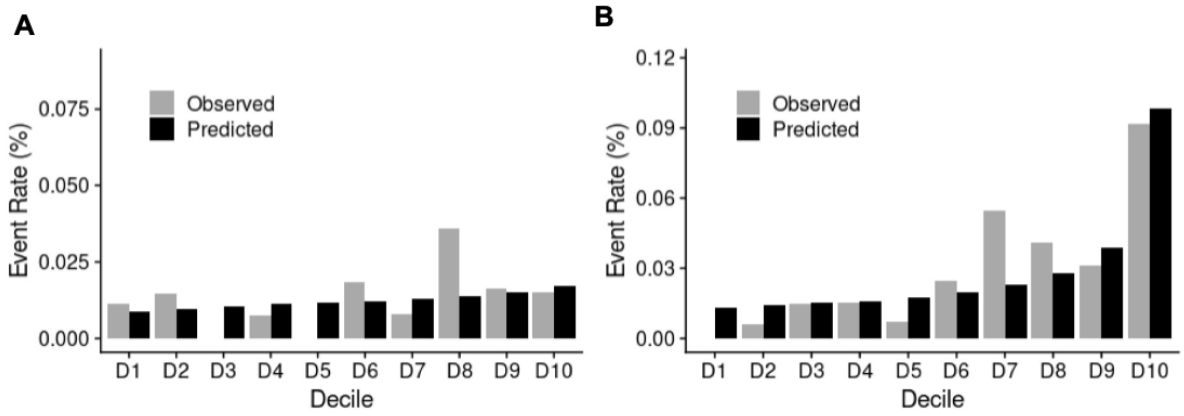
eFigure 4. Observed vs. predicted probability of in-hospital mortality for the race-specific ML models in **A)** Black and **B)** non-Black patients in the internal validation cohort and in **C)** Black and **D)** non-Black patients in the validation cohort with up to 50% missingness.



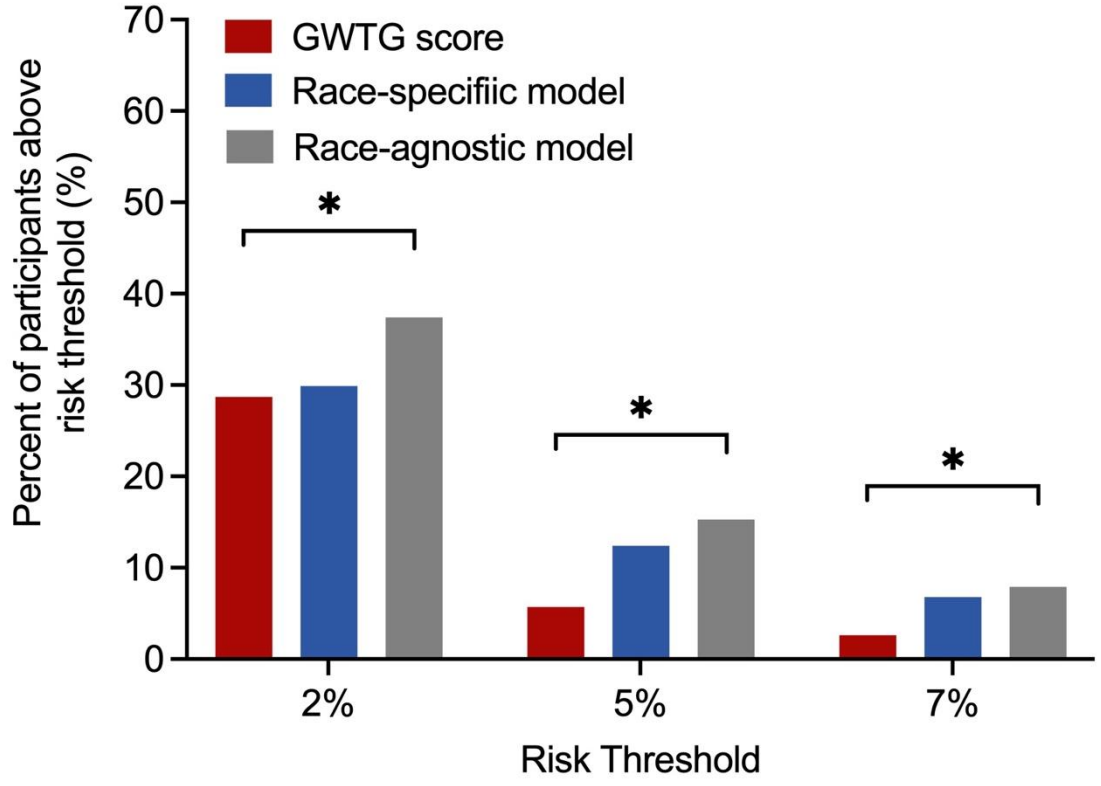
eFigure 5. Observed vs. predicted probability of in-hospital mortality for the race-agnostic ML models in **A)** Black and **B)** non-Black patients in the internal validation cohort and in **C)** Black and **D)** non-Black patients in the validation cohort with up to 50% missingness.



eFigure 6. Observed vs. predicted probability of in-hospital mortality for the GWTG-HF risk score in **A)** Black and **B)** non-Black patients in the ARIC external validation cohort.



eFigure 7. Among Black participants in the ARIC external validation cohort, percentage of participants with a predicted risk above the specific risk thresholds between the original GWTG risk score and the race-specific ML model. Asterisks indicate a chi-square p-value < 0.05.



eFigure 8. Observed vs. predicted probability of in-hospital mortality for the race-specific ML + social determinants of health models in **A)** Black and **B)** non-Black patients and race-agnostic ML + social determinants of health models in **C)** Black and **D)** non-Black patients in the internal validation cohort with up to 50% missingness.