ESC HEART FAILURE
ESC Heart Failure 2024; **11**: 2490–2498
Published online 18 April 2024 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/ehf2.14796

REVIEW

Predicting in-hospital mortality among patients admitted with a diagnosis of heart failure: a machine learning approach

Zina Jawadi[1], Rosemary He[2], Pratyaksh K. Srivastava[3], Gregg C. Fonarow[3], Suzan O. Khalil[3], Srikanth Krishnan[4], Eleazar Eskin[2,5], Jeffrey N. Chiang[5,6] and Ali Nsair[3]*

[1]UCLA David Geffen School of Medicine, Los Angeles, CA, USA; [2]Department of Computer Science, UCLA, Los Angeles, CA, USA; [3]Ahmanson-UCLA Cardiomyopathy Center, Ronald Reagan-UCLA Medical Center, MRL 3-760, 675 C.E. Young Dr., Los Angeles, CA 90095-1760, USA; [4]Division of Cardiology, Lundquist Institute for Biomedical Innovation, Harbor-UCLA Medical Center, Los Angeles, CA, USA; [5]Department of Computational Medicine, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA; and [6]Department of Neurosurgery, David Geffen School of Medicine, UCLA, Los Angeles, CA, USA

Abstract

Existing risk prediction models for hospitalized heart failure patients are limited. We identified patients hospitalized with a diagnosis of heart failure between 7 May 2013 and 26 April 2022 from a large academic, quaternary care medical centre (training cohort). Demographics, medical comorbidities, vitals, and labs were collected and were used to construct random forest machine learning models to predict in-hospital mortality. Models were compared with logistic regression, and to commonly used heart failure risk scores. The models were subsequently validated in patients hospitalized with a diagnosis of heart failure from a second academic, community medical centre (validation cohort). The entire cohort comprised 21 802 patients, of which 14 539 were in the training cohort and 7263 were in the validation cohort. The median age (25th–75th percentile) was 70 (58–82) for the entire cohort, 43.2% were female, and 6.7% experienced inpatient mortality. In the overall cohort, 7621 (35.0%) patients had heart failure with reduced ejection fraction (EF $\leq$ 40%), 1271 (5.8%) had heart failure with mildly reduced EF (EF 41–49%), and 12 910 (59.2%) had heart failure with preserved EF (EF $\geq$ 50%). Random forest models in the validation cohort demonstrated a *c*-statistic (95% confidence interval) of 0.96 (0.95–0.97), sensitivity (SN) of 87.3%, and specificity (SP) of 90.6% for the prediction of in-hospital mortality. Models for those with HFrEF demonstrated a *c*-statistic of 0.96 (0.94–0.98), SN 88.2%, and SP 91.0%, and those for patients with HFpEF showed a *c*-statistic of 0.95 (0.93–0.97), SN 87.4%, and SP 89.5% for predicting in-hospital mortality. The random forest model significantly outperformed logistic regression (*c*-statistic 0.87, SN 75.9%, and SP 86.9%), and current existing risk scores including the Acute Decompensated Heart Failure National Registry risk score (*c*-statistic of 0.70, SN 69%, and SP 62%), and the Get With the Guidelines-Heart Failure risk score (*c*-statistic 0.69, SN 67%, and SP 63%); $P < 0.001$ for comparison. Machine learning models built from commonly recorded patient information can accurately predict in-hospital mortality among patients hospitalized with a diagnosis of heart failure.

Keywords Machine learning; Heart failure; Risk prediction

Received: 28 August 2023; Revised: 31 January 2024; Accepted: 25 March 2024

**Correspondence to: Ali Nsair, Ahmanson-UCLA Cardiomyopathy Center, Ronald Reagan-UCLA Medical Center, MRL 3-760, 675 C.E. Young Dr., Los Angeles, CA 90095-1760, USA. Email: ansair@mednet.ucla.edu*
Zina Jawadi, BS, MS; Rosemary He, BA, MS; and Pratyaksh K. Srivastava, MD, contributed equally to this work.

Introduction

Heart failure (HF) patients experience frequent hospital admissions with in-hospital mortality rates ranging from 2–5%.[1,2] Over the years, there has been a significant effort devoted to understanding the factors that contribute to morbidity and mortality in this high-risk population. This important work has led to the development of a number of risk prediction tools, including scores such as Get With the Guidelines Heart Failure (GWTG-HF) and Acute Decompensated Heart Failure National Registry (ADHERE), that have significantly improved our ability to predict adverse outcomes.[3–5] While these models have accurately predicted in-hospital mortality with good discrimination, there still exists room for additional optimization.

Machine learning (ML), a domain of artificial intelligence, utilizes computer software to recognize complex patterns in large datasets.[6] In medicine, its use has been diverse, with applications ranging from image interpretation to risk prediction.[7–9] The main benefit of ML derives from its ability to make multi-level correlations using large amounts of data, and in some circumstances, it has been shown to outperform existing risk prediction algorithms.[6,7]

In an attempt to improve on currently existing risk prediction models, we propose a novel ML model to predict in-hospital mortality among patients hospitalized with a diagnosis of HF and compare it to existing HF risk prediction tools.

Methods

The study population included all adult (age$>$18 years) patients admitted with a diagnosis of HF who had an ejection fraction (EF) available, and who had a hospital stay of $>$ 24 hours from two large medical centres between 7 May 2013 and 26 April 2022. The first centre is a large 520 bed quaternary care centre located on a university campus (centre 1). The second is a 281 bed tertiary care hospital in a community setting (centre 2). Patient data were extracted from the electronic medical record into a structured query language database. The database contains deidentified information on all patient records available in these two centres. HF diagnosis and medical comorbidities were extracted using International Classification of Diseases (ICD) 9 and 10 codes. Patients were stratified by in-hospital mortality status, and patient information, including demographics, medical comorbidities, vitals on admission, and labs, were compared using Kruskal–Wallis and Chi-squared tests for categorical and continuous variables, respectively.

ML predictive models were created utilizing random forest modelling to predict in-hospital mortality among the cohort. Random forest is a bagging ML model, which makes predictions based on the results from a set of decision trees. Each decision tree is a predictor on a subspace of the entire dataset chosen randomly and independently, and random forest models take the majority prediction among all decision trees. We trained our models with 256 trees and a maximum depth of 1000. We take several approaches to detect and prevent overfitting, including k-fold validation and a validation on a validation cohort. For training, we take 10-fold cross validation, which helps assess generalization of the model to a random subset of the samples and reduces variance associated with a single train-test split.

Models were trained and optimized on data prior to 1 June 2020 from centre 1. Data recorded after 1 June 2020 were used for model evaluation. Data from centre 2 were used for external validation.

Patient demographics, characteristics, medical comorbidities, vitals, and labs throughout the hospitalization were included in the ML models (*Supporting information*, *Table S1*). The admission was split temporally into 5 times points, with labs and vitals recorded at each point and included in the model. Time point 1 represented labs/vitals closest to admission, time point 2 values closest to 25% of the way into the admission, time point 3 values closest to 50% of the way into the admission, time point 4 values closest to 75% of the way into admission, and time point 5 values closest to time of discharge or time of death. Missing values were denoted with a special numerical indicator and included in the model. Missingness of the cohort is shown in *Table S2*. For certain labs (total cholesterol, low density lipoprotein, high density lipoprotein, triglycerides, haemoglobin A1c), values were pulled from prior admission if no values from the patient's current admission were available. Values were input into the model up to 12 hours prior to the outcome.

Random forest ML models were created first utilizing the entire training cohort and then specifically for those with HF with reduced EF (HFrEF, EF$\leq$40%), HF with mildly reduced EF (HFmrEF, EF 41–49%), and HF with preserved EF (EF $\geq$ 50%). Patients were stratified into these three groups based on the EF on an echocardiogram obtained during the hospitalization from which data were pulled.

For comparison, a logistic regression model was created utilizing the same variables from the ML Random Forest model (*Table S1*). The models were then compared with two existing risk scores: the Get With the Guidelines-Heart Failure (GWTG-HF) risk score and the Acute Decompensated Heart Failure National Registry (ADHERE) risk score. The GWTG-HF risk score is composed of the following variables: systolic blood pressure (mmHg), blood urea nitrogen (mg/dL), sodium (mEq/L), age (years), heart rate (beats per minute), chronic obstructive pulmonary artery disease history, and Black race. These variables were used to calculate a GWTG-HF risk score using an existing previously published calculator.[4] This score was then used as a continuous variable in a logistic regression predicting in-hospital mortality. For the ADHERE risk score comparison, the variables that make up the score [systolic blood pressure (mm Hg), blood urea nitrogen (mg/dL), age (years), and heart rate (beats per minute)] were used in a logistic regression model.

In order to evaluate the predictive abilities of the models above, receiver operating curves were created, and f1 scores and *c*-statistics were determined for each model. We used DeLong tests with an alpha value of 95% to determine the confidence intervals for the *c*-statistics and f1 scores. Thresholds were set by calculating the *g*-mean and Youden's *J* statistic, and subsequent model sensitivity and specificity were obtained. Precision–recall curves were also created, and an average precision score was calculated. Top predictors of the models were also derived. These predictors were computed on the centre 1 test set data.

In order to validate the models, the analysis above was repeated on a validation cohort at a second medical centre (centre 2). Medical data were extracted from EPIC (Epic Systems Corporation). Data analysis was performed using Python. A two-sided *P* value of <0.05 was used to determine statistical significance. The study was approved by a local institutional review board.

Table 1 General characteristics of the cohort stratified by in-hospital mortality

	Overall *N* = 21 802	No in-hospital mortality *N* = 20 348	In-hospital mortality *N* = 1454	*P* value[a]
Demographics				
Age, years	70 (58–82)	70 (58–82)	70 (58–81)	0.84
Female, *n* (%)	9428 (43.2%)	8841 (43.4%)	587 (40.4%)	0.02
Body mass index, kg/m^2	25.8 (22.3–30.4)	25.9 (22.3–30.4)	25.6 (22.1–30.4)	0.24
Race, *n* (%)				<0.001
Asian and Pacific Islander	1768 (8.1%)	1625 (8.0%)	143 (9.8%)	
Indigenous	78 (0.4%)	72 (0.4%)	6 (0.4%)	
Latino	3964 (18.2%)	3664 (18.0%)	300 (20.6%)	
Non-Hispanic Black	3185 (14.6%)	3022 (14.9%)	163 (11.2%)	
Non-Hispanic White	10 957 (50.3%)	10 262 (50.4%)	695 (47.8%)	
Unknown	1850 (8.5%)	1703 (8.4%)	147 (10.1%)	
Heart failure, *n* (%)				<0.001
Heart failure with reduced ejection fraction (EF ≤ 40%)	7621 (35.0%)	6979 (34.3%)	642 (44.2%)	
Heart failure with mid-range ejection fraction (EF 41–49%)	1271 (5.8%)	1191 (5.9%)	80 (5.5%)	
Heart failure with preserved ejection fraction (EF ≥ 50%)	12 910 (59.2%)	12 178 (59.8%)	732 (50.3%)	
Location, *n* (%)				<0.001
Centre 1 cohort	14 539 (66.7%)	13 479 (66.2%)	1060 (72.9%)	
Centre 2 validation cohort	7263 (33.3%)	6869 (33.8%)	394 (27.1%)	
Medical comorbidities/history, *n* (%)				
Aortic valve disorders	4008 (18.4%)	3793 (18.6%)	215 (14.8%)	<0.001
Other valvular heart disease	9377 (43.0%)	8817 (43.3%)	560 (38.5%)	<0.001
Atrial fibrillation/flutter	8781 (40.3%)	8216 (40.4%)	565 (38.9%)	0.27
Cardiac transplant	1338 (6.1%)	1305 (6.4%)	33 (2.3%)	<0.001
Other organ transplant	1788 (8.2%)	1633 (8.0%)	155 (10.7%)	<0.001
Cerebrovascular disease	5022 (23.0%)	4686 (23.0%)	336 (23.1%)	0.97
Chronic obstructive pulmonary disease (COPD)	8434 (38.7%)	7923 (38.9%)	511 (35.1%)	0.01
Chronic kidney disease	9551 (43.8%)	8884 (43.7%)	667 (45.9%)	0.10
Coagulopathy	5483 (25.1%)	4910 (24.1%)	573 (39.4%)	<0.001
History of coronary artery bypass graft	2624 (12.0%)	2455 (12.1%)	169 (11.6%)	0.62
Coronary artery disease/acute coronary syndrome	12 072 (55.4%)	11 324 (55.7%)	748 (51.4%)	0.002
Congenital heart disease	1855 (8.5%)	1757 (8.6%)	98 (6.74%	0.01
Dementia	674 (3.1%)	640 (3.1%)	34 (2.34%)	0.09
Depression/bipolar disorder	5532 (25.4%)	5208 (25.6%)	324 (22.3%)	0.01
Diabetes mellitus	8025 (36.8%)	7525 (37.0%)	500 (34.4%)	0.05
Dialysis/history of dialysis	1930 (8.9%)	1753 (8.6%)	177 (12.2%)	<0.001
Drug use	2159 (9.9%)	2044 (10.0%)	115 (7.9%)	0.01
Ejection fraction	55.0 (35.0–60.0)	55.0 (35.0–60.0)	50.0 (25.0–60.0)	<0.001
Ever smoker, *n* (%)	9548 (43.8%)	9017 (44.3%)	531 (36.5%)	<0.001
Human immunodeficiency virus (HIV)	162 (0.7%)	153 (0.8%)	9 (0.6%)	0.68
Hypertension	15 280 (70.1%)	14 410 (70.8%)	870 (59.8%)	<0.001
Hyperlipidaemia	11 729 (53.8%)	11 099 (54.5%)	630 (43.3%)	<0.001
Iron-deficiency anaemia	4311 (19.8%)	4056 (19.9%)	255 (17.5%)	0.03
Liver disease	5058 (23.2%)	4521 (22.2%)	537 (36.9%)	<0.001
Obesity	4393 (20.1%)	4177 (20.5%)	216 (14.9%)	<0.001
Pacemaker/ICD/CRT	4025 (18.5%)	3788 (18.6%)	237 (16.3%)	0.03
History of percutaneous coronary intervention	3266 (15.0%)	3054 (15.0%)	212 (14.6%)	0.66
Peripheral vascular disease/vasculitis	8448 (38.7%)	7923 (38.9%)	525 (36.1%)	0.03
Psychosis/schizophrenia	916 (4.2%)	855 (4.2%)	61 (4.2%)	0.99
Rheumatoid arthritis/collagen vascular/connective tissue disease	2157 (9.9%)	2025 (10.0%)	132 (9.1%)	0.28
Venous thromboembolism	3308 (15.2%)	3030 (14.9%)	278 (19.1%)	<0.001
Ventricular tachycardia	3701 (17.0%)	3409 (16.8%)	292 (20.1%)	0.001

Abbreviations: CRT, cardiac resynchronization therapy; ICD, implantable cardioverter defibrillator.

[a] Continuous variables presented as median (25th–75th percentile). Categorical variables presented as *n* (%).Continuous and categorical variables compared by Kruskal–Wallis and Chi-squared tests, respectively.

Results

The overall cohort consisted of 21 802 patients, of which 14 539 were from centre 1 and 7263 were from centre 2 (validation cohort). In the overall cohort, 7621 (35.0%) patients had HFrEF, 1271 (5.8%) had HFmrEF, and 12 910 (59.2%) had HFpEF. The median age (25th–75th percentile) of the cohort was 70 (58–82) years, 43.2% of the group was female, and 6.7% experienced in hospital patient mortality (*Table 1*). The cohort was then stratified by in-hospital mortality status. Compared with patients who survived their hospitalization, patients admitted with a diagnosis of HF who experienced in-hospital mortality were less likely to be female (40.4% vs. 43.4%), more likely to be Asian/Pacific Islander (9.8% vs. 8.0%), and less likely to be non-Hispanic Black (11.2% vs. 14.9%). Patients who experienced in-hospital mortality were more likely to have HFrEF (44.2% vs. 34.3%). They were further less likely to have certain comorbidities, including valvular disease, cardiac transplant, and chronic obstructive pulmonary disease, and more likely to have coagulopathy, history of dialysis, liver disease, malignancy, and ventricular tachycardia when compared with patients without in-hospital mortality (*Table 1*; *P* values for all comparisons above <0.05). When comparing admission characteristics, patients with in-hospital mortality were found to have a higher heart rate and lower blood pressure, as well as a higher white blood cell count, creatinine, blood glucose, and B-type natriuretic peptide. They were found to have lower haemoglobin, platelets, sodium, and cholesterol values (*Table 2*; all *P* values <0.05).

To evaluate the efficacy of our ML random forest model, we generated receiver operating characteristic curves (ROC) for both cohorts (*Table 3*). The evaluation metrics for centre 1 (training cohort) are shown in *Figures S1* and *S2* and *Table 3*. We highlight the performance of the constructed models using the validation cohort (centre 2) in *Figure 1* and *Table 3*. In the validation cohort, the random forest models were found to have a *c*-statistic (95% confidence interval) of 0.96 (0.95–0.97). At a threshold of 0.07, the sensitivity of the model was 87.3% and specificity 90.6% for predicting in-hospital mortality (*Figure 1*). The model was also found to have a precision score of 79.9% for the prediction of in-hospital mortality. *Figure 1* also demonstrates ROC and precision–recall curves for simple logistic regression using the same variables as the ML model (full logistic regression) as well as performance characteristics for the GWTG-HF and ADHERE risk scores. A DeLong test demonstrated a statistically significant difference between the four models (*Table 3*).

Figure 2 demonstrates the ROC curves for the validation cohort models stratified by EF. For those with HFrEF, the *c*-statistic for the random forest ML model was 0.96

Table 2 Vitals and labs of the cohort stratified by in-hospital mortality

	Overall *N* = 21 802	No in-hospital mortality *N* = 20 348	In-hospital mortality *N* = 1454	*P* value[a]
Vitals on admission				
Temperature (°C)	98.0 (97.5–98.5)	98.0 (97.5–98.5)	97.9 (97.3–98.5)	<0.001
Heart rate (beats per minute)	83.0 (70.0–100.0)	83.0 (70.0–99.0)	91.0 (76.0–108.0)	<0.001
Systolic blood pressure (mmHg)	130.0 (112.0–150.0)	131.0 (113.0–150.0)	116.0 (100.0–136.0)	<0.001
Diastolic blood pressure (mmHg)	74.0 (64.0–86.0)	75.0 (64.0–87.0)	69.0 (58.0–81.0)	<0.001
Mean arterial pressure (mmHg)	112.0 (97.7–127.7)	112.7 (98.3–128.3)	100.3 (87.3–116.7)	<0.001
Supplementary O_2, *n* (%)	97.0 (95.0–99.0)	97.0 (95.0–99.0)	97.0 (94.0–99.0)	<0.001
Labs on admission				
White blood cell count (K/μL)	8.33 (6.3–11.3)	8.3 (6.3–11.1)	9.8 (6.6–14.4)	<0.001
Haemoglobin (g/dL)	11.5 (9.7–13.2)	11.6 (9.8–13.3)	10.3 (8.8–12.3)	<0.001
Haematocrit (%)	35.3 (30.5–40.1)	35.5 (30.8–40.2)	32.3 (27.8–37.5)	<0.001
Platelets (K/μL)	202.0 (151.0–263.0)	204.0 (153.0–264.0)	172.0 (101.3–247.8)	<0.001
Sodium	139.0 (135.0–141.0)	139.0 (136.0–141.0)	137.0 (133.0–140.0)	<0.001
Potassium	4.2 (3.9–4.6)	4.2 (3.9–4.6)	4.3 (3.9–4.9)	<0.001
Chloride	101.0 (97.0–105.0)	102.0 (98.0–105.0)	99.0 (95.0–103.0)	<0.001
Bicarbonate/carbon dioxide	24.0 (21.0–26.0)	24.0 (21.0–26.0)	22.0 (19.0–26.0)	<0.001
Blood urea nitrogen (BUN)	24.0 (16.0–38.0)	23.0 (16.0–36.0)	33.0 (20.0–53.0)	<0.001
Creatinine	1.2 (0.9–1.8)	1.2 (0.9–1.8)	1.6 (1.0–2.7)	<0.001
Glucose	122.0 (102.0–160.0)	122.0 (102.0–159.0)	129.5 (105.0–179.3)	<0.001
Calcium	8.9 (8.5–9.3)	9.0 (8.6–9.3)	8.8 (8.2–9.2)	<0.001
Magnesium	1.7 (1.5–1.9)	1.7 (1.5–1.8)	1.7 (1.5–2.0)	<0.001
B-type natriuretic peptide (BNP)	479.0 (205.0–1067.0)	466.0 (199.0–1024.0)	768.0 (337.0–1871.3)	<0.001
Labs				
Total cholesterol	143.0 (115.0–174.0)	144.0 (116.0–175.0)	124.0 (93.0–156.0)	<0.001
Low-density lipoprotein	71.0 (50.0–96.0)	72.0 (51.0–96.0)	56.0 (37.0–81.0)	<0.001
High-density lipoprotein	44.0 (34.0–57.0)	45.0 (35.0–57.0)	37.0 (23.0–52.0)	<0.001
Triglycerides	100.0 (72.0–143.0)	100.0 (72.0–143.0)	101.0 (72.0–152.0)	0.13
Haemoglobin A1c	6.0 (5.5–6.7)	6.0 (5.5–6.7)	5.9 (5.5–6.7)	0.01

[a] Continuous variables presented as median (25th–75th percentile). Categorical variables presented as *n* (%).Continuous and categorical variables compared by Kruskal–Wallis and Chi-squared tests, respectively.

Table 3 Model performance parameters

	AUC	Sensitivity	Specificity	Precision score	DeLong*
Centre 1 (training cohort) all ejection fraction					
Random forest	0.99 (0.98–0.99)	93.3%	97.3%	93.8%	N/A
Logistic regression	0.93 (0.91–0.95)	81.2%	95.0%	82.2%	<0.001
ADHERE	0.68 (0.65–0.71)	75.9%	49.4%	15.0%	<0.001
GWTG-HF	0.69 (0.66–0.72)	64.5%	63.0%	14.6%	<0.001
Centre 1 (training cohort) heart failure with reduced ejection fraction (ejection fraction ≤40%)					
Random forest	0.99 (0.98–1.00)	94.1%	96.8%	95.1%	N/A
Logistic regression	0.90 (0.86–0.94)	83.9%	86.9%	71.1%	3.11E-03
ADHERE	0.68 (0.63–0.72)	59.3%	66.8%	18.1%	<0.001
GWTG-HF	0.64 (0.59–0.69)	52.5%	68.5%	15.8%	<0.001
Centre 1 (training cohort) heart failure with mildly reduced ejection fraction (ejection fraction 41–49%)					
Random forest	1.00 (1.00–1.00)	100.0%	99.6%	99.5%	N/A
Logistic regression	0.73 (0.59–0.87)	79.0%	67.6%	31.6%	2.33E-02
ADHERE	0.67 (0.54–0.80)	79.0%	49.6%	18.7%	1.59E-03
GWTG-HF	0.66 (0.53–0.80)	68.4%	62.3%	15.6%	2.59E-03
Centre 1 (training cohort) heart failure with preserved ejection fraction (ejection fraction ≥50%)					
Random forest	0.98 (0.97–1.00)	94.5%	93.9%	92.5%	N/A
Logistic regression	0.91 (0.88–0.95)	77.9%	94.7%	77.1%	5.93E-03
ADHERE	0.68 (0.64–0.72)	60.0%	66.7%	13.2%	<0.001
GWTG-HF	0.57 (0.53–0.62)	44.1%	71.1%	9.0%	<0.001
Centre 2 (validation cohort) all ejection fraction					
Random forest	0.96 (0.95–0.97)	87.3%	90.6%	79.9%	N/A
Logistic regression	0.87 (0.85–0.89)	75.9%	86.9%	63.4%	<0.001
ADHERE	0.70 (0.67–0.72)	69.0%	61.7%	11.4%	<0.001
GWTG-HF	0.69 (0.67–0.72)	66.8%	62.9%	11.1%	<0.001
Centre 2 (validation cohort) heart failure with reduced ejection fraction (ejection fraction ≤40%)					
Random forest	0.96 (0.94–0.98)	88.2%	91.0%	81.6%	N/A
Logistic regression	0.85 (0.81–0.89)	64.6%	90.1%	54.8%	<0.001
ADHERE	0.69 (0.65–0.74)	70.8%	60.9%	15.8%	<0.001
GWTG-HF	0.55 (0.50–0.60)	41.0%	68.5%	8.3%	<0.001
Centre 2 (validation cohort) heart failure with mildly reduced ejection fraction (ejection fraction 41–49%)					
Random forest	0.95 (0.88–1.00)	90.0%	95.3%	83.1%	N/A
Logistic regression	0.80 (0.69–0.90)	75.0%	77.9%	20.2%	1.72E-01
ADHERE	0.70 (0.58–0.81)	70.0%	72.6%	9.8%	2.17E-02
GWTG-HF	0.71 (0.60–0.83)	75.0%	67.3%	10.7%	3.53E-02
Centre 2 (validation cohort) heart failure with preserved ejection fraction (ejection fraction ≥50%)					
Random forest	0.95 (0.93–0.97)	87.4%	89.5%	76.9%	N/A
Logistic regression	0.84 (0.81–0.88)	73.0%	84.5%	54.0%	<0.001
ADHERE	0.70 (0.66–0.73)	66.5%	62.8%	9.9%	<0.001
GWTG-HF	0.59 (0.55–0.63)	35.2%	79.7%	7.7%	<0.001

**P* value, compared with random forest.
Abbreviations: ADHERE, Acute Decompensated Heart Failure National Registry; AUC, area under the curve; GWTG-HF, Get With the Guidelines Heart Failure.

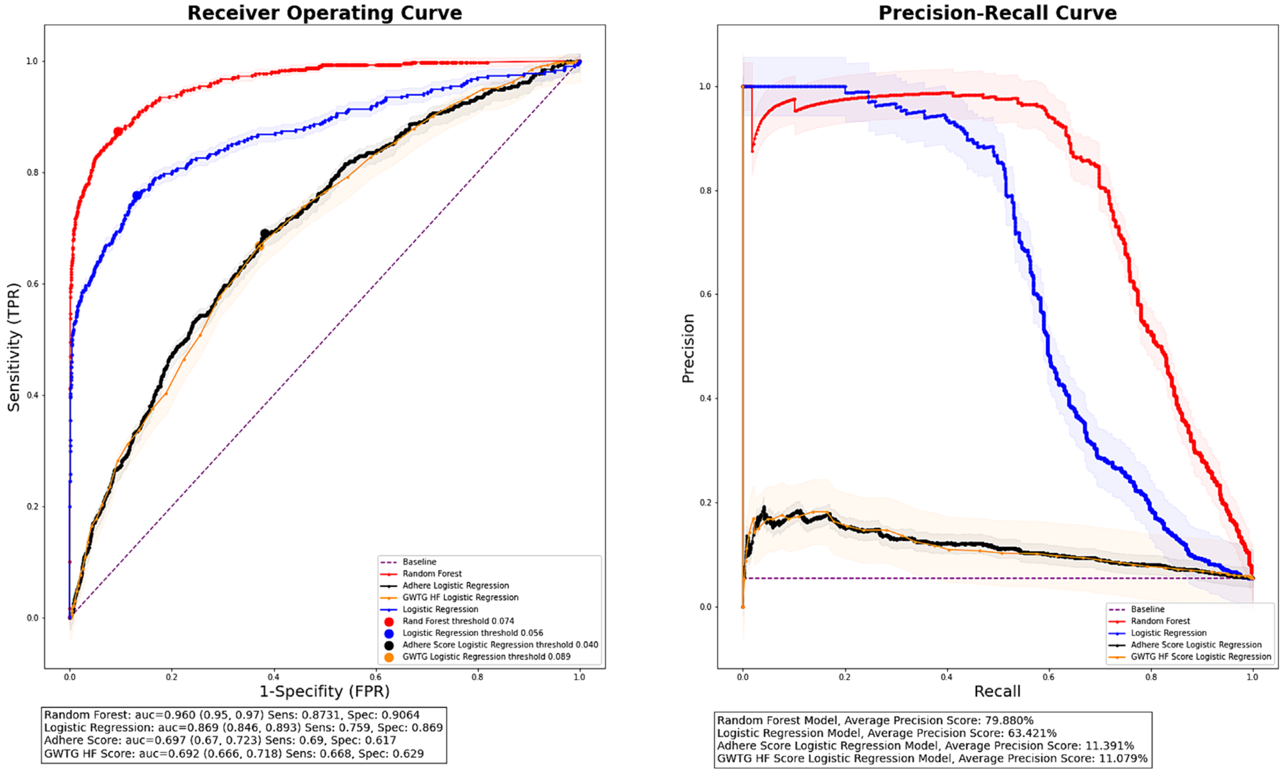
(0.94–0.98), the average precision 81.6%, the sensitivity 88.2%, and the specificity 91% at a threshold of 0.10 for the outcome of in-hospital mortality. For those with HFmrEF, the *c*-statistic of the ML model was 0.95 (0.88–1.00), the average precision 83.1%, and the sensitivity and specificity 90% and 95%, respectively, at a threshold of 0.14. For those with HFpEF, the *c*-statistic of the ML model was 0.95 (0.93–0.97), the average precision 76.9%, and the sensitivity and specificity 87.4% and 89.5%, respectively, at a threshold of 0.06. For those with HFrEF, HFmrEF, and HFpEF, the *c*-statistic s for the ML models were statistically different than those of the full logistic regression, GWTG-HF, and ADHERE risk prediction models. A comparison of the *c*-statistics, SN, SP, precision scores, and DeLong statistics for each of the models stratified by EF is shown in *Table 3*. Top predictors of the random forest models at centre 1 stratified by EF are shown in *Table S3*.

Discussion

HF hospitalization is associated with significant morbidity and mortality, and outcomes among this high-risk population are not optimally predicted by current risk scores. Using a cohort of 21 802 patients (14 439 initial cohort, 7 263 validation cohort) hospitalized with a diagnosis of HF between 7 May 2013 and 26 April 2022, we constructed ML models to predict in-hospital mortality, and compared them with currently existing in-hospital HF risk prediction scores. We demonstrate high sensitivity, specificity, and accuracy of our ML models to predict in-hospital mortality among those admitted with a diagnosis of HF across the entire spectrum of EF and show improvement when compared with both simple logistic regression and with currently existing risk prediction scores.

ML is a form of artificial intelligence that centres on the creation of algorithms with the ability to learn through the

Figure 1 Receiver operating and precision–recall curves for centre 2 (validation cohort) by risk prediction method. Abbreviations: ADHERE, Acute Decompensated Heart Failure National Registry; AUC, area under the curve; FPR, false positive rate; GWTG HF, Get With the Guidelines Heart-Failure; Rand, Random Forest; TPR, true positive rate.

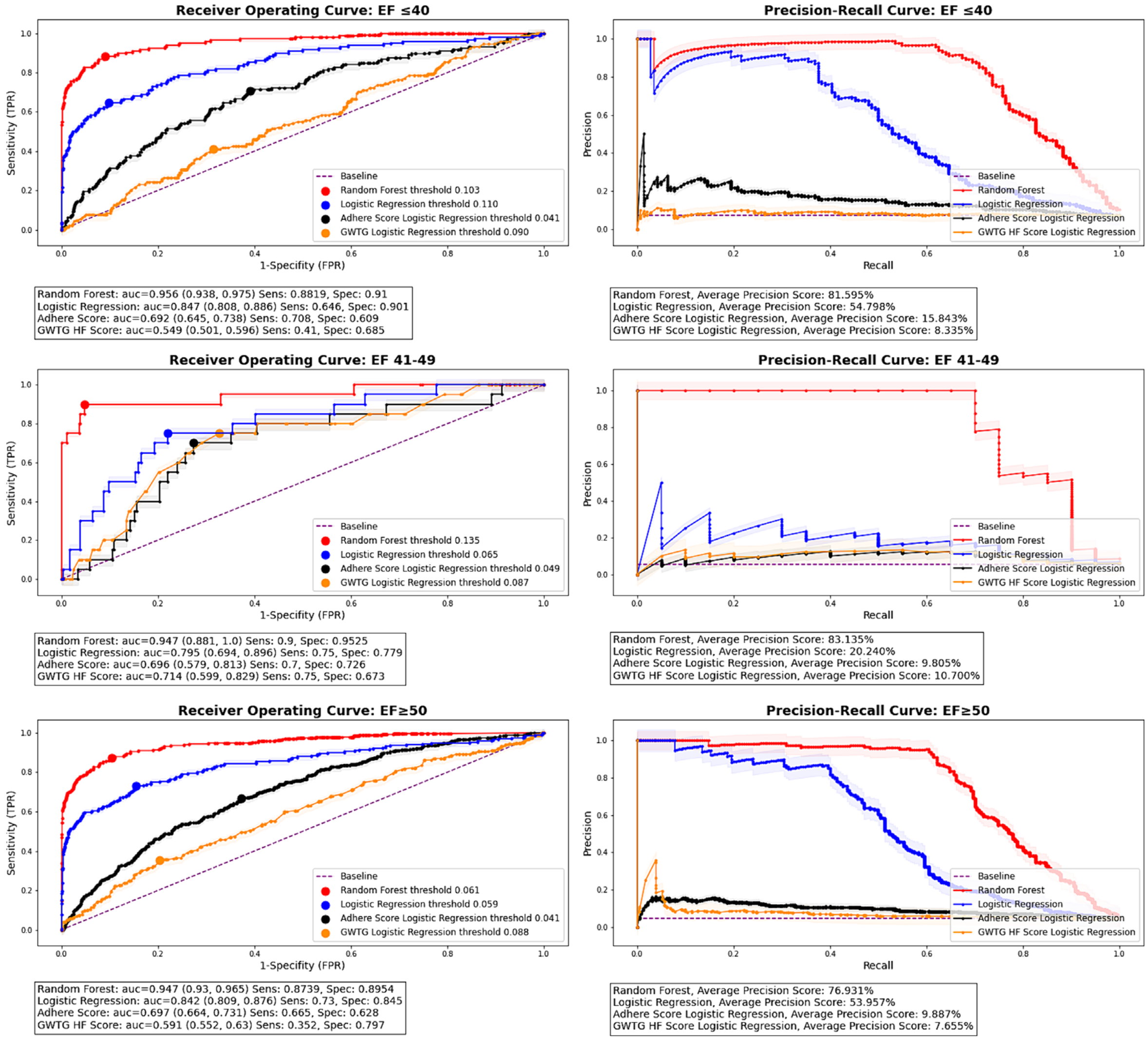


discovery of patterns.[6,7] ML algorithms can be divided into supervised learning, where models are created to predict specific outcomes, and unsupervised learning, where large groups of data are analysed with the intent of finding patterns.[7,10] By analysing large data sets, these computer-based algorithms are able to evaluate complex correlations to ultimately develop associations between a large number of variables.[11] For example, when creating a model, variables may not be individually predictive, yet may be able to accurately model an outcome when taken together in different combinations with each other. The evaluation of such complex patterns is often beyond the scope of traditional statistical methodologies, such as logistic regression, and therefore, ML presents an attractive opportunity to improve medical risk prediction.[7]

ML has been applied to a wide range of medical applications ranging from electrocardiogram and echocardiogram interpretation,[12–15] to detection of possible malignancies in chest radiographs,[16–18] and to prediction of outcomes in patients with various disease processes.[19] In the field of HF, ML has also been shown to be of value.[20] In one study of HFpEF patients, an unsupervised ML model was shown to predict survival and was shown to improve on current commonly used models for risk assessment.[21] In another study of both admitted patients and outpatients at a large academic medical centre, an ML algorithm was able to accurately predict mortality and was used to generate a mortality risk score that has been shown to be more accurate than previously existing risk prediction models.[11]

In our study, we demonstrate the power of ML to accurately predict in-hospital mortality among patients admitted with a diagnosis of HF regardless of EF. We also show an improvement when comparing ML to traditional logistic regression and to logistic regression using variables from previously existing risk scores such as GWTG-HF and ADHERE in both the initial and validation cohorts. Although prior ML models have been created to predict outcomes in those hospitalized with HF, ours is unique in the utilization of temporal changes in vital signs and labs (i.e., inputting labs and vital signs from five distinct time points during every admission). This input helps capture important information on patient trajectories while admitted and, in doing so, is able to add an additional dimension to risk prediction not present in the majority of previously constructed HF ML models. This added dimension may help explain why our model appears to have superior ability (*c*-statistic and sensitivity) in predicting mortality compared with nearly 30other HF models evaluated in different cohorts in a recent meta-analysis performed by Mpanya *et al*.[20] The utilization of temporal variables throughout a patient's hospitalization may allow for the ultimate creation of

Figure 2 Receiver operating and precision–recall curves for centre 2 (validation cohort) stratified by ejection fraction. Abbreviations: ADHERE, Acute Decompensated Heart Failure National Registry; AUC, area under the curve; EF, ejection fraction; FPR, false positive rate; GWTG HF, Get With the Guidelines-Heart Failure; Rand, random forest; TPR, true positive rate.



Random Forest: auc=0.956 (0.938, 0.975) Sens: 0.8819, Spec: 0.91
Logistic Regression: auc=0.847 (0.808, 0.886) Sens: 0.646, Spec: 0.901
Adhere Score: auc=0.692 (0.645, 0.738) Sens: 0.708, Spec: 0.609
GWTG HF Score: auc=0.549 (0.501, 0.596) Sens: 0.41, Spec: 0.685

Random Forest, Average Precision Score: 81.595%
Logistic Regression, Average Precision Score: 54.798%
Adhere Score Logistic Regression, Average Precision Score: 15.843%
GWTG HF Score Logistic Regression, Average Precision Score: 8.335%

Random Forest: auc=0.947 (0.881, 1.0) Sens: 0.9, Spec: 0.9525
Logistic Regression: auc=0.795 (0.694, 0.896) Sens: 0.75, Spec: 0.779
Adhere Score: auc=0.696 (0.579, 0.813) Sens: 0.7, Spec: 0.726
GWTG HF Score: auc=0.714 (0.599, 0.829) Sens: 0.75, Spec: 0.673

Random Forest, Average Precision Score: 83.135%
Logistic Regression, Average Precision Score: 20.240%
Adhere Score Logistic Regression, Average Precision Score: 9.805%
GWTG HF Score Logistic Regression, Average Precision Score: 10.700%

Random Forest: auc=0.947 (0.93, 0.965) Sens: 0.8739, Spec: 0.8954
Logistic Regression: auc=0.842 (0.809, 0.876) Sens: 0.73, Spec: 0.845
Adhere Score: auc=0.697 (0.664, 0.731) Sens: 0.665, Spec: 0.628
GWTG HF Score: auc=0.591 (0.552, 0.63) Sens: 0.352, Spec: 0.797

Random Forest, Average Precision Score: 76.931%
Logistic Regression, Average Precision Score: 53.957%
Adhere Score Logistic Regression, Average Precision Score: 9.887%
GWTG HF Score Logistic Regression, Average Precision Score: 7.655%

dynamic risk prediction that is able to continuously monitor for changes in a patient's clinical condition during hospitalization to provide updated risk metrics as the patient's medical course progresses. This may help flag hospitalized HF patients at increased risk of adverse outcomes and may be useful in triggering interventions such as cardiology or advanced HF consultation or upgrade to a higher level of care.

Limitations

Although the models were validated at a second hospital, overall generalizability may still be limited, and further validation in a diverse set of populations is needed. Our cohort is also limited to those who had an EF obtained during their hospitalization, which further limits generalizability. The patients included in this study were hospitalized with a diagnosis of HF, but due to the design of the database, we were unable to confirm that this was the primary diagnosis. The models created require input from many variables and so would have to be integrated into an electronic medical record and calculated by computer, as manually inputting such a large amount of information would not be feasible.

Conclusions

Here we demonstrate the creation of accurate, sensitive, and specific ML models to predict in-hospital mortality among patients hospitalized for HF irrespective of EF and demonstrate improvement when compared with traditional logistic regression and commonly used risk prediction scores. These models, when used in the appropriate context, may help flag patients hospitalized with a diagnosis of HF at high risk for inpatient mortality.

Code availability

We include the Python scripts used for defining and training the models, and more specific data input formats in the public GitHub repository (link).

Conflict of interest

Dr. Fonarow reports consulting from Abbott, Amgen, AstraZeneca, Bayer, Cytokinetics, Eli Lilly, Johnson & Johnson, Medtronic, Merck, Novartis, and Pfizer.

Supporting information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Figure S1: Receiver Operating and Precision-Recall Curves for Center 1 by Risk Prediction Method.
Abbreviations: AUC, Area Under the Curve; ADHERE, Acute Decompensated Heart Failure National Registry; FPR, False Positive Rate; GWTG HF, Get With the Guidelines Heart-Failure; Rand, Random Forest; TPR, True Positive Rate.
Figure S2: Receiver Operating and Precision-Recall Curves for Center 1 Stratified by Ejection Fraction.
Abbreviations: AUC, Area Under the Curve; ADHERE, Acute Decompensated Heart Failure National Registry; EF, Ejection Fraction; FPR, False Positive Rate; GWTG HF, Get With the Guidelines-Heart Failure; Rand, Random Forest; TPR, True Positive Rate.
Table S1: List of Variables Used in Logistic Regression and Machine Learning Models.
Table S2. Missingness of the Cohort.
Table S3. Top Predictors of Random Forest Models at Center 1 Stratified by Ejection Fraction.

References

1. Salah HM, Minhas AMK, Khan MS, Khan SU, Ambrosy AP, Blumer V, *et al*. Trends and characteristics of hospitalizations for heart failure in the United States from 2004 to 2018. *ESC Heart Fail* 2022;**9**: 947-952. doi:10.1002/ehf2.13823
2. Abraham WT, Adams KF, Fonarow GC, Costanzo MR, Berkowitz RL, LeJemtel T, *et al*. In-hospital mortality in patients with acute decompensated heart failure requiring intravenous vasoactive medications: an analysis from the acute decompensated heart failure National Registry (ADHERE). *J Am Coll Cardiol* 2005;**46**: 57-64. doi:10.1016/j.jacc.2005.03.051
3. Lagu T, Pekow PS, Shieh MS, Stefan M, Pack QR, Kashef MA, *et al*. Validation and comparison of seven mortality prediction models for hospitalized patients with acute decompensated heart failure. *Circ Heart Fail* 2016;**9**:9. doi:10.1161/CIRCHEARTFAILURE.115.002912
4. Peterson PN, Rumsfeld JS, Liang L, Albert NM, Hernandez AF, Peterson ED, *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. doi:10.1161/CIRCOUTCOMES.109.854877
5. Fonarow GC, Adams KF Jr, Abraham WT, Yancy CW, Boscardin WJ, ADHERE Scientific Advisory Committee, Study Group, and Investigators. Risk stratification for in-hospital mortality in acutely decompensated heart failure: classification and regression tree analysis. *JAMA* 2005;**293**:572-580. doi:10.1001/jama.293.5.572
6. Rowe M. An introduction to machine learning for clinicians. *Acad Med* 2019;**94**:1433-1436. doi:10.1097/ACM.0000000000002792
7. Deo RC. Machine learning in medicine. *Circulation* 2015;**132**:1920-1930. doi:10.1161/CIRCULATIONAHA.115.001593
8. May M. Eight ways machine learning is assisting medicine. *Nat Med* 2021;**27**: 2-3. doi:10.1038/s41591-020-01197-2
9. Sidey-Gibbons JAM, Sidey-Gibbons CJ. Machine learning in medicine: a practical introduction. *BMC Med Res Methodol* 2019;**19**:64. doi:10.1186/s12874-019-0681-4
10. Jovel J, Greiner R. An introduction to machine learning approaches for biomedical research. *Front Med (Lausanne)* 2021;**8**:771607. doi:10.3389/fmed.2021.771607
11. Adler ED, Voors AA, Klein L, Macheret F, Braun OO, Urey MA, *et al*. Improving risk prediction in heart failure using machine learning. *Eur J Heart Fail* 2020;**22**: 139-147. doi:10.1002/ejhf.1628
12. Gorodeski EZ, Ishwaran H, Kogalur UB, Blackstone EH, Hsich E, Zhang ZM, *et al*. 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. doi:10.1161/CIRCOUTCOMES.110.959023
13. Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, *et al*. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nat Commun* 2020;**11**:1760. doi:10.1038/s41467-020-15432-4
14. Aziz S, Ahmed S, Alouini MS. ECG-based machine-learning algorithms for heartbeat classification. *Sci Rep* 2021; **11**:18738. doi:10.1038/s41598-021-97118-5
15. Ouyang D, He B, Ghorbani A, Yuan N, Ebinger J, Langlotz CP, *et al*. Video-based AI for beat-to-beat assessment of cardiac function. *Nature* 2020;**580**:252-256. doi:10.1038/s41586-020-2145-8
16. Lu MT, Raghu VK, Mayrhofer T, Aerts H, Hoffmann U. Deep learning using chest radiographs to identify high-risk smokers for lung cancer screening computed tomography: development and validation of a prediction model. *Ann Intern Med* 2020;**173**: 704-713. doi:10.7326/M20-1868

17. Yoo H, Lee SH, Arru CD, Doda Khera R, Singh R, Siebert S, *et al*. AI-based improvement in lung cancer detection on chest radiographs: results of a multi-reader study in NLST dataset. *Eur Radiol* 2021;**31**:9664-9674. doi:10.1007/s00330-021-08074-7
18. Shimazaki A, Ueda D, Choppin A, Yamamoto A, Honjo T, Shimahara Y, *et al*. Deep learning-based algorithm for lung cancer detection on chest radiographs using the segmentation method. *Sci Rep* 2022;**12**:727. doi:10.1038/s41598-021-04667-w
19. Pappada SM. Machine learning in medicine: it has arrived, let's embrace it. *J Card Surg* 2021;**36**:4121-4124. doi:10.1111/jocs.15918
20. Mpanya D, Celik T, Klug E, Ntsinjana H. Predicting mortality and hospitalization in heart failure using machine learning: a systematic literature review. *Int J Cardiol Heart Vasc* 2021;**34**:100773. doi:10.1016/j.ijcha.2021.100773
21. Katz DH, Deo RC, Aguilar FG, Selvaraj S, Martinez EE, Beussink-Nelson L, *et al*. Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. *J Cardiovasc Transl Res* 2017;**10**:275-284. doi:10.1007/s12265-017-9739-z