



Electronic Health Record-Based Deep Learning Prediction of Death or Severe Decompensation in Heart Failure Patients

Martha M.O. McGilvray, MSt, MD,^a Jeffrey Heaton, PhD,^b Aixia Guo, PhD,^c M. Faraz Masood, MD,^a Brian P. Cupps, PhD,^a Marci Damiano, RN, MSN,^a Michael K. Pasque, MD,^a Randi Foraker, PhD^d

ABSTRACT

BACKGROUND Surgical mechanical ventricular assistance and cardiac replacement therapies, although life-saving in many heart failure (HF) patients, remain high-risk. Despite this, the difficulty in timely identification of medical therapy nonresponders and the dire consequences of nonresponse have fueled early, less selective surgical referral. Patients who would have ultimately responded to medical therapy are therefore subjected to the risk and life disruption of surgical therapy.

OBJECTIVES The purpose of this study was to develop deep learning models based upon commonly-available electronic health record (EHR) variables to assist clinicians in the timely and accurate identification of HF medical therapy nonresponders.

METHODS The study cohort consisted of all patients (age 18 to 90 years) admitted to a single tertiary care institution from January 2009 through December 2018, with International Classification of Disease HF diagnostic coding. Ensemble deep learning models employing time-series and densely-connected networks were developed from standard EHR data. The positive class included all observations resulting in severe progression (death from any cause or referral for HF surgical intervention) within 1 year.

RESULTS A total of 79,850 distinct admissions from 52,265 HF patients met observation criteria and contributed >350 million EHR datapoints for model training, validation, and testing. A total of 20% of model observations fit positive class criteria. The model C-statistic was 0.91.

CONCLUSIONS The demonstrated accuracy of EHR-based deep learning model prediction of 1-year all-cause death or referral for HF surgical therapy supports clinical relevance. EHR-based deep learning models have considerable potential to assist HF clinicians in improving the application of advanced HF surgical therapy in medical therapy nonresponders. (J Am Coll Cardiol HF 2022;10:637–647) © 2022 the American College of Cardiology Foundation. Published by Elsevier. All rights reserved.

Heart failure (HF) patients comprise the largest, most rapidly growing, and most expensive subset of patients with cardiovascular disease ([Central Illustration](#)).¹ Annual HF patient mortality in the United States ranges from 22%–30%,^{2,3} and among patients with New York Heart Association

functional class IV, annual mortality is >50%.⁴ The clinical presentation of HF commonly includes alarming congestive symptoms with severely impaired left ventricular function. Despite the gravity of their presentation, most HF patients will nonetheless demonstrate a favorable response to medical therapy.^{5–8}

From the ^aDivision of Cardiothoracic Surgery, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA;

^bSever Institute, McKelvey School of Engineering, Washington University in St. Louis, St. Louis, Missouri, USA; ^cVeterans Affairs St. Louis Health Care System—St. Louis, St. Louis, Missouri, USA; and the ^dInstitute for Informatics, Division of General Medical Sciences, School of Medicine, Washington University in St. Louis, St. Louis, Missouri, USA.

The authors attest they are in compliance with human studies committees and animal welfare regulations of the authors' institutions and Food and Drug Administration guidelines, including patient consent where appropriate. For more information, visit the [Author Center](#).

Manuscript received February 15, 2022; revised manuscript received May 12, 2022, accepted May 18, 2022.

ABBREVIATIONS AND ACRONYMS

AUC	= area under the curve
BNP	= brain natriuretic peptide
BP	= blood pressure
ECMO	= extracorporeal membrane oxygenation
EF	= ejection fraction
EHR	= electronic health record
HF	= heart failure
LOS	= length of stay
LSTM	= long short-term memory
PR	= precision recall
VAD	= ventricular assist device

Conversely, many medical therapy nonresponders progress to precipitous deterioration. This decline is often so rapid and unheralded that the consequences, sudden death or severe end-organ injury, preclude further salvage efforts.⁹ The difficulty in the timely identification of these medical therapy nonresponders and the dire consequences of this nonresponse lowers the clinician threshold for immediate referral for invasive surgical therapy.¹⁰

There are many surgical options in decompensated HF, including use of intra-aortic balloon pump assist, extracorporeal membrane oxygenator (ECMO) support, ventricular assist device (VAD) placement, and subsequent listing of appropriate candidates for cardiac transplantation. Although an aggressive therapeutic paradigm may assist in making this difficult referral decision, the inaccuracy inherent to a lowered threshold for highly invasive intervention results in many patients who would have responded favorably on medical therapy having unnecessary exposure to life-disruptive high-risk surgical therapy.^{9–11}

SEE PAGE 648

Unfortunately, frequently applied prognostic metrics, such as left ventricular ejection fraction (EF), catheter-based hemodynamics, and focused laboratory testing, do not independently demonstrate the predictive accuracy necessary to direct patient-specific clinical decision-making.^{5,12–18} Even the accuracy of some of the most predictive clinical variables, such as blood urea nitrogen and serum sodium,¹⁶ extends only to large subpopulation comparisons, not achieving—as lone predictors—sufficient accuracy to drive individual patient decision-making. Perhaps the most commonly used single clinical variable is echocardiography-based EF, which musters a 1-year mortality prediction C-statistic of only 0.52.⁵

In a meta-analysis including 39,372 HF patients with both reduced and preserved EF,¹⁸ 13 highly significant independent predictors of mortality were identified. All of these variables, along with many others (such as blood urea nitrogen, brain natriuretic peptide [BNP], and serum sodium),¹⁶ are routinely acquired in almost all HF patients and are readily accessible by machine learning models. The application of machine learning in large electronic health record (EHR) data sets, therefore, offers an attractive avenue for testing the intuitive hypothesis that combining the predictive power of these variables

might improve accuracy of personalized HF therapeutic decision-making.

To this end, we developed EHR-based deep learning models from HF patient observations gathered over a decade of admissions to a single institution. To achieve model applicability to most HF venues, model development was focused on standard EHR variables that are common to most HF admissions.^{5,12–24} To further improve model relevance to clinical decision-making, outcome class definitions were expanded beyond those used in prior predictive modeling where mortality and/or hospital readmission predominate.^{5,12,13,15,16,18,20–24} Specifically, referral for HF surgical intervention,²⁵ including mechanical ventricular support and cardiac transplantation, was added to all-cause mortality in defining the positive class (medical therapy nonresponders). This expanded positive class definition more appropriately targets the power of deep learning toward a specific clinical need—the timely and accurate identification of medical therapy nonresponders for consideration of high-risk HF surgical options.

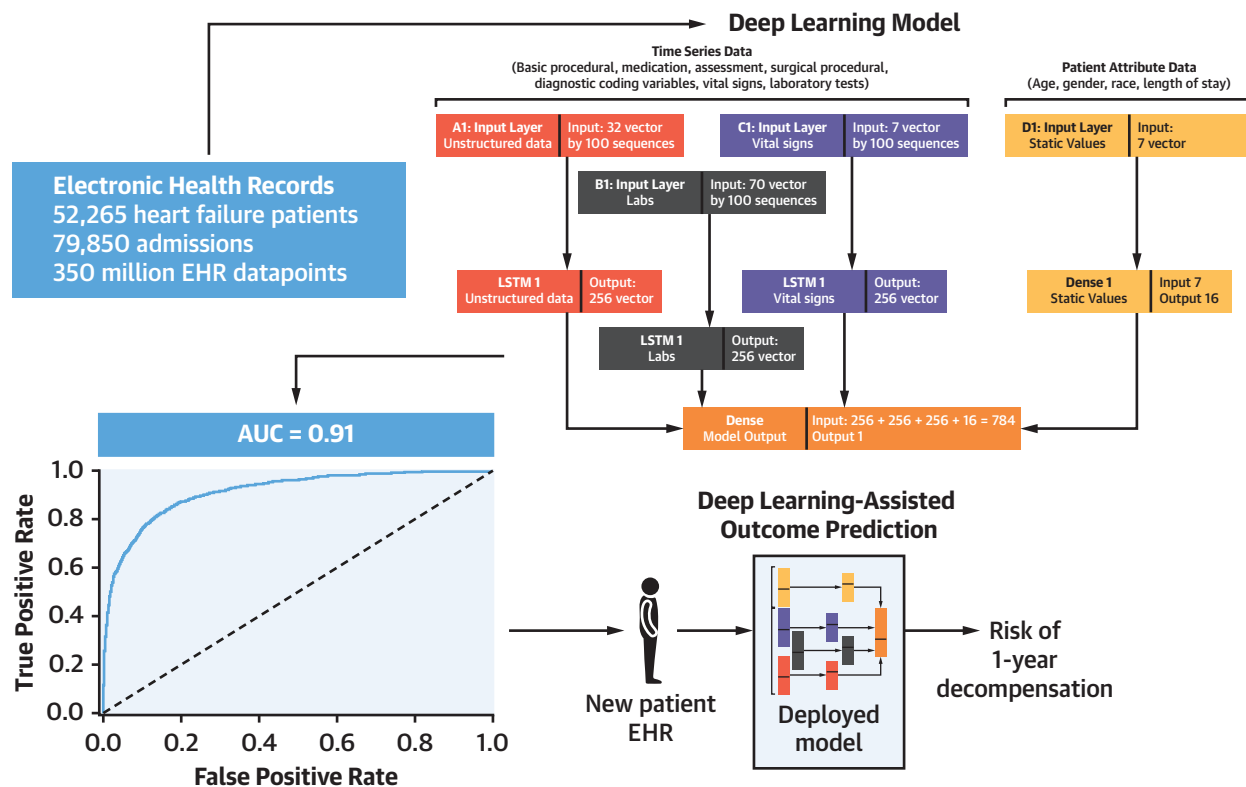
METHODS

DATA SOURCE. This study was approved by the Washington University School of Medicine Human Studies Institutional Review Board. All clinical variables were derived from individual patient EHR records at a single institution: Barnes-Jewish Hospital at Washington University Medical Center in St. Louis, Missouri.

Our deep learning model's positive class definition was expanded beyond isolated all-cause mortality within 1 year to include observation admissions for patients who reached institutional clinical thresholds for referral for invasive surgical intervention, including mechanical ventricular support (ECMO/VAD) and cardiac transplantation.²⁵

COHORT AND STUDY DESIGN. All adult patients admitted with an HF International Classification of Diseases diagnostic code from January 2009 through December 2018 comprised our initial study population. From these 52,806 patients, hospital admissions from 52,265 patients who fit age criteria (age ≥18 and <90 years) were examined. Only those admissions with at least 1 vital sign reading; 1 laboratory value; and 5 medication, assessment, procedural, or diagnostic coding data records were included as study observations. All admissions meeting these observation criteria were included, allowing patients to contribute more than 1 observation over the natural history of their clinical course. A total of 79,850 admissions met observation criteria and contributed EHR data for model development.

CENTRAL ILLUSTRATION Electronic Health Record-Based Deep Learning Model Predicts Heart Failure Decompensation



McGilvray MMO., et al. J Am Coll Cardiol HF. 2022;10(9):637–647.

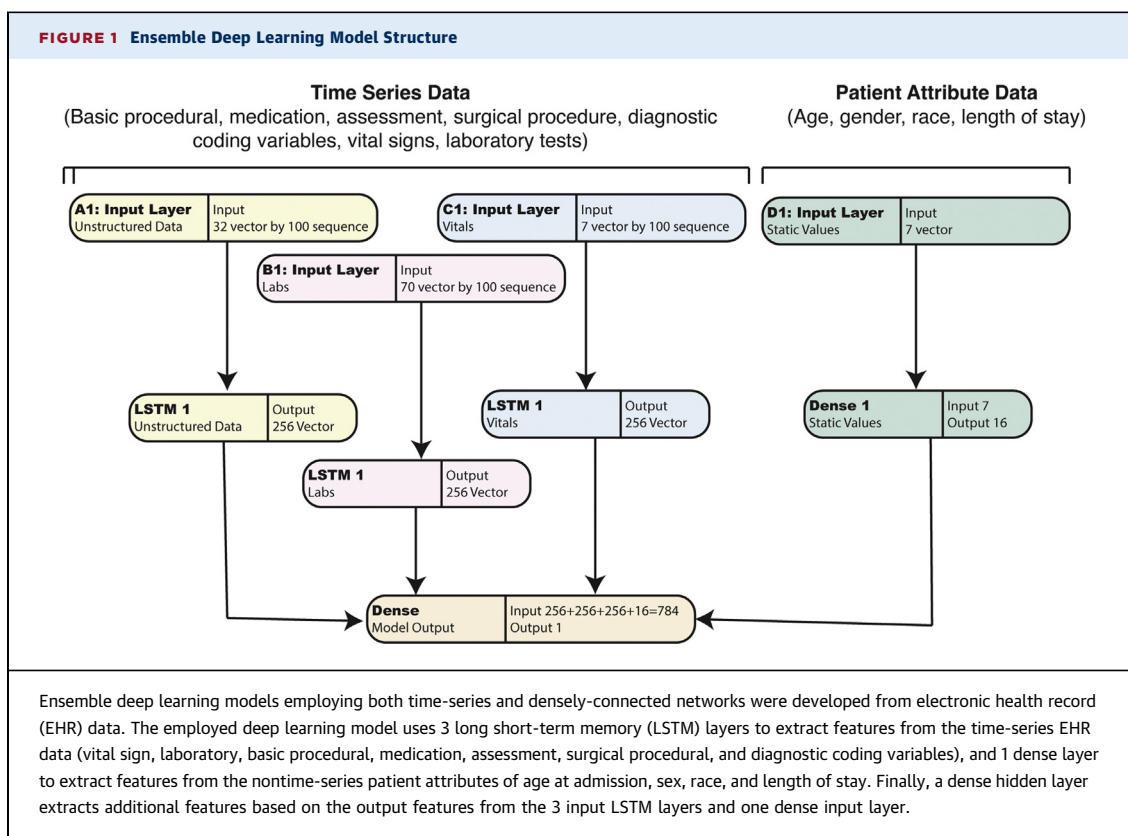
A total of 350 million electronic health record (EHR) datapoints were used to develop a deep learning model predicting 1-year death or severe decompensation in heart failure patients. The model's receiver-operator characteristic curve had an area under the curve (AUC) of 0.91, supporting clinical relevance. LSTM = long short-term memory.

Each observation's admission date triggered an observation-specific 1-year clock against which all-cause mortality and HF surgical referral were referenced for determination of positive and negative classes. An observation was classified as positive if the corresponding patient had a recorded date within 1 year of the following: 1) all-cause death; 2) ECMO/VAD; or 3) listing for cardiac transplantation. All other HF patient admission observations were classified as negative.

FEATURE EXTRACTION AND DEEP LEARNING NETWORK DESIGN. The employed deep learning model (Figure 1) uses 3 long short-term memory (LSTM) layers to extract features from the time-series EHR data (vital sign, laboratory, basic procedural, medication, assessment, surgical procedural, and diagnostic coding variables),²⁶ and 1 dense layer to extract features from the nontime-series patient attributes of age at admission, sex, race, and length of

stay (LOS). Finally, a dense hidden layer extracts additional features based on the output features from the 3 input LSTM layers and 1 dense input layer. The 3 LSTM layers learn to extract 3 256-length vectors as intermediate features from the input, while the single dense input layer takes age, sex, race, and LOS hours and learns to extract a 16-length vector as an intermediate feature. The hidden layers containing these 4 vectors flow to a hidden layer consisting of 16 neurons plus a bias neuron. This final dense hidden layer extracts features from all the preceding inputs and presents a 16-length vector to the final output neuron that performs the actual classification. All dense hidden layers make use of a rectified linear unit activation function. The final classification output neuron utilizes a sigmoid activation function.

Demographic values exist outside of the time-series flow of a medical record. Because absent demographic values usually result from collection



errors, their impact upon the model is minimized by treatment as missing data. If age was missing, then a default age of 61 years (the data set mean age) was substituted. Gender was encoded to 1 for male, –1 for female, and 0 for missing or unknown. Race was reduced into 4 dummy variables: White, Black, Asian, and other. If race was unknown, then all 4 race dummy variables were set to zero. Of note, any records in the training set that contained missing demographic values were completely removed.

The potential predictive impact of patient-specific clinical information accumulating before observation admission dates was accounted for by model inclusion of all available time-series EHR data (excludes demographics) leading up to each admission, along with the EHR data accumulated during the actual observation hospital stay. The model's 3 LSTM layers each accept sequences of up to a maximum of 100 time slices. The time slices do not conform to a fixed timespan, instead representing the last 100 values. If <100 days of values were available, data set mean values were used to complete the data array. This encoding therefore results in a 100×7 tensor for vital signs; 100×70 for laboratory test values; and 100×32 for the basic procedural, medication, assessment, surgical

procedural, and diagnostic coding variables. Although outpatient EHR data is clearly included in the 100 time slices of data accrued before hospital admission, outpatient EHR data cannot enter the model as an observation without association with hospital admission. The model has not been specifically developed to make predictions on isolated outpatient EHR data.

Laboratory values (Table 1), which are encoded as a vector of 70 values, initially start with the mean data set value for each laboratory test. As laboratory test values are obtained over time, the corresponding test value is updated and holds this value in the tensor for that date and subsequent days until a new laboratory value occurs again in the future. Thus, abnormal laboratory test results remain indefinitely until a different value occurs in the patient's EHR. Laboratory tests are of high cardinality, such that most patients have only a subset of the 70 test values we identified. Similarly, vital signs (Table 1) are encoded to a vector of 7 values, and missing data values are also substituted With the data set mean value of that particular vital sign. Each column in the sequence holds either the most recent vital sign value or the data set mean value if none are available. of note, time slices in the vital sign sequences correspond to

TABLE 1 Feature Importance for Vital Signs, Laboratory Values, and Static Variables

	AUC Without Variable	Importance
Vital signs		
Systolic blood pressure	0.894748	1.000000
Weight	0.901138	0.939287
Diastolic blood pressure	0.901781	0.933175
Pulse	0.903363	0.918146
Height	0.907861	0.875409
O ₂ saturation	0.909794	0.857045
O ₂ flow rate	0.909884	0.85619
Laboratory values		
Brain natriuretic peptide	0.90653234	1.000000
Ammonia	0.910818	0.95414821
Cholesterol	0.91148733	0.94698705
Aspartate aminotransferase	0.91176035	0.9440661
Fibrinogen	0.91214804	0.93991826
pH	0.91256713	0.93543438
D-dimer	0.9126234	0.93483235
Oxygen gradient	0.91269069	0.93411247
Benzodiazepines	0.91270801	0.93392719
Alanine aminotransferase	0.91276195	0.93335011
Triglycerides	0.91302891	0.93049393
C-reactive-protein	0.91312547	0.92946082
Cocaine	0.913161	0.9290807
Oxygen saturation	0.91318655	0.92880729
Platelets	0.91320159	0.92864644
Partial thromboplastin time	0.91321558	0.92849672
Prothrombin time	0.91327815	0.92782727
Lipase	0.91329225	0.9276765
Pulmonary artery oxyhemoglobin	0.9133117	0.92746839
Chloride	0.91339829	0.92654199
Lymphocytes	0.91347624	0.92570797
Cannabinoids	0.91351971	0.92524291
Fasting glucose	0.91352675	0.92516752
Creatine kinase	0.91354085	0.92501675
Calculated oxygen saturation	0.91354099	0.92501516
Bicarbonate	0.91357791	0.92462018
Hematocrit	0.9135801	0.92459682
Glucose	0.91359945	0.92438977
Sodium	0.91362416	0.92412539
FiO ₂	0.9136578	0.92376544
High density lipoprotein	0.91371526	0.92315068
Lactate dehydrogenase	0.91377863	0.92247273
Partial pressure of oxygen	0.91378528	0.92240159
Monocytes	0.91380344	0.92220729
Partial pressure of carbon dioxide	0.91382835	0.92194078
Alkaline phosphatase	0.9138284	0.92194025
Hemoglobin A _{1c}	0.91387152	0.92147891
Calcium	0.91387782	0.92141148
Hemoglobin	0.9139326	0.92082538
Uric acid	0.91393786	0.92076911
Plasma phosphorus	0.91399572	0.92015009
Pulmonary artery oxygen saturation	0.91400545	0.92004604
Potassium	0.91402986	0.91978484
Carbon dioxide	0.91410067	0.91902726
Hemoglobin total	0.91410335	0.9189986
Creatinine	0.91411476	0.91887649

Continued in the next column

TABLE 1 Continued

	AUC Without Variable	Importance
Amylase	0.91413163	0.91869599
Anion gap	0.91413302	0.91868112
Mixed venous oxygen	0.91413615	0.91864768
Plasma protein	0.91413635	0.91864555
Carboxyhemoglobin	0.91416299	0.91836047
Ketones	0.91416324	0.91835781
Eosinophils	0.91416662	0.91832171
Mixed venous oxygen saturation	0.91416716	0.91831587
Plasma potassium	0.91417297	0.91825376
Indirect Coombs	0.91417396	0.91824314
Base excess	0.91418577	0.91811679
Thyroid stimulating hormone	0.91419118	0.91805892
Bilirubin	0.9142014	0.91794956
Troponin I	0.91420735	0.91788585
Basophils	0.9142206	0.91774411
Magnesium	0.91422591	0.9176873
Lactic acid	0.91423033	0.91764005
Methemoglobin	0.91423405	0.91760023
International normalized ratio	0.91423465	0.91759386
Urine specific gravity	0.91424144	0.91752113
Direct bilirubin	0.91426635	0.91725463
Albumin	0.91427449	0.91716756
Neutrophils	0.91428045	0.91710385
Blood urea nitrogen	0.91432183	0.91666109
Static variables		
Length of stay hours	0.90074014	1.000000
Age at admission	0.90916532	0.91511991
Vital sign row count	0.90925047	0.91426206
Unstructured row count	0.91011626	0.90553964
Laboratory test row count	0.91060244	0.90064153
Gender	0.91381034	0.8683234
Race	0.91410881	0.86531645

AUC = area under the curve; FiO₂ = fraction of inspired oxygen.

the mean vital sign values for each of the last 100 days.

All basic procedural, medication, assessment, surgical procedural, and diagnostic coding variables were mapped to 100-length sequences of a 32-dimensional vector space generated by Python Genism Word2Vec before input into the LSTM layer.²⁷ These values are high-cardinality and sparse. Because their data are encoded to 32-number vectors that occur over time, there is no need to handle missing data. A particular value is encoded when present. Values not present are not considered, because they are not encoded into the time series.

The admission observations were randomly split into training (80%), validation (10%), and testing (10%) data sets. The primary metric of model performance is area under the receiver-operator

TABLE 2 Model Performance Results

ROC AUC	0.91
Log loss	0.35
PR AUC	0.78
Brier score	0.09
PR = precision recall; ROC = receiver-operating characteristic; other abbreviation as in Table 1.	

characteristic curve (AUC) (or C-statistic). Other performance metrics include log loss, precision recall (PR) AUC, and Brier score (Table 2). The permutation feature importance method²⁸ was used for all feature importance calculations (Table 1). This method evaluates feature importance on an already trained neural network by permuting select inputs and observing the resulting change in AUC. Permuting multiple features provides their collaborative importance.

RESULTS

The patient demographic characteristics of all study observations are shown in Table 3. According to our classification definitions, 15,833 (20%) and 64,017 (80%) of the 79,850 observations were labeled positive and negative, respectively. The average observation patient age was 62 years, and the majority of observation patients were White and men. As expected, a strong majority (97%) of the positive class observations (n = 15,293) in our study cohort were all-cause mortality, whereas 3% (n = 540) were classified as positive secondary to HF medical therapy nonresponse that met institutional guidelines for placement of mechanical ventricular support and/or cardiac transplantation. Descriptive statistics per positive and negative observation are shown in Table 4, including mean values of LOS hours, vital sign data count, laboratory value count, and unstructured data count.

Table 1 demonstrates the univariate feature importance of vital sign, laboratory test, and static variables. Systolic blood pressure (BP), BNP, and observation LOS were the most important features at a univariate level.

Multivariate exploration of feature importance can uncover complex EHR-data relationships with potential to not only predict response to medical therapy, but also generate mechanistic hypotheses. Figure 2 is a heat map demonstrating the importance interaction between vital sign and laboratory test variables. On the x- and y-axes, respectively, are the vital sign and laboratory test features ordered by univariate importance. A dominant impact upon

TABLE 3 Model Observation Characteristics

	Negative	Positive	Total
Number of observations	64,017 (80.2)	15,833 (19.8)	79,850
Age at observation, y	61 ± 15	65 ± 14	62 ± 15
Gender of observation patient			
Male	33,981 (53.1)	9,168 (57.9)	43,149 (54.0)
Female	30,034 (46.9)	6,661 (42.1)	36,695 (46.0)
Unknown	2 (0)	4 (0)	6 (0)
Race of observation patient			
White	25,233 (39.4)	7,139 (45.1)	32,372 (40.5)
Black	25,936 (56.1)	6,752 (42.6)	42,688 (53.5)
Asian	313 (0.4)	78 (0.5)	391 (0.5)
Other	2,535 (4.1)	1,864 (11.8)	4,399 (5.5)
Values are n (%) or mean ± SD.			

model prediction is demonstrated for systolic BP and BNP. The rapid increase in the heat map horizontal gradient demonstrates the marked impact of systolic BP, weight, pulse, and diastolic BP (in descending order). The less well-defined vertical gradient demonstrates the individual strength of the 70 laboratory test values whose acquisition is common in HF patient management.

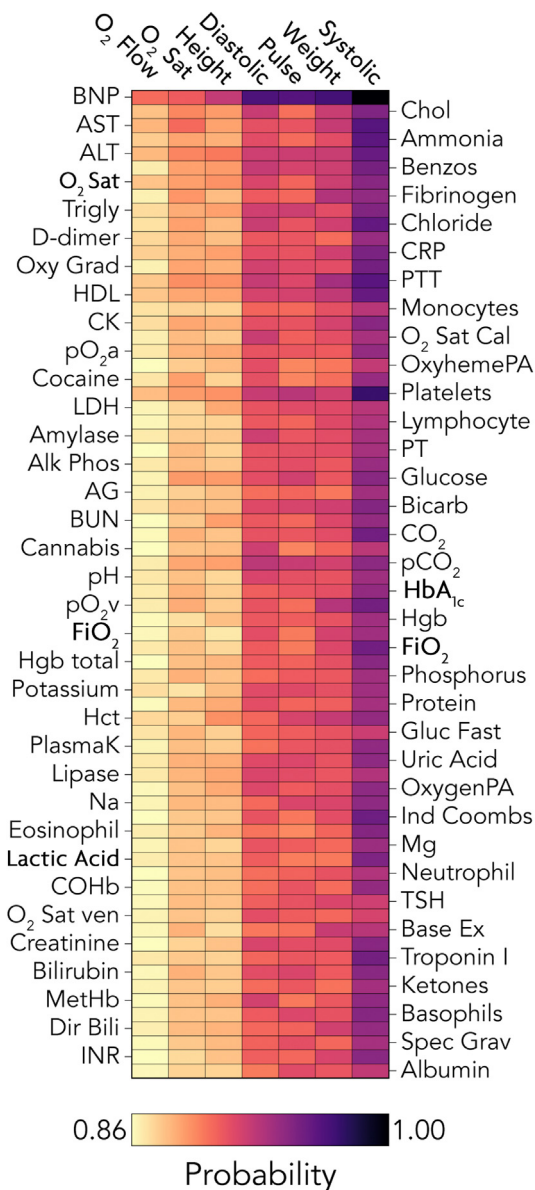
MODEL PERFORMANCE EVALUATION. Figure 3 shows the model performance by receiver-operator characteristic curve. The AUC (C-statistic) for the deep learning model test set was 0.91 with 95% CI: 0.91–0.92. The AUC for our training set was 0.98. We utilized a validation set to determine when to stop training. This validation set had an AUC of 0.91.

Additional performance evaluation metrics are demonstrated in Table 2. The PR AUC was 0.77 with a 95% CI: 0.75–0.80. The PR AUC referenced in Table 2 is shown in Figure 4.

TABLE 4 Descriptive Statistics per Observation

	Negative	Positive	Intervention	Overall
Mean LOS hours	97	174	302	112
Mean vital sign data count	30	41	25	32
Mean laboratory data count	43	53	33	45
Mean unstructured data count	58	60	49	59
Age	61	65	51	62
O ₂ SAT	97	97	97	97
Height	169	170	173	169
Weight	90	84	92	89
Pulse	79	83	81	79
Systolic	128	121	106	127
Diastolic	72	68	68	71
O ₂ flow	0.72	1.29	0.28	0.83
LOS = length of stay; SAT = saturation.				

FIGURE 2 Heat Map of Vital Sign Versus Laboratory Test Values



This heat map demonstrates the *importance* interaction between vital sign and laboratory test variables, with vital signs on the x-axis and laboratory values on the y-axis. Variables are presented in order of univariate importance. Systolic blood pressure and brain natriuretic peptide had dominant effects on model prediction. There is a steep increase in the horizontal gradient, demonstrating the particular importance of systolic blood pressure, weight, pulse, and diastolic blood pressure (in descending order). AG = anion gap; Alk Phos = alkaline phosphatase; ALT = alanine aminotransferase; AST = aspartate aminotransferase; Base Ex = base excess; Benzos = benzodiazepines; Bicarb = bicarbonate; BNP = brain natriuretic peptide; BUN = blood urea nitrogen; Chol = cholesterol; CK = creatine kinase; CO₂ = carbon dioxide; COHb = carboxyhemoglobin; CRP = c-reactive protein; Diastolic = diastolic blood pressure; Dir Bili = direct bilirubin; FiO₂ = fraction of inspired oxygen; Gluc Fast = fasting glucose; HbA_{1c} = glycosylated hemoglobin; Hct = hematocrit; HDL = high-density lipoprotein; Hgb = hemoglobin; Ind Coombs = indirect Coombs; INR = international normalized ratio; LDH = lactate dehydrogenase; MethHb = methemoglobin; Mg = magnesium; PlasmaK = plasma potassium; Na = sodium; O₂ Flow = oxygen flow; O₂ Sat = oxygen saturation; O₂ Sat Cal = calculated oxygen saturation; O₂ Sat ven = mixed venous oxygen saturation; OxygenPA = pulmonary artery oxygen saturation; Oxy grad = oxygen gradient; OxyhemePA = pulmonary artery oxyhemoglobin; pCO₂ = carbon dioxide level; pO_{2a} = arterial oxygen partial pressure; pO_{2v} = mixed venous oxygen partial pressure; PT = prothrombin time; PTT = partial thromboplastin time; Spec Grav = urine specific gravity; Systolic = systolic blood pressure; Trigly = triglycerides; TSH = thyroid stimulating hormone.

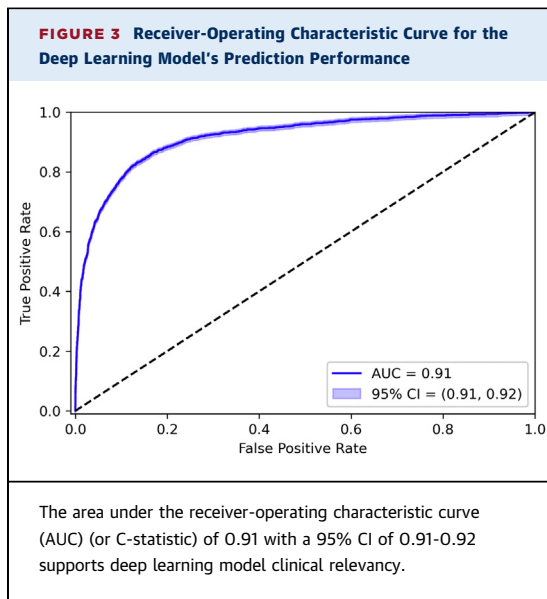


Table 4 demonstrates mean descriptive statistics referenced to positive and negative observations, and **Table 5** summarizes the model's AUC for a broad range of hospital admission LOS (hours). The model predicted well across the spectrum of observation admission LOS, even for those observations based upon very short hospital admissions.

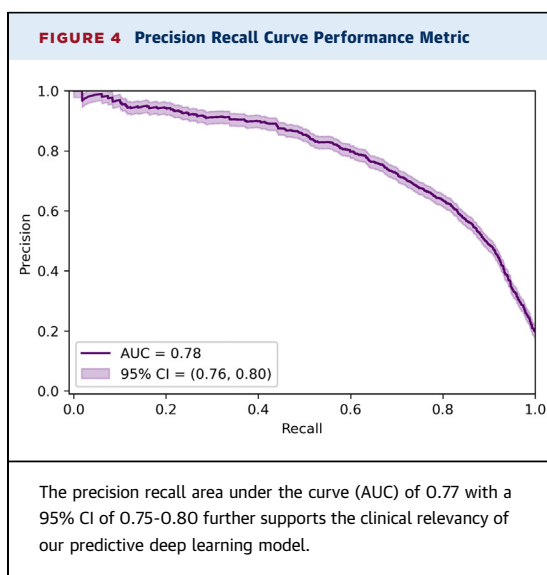
DISCUSSION

Deep neural networks have proven successful at extracting features from EHR data in HF patient populations.^{5,12-24} In this study, EHR-based deep learning models demonstrated potential to assist

clinicians in more accurately directing the application of highly invasive, life-disruptive, yet potentially life-saving HF surgical therapy. To improve generalization, the model was trained, validated, and tested using >350 million EHR datapoints accumulated over a 10-year period from 79,850 patient admission observations in a final study cohort of 52,265 unique HF patients. An ensemble deep learning model using both LSTM time-series and densely connected algorithms accurately predicted a positive observation class whose patients experienced severe deterioration within 1 year with a C-statistic of 0.91. These results support the hypothesis that strong predictive relationships exist in EHR data routinely acquired during HF patient management, and that this predictive power is potentially exploitable by machine learning algorithms to predict real-world nonresponse to HF medical therapy.

Recent publications support the potential contribution of machine learning in the enhancement of clinical decision-making when compatible algorithms are judiciously matched to appropriate clinical problems.^{5,12-14,16,17,19-26,29,30} For instance, when deep convolutional neural networks, which are optimal for image analysis, are enhanced by a combination of pretraining, data-preprocessing, model ensembling, and radiologist-augmented workflow, previously unimaginable results (AUC: 0.99) have been obtained in the automated classification of tuberculosis by chest radiography.²⁹ Most machine learning applications in HF patient populations have focused upon commonly available EHR variables and have been aimed at improving prognostic modeling to assist clinicians in the accurate identification of patients nearing decompensation, with primary outcome definitions centered upon mortality and/or hospital readmission.^{5,12,13,15-17,20-24,31}

The current investigation builds upon this progress by focusing the predictive power of deep learning on a single and yet critical patient management issue: the timing and accuracy of HF patient referral for high-risk mechanical ventricular support or cardiac replacement therapy. Accurate timing of patient referral for advanced HF intervention represents a prototypical machine learning binary classification problem to which deep learning algorithms are especially applicable. The relevance of prior machine learning investigations that used only mortality or hospital readmission to define the positive class is difficult to assess. Use of a mortality-based positive class definition that does not include patients referred for advanced intervention builds a small inherent error into any



potential clinical application in our target population. Although 97% of the positive class in our study cohort experienced all-cause mortality, the remaining 3% (540 patient observations) would be classified in the negative class if all-cause mortality was the sole classification descriptor.

Our use of machine learning algorithms sensitive to the predictive potential of temporal patterns of time-sensitive EHR variables is critical to the demonstrated model accuracy. Detection of subtle trends in disease progression or remission, as indicated by progressive time-based deterioration or improvement of EHR-based HF metrics, clearly boosted model predictive capability, just as it has intuitively contributed to routine clinical HF management for decades. In our model development, time-series^{22,26,30} machine learning algorithms (ie, LSTM networks) were used in the analysis of all variables, except nontime-dependent, fixed variables (ie, demographics).

Many electronic medical record systems, such as EPIC and Cerner, allow application programming interfaces in HL7 FIHR (Fast Healthcare Interoperability Resources) format to automatically populate embedded machine learning models with patient-specific EHR data. With further model development, these capabilities may allow advanced machine learning models, such as those discussed in this paper, to assist clinicians in real-time patient management decisions by rendering patient-specific heart failure outcome predictions.

STUDY LIMITATIONS. Potential limitations of this investigation include its use of EHR data retrospectively gathered from a single institution, possibly frustrating our emphasis upon generalization to other HF populations and venues. Further, all investigations employing machine learning models based upon retrospectively obtained sequentially occurring EHR database variables may have inadvertent retrograde leakage into the model of data informative of subsequent occurrence of HF classification endpoints. This information may unfairly inform the model and falsely improve model test set accuracy. We have manually and analytically searched the EHR data for all structured and unstructured mention of endpoints and have been assured by these efforts that no systematic leakage occurred.

To improve model generalization to real-world application, all admissions that met our observation criteria were included, allowing many patients to contribute more than 1 observation. Although this may have skewed model training toward the

TABLE 5 ROC AUC on Test Dataset for LOS Hours

LOS, h	Observations	AUC	Positive, %
≤5	949	0.93	12
≤10	2,359	0.93	12
≤20	3,573	0.94	12
≤30	4,467	0.94	14
≤40	4,938	0.94	14
≤50	5,234	0.93	14
≤60	5,417	0.93	14
≤70	5,702	0.93	15
≤80	5,884	0.93	15
≤90	6,020	0.93	15
≤100	6,237	0.93	16
≤200	7,181	0.92	17
≤300	7,498	0.92	18
≤400	7,658	0.92	19
≤500	7,749	0.92	19
≤600	7,806	0.92	19

AUC = area under the curve; LOS = length of stay; ROC = receiver-operating characteristic.

sickest patients, thereby compromising observation independence, the model's predictive potential is of primary importance. All observations were included to capture the important predictive power of temporal EHR patterns during disease progression.

The accurate determination of death in any medical population remains problematic. As a result of the considerable efforts of our hospital billing services, our hospital EHR itself remains the most reliable source of mortality statistics for our study population. Similarly, our use of “all-cause” mortality in the definition of our positive class, although reasonable, is not ideal, because it is inclusive of non-HF-dependent deaths. Our target HF patient population is comprised of medically complex patients with equally complicated, multifactorial, and often obscure causes contributing to their deaths. Sorting them out with any degree of precision is clinically impossible, making all-cause mortality the optimal, albeit compromised, choice.

CONCLUSIONS

The identification of HF patients approaching decompensation would facilitate the accurate targeting of these patients with available surgical options, while sparing the rest from the associated risk, expense, and uncertainty of these life-disrupting yet potentially life-saving technologies. An EHR-based deep learning model that predicts

death or referral for HF surgical intervention was developed in a large single-institution cohort of HF patients with accuracy supporting clinical relevance. Future ensemble models employing not only EHR-based variables, but also advanced patient-specific HF physiological metrics, such as functional testing^{25,30} or regional contractile indexes,^{32,33} offer considerable promise in achieving predictive accuracy that engenders uncontested clinical relevance. As further clinical application increases model accuracy, automated EHR-embedded deep learning models may have potential to directly assist HF clinicians in improving the application of advanced HF surgical therapy.

FUNDING SUPPORT AND AUTHOR DISCLOSURES

The authors have reported that they have no relationships relevant to the contents of this paper to disclose.

ADDRESS FOR CORRESPONDENCE: Dr Michael K. Pasque, Washington University School of Medicine, Campus Box 8234, 660 South Euclid Avenue, St. Louis, Missouri 63110, USA. E-mail: pasquem@wustl.edu.

PERSPECTIVES

COMPETENCY IN MEDICAL KNOWLEDGE:

The demonstrated accuracy of EHR-based deep learning model prediction of 1-year all-cause death or referral for HF surgical therapy supports clinical relevance. EHR-based deep learning models have considerable potential to assist HF clinicians in improving the application of advanced HF surgical therapy in medical therapy nonresponders.

TRANSLATIONAL OUTLOOK: A deep learning model developed from 350 million EHR datapoints from 52,265 heart failure patients predicted 1-year death or severe decompensation with a C-statistic of 0.91. Deployment of this model is aimed at assisting clinicians in the early identification of at-risk heart failure patients and timing of individual referral for advanced surgical intervention, such as mechanical ventricular assistance and cardiac transplantation.

REFERENCES

- Virani SS, Alonso A, Aparicio HJ, Benjamin EJ. Heart disease and stroke statistics—2021 update: a report from the American Heart Association. *Circulation*. 2021;143:e254–e743. <https://www.ahajournals.org/doi/abs/10.1161/CIR.0000000000000950>
- Chen J, Normand S-LT, Wang Y, Krumholz HM. National and regional trends in heart failure hospitalization and mortality rates for Medicare beneficiaries, 1998–2008. *JAMA*. 2011;306:1669–1678.
- Loehr LR, Rosamond WD, Chang PP, Folsom AR, Chambless LE. Heart failure incidence and survival (from the Atherosclerosis Risk in Communities study). *Am J Cardiol*. 2008;101:1016–1022.
- Lindenfeld J, Feldman AM, Saxon L, Boehmer J, Carson P, Ghali JK, et al. Effects of cardiac resynchronization therapy with or without a defibrillator on survival and hospitalizations in patients with New York Heart Association class IV heart failure. *Circulation*. 2007;115:204–212.
- Ahmad T, Lund LH, Rao P, Ghosh R, Warier P, Vaccaro B, et al. Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc*. 2018;7:e008081.
- Gayat E, Arrigo M, Littnerova S, et al. Heart failure oral therapies at discharge are associated with better outcome in acute heart failure: a propensity-score matched study. *Eur J Heart Fail*. 2018;20:345–354.
- McDonagh TA, Metra M, Adamo M. Guidelines for the diagnosis and treatment of acute and chronic heart failure. *Eur Heart J*. 2021;42(36):3599–3726. <https://doi.org/10.1093/eurheartj/ehab368>
- van der Meer P, Gaggin HK, Dec GW. ACC/AHA versus ESC guidelines on heart failure: JACC guideline comparison. *J Am Coll Cardiol*. 2019;73:2756–2768.
- Rossignol P, Hernandez AF, Solomon SD, Zannad F. Heart failure drug treatment. *Lancet*. 2019;393:1034–1044.
- Gustafsson F, Rogers JG. Left ventricular assist device therapy in advanced heart failure: patient selection and outcomes. *Eur J Heart Fail*. 2017;19:595–602.
- Csepe TA, Kilic A. Advancements in mechanical circulatory support for patients in acute and chronic heart failure. *J Thorac Dis*. 2017;9:4070–4083.
- 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.
- Angraal S, Mortazavi BJ, Gupta A, et al. Machine learning prediction of mortality and hospitalization in heart failure with preserved ejection fraction. *J Am Coll Cardiol HF*. 2020;8:12–21.
- Samuel OW, Asogbon GM, Sangaiah AK, Fang P, Li G. An integrated decision support system based on ANN and Fuzzy_AHP for heart failure risk prediction. *Expert Syst Appl*. 2017;68:163–172.
- Eapen ZJ, Liang L, Fonarow GC, et al. Validated, electronic health record deployable prediction models for assessing patient risk of 30-day rehospitalization and mortality in older heart failure patients. *J Am Coll Cardiol HF*. 2013;1:245–251.
- Ouwkerk W, Voors AA, Zwinderman AH. Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure. *J Am Coll Cardiol HF*. 2014;2:429–436.
- Levy D, Kenchaiah S, Larson MG, et al. Long-term trends in the incidence of and survival with heart failure. *N Engl J Med*. 2002;347:1397–1402.
- Pocock SJ, Ariti CA, McMurray JJV, et al. Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies. *Eur Heart J*. 2013;34:1404–1413.
- Al'Aref SJ, Anouchke K, Singh G, et al. Clinical applications of machine learning in cardiovascular disease and its relevance to cardiac imaging. *Eur Heart J*. 2019;40:1975–1986.
- Mortazavi BJ, Downing NS, Bucholz EM, et al. Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes*. 2016;9:629–640.
- Taslimitehrani V, Dong G, Pereira NL, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. *J Biomed Inform*. 2016;60:260–269.
- Jiang W, Siddiqui S, Barnes S, et al. Readmission risk trajectories for patients with heart failure using a dynamic prediction approach:

retrospective study. *JMIR Med Inform*. 2019;7:e14756.

23. Frizzell JD, Liang L, Schulte PJ, et al. Prediction of 30-day all-cause readmissions in patients hospitalized for heart failure: comparison of machine learning and other statistical approaches. *JAMA Cardiol*. 2017;2:204–209.

24. Blecker S, Katz SD, Horwitz LJ, et al. Comparison of Approaches for Heart Failure Case Identification From Electronic Health Record Data. *JAMA Cardiol*. 2016;1:1014–1020.

25. Hearn J, Ross HJ, Mueller B, et al. Neural networks for prognostication of patients with heart failure. *Circ Heart Fail*. 2018;11:e005193.

26. Mpanya D, Celik T, Klug E, Ntsinjana H. Machine learning and statistical methods for predicting mortality in heart failure. *Heart Fail Rev*. 2021;26:545–552.

27. Choi E, Xiao C, Stewart WF, Sun J. MiME: multilevel medical embedding of electronic health records for predictive healthcare. *arXiv [cs.LG]*. 2018. Accessed June 1, 2022. <http://arxiv.org/abs/1810.09593>

28. Breiman L. Random Forests. *Mach Learn*. 2001;45:5–32.

29. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology*. 2017;284:574–582.

30. Tabassian M, Sunderji I, Erdei T, et al. Diagnosis of heart failure with preserved ejection fraction: machine learning of spatio-temporal variations in left ventricular deformation. *J Am Soc Echocardiogr*. 2018;31:1272–1284. e9.

31. Kwon J-M, Kim K-H, Jeon K-H, et al. Artificial intelligence algorithm for predicting mortality of patients with acute heart failure. *PLoS One*. 2019;14:e0219302.

32. MacGregor RM, Guo A, Masood MF, et al. Machine learning outcome prediction in dilated cardiomyopathy using regional left ventricular multiparametric strain. *Ann Biomed Eng*. 2021;49:922–932.

33. Henn MC, Lawrance CP, Kar J, et al. Dilated cardiomyopathy: normalized multiparametric myocardial strain predicts contractile recovery. *Ann Thorac Surg*. 2015;100:1284–1291.

KEY WORDS heart failure (HF), machine learning, mechanical circulatory support, ventricular assist device (VAD)