Title

Artificial intelligence approaches for phenotyping heart failure in U.S. Veterans Health Administration electronic health record.

Permalink

https://escholarship.org/uc/item/15t4d84w

Journal

ESC Heart Failure, 11(5)

Authors

Shao, Yijun

Zhang, Sijian

Raman, Venkatesh

et al.

Publication Date

2024-10-01

DOI

10.1002/ehf2.14787

Peer reviewed

ESC HEART FAILURE
ESC Heart Failure 2024; **11**: 3155–3166
Published online 14 June 2024 in Wiley Online Library (wileyonlinelibrary.com) **DOI:** 10.1002/ehf2.14787

ORIGINAL ARTICLE

Artificial intelligence approaches for phenotyping heart failure in U.S. Veterans Health Administration electronic health record

Yijun Shao[1,2]*, Sijian Zhang[1,2], Venkatesh K. Raman[1,3], Samir S. Patel[1,2], Yan Cheng[1,2], Anshul Parulkar[4,5], Phillip H. Lam[1,3,6], Hans Moore[1,2,3,7], Helen M. Sheriff[1,2], Gregg C. Fonarow[8], Paul A. Heidenreich[9,10], Wen-Chih Wu[4,5], Ali Ahmed[1,2,3]* and Qing Zeng-Treitler[1,2]*

[1]Center for Data Science and Outcomes Research, Veterans Affairs Medical Center, Washington, DC, USA; [2]George Washington University, Washington, DC, USA; [3]Georgetown University, Washington, DC, USA; [4]Veterans Affairs Medical Center, Providence, RI, USA; [5]Brown University, Providence, RI, USA; [6]MedStar Washington Hospital Center, Washington, DC, USA; [7]Uniformed Services University, Bethesda, MD, USA; [8]University of California, Los Angeles, CA, USA; [9]Veterans Affairs Palo Alto Health Care System, Palo Alto, CA, USA; and [10]Stanford University School of Medicine, Stanford, CA, USA

Abstract

Aims Heart failure (HF) is a clinical syndrome with no definitive diagnostic tests. HF registries are often based on manual reviews of medical records of hospitalized HF patients identified using International Classification of Diseases (ICD) codes. However, most HF patients are not hospitalized, and manual review of big electronic health record (EHR) data is not practical. The US Department of Veterans Affairs (VA) has the largest integrated healthcare system in the nation, and an estimated 1.5 million patients have ICD codes for HF (HF ICD-code universe) in their VA EHR. The objective of our study was to develop artificial intelligence (AI) models to phenotype HF in these patients.

Methods and results The model development cohort (n = 20 000: training, 16 000; validation 2000; testing, 2000) included 10 000 patients with HF and 10 000 without HF who were matched by age, sex, race, inpatient/outpatient status, hospital, and encounter date (within 60 days). HF status was ascertained by manual chart reviews in VA's External Peer Review Program for HF (EPRP-HF) and non-HF status was ascertained by the absence of ICD codes for HF in VA EHR. Two clinicians annotated 1000 random snippets with HF-related keywords and labelled 436 as HF, which was then used to train and test a natural language processing (NLP) model to classify HF (positive predictive value or PPV, 0.81; sensitivity, 0.77). A machine learning (ML) model using linear support vector machine architecture was trained and tested to classify HF using EPRP-HF as cases (PPV, 0.86; sensitivity, 0.86). From the 'HF ICD-code universe', we randomly selected 200 patients (gold standard cohort) and two clinicians manually adjudicated HF (gold standard HF) in 145 of those patients by chart reviews. We calculated NLP, ML, and NLP + ML scores and used weighted F scores to derive their optimal threshold values for HF classification, which resulted in PPVs of 0.83, 0.77, and 0.85 and sensitivities of 0.86, 0.88, and 0.83, respectively. HF patients classified by the NLP + ML model were characteristically and prognostically similar to those with gold standard HF. All three models performed better than ICD code approaches: one principal hospital discharge diagnosis code for HF (PPV, 0.97; sensitivity, 0.21) or two primary outpatient encounter diagnosis codes for HF (PPV, 0.88; sensitivity, 0.54).

Conclusions These findings suggest that NLP and ML models are efficient AI tools to phenotype HF in big EHR data to create contemporary HF registries for clinical studies of effectiveness, quality improvement, and hypothesis generation.

Keywords Heart failure; Artificial intelligence; Machine learning; Natural language processing; Phenotyping; Big data; Electronic health record

Received: 11 January 2024; Revised: 23 February 2024; Accepted: 15 March 2024
**Correspondence to: Yijun Shao, Ali Ahmed, and Qing Zeng-Treitler, Center for Data Science and Outcomes Research, Veterans Affairs Medical Center, 50 Irving St. NW, Washington, DC 20422, USA. Email: yshao@gwu.edu, zengq@gwu.edu, and ali.ahmed@va.gov*
Yijun Shao, Sijian Zhang, and Venkatesh K. Raman provided equal contribution.
Wen-Chih Wu, Ali Ahmed, and Qing Zeng-Treitler provided equal contribution.

Introduction

Heart failure (HF) is a leading cause of morbidity and mortality.[1,2] Evidence from randomized controlled trials (RCTs) provide the foundation for guideline-directed medical therapy (GDMT) for patients with HF that contributes to improving clinical outcomes.[3] Post hoc analyses of data from RCTs provide important insights into risk stratification, treatment optimization, and hypothesis generation.[4–14] However, patients enrolled in RCTs are often not representative of those seen in clinical practice.[15] HF registries include real-world patients from clinical practice, but are often based on quality improvement initiatives in hospital settings with limited follow-up data.[16–32] The Organized Program to Initiate Lifesaving Treatment in Hospitalized Patients With Heart Failure (OPTIMIZE-HF) began as a quality improvement initiative that included nearly 50 000 hospitalizations, but the registry component included about 6000 patients with 60–90 days of follow up.[28,33,34] The hospitalization records of OPTIMIZE-HF was later linked to Medicare inpatient claims data using indirect identifiers to obtain long-term follow-up data of unique patients.[17,35] The Get With The Guidelines (GWTG) HF data was similarly linked to Medicare for long-term outcomes data.[29,30]

HF registries contain extensive data including admission and discharge medications, which allow the use of new-user design to minimize prevalent-user bias.[36] However, these registries are based on manual abstraction of charts of patients identified by International Classification of Diseases (ICD) codes for HF. Because charts are often abstracted at the local hospital level, they are also subject to potential bias due to inter-abstractor and inter-hospital variabilities. These registries are often limited to index hospitalizations, with limited or no access to prior medical records, and no access to any medical record after data collection is completed. The emergence of electronic heath records (EHRs) and the advances in the field of artificial intelligence (AI) have created an opportunity to automate the creation of contemporary clinical registries with longitudinal data for both quality improvement and outcomes research. However, an efficient use of these big datasets requires the development of a uniform data abstraction process using AI approaches. While several studies have used AI approaches to define HF phenotype in EHR,[37–42] most focused on natural language processing (NLP) and a few utilized machine learning (ML), in conjunction with ICD codes. The past NLP and ML approaches relied heavily on supervised learning or expert-crafted rule. We propose an approach that will combine NLP and weakly supervised ML to create a phenotype that is not dependent on ICD. The Veterans Health Administration (VHA) of the Department of Veterans Affairs (VA) is the largest integrated healthcare system in the United States, which is enriched by large genomic and phenomic databases.[43–45] The VA EHR is one of the largest integrated EHR in the world and nearly 1.5 million patients have ICD codes for HF in their EHR. Thus, an HF registry based on the VA EHR in which HF has been adjudicated based on medical record would be a valuable resource for clinical studies of effectiveness, quality improvement, and hypothesis generation. Thus, the objective of our study is to develop ML and NLP models to phenotype HF in patients with an ICD code for HF in national VHA EHR data.[46]

Methods

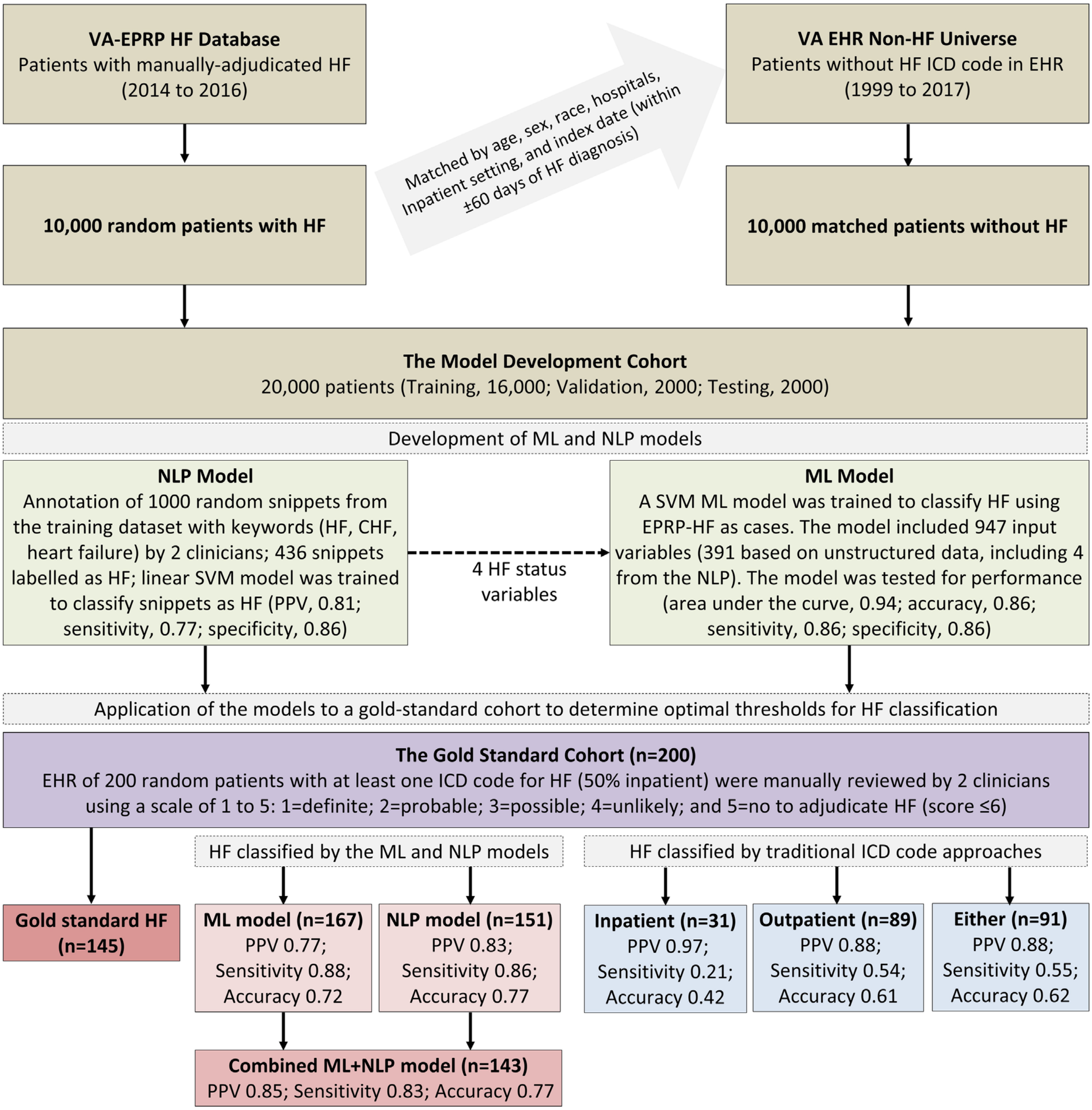
Data source and study population for model development

We used VA's national EHR data available at the VA Informatics and Computing Infrastructure platform and VA's External Peer Review Program for HF (EPRP-HF) available from EPRP. EPRP is VA's quality improvement initiative for monitoring hospital performance by manual abstraction of randomly selected medical records by external medical professional abstractors. Patients are selected for EPRP review if they have used VA healthcare system at least once in the 2 years before and had at least 1 primary care or specialty medical visit in the month being sampled for the year being evaluated.[47–50] We began by randomly selecting 10 000 patients with HF from the EPRP-HF (2014–2016) and 10 000 patients from VA EHR who had no ICD code for HF in their EHR, matching them by age, sex, race, setting (inpatient vs. outpatient), location (medical center), and within 60 days of index date (HF diagnosis date in EPRP-HF data), thus assembling a population of 20 000 patients (*Figure 1*). The cohort was divided into three subsets: training (n = 16 000), validation (n = 2000), and testing (n = 2000), each subset containing the same number of cases and controls. Baseline characteristics of these patients were examined and compared using absolute standardized difference (ASD). Unlike Chi-square or t-tests, ASD is not influenced by sample size.[51] ASD values $<$10% suggest that two study groups are comparable.

Model development

We used both NLP and ML models. A linear support vector machine (SVM) NLP model was trained to determine the HF status based on the textual data. We started by determining a set of keywords (viz. HF, CHF, and heart failure) in the notes and creating snippets (30 words before and 30 words after the keyword) to determine if the context surrounding those keywords can classify the HF status (yes, no, or uncertain)

Figure 1 Phenotyping HF in VA national EHR using ML and NLP models. After developing and testing NLP and ML models in the model development cohort of 20,000 patients, we applied the models to the gold standard cohort of 200 patients. We calculated NLP and ML scores and derived NLP + ML scores for each of the 200 patients and estimated the best threshold scores for each model using the highest F score values. Using threshold scores for each model, we classified HF in the gold standard cohort. Th NLP + ML model had the highest PPV and was chosen as the best-performing model. All 3 models performed better than the traditional ICD code approaches for identifying HF cohorts. The presence of ≥1 ICD code of HF as the principal hospital discharge diagnosis was used to defined the 'Inpatient' HF cohort and ≥2 ICD codes as primary outpatient encounter diagnoses were used to define the 'Outpatient' HF cohort ('Either' included ≥1 inpatient and ≥2 outpatient HF diagnoses). **Abbreviations**: AI, artificial intelligence; EHR, electronic health record; EPRP, External Peer Review Program; HF, heart failure; ML, machine learning; NLP, natural language processing; PPV, positive predictive value; SVM, support vector machine; VA, Veterans Affairs.


VA-EPRP HF Database
Patients with manually-adjudicated HF
(2014 to 2016)
Matched by age, sex, race, hospitals,
Inpatient setting, and index date (within
±60 days of HF diagnosis)
VA EHR Non-HF Universe
Patients without HF ICD code in EHR
(1999 to 2017)
10,000 random patients with HF
10,000 matched patients without HF
The Model Development Cohort
20,000 patients (Training, 16,000; Validation, 2000; Testing, 2000)
Development of ML and NLP models
NLP Model
Annotation of 1000 random snippets from the training dataset with keywords (HF, CHF, heart failure) by 2 clinicians; 436 snippets labelled as HF; linear SVM model was trained to classify snippets as HF (PPV, 0.81; sensitivity, 0.77; specificity, 0.86)
4 HF status variables
ML Model
A SVM ML model was trained to classify HF using EPRP-HF as cases. The model included 947 input variables (391 based on unstructured data, including 4 from the NLP). The model was tested for performance (area under the curve, 0.94; accuracy, 0.86; sensitivity, 0.86; specificity, 0.86)
Application of the models to a gold-standard cohort to determine optimal thresholds for HF classification
The Gold Standard Cohort (n=200)
EHR of 200 random patients with at least one ICD code for HF (50% inpatient) were manually reviewed by 2 clinicians using a scale of 1 to 5: 1=definite; 2=probable; 3=possible; 4=unlikely; and 5=no to adjudicate HF (score ≤6)
HF classified by the ML and NLP models
HF classified by traditional ICD code approaches
Gold standard HF (n=145)
ML model (n=167)
PPV 0.77;
Sensitivity 0.88;
Accuracy 0.72
NLP model (n=151)
PPV 0.83;
Sensitivity 0.86;
Accuracy 0.77
Inpatient (n=31)
PPV 0.97;
Sensitivity 0.21;
Accuracy 0.42
Outpatient (n=89)
PPV 0.88;
Sensitivity 0.54;
Accuracy 0.61
Either (n=91)
PPV 0.88;
Sensitivity 0.55;
Accuracy 0.62
Combined ML+NLP model (n=143)
PPV 0.85; Sensitivity 0.83; Accuracy 0.77

of the patient. Two clinician authors (A. A. and S. P.) performed the annotation of 1000 random notes from a total of 70 159 snippets derived from 135 856 notes from the 20 000 patients. For outpatient visits, we used notes from the index date and for inpatient stays, we used discharge summary notes and eight randomly selected notes between admission and discharge dates. The annotation results were used as the gold standard for the training and testing of the NLP model.

The ML was developed through supervised learning while also utilized an unsupervised learning method called topic modelling for the extraction of features from text notes. For ML, we used linear SVM which has several advantages relevant to this study, including excellent prediction performance, fast training speed on datasets of large sample sizes with large number of features, and less prone to overfitting.[52,53] Variables used in ML were obtained from both structured and unstructured data. Features from structured data included 43 manually crafted variables (viz. comorbidities, laboratory values, and vital signs) and additional 513 variables that included 130 medications, 106 procedure codes, 258 laboratory test orders, and 19 note titles (*Table S1*). The 513 variables were selected based on Chi-square (≥10) and prevalence (≥10%) in the training data.

Topic modelling was used to extract topic variables from the unstructured text data. For topic modelling, we used the method of latent Dirichlet allocation (LDA), which is an unsupervised ML method for uncovering the hidden topics within a large number of textual documents. The topic model was trained on the above-mentioned 135 856 clinical notes that generated 1694 stable topic variables. We selected 387 topic variables using the same approach described above (*Table S1*). In addition, we included the results from the NLP model as features. Details of the model development are presented in *Supporting information*. The final ML model used a total of 947 variables, which included all the variables extracted from structured and unstructured data (*Table S1*).

The linear SVM model we used (for both NLP and ML) had only one meta-parameter: C, which was set as 0.03 for NLP model and 0.0002 for ML model. The LDA model had only one meta-parameter: total number of topics, which was set as 2000. The input feature values for the ML model were all normalized to have zero mean and unit standard deviation while the input feature values for the NLP model were kept unchanged before they were fed into the model. These models were used to phenotype HF in VA's Centralized Interactive Phenomics Resource (CIPHER) and the source codes used are shared in the publicly-available GitHub platform.[54] CIPHER is a knowledgebase of computable EHR-based phenotypes, designed to optimize the use of VA's EHR data for use in research and clinical operations. The software library for developing the SVM models (NLP and ML) was the Python (version 3.7) library named Scikit-Learn (version 0.22.1), and the library for developing the LDA topic model was the Java (version 1.8) library named MALLET (version 2.0).

Assembly of the gold standard HF and ICD HF cohorts

We randomly selected 200 patients (100 inpatient) from the HF universe (at least 1 ICD code for HF) in the VA EHR (n = 1 446 053). To identify gold standard HF, two clinician authors (A. A. and V. R.) manually reviewed the charts, scoring HF status using a scale of 1 to 5: 1 = *definite*; 2 = *probable*; 3 = *possible*; 4 = *unlikely*; and 5 = *no*. Each patient could have a combined score between 2 (definite, by both reviewers) and 10 (no HF, by both reviewers). Patients were considered to have HF if they had a combined score of ≤6 by both reviewers. Patients with classified as 'no HF' (score of 5) by one reviewer were not considered as HF regardless of the score (e.g., '1 + 5'). In addition to the gold standard based on manual review described above, we used three other cohorts based on ICD codes for HF: (i) at least one principal hospital discharge diagnosis, (ii) at least two primary outpatient encounter diagnoses, and (iii) either one principal hospital discharge diagnosis or two primary outpatient encounter diagnoses.

Calculation of model scores and selection of optimal threshold score

We calculated NLP and ML scores for all 200 patients. Using the minimum and maximum values as boundaries, we created candidate threshold values for HF classification for both scores. For example, for the NLP model, the NLP scores for the 200 patients ranged between −0.30 and 2.76. We then created 307 candidate NLP threshold values starting with −0.30, −0.29, −0.28, and −0.27 and ending with 2.73, 2.74, 2.75, and 2.76 (using an arbitrary increment of 0.01). Respective ML scores for the 200 patients ranged between −0.80 and 4.27, which generated 508 candidate threshold values. Finally, we used a simple logistic regression model and the NLP and ML scores to generate combined scores, denoted as NLP + ML, for each of the 200 patients. The combined NLP + ML score had a range between −2.49 and 3.44, and 594 candidate threshold values. To identify the optimal threshold values, for each candidate threshold value of NLP, ML, and NLP + ML score, we calculated positive predictive value (PPV) and sensitivity. Using PPVs and sensitivities, we calculated F scores, which are harmonic means of the two (PPV and sensitivity). To limit inclusion of false-positives in the HF cohort to be classified by the models (higher PPV or precision) over capturing all true HF patients

Table 1 Baseline characteristics by 20 000 patients with and without heart failure used for model development

n (%) or mean (±standard deviation)	No heart failure (*n* = 10 000)	Heart failure (*n* = 10 000)	ASD (%)	*P* value
Age, years	67.5 (±11.2)	67.5 (±11.2)	0	1.00
Female	230 (2.3%)	230 (2.3%)	0	1.00
Race				
White	8080 (80.1%)	8080 (80.1%)	0	
African American	1913 (19.1%)	1913 (19.1%)	0	1.00
Others	79 (0.8%)	79 (0.8%)	0	
Left ventricular ejection fraction (%)				
≤40%	488 (4.9%)	3183 (31.8%)	74.3	<0.001
41 to 49%	198 (2%)	864 (8.6%)	30	<0.001
≥50%	2345 (23.5%)	4222 (42.2%)	40.8	<0.001
Unknown	6969 (69.7%)	1731 (17.3%)	124.4	<0.001
Hospitalization in prior 1 year	5688 (56.9%)	6644 (66.4%)	19.8	<0.0001
Smoking history	2466 (24.7%)	2790 (27.9%)	7.4	<0.0001
Hypertension	7415 (74.2%)	8952 (89.5%)	40.7	<0.0001
Coronary artery disease	3110 (31.1%)	6367 (63.7%)	69	<0.0001
Acute myocardial infarction	767 (7.7%)	2459 (24.6%)	47.3	<0.0001
Coronary artery bypass graft surgery	236 (2.4%)	710 (7.1%)	22.5	<0.0001
Percutaneous coronary intervention	573 (5.7%)	1241 (12.4%)	23.4	<0.0001
Defibrillator	70 (0.7%)	267 (2.7%)	15.4	<0.0001
Pacemaker	227 (2.3%)	554 (5.5%)	16.9	<0.0001
Atrial fibrillation	937 (9.4%)	2684 (26.8%)	46.6	<0.0001
Lipid disorder	6692 (66.9%)	7826 (78.3%)	25.6	<0.0001
Diabetes mellitus	3375 (33.8%)	5193 (51.9%)	37.4	<0.0001
Stroke	297 (3%)	460 (4.6%)	8.5	<0.0001
Peripheral arterial disease	1721 (17.2%)	2871 (28.7%)	27.6	<0.0001
Chronic obstructive pulmonary disease	2429 (24.3%)	3884 (38.8%)	31.7	<0.0001
Asthma	660 (6.6%)	1059 (10.6%)	14.3	<0.0001
Autoimmune disease	1635 (16.4%)	1789 (17.9%)	4.1	0.0038
Liver disease	947 (9.5%)	1081 (10.8%)	4.4	0.0017
Renal failure and/or dialysis	1376 (13.8%)	2653 (26.5%)	32.3	<0.0001
Cancer	4956 (49.6%)	5415 (54.2%)	9.2	<0.0001
Anaemia of deficiency	939 (9.4%)	1519 (15.2%)	17.7	<0.0001
Anaemia of chronic disease	328 (3.3%)	664 (6.6%)	15.5	<0.0001
Osteoarthritis	4124 (41.2%)	4713 (47.1%)	11.9	<0.0001
Depression	3690 (36.9%)	4257 (42.6%)	11.6	<0.0001
Dementia	632 (6.3%)	535 (5.4%)	4.1	0.0034
Body mass index, kg/m^2	28.6 (±6.6)	31.5 (±7.6)	41.1	<0.0001
Pulse, beat/min	77.1 (±14.9)	80.6 (±17.8)	21.1	<0.0001
Systolic blood pressure, mmHg	134.4 (±18.7)	136.0 (±21.7)	7.7	<0.0001
Diastolic blood pressure, mmHg	76.3 (±11.2)	77.0 (±13.5)	5.4	0.0002
Serum creatinine, mg/dL	1.2 (±0.9)	1.4 (±1.1)	17.1	<0.0001
Serum glucose, mg/dL	150.5 (±63.2)	157.8 (±66.9)	11.2	<0.0001
Serum total cholesterol, mg/dL	170.9 (±43.1)	161.8 (±44.8)	20.7	<0.0001
Serum sodium, mEq/L	138.6 (±3.7)	138.6 (±3.8)	0.6	0.7102
Serum potassium, mEq/L	4.2 (±0.5)	4.2 (±0.5)	2.3	0.1301
Haemoglobin, g/dL	13.6 (±2.1)	13.0 (±2.2)	26	<0.0001
White blood cell, 10^9/L	8.4 (±6.5)	8.5 (±5.2)	1.6	0.2993
Platelets, 10^9/L	226.6 (±81.7)	221.3 (±82.1)	6.5	<0.0001

Of the 3671 patients with left ventricular ejection fraction ≤40%, 86.7% (3183/3671) were in the group with heart failure. Respective proportions for those with ejection fraction 41 to 49%, ≥50% and unknown were 81.4%, 64.3%, and 19.9%.
ASD, absolute standardized difference.

from the 'HF ICD-code universe' (higher sensitivity or recall), we used a weighted *F* score $F_\beta = \frac{(1+\beta^2)\cdot \text{PPV}\cdot \text{Sensitivity}}{\beta^2\cdot \text{PPV} + \text{Sensitivity}}$ where β is a real valued number indicating the importance of sensitivity relative to PPV, specifically we used $\beta = 0.5$. The maximal weighted *F* score values were used to determine the optimal threshold values of each of NLP, ML, and NLP + ML scores to classify HF. Model with the best performing PPV was used to classify HF.

Patient characteristics and outcomes

We compared patient characteristics at baseline of the HF patients classified by the best performing model with those with manually adjudicated gold-standard HF as well as those identified by the three ICD code approaches. The ICD approaches included (i) one principal hospital discharge diagnosis of HF; (ii) two primary outpatient encounter diagnoses for HF; and (iii) either one principal inpatient or two primary

Table 2 Baseline characteristics by 200 patients with and without gold standard heart failure

n (%) or mean (±standard deviation)	No heart failure (*n* = 55)	Heart failure (*n* = 145)	ASD (%)	*P* value
Age, years	67.1 (±12.2)	72.7 (±11.4)	47.4	0.0027
Female	2 (3.6%)	5 (3.4%)	1.0	0.9485
Race				0.0089
White	43 (78.2%)	126 (86.9%)	23.1	
African American	6 (10.9%)	17 (11.7%)	2.6	
Others	6 (10.9%)	2 (1.4%)	40.9	
Left ventricular ejection fraction (%)[a]				
≤40%	4 (7.3%)	37 (25.5%)	50.8	0.0003
41 to 49%	3 (5.5%)	11 (7.6%)	8.6	
≥50%	27 (49.1%)	30 (20.7%)	62.4	
Unknown	21 (38.2%)	67 (46.2%)	16.3	
Hospitalization in prior 1 year	30 (54.5%)	43 (29.7%)	52.1	0.0011
Smoking history	12 (21.8%)	18 (12.4%)	25.2	0.0963
Hypertension	43 (78.2%)	119 (82.1%)	9.8	0.5315
Coronary artery disease	28 (50.9%)	77 (53.1%)	4.4	0.7814
Acute myocardial infarction	8 (14.5%)	23 (15.9%)	3.7	0.8183
Coronary artery bypass graft surgery	1 (1.8%)	4 (2.8%)	6.3	0.7037
Percutaneous coronary intervention	3 (5.5%)	11 (7.6%)	8.6	0.5978
Defibrillator	0 (0%)	4 (2.8%)	23.8	0.2134
Pacemaker	1 (1.8%)	12 (8.3%)	29.8	0.0981
Atrial fibrillation	12 (21.8%)	40 (27.6%)	13.4	0.4063
Lipid disorder	33 (60%)	83 (57.2%)	5.6	0.7241
Diabetes mellitus	23 (41.8%)	58 (40%)	3.7	0.8151
Stroke	1 (1.8%)	3 (2.1%)	1.8	0.9099
Peripheral arterial disease	11 (20%)	32 (22.1%)	5.1	0.7505
Chronic obstructive pulmonary disease	18 (32.7%)	47 (32.4%)	0.7	0.9663
Asthma	5 (9.1%)	10 (6.9%)	8.1	0.5988
Autoimmune disease	3 (5.5%)	20 (13.8%)	28.6	0.0988
Liver disease	10 (18.2%)	5 (3.4%)	48.8	0.0004
Renal failure and/or dialysis	11 (20%)	24 (16.6%)	8.9	0.5666
Cancer	26 (47.3%)	47 (32.4%)	30.7	0.0513
Anaemia of deficiency	5 (9.1%)	13 (9%)	0.4	0.9779
Anaemia of chronic disease	4 (7.3%)	4 (2.8%)	20.8	0.1458
Osteoarthritis	20 (36.4%)	44 (30.3%)	12.8	0.4152
Depression	24 (43.6%)	24 (16.6%)	61.8	<0.0001
Dementia	10 (18.2%)	11 (7.6%)	32	0.0291
Body mass index, kg/m^2	30.2 (±6.9)	30.2 (±7.2)	0.8	0.9639
Pulse, beat/min	77.1 (±17.6)	79.0 (±16.8)	11.5	0.4801
Systolic blood pressure, mmHg	127.6 (±24.0)	134.4 (±22.8)	29.2	0.0884
Diastolic blood pressure, mmHg	73.2 (±12.9)	72.6 (±14.4)	4.5	0.8000
Serum creatinine, mg/dL	1.3 (±0.8)	1.6 (±1.6)	21.0	0.2959
Serum glucose, mg/dL	141.0 (±62.3)	144.6 (±44.8)	6.6	0.8123
Serum total cholesterol, mg/dL	151.0 (±49.5)	165.6 (±39.3)	32.6	0.0757
Serum sodium, mEq/L	138.1 (±3.7)	139.7 (±4.1)	41.5	0.0481
Serum potassium, mEq/L	4.2 (±0.5)	4.3 (±0.6)	15.2	0.4811
Haemoglobin, g/dL	12.7 (±2.7)	13.1 (±2.0)	13.9	0.4008
White blood cell, 10^9/L	7.3 (±2.6)	8.9 (±4.4)	45.7	0.0159
Platelets, 10^9/L	190.8 (±78.5)	228.8 (±79.5)	48.1	0.0064

ASD, absolute standardized difference.

[a] Of the 41 patients with left ventricular ejection fraction ≤40%, 90.2% (37/41) were in the group with heart failure. Respective proportions for those with ejection fraction 41 to 49%, ≥50% and unknown were 78.6%, 52.6%, and 76.1%.

outpatient HF diagnoses. Baselines are based on the dates of the first mention of HF in VA EHR. Baseline comorbidities were defined using ICD codes any time before baseline. For all other baseline characteristics, we used data up to 1 year before baseline. We also examined all-cause mortality, HF hospitalization, and all-cause hospitalization in these patients during 1 and 5 years of follow up, up to 31 December 2022. We also examined all-cause mortality, HF hospitalization, and all-cause hospitalization in these patients during 1 and 5 years of follow up.

Results

Baseline characteristics of the model development cohort

Patients with and without HF used for the model development had a mean age of 67.5 ± 11.2 years, 230 (2.3%) were women, 1913 (19.1%) were African American, and 5000 (50.0%) were inpatient (*Table 1*). Patients with HF had a higher prevalence of cardiovascular risk factors and morbidities such as

Table 3 Performance of various models and traditional ICD code approaches to define and identify patients with HF among 200 patients with and without gold standard HF where gold standard HF that *included* possible HF

	Gold standard HF	HF classified by models or ICD codes	Number of gold standard HF within the HF cohort	Precision (PPV)	Recall (sensitivity)	*F* score	Accuracy
1. NLP	145	151	125	82.8%	86.2%	83.6%	77.0%
2. ML	145	167	128	76.6%	88.3%	79.0%	72.0%
3. NLP + ML	145	143	121	84.6%	83.4%	84.7%	77.0%
4. ICD codes							
a: One or more principal hospital discharge diagnosis	145	31	30	96.8%	20.7%	55.8%	42.0%
b. Two or more primary outpatient encounter diagnoses	145	89	78	87.6%	53.8%	77.8%	61.0%
c. One principal hospital discharge diagnosis or two primary outpatient encounter diagnoses	145	91	80	87.9%	55.2%	78.6%	62.0%

HF, heart failure; ICD, International Classification of Diseases; ML, machine learning; NLP, natural language processing.

hypertension, coronary artery disease, atrial fibrillation, and diabetes mellitus, as well as non-cardiovascular morbidity such as cancer, osteoarthritis, and depression (*Table 1*).

Baseline characteristics of the gold standard cohort

The gold standard cohort of 200 patients included 145 manually adjudicated as HF. Of these, 47 had a score of 2 (definite HF by both reviewers), 9 patients had a score of 3 (definite HF by one reviewer and probable HF by the other), 25 patients had a score of 4 (probable HF by both reviewers, except 3 who had possible HF by one reviewer), 26 had a score of 5 (probable HF by one reviewer and possible HF by the other), and 37 patients had a score of 6 (possible HF by both reviewers, except 1 had unlikely HF by one reviewer). Unlike the model development cohort of 20 000 patients in which non-HF patients had no ICD code for HF, in the gold standard cohort of 200 patients, those without HF had at least one ICD code for HF but a diagnosis of HF could not be confirmed by manual adjudication by two reviewers. Thus, the distribution of cardiovascular risk factors and morbidities such as hypertension, coronary artery disease, and diabetes mellitus were relatively balanced based on ASD $< 10\%$ (*Table 2*).

Evaluation of model performance

The NLP approach used human annotation as the supervision for learning whether the occurrence of a HF keyword within a note was a positive or negative indication of HF based on the context around the keyword. The NLP contributed four binary variables from its classifications: (i) yes HF, (ii) no HF, (iii) HF uncertain, and (iv) no keyword. The model achieved a PPV 0.81 and sensitivity 0.77 in classifying 'yes' compared with other categories; a PPV 0.81 and sensitivity 0.70 in classifying 'uncertain' compared with other categories; and a PPV 0.83 and sensitivity 0.71 in classifying 'no' compared with other categories. The concordance index of the model was 0.89. The concordance index is a performance metric for ordinal classification and can be considered an extension of the area under ROC curve (AUC) for binary classification. The final ML model with 947 variables achieved AUC of 0.94, accuracy of 0.86, sensitivity of 0.86. and specificity of 0.86 on the testing set.

Classification of HF using the optimal threshold score

The highest F scores for the NLP, ML, and NLP + ML models were 0.84, 0.79, and 0.85, respectively (*Table 3*). Corresponding respective optimal threshold NLP, ML, and NLP + ML models were 1.23, 0.55, and 0.63. At the threshold of 0.63, the NLP + ML model identified 143 patients as HF, of whom 121 had gold standard HF (PPV, 0.85), who came from the 145 manually-adjudicated HF patients in the gold standard cohort (sensitivity, 0.83; *Table 3*). PPV and sensitivity of the three ICD code approaches are displayed in *Table 3*.

Patients characteristics and outcomes

Baseline characteristics of 143 patients whose HF was classified by the model were comparable with those with the 145 patients with gold standard HF as well as those identified by ICD codes: 31 with a principal discharge diagnosis of HF, 89 with two primary outpatients encounters for HF, and 91 with either (*Table 4*). During 1 year of follow up from study baseline, 15.9% of the patients from the manually-adjudicated HF cohort and 15.4% of the patients from the model-classified HF cohort died due to all causes (*Table 5*). Respective rates for HF hospitalizations were 9.0% and 7.7%. Other outcomes of the model-classified HF patients and those with gold standard HF and identified by ICD codes are presented in *Table 5*.

Table 4 Baseline characteristics of patients identified as HF using various approaches from a cohort of 200 patients with at least one ICD code for HF.

			HF diagnosed using ICD codes		
n (%) or mean (±standard deviation)	Gold standard HF (n=145)	Model (NLP+ML) classified HF (n=143)	One principal hospital discharge diagnosis of HF (n=31)	Two primary outpatient encounter diagnosis of HF (n=89)	Either one inpatient or two outpatient diagnosis of HF (n=91)
Age, years	72.7 (±11.4)	71.3 (±11.7)	73.5 (±12.6)	71.2 (±12.6)	71.4 (±12.7)
Female	5 (3.4%)	4 (2.8%)	2 (6.5%)	5 (5.6%)	5 (5.5%)
Race					
White	126 (86.9%)	124 (86.7%)	28 (90.3%)	76 (85.4%)	78 (85.7%)
African American	17 (11.7%)	15 (10.5%)	3 (9.7%)	10 (11.2%)	10 (11%)
Others	2 (1.4%)	4 (2.8%)	. (.%)	3 (3.4%)	3 (3.3%)
Left ventricular ejection fraction (%)					
≤40%	37 (25.5%)	33 (23.1%)	10 (32.3%)	21 (23.6%)	21 (23.1%)
41 to 49%	11 (7.6%)	12 (8.4%)	6 (19.4%)	10 (11.2%)	11 (12.1%)
≥50%	30 (20.7%)	39 (27.3%)	11 (35.5%)	24 (27%)	24 (26.4%)
Unknown	67 (46.2%)	59 (41.3%)	4 (12.9%)	34 (38.2%)	35 (38.5%)
Hospitalization in prior one year	43 (29.7%)	48 (33.6%)	20 (64.5%)	26 (29.2%)	28 (30.8%)
Smoking history	18 (12.4%)	18 (12.6%)	8 (25.8%)	16 (18%)	16 (17.6%)
Hypertension	119 (82.1%)	116 (81.1%)	26 (83.9%)	73 (82%)	75 (82.4%)
Coronary artery disease	77 (53.1%)	77 (53.8%)	20 (64.5%)	44 (49.4%)	45 (49.5%)
Acute myocardial infarction	23 (15.9%)	21 (14.7%)	3 (9.7%)	11 (12.4%)	11 (12.1%)
Coronary artery bypass graft surgery	4 (2.8%)	4 (2.8%)	0 (0%)	2 (2.2%)	2 (2.2%)
Percutaneous coronary intervention	11 (7.6%)	9 (6.3%)	2 (6.5%)	4 (4.5%)	4 (4.4%)
Defibrillator	4 (2.8%)	4 (2.8%)	1 (3.2%)	2 (2.2%)	2 (2.2%)
Pacemaker	12 (8.3%)	12 (8.4%)	3 (9.7%)	8 (9%)	8 (8.8%)
Atrial fibrillation	40 (27.6%)	34 (23.8%)	10 (32.3%)	21 (23.6%)	23 (25.3%)
Lipid disorder	83 (57.2%)	87 (60.8%)	16 (51.6%)	49 (55.1%)	50 (54.9%)
Diabetes mellitus	58 (40%)	64 (44.8%)	15 (48.4%)	33 (37.1%)	34 (37.4%)
Stroke	3 (2.1%)	3 (2.1%)	2 (6.5%)	2 (2.2%)	2 (2.2%)
Peripheral arterial disease	32 (22.1%)	31 (21.7%)	7 (22.6%)	15 (16.9%)	15 (16.5%)
Chronic obstructive pulmonary disease	47 (32.4%)	44 (30.8%)	10 (32.3%)	25 (28.1%)	25 (27.5%)
Asthma	10 (6.9%)	12 (8.4%)	2 (6.5%)	9 (10.1%)	9 (9.9%)
Autoimmune disease	20 (13.8%)	15 (10.5%)	4 (12.9%)	8 (9%)	8 (8.8%)
Liver disease	5 (3.4%)	7 (4.9%)	3 (9.7%)	5 (5.6%)	5 (5.5%)
Renal failure and/or dialysis	24 (16.6%)	26 (18.2%)	8 (25.8%)	16 (18%)	16 (17.6%)
Cancer	47 (32.4%)	52 (36.4%)	14 (45.2%)	30 (33.7%)	31 (34.1%)
Anemia of deficiency	13 (9%)	14 (9.8%)	7 (22.6%)	8 (9%)	8 (8.8%)
Anemia of chronic disease	4 (2.8%)	5 (3.5%)	0 (0%)	1 (1.1%)	1 (1.1%)
Osteoarthritis	44 (30.3%)	41 (28.7%)	11 (35.5%)	25 (28.1%)	27 (29.7%)
Depression	24 (16.6%)	28 (19.6%)	9 (29%)	19 (21.3%)	19 (20.9%)
Dementia	11 (7.6%)	11 (7.7%)	1 (3.2%)	5 (5.6%)	5 (5.5%)
Body mass index, kg/m^2	30.2 (±7.2)	30.3 (±7.3)	30.4 (±5.6)	31.0 (±8.0)	31.0 (±8.0)
Pulse, beat/min	79.0 (±16.8)	77.9 (±16.5)	79.9 (±14.2)	79.1 (±16.1)	78.9 (±16.1)
Systolic blood pressure, mmHg	134.4 (±22.8)	133.9 (±23.0)	133.3 (±22.2)	134.1 (±23.5)	134.4 (±23.3)
Diastolic blood pressure, mmHg	72.6 (±14.4)	73.3 (±14.6)	74.8 (±14.5)	73.8 (±15.5)	74.1 (±15.5)
Serum creatinine, mg/dL	1.6 (±1.6)	1.5 (±1.4)	1.4 (±0.5)	1.4 (±0.6)	1.4 (±0.6)
Serum glucose, mg/dL	144.6 (±44.8)	149.2 (±56.2)	145.0 (±55.5)	148.1 (±49.1)	147.2 (±48.1)
Serum total cholesterol, mg/dL	165.6 (±39.3)	161.4 (±40.4)	144.7 (±45.5)	160.3 (±42.9)	160.0 (±42.6)
Serum sodium, mEq/L	139.7 (±4.1)	139.4 (±4.1)	139.3 (±4.0)	139.7 (±3.8)	139.6 (±3.9)
Serum potassium, mEq/L	4.3 (±0.6)	4.2 (±0.6)	4.3 (±0.7)	4.2 (±0.6)	4.2 (±0.6)
Hemoglobin, g/dL	13.1 (±2.0)	12.9 (±2.2)	12.8 (±2.5)	13.1 (±2.2)	13.1 (±2.2)
White blood cell, 10 ^ 9/L	8.9 (±4.4)	8.5 (±4.2)	7.9 (±2.0)	8.6 (±4.5)	8.6 (±4.4)
Platelets, 10 ^ 9/L	228.8 (±79.5)	222.5 (±75.3)	228.5 (±87.4)	225.5 (±75.6)	224.0 (±75.2)

HF, heart failure; ICD, International Classification of Diseases; ML, machine learning; NLP, natural language processing.

Classification of HF using the alternate threshold score

The use of a stricter criteria to defined gold standard HF that excluded possible HF identified 80 patients from the gold standard cohort of 200 patients. The highest *F* scores for the NLP, ML, and NLP + ML models were 0.70, 0.71, and 0.76, respectively (*Table 6*). Corresponding respective optimal threshold NLP, ML, and NLP + ML models were 1.91, 1.72, and 0.13. At the threshold of 0.13, the NLP + ML model identified 64 patients as HF, of whom 51 had gold standard HF (PPV, 0.80), who came from the 80 gold standard HF patients (sensitivity, 0.64). PPV and sensitivity of the three ICD code approaches are displayed in *Table 6*.

Discussion

The findings from our study demonstrate that an approach based on NLP and ML algorithms to classify HF from a pool of patients with ICD codes for HF in the VA EHR performed

Table 5 Outcomes of patients identified as HF using various approaches from a cohort of 200 patients with at least one ICD code for HF.

			HF diagnosed using ICD codes		
Outcomes	Gold standard HF (n=145)	Model (NLP+ML) classified HF (n=143)	One principal hospital discharge diagnosis of HF (n=31)	Two primary outpatient encounter diagnosis of HF (n=89)	Either one inpatient or two outpatient diagnosis of HF (n=91)
All-cause mortality					
1 Year	23 (15.9%)	22 (15.4%)	1 (3.2%)	11 (12.4%)	11 (12.1%)
5 Years	82 (56.6%)	76 (53.1%)	18 (58.1%)	44 (49.4%)	46 (50.5%)
HF Hospitalization					
1 Year	13 (9.0%)	11 (7.7%)	9 (29.0%)	11 (12.4%)	11 (12.1%)
5 Years	34 (23.4%)	31 (21.7%)	21 (67.7%)	27 (30.3%)	28 (30.8%)
All-cause hospitalization					
1 Year	64 (44.1%)	71 (49.7%)	17 (54.8%)	44 (49.4%)	45 (49.5%)
5 Years	101 (69.7%)	104 (72.7%)	26 (83.9%)	68 (76.4%)	70 (76.9%)

HF, heart failure; ICD, International Classification of Diseases; ML, machine learning; NLP, natural language processing.

Table 6 Performance of various models and traditional ICD code approaches to define and identify patients with HF among 200 patients with and without gold standard HF where gold standard HF that *excluded* possible HF

	Gold standard HF	HF classified by models or ICD codes	Number of gold standard HF within the HF cohort	Precision (PPV)	Recall (Sensitivity)	F score	Accuracy
1. NLP	80	52	40	76.9%	50.0%	69.5%	74.0%
2. ML	80	73	53	72.6%	66.3%	71.5%	77.0%
3. NLP + ML	80	64	51	79.6%	63.7%	76.1%	79.0%
4. ICD codes	80						
a: One or more principal hospital discharge diagnosis	80	31	26	83.9%	32.5%	63.7%	70.5%
b. Two or more primary outpatient encounter diagnoses	80	89	53	59.6%	66.3%	60.8%	68.5%
c. One principal hospital discharge diagnosis or two primary outpatient encounter diagnoses	80	91	55	60.4%	68.8%	61.9%	69.5%

HF, heart failure; ICD, International Classification of Diseases; ML, machine learning; NLP, natural language processing; PPV, positive predictive value.

well and that NLP and ML models, alone or in combination, performed better than approaches based on ICD codes alone. These findings also shows that model-identified HF patients were characteristically and prognostically similar to those whose HF was manually adjudicated by extensive chart review by two clinicians. These findings suggest that HF registries created from big EHR data using AI-based approaches are representative of HF populations that are adjudicated by manual chart review and can be useful for answering clinical questions regarding quality and outcomes of care, effectiveness of therapy, and generation of hypothesis.

Accurately phenotyping HF on a population level is challenging yet highly impactful. As illustrated by national registry initiatives such as OPTIMIZE-HF and GWTG-HF, adherence to guideline directed medical therapy in HF can translate directly to mortality benefits and a reduction in costly readmissions.[28,55] In this study, we show that an approach combining NLP and ML techniques can allow for accurate identification of HF using EHR data. This approach is novel in its use of an architecture that combines structured and unstructured data elements into a classification algorithm to boost performance. A distinctive feature of our algorithm is selection of features from unstructured text data through both linear SVM classification of text snippets (supervised) and topic modelling (unsupervised). When topic variables were incorporated alongside standard ML trained on structured variables, the model achieved an AUC of 0.92, which improved further, albeit modestly, to 0.94, when variables from the NLP model were included. Our approach to use text data using a combination of topic modelling and clustering through snippets can be implemented within a different institution and take advantage of subtle differences in documentation practices.

Given the large pool of patients with one or more ICD code for HF in large EHR systems such as that of the VA's, our emphasis was not on identifying all possible true HF cases. Instead, our focus was on including fewer non-HF patients into the HF cohort classified by the model. Thus, we chose an *F* score that was associated with the greatest PPV and sensitivity of all the possible combinations. Even though the *F* score of the NLP model was similar to that of the NLP + ML model, we chose the NLP + ML model as it had a higher PPV. A 2% difference in PPV may translate into additional 20 000 non-HF patients in a cohort of 1 million patients classified by the model. Excluding possible HF cases from the gold standard HF would be expected to make the adjudication more stringent and the resultant cohort to include a higher proportion of true HF. However, we observed that the model did not perform as well when the gold standard HF cohort excluded possible HF cases. Of note, prior studies on HF phenotyping, especially in the ambulatory setting, have included 'possible HF' as gold standard.[56]

All models in both gold standard HF (including and excluding possible HF) seem to have performed better than the traditional ICD code approaches. When HF is diagnosed using a principal hospital discharge diagnosis with a HF ICD code, the PPV was the highest (0.88) but also had the lowest sensitivity (0.31). We found that when the gold standard HF includes possible HF, when both NLP and ML models are combined to calculate an NLP + ML score, and when a weighted *F* score is used to identify the optimal threshold score, the phenotyping process was most efficient, and the assembled HF cohort is characteristically and prognostically similar to those of a manually-adjudicated HF cohort. While it is not surprising given the considerable overlap between the two populations, considering that both manually-adjudicated and model-classified HF cohorts included 122 gold standard HF patients from the gold standard cohort of 200 patients, this also highlights the success of our models to classify HF from a pool of patients with an ICD code for HF in EHR.

Several limitations of our study need to be acknowledged. Our gold standard dataset for estimation of threshold for HF classification was relatively small. We did not compare our algorithm with other published algorithms as most are not readily available as free-standing tools or rely on ICD codes. Our use of 947 variable specifically available in the VA EHR also limits generalizability to other EHR data. However, the objective of our study was to develop AI models to phenotype HF within the VA EHR and was not meant for non-VA EHR. It is akin to the logistic regression models used to calculate propensity scores for a specific patient population and are not meant to be used for other patient populations.[57] Patients in our study are predominantly male US veteran patients which may limit generalizability to other populations.

In conclusion, ML and NLP models performed better than the traditional ICD code-based approaches in identifying patients with HF from big VA EHR data, and these AI-phenotyped HF patients were characteristically and prognostically similar to those manually adjudicated to have HF. These findings suggest that AI approaches are useful in developing HF registries from big EHR data, which can serve as contemporary resources for studies of quality of care and outcome, clinical effectiveness, and hypotheses generation.

Funding

This work was supported by a grant (I01HX002422) from the Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, Health Services Research and Development Service to the Washington DC VA Medical Center. The funding organization or sponsor played no role in the design, analysis, or interpretation of the current study or in the preparation, review, approval, or the decision to submit the manuscript for publication. Disclaimer: The views expressed in this article are those of the authors and do not necessarily reflect the position or policy of the Department of Veterans Affairs or the United States government.

Conflict of interest

We verify and confirm that all conflict-of-interest disclosure information for all co-authors is accurate, complete, up-to-date, and reported in the Acknowledgment section of the manuscript in a manner that is consistent with that reported in the disclosures of potential conflicts of interest section of the journal's authorship form. Dr Fonarow reports consulting for 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.

Data S1. Supporting Information.
Table S1. List of 947 variables used in the final ML model.

References

1. Roger VL. Epidemiology of heart failure: a contemporary perspective. *Circ Res* 2021;**128**:1421–1434. doi:10.1161/CIRCRESAHA.121.318172
2. Tsao CW, Aday AW, Almarzooq ZI, Anderson CA, Arora P, Avery CL,

et al. Heart disease and stroke statistics-2023 update: a report from the American Heart Association. *Circulation* 2023;**147**:e93–e621. doi:10.1161/CIR.0000000000001123

3. Heidenreich PA, Bozkurt B, Aguilar D, Allen LA, Byun JJ, Colvin MM, *et al*. 2022 AHA/ACC/HFSA guideline for the management of heart failure: executive summary: a report of the American College of Cardiology/American Heart Association Joint Committee on clinical practice guidelines. *Circulation* 2022; **145**:e876–e894. doi:10.1161/CIR.0000000000001062
4. Ahmed A, Zannad F, Love TE, Tallaj J, Gheorghiade M, Ekundayo OJ, *et al*. A propensity-matched study of the association of low serum potassium levels and mortality in chronic heart failure. *Eur Heart J* 2007;**28**:1334–1343. doi:10.1093/eurheartj/ehm091
5. Campbell RC, Sui X, Filippatos G, Love TE, Wahle C, Sanders PW, *et al*. Association of chronic kidney disease with outcomes in chronic heart failure: a propensity-matched study. *Nephrol Dial Transplant* 2009;**24**:186–193. doi:10.1093/ndt/gfn445
6. Ahmed A, Rich MW, Love TE, Lloyd-Jones DM, Aban IB, Colucci WS, *et al*. Digoxin and reduction in mortality and hospitalization in heart failure: a comprehensive post hoc analysis of the DIG trial. *Eur Heart J* 2006;**27**: 178–186. doi:10.1093/eurheartj/ehi687
7. Bourge RC, Fleg JL, Fonarow GC, Cleland JGF, McMurray JJV, van Veldhuisen DJ, *et al*. Digoxin reduces 30-day all-cause hospital admission in older patients with chronic systolic heart failure. *Am J Med* 2013;**126**:701–708. doi:10.1016/j.amjmed.2013.02.001
8. Ahmed A, Rich MW, Fleg JL, Zile MR, Young JB, Kitzman DW, *et al*. Effects of digoxin on morbidity and mortality in diastolic heart failure: the ancillary digitalis investigation group trial. *Circulation* 2006;**114**:397–403. doi:10.1161/CIRCULATIONAHA.106.628347
9. Bowling CB, Sanders PW, Allman RM, Rogers WJ, Patel K, Aban IB, *et al*. Effects of enalapril in systolic heart failure patients with and without chronic kidney disease: insights from the SOLVD treatment trial. *Int J Cardiol* 2013;**167**:151–156. doi:10.1016/j.ijcard.2011.12.056
10. Lam PH, Dooley DJ, Fonarow GC, Butler J, Bhatt DL, Filippatos GS, *et al*. Similar clinical benefits from below-target and target dose enalapril in patients with heart failure in the SOLVD treatment trial. *Eur J Heart Fail* 2018;**20**:359–369. doi:10.1002/ejhf.937
11. Lam PH, Packer M, Fonarow GC, Faselis C, Allman RM, Morgan CJ, *et al*. Early effects of starting doses of enalapril in patients with chronic heart failure in the SOLVD treatment trial. *Am J Med* 2020;**133**:e25–e31. doi:10.1016/j.amjmed.2019.06.053
12. Lam PH, Keramida K, Filippatos GS, Gupta N, Faselis C, Deedwania P, *et al*. Right ventricular ejection fraction and Beta-blocker effect in heart failure with reduced ejection fraction. *J Card Fail* 2022;**28**:65–70. doi:10.1016/j.cardfail.2021.07.026
13. Meyer P, Filippatos GS, Ahmed MI, Iskandrian AE, Bittner V, Perry GJ, *et al*. Effects of right ventricular ejection fraction on outcomes in chronic systolic heart failure. *Circulation* 2010;**121**:252–258. doi:10.1161/CIRCULATIONAHA.109.887570
14. Smith A, Kumar S, Moore HJ, *et al*. Imaging modality for left ventricular ejection fraction estimation and effect of implantable cardioverter defibrillator on mortality in patients with heart failure. *Heart Rhythm* 2023;**20**:886–890. doi:10.1016/j.hrthm.2023.03.010
15. D'Agostino RB Jr, D'Agostino RB Sr. Estimating treatment effects using observational data. *JAMA* 2007;**297**:314–316. doi:10.1001/jama.297.3.314
16. Fonarow GC, Corday E, Committee ASA. Overview of acutely decompensated congestive heart failure (ADHF): a report from the ADHERE registry. *Heart Fail Rev* 2004;**9**:179–185. doi:10.1007/s10741-005-6127-6
17. Zhang Y, Kilgore ML, Arora T, *et al*. Design and rationale of studies of neurohormonal blockade and outcomes in diastolic heart failure using OPTIMIZE-HF registry linked to Medicare data. *Int J Cardiol* 2013;**166**:230–235. doi:10.1016/j.ijcard.2011.10.089
18. Feller MA, Mujib M, Zhang Y, Ekundayo OJ, Aban IB, Fonarow GC, *et al*. Baseline characteristics, quality of care, and outcomes of younger and older Medicare beneficiaries hospitalized with heart failure: findings from the Alabama Heart Failure Project. *Int J Cardiol* 2012;**162**:39–44. doi:10.1016/j.ijcard.2011.05.003
19. Lam PH, Dooley DJ, Deedwania P, *et al*. Heart rate and outcomes in hospitalized patients with heart failure with preserved ejection fraction. *J Am Coll Cardiol* 2017;**70**:1861–1871. doi:10.1016/j.jacc.2017.08.022
20. Tsimploulis A, Lam PH, Arundel C, *et al*. Systolic blood pressure and outcomes in patients with heart failure with preserved ejection fraction. *JAMA Cardiol* 2018;**3**:288–297. doi:10.1001/jamacardio.2017.5365
21. Arundel C, Lam PH, Gill GS, Patel S, Panjrath G, Faselis C, *et al*. Systolic blood pressure and outcomes in patients with heart failure with reduced ejection fraction. *J Am Coll Cardiol* 2019;**73**: 3054–3063. doi:10.1016/j.jacc.2019.04.022
22. Bayoumi E, Lam PH, Dooley DJ, Singh S, Faselis C, Morgan CJ, *et al*. Spironolactone and outcomes in older patients with heart failure and reduced ejection fraction. *Am J Med* 2019;**132**:71–80 e1. doi:10.1016/j.amjmed.2018.09.011
23. Faselis C, Arundel C, Patel S, Lam PH, Gottlieb SS, Zile MR, *et al*. Loop diuretic prescription and 30-day outcomes in older patients with heart failure. *J Am Coll Cardiol* 2020;**76**:669–679. doi:10.1016/j.jacc.2020.06.022
24. Qamer SZ, Malik A, Bayoumi E, Lam PH, Singh S, Packer M, *et al*. Digoxin use and outcomes in patients with heart failure with reduced ejection fraction. *Am J Med* 2019;**132**:1311–1319. doi:10.1016/j.amjmed.2019.05.012
25. Malik A, Masson R, Singh S, Wu WC, Packer M, Pitt B, *et al*. Digoxin discontinuation and outcomes in patients with heart failure with reduced ejection fraction. *J Am Coll Cardiol* 2019;**74**:617–627. doi:10.1016/j.jacc.2019.05.064
26. Fonarow GC, Stough WG, Abraham WT, Albert NM, Gheorghiade M, Greenberg BH, *et al*. Characteristics, treatments, and outcomes of patients with preserved systolic function hospitalized for heart failure: a report from the OPTIMIZE-HF registry. *J Am Coll Cardiol* 2007;**50**: 768–777. doi:10.1016/j.jacc.2007.04.064
27. Shah KS, Xu H, Matsouaka RA, Bhatt DL, Heidenreich PA, Hernandez AF, *et al*. Heart failure with preserved, borderline, and reduced ejection fraction: 5-year outcomes. *J Am Coll Cardiol* 2017;**70**:2476–2486. doi:10.1016/j.jacc.2017.08.074
28. Fonarow GC, Abraham WT, Albert NM, Gattis WA, Gheorghiade M, Greenberg B, *et al*. Organized program to initiate lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF): rationale and design. *Am Heart J* 2004;**148**:43–51. doi:10.1016/j.ahj.2004.03.004
29. Smaha LA, American HA. The American Heart Association get with the guidelines program. *Am Heart J* 2004;**148**:S46–S48. doi:10.1016/j.ahj.2004.09.015
30. Ellrodt AG, Fonarow GC, Schwamm LH, Albert N, Bhatt DL, Cannon CP, *et al*. Synthesizing lessons learned from get with the guidelines: the value of disease-based registries in improving quality and outcomes. *Circulation* 2013; **128**:2447–2460. doi:10.1161/01.cir.0000435779.48007.5c
31. Havranek EP, Masoudi FA, Westfall KA, Wolfe P, Ordin DL, Krumholz HM. Spectrum of heart failure in older patients: results from the National Heart Failure project. *Am Heart J* 2002;**143**:412–417. doi:10.1067/mhj.2002.120773
32. Havranek EP, Masoudi FA, Smith GL, Wolfe P, Ralston DL, Krumholz HM, *et al*. Lessons learned from the national heart failure project: a center for Medicare and Medicaid services initiative to improve the care of Medicare beneficiaries with heart failure. *Congest Heart Fail* 2001;**7**:334–336. doi:10.1111/j.1527-5299.2001.00273.x
33. Fonarow GC, Abraham WT, Albert NM, Gattis Stough W, Gheorghiade M,

Greenberg BH, *et al*. Influence of a performance-improvement initiative on quality of care for patients hospitalized with heart failure: results of the organized program to initiate lifesaving treatment in hospitalized patients with heart failure (OPTIMIZE-HF). *Arch Intern Med* 2007;**167**:1493–1502. doi:10.1001/archinte.167.14.1493

34. Fonarow GC, Abraham WT, Albert NM, Stough WG, Gheorghiade M, Greenberg BH, *et al*. Influence of beta-blocker continuation or withdrawal on outcomes in patients hospitalized with heart failure: findings from the OPTIMIZE-HF program. *J Am Coll Cardiol* 2008;**52**:190–199. doi:10.1016/j.jacc.2008.03.048
35. Hammill BG, Hernandez AF, Peterson ED, Fonarow GC, Schulman KA, Curtis LH. Linking inpatient clinical registry data to Medicare claims data using indirect identifiers. *Am Heart J* 2009;**157**:995–1000. doi:10.1016/j.ahj.2009.04.002
36. Aguirre Dávila L, Weber K, Bavendiek U, Bauersachs J, Wittes J, Yusuf S, *et al*. Digoxin-mortality: randomized vs. observational comparison in the DIG trial. *Eur Heart J* 2019;**40**:3336–3341. doi:10.1093/eurheartj/ehz395
37. Bielinski SJ, Pathak J, Carrell DS, Takahashi PY, Olson JE, Larson NB, *et al*. A robust e-epidemiology tool in phenotyping heart failure with differentiation for preserved and reduced ejection fraction: the Electronic Medical Records and Genomics (eMERGE) network. *J Cardiovasc Transl Res* 2015;**8**: 475–483. doi:10.1007/s12265-015-9644-2
38. Patel YR, Robbins JM, Kurgansky KE, *et al*. Development and validation of a heart failure with preserved ejection fraction cohort using electronic medical records. *BMC Cardiovasc Disord* 2018;**18**:128. doi:10.1186/s12872-018-0866-5
39. Cunningham JW, Singh P, Reeder C, Claggett B, Marti-Castellote PM, Lau ES, *et al*. Natural language processing for adjudication of heart failure in a multicenter clinical trial: a secondary analysis of a randomized clinical trial. *JAMA Cardiol* 2023;**9**:174–181. doi:10.1001/jamacardio.2023.4859
40. Cunningham JW, Singh P, Reeder C, *et al*. Natural language processing for adjudication of heart failure in the electronic health record. *JACC Heart Fail* 2023;**11**:852–854. doi:10.1016/j.jchf.2023.02.012
41. Garan AR, Monda KL, Dent-Acosta RE, Riskin DJ, Gluckman TJ. Retrospective comparison of traditional and artificial intelligence-based heart failure phenotyping in a US health system to enable real-world evidence. *BMJ Open* 2023; **13**:e073178. doi:10.1136/bmjopen-2023-073178
42. Levinson R, Malinowski J, Bielinski S, et al. Identifying heart failure from electronic health records: a systematic evidence review *medRxiv* 2022. 10.1101/2021.02.01.21250933
43. Gaziano JM, Concato J, Brophy M, Fiore L, Pyarajan S, Breeling J, *et al*. Million veteran program: a mega-biobank to study genetic influences on health and disease. *J Clin Epidemiol* 2016;**70**:214–223. doi:10.1016/j.jclinepi.2015.09.016
44. Sofer T, Kurniansyah N, Murray M, *et al*. Genome-wide association study of obstructive sleep apnoea in the million veteran program uncovers genetic heterogeneity by sex. *EBioMedicine* 2023;**90**:104536. doi:10.1016/j.ebiom.2023.104536
45. Honerlaw J, Ho YL, Fontin F, *et al*. Framework of the centralized interactive phenomics resource (CIPHER) standard for electronic health data-based phenomics knowledgebase. *J Am Med Inform Assoc* 2023;**30**:958–964. doi:10.1093/jamia/ocad030
46. Rajeevan N, Niehoff KM, Charpentier P, Levin FL, Justice A, Brandt CA, *et al*. Utilizing patient data from the veterans administration electronic health record to support web-based clinical decision support: informatics challenges and issues from three clinical domains. *BMC Med Inform Decis Mak* 2017;**17**:111. doi:10.1186/s12911-017-0501-x
47. Deswal A, Petersen NJ, Urbauer DL, Wright SM, Beyth R. Racial variations in quality of care and outcomes in an ambulatory heart failure cohort. *Am Heart J* 2006;**152**:348–354. doi:10.1016/j.ahj.2005.12.004
48. Wu WC, Jiang L, Friedmann PD, Trivedi A. Association between process quality measures for heart failure and mortality among US veterans. *Am Heart J* 2014;**168**:713–720. doi:10.1016/j.ahj.2014.06.024
49. Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, *et al*. Automating quality measures for heart failure using natural language processing: a descriptive study in the Department of Veterans Affairs. *JMIR Med Inform* 2018;**6**:e5. doi:10.2196/medinform.9150
50. Goulet JL, Erdos J, Kancir S, Levin FL, Wright SM, Daniels SM, *et al*. Measuring performance directly using the veterans health administration electronic medical record: a comparison with external peer review. *Med Care* 2007;**45**:73–79. doi:10.1097/01.mlr.0000244510.09001.e5
51. Austin PC. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Stat Med* 2009;**28**:3083–3107. doi:10.1002/sim.3697
52. Shao Y, Todd K, Shutes-David A, Millard SP, Brown K, Thomas A, et al. Identifying probable dementia in undiagnosed Black and White Americans using machine learning in Veterans Health Administration electronic health records. *medRxiv*. 2023. doi:10.1101/2023.02.08.23285540
53. Redd DF, Shao Y, Zeng-Treitler Q, Myers LJ, Barker BC, Nelson SJ, *et al*. Identification of colorectal cancer using structured and free text clinical data. *Health Informatics J* 2022; **28**:14604582221134406. doi:10.1177/14604582221134406
54. The U.S. Department of Veterans Affairs Centralized Interactive Phenomics Resource (CIPHER). Heart failure (Ahmed and Zeng). Accessed February 14, 2024. https://phenomics.va.ornl.gov/web/cipher/phenotype-viewer?uqid=8964857ed16845bfbc71cc63b3bb8986&name=Heart_Failure__Ahmed_and_Zeng_
55. Golwala H, Pandey A, Ju C, Butler J, Yancy C, Bhatt DL, *et al*. Temporal trends and factors associated with cardiac rehabilitation referral among patients hospitalized with heart failure: findings from get with the guidelines-heart failure registry. *J Am Coll Cardiol* 2015;**66**:917–926. doi:10.1016/j.jacc.2015.06.1089
56. Rosamond WD, Chang PP, Baggett C, Johnson A, Bertoni AG, Shahar E, *et al*. Classification of heart failure in the atherosclerosis risk in communities (ARIC) study: a comparison of diagnostic criteria. *Circ Heart Fail* 2012;**5**:152–159. doi:10.1161/CIRCHEARTFAILURE.111.963199
57. Ekundayo OJ, Allman RM, Sanders PW, Aban I, Love TE, Arnett D, *et al*. Isolated systolic hypertension and incident heart failure in older adults: a propensity-matched study. *Hypertension* 2009;**53**:458–465. doi:10.1161/HYPERTENSIONAHA.108.119792