Enhancing ICD-Code-Based Case Definition for Heart Failure Using Electronic Medical Record Data

YUAN XU, MD, PhD,[1,2,3,4] SEUNGWON LEE, MSc,[3,4,5] ELLIOT MARTIN, PhD,[4,5] ADAM G. D'SOUZA, PhD,[4,5] CHELSEA T.A. DOKTORCHIK, MSc,[3,4] JASON JIANG, MSc,[4,5] SANGMIN LEE, MSc,[3,4] CATHY A. EASTWOOD, RN, PhD,[3,4] NOWELL FINE, MD, MSc,[6] BRENDA HEMMELGARN, MD, PhD,[3] KATHRYN TODD, PhD,[5,7] AND HUDE QUAN, MD, PhD[3,4]

Calgary, and Edmonton, Canada

ABSTRACT

Background: Surveillance and outcome studies for heart failure (HF) require accurate identification of patients with HF. Algorithms based on *International Classification of Diseases* (ICD) codes to identify HF from administrative data are inadequate owing to their relatively low sensitivity. Detailed clinical information from electronic medical records (EMRs) is potentially useful for improving ICD algorithms. This study aimed to enhance the ICD algorithm for HF definition by incorporating comprehensive information from EMRs.

Methods: The study included 2106 inpatients in Calgary, Alberta, Canada. Medical chart review was used as the reference gold standard for evaluating developed algorithms. The commonly used ICD codes for defining HF were used (namely, the ICD algorithm). The performance of different algorithms using the free text discharge summaries from a population-based EMR were compared with the ICD algorithm. These algorithms included a keyword search algorithm looking for HF-specific terms, a machine learning−based HF concept (HFC) algorithm, an EMR structured data based algorithm, and combined algorithms (the ICD and HFC combined algorithm).

Results: Of 2106 patients, 296 (14.1%) were patients with HF as determined by chart review. The ICD algorithm had 92.4% positive predictive value (PPV) but low sensitivity (57.4%). The EMR keyword search algorithm achieved a higher sensitivity (65.5%) than the ICD algorithm, but with a lower PPV (77.6%). The HFC algorithm achieved a better sensitivity (80.0%) and maintained a reasonable PPV (88.9%) compared with the ICD algorithm and the keyword algorithm. An even higher sensitivity (83.3%) was reached by combining the HFC and ICD algorithms, with a lower PPV (83.3%). The structured EMR data algorithm reached a sensitivity of 78% and a PPV of 54.2%. The combined EMR structured data and ICD algorithm had a higher sensitivity (82.4%), but the PPV remained low at 54.8%. All algorithms had a specificity ranging from 87.5% to 99.2%.

Conclusions: Applying natural language processing and machine learning on the discharge summaries of inpatient EMR data can improve the capture of cases of HF compared with the widely used ICD algorithm. The utility of the HFC algorithm is straightforward, making it easily applied for HF case identification. (*J Cardiac Fail 2020;26:610−617*)

Key Words: Electronic medical record, heart failure, case definition, machine learning, natural language processing.

From the [1]Department of Oncology, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [2]Department of Surgery, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [3]Department of Community Health Sciences, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [4]Centre for Health Informatics, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada; [5]Alberta Health Services, Calgary, Alberta, Canada; [6]Division of Cardiology, Department of Cardiac Sciences, Libin Cardiovascular Institute of Alberta, Cumming School of Medicine, University of Calgary, Calgary, Alberta, Canada and [7]Neurochemical Research Unit, Department of Psychiatry, University of Alberta, Edmonton, Alberta, Canada.

Manuscript received January 10, 2020; revised manuscript received March 7, 2020; revised manuscript accepted April 1, 2020.

Reprint requests: Yuan Xu, 3280 Hospital Drive NW, Calgary, Alberta, Canada, T2N4Z6. Office phone: 403-210-9554. E-mail: yuxu@ucalgary.ca

Funding: Dr. Yuan Xu received the funding support from Canadian Institutes of Health Research-Health System Impact Fellow (CIHR-HSIF) award. The funder has no role in the design and conduct of the study, in the collection, analysis, and interpretation of the data, and in the preparation, review, or approval of the manuscript.

1071-9164/$ - see front matter

https://doi.org/10.1016/j.cardfail.2020.04.003

Heart failure (HF) is a chronic condition with immense economic impact and disease burden for patients across health care systems. An estimated 26 million people worldwide are affected by HF,[1] with more than 1 million hospitalizations per year in both the United States and Europe.[2,3] In Canada (2004−2013), symptomatic HF led to more than 42,000 hospitalizations annually with a median average acute length of stay of 6 days, and consumed an average of $10,000 per patient for direct services per year.[4] Decreasing health care use and costs while improving the quality of care for patients with HF is a top research and health system delivery priority. Previous research has studied early detection of HF,[5,6] as well as prediction of HF prognosis and complications.[7] Such work requires accurate HF identification from a population-based data source to capture the true prevalence of HF within a population.

HF is commonly identified in research and surveillance studies using *International Classification of Diseases* (ICD) codes in administrative databases.[8] However, certain diseases, such as HF, can be undercoded in administrative databases (eg, the Canadian Discharge Abstract Database [DAD]) owing to time and resource (eg, incomplete clinical documentation) constraints. These constraints are placed on coders who abstract the diagnostic codes from medical charts, which are comprehensive records of a patient's medical history and clinical information and usually in paper and electronic format. As such, ICD-based algorithms used to identify cases of HF in administrative databases often have poor sensitivity.[8,9] Chart review or clinician-driven prospective data collection are regarded as the most accurate methods for ascertaining HF.[10,11] However, these methods are extremely laborious and time consuming, and are not feasible for large studies. Data from electronic medical records (EMR), an electronic format of medical charts, may improve the accuracy of ICD-based HF case definitions and decrease the amount of manual labor required. Additionally, machine learning (ML) and natural language processing (NLP) have been increasingly applied in the identification of cases of HF from the EMR because of their ability to deal with nonstructured data types, including free text data. Previously published EMR-based HF case definitions using ML methods have better validity than ICD algorithms.[6,12,13] However, the previous studies developed the HF case definitions using an EMR from a single center or selected cohort, and their ML methods are not completely data driven, making them not easily applicable to other settings. Moreover, to our knowledge, there is no study to date using population-based inpatient EMR data to identify cases of HF in Canada.

The city of Calgary in the province of Alberta, Canada, has maintained a city-wide EMR system since 2006. This system covers all acute care inpatient facilities and serves a population of more than 1 million.[14] EMRs represent a comprehensive source of clinical information compared with administrative data. However, EMRs include data in formats that are not amenable to traditional statistical analysis, such as free text. Leveraging the EMR data fully for research purposes requires methods for extracting key clinical information from free text data. The purpose of this study was to develop and provide validity support for various algorithms incorporating EMR data to enhance the identification of HF using chart review data as the reference standard. The developed EMR algorithms include (1) a search algorithm looking for HF-specific keywords, (2) an ML-based HF concept (HFC) algorithm, (3) an algorithm based on structured EMR data, and (4) combined algorithms such as the ICD and HFC combined algorithm. The performance of EMR and ICD algorithms in identifying HF was assessed by comparing against the chart review data (gold standard); then, the performance of these algorithms was compared with each other.

Methods

Study Cohort

The study cohort included a randomly selected sample of 3043 adult patients (≥18 years of age) who were admitted to 1 of 3 major acute care hospitals in Calgary between January 1 and June 30, 2015. All-cause admissions were included except live births (obstetrics). The 3043 inpatients' chart review data were linked to the Sunrise Clinical Manager EMR and DAD. From the 3043 patients, 2118 patients who had complete discharge summaries were included. The discharge summary is an essential component of the EMR, providing a medical narrative of inpatient admission encounters. The missing electronic discharge summaries were further investigated through an extended chart review currently ongoing in a separate project using the same dataset, to mitigate against potential bias introduced by missing records. This study was reviewed and approved by the Conjoint Health Research Ethics Board, University of Calgary (Ethics ID: REB17-1898).

Data Sources

Clinical Chart Review Data. Six nurses underwent training by 1 clinician researcher (CAE) before conducting chart review. These chart reviewers were clinical staff nurses with 1−6 years of nursing experience. Two independent reviewers assessed the presence or absence of HF in a subset of 60 patient charts. Reviewers achieved an acceptable level of consensus, with an final interrater agreement (Cohen's kappa) of greater than 88%. Cases with unclear data or ambiguity at the time of primary data entry were reviewed by the chart reviewers, with outcomes agreed upon by consensus. The chart reviewers then manually reviewed the 3043 patients' entire paper charts (1 admission per patient) including the illness history, test results, medications, and discharge summary. HF was defined as "includes acute and chronic systolic or diastolic heart failure (HF); includes left, right, and biventricular heart failure with reduced or preserved ejection fraction. Includes HF from congenital deformities, valvular disease, hypertension, or pregnancy; includes pulmonary edema with heart failure; includes cardiomyopathy (any kind); cardiomegaly if HF is also listed; if pulmonary hypertension, also look for right heart failure. Various forms of edema or anasarca can be due to HF; so, can be portal hypertension and chronic or end-stage kidney disease." The reviewers classified HF status as "yes," "no," or "possible." To create a binary variable denoting cases of HF for validation purposes, patients ($n = 12$ [0.05%] out of a total of 2118 patients with electronic discharge summaries) who were classified as possible HF by reviewers were excluded, leading to a final study data set that included 2106 patients.

DAD. In Canada, administrative data provide an abstract of the primary clinical data (such as the EMR). Administrative data are coded by professionally trained coders using ICD codes or other coding systems, with a national standard. In the context of this article, administrative data mainly refer to the ICD-coded data specifically the Canadian DAD. The DAD is a national administrative database in Canada that covers all inpatient hospital admissions and captures up to 25 ICD, 10th version (ICD-10) coded discharge diagnoses and 20 Canadian Classification for Health Intervention coded procedures, in addition to patient demographic and administrative information. The codes are assigned by professional data collection specialists (coders), based on the clinical documentation in the patients' charts at the time of discharge. These coding specialists are certified by the Canadian College of Health Information Management and are employed by Alberta Health Services, the single health authority in the province of Alberta. The DAD is commonly used in health services research owing to its very broad coverage of the inpatient population.

EMR Data. The population-based EMR data were used, which included the entire population in a distinct area and minimizes the selection bias. The population-based EMR data in this study refer to the city-wide EMR data, which include all eligible subjects in a defined geographical region (ie, the city of Calgary). The AllScripts Sunrise Clinical Manager EMR is the EMR system used in all Calgary inpatient hospitals, containing more than 5.5 million individual patients over a 10-year period. Physicians, nurses, and other clinicians routinely enter data into the system in structured or unstructured (narrative) formats. Structured data come from discrete fields (eg, a laboratory value, a pull-down box option), and unstructured data come from free text boxes. Extensive clinical information from nursing documentation, physician dictations, imaging reports, multidisciplinary progress records, clinician orders, and laboratory results are retained by Alberta Health Services. Dictated documents from the transcription interface (such as dictated discharge summaries, consultation, and radiologist reports) are stored as multiple copies, with line numbers for easy restructuring of data based on analytical need.

Development of HF Case-Finding Algorithms

ICD Algorithm. The previously established ICD-10 algorithm[15] (algorithm #1 in Table 1) was used to identify HF from DAD data, and its performance was determined by comparing against the chart review data (gold standard). The performance of the ICD-10 algorithm was then compared with that of the alternative proposed case definitions.

EMR-Based Algorithms. A set of EMR-based definitions for HF (algorithms #2, 4, and 6 in Table 1) was created, varying the methodology (keywords vs HFC) and components of the EMR used. To develop the case definitions for HF using EMR, the domain experts in the study team, including a cardiologist (NF) and a nurse specialized in cardiovascular disease (CAE), identified clinical information indicative of a patient having HF, including medications, laboratory or imaging tests, diagnostic reports, progress reports, and discharge summaries. The study team extracted needed data elements from Sunrise Clinical Manager EMR and designed algorithms.

Algorithm Using Only the Structured EMR Data. The following structured data elements were extracted: the B-type natriuretic peptide (BNP) or N-terminal pro BNP test results from laboratory data, the inotropic and intravenous diuretic medications from the medication data. The structured EMR data algorithm identified cases as HF if the BNP test was outside normal range (BNP was >50 pg/mL or N-terminal pro BNP was >125 pg/mL) or inotropic or intravenous diuretic medications were administered (algorithm #6 in Table 1).

Keyword Algorithms. HF-specific keywords (see Supplementary Information) were developed for extracting information from the discharge summary. All the HF related terms from the unified medical language system (UMLS) metathesaurus were pulled that was provided by the United States National Library of Medicine, by using MetaMap[16] in term processing mode with the relaxed model and the "ignore word order" option enabled, to identify synonyms for the term "heart failure." UMLS maps synonymous terms to the same underlying concept and accounts for variations in terminology between providers. For example, in UMLS, the clinical concept of "heart failure" is assigned a concept unique identifier (CUI) of C0018801. The 2018AB release contains 62 synonyms for this clinical concept, including "cardiac failure," "weak heart," "insufficiency cardiac," and "myocardial failure," among others. All these synonyms map to the same CUI. Then, the clinical experts in the study team combined some of the UMLS terms because they may be captured by 1 search keyword (eg, research "heart failure" can cover both "heart failure" and "congestive heart failure") to make sure the final research keywords (see Supplementary File) covered all the UMLS terms for HF. The keyword algorithm (algorithm #2 in Table 1) then labeled the discharge summary where it contained any of these inclusion keywords as positive unless it also contained an exclusion keyword (eg, no, not, or denied). The selected keywords and negation terms were trained in randomly selected 80% of the cohort, and its validity was evaluated in the remaining 20% cohort. The complete list of inclusion and exclusion keywords is available in the Supplementary File.

HFC Algorithm. The HFC algorithm is defined as follows. The raw, free text discharge summaries from 2106

Table 1. Description of the Algorithms Used to Define Heart Failure

Algorithm	Description of the Algorithm	Data Source
#1 ICD - 10	Present in any of following ICD-10 codes: I09.9, I11.0, I13.0, I13.2, I25.5, I42.0, I42.5−I42.9, I43.x, I50.x, P29.0	DAD
#2 Keyword	Keywords*	Discharge summary
#3 Combined #1 and #2	ICD - 10 + keywords*	DAD + discharge summary
#4 ML	ML	Discharge summary
#5 Combined #1 and #4	ICD - 10 + ML	DAD + discharge summary
#6 EMR structured data	Structured EMR	Laboratory result (blood BNP level), medications (inotropics and intravenous diuretics)
#7 Combined #1 and #6	ICD - 10 + structured EMR	DAD + structured EMR
#8 Combined #1, #2 and #6	ICD - 10 + keywords* + structured EMR	DAD + discharge summary + structured EMR

ICD-10, *International Classification of Disease*, 10th version; DAD, discharge abstract data; EMR, electronic medical record; ML, machine learning; BNP, brain natriuretic peptide.

*The keywords are listed in the supplementary file.

patients of the study cohort were obtained. These raw summaries were processed to extract clinical concepts as covariates. XGBoost[17] was used for feature selection and dimensionality reduction. This step helped with model interpretation. Specifically, the XGBoost model was trained to predict the target outcome (HF status) with UMLS CUIs as input features. The feature importance of each CUI for identifying HF was measured, and the 6 top performing features were retained. The 6 selected features were then used as indicators to build an interpretable classifier algorithm with HF status labels from the chart review as the outcome target. The complete, reproducible details of the ML analysis used to develop the HFC algorithm can be found in the Supplementary File.

Combined ICD- and EMR-Based Algorithms. The combined algorithms were created by applying the ICD-10−based algorithm and the EMR-based algorithms separately to find the cases of HF (algorithms #3, 5, 7, and 8 in Table 1). The presence of HF was defined through Boolean logic “or.” In other words, the patient had HF if found by either algorithm; otherwise, it was defined as a non-HF case. Following this rule, the ICD-10 and keyword combined algorithm (algorithm #3), ICD-10 and HFC combined algorithm (algorithm #5), ICD-10 and EMR structured data combined algorithm (algorithm #7), and ICD-10, keyword, and EMR structured data combined algorithm (algorithm #8) were constructed.

Validation and Statistical Analysis

Descriptive statistics were conducted to summarize the baseline characteristics of the cohort. Sensitivity, specificity, positive predictive value (PPV), negative predictive value, and F1 score (ie, 2 × [(sensitivity × PPV)/(sensitivity + PPV)]) were calculated using the chart review as the gold standard, for the keyword algorithms, ICD algorithm, and HFC algorithm (at the threshold that maximized accuracy in the latter case). The 95% confidence intervals for the validation metrics were calculated assuming a binomial distribution. The analyses were performed using SAS 9.4 (SAS Institute Inc., Cary, NC) for descriptive analyses, MetaMap for keyword selection, and Python 3.6 and cTAKES 3.2 for the HFC algorithm.

Results

Baseline Characteristics of Study Cohort

In total, 2106 patients were included after the chart review cohort. Among the 2106 patients, 296 (14.1%) were patients with HF determined by chart review. The median age was 64 years (interquartile range 49−78 years), and 48.6% ($n = 1023$) were older than 64 years. Among the entire cohort, 50.2% ($n = 1057$) were female, and 53.9% ($n = 1136$) had a Charlson comorbidity score of greater than or equal to 2 points. The Charlson comorbidity score includes 17 chronic conditions with each condition category associating a weight, and the sum of all the weights results in a score (ranging from 0 to 33). A higher score indicates a greater comorbidity burden. Patients who were admitted to a surgical hospital service accounted for 24.8% ($n = 526$) of the study cohort. The mortality rate at discharge was 2.3% ($n = 48$) (Table 2). Although there was no significant difference in age and sex between patients with and without a discharge summary, patients with missing discharge summaries (paper discharge summaries were often present but not usable for this study) were more likely to be surgical patients.

Comparing the Performance of the Algorithms

All algorithms had a specificity ranging from 87.5% to 99.2% and a negative predictive value or more than 93% (Table 3). The ICD algorithm for DAD data also had a high PPV (92.4%; 95% confidence interval [CI] 88.6−96.2%) but low sensitivity (57.4%; 95% CI 51.8%−63.0%). The algorithm using keywords to search the discharge summary (algorithm #2 in Tables 1 and 3) achieved a higher sensitivity (65.5%; 95% CI 60.1%−71.0%) compared with the ICD algorithm, but compromised PPV (77.6%; 95% CI 72.4%−82.8%). The combined ICD code and keyword

Table 2. Characteristics of the Study Cohort ($N = 2106$)

Characteristic	Frequency, N (%)
Age >64 years	1023 (48.6)
Female sex	1057(50.2)
HF present	296 (14.1)
Abnormal blood BNP level	331 (15.7)
Inotropics or IV diuretics administered	220 (10.4)
Myocardial infarction	92 (4.4)
Peripheral vascular disease	112 (5.3)
Cerebrovascular disease	295 (14)
Dementia	154 (7.3)
Chronic pulmonary disease	355 (16.9)
Connective tissue/rheumatic disease	119 (5.7)
Peptic ulcer disease	717 (34)
Mild liver disease	182 (8.6)
Diabetes without complications	434 (20.6)
Diabetes with complications	253 (12)
Paraplegia and hemiplegia	45 (2.1)
Renal disease	342 (16.2)
Cancer	278 (13.2)
Moderate or severe liver disease	21 (1)
Metastatic carcinoma	111 (5.3)
AIDS/HIV	9 (0.4)
Charlson comorbidity score (excluding HF)*	
0	535 (25.4)
1	435 (20.7)
≥2	1136 (53.9)
Surgical patient	526 (24.8)
Length of stay (days)	
≤3	771 (36.6)
4−7	716 (34.0)
8−11	246 (11.7)
≥12	385 (18.3)
In-hospital death	48 (2.3)

BNP, brain natriuretic peptide; HF, heart failure; HIV, human immunodeficiency virus; IV, intravenous.

*The Charlson comorbidity score includes 17 chronic conditions with each condition category associating a weight, and the sum of all the weights results in a score (ranging from 0 to 33). A higher score indicates a greater comorbidity burden.

algorithm achieved significant improvement in sensitivity (from 57.4% to 77.0%); however, the PPV decreased to 78.4%. The algorithm using only structured EMR data (ie, BNP results, and intravenous diuretic and inotropic medication use) obtained a higher sensitivity (78%; 95% CI 73.3%−82.8%), but a much lower PPV (54.2%; 95% CI 49.5%−59.0%) than the keyword algorithm and the ICD algorithm. The combined algorithm incorporating the structured data, keyword, and/or ICD code algorithms reached a high sensitivity (82.4%−88.2%), but retained a low PPV (53.5%−54.8%).

Applying XGBoost resulted in a mean sensitivity of 74.3% and a mean PPV of 87.4% across the 5 folds. The 6 features that occurred in the top 10 most important features of all 5 folds were C0018801 (heart failure), C0018802 (congestive heart failure), C0054836 (Carvedilol, a drug could be used for treating HF), C0277785 (functional disorder), C0016860 (furosemide, a drug could be used for treating HF), C0699992 (Lasix, the brand name for furosemide). After fitting a decision tree using these features, the result showed that the only two that separated the cases into cases of HF and cases without HF were C0018801 (heart failure) and C0018802 (congestive heart failure). Therefore, the final ML model (decision tree) was fitted using only these 2 features (algorithm #4 in Tables 1 and 3). On the test set (a 20% random sample of the cohort), this model achieved a sensitivity of 80% (95% CI 69.9%−90.1%), and a PPV of 88.9% (95% CI 80.5%−97.3%). An even higher sensitivity (83.3%; 95% CI 73.9%−92.8%) was reached by combining ICD code and HFC algorithm, while the PPV slightly decreased to 83.3% (95% CI 73.9%−92.8%).

The F1 score also shows that the algorithm with the best balance of sensitivity and PPV was the HFC algorithm with F1 84.2% (95% CI 82.4%−86%), which was significantly better than the ICD algorithm (F1 70.8%; 95% CI 68.6%−73%) and keyword algorithm (F1 71.0%; 95% CI 68.8%−73.2%). The algorithm combining ICD and HFC did not improve the balanced performance on HFC alone based on F1 (83.3% vs 84.2%).

The stratified analyses showed that in the nonsurgical cohort, the HFC algorithm still performed better than ICD code regarding the sensitivity (79.7% vs 60.0%) and obtained a PPV of 88.7%. The sensitivity of the combined ICD and HFC algorithm was approximately the same in the nonsurgical patient cohort compared with that of the entire cohort (83.1% vs 83.3%). According to F1 score, the rank of the algorithms (HFC, combined HFC and ICD, keyword, and ICD) did not change in the stratified analyses (Table 4).

Discussion

This study developed and provided validity support for EMR-based algorithms for defining HF using population-

Table 3. Validity of the EMR-Based Algorithms for Identifying HF

Algorithm	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	F1 score, % (95% CI)
#1 ICD-10	57.4 (51.8−63.0)	99.2 (98.8−99.6)	92.4 (88.6−96.2)	93.4 (92.3−94.5)	70.8 (68.6−73)
#2 Keyword	65.5 (60.1−71)	96.9 (96.1−97.7)	77.6 (72.4−82.8)	94.5 (93.5−95.5)	71 (68.8−73.2)
#3 Combined #1 and #2	77.0 (72.2−81.8)	98.4 (97.1−99.7)	78.4 (73.6−83.1)	96.3 (95.4−97.1)	77.7 (75.7−79.7)
#4 ML	80.0 (69.9−90.1)	97.5 (95.9−99.1)	88.9 (80.5−97.3)	96.8 (95.0−98.6)	84.2 (82.4−86)
#5 Combined #1 and #4	83.3 (73.9−92.8)	97.3 (95.6−98.9)	83.3 (73.9−92.8)	97.3 (95.6−98.9)	83.3 (81.5−85.1)
#6 EMR structured data	78.0 (73.3−82.8)	89.2 (87.8−90.7)	54.2 (49.5−59.0)	96.1 (95.2−97.1)	64 (61.6−66.3)
#7 Combined #1 and #6	82.4 (78.1−86.8)	88.9 (87.5−90.3)	54.8 (50.2−59.5)	96.9 (96−97.7)	65.8 (63.5−68.1)
#8 Combined #1, #2 and #6	88.2 (84.5−91.9)	87.5 (85.9−89)	53.5 (49.1−57.9)	97.8 (97.1−98.6)	66.6 (64.3−68.9)

CI, confidence interval; EMR, electronic medical record; ICD, *International Classification of Diseases*; ML, machine learning.

Table 4. Validity of the Algorithms for Identifying HF Among A Nonsurgical Patient Cohort ($N = 1580$)

Algorithm	Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	NPV, % (95% CI)	F1 score, % (95% CI)
ICD-10	60.0 (54.2−65.8)	99.1 (98.6−99.6)	93.2 (89.5−96.9)	92.1 (90.7−93.5)	73.0 (70.8−75.2)
Keywords	69.5 (64.0−74.9)	96.2 (95.1−97.2)	79.3 (74.1−84.4)	93.7 (92.4−95.0)	74.1 (71.9−76.3)
ML	79.7 (69.4−89.9	97.7 (95.8−99.5)	88.7 (80.1−97.2)	95.5 (92.9−97.0)	84.0 (82.1−85.8)
Combined ML and ICD-10	83.1 (73.5−92.6)	96.1 (93.8−98.5)	83.1 (73.5−92.6)	96.1 (93.8−98.5)	83.1 (81.2−85.0)

AUC, area under the curve; CI, confidence interval; HF, heart failure; ML, machine learning; PPV, positive predictive value; NPV, negative predictive value.

based inpatient EMR data in Canada. The results have demonstrated that an ICD-10−based HF case finding algorithm can be significantly improved using EMR data and by integrating data-driven techniques that leverage free text data such as NLP and ML. The HFC algorithm has maintained a reasonable PPV, while improving sensitivity significantly. This remained the case, even upon stratifying by surgical vs nonsurgical cohorts.

Rationale for Developing EMR-Based Algorithms

A keyword algorithm was implemented using keywords selected from the UMLS metathesaurus. Although this method is the simplest NLP technique to apply, it did not show superior performance to the ICD algorithm in the present study. This may be due to the clinical complexity of a HF diagnosis that requires not only objective examination, but also the physician's clinical judgement. Therefore, the keyword search methodology is insufficient for identifying HF from the EMR discharge summary. The subsequent analysis applied more sophisticated NLP methods. The discharge summary was processed using cTAKES,[18] which incorporated aspects of NLP, such as negation and tagging, to obtain CUIs from the UMLS metathesaurus. Although the "preferred terms" for the final CUIs used in our decision tree are "heart failure" and "congestive heart failure," these CUIs also encompass all synonyms associated with these concepts in UMLS. By extracting these CUIs with cTAKES, we also excluded negated versions of these terms, as well as those not referring to the patient. These steps were automated, which allowed us to process these EMR-based free text documents at a fast rate. These CUIs were used as features for an ML algorithm (XGBoost gradient boosting for feature selection and decision-tree for model output). This methodology ensured the clinical knowledge extracted through NLP with ML methodologies are easily interpretable. This model performed better than the ICD algorithm. Our automated methodology does align with what is expected in the chart review. In the Canadian health system and other jurisdictions around the world, ICD coding through manual chart review is a resource-intensive process and issues such as insufficient time allocation per chart have been noted. Administrative data (mainly ICD coded data) remain the cornerstone of health services planning. Health systems operate based on information obtained through administrative data. Additionally, the health data research paradigm heavily uses administrative data for research activities. Therefore, combining the results of the administrative data algorithm with those obtained from EMR-based methods was deemed the most sensible choice. Thus, a combined algorithm that included DAD and EMR data was developed. The F1 score was calculated for each algorithm as an objective approach to compare overall performance, because it reflects both the sensitivity and PPV in 1 measure. However, depending on the user's application of the algorithm, the F1 score may not reflect the best algorithm for their needs.

How to Choose Different Algorithms for Applications

There are several possible clinical and research applications of the proposed algorithms. First, they can lead to building a more accurate HF cohort for outcome or health service research such as inpatient mortality of patients with HF. Second, the algorithms may assist in decreasing misclassification bias when looking at HF as one of the risk factors of interest in risk adjustment analysis. Third, the algorithms may lead toward assessment of health system performance on HF outcomes by incorporating the algorithm in the reporting system (eg, dashboard) for near real-time performance reporting, such as readmission of HF among different health care facilities. For instance, the health system in Alberta is currently transitioning to a province-wide EPIC EMR system and there is keen interest in developing decision support system using EMR data directly.

The purpose and objective of any particular study should dictate the choice of algorithm. If the purpose of a study is to ensure all identified cases are truly positive, then achieving a high PPV through the ICD algorithm becomes important. For example, if one is interested in investigating patterns of care of patients with HF, then accurate recruitment of the cohort of patients with HF becomes of primary importance. Alternatively, in studies focused on HF adverse event surveillance and reporting, a high sensitivity algorithm (eg, the combined ICD, keyword, and structured EMR algorithm) would be useful. In studies focused on studying disease (eg, HF) incidence or prevalence in a population, an algorithm that balances sensitivity and PPV (eg, the combined ICD and HFC algorithm) would be suitable. Such algorithms perform well in capturing cases of HF while keeping the false-positive rate low.

There is a tradeoff between the performance and the generalizability of the algorithms. Sophisticated methods such as ML typically perform better in classification modelling, but its application can be challenging and may require a lengthy development process. The keyword search algorithm is easy to implement but performs sub-optimally. The clinical complexity associated with HF limits accurate identification through the ICD algorithm. The ICD algorithm is also subject to potential misclassification bias owing to coding quality issues (such as undercoding).[15]

Comparing the Developed Algorithms With Other Published Algorithms

The developed EMR-based algorithms in this study leveraged the narrative free text from the EMR, similar to what have been done in previously published works.[12,19,20] However, some of the published case finding algorithms for HF are not straightforward to apply, require intensive clinical knowledge, or may require accessing data that are not readily available (eg, echocardiogram and cardiac imaging data). All of these requirements are potential impediments to adoption. One of the key objectives of this study was to create a HF case-finding algorithm that is relatively easy to implement. All the required tools used in the developed HFC algorithm are open source and available for free. To implement the HFC algorithm, the first step is to get a working installation of cTAKES and apply it to the local data to generate the features (CUIs) outputs. Then, the data for the XGBoost and CART classifiers can be realized by a straightforward parsing of the cTAKES output to label the individual patients as HF or not. Free text data used in this study were limited to discharge summaries, which is common to all inpatient EMR systems and is therefore widely accessible. Some structured laboratory and medication data were also considered in the present study, but the results were not promising. This finding may be due to the lack of standard diagnostic biomarkers for HF. Further, the medications used for managing HF are not specific to that condition and are also used to treat other diseases. For example, diuretics are sometimes used for the management of hypertension.

Stratification Analyses

Data quality plays a role in the performance of any model. A previous study suggested that data quality (eg, completeness and granularity) varies by hospital department.[21] The chart review confirmed that the surgical cohort had a significant number of discharge summaries missing. This finding suggests that algorithm development needs to consider variations in clinical practices between hospital departments. A stratified analysis was conducted, dividing the study sample into surgical and nonsurgical patients. The results suggested that, although the combined algorithm worked slightly better in the nonsurgical cohort, the main findings had no significant difference between nonsurgical cohort and entire cohort.

Study Limitations and Strengths

This study has several limitations. First, cardiovascular imaging data were not included. The hypothesis is that adding these data elements (eg, left ventricular ejection fraction) will improve the performance of the algorithms, but these data were not accessible. Second, excluding patients without a discharge summary may have introduced potential selection bias. However, the distribution of demographics (ie, age and sex) between patients with and without discharge summaries was similar. Third, inpatient data may be not ideal to identify patients with mild HF who may be managed through primary care and outpatient clinics. Last, even though efforts were made to minimize overfitting (internal validation, generating searching terms without considering local data), it is not clear how well the algorithms will work in different dataset from a different health system owing to lack of external validation.

Although there are limitations, the study has several strengths worth noting. First, the chart review (gold standard) dataset is large. Previous studies had a relatively small chart ($n = 200-400$) review of patients with HF as the gold standard validation.[5,12,13,19] Second, it was systematically shown how to apply a variety of case-finding algorithms leveraging free text data, from keywords derived from UMLS using MetaMap, to combined ML and cTAKES NLP algorithms. UMLS, MetaMap, and cTAKES are all publicly available tools, so the proposed methodology can be adopted widely. Third, the comprehensive approach employed to detect curated synonyms from different ontologies (eg, SNOMED) to the same concept is generalizable for researchers using data from other populations and geographic locations. This work therefore enhances and builds on the previously completed works on this topic.[6,12,13] Finally, the current study is based on population-level inpatient EMR data, which is the first Canadian study of its kind.

Conclusions

The widely used ICD algorithm for ascertaining HF from administrative data can be improved by incorporating discharge summaries from inpatient EMRs through application of ML. With the increasing use of EMRs nationally and internationally, the findings of the current study are timely, and offer a time- and cost-efficient approach to improved HF case identification leveraging EMR data.

Declaration of Competing Interest

None declared.

Supplementary materials

Supplementary material associated with this article can be found in the online version at doi:10.1016/j.cardfail.2020.04.003.

References

1. Savarese G, Lund LH. Global Public Health Burden of Heart Failure. Card Fail Rev 2017;3:7–11.
2. CW Yancy, Jessup M, Bozkurt B, Butler J, Casey DE Jr, Drazner MH, et al. 2013 ACCF/AHA guideline for the

management of heart failure: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines. J Am Coll Cardiol 2013;62: e147–239.

3. Ponikowski P, Voors AA, Anker SD, Bueno H, Cleland JG, Coats AJ, et al. 2016 ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure: The Task Force for the diagnosis and treatment of acute and chronic heart failure of the European Society of Cardiology (ESC). Developed with the special contribution of the Heart Failure Association (HFA) of the ESC. Eur J Heart Fail 2016;18:891–975.
4. Tran DT, Ohinmaa A, Thanh NX, Howlett JG, Ezekowitz JA, McAlister FA, et al. The current and future financial burden of hospital admissions for heart failure in Canada: a cost analysis. CMAJ Open 2016;4:E365–70.
5. Ng K, Steinhubl SR, deFilippi C, Dey S, Stewart WF. Early detection of heart failure using electronic health records: practical implications for time before diagnosis, data diversity, data quantity, and data density. Circ Cardiovasc Qual Outcomes 2016;9:649–58.
6. Cox ZL, Lewis CM, Lai P, Lenihan DJ. Validation of an automated electronic algorithm and "dashboard" to identify and characterize decompensated heart failure admissions across a medical center. Am Heart J 2017;183:40–8.
7. Pike MM, Decker PA, Larson NB, St Sauver JL, Takahashi PY, Roger VL, et al. Improvement in cardiovascular risk prediction with electronic health records. J Cardiovasc Transl Res 2016;9:214–22.
8. Quach S, Blais C, Quan H. Administrative data have high variation in validity for recording heart failure. Can J Cardiol 2010;26:306–12.
9. Kaspar M, Fette G, Guder G, Seidlmayer L, Ertl M, Dietrich G, et al. Underestimated prevalence of heart failure in hospital inpatients: a comparison of ICD codes and discharge letter information. Clin Res Cardiol 2018;107:778–87.
10. Lee DS, Donovan L, Austin PC, Gong Y, Liu PP, Rouleau JL, et al. Comparison of coding of heart failure and comorbidities in administrative and clinical data for use in outcomes research. Med Care 2005;43:182–8.
11. Gioli-Pereira L, Bernardez-Pereira S, Goulart Marcondes-Braga F, Rocha Spina JM, Muniz Miranda da Silva R, Evangelista Ferreira N, et al. Genetic and ElectroNic medIcal records to predict oUtcomeS in Heart Failure patients (GENIUS-HF) - design and rationale. BMC Cardiovasc Disord 2014;14:32.
12. Blecker S, Katz SD, Horwitz LI, Kuperman G, Park H, Gold A, et al. Comparison of approaches for heart failure case identification from electronic health record data. JAMA Cardiol 2016;1:1014–20.
13. Patel YR, Robbins JM, Kurgansky KE, Imran T, Orkaby AR, McLean RR, et al. Development and validation of a heart failure with preserved ejection fraction cohort using electronic medical records. BMC Cardiovasc Disord 2018;18:128.
14. Lee S, Xu Y, D'Souza AG, Martin EA, Doktorchik C, Zhang Z, et al. Unlocking the potential of electronic health records for health research. Int J Popul Data Sci 2020:5.
15. Quan H, Sundararajan V, Halfon P, Fong A, Burnand B, Luthi JC, et al. Coding algorithms for defining comorbidities in ICD-9-CM and ICD-10 administrative data. Med Care 2005;43:1130–9.
16. Aronson AR, Lang FM. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc 2010;17:229–36.
17. Tianqi Chen CG. XGBoost: a scalable tree boosting system. In: Presented at the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA; 2016.
18. Savova GK, Masanz JJ, Ogren PV, Eng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc 2010;17:507–13.
19. Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 2014;83:983–92.
20. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. J Am Med Inform Assoc 2017;24:361–70.
21. Adeleke IT, Adekanye AO, Onawola KA, Okuku AG, Adefemi SA, Erinle SA, et al. Data quality assessment in healthcare: a 365-day chart review of inpatients' health records at a Nigerian tertiary hospital. J Am Med Inform Assoc 2012;19:1039–42.