BMJ Open

Ascertaining Framingham heart failure phenotype from inpatient electronic health record data using natural language processing: a multicentre Atherosclerosis Risk in Communities (ARIC) validation study

Carlton R Moore [ID],[1] Saumya Jain [ID],[2] Stephanie Haas,[3] Harish Yadav,[3] Eric Whitsel,[2] Wayne Rosamand,[2] Gerardo Heiss,[2] Anna M Kucharska-Newton[2]

ABSTRACT

Objectives Using free-text clinical notes and reports from hospitalised patients, determine the performance of natural language processing (NLP) ascertainment of Framingham heart failure (HF) criteria and phenotype.

Study design A retrospective observational study design of patients hospitalised in 2015 from four hospitals participating in the Atherosclerosis Risk in Communities (ARIC) study was used to determine NLP performance in the ascertainment of Framingham HF criteria and phenotype.

Setting Four ARIC study hospitals, each representing an ARIC study region in the USA.

Participants A stratified random sample of hospitalisations identified using a broad range of International Classification of Disease, ninth revision, diagnostic codes indicative of an HF event and occurring during 2015 was drawn for this study. A randomly selected set of 394 hospitalisations was used as the derivation dataset and 406 hospitalisations was used as the validation dataset.

Intervention Use of NLP on free-text clinical notes and reports to ascertain Framingham HF criteria and phenotype.

Primary and secondary outcome measures NLP performance as measured by sensitivity, specificity, positive-predictive value (PPV) and agreement in ascertainment of Framingham HF criteria and phenotype. Manual medical record review by trained ARIC abstractors was used as the reference standard.

Results Overall, performance of NLP ascertainment of Framingham HF phenotype in the validation dataset was good, with 78.8%, 81.7%, 84.4% and 80.0% for sensitivity, specificity, PPV and agreement, respectively.

Conclusions By decreasing the need for manual chart review, our results on the use of NLP to ascertain Framingham HF phenotype from free-text electronic health record data suggest that validated NLP technology holds the potential for significantly improving the feasibility and efficiency of conducting large-scale epidemiologic surveillance of HF prevalence and incidence.

Strengths and limitations of this study

► The article describes the first study to evaluate performance of natural language processing (NLP) using free-text clinical notes and reports stored in electronic health records to ascertain Framingham heart failure phenotype in multiple regionally dispersed hospitals in the USA with different health systems.

► NLP performances (sensitivity, specificity, positive-predictive value and agreement) are assessed with the reference standard being manual extraction of prespecified information by trained and certified abstractors, using a highly standardised protocol, with quality assurance programmes in place that monitored accuracy, completeness and repeatability of the process.

► The NLP programme used open-source software (clinical Text Analysis Knowledge Extraction System and Python).

► A limitation to the study is that it only includes a subset of hospitalised patients at risk for acute decompensated heart failure based on diagnostic codes (International Classification of Disease, ninth revision) and therefore is not representative of the general hospitalised population.

INTRODUCTION

Since the passage of the Health Information Technology for Economic and Clinical Health Act in 2009,[1] the use of electronic health records (EHRs) in hospital settings has become nearly ubiquitous. Although in 2008, approximately 9% of hospitals were using EHRs, by 2020 the adoption of EHR use among hospitals is approaching 100%.[2] This creates unprecedented opportunities for researchers to automate the process

of extracting clinical phenotype from patient medical records through electronic search methods.

Scientific organisations and experts promote leveraging electronic data as beneficial to the future of research, public health surveillance and quality improvement initiatives.[3] The Working Group on Epidemiology and Population Sciences established by the National Heart, Lung and Blood Institute identified e-epidemiology as a strategic priority for research, with recommendations for studies to 'determine the validity, reliability and scalability of electronic tools for data collection'.[4] Clinical phenotypes can be efficiently and accurately extracted from EHRs through the application of algorithms integrating structured data elements such as diagnostic codes, clinical laboratory data and medication lists.[5] Less well-studied is the use of natural language processing (NLP) of free-text clinical notes stored in EHRs for the ascertainment of complex clinical phenotypes and syndromes.

We focused this study on the use of NLP for the ascertainment of heart failure (HF), a leading cause of hospital admissions and mortality among older adults in the USA.[6] HF is a complex clinical syndrome characterised by the heart's inability to supply blood flow sufficient to meet the needs of the body. It is estimated to affect 5.7 million American adults and its prevalence is expected to rise to 8.4 million by 2030.[7] Reflecting the heterogenous nature of HF syndromes, there is no universally accepted diagnostic schema for HF that adequately classifies all patients across this syndrome's pathophysiology, ranging from HF with reduced left ventricular ejection fraction (LVEF) to HF with preserved LVEF (diastolic dysfunction). Signs and symptoms of HF may differ from patient to patient and clinical judgement is typically required to establish a diagnosis of HF for a given patient. The goal of this study is to determine the extent to which accurate EHR-based extraction of Framingham HF criteria phenotypes and HF event classification[8] can be performed in an automated fashion from clinical notes. We sampled inpatient EHR at four geographically dispersed hospitals with disparate healthcare systems for automated processing and used as a benchmark for our performance an established, standardised protocol of record abstraction and classification.[7]

METHODS
Study population
From 2005 through 2014, the Atherosclerosis Risk in Communities (ARIC) study[9] conducted community surveillance of HF hospitalisations, classified according to the Framingham schema,[8] for residents aged 55–84 years in four regions in the USA.[10 11] To produce annual event rates of HF, eligible hospitalisations from a sample of discharges from acute care hospitals located in ARIC study communities were manually abstracted and events classified according to the presence of the Framingham HF classification criteria.[8] A hospitalisation was considered eligible for inclusion based on specific primary or secondary International Classification of Disease, ninth revision, Clinical Modification codes (HF: 428; rheumatic heart disease: 398.91; hypertensive heart disease with congestive heart failure: 402.0, 402.11 or 402.91; hypertensive heart disease and renal failure with HF: 404.01, 404.03, 404.13, 404.91 or 404.93; acute cor-pulmonale: 415.0; chronic pulmonary heart disease, unspecified: 416.9; other primary cardiomyopathies: 425.4; acute oedema of lung, unspecified: 518.4; dyspnoea and respiratory abnormalities: 786.0). Extraction of prespecified information was performed manually by trained and certified abstractors, using a highly standardised protocol, with quality assurance programmes in place that monitored accuracy, completeness and repeatability of the process.[12] A stratified random sample of these hospitalisations occurring during 2015 in four ARIC study hospitals in different ARIC study regions was drawn for the study. A randomly selected set of 394 records was employed as the derivation dataset; the remainder was set aside as the validation set (table 1, N=406). There were no statistically significant differences in patient demographics between the derivation and validation datasets.

Patient and public involvement
HF is a leading cause of hospital admissions and mortality among older adults in the USA.[6] It is estimated to affect 5.7 million American adults and its prevalence is expected to rise to 8.4 million by 2030.[7] Therefore, the study outcomes are likely to be a high priority for patients. However, patients were not directly involved in the study design, conduct or outcomes of the research project.

Study design
The primary goal of this study is to determine the accuracy with which EHR-based NLP algorithms can be used to (1) extract Framingham HF criteria variables (table 2) from free-text clinical notes and (2) ascertain the HF phenotype according to the Framingham schema.[8] As shown in table 2, HF is present if at least two major Framingham criteria are met, or one major and two minor criteria are met. The study also seeks to assess NLP performance reproducibility in ascertainment of Framingham HF phenotype across the four study hospitals.

Data manually extracted by certified ARIC abstractors following a standardised protocol[13] were used as the reference standard to assess the EHR-based performance of NLP. We used the derivation dataset of 394 records from the four study hospitals to develop the NLP algorithms to extract Framingham HF criteria variables. Once the NLP algorithms were optimised, we assessed NLP performance using a separate validation dataset of 406 unique patient records (table 1).

Figure 1 summarises the study design in which analysis of free-text clinical notes stored in EHRs was compared with manually abstracted Framingham HF phenotype criteria variables (reference standard) from hospitalisations occurring in 2015 at four study hospitals enrolled in the ARIC study (table 1). EHR clinical note types

Table 1 Hospital and patient characteristics for the validation dataset (N=406)

	Hospital ID			
	A	B	C	D
Hospital characteristic				
EHR vendor	Epic	Epic	Epic	Allscripts
Region	South-east	South	North	East
Status	Academic	Academic	Academic	Non-academic
Hospital bed size	873	700	385	247
Abstracted records (N)	122	46	117	121
Patient characteristics				
Age, mean (SD)	73.2 (10.6)	70.4 (10.5)	77.4 (9.9)	78.6 (10.3)
Female, %	43.2	39.1	61.5	59.0
Identified as white, %	54.4	8.7	90.6	93.4
No health insurance, %	0.0	0.0	0.9	9.8
Medicaid insurance, %	4.0	10.9	13.7	1.6

EHR, electronic health record.

used for the analysis included emergency department notes, hospital admission notes, discharge summaries and imaging studies, when available. A structured data element (≥ 4.5 kg weight change during the hospitalisation) was also included.

Extracting HF phenotype criteria from clinical notes in EHRs using NLP

We developed an NLP system using the open-source Apache clinical Text Analysis and Knowledge Extraction Tool[14] (clinical Text Analysis Knowledge Extraction System (cTAKES)) and Python[15] programming software. cTAKES is an NLP programme specifically designed to analyse free-text clinical notes. It includes specific modules for clinical concept coding and negation status. Concept coding from the Unified Medical Language System[16] was used to identify HF phenotype criteria (table 2), such as 'paroxysmal nocturnal dyspnoea', and associate them with standardised concept unique identifiers (CUI), such as 'C1956415', that can easily be referenced in a Python programme. The cTAKES programme also assigns a negation status to each concept it identifies in the electronic clinical text. For example, the HF criterion 'no paroxysmal nocturnal dyspnoea' is processed by cTAKES by first assigning the CUI 'C1956415' to the 'paroxysmal nocturnal dyspnoea' concept and then assigning a negation flag to the concept if it identifies predefined negation terms, such as 'no' or 'denies', associated with the concept. Our study required an additional layer of Python code to identify non-standard documentation of HF criteria (eg, the abbreviation 'PND' for 'paroxysmal nocturnal dyspnoea'), as well as augmented negation so that HF signs and symptoms not described as new or

Table 2 Framingham[8] HF phenotype criteria variables

Framingham HF phenotype criteria variables	Criteria
4.5 kg weight change over 5 days during hospitalisation	Major
Jugular venous distension	Major
Hepatojugular reflux	Major
Paroxysmal nocturnal dyspnoea	Major
Orthopnea	Major
Pulmonary basilar rales	Major
S3 gallop	Major
Alveolar/pulmonary oedema on chest X-ray	Major
Cardiomegaly on chest X-ray	Major
Lower extremity oedema	Minor
Hepatomegaly	Minor
Dyspnoea—exertion	Minor
Bilateral pleural effusion	Minor

HF is diagnosed if the following are present: (1) two major criteria or (2) one major and two minor criteria.
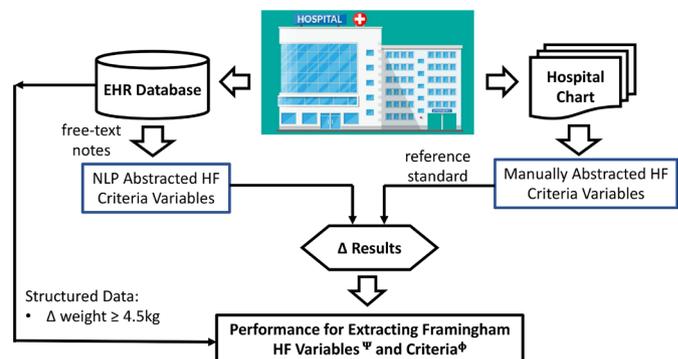HF, heart failure.

Figure 1 Study design. [Ψ]Recall (sensitivity), precision (positive-predictive value); [φ]recall, precision, Δ estimated HF prevalence, % agreement; EHR, electronic health record; HF, heart failure; NLP, natural language processing.
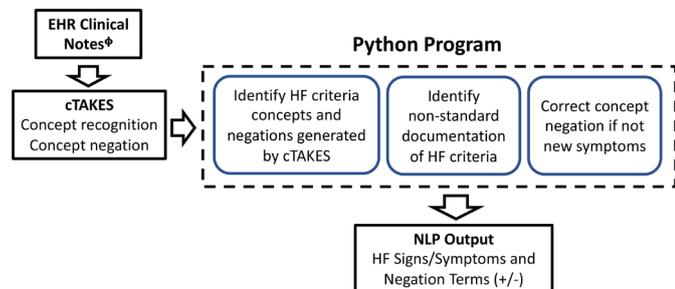
Figure 2 NLP pipeline. ⁺Emergency department notes, hospital admission notes and discharge summaries; cTAKES, clinical Text Analysis Knowledge Extraction System; EHR, electronic health record; HF, heart failure; Negation, determination of whether an HF criteria is negated (eg, patient has oedema vs patient has no oedema); NLP, natural language processing.

worsening were also negated. Figure 2 shows an overview of the NLP pipeline used to extract Framingham HF phenotype criteria from free-text clinical notes stored in study hospital EHRs. For details of the NLP programme (see online supplemental appendix 1).

Data analysis

We computed sensitivity, specificity and positive-predictive value (PPV) as performance metrics to compare EHR-based HF phenotype criteria with the reference standards (manual review by trained ARIC chart abstractors). Using EHR-based NLP Framingham HF phenotype ascertained criteria (table 2), we then calculated the presence or absence of the HF phenotype according to the Framingham[8] HF schema for the study population, and compared results with Framingham HF phenotype

calculated using manually abstracted Framingham HF criteria from the ARIC study (reference standard). χ^2 and Fisher's exact tests on weighted proportions were used to calculate 95% CIs and p values for EHR-based NLP performance characteristics. All analyses were performed using SAS V.9.4 and Stata/SE V.15.0 software.

RESULTS

EHR performance for extraction of Framingham HF phenotype criteria variables

Table 3 shows the performance of EHR-based NLP abstraction of Framingham HF phenotype criteria from free-text clinical notes, compared with manual chart abstraction for the validation data (see online supplemental appendix 2 for results using derivation data). Cardiomegaly and dyspnoea on exertion showed the best performance at PPV 96.7% and 94.5%, respectively. Conversely, hepatojugular reflux and S3 gallop had the lowest PPVs (0.0% and 11.8%, respectively). A major factor in the poor performance was the low frequency of these variables in the patient sample, 0 and 5 occurrences for hepatojugular reflux and S3 gallop, respectively. Pulmonary oedema demonstrated the best sensitivity (91.7%) and hepatomegaly demonstrated the best specificity (99.0%). See online supplemental appendix 3 for performance of NLP in ascertaining Framingham HF phenotype criteria variables for each study hospital.

NLP performance in the ascertainment of the Framingham HF phenotype from EHR data

Overall, performance of EHR-based ascertainment of Framingham[8] HF phenotype in the validation dataset was good,

Table 3	NLP performance for abstracting Framingham HF phenotype criteria from EHRs. Validation dataset (N=406)			
HF criteria variables (n)*	**Sensitivity, % (95% CI)**	**Specificity, % (95% CI)**	**PPV, % (95% CI)**	**Note types used**
Weight loss ≥4.5 kg† (27)	81.5 (61.9 to 93.7)	96.0 (93.6 to 97.8)	59.5 (43.7 to 75.3)	Structured data
Jugular venous distension (56)	60.7 (46.8 to 73.5)	91.7 (87.6 to 94.8)	61.8 (49.0 to 74.6)	ED, AN
Hepatojugular reflux (0)	N/A	99.7 (98.2 to 100.0)	0.00	ED, AN
PND (27)	55.6 (35.3 to 74.5)	89.4 (85.2 to 92.7)	33.3 (19.2 to 46.7)	ED, AN, DC
Orthopnea (64)	59.4 (46.4 to 71.5)	92.7 (88.7 to 95.6)	67.9 (55.7 to 80.1)	ED, AN, DC
Pulmonary basilar rales (93)	61.3 (50.6 to 71.2)	66.4 (59.7 to 72.6)	43.8 (35.3 to 52.3)	ED, AN, DC
S3 gallop (5)	40.0 (5.3 to 85.3)	95.1 (92.0 to 97.2)	11.8 (0.00 to 27.14)	ED, AN, DC
Pulmonary oedema (48)	91.7 (80.0 to 97.7)	51.0 (44.5 to 57.5)	27.3 (20.4 to 34.2)	ED, AN, DC, IR
Cardiomegaly (162)	54.3 (46.3 to 62.2)	96.0 (90.9 to 98.7)	96.7 (93.0 to 100.0)	ED, AN, DC, IR
Lower extremity oedema (163)	74.8 (67.5 to 81.3)	75.5 (67.7 to 82.2)	77.2 (70.7 to 83.7)	ED, AN, DC
Hepatomegaly (3)	33.3 (0.8 to 90.6)	99.0 (97.2 to 99.8)	33.3 (0.00 to 86.2)	ED, AN, IR
Dyspnoea on exertion (263)	79.1 (73.7 to 83.8)	74.5 (59.7 to 86.1)	94.5 (91.5 to 97.5)	ED, AN, DC
Bilateral pleural effusion (79)	75.9 (65.0 to 84.9)	73.1 (66.5 to 79.0)	51.7 (42.6 to 60.8)	ED, AN, DC, IR

*Instances in total cohort that criteria were identified by manual ARIC abstractors (reference standard).
†Weight loss during hospitalisation based on structured daily patient weight data.
AN, admission note; ARIC, Atherosclerosis Risk in Communities; DC, discharge summary; ED, emergency department ; EHRs, electronic health records; HF, heart failure; IR, imaging report; NLP, natural language processing; PND, paroxysmal nocturnal dyspnoea; PPV, positive-predictive value.

Table 4 Performance of NLP-based ascertainment of Framingham HF phenotype

Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95% CI)	Agreement, % (95% CI)
78.8 (72.8 to 83.9)	81.7 (75.2 to 87.0)	84.4 (79.5 to 89.3)	80.0 (75.8 to 83.8)

Note types: emergency department notes, hospital admission notes and discharge summaries.
HF, heart failure; NLP, natural language processing; PPV, positive-predictive value.

with 78.8%, 81.7%, 84.4% and 80.0% as sensitivity, specificity, PPV and agreement metrics, respectively (table 4).

Performance of NLP-based ascertainment of Framingham HF phenotype across study hospitals

Figure 3 shows EHR-based performance in the ascertainment of Framingham HF phenotype for each of the four study hospitals. There was good reproducibility of NLP performance and no meaningful differences in NLP performance across hospitals for the three performance measures of sensitivity, specificity and agreement (all 95% CIs overlap between hospitals for each performance measure).

DISCUSSION

Here, we report on the derivation and validation of an open-source software NLP application that uses EHR data to ascertain HF according to the established Framingham schema in patients hospitalised in dispersed regions of the USA. EHR-based identification of the Framingham HF phenotype had very good performance characteristics (sensitivity: 78.8%, specificity: 81.7%, PPV: 84.4% and agreement: 80.0%) and was reproducible across the four study hospitals.

Several studies have investigated the use of billing codes and lab results to ascertain Framingham HF phenotype in inpatient settings within single healthcare systems.[17 18] To our knowledge, this is the first study to describe the performance of EHR-based NLP tools to ascertain Framingham HF phenotype in inpatients from

Figure 3 EHR-based performance for Framingham HF phenotype by hospital with 95% CIs. EHR, electronic health record; HF, heart failure.

multiple geographically diverse hospitals from different healthcare systems. Our results compare favourably with studies using ICD-9, diagnosis related group (DRG) codes and lab results to ascertain Framingham HF phenotype. Using ICD-9 and DRG codes, Presley et al[17] ascertained Framingham HF phenotype for hospitalised patients in the Veterans Administration (VA) healthcare system.[17] The VA study demonstrated sensitivity of 45.1%, specificity of 99.4% and a PPV of 89.7% for Framingham HF phenotype in population that was homogenous with respect to gender (98.8% male). Using ICD-9 codes, HF medications and lab results, Tison et al[18] ascertained Framingham HF phenotype for inpatients within a single healthcare system in Minnesota. Of the multiple study algorithms used in the study, the one with the highest PPV (86.5%) had a sensitivity of only 41.6%.

Our study adds to the growing body of evidence which suggests that NLP has the potential to improve the cost-effectiveness and timeliness of phenotyping in clinical and epidemiological studies by reducing the need for manual chart abstraction.

In this first step towards the development of a robust protocol for EHR-based NLP surveillance of hospitalised HF patients, we designed a prototype system that had good performance in ascertaining Framingham HF phenotype that was reproducible across four hospitals selected to be geographically dispersed. Underlying this reproducibility, however, was considerable effort required to harmonise a single NLP algorithm that accurately and consistently performed well (figure 3) across the four hospitals.

Evaluation of our results revealed several lessons learnt in the extraction of HF phenotype criteria. First, having complete sets of clinical note and report types from hospitals likely had a significant impact on performance. Our study used NLP to process emergency department notes, admission notes, discharge summaries and imaging study reports. Given the notable lack of standardisation of note type nomenclature across hospitals, we found significant variability between the four study hospitals in nomenclature used to identify specific clinical note types. For example, participating hospitals designated discharge summaries as 'Discharge Summaries', 'Discharge PN', 'PMNDIS' and instances in which the discharging physicians name was concatenated with 'Discharge PN' (eg, Smith Discharge PN). To properly capture phenotypes and clinical outcomes from EHRs requires overcoming a lack of standardised nomenclature, variability in standards for defining and recording data elements, and uncertain collection of longitudinal information or data
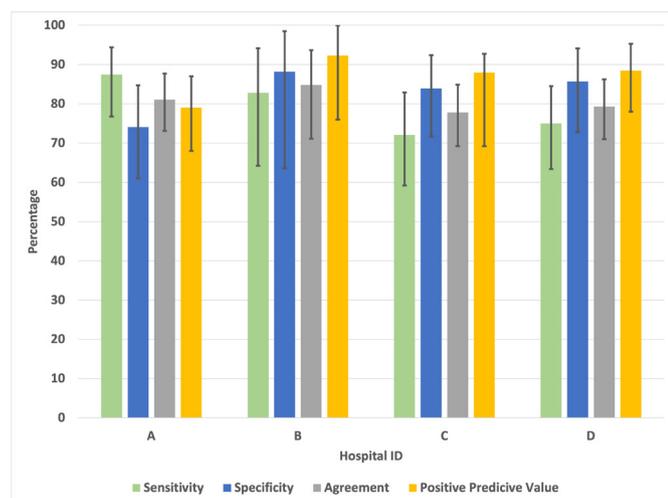
across settings of care. In contrast, these are all features embedded in the standardised community surveillance registry that systematically gathers data entered by many clinicians in numerous hospitals, and served as the benchmark to validate our HF phenotype identification and event classification from EHR. As is typically the case for dedicated registries, ARIC's data element extraction from records is performed by trained abstractors according to specific definitions, standardised procedures, and use of specialised forms leading to highly reliable and valid information under quality control monitoring. Data in EHRs by contrast are captured in the process of patient care by various members of the clinical team, for purposes other than event ascertainment or analysis. Although several efforts exist to establish common data models for EHR data,[19 20] such models are not yet in widespread use and standardised definitions when documenting patient care are uncommon.

The second lesson learnt from the study was the challenge in optimising NLP performance to accurately determine negation for Framingham HF phenotype criteria variables documented in clinical notes. We observed multiple instances in which clinicians documented negative HF signs and symptoms phrased as 'patient denies cough, fever, abdominal pain, chest pain, dyspnoea'. In this example, it was often difficult to accurately assess whether an HF phenotype criteria variable was negated by 'denies'. Similarly, formatting of negation terms often varied by clinician and hospital and included terms such as 'no', 'denies', 'negative', 'neg', '(−)', '−', 'patient does not report'; among other idiosyncratic terminology. Another challenge was establishing negation when clinicians described conditions in discharge summaries under which it was appropriate for patients to take a given medication. For example, 'use albuterol inhaler four times daily as needed for dyspnoea'. In this case, the 'dyspnoea' Framingham HF criterion should be negated because the patient is not currently experiencing dyspnoea, a conditional symptom in which a particular medication should be used.

There are limitations to our study results. The study population represents a sample of hospitalised patients selected for the likelihood of having congestive HF based on ICD-9-CM codes (the prevalence of Framingham HF was 52.0% for NLP and 55.8% for manual chart abstraction). However, this limitation can be mitigated by automated screening of patients using the same ICD-9-CM codes before using NLP ascertainment of Framingham HF criteria. Generalisability of study findings to other populations has not been tested. Furthermore, among the metrics used to ascertain NLP performance, estimated PPV is influenced by the prevalence of the condition. Lastly and not unexpectedly, PPV performed poorly for Framingham HF criteria that occurred infrequently in the patient population. Examples of those were hepatojugular reflux (n=0/406), hepatomegaly (n=3/406), S3 gallop (n=5/406) and PND (n=27/406) had PPVs of 0.0%, 33.3%, 11.8% and 33.3%; respectively (table 3).

Nonetheless, because of their low prevalence in the study population, these criteria likely had a relatively small impact on the determination of the Framingham HF phenotype prevalence.

The means to assess the population burden of HF and the impact of medical interventions and public health policies on these metrics are limited, and largely rely on efforts by professional organisations such as the American Heart Association[21] drawn from various NIH-supported observational studies. Our data suggest that NLP has good performance characteristics in determining Framingham HF phenotype in hospitals from four distinct regions of the country. Such estimates do not substitute for comprehensive population data, nor are they regionally (or nationally) representative, and they do not lend themselves to estimation of population burden metrics or temporal trends. A 2011 report from the Institute of Medicine[22] recommended a national surveillance programme to be put in place funded by the Affordable Care Act,[22] but questions persist about the feasibility of community surveillance that can efficiently incorporate EHR capabilities for accurate estimates of disease burden and to monitor trends in cardiovascular diseases. To accomplish this, such surveillance should be able to link EHR resources to population denominators, harmonise diverse EHRs and implement information extraction tools of known validity and portability, while safeguarding patient privacy and be robust to changes in diagnostic fashion, technologies and coding practices. Such challenges need careful attention to realise the potential of EHR-enabled community surveillance. The alternative—the current inability to monitor population burden and trends—represents a significant impediment to the ability to gauge the impact of heath care and public health initiatives on the burden of, and trends in the most prominent contributors to morbidity, mortality and healthcare expenditures in the USA.

Importantly, the lack of community surveillance programmes encumbers the progress in the understanding of and in reducing health disparities in the incidence of the major cardiovascular health events and their outcomes.[23] Regional epidemiologic surveillance programmes, such as ARIC's, indicate that during the years 2005–2012, annual rates of incident hospitalised HF increased in all race–gender groups, but markedly so for black women. Ongoing HF surveillance efforts are therefore needed to identify vulnerable population subgroups and develop effective prevention strategies.

Future directions for our project include developing a user-friendly interface to adjust NLP algorithms based on institution-specific patterns in documentation of negations, as well as investigating the use of machine-learning technology to optimise performance of the current rule-based NLP system.[24] Specifically, our goal is to approach 100% sensitivity while optimising specificity and PPV.

In conclusion, by decreasing the need for manual chart review, our results on the use of NLP to ascertain Framingham HF phenotype from free-text EHR data suggest

that validated NLP technology holds the potential for significantly improving the feasibility and efficiency of conducting large-scale epidemiologic surveillance of HF prevalence and incidence.

Author affiliations
[1]Medicine, University of North Carolina at Chapel Hill School of Medicine, Chapel Hill, North Carolina, USA
[2]Epidemiology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA
[3]School of Information and Library Science, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina, USA

ORCID iDs
Carlton R Moore http://orcid.org/0000-0001-7732-4829
Saumya Jain http://orcid.org/0000-0002-4083-3918

REFERENCES

1 Bowes WA. Assessing readiness for meeting meaningful use: identifying electronic health record functionality and measuring levels of adoption. *AMIA Annual Symposium proceedings / AMIA Symposium AMIA Symposium*, 2010:66–70.
2 et alHenry J, Pylypchuk Y, Searcy T. Adoption of electronic health record systems among U.S. Non-Federal acute care hospitals, 2016. Available: https://dashboard.healthit.gov/evaluations/data-briefs/non-federal-acute-care-hospital-ehr-adoption-2008-2015.php [Accessed 28 Jan 2020].
3 IOM. *To err is human: building a safer health system*. Washigon, DC: Institute of Medicine, 2000.
4 Boerwinkle E, Crapo JD, Douglas PS. *Strategic transformation of population studies. NHLBI Advisory Council*, 2014.
5 Ritchie MD, Denny JC, Crawford DC, *et al*. Robust replication of genotype-phenotype associations across multiple diseases in an electronic medical record. *Am J Hum Genet* 2010;86:560–72.
6 Virani SS, Alonso A, Benjamin EJ, *et al*. Heart disease and stroke Statistics-2020 update: a report from the American heart association. *Circulation* 2020;141:e139–596.
7 Heidenreich PA, Albert NM, Allen LA, *et al*. Forecasting the impact of heart failure in the United States: a policy statement from the American heart association. *Circ Heart Fail* 2013;6:606–19.
8 Ho KK, Pinsky JL, Kannel WB, *et al*. The epidemiology of heart failure: the Framingham study. *J Am Coll Cardiol* 1993;22:6A–13.
9 The Atherosclerosis risk in communities (ARIC) study: design and objectives. The ARIC Investigators. *Am J Epidemiol* 1989;129:687–702.
10 White AD, Folsom AR, Chambless LE, *et al*. Community surveillance of coronary heart disease in the Atherosclerosis risk in communities (ARIC) study: methods and initial two years' experience. *J Clin Epidemiol* 1996;49:223–33.
11 Rosamond WD, Chang PP, Baggett C, *et al*. Classification of heart failure in the Atherosclerosis risk in communities (ARIC) study: a comparison of diagnostic criteria. *Circ Heart Fail* 2012;5:152–9.
12 Loehr LR, Agarwal SK, Baggett C, *et al*. Classification of acute decompensated heart failure: an automated algorithm compared with a physician reviewer panel: the Atherosclerosis risk in Communities study. *Circ Heart Fail* 2013;6:719–26.
13 ARIC. Surveillance of heart failure manual of operation. ARIC coordinating center, 2011. Available: https://sites.cscc.unc.edu/aric/sites/default/files/public/manuals/Man3a_updated_6172011_0.pdf [Accessed 26 Aug 2020].
14 Savova GK, Masanz JJ, Ogren PV, *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.
15 Python. Python. python software Foundation, 2016. Available: https://www.python.org/doc/ [Accessed 29 Jan 2019].
16 NIH/NLM. Unified medical language system (UMLS). National library of medicine, 2016. Available: https://www.nlm.nih.gov/research/umls/ [Accessed 2 Jun 2016].
17 Presley CA, Min JY, Chipman J, *et al*. Validation of an algorithm to identify heart failure hospitalisations in patients with diabetes within the Veterans health administration. *BMJ Open* 2018;8:e020455.
18 Tison GH, Chamberlain AM, Pletcher MJ, *et al*. Identifying heart failure using EMR-based algorithms. *Int J Med Inform* 2018;120:1–7.
19 Stang PE, Ryan PB, Racoosin JA, *et al*. Advancing the science for active surveillance: rationale and design for the observational medical outcomes partnership. *Ann Intern Med* 2010;153:600–6.
20 ONC, Interoperability. Office of the National coordinator for health information technology (onc), 2019. Available: https://www.healthit.gov/topic/interoperability [Accessed 17 Sep 2020].
21 Writing Group Members, Mozaffarian D, Benjamin EJ, *et al*. Heart disease and stroke Statistics-2016 update: a report from the American heart association. *Circulation* 2016;133:e38–60.
22 IOM. *The future of nursing: leading change, advancing health*. Washington, DC: National Acedemy of Sciences, 2011.
23 Nih morbidity and mortality chart book, 2012. Available: http://www.nhlbi.nih.gov/research/reports/2012-mortality-chart-book.
24 Pradhan S, Elhadad N, South BR, *et al*. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J Am Med Inform Assoc* 2015;22:143–54.

APPENDIX 1
(details of natural language processing)

Natural Language Processing (NLP)

cTAKES NLP Program

The clinical Text Analysis and Knowledge Extraction Tool (cTAKES) is open source natural language processing (NLP) package that is specifically designed to analyze clinical free-text notes and reports; cTAKES uses specific modules for clinical concept coding, as well as negation status of identified clinical concepts. Concept coding based on standardized electronic dictionaries (UMLS Metathesaurus) are used by cTAKES to identify clinical terms of interest, such as 'dyspnea', and assign to them the appropriate concept unique identifier (CUI), such as 'C0013404', that can be identified by a computer algorithm from cTAKES output files. The cTAKES program also assigns a negation status to each concept it identifies in the electronic text. For example, the term 'no dyspnea' is processed by cTAKES by first assigning the concept unique identifier (CUI), 'C0013404', to the 'dyspnea' term and then assigns a negation flag to the concept.

The project required an additional layer of negation so atypical negations not captured by cTAKES were identified. Therefore, we developed a program written in Python that processes the cTAKES output (and raw text) and negates relevant concepts not captured by cTAKES (Figure 1).

Figure 1. ARIC Natural Language Processing Pipeline (NLP)
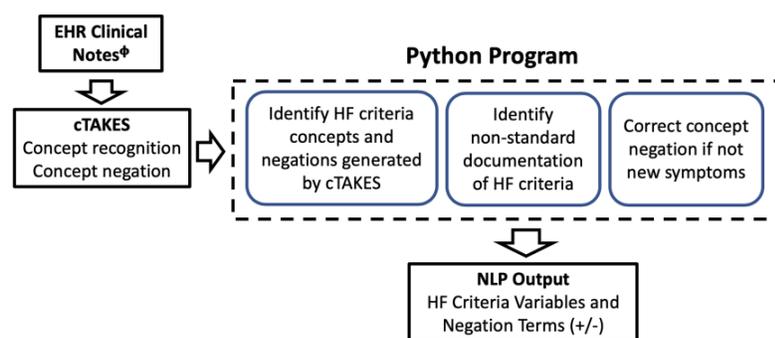
ϕ emergency department notes, hospital admission notes, and discharge summaries
cTAKES: clinical Text Analysis Knowledge Extraction System
Negation: determination of whether a HF criteria is negated (e.g., patient has edema vs. patient has no edema)

For example, the 'dyspnea' concept in the text 'the patient has chronic dyspnea' is negated by the Python program because the symptoms are not new onset or worsening. In this example, the term 'chronic' is recognized by the Python program as negating 'dyspnea'.

Modification of cTAKES Negation of Framingham HF Criteria Variables

Table 1 shows the CUIs and negation terms (added to cTAKES negation) used to identify relevant Framingham criteria HF variables. For example, dyspnea on exertion, three CUIs identify this clinical symptom; 'C0013404', 'C0231835', and 'C1145670'. These CUIs represent 'shortness of breath', 'tachypnea', and 'respiratory distress'; respectively. These terms are all synonymous for 'dyspnea on exertion'. Their respective negation terms are shown in the last column. Again, these are negations that augment cTAKES negation. The program allows the user to specify the number of characters before, and after, the relevant Framingham HF criteria term to search for the relevant negation term.

Table 1. List of relevant CUI codes for Framingham HF Criteria Variables and Negation Terms

Description	CUI	CUI Description	Negation Terms (added to cTAKES negation)
Dyspnea on exertion	C0013404	Dyspnea, shortness of breath	'negative', 'neg', 'denies', 'denied', 'needed', 'stable', 'prn', 'chronic', 'without', 'or', 'if', 'mild', 'history', 'hx', 'allergy'
	C0231835	tachypnea	'neg', 'denies', 'denied'
	C1145670	Respiratory failure	'neg'
Lower extremity edema	C0013604	Edema	'negative', 'neg', 'denies', 'denied', 'history', 'hx', 'h/o', 'chronic', 'stable', 'mild', 'trace', 'if', 'or', 'infiltration', 'related', 'cxr', 'lobe', 'alveolar', 'facial', 'for', 'pulmonary', 'interstitial', 'mass', 'pulm', 'hemorrhage', 'represent', 'mesenteric', 'residual', 'arm', 'vasogenic', 'gallbladder', 'w/o'
	C0239340	Lower extremity edema	'negative', 'neg', 'denies', 'denied', 'trace', 'pulmonary', 'resolved', 'chronic', 'stable', 'needed', 'prn', 'without', 'history', 'hx', 'h/o'
	C0581394	Lower extremity swelling	'negative', 'neg', 'denies', 'denied', 'trace', 'pulmonary'
	C0235439	Swollen ankles	'negative', 'neg', 'denies', 'denied'
	C0151603	Anasarca	'negative', 'neg'
Jugular venous distention (JVD)	C0425687	jugular venous distention	'negative', 'neg'
Hepatojugular reflux	C0239949	hepatojugular reflux	'neg'
Hepatomegaly	C0019209	hepatomegaly	'neg'
Paroxysmal nocturnal dyspnea	C1956415	paroxysmal nocturnal dyspnea	'negative', 'neg', 'denies', 'denied'
	C0344357	nocturnal dyspnea	'negative', 'neg', 'denies', 'denied'
Orthopnea	C0085619	orthopnea	'negative', 'neg', 'denies', 'denied'
Pulmonary basilar rales	C0034642	rales	'negative', 'neg', 'denies', 'denied'
S3 gallop	C0232278	S3 gallop	'neg'
Alveolar/Pulmonary edema on chest x-ray	C0034063	Pulmonary edema	'versus', 'vs', 'neg', 'history', 'mild', 'may', 'could', 'possible'
	C0748120	Pulmonary interstitial edema	'neg'
	C2939069	Alveolar edema	'neg'
	C4313225	Interstitial infiltrates	'neg'
	C2750120	Interstitial thickening	'neg'
	C3670573	Perivascular edema	'neg'
	C0746171	Lower lung interstitial markings increased	'neg'
Cardiomegaly	C0018800	cardiomegaly	'neg'
Bilateral pleural effusion	C0032227	pleural effusion	'resolution', 'possible'

Handling of Non-traditional ways of Documenting Framingham HF Criteria Variables in Clinical Text

Review of the clinical notes and reports revealed several idiosyncratic ways in which clinicians document HF signs and symptoms. In these instances, cTAKES does not recognize the terms as clinical concepts and therefore does not assign CUIs. For example, 'jugular venous distention' is often abbreviated 'JVD' in clinical text. However, 'JVD' is not recognized by cTAKES as a clinical concept and not assigned a CUI. Therefore, we build a module in the Python program that used regular expression to identify non-traditional terms and assign them to the correct CUI.

Detail of this process are as follows:

1. Jugular venous distention (JVD) is sometimes documented as increased jugular venous pressure (JVP). JVP does not have a CUI and therefore regular expressions were used to identify this sign. When the term JVP is identified, the python program assigns it the JVD CUI (C0425687) and then performs negation as follows:

   Negation_terms_left_jvd and negation_terms_right_jvd applied on top of negation_terms_left/right terms to jvd/jvp specific CUI extracted manually i.e. not generated by cTAKES

   - negation_terms_left_jvd = ['no jvp', 'unable to', 'could not', 'normal jvp', 'negative', 'neg', 'no jvd', 'unable to', 'could not', 'normal jvd', 'negative', 'neg']
   - negation_terms_right_jvd = ['jvp not', 'unelevated', 'jvp normal', "jvp doesn't", 'jvp does not', 'jvd not', 'unelevated', 'jvd normal',"jvd doesn't", 'jvd does not']

2. Paroxysmal nocturnal dyspnea is sometimes documented as PND. cTAKES does not recognize PND (no CUI assigned); therefore, regular expressions were used to identify this term. When the term PND is identified, the python program assigns it the paroxysmal nocturnal dyspnea CUI and then performs negation as follows:

   Negation_terms_left_pnd and negation_terms_left_negs applies on top of negation_terms_left/right terms to pnd specific CUI extracted manually i.e. not generated by cTAKES

   - negation_terms_left_pnd = ['no pnd', 'no paroxysmal nocturnal dyspnea', 'denies pnd', 'denies paroxysmal nocturnal dyspnea', 'denied pnd', 'denied paroxysmal nocturnal dyspnea']
   - negation_terms_left_negs = ['neg']

   Additionally, if the program identifies a 'shortness of breath' CUI, it searches for the following terms in close proximity: 'woke', 'overnight', 'sleep', and 'awakened'. If it finds any of these terms, then 'paroxysmal nocturnal dyspnea' is considered to be present.

APPENDIX 2
(Analysis of Derivation Data)

Table 1. Hospital and Patient Characteristics for Derivation Dataset (N = 797)

Hospital Characteristic	Hospital ID			
	A	**B**	**C**	**D**
EHR vendor	Epic	Epic	Epic	Allscripts
Geographic location	South East	South	North	East
Status	Academic	Academic	Academic	Non-academic
Hospital bed size	873	700	385	247
Abstracted records (N)	262	66	199	270
Patient Characteristics	**Hospital ID**			
	A	**B**	**C**	**D**
Age, mean (S.D.)	68.6 (12.4)	67.0 (12.1)	75.3 (12.8)	80.0 (10.9)
Female, %	43.6	50.2	57.3	54.0
White race, %	54.3	10.1	88.9	95.9
No health insurance, %	0.0	5.9	1.0	6.0
Medicaid insurance, %	8.1	13.0	19.2	4.2

EHR: electronic health record

Table 2. Performance for EHR-based Framingham HF Phenotype Criteria (n=394)

HF Criteria Variables (n[†])	Sensitivity, % (95%CI)	Specificity, % (95%CI)	PPV, % (95% CI)
Weight loss ≥ 4.5 kg[††] (38)	60.5 (89.0 – 94.6)	95.5 (92.8 – 97.4)	59.0 (44.6 – 74.4)
Jugular venous distension (55)	60.0 (45.9 – 73.0)	82.5 (77.2 – 87.1)	43.4 (32.3 – 54.5)
Hepatojugular reflux (1)	100.0 (25.0 – 100.0)	99.3 (97.6 – 99.9)	33.3 (0.0 – 86.6)
PND (32)	68.8 (50.0 – 83.9)	88.1 (83.6 – 91.7)	40.7 (27.6 – 53.8)
Orthopnea (55)	56.4 (42.3 – 69.7)	93.5 (89.7 – 96.2)	66.0 (52.4 – 79.5)
Pulmonary basilar rales (72)	65.3 (53.1 – 76.1)	62.9 (56.3 – 69.2)	35.6 (27.4 – 43.8)
S3 gallop (5)	20.0 (5.0 – 71.6)	95.6 (92.6 – 97.6)	7.1 (0.0 – 22.2)
Pulmonary edema (53)	83.0 (70.2 – 91.9)	48.0 (41.4 – 54.7)	27.0 (20.2 – 33.8)
Cardiomegaly (157)	49.7 (41.6 – 57.8)	95.2 (89.8 – 98.2)	92.9 (87.3 – 98.4)
Lower extremity edema (163)	74.8 (67.5 – 81.3)	84.8 (77.7 – 90.3)	85.3 (79.5 – 91.1)
Hepatomegaly (0)	0.0	99.3 (97.6 – 99.9)	0.0 (0.0 – 0.0)
Dyspnea on exertion (256)	78.5 (73.0 – 83.4)	62.2 (46.5 – 76.2)	92.2 (88.6 – 95.8)
Bilateral pleural effusion (81)	67.9 (56.6 – 77.8)	80.6 (74.4 – 85.8)	58.5 (48.6 – 69.5)

[†] instances in total cohort that criteria were identified by manual ARIC abstractors (reference standard)

[††] weight loss during hospitalization based on structured daily patient weight data, PPV: positive-predictive value, CI: confidence interval

Table 3. Performance of EHR-based Ascertainment of Framingham HF Phenotype (n = 394)

Sensitivity, % (95% CI)	Specificity, % (95% CI)	PPV, % (95%CI)	P_{EHR}	P_{Ref}	P_{Ref} - P_{EHR}	% Agreement (95%CI)
75.0 (68.7 - 80.6)	74.1 (67.0 – 80.5)	78.6 (73.1 - 84.1)	0.533	0.558	0.025 (0.01-0.04)	74.6 (70.3 - 78.9)

[†] reference is calculations based on ARIC manually abstracted Framingham criteria,

Note types: emergency department notes, hospital admission notes, and discharge summaries

P_{EHR}: EHR-based ascertainment of HF prevalence, P_{Ref}: prevalence reference standard based on manual chart abstraction

APPENDIX 3
(Performance of Framingham Heart Failure (HF) Criteria by Study Hospital)

Table 1. Hospital A: Performance for EHR-based Framingham HF Phenotype Criteria Variables.

HF Criteria Variable	Number that agree[*] (%, 95% CI)	Sensitivity[+] (%, 95% CI)	Specificity[+] (%, 95% CI)
Weight loss >= 4.5 kg	118 (96.7%, 0.918 - 0.991)	12 (100.0%, 0.735 - 1.000)	110 (96.4%, 0.910 - 0.990)
Lower extremity edema	69 (56.6%, 0.473 - 0.655)	47 (80.9%, 0.667 - 0.909)	39 (74.4%, 0.579 - 0.870)
Jugular venous distension (JVD)	73 (59.8%, 0.506 - 0.686)	18 (44.4%, 0.215 - 0.692)	68 (92.6%, 0.837 - 0.976)
Hepatojugular reflux	85 (69.7%, 0.607 - 0.777)	0	84 (98.8%, 0.935 - 1.000)
Hepatomegaly	86 (70.5%, 0.616 - 0.784)	0	86 (97.7%, 0.919 - 0.997)
Dyspnea (Climbing or exertion)	75 (61.5%, 0.522 - 0.701)	73 (86.3%, 0.762 - 0.932)	13 (76.9%, 0.462 - 0.950)
Paroxysmal nocturnal dyspnea	76 (62.3%, 0.531 - 0.709)	10 (70.0%, 0.348 - 0.933)	76 (88.2%, 0.787 - 0.944)
Orthopnea	78 (63.9%, 0.547 - 0.724)	15 (66.7%, 0.384 - 0.882)	71 (93.0%, 0.843 - 0.977)
Pulmonary basilar rales	53 (43.4%, 0.345 - 0.527)	19 (36.8%, 0.163 - 0.616)	67 (65.7%, 0.531 - 0.768)
S3 (gallop)	87 (71.3%, 0.624 - 0.791)	0	85 (100.0%, 0.958 - 1.000)
Alveolar/pulmonary edema	53 (43.4%, 0.345 - 0.527)	20 (100.0%, 0.832 - 1.000)	62 (48.4%, 0.355 - 0.614)
Cardiomegaly	71 (58.2%, 0.489 - 0.671)	46 (73.9%, 0.589 - 0.857)	36 (94.4%, 0.813 - 0.993)
Bilateral pleural effusion	70 (57.4%, 0.481 - 0.663)	28 (92.9%, 0.765 - 0.991)	54 (75.9%, 0.624 - 0.865)
Framingham Criteria	99 (81.1%, 0.731 - 0.877)	64 (87.5%, 0.768 - 0.944)	58 (74.1%, 0.610 - 0.847)

Table 2. Hospital B: Performance for EHR-based Framingham HF Phenotype Criteria Variables.

Variable Description	Number that agree[*] (%, 95% CI)	Sensitivity[+] (%, 95% CI)	Specificity[+] (%, 95% CI)
Weight loss >= 4.5 kg	44 (95.7%, 0.852 - 0.995)	6 (66.7%, 0.223 - 0.957)	40 (100.0%, 0.912 - 1.000)
Lower extremity edema	33 (71.7%, 0.565 - 0.840)	22 (86.4%, 0.651 - 0.971)	19 (73.7%, 0.488 - 0.909)
Jugular venous distension (JVD)	35 (76.1%, 0.612 - 0.874)	16 (75.0%, 0.476 - 0.927)	25 (92.0%, 0.740 - 0.990)
Hepatojugular reflux	41 (89.1%, 0.764 - 0.964)	0	41 (100.0%, 0.914 - 1.000)
Hepatomegaly	41 (89.1%, 0.764 - 0.964)	1 (100.0%, 0.025 - 1.000)	40 (100.0%, 0.912 - 1.000)
Dyspnea (Climbing or exertion)	30 (65.2%, 0.498 - 0.786)	37 (70.3%, 0.530 - 0.841)	4 (100.0%, 0.398 - 1.000)
Paroxysmal nocturnal dyspnea	32 (69.6%, 0.542 - 0.823)	4 (50.0%, 0.068 - 0.932)	37 (81.1%, 0.648 - 0.920)
Orthopnea	35 (76.1%, 0.612 - 0.874)	14 (71.4%, 0.419 - 0.916)	27 (92.6%, 0.757 - 0.991)
Pulmonary basilar rales	25 (54.3%, 0.390 - 0.691)	13 (53.8%, 0.251 - 0.808)	28 (64.3%, 0.441 - 0.814)
S3 (gallop)	34 (73.9%, 0.589 - 0.857)	3 (66.7%, 0.094 - 0.992)	38 (84.2%, 0.687 - 0.940)
Alveolar/pulmonary edema	19 (41.3%, 0.270 - 0.568)	5 (100.0%, 0.478 - 1.000)	32 (43.8%, 0.264 - 0.623)
Cardiomegaly	30 (65.2%, 0.498 - 0.786)	15 (66.7%, 0.384 - 0.882)	22 (90.9%, 0.708 - 0.989)
Bilateral pleural effusion	22 (47.8%, 0.329 - 0.631)	4 (75.0%, 0.194 - 0.994)	33 (57.6%, 0.392 - 0.745)
FRAMINGHAM	39 (84.8%, 0.711 - 0.937)	29 (82.8%, 0.642 - 0.942)	17 (88.2%, 0.636 - 0.985)

Supplemental material

BMJ Publishing Group Limited (BMJ) disclaims all liability and responsibility arising from any reliance
placed on this supplemental material which has been supplied by the author(s)

BMJ Open

Table 3. Hospital C: Performance for EHR-based Framingham HF Phenotype Criteria Variables.

Variable Description	Number that agree* (%, 95% CI)	Sensitivity+ (%, 95% CI)	Specificity+ (%, 95% CI)
Weight loss >= 4.5 kg	111 (94.9%, 0.892 - 0.981)	3 (66.7%, 0.094 - 0.992)	114 (95.6%, 0.901 - 0.986)
Lower extremity edema	76 (65.0%, 0.556 - 0.735)	45 (73.3%, 0.581 - 0.854)	46 (87.0%, 0.737 - 0.951)
Jugular venous distension (JVD)	86 (73.5%, 0.645 - 0.812)	10 (50.0%, 0.187 - 0.813)	81 (96.3%, 0.896 - 0.992)
Hepatojugular reflux	93 (79.5%, 0.710 - 0.864)	0	90 (100.0%, 0.960 - 1.000)
Hepatomegaly	91 (77.8%, 0.692 - 0.849)	0	89 (98.9%, 0.939 - 1.000)
Dyspnea (Climbing or exertion)	77 (65.8%, 0.565 - 0.743)	78 (79.5%, 0.688 - 0.878)	13 (92.3%, 0.640 - 0.998)
Paroxysmal nocturnal dyspnea	84 (71.8%, 0.627 - 0.797)	8 (50.0%, 0.157 - 0.843)	83 (92.8%, 0.849 - 0.973)
Orthopnea	77 (65.8%, 0.565 - 0.743)	19 (47.4%, 0.244 - 0.711)	72 (90.3%, 0.810 - 0.960)
Pulmonary basilar rales	70 (59.8%, 0.504 - 0.688)	32 (68.8%, 0.500 - 0.839)	59 (76.3%, 0.634 - 0.864)
S3 (gallop)	91 (77.8%, 0.692 - 0.849)	0	90 (97.8%, 0.922 - 0.997)
Alveolar/pulmonary edema	52 (44.4%, 0.353 - 0.539)	12 (75.0%, 0.428 - 0.945)	68 (55.9%, 0.433 - 0.679)
Cardiomegaly	62 (53.0%, 0.435 - 0.623)	52 (55.8%, 0.413 - 0.695)	28 (100.0%, 0.877 - 1.000)
Bilateral pleural effusion	66 (56.4%, 0.469 - 0.656)	21 (66.7%, 0.430 - 0.854)	59 (79.7%, 0.672 - 0.890)
FRAMINGHAM	91 (77.8%, 0.692 - 0.849)	61 (72.1%, 0.592 - 0.829)	56 (83.9%, 0.717 - 0.924)

Table 4. Hospital D: Performance for EHR-based Framingham HF Phenotype Criteria Variables.

Variable Description	Number that agree* (%, 95% CI)	Sensitivity+ (%, 95% CI)	Specificity+ (%, 95% CI)
Weight loss >= 4.5 kg	113 (93.4%, 0.874 - 0.971)	6 (66.7%, 0.223 - 0.957)	115 (94.8%, 0.890 - 0.981)
Lower extremity edema	63 (52.1%, 0.428 - 0.612)	49 (65.3%, 0.504 - 0.783)	43 (65.1%, 0.491 - 0.790)
Jugular venous distension (JVD)	81 (66.9%, 0.578 - 0.752)	12 (75.0%, 0.428 - 0.945)	80 (86.3%, 0.767 - 0.929)
Hepatojugular reflux	95 (78.5%, 0.701 - 0.855)	0	92 (100.0%, 0.961 - 1.000)
Hepatomegaly	95 (78.5%, 0.701 - 0.855)	0	92 (100.0%, 0.961 - 1.000)
Dyspnea (Climbing or exertion)	69 (57.0%, 0.477 - 0.660)	75 (76.0%, 0.647 - 0.851)	17 (52.9%, 0.278 - 0.770)
Paroxysmal nocturnal dyspnea	84 (69.4%, 0.604 - 0.775)	5 (40.0%, 0.053 - 0.853)	87 (90.8%, 0.827 - 0.959)
Orthopnea	84 (69.4%, 0.604 - 0.775)	16 (56.3%, 0.299 - 0.802)	76 (94.7%, 0.871 - 0.985)
Pulmonary basilar rales	61 (50.4%, 0.412 - 0.596)	29 (72.4%, 0.528 - 0.873)	63 (58.7%, 0.456 - 0.710)
S3 (gallop)	88 (72.7%, 0.639 - 0.804)	0	92 (92.4%, 0.849 - 0.969)
Alveolar/pulmonary edema	54 (44.6%, 0.356 - 0.539)	11 (90.9%, 0.587 - 0.998)	77 (51.9%, 0.403 - 0.635)
Cardiomegaly	57 (47.1%, 0.380 - 0.564)	49 (30.6%, 0.183 - 0.454)	39 (97.4%, 0.865 - 0.999)
Bilateral pleural effusion	66 (54.5%, 0.452 - 0.636)	26 (65.4%, 0.443 - 0.828)	62 (72.6%, 0.598 - 0.831)
FRAMINGHAM	96 (79.3%, 0.710 - 0.862)	72 (75.0%, 0.634 - 0.845)	49 (85.7%, 0.728 - 0.941)