Journal of the American Medical Informatics Association, 24(e1), 2017, e40–e46
doi: 10.1093/jamia/ocw097
Advance Access Publication Date: 12 July 2016
Research and Applications

Research and Applications

Congestive heart failure information extraction framework for automated treatment performance measures assessment

Stéphane M Meystre,[1,2] Youngjun Kim,[2,3] Glenn T Gobbel,[4] Michael E Matheny,[4] Andrew Redd,[2] Bruce E Bray,[1] and Jennifer H Garvin[1,2]

[1]Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA, [2]Salt Lake City VA Healthcare System, Salt Lake City, UT, USA, [3]School of Computing, University of Utah, Salt Lake City, UT, USA and [4]Department of Biomedical Informatics, Vanderbilt University, Nashville, TN, USA

Corresponding Author: Stephane M Meystre, MD, PhD., Department of Biomedical Informatics, University of Utah, 421 Wakara Way, Suite 140 Salt Lake City, UT 84108, USA. E-mail: stephane.meystre@hsc.utah.edu; Tel.: 801-581-4080; Fax: 801-581-4297

Received 16 December 2015; Revised 11 May 2016; Accepted 24 May 2016

ABSTRACT

Objective: This paper describes a new congestive heart failure (CHF) treatment performance measure information extraction system – CHIEF – developed as part of the Automated Data Acquisition for Heart Failure project, a Veterans Health Administration project aiming at improving the detection of patients not receiving recommended care for CHF.

Design: CHIEF is based on the Apache Unstructured Information Management Architecture framework, and uses a combination of rules, dictionaries, and machine learning methods to extract left ventricular function mentions and values, CHF medications, and documented reasons for a patient not receiving these medications.

Measurements: The training and evaluation of CHIEF were based on subsets of a reference standard of various clinical notes from 1083 Veterans Health Administration patients. Domain experts manually annotated these notes to create our reference standard. Metrics used included recall, precision, and the F_1-measure.

Results: In general, CHIEF extracted CHF medications with high recall (>0.990) and good precision (0.960–0.978). Mentions of Left Ventricular Ejection Fraction were also extracted with high recall (0.978–0.986) and precision (0.986–0.994), and quantitative values of Left Ventricular Ejection Fraction were found with 0.910–0.945 recall and with high precision (0.939–0.976). Reasons for not prescribing CHF medications were more difficult to extract, only reaching fair accuracy with about 0.310–0.400 recall and 0.250–0.320 precision.

Conclusion: This study demonstrated that applying natural language processing to unlock the rich and detailed clinical information found in clinical narrative text notes makes fast and scalable quality improvement approaches possible, eventually improving management and outpatient treatment of patients suffering from CHF.

Key words: Natural Language Processing (NLP) [MeSH L01.224.050.375.580], Heart Failure [C14.280.434], Left Ventricular Ejection Fraction, Medical Informatics [L01.313.500], Quality Indicators, Health Care [N04.761.789]

INTRODUCTION

Congestive Heart Failure (CHF) is a common condition causing substantial morbidity and mortality, especially in the older population. In 2010, CHF was the most common cause of hospitalization for patients aged 65 or more,[1] and the most frequent discharge diagnosis among Veterans Health Administration (VHA) patients.[2] Unlike almost all cardiovascular disorders, the prevalence of CHF is increasing, from about 6.6 million adults in the USA in 2010, to a forecasted 9.3 million in 2030.[3] CHF is also a costly condition, with total costs forecasted to rise from $31 billion in 2012, to $91 billion in 2030.[3] The evolution of this chronic disease often is comprised of acute exacerbations that accounted for 55% of potentially preventable hospitalizations in a 2003 study.[4] The frequency and severity of these exacerbations could be reduced with adequate treatment and outpatient management.

Recommendations for CHF treatment have been published by the American College of Cardiology Foundation and American Heart Association Task Force on Practice Guidelines.[5] They include dietary and physical activity therapies and invasive therapies, but pharmacologic therapies are the most common. Among pharmacologic therapies, angiotensin converting enzyme inhibitors (ACEIs) are a mainstay of treatment in patients who can tolerate them; for patients who cannot take these drugs, angiotensin receptor blockers (ARB) agents offer an alternative. Evaluation of the adherence to these recommendations can be assessed with treatment performance measures, such as the Heart Failure Performance Measurement Set published by the American College of Cardiology Foundation, American Heart Association, and Physician Consortium for Performance Improvement®.[6] The information needed to calculate these measures is found in patient electronic health records (EHRs), but this information is often recorded in narrative text notes. Extraction of this information from text notes is an expensive and time-consuming effort when performed by trained chart abstractors. Automated approaches based on Natural Language Processing (NLP) have allowed for more efficient and scalable extraction of information from clinical notes.

NLP has been used to extract various types of clinical information from diverse sources of narrative text.[7] In the domain of CHF, most applications of NLP focused on secondary use of clinical information for research purposes. Friedlin and colleagues[8] developed a NLP application to extract imaging observations from chest radiology reports. Pakhomov and colleagues[9] evaluated NLP and predictive modeling to identify patients with CHF using clinical notes. Byrd et al.[10] and Vijayakrishnan and colleagues[11] both developed NLP applications to extract CHF signs and symptoms from clinical notes. The automatic extraction of medication information was the main task of the 2009 i2b2 NLP challenge.[12] It focused on identifying medications and attributes (i.e., dosage, frequency, treatment duration, mode of administration, and reason for the administration of the medication) in clinical notes. Almost twenty teams participated in this challenge, and Meystre and colleagues[13] built a system called Textractor that reached a performance of about 0.72 recall and 0.83 precision. Patrick and Li[14] trained a sequence-tagging model using conditional random fields with various lexical, morphological, and gazetteer features. Their tagger reached about 0.86 recall and 0.91 precision (ranked first in the challenge).[14]

This study was undertaken as part of the Automated Data Acquisition for Heart Failure project, a VHA project aimed at improving the detection of patients not receiving recommended care for CHF. To this end, we developed and evaluated an automated treatment performance measure extraction and classification system: Congestive Heart failure Information Extraction Framework (CHIEF). This system uses NLP to automatically extract left ventricular function assessment information,[15–17] medications (ACEIs and ARBs), and reasons the medications may not be prescribed (e.g., contra-indications). Adding the temporal and contextual analysis (e.g., negation) of this extracted information, a patient-level classification completes the process, classifying patients as either meeting the treatment performance measure, or not. Healthcare providers can then be alerted about the latter, to improve treatment and follow-up of their patients suffering from CHF.

MATERIALS AND METHODS

Study setting and patient population

This study was based on a cohort of 1083 inpatients diagnosed with CHF who were discharged from eight VHA medical centers in 2008–2009. Clinical notes of select types were extracted from the EHR of our study cohort from the VHA Corporate Data Warehouse[18] and stored within the VA Informatics and Computing Infrastructure,[19] which facilitates research and analysis of data in a secure environment within the VHA.

Reference standard development

To develop our reference standard for training and testing, each clinical note was manually annotated by domain experts. They referred to pilot tested and iteratively developed annotation guidelines describing the information to annotate in details. This included mentions and values of the left ventricular ejection fraction (LVEF), mentions of ACEI or ARB medications, and documented reasons for why the patient was not treated with the aforementioned medications (e.g., contra-indications, allergy) Reasons Not treated with Medication (RNM). These reasons medications may not be prescribed were only classified as such when explicitly written in the clinical record. For example, in "Patient was not prescribed an ACEI or ARB due to coughing," the phrase "due to coughing" would be classified as a RNM. In contrast, the sentence, "The patient coughs when on ACEI or ARB medications" does not contain a RNM phrase even though coughing is a contraindication for ACEIs and ARBs. In addition, relations between mentions of LVEF and the annotated quantitative or qualitative values were also annotated, as well as the prescription status of the medications (active, inactive/discontinued, negative) and their relations with possible reasons not to take them. All this annotated information was then summarized at the clinical note and patient-level to eventually automatically classify patients as meeting the CHF treatment performance measure, or not. Details of this process and its patient-level evaluation are available in another publication (Garvin et al., Submitted for publication)

The clinical notes for our study cohort were organized into batches (one per patient). Of these batches, 314 were used for training and 769 for testing. Because of the low prevalence of RNM mentions in the training corpus, annotation of RNM for training was confined to some document types (e.g., history and physical, progress notes, discharge summaries). This RNM focused dataset corresponded to 404 documents in 171 batches within the general training corpus. Annotators used the Knowtator Protégé plug-in to annotate our clinical notes.[20] To facilitate the annotation process, two categories of information were pre-annotated: mentions and quantitative values of LVEF, and medications. The former were

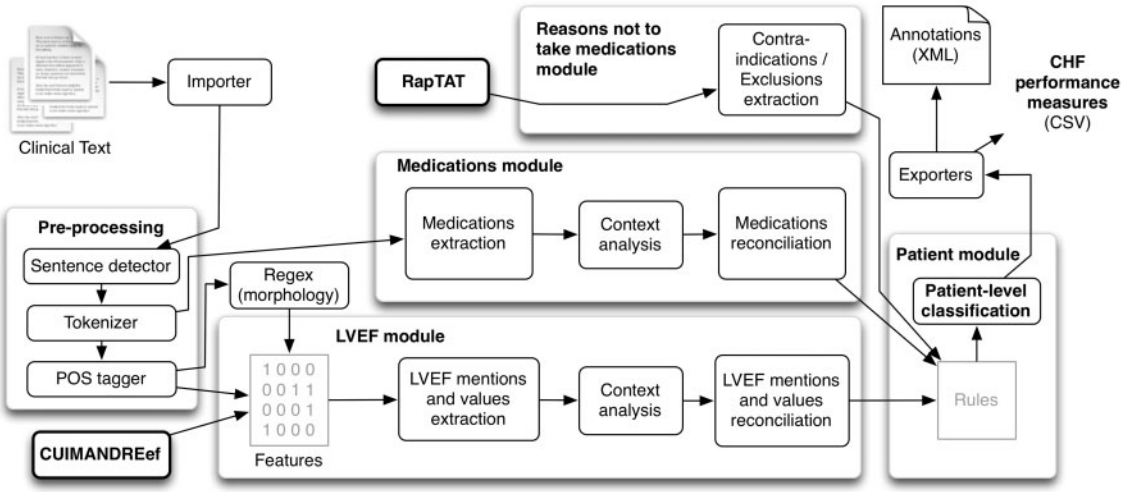

Figure 1. CHIEF general architecture.

pre-annotated with CUIMANDREef,[21] a rule-based application using regular expressions, and the latter with eHOST (Extensible Human Oracle Suite of Tools[22]), a text annotation tool that includes dictionary-based pre-annotation features. Two reviewers independently annotated each clinical note, and disagreements were then adjudicated by a domain expert (cardiologist).

Evaluation metrics

Metrics used to evaluate performance of the CHIEF system were based on counts of each annotation as true positive (system output matches the reference standard), false positive (system output without match in the reference standard), and false negative (reference standard annotation not found in the system output). Comparisons were done as partial matches (any overlap between the reference standard and the system output with the same information category) and exact matches. We then computed recall (i.e., sensitivity), precision (i.e., positive predictive value), and the F_1-measure, a harmonic mean of recall and precision (giving equal weight to each).[23] Each metric was micro-averaged across each mention in clinical notes (i.e., calculated from a confusion matrix combining all mentions in the corpus).

CHIEF system general architecture

CHIEF is based on the Apache Unstructured Information Management Architecture (UIMA) framework for robustness and scalability.[24] As depicted in Figure 1, it includes modules for clinical text pre-processing (detecting sentences and tokens, and syntactic analysis), for extracting mentions of LVEF as well as quantitative and qualitative values, and for extracting mentions of medications (ACEI and ARBs). Explicit documentation of reasons patients were not treated with those medications are extracted with Rapid Text Annotation Tool (RapTAT),[25] a separate NLP application, and integrated in CHIEF. Finally, all extracted information is compared and combined at the clinical note and at the patient level to assess CHF treatment performance measures. All four components were needed to assess the performance measures automatically. Information from each component was used sequentially, applying rules for patient level classification.

For text pre-processing, segmentation of the text in sentences and syntactic analysis (part-of-speech tagging) are both based on OpenNLP modules,[26] with trained models from cTAKES.[27] OpenNLP modules are both based on maximum entropy machine learning algorithms.[28] We used a simple rule-based approach for text tokenization, adapted from cTAKES. A separate NLP application, CUIMANDREef,[21] is also used. It is based on a set of regular expressions targeting mentions and values of LVEF and we used it for text pre-processing, providing the LVEF classifier with some features. LVEF, medications, and reasons not to take medications extraction are detailed in separate sections below.

For the patient-level classification, we developed a set of rules that implement our annotation guideline instructions. These rules start at the clinical note level, allowing selecting the clinically relevant information. For LVEF mentions and values, priority was given to information extracted from echocardiogram reports. Since note types are not consistently recorded in VHA systems, we developed an automatic note type classifier that reached about 0.92 accuracy on average.[29] The most current mentions and values from these notes were selected. For medications, the best information whether the patient is treated with an ACEI or ARB at time of discharge is selected. Possible reasons why the patient is not treated with an ACEI or ARB at time of discharge are also selected. With all information combined at the document level, the following questions are then assessed at the patient level: Has LVEF been measured? Is LVEF less than 40%? Is the patient treated with an ACEI or ARB? If no, did the provider explicitly state a reason for the patient not to receive these medications? Finally, patients with LVEF measured below 40% and treated with an ACEI or ARB, or with a reason not to take these medications, are classified as meeting the CHF treatment quality measure (Garvin et al., Submitted for publication)

Left ventricular function information extraction

This module focuses on the extraction of mentions of LVEF (e.g., "Estimate of LVEF," "EF") as well as related quantitative values (e.g., "~0.60–0.65", "45%").[15] Two versions of this module were initially developed. The first was based on CUIMANDREef,[21] and the second was based on machine learning named entity recognition. The latter uses Miralium,[30] a Java implementation of the Margin Infused Relaxed Algorithm,[31] with morphological, lexical, syntactic (i.e., part-of-speech tags), and semantic features (i.e., output of CUIMANDREef) in a window of four preceding and following tokens.

Table 1. ACEI and ARB medications extracted

Medication name	Med. class	Examples
Angiotensin-converting enzyme Inhibitor(s)	ACEI	ACE Inhibitors, ACEi, acei
Ramipril	ACEI	Altace, ramipril
Enalapril (± hydrochlorothiazide, diltiazem, felodipine)	ACEI	Vaseretic, Vasotec, enalapril, Enalaprilat, Teczem, Lexxel
Fosinopril	ACEI	Fosinopril, Fos, Monopril
Lisinopril (± hydrochlorothiazide)	ACEI	Zestoretic, Prinzide, Zestril, lisinopril, LIS, Prinivil
Perindopril	ACEI	Aceon, perindopril
Trandolapril	ACEI	Mavik, trandolapril
Benazepril (± hydrochlorothiazide or amlodipine)	ACEI	Lotensin, benazepril, Lotensin HCT, Lotrel
Captopril (± hydrochlorothiazide)	ACEI	Capozide, captopril, cap, Capoten
Moexipril (± hydrochlorothiazide)	ACEI	Univasc, moexipril, Uniretic
Quinapril	ACEI	Accupril, quinapril
Tradolapril (+verapamil)	ACEI	Tarka
Angiotensin II receptor blocker(s)	ARB	ARBs, arb, sartans
Candesartan	ARB	Atacand, candesartan
Eprosartan	ARB	Teveten, eprosartan
Irbesartan	ARB	Avapro, irbesartan
Telmisartan	ARB	Micardis, telmisartan
Valsartan (± hydrochlorothiazide)	ARB	Diovan, valsartan, Diovan HCT
Losartan (± hydrochlorothiazide)	ARB	Cozaar, losartan, Nyzaar
Olmesartan	ARB	Benicar, olmesartan, OLM

The window size was inspired by earlier research indicating that a size of 3–4 tokens or more was optimal.[32,33]

Morphological features included word "shape" (e.g., "EF" normalized to "AA"), prefixes, suffixes, and orthographic features (e.g., alphanumeric characters, punctuation). Lexical features included the words themselves and bi-grams of words.[17] In an earlier study, both versions were compared using an existing corpus of 765 echocardiogram reports from seven VA medical centers.[21] CUIMANDREef reached an overall F_1-measure of 0.891, and the Margin Infused Relaxed Algorithm-based version reached an F_1-measure of 0.95.[15] The latter was therefore selected for use in CHIEF.

CHF medications extraction

As already explained, 2 classes of medications were targeted for extraction: ACEI and ARBs. In each class, a comprehensive selection of medications and mentions of the classes in general were targeted (Table 1). A first version of the module was based on a dictionary of medication terms (generic and brand names, abbreviations, and class names). This dictionary was manually built with terms from RxNorm[34] and from clinical experts' experience with clinical text. To also extract misspelled medication names (e.g., we found 21 different misspellings of "Lisinopril" in our corpus), we used fuzzy text string matching based on the edit distance (or Levenshtein distance[35]) and re-assembled medication names that had been split by a newline character (i.e., combining 2 medication name annotations separated by only a newline character into one unique annotation, such as "lot" \n "ensin" into "lotensin"). A second version of the module implemented a token-based classifier based on LIBLINEAR[36] with a linear support vector machine (SVM) classifier to train our token-based model. This second version only allowed for minimal accuracy improvements[37] and was therefore not used in the final version of CHIEF.

To classify the extracted medications as Active, Inactive/Discontinued, or Negative, we used a random subset of our corpus (3000 notes from our training corpus) to train and test another linear SVM classifier we built based on earlier work.[38] We used lexical features (medication name, 5 words preceding it, and 2 words following it) to make the classifier learn the cue words in the context windows surrounding the medication automatically.

Reasons medications are not prescribed extraction

Documentation of the reasons the patient was not taking ACEI or ARB medications was extracted by another application – RapTAT – and then integrated in CHIEF. The RapTAT system was trained to identify RNM mentions based on the annotated training subset described above. Due to the low prevalence of mentions even in the RNM focused data set, training was supplemented by generating a synthetic document containing annotated mentions of RNM from within a dictionary of RNM phrases. The dictionary consisted of multiple synonymous rephrasings of statements of non-compliance within the training documents as well as known contraindications to ACEI and ARB medications. Synonymous phrases for contraindications were generated based on terms within the 2013 Unified Medical Language System® Metathesaurus®.[39] The initial system version often incorrectly identified symptoms that could be contraindications to prescribing ACEI and ARB medications but were not explicitly stated as such. We therefore added a rule to the system to reduce the number of incorrectly identified RNM mentions and improve system precision. The rule was that putative RNM mentions were only marked as true mentions when they occurred in a sentence containing at least one mention of an ACEI and ARB medication. Leave-one-out cross-validation was used to determine the optimal settings for RapTAT to achieve maximal performance in terms of F_1-measure. Documents were grouped by patient for cross-validation, so all documents for a single patient constituted a single training unit.

RESULTS

Study population and reference standard development

Clinical notes (45 703) were retrieved from the EHR of the 1083 VHA patients in our cohort. The clinical notes were grouped in folders, one folder per patient discharge and in chronological order. Patients in the cohort were then randomly assigned to the training set (314 patients) or testing set (769 patients). In the testing set, all patient records were annotated at the document and patient level to

Table 2. CHIEF information extraction results with exact matches (with 0.95 binomial exact confidence intervals)

Information extracted	N	Recall	Precision	F_1-measure
Mentions of LVEF	2276	0.978 (0.971–0.984)	0.986 (0.980–0.990)	0.982
LVEF quantitative values	2200	0.910 (0.897–0.921)	0.939 (0.928–0.949)	0.924
ACEI medications	2949	0.994 (0.990–0.996)	0.976 (0.970–0.981)	0.985
ARB medications	591	0.978 (0.963–0.988)	0.960 (0.941–0.974)	0.969
Reasons not to take ACEI/ARB	483	0.311 (0.270–0.354)	0.247 (0.213–0.283)	0.275
Overall (micro-average)	8499	0.928 (0.922–0.933)	0.917 (0.911–0.923)	0.922

Table 3. CHIEF information extraction results with partial matches (with 0.95 binomial exact confidence intervals)

Information extracted	N	Recall	Precision	F_1-measure
Mentions of LVEF	2276	0.986 (0.980–0.990)	0.994 (0.990–0.997)	0.990
LVEF quantitative values	2200	0.945 (0.934–0.954)	0.976 (0.968–0.982)	0.960
ACEI medications	2949	0.996 (0.993–0.998)	0.978 (0.972–0.983)	0.987
ARB medications	591	0.997 (0.988–1.000)	0.978 (0.963–0.988)	0.987
Reasons not to take ACEI/ARB	483	0.404 (0.360–0.449)	0.321 (0.284–0.359)	0.358
Overall (micro-average)	8499	0.946 (0.941–0.951)	0.935 (0.930–0.940)	0.940

develop our reference standard, but only 209 patient records (4724 clinical notes) were annotated at the concept level and used for the evaluation presented here. Each clinical note was classified as one of 10 medico-legal note types: history and physical, progress notes, cardiology consult, echocardiogram, pharmacy medication reconciliation, pharmacy other, other consult, discharge summary, nursing note, and overall other.

Agreement between annotators was assessed at the patient level using Cohen's kappa to determine reliability of the reference standard. This agreement was found to be excellent, reaching 0.910 in a pair-wise comparison of all annotations.

Left ventricular function information extraction

As already explained, two types of information were extracted in this module: mentions of LVEF, and quantitative values of LVEF. When considering exact matches between the system output and the reference standard, recall reached 0.980 for mentions of LVEF and 0.91 for quantitative values, with a precision (i.e., positive predictive value) of about 0.940–0.990 (Table 2).

When considering partial matches (i.e., any overlap), performance was significantly better for quantitative LVEF values. Recall and precision increased about 3–4% (from 0.910 to 0.945 and from 0.939 to 0.976) (Table 3).

CHF medications extraction

CHIEF extracted medications with high accuracy, reaching a recall of about 0.980–0.990 for mentions of ACEI and ARB, with an F_1-measure of 0.969–0.985 (Table 2). When considering partial matches, recall was close to 1.000 (Table 3).

Each annotated medication was also assigned a status category: active, discontinued, or negative. In the random sample of 3000 notes with 6007 annotated medications, 74.8% were active, 19.8% were discontinued, and 5.4% were negative. The 5 most frequently mentioned medications in the testing set were Lisinopril (52.6% of all medication annotations), ACEI in general (16.6%), Losartan (6.4%), ARB in general (5.1%), and Benazepril (4.9%).

We used a 5-fold cross validation with annotated medications to measure performance of medication prescription status classification. The overall accuracy reached 0.955. Precision of each status was above 0.900, and recall of the *discontinued* status was 0.862. Recall was higher than precision with the *negative* status, even though they corresponded to only 5.41% of the annotated medications in our corpus. A total of 230 (71 + 159) *active* or *discontinued* cases were misclassified as the other class (Table 4).

Reasons medications are not prescribed extraction

The RapTAT application was independently trained and tested with the same clinical note sets than the LVEF and medication extraction modules. As noted above, RapTAT was trained using a focused subset of 404 documents from 171 patients. There were 215 mentions of RNM in 77 of the files, and the F_1-measure for inter annotator agreement was 0.720. Based on cross-validation using the training set, optimal RapTAT performance was achieved when certain generally utilized text pre-processing steps were excluded, namely lemmatization, part-of-speech tagging, and removal of stop words. These pre-processing steps were subsequently excluded from system training and testing. Performance with cross-validation reached 0.404 recall, 0.688 precision, and 0.509 F_1-measure. As noted in Table 2, recall was lower when evaluated with the testing set, measured at 0.311 with RapTAT identifying RNM mentions, whereas precision was at 0.247. When considering partial matches, performance was significantly better. Recall and precision increased of about 30% (from 0.311 to 0.404 and from 0.247 to 0.321) (Table 3).

DISCUSSION

Results discussion

In general, CHIEF extracted CHF treatment performance measure information with high recall (0.946 with partial matches, or 0.981 if ignoring reasons not to take CHF medications) and precision (0.935 with partial matches, or 0.982 if ignoring reasons not to take CHF medications). Medications were extracted with very high recall (0.996–0.997) and good precision (0.978). Mentions of LVEF were also extracted with high recall (0.986) and precision (0.994), and quantitative values of LVEF were found with 0.945 recall and high precision (0.976). Reasons not to take CHF medications were more difficult to extract, only reaching fair accuracy with about 0.404 recall and 0.321 precision.

These results compare favorably with previous similar research. When extracting medication names during the 2009 i2b2 NLP chal-

Table 4. CHF medications status classification results

Status	Classified as			Count	Recall	Precision	F_1-measure
	Active	Discont.	Negative				
Active	4403	71	17	4491	0.980	0.963	0.971
Discontinued	159	1027	5	1191	0.862	0.929	0.895
Negative	12	7	306	325	0.942	0.933	0.937
Overall	4574	1105	328	6007	0.955	0.955	0.955

lenge,[12] the ten best-performing applications reached F_1-measures between 0.759 and 0.884. CHIEF reached an F_1-measure of 0.987 for the same task, although with a different set of clinical notes. CUIMANDREef reached a recall of 0.895 and a precision of 0.909 when extracting mentions of LVEF, and a recall of 0.910 and precision of 0.955 with quantitative values of LVEF. No previous study extracted reasons not to take CHF medications from clinical notes.

CHIEF errors analysis

The extraction of documented reasons for not prescribing ACEI or ARB medications reached the lowest accuracy. Difficulty in precisely defining when a potential contraindication qualified as an explicit justification appeared to be a critical constraint on performance for both the system and the reference standard. False negatives commonly occurred because detecting such mentions required integration of information from distinct parts of the text, which was not part of the RapTAT application. A detailed analysis of a sample of false positives suggested that approximately half of the 2 most common reasons given for not prescribing ACEI or ARB medications, allergies and hypotension, were probably missed by the annotators when developing the reference standard. The particularly low prevalence of the concept was also a substantial contributor to both false negative and false positive instances; it resulted in poor estimates of the likelihood of particular words or phrases mapping to the concept. To mitigate low prevalence, training of the system was supplemented with a manually generated dictionary of phrases that might be used to justify non-prescription of ACEI or ARB medications. This did reduce overall errors and improve recall as measured by cross-validation on the training data, but, not unexpectedly, this approach also reduced precision.

Among other information for CHF treatment performance measures, LVEF quantitative values had the lowest recall, missing 122 of them. These false negatives were mostly caused by values found far from the associated LVEF mention. Fifty LVEF values were spurious findings by CHIEF, mostly when another numeric value had a format similar to common LVEF value formats (e.g., "FEVI of 55%"). Mentions of LVEF were only missed 32 times, often because some white-space character was found in the middle of a mention of LVEF, such as "Ejection fr_action," "LV \n Function." Only 14 mentions of LVEF were false positives, and most were found to be errors in the reference standard (i.e., missed by both annotators). One example was terms split by newline characters, such as "pleural ef \n fusion."

ACEI and ARB medications were rarely missed (14 false negatives together). These false negatives were mostly misspellings with a longer edit distance than our threshold ($\leq$2; e.g., "lisnioril" for "lisinopril"). A total of 80 ACEIs and ARBs were false positives, and part of them was caused by expressions that included medication category terms. For example, "ace wrap" (an elastic bandage) or "prednisolone ace 1%" (abbreviation for "prednisolone acetate") were detected because of the term "ace." Others were counted as false positives but were actually missing in the reference standard (i.e., missed by both annotators).

Study limitations

The sample size for our evaluation was limited, although sufficient for reliable measurements of recall and precision with the most prevalent types of information (i.e., LVEF mentions and values, and ACEI medications). The sample size was relatively small for reasons not to take CHF medications, allowing for confidence intervals with a width of 0.075–0.089.

The reference standard of clinical note annotations was of good quality but not perfect. The agreement between independent annotators was excellent (Cohen's kappa of 0.91), but the detailed errors analysis of CHIEF's output revealed that several false positive errors were actually missing annotations in the reference standard.

The evaluation did not include a baseline for practical reasons: our aim was the development of an application supporting the CHF treatment quality improvement effort, not demonstrating information extraction accuracy improvements. This prevents ruling out that the task at hand was relatively easy and that the high accuracy can partly be attributed to that.

Finally, our reference standard was built from a selection of various types of clinical notes from eight VHA medical centers, allowing for good a variety of clinical text formats and content. But generalization of our results to other VHA medical centers or clinical note types could be limited, and even more so if generalizing to other healthcare organizations.

CONCLUSION

As demonstrated in this study, applying NLP to unlock the rich and detailed clinical information found in clinical narrative text notes makes fast and scalable quality improvement approaches possible. The automatically extracted treatment performance measures could improve management and outpatient treatment of patients suffering from CHF. CHIEF allows fast and scalable detection of CHF patients not benefiting from recommended treatment. We initiated implementation of CHIEF in two VA medical center settings by using the extracted data within a succinct message that delivers actionable information to clinicians at the point of care. We are currently finishing the user-centered design phase. We plan to refine the delivery of the information, develop needed infrastructure to support the deployment and subsequently evaluate our succinct message within the next several years.

FUNDING

This work was supported by VA HSR&D grants number IBE 09-069, HIR 08-374 (Consortium for Healthcare Informatics Research) and HIR 09-007 (Translational Use Case – Ejection Fraction).

COMPETING INTERESTS

The authors have no competing interests to declare.

CONTRIBUTORS

All authors made substantial contributions to the conception of the work or analysis and interpretation of data. S.M.M. conceived the CHIEF system and led its development. Y.K. was responsible for most development work. G.T.G. was responsible for the development of RapTAT, under the supervision of M.E.M. A.R. provided statistical expertise for data analysis, and B.E.B. provided cardiology domain expertise. All authors drafted the work or revised it crit-

ically. S.M.M. drafted the initial manuscript. Y.K., G.T.G., M.E.M., A.R., B.E.B., and J.H.G. provided critical revision of the manuscript. All authors gave final approval of the version to be published. All authors agree to be accountable for all aspects of the work in ensuring that questions related to the accuracy or integrity of any part of the work are appropriately investigated and resolved.

ACKNOWLEDGMENTS

We thank Julia Heavirland, Natalie Kelly, and Jenifer Williams for their significant contributions to this work, organizing and managing the reference standard development process.

The views expressed in this article are those of the authors and do not necessarily represent the views of the Department of Veterans Affairs, their academic affiliates or the University of Utah School of Medicine.

REFERENCES

1. Pfuntner A, Wier LM, Stocks C. *Most Frequent Conditions in U.S. Hospitals*, 2010. 2013. http://www.hcup-us.ahrq.gov/reports/statbriefs/sb148.pdf. Accessed May 1, 2016.
2. *Health Services Research and Development Service Quality Enhancement Research Initiative (QUERI) Chronic Heart Failure*. 2011. http://www.queri.research.va.gov/about/factsheets/chf_factsheet.pdf. Accessed May 1, 2016.
3. Heidenreich PA, Trogdon JG, Khavjou OA, *et al*. Forecasting the future of cardiovascular disease in the United States: a policy statement from the American Heart Association. *Circulation*. 2011;123(8):933–944.
4. Braunstein JB, Anderson GF, Gerstenblith G, *et al*. Noncardiac comorbidity increases preventable hospitalizations and mortality among Medicare beneficiaries with chronic heart failure. *J Am Coll Cardiol* 2003;42(7):1226–1233.
5. Yancy CW, Jessup M, Bozkurt B, *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. *Circulation*. 2013;128(16): e240–e327.
6. American College of Cardiology Foundation, American Heart Association, Physician Consortium for Performance Improvement. *Heart Failure*. 2012. http://www.ama-assn.org/ama1/pub/upload/mm/pcpi/hfset-12-5.pdf. Accessed May 1, 2016.
7. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008;128–144.
8. Friedlin J, McDonald CJ. A natural language processing system to extract and code concepts relating to congestive heart failure from chest radiology reports. *AMIA Annu Symp Proc* 2006;269–273.
9. Pakhomov S, Weston SA, Jacobsen SJ, Chute CG, Meverden R, Roger VL. Electronic medical records for clinical research: application to the identification of heart failure. *Am J Manag Care* 2007;13 (6 Part 1):281–288.
10. 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(12):983–992.
11. Vijayakrishnan R, Steinhubl SR, Ng K, *et al*. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *J Cardiac Failure* 2014;20(7):459–464.
12. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. *J Am Med Inform Assoc* 2010;17(5):514–518.
13. Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *J Am Med Inform Assoc* 2010;17(5):559–562.
14. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. *J Am Med Inform Assoc* 2010;17(5):524–527.
15. Meystre SM, Kim J, Garvin J. Comparing methods for left ventricular ejection fraction clinical information extraction. *AMIA Clinical Research Informatics Summit*. San Francisco, CA; 2012:138.
16. Kim Y, Garvin JH, Heavirland J, Meystre SM. Relatedness analysis of LVEF qualitative assessments and quantitative values. *AMIA Clinical Research Informatics Summit*. San Francisco, CA; 2013:123.
17. Kim Y, Garvin JH, Heavirland J, Meystre SM. Improving heart failure information extraction by domain adaptation. *Stud Health Technol Inform* 2013;192:185–189.
18. VA Corporate Data Warehouse (CDW). http://www.hsrd.research.va.gov/for_researchers/vinci/cdw.cfm. Accessed May 1, 2016.
19. VA Informatics and Computing Infrastructure (VINCI). http://www.hsrd.research.va.gov/for_researchers/vinci/. Accessed May 1, 2016.
20. Ogren PV. Knowtator. http://bionlp.sourceforge.net/Knowtator/. Accessed May 1, 2016.
21. Garvin JH, DuVall SL, South BR, *et al*. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. *J Am Med Inform Assoc* 2012;19(5):859–866.
22. South BR. *eHOST*. http://code.google.com/p/ehost/. Accessed May 1, 2016.
23. van Rijsbergen CJ. *Information Retrieval*. Oxford, UK: Butterworth; 1979.
24. Apache UIMA. http://uima.apache.org. Accessed May 1, 2016.
25. Gobbel GT, Reeves R, Jayaramaraja S, *et al*. Development and evaluation of RapTAT: a machine learning system for concept mapping of phrases from medical narratives. *J Biomed Inform* 2014;48:54–65.
26. Apache OpenNLP. http://opennlp.apache.org. Accessed May 1, 2016.
27. 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(5):507–513.
28. Berger AL, Pietra VJD, Pietra SAD. A maximum entropy approach to natural language processing. *Computational Linguistics*. MIT Press; 199622(1):39–71.
29. Kim Y, Garvin J, Heavirland J, Meystre SM. Automatic clinical note type classification for heart failure patients. *AMIA Clinical Research Informatics Summit*. San Francisco, CA; 2014:182.
30. Miralium. http://code.google.com/p/miralium/. Accessed May 1, 2016.
31. Crammer K, Singer Y. Ultraconservative online algorithms for multiclass problems. *J Machine Learning Res* 2003;3.
32. Kim Y, Hurdle J, Meystre SM. Using UMLS lexical resources to disambiguate abbreviations in clinical text. *AMIA Annu Symp Proc* 2011;2011:715–722.
33. Patrick JD, Nguyen DHM, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. *J Am Med Inform Assoc* 2011;18(5):574–579.
34. RxNorm. *U.S. National Library of Medicine*. https://www.nlm.nih.gov/research/umls/rxnorm/. Accessed May 1, 2016.
35. Levenshtein VI. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady* 1966;10:707.
36. Chang C-C, Lin C-J. *LIBSVM: a Library for Support Vector Machines*. 2001. http://www.csie.ntu.edu.tw/~cjlin/libsvm. Accessed May 1, 2016.
37. Meystre SM, Kim Y, Heavirland J, Williams J, Bray BE, Garvin JH. Heart failure medications detection and prescription status classification in clinical narrative documents. *Stud Health Technol Inform* 2015;216:609–613.
38. Kim Y, Riloff E, Meystre SM. Improving classification of medical assertions in clinical notes. *HLT '11 Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. 2011:311–316. http://www.aclweb.org/anthology/P/P11/P11-2054.pdf. Accessed May 1, 2016.
39. McCray AT, Aronson AR, Browne AC, Rindflesch TC, Razi A, Srinivasan S. UMLS knowledge for biomedical language processing. *Bull Med Libr Assoc* 1993;81(2):184–194.