Journal of Biomedical Informatics 67 (2017) 42–48

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

journal homepage: www.elsevier.com/locate/yjbin

Extraction of left ventricular ejection fraction information from various types of clinical reports

Youngjun Kim [a,b,*], Jennifer H. Garvin [b,c], Mary K. Goldstein [d,e], Tammy S. Hwang [d], Andrew Redd [b,f], Dan Bolton [b,f], Paul A. Heidenreich [d,e], Stéphane M. Meystre [c,g]

[a] School of Computing, University of Utah, Salt Lake City, UT, USA
[b] VA Health Care System, Salt Lake City, UT, USA
[c] Department of Biomedical Informatics, University of Utah, Salt Lake City, UT, USA
[d] VA Palo Alto Health Care System, Palo Alto, CA, USA
[e] Stanford University, Stanford, CA, USA
[f] Division of Epidemiology, University of Utah, Salt Lake City, UT, USA
[g] Medical University of South Carolina, Charleston, SC, USA

ARTICLE INFO

Article history:
Received 15 September 2016
Revised 16 December 2016
Accepted 31 January 2017
Available online 2 February 2017

Keywords:
Heart failure
Ventricular ejection fraction
Medical informatics
Natural language processing

ABSTRACT

Efforts to improve the treatment of congestive heart failure, a common and serious medical condition, include the use of quality measures to assess guideline-concordant care. The goal of this study is to identify left ventricular ejection fraction (LVEF) information from various types of clinical notes, and to then use this information for heart failure quality measurement. We analyzed the annotation differences between a new corpus of clinical notes from the Echocardiography, Radiology, and Text Integrated Utility package and other corpora annotated for natural language processing (NLP) research in the Department of Veterans Affairs. These reports contain varying degrees of structure. To examine whether existing LVEF extraction modules we developed in prior research improve the accuracy of LVEF information extraction from the new corpus, we created two sequence-tagging NLP modules trained with a new data set, with or without predictions from the existing LVEF extraction modules. We also conducted a set of experiments to examine the impact of training data size on information extraction accuracy. We found that less training data is needed when reports are highly structured, and that combining predictions from existing LVEF extraction modules improves information extraction when reports have less structured formats and a rich set of vocabulary.

1. Introduction

Heart Failure (HF) is a common but serious medical condition associated with high healthcare costs [1]. Efforts to improve the treatment of HF and reduce associated costs are facilitated by quality measures to assess treatment guideline-concordant care [2]. The American College of Cardiology Foundation, the American Heart Association, and the Physician Consortium for Performance Improvement published such performance measures, including a measure for "Left Ventricular Ejection Fraction (LVEF) Assessment" [3]. The LVEF measures the percentage of blood in the left ventricle expelled during the systolic contraction, and is not captured in a structured and coded format in many electronic health records, making its use for quality improvement and healthcare management difficult. The "Left Ventricular Ejection Fraction (LVEF) Assessment" measure describes the number of patients with a principal diagnosis of HF who have documentation of a planned or completed LVEF assessment. The LVEF is important because patients will benefit from life prolonging treatment only if their LVEF is low (e.g. <40%).

In order to make LVEF information available from text, we undertook two different projects to automate extraction of LVEF in the Department of Veterans Affairs (VA). Automated Data Acquisition for Heart Failure (ADAHF) [4] project automated the extraction of information needed for the inpatient HF quality measure from clinical notes, including LVEF and its values, and guideline-concordant HF medications, Angiotensin-Converting Enzyme Inhibitors (ACEI) and Angiotensin Receptor Blockers (ARB), and reasons why the patient was not prescribed these medications. The Consortium for Healthcare Informatics Research (CHIR) [5] Translational

* Corresponding author at: School of Computing, University of Utah, Salt Lake City, UT, USA.
E-mail address: youngjun@cs.utah.edu (Y. Kim).

http://dx.doi.org/10.1016/j.jbi.2017.01.017
1532-0464/

Use-Case Project for Ejection Fraction (TUCP-EF) [6] aimed at the automated extraction of LVEF mentions, LVSF (left ventricular systolic function) mentions, and their associated qualitative assessments and quantitative values:

- LVEF mentions (e.g., "left ventricular ejection fraction", "VISUAL ESTIMATE OF LVEF", "EF")
- LVEF quantitative values (e.g., "~0.60–0.65", "0.45", "50%")
- LVEF or LVSF qualitative assessments (e.g., "NORMAL", "mildly decreased", "SEVERE")
- LVSF mentions (e.g., "Global left ventricular systolic function", "systolic dysfunction", "LVSF")

Recent studies have used natural language processing (NLP) approaches to extract medical concepts related to HF from clinical notes. Chung and Murphy [7] developed a rule-based information extraction (IE) system to extract concepts and their associated values from echocardiogram reports. The concepts targeted for their study included Ejection Fraction (EF), Mitral Valve Insufficiency, Pericardial Effusion, etc. and the concept-value pairs related to the study conditions are identified using string matching, concept pattern matching and predefined tagging methods. Friedlin and McDonald [8] developed a specialized IE system, the REgenstrief eXtraction tool (REX) to extract HF related concepts (congestive heart failure, Kerley B lines, cardiomegaly, prominent pulmonary vasculature, pulmonary edema, and pleural effusion) from chest radiology reports. For patient level classification, to identify patients with HF, Pakhomov et al. [9] implemented two different methods based on rule-based and machine learning-based approaches.

Garvin et al. [6] developed a rule-based system within the UIMA (Unstructured Information Management Architecture) [10,11] framework, named "Capture with UIMA of Needed Data using Regular Expressions for Ejection Fraction" (CUIMANDREef), to extract LVEF related concepts and their associated values from VA echocardiogram reports. Subsequent research by Meystre et al. [12] showed that a sequence-tagging model achieved better performance than a rule-based approach on same corpus (*TUCP EF year 1 corpus*). Gobbel et al. [13] used the Rapid Text Annotation Tool (RapTAT) to assist annotation by interactively and iteratively pre-annotating HF related concepts. Their iterative training reduced the annotation time and produced more consistent annotation.

We developed an information extraction application based on rules and machine learning approaches to detect LVEF mentions and associated values in clinical notes as well as other information for HF treatment performance measures [4,6,12,14]. This application was named Congestive Heart Failure Information Extraction Framework (CHIEF) [12,14–16]. The CHIEF application is based on the Apache UIMA [10,11] framework with modules extracting LVEF information, medications (ACEIs and ARBs), and reasons not to administer these medications, with general linguistic analysis functionalities and patient-level analysis. Three versions of the LVEF extraction modules were developed:

- CUIMANDREef: a rule-based application using regular expressions to capture lexical patterns and a lexicon, trained with a specific corpus of VA echocardiogram reports (we will call it '*TUCP EF year 1 corpus*') [6].
- CHIEF EF: a machine learning-based version with sequential tagging using morphological, lexical, and syntactic information, trained with the *TUCP EF year 1 corpus* [12,14].
- CHIEF ADAHF: an adaptation of the CHIEF EF machine learning-based version, but trained on a corpus of VA clinical notes of different types (e.g., progress notes, discharge summaries, consultation notes, echocardiogram reports, etc.) (we will call it '*ADAHF corpus*') [4].

In this study, we used a new data set, a national corpus of VA clinical echocardiography reports described below (we will refer to it as the '*TUCP EF year 3 corpus*') to study the generalizability of our LVEF extraction modules. We start with an analysis of differences between the new *TUCP EF year 3* corpus and other corpora annotated for previous studies in Section 2.1. We then describe our NLP methods based on a machine learning sequential tagger and discuss features used for identifying LVEF information (Section 2.2).

The contributions of our work are twofold. First, we examined how three existing LVEF extraction modules perform with the new *TUCP EF year 3* corpus, a corpus composed of different clinical notes than the ones used for the original development and training of the LVEF extraction modules (Section 3.1). To investigate the possible contributions of these existing LVEF extraction modules, we trained machine learning algorithms with the new corpus, with or without predictions from the existing LVEF extraction modules, and examined if this approach allowed for improved accuracy in Section 3.2. Second, we examined the impact of training data quantity on the new NLP modules accuracy (Section 3.3). We trained models with or without predictions from the existing LVEF extraction modules, with different amounts of training data, and conducted a set of experiments to evaluate the accuracy of each combination.

2. Materials and methods

2.1. Annotation data sets and concepts

2.1.1. TUCP EF year 1 corpus

The *TUCP EF year 1* corpus consisted of 765 echocardiogram reports, obtained from the Text Integration Utility (TIU) [17] section of the Veterans Health Information Systems and Technology Architecture (VistA) [18], from seven VA medical centers [6]. Two human annotators annotated these reports independently using an annotation schema and guideline, and a third annotator adjudicated differences between the two original annotators. We used 275 notes for training ('*TUCP EF year 1* ***train***') and 490 notes for testing ('*TUCP EF year 1* ***test***').

2.1.2. ADAHF corpus

The *ADAHF* corpus [4], obtained from the TIU files, includes clinical notes from inpatients with HF discharged from a selection of 8 VA medical centers in 2008. It includes 18,397 notes that were manually annotated for LVEF and its values, for a selection of HF medications (ACEIs and ARBs), and for reasons not to administer these medications. The corpus was split in a training set of 13,673 notes ('*ADAHF* ***train***') and a testing set of 4724 notes ('*ADAHF* ***test***'). The *ADAHF corpus* contains various clinical document types (progress notes, discharge summaries, history and physical notes, cardiology consultation notes, etc.) and echocardiogram reports represented about 1.5% of the notes found in the *ADAHF* corpus.

2.1.3. TUCP EF year 3 corpus

The new dataset, *TUCP EF year 3* corpus includes 3060 clinical notes from three different storage locations and associated software ("packages") used within the VistA systems: Echocardiography (1140 reports, from 16 VA Health Care sites sampled at random), Radiology (720 reports, from 5 sites sampled at random), and TIU (1200 reports, from 17 sites sampled at random). Among these 3060 reports, 1465 containing at least one of our concepts of interest were selected for this project. Two reviewers independently annotated each clinical note with eHOST (Extensible Human Oracle Suite of Tools) [19] and then a domain expert adjudicated

the disagreements between the two primary annotators. The F_1-measure for inter annotator agreement was 0.9099.

For this study, we divided the *TUCP EF year 3 corpus* into training and test set by randomly assigning notes from the 3 different packages. We used 732 notes for training from the TUCP-EF Year 3 ('*TUCP EF year 3* ***train***') and 733 notes for testing ('*TUCP EF year 3* ***test***'). Table 1 shows the number of concepts and notes found in each corpus.

The format of the reports varied depending on the package they came from. Reports from the Echocardiography package tended to be shorter and have highly structured and simple data formats (e.g., "EstimatedEF: 60%"). Reports from the Radiology package tended to have more unstructured text and technical measurements. Notes from the TIU package tended to have a combination of structured and unstructured text.

The three corpora were manually annotated with a few differences in the types and definitions of information captured, as shown in Table 1. These corpora were developed for different projects, with annotation differences as well as differences in clinical note type and format. A key difference between the corpora annotations is how LVSF mentions were captured. In the *TUCP EF year 1* corpus, LVSF mentions and LVEF mentions were annotated as one "LVEF mentions" category. LVSF mentions were not annotated in the *ADAHF* corpus. In the *TUCP EF year 3* corpus, LVSF mentions were annotated as a separate category. Because of these differences, the direct application of CUIMANDREef or CHIEF EF to TUCP was not possible and required adaptations described below.

2.2. LVEF information extraction modules

To examine the generalizability of our modules to the new *TUCP EF year 3 corpus*, we carried out several experiments with three versions of the LVEF extraction modules and a new version trained on the *TUCP EF year 3 training* set.

2.2.1. Existing LVEF extraction modules

As described above, we previously developed three versions of the LVEF extraction modules: CUIMANDREef [6], CHIEF EF [12,14], and CHIEF ADAHF [4] to detect LVEF mentions and associated values in clinical notes. CUIMANDREef is a rule-based application using regular expressions to capture lexical patterns targeting specific VA echocardiogram reports. CHIEF EF and CHIEF ADAHF are machine learning-based applications using morphological, lexical, and syntactic information. The output of these three existing LVEF extraction modules was used with automated post-processing, examining all extracted LVEF mentions and filtering out LVSF mentions by searching for keywords compiled from the '*TUCP EF year 1 train*' corpus (e.g., "LVSF", "LVSD", "systolic", "dysfunction", "function"). This filtering was required because of the aforementioned LVSF annotation differences.

2.2.2. New TUCP LVEF extraction modules

Two different approaches were used, all based on a machine learning sequential tagger using *Miralium* [20], a Java implementation of the Margin Infused Relaxed Algorithm (MIRA) [21], with or without the outputs from existing LVEF extraction modules. We reformatted the training instances with BIO tags (B: at the beginning, I: inside, or O: outside of a concept) and then trained the models to produce BIO labels for all four concept types (*LVEF*, *LVSF*, *Qualitative value*, and *Quantitative value*). All training ended after ten iterations and no feature pruning was done for this study.

(1) The first version, simply called '*TUCP*', was trained with the *TUCP EF year 3 corpus* using the same feature set than CHIEF EF and CHIEF ADAHF. These features were obtained from text analysis and included words (current word, four preceding words, and four following words), bi-grams of words, part-of-speech tags (current word part-of-speech tag, part-of-speech tags of four preceding words, and four following words), bi-grams of part-of-speech tags, word morphology (alpha-numeric characters, punctuations, etc.), word shape information (e.g., "LVSF" normalized to "AAAA", "ef" to "aa"; for the current word, three preceding words, and three following words), infixes (prefix and suffix), and the output of CUIMANDREef (for the current word, two preceding and two following words).

(2) The second version, called '*TUCP + Prediction*', combined features from *TUCP* and the predictions from CUIMANDREef, CHIEF EF, and CHIEF ADAHF to allow the sequence tagger trained with the *TUCP EF year 3* corpus to benefit from existing LVEF extraction modules [22]. We first processed the complete *TUCP EF year 3* corpus with the three existing LVEF extraction modules, and then reformatted their output (concept type and text spans) with BIO tags and used it as new features to train the new *TUCP + Prediction* module. The architecture of *TUCP + Prediction* is depicted in Fig. 1.

3. Results

3.1. Evaluation metrics

We used three metrics to measure information extraction accuracy: recall (equivalent to sensitivity), precision (positive predictive value) and 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 reports for compar-

Table 1
Corpora characteristics and annotated information differences.

	TUCP EF year 1		ADAHF		TUCP EF year 3	
	Train	Test	Train	Test	Train	Test
Corpora characteristics						
- Number of notes	279	490	13,673	4724	732	733
- Number of notes with concepts	275 (98.6%)	487 (99.4%)	3364 (24.6%)	1224 (25.9%)	732 (100%)	733 (100%)
Number of concepts						
- LVEF mentions	723	1250	6561	2276	1122	1124
- LVSF mentions	0	0	0	0	609	628
- Qualitative values	439	759	0	0	630	648
- Quantitative values	430	746	5939	2200	1100	1085
Annotation differences						
- LVEF mentions		Annotated		Annotated		Annotated
- LVSF mentions		Annotated as "LVEF"		Not annotated		Annotated
- Qualitative values		Annotated		Not annotated		Annotated
- Quantitative values		Annotated		Annotated		Annotated

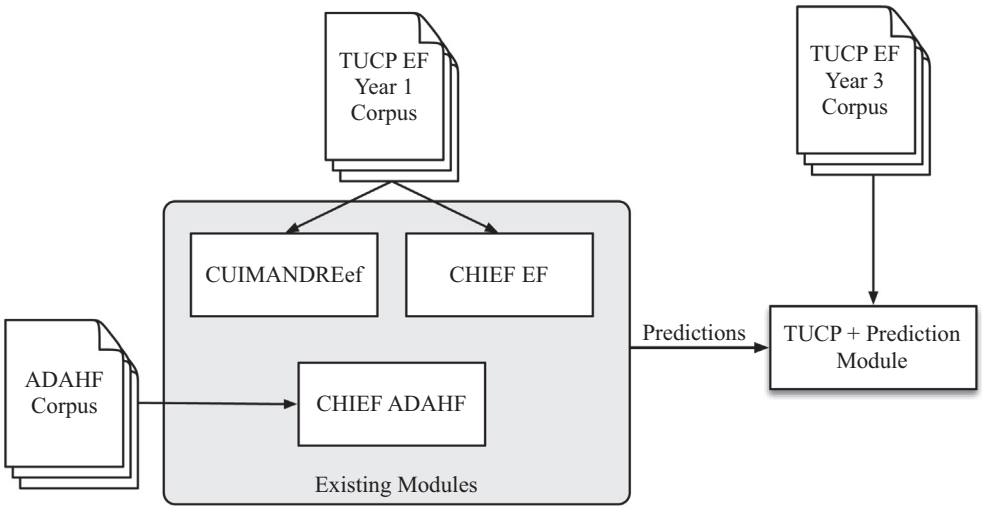

Fig. 1. Architecture of TUCP + Prediction module.

ison of exact matches (perfect text span match between the annotated reference standard and the system output) and inexact matches (overlap of annotated reference standard text spans with at least one word of the system output).

3.2. Performance of existing LVEF extraction modules

When trained and tested with the same corpus (trained with the training subset, and tested with the testing subset), CHIEF EF reached quite good performance with a 95.5% F_1-measure [14]. Precision was about 3% higher than recall on average, except with quantitative values. CHIEF ADAHF worked very well with LVEF mentions (98.2% F_1-measure) [4]. For quantitative values, the recall of CHIEF ADAHF was lower than CHIEF EF, reflecting the increased difficulty of capturing values in non-echocardiogram notes. CHIEF ADAHF reached overall 95.4% F_1-measure (Table 2). For detailed results, refer to [14,4] for CHIEF EF and CHIEF ADAHF respectively.

When trained with their original corpus, and then tested with the *TUCP EF year 3 corpus*, without any feature modification or re-training, CUIMANDREef, CHIEF EF, and CHIEF ADAHF did not perform well with exact matches, but obtained reasonable results with inexact matches (Table 3).

CUIMANDREef obtained the highest recall (62.7%) for qualitative values and the best precision (93.5%) for quantitative values with exact matches. CHIEF EF showed overall good performance, especially for quantitative values (85.2% F_1-measure) and LVSF mentions (74.5% F_1-measure) with exact matches. CHIEF ADAHF obtained the best performance for LVEF mentions. When counting

Table 2
Exact match results of existing modules trained and tested with the same corpus (percentages).

	CHIEF EF tested with 'TUCP EF year 1 test' [14]			CHIEF ADAHF tested with 'ADAHF test' [4]		
	R	P	F	R	P	F
LVEF mentions	93.8	97.1	95.4	97.8	98.6	98.2
Qualitative values	91.8	96.7	94.2	–	–	–
Quantitative values	97.1	96.9	97.0	91.0	93.9	92.4
All concepts	94.1	96.9	95.5	94.4	96.3	95.4

Recall (R), precision (P), and F_1-measure (F).

Table 3
Results of existing modules tested with the *TUCP EF year 3 corpus*.

	CUIMANDREef			CHIEF EF			CHIEF ADAHF		
	R	P	F	R	P	F	R	P	F
Exact match									
LVEF mentions	30.3	39.2	34.2	36.6	44.9	40.3	**54.7**	**46.9**	**50.5**
LVSF mentions	60.7	82.8	70.0	**65.6**	**86.2**	**74.5**	–	–	–
Qualitative values	**62.7**	71.6	**66.8**	36.4	**72.0**	48.4	–	–	–
Quantitative values	68.7	**93.5**	79.2	**81.0**	89.8	**85.2**	73.7	89.6	80.9
All concepts	53.7	69.5	60.6	**55.6**	**71.8**	**62.7**	40.6	64.2	49.7
Inexact match									
LVEF mentions	70.6	91.3	79.6	76.7	**94.1**	84.5	**97.2**	83.4	**89.8**
LVSF mentions	63.5	86.7	73.3	**70.5**	**92.7**	**80.1**	–	–	–
Qualitative values	**65.4**	74.8	**69.8**	38.7	**76.5**	51.4	–	–	–
Quantitative values	71.3	97.1	82.3	**86.6**	96.0	**91.1**	80.4	**97.6**	88.2
All concepts	68.6	88.8	77.4	**71.6**	**92.4**	**80.7**	56.4	89.2	69.1

Recall (R), precision (P), and F_1-measure (F) with highest recall, precision, and F_1-measure for each concept bolded.

only LVEF mentions and quantitative values, the results of CHIEF ADAHF had an F_1-measure of 64.1% with exact matches, and 89.1% with inexact matches. The overall F_1-measure with exact matches was about 18% lower than with inexact matches on average. For LVEF mentions, there was about 40% difference between exact and inexact matches. One reason for this difference arose from including (or excluding) LVEF mention modifiers in the reference standard annotations. For example, in the phrase "ESTIMATED EF", the whole phrase had to be annotated in the reference standard for CHIEF EF, but only "EF" was annotated as LVEF mention in the *TUCP EF year 3 corpus*.

3.3. Performance of the new TUCP LVEF extraction modules

When training and testing with the *TUCP EF year 3 corpus*, performance was much improved with the new TUCP LVEF extraction modules. All results are listed in Table 4 (exact match and inexact match). *TUCP + Prediction* slightly outperformed *TUCP*. We confirmed that the feature set used for CHIEF EF and CHIEF ADAHF could be successfully utilized with the *TUCP EF year 3* corpus.

Fig. 2 shows exact match results with each category of clinical notes (package). In each graph, the first bar represents results with the *TUCP* model, and the second bar with the *TUCP + Prediction*

Table 4
Results of TUCP models.

	TUCP			TUCP + Prediction		
	R	P	F	R	P	F
Exact match						
LVEF mentions	**96.8**	96.5	**96.6**	96.1	**96.6**	96.3
LVSF mentions	91.6	**92.6**	**92.1**	**91.7**	91.6	91.6
Qualitative values	74.4	87.8	80.5	**79.0**	**89.5**	**83.9**
Quantitative values	93.6	93.0	93.3	**93.8**	**93.5**	**93.7**
All concepts	90.7	93.2	92.0	**91.4**	**93.5**	**92.4**
Inexact match						
LVEF mentions	**98.9**	98.6	**98.8**	98.4	**98.9**	98.7
LVSF mentions	**95.1**	**96.1**	**95.6**	94.6	94.4	94.5
Qualitative values	78.1	92.2	84.5	**82.3**	**93.2**	**87.4**
Quantitative values	**99.3**	**98.6**	**98.9**	98.9	98.5	98.7
All concepts	94.5	**97.1**	95.8	**94.9**	97.0	**95.9**

Recall (R), precision (P), and F_1-measure (F) with highest recall, precision, and F_1-measure for each concept bolded.

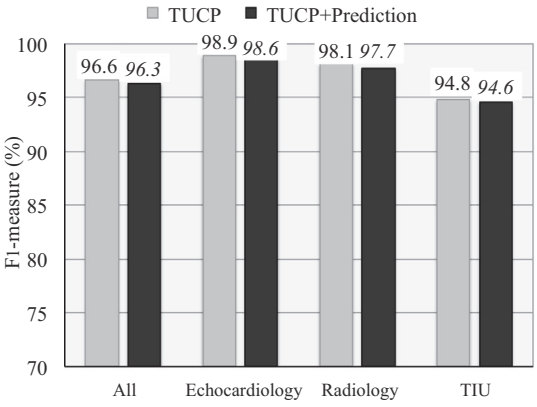

Figure 2.1. LVEF mentions

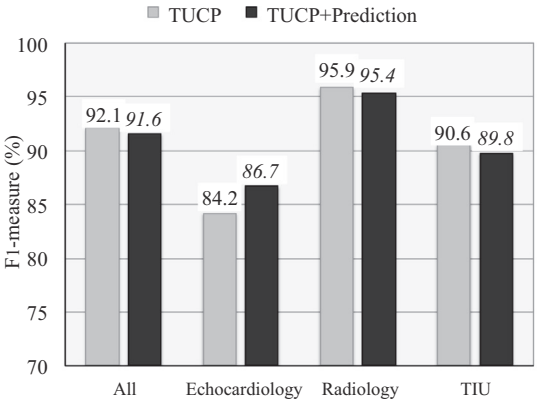

Figure 2.2. LVSF mentions

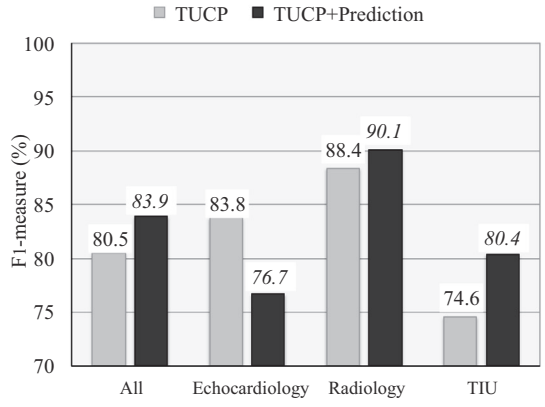

Figure 2.3. Qualitative values

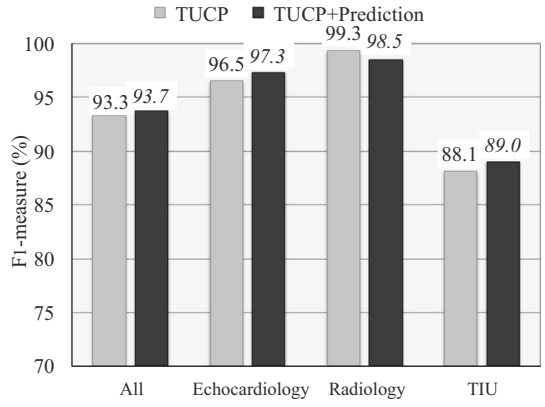

Figure 2.4. Quantitative values

Fig. 2. Exact match results with each package reports.

model. Note that the y-axis scale in each graph does not start at zero to focus on the value ranges of interest.

With Echocardiography and Radiology package reports, *TUCP + Prediction* did not always allow for any significant improvement over *TUCP*, probably because of the small vocabulary of our target concepts. In Echocardiography package reports, *TUCP + Prediction* performed better than *TUCP* when extracting quantitative values and LVSF mentions, but worse with qualitative values. In Radiology package reports, *TUCP + Prediction* only performed better when extracting qualitative values. In TIU package reports, *TUCP + Prediction* showed better performance in overall recall and precision when compared with *TUCP*, with the largest improvement for qualitative values.

3.4. Performance comparison with different quantities of training data

We designed additional experiments to examine the impact of the quantity of training data on the accuracy of our modules. We created four different training data subsets by randomly selecting sets of 20% of training data and then combining them to have subsets of 20%, 40%, 60%, 80%, and 100% of our training data.

We trained the *TUCP* and *TUCP + Prediction* modules with each training data subset, and then tested them with our original testing set. Fig. 3 shows the measured F_1-measures with different amounts of training data. In each figure, the dotted line represents results with *TUCP*, and the solid line with *TUCP + Prediction*.

On all test data, *TUCP + Prediction* consistently outperformed *TUCP* (about 1% higher) when using 20–80% of the training data. When using all training data, *TUCP + Prediction* achieved an F_1-measure of 92.4%, only about 0.4% higher than *TUCP*. As explained above, with Echocardiography and Radiology package reports, *TUCP + Prediction* did not affect performance much. *TUCP* obtained better performance than *TUCP + Prediction* when using most training data subsets. With TIU package reports, *TUCP + Prediction* performed better than *TUCP*, producing about 2% higher F_1-measures with most subsets. Our analysis revealed that *TUCP + Prediction* excels at extracting LVEF mentions and values from the notes having less structured formats and a rich set of vocabulary, like TIU notes.

Statistical comparisons of difference between *TUCP* and *TUCP + Prediction* were based on the z-test between two proportions. This analysis demonstrated that the F1-measures of *TUCP + Prediction* using 60% and 80% of the training data were significantly better than *TUCP* with TIU package reports at the $p < 0.05$ significance level, but no statistically significant difference was found between the two models in other scenarios.

Overall, using more training data allowed for better performance. When using all test data, *TUCP* obtained an F_1-measure of 87.2% with 20% of the training data and an F_1-measure of 92.0%

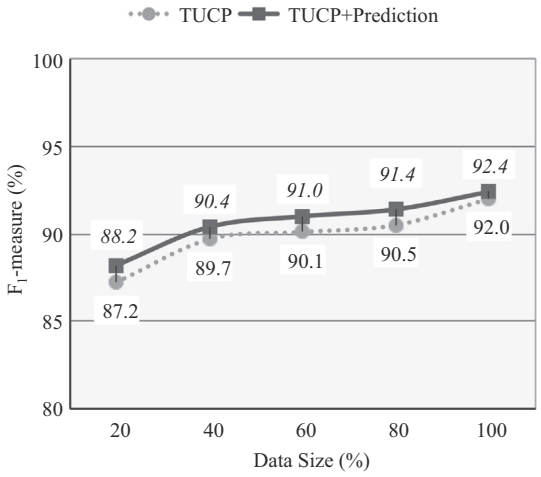

Figure 3.1. All of test data

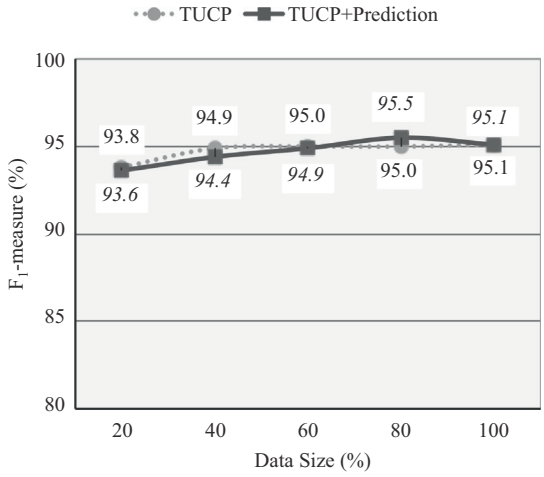

Figure 3.2. Echocardiography

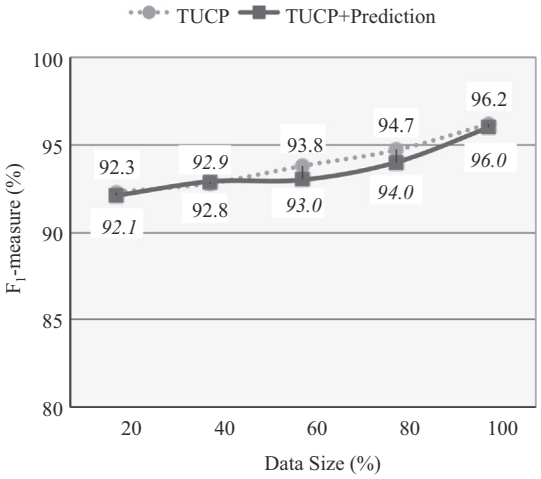

Figure 3.3. Radiology

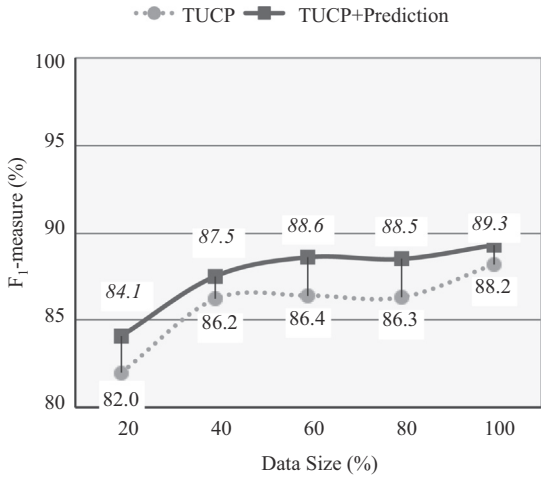

Figure 3.4. TIU

Fig. 3. Performance comparison between *TUCP* and *TUCP + Prediction* with different size of training data.

with all training data, about 5% higher. However, with Echocardiography package reports, only using 40% of training data already allowed for an F_1-measure of 94.9%. With Radiology package reports, the F_1-measure increased regularly with the amount of training data. With TIU package reports, *TUCP* obtained an F_1-measure of 82.0% when using only 20% of the training data and an F_1-measure of 88.2% with all training data (about 6% higher). The F_1-measure increased rapidly when using 40% or more of the training data. In general, we observed that less training data would be needed when using highly structured notes with less complicated data formats such as Echocardiography and Radiology package reports.

4. Discussion

The new TUCP LVEF extraction modules performed much better than existing modules with the *TUCP EF year 3* corpus. When clinical reports are highly structured, such as Echocardiography and Radiology package reports, performance did not change much. However, *TUCP + Prediction* performed better than *TUCP* when reports have less structured formats and a rich set of vocabulary, such as TIU package reports. With small amounts of training data, *TUCP + Prediction* performed significantly better than *TUCP* with TIU package reports.

We manually inspected all false negative and false positive errors of our modules. Among false negative errors (i.e., missed information) by *TUCP + Prediction*, we observed that some LVEF mentions were missed when found in a different sentence than their associated values. Also, some errors were caused by incorrect boundaries for LVEF mentions. For example, in "*estimated ejection fraction*", our system only identified "*ejection fraction.*" We observed similar term boundary errors with qualitative or quantitative values. For example, in "*overall preserved*", our system only identified "*preserved*"; in "*plus – minus 56%*" and "*range of 50% to 55%*", only "*56%*" and "*50% to 55%*" were detected. Other false negative errors were due to rare terms in our training data. For example, with "*EF was slightly diminished to about 40%*", our system missed a qualitative value, "*slightly diminished*", because "*diminished*" occurred only once in the training data. Finally, some LVSF mention false negative errors were caused by incorrect tokenization. For example, in "*Grossly normal L V*", "*LV*" was missed when a whitespace character was found within the term.

5. Conclusion

This study showed that our LVEF information extraction application, CHIEF, could be successfully applied to a new data set, reaching good or excellent recall and precision. We observed that our application performed well with the new corpus when applying the same original feature set used for training existing modules. The large number of different note types (e.g., echocardiogram report versus radiology note) and differences between medical centers caused a great variety of clinical note formats and content. We observed that less structured notes containing various formats of data may need more training data than highly structured notes, and that better accuracy can be achieved by leveraging existing applications trained with different clinical data sets. An accurate method for LVEF abstraction will be valuable to the VA and other health care systems in their efforts to measure and improve heart failure care.

Funding

This publication was supported by the United States (U.S.) Department of Veterans Affairs, Veterans Health Administration, Office of Research and Development, IDEAS 2.0 HSR&D Research Center, Grant Nos. IBE 09-069 and HIR 08-374 (Consortium for Healthcare Informatics Research) and HIR 09-007 (Translational Use Case).

Competing interests

The authors have no competing interests to declare.

Acknowledgments

The contents do not represent the views of the Department of Veterans Affairs, their academic affiliates, the University of Utah School of Medicine, or the United States Government.

References

[1] P.A. Heidenreich, J.G. Trogdon, O.A. Khavjou, et al., Forecasting the Future of Cardiovascular Disease in the United States: A Policy Statement from the American Heart Association, Lippincott Williams & Wilkins, 2011. p. 933–944.
[2] S.A. Hunt, W.T. Abraham, M.H. Chin, 2009 focused update incorporated into the ACC/AHA 2005 guidelines for the diagnosis and management of heart failure in adults a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines: developed in collaboration with the international society for heart and lung transplantation, Circulation 119 (2009) e391–e479.
[3] American College of Cardiology Foundation, American Heart Association, Physician Consortium for Performance Improvement. Heart Failure, 2012.
[4] S.M. Meystre, Y. Kim, G.T. Gobbel, et al., Congestive heart failure information extraction framework for automated treatment performance measures assessment, JAMIA (2016), http://dx.doi.org/10.1093/jamia/ocw097.
[5] The Consortium for Healthcare Informatics Research (CHIR), <http://www.research.va.gov/funding/solicitations/docs/Consortium-Healthcare-Informatics.pdf>.
[6] J.H. Garvin, S.L. Duvall, B.R. South, et al., Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure, JAMIA 19 (5) (2012) 859–866, http://dx.doi.org/10.1136/amiajnl-2011-000535.
[7] J. Chung, S. Murphy, Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports, AMIA Annu. Symp. Proc. (2005) 31–135.
[8] J. Friedlin, C.J. McDonald, 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] S. Pakhomov, S.A. Weston, S.J. Jacobsen, et al., Electronic medical records for clinical research: application to the identification of heart failure, Am. J. Manage. Care (2007) 281–288.
[10] Apache UIMA 2008, Available at <http://uima.apache.org>.
[11] D. Ferrucci, A. Lally, UIMA: an architectural approach to unstructured information processing in the corporate research environment, J. Nat. Lang. Eng. 10 (3–4) (2004) 327–348.
[12] S.M. Meystre, Y. Kim, J.H. Garvin, Comparing methods for left ventricular ejection fraction clinical information extraction, AMIA Summits Transl. Sci. Proc. CRI (2012) 138.
[13] G.T. Gobbel, J. Garvin, R. Reeves, et al., Assisted annotation of medical free text using RapTAT, JAMIA 21 (5) (2014) 833–841.
[14] Y. Kim, J.H. Garvin, J. Heavirland, S.M. Meystre, Improving heart failure information extraction by domain adaptation, Stud. Health Technol. Inform. 192 (2013) 185–189.
[15] Y. Kim, J.H. Garvin, J. Heavirland, S.M. Meystre, Relatedness analysis of LVEF qualitative assessments and quantitative values, AMIA Summits Transl. Sci. Proc. CRI (2013).
[16] Y. Kim, J.H. Garvin, J. Heavirland, S.M. Meystre, Automatic Clinical Note Type Classification for Heart Failure Patients, AMIA Summits Transl Sci Proc, CRI. (2014).
[17] Text Integration Utilities (TIU) Technical Manual; Version 1.0., 1997. <http://www.va.gov/vdl/documents/Clinical/CPRS-Text_Integration_Utility_(TIU)/tiutm.doc> (Revised June 2010).
[18] VistA Monograph, <http://www.ehealth.va.gov/VistA_Monograph.asp>.
[19] B. South, S. Shen, J. Leng, et al., A prototype tool set to support machine-assisted annotation, in: Proc. Conf. BioNLP, 2012, pp. 130–139.
[20] Miralium, <http://code.google.com/p/miralium/>.
[21] K. Crammer, Y. Singer, Ultraconservative online algorithms for multiclass problems, J. Mach. Learn. Res. 3 (2003) 951–991.
[22] R. Florian, H. Hassan, A. Ittycheriah, et al., A statistical model for multilingual entity detection and tracking, in: Proc. Conf. NAACL and HLT, 2004.
[23] C.J. van Rijsbergen, Information Retrieval, Butterworth, Oxford, UK, 1979.