Evaluating the Portability of an NLP System for Processing Echocardiograms: A Retrospective, Multi-site Observational Study

Prakash Adekkanattu, PhD[1], Guoqian Jiang, MD PhD[2], Yuan Luo, PhD[3], Paul R. Kingsbury, PhD[2], Zhenxing Xu, PhD[1], Luke V. Rasmussen, MS[3], Jennifer A. Pacheco, MS[3], Richard C. Kiefer, MS[2], Daniel J. Stone, MS[2], Pascal S. Brandt, MS[4], Liang Yao, PhD[3], Yizhen Zhong, BS[3], Yu Deng, BS[3], Fei Wang, PhD[1], Jessica S. Ancker, MPH PhD[1], Thomas R. Campion, Jr, PhD[1], Jyotishman Pathak, PhD[1]

[1]Weill Cornell Medicine, New York, NY, [2]Mayo Clinic, Rochester, MN, [3]Northwestern University, Chicago, IL, [4]University of Washington, Seattle, WA

Abstract

While natural language processing (NLP) of unstructured clinical narratives holds the potential for patient care and clinical research, portability of NLP approaches across multiple sites remains a major challenge. This study investigated the portability of an NLP system developed initially at the Department of Veterans Affairs (VA) to extract 27 key cardiac concepts from free-text or semi-structured echocardiograms from three academic medical centers: Weill Cornell Medicine, Mayo Clinic and Northwestern Medicine. While the NLP system showed high precision and recall measurements for four target concepts (aortic valve regurgitation, left atrium size at end systole, mitral valve regurgitation, tricuspid valve regurgitation) across all sites, we found moderate or poor results for the remaining concepts and the NLP system performance varied between individual sites.

Introduction

Echocardiography is currently the most widely used non-invasive cardiac imaging in the diagnosis, treatment, and follow-up of patients with cardiovascular diseases. Various measurements in echocardiography provide key insights into the mechanisms and health of a patient's cardiovascular system[1]. Aortic, mitral and tricuspid valves are assessed for abnormal structure, thickness, and valve dysfunction (regurgitation, stenosis). Left ventricular systolic performance indicates severity of heart disease, and most laboratories quantify Left Ventricular Ejection Fraction (LVEF) to a value ranging from 5-75 percent of total blood volume. Echocardiography measurements are generally stored as an echocardiogram (echo) report in electronic health record (EHR) systems, which is frequently an unstructured or semi-structured text document. This limits the downstream use of the report data by automated systems for clinical research (e.g., phenotyping), care management and point-of-care clinical decision support[2].

Natural language processing (NLP) of unstructured clinical narratives, such as echo reports, holds tremendous potential to extract meaningful information and fill this knowledge gap[3-6]. NLP utilizes various algorithms to automatically extract relevant clinical information from free text and semi-structured data sources. However, a recent review of the literature by Demner-Fushman and Elhadad suggests that NLP remains an "emerging technology", with a significant gap between promise and reality[7]. One of the major challenges in broader adoption of NLP systems and technologies is their portability across multiple EHR systems. While several studies, including work done by our study team, have demonstrated varying levels of success in portability of NLP technologies across institutions[8-12], recent work by Carrell et al. argue that there remain significant challenges in adapting NLP systems across multiple sites, which include assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions[13]. These barriers and the requirement for significant upfront "investment" limit wide-scale adoption of NLP tools and systems across health systems, particularly those in low-resource settings and environments.

To understand these barriers further, in this study we investigated the implementation of an existing NLP system - Leo, developed by the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI) - across three academic medical centers. The goal of the current study was to assess the portability of *EchoExtractor* by installing, without any system modifications whatsoever, and evaluating its performance at three non-VA medical centers. This project, as part of the Phenotype Execution and Modeling Architecture (PhEMA), was motivated by the fact that EHR based computational phenotyping using both structured and unstructured data is a key requirement for deriving high quality phenotypes from EHRs[14], and a portable NLP system that can be run across multiple institutions is much needed[15,16].

Materials and Methods

<u>System Description and Target Concepts</u>

Leo provides a set of services and libraries that facilitate the rapid creation and deployment of NLP analysis tools. A major strength of this approach is the use of Unstructured Information Management Architecture (UIMA), which provides a set of standards for creating, discovering, composing and deploying a broad range of text analytics capabilities and integrating them with search technologies[17]. This standard is very attractive for NLP research because it is open-source (available through Apache[18]), and enables incorporation of NLP tools developed elsewhere. The distribution package for *EchoExtractor* was made available through a VA GitHub repository[19]. Installation of *EchoExtractor* using UIMA-AS 2.8.1 was deemed to be relatively straightforward, and was completed within a day at WCM and Northwestern. Installation at Mayo Clinic was performed using installation instructions found for Leo[20] where an alternate version of UIMA (UIMA-AS 2.9.0) was installed to support Leo, yet the installation process was still also completed within a day. We tested the performance of the pipeline without any custom modification to evaluate the true portability of the NLP system.

While previous studies have shown success in extracting echocardiography measurements from Echo reports[12] using NLP methods, the focus has been primarily on extracting one key concept - LVEF. The VINCI team created a specific instance of Leo named *EchoExtractor* to extract 27 different important cardiac concepts from echo and radiology reports[21]. A preliminary analysis of results obtained after processing 100 echo reports from WCM center through the NLP pipeline revealed that not all target concepts were present. Concepts such as left ventricular contractility, left ventricular hypertrophy and aortic valve max pressure gradient were not identified. Therefore, we focused on the remaining 24 elements in our extended evaluation study. Later, we observed that the above three concepts were in fact present in echo reports originated from other centers, however we did not include them in our performance evaluation. The 24 target concepts we investigated in this study is listed in Table 1. For a description of these concepts, we refer readers to Patterson, et al.[21].

<u>Data collection</u>

We obtained Echo reports from all three medical centers participating in this study. Weill Cornell Medicine (WCM) uses the EpicCare® Ambulatory EHR platform to document clinical care, from which we collected all echocardiograms used as the basis of this study. We obtained 200 echocardiograms for 200 patients who were 18 years or older at the time of the echocardiogram results and one of the study team members (PA) manually annotated the data elements extracted by the *EchoExtractor* to create the gold standard at WCM. These reports in general formatted into multiple sections containing semi-structured, and unstructured data. Reports have a subject meta-data section, one or more tabular sections, and a comment section. The tabular sections contain quantitative measures such as wall thicknesses, chamber dimensions, or flow velocities. Unstructured fields contain descriptions of clinically relevant findings as interpreted by the technician. For example, the descriptions could say "Normal aortic valve and aortic root", "Normal left ventricular size and function", "There is trace mitral regurgitation". Furthermore, left ventricular function is often subjectively quantified as "mild", "moderate", "mildly to moderately reduced" or "severely reduced".

Mayo Clinics implemented an Unified Data Platform (UDP) to provide practical data solutions that creates a combined view of multiple heterogeneous EHR data sources (including Epic) through effective data orchestration, along with a number of data marts based on common data models. We retrieved the echocardiogram reports from the Mayo Clinic UDP platform from patients over 18 years old at the time of the echocardiogram results after verifying authorization to use the data in accordance with the Minnesota Health Records Act. Patient records were found in two formats: Rich Text Format (2000 to 2012), and Portable Document Format (PDF) from 2012 to 2018. For the echocardiograms generated between 2012-18, using a custom-built Python application, original text from the PDF documents was extracted. We randomly selected 200 echocardiogram reports for this study from each document format. The text portions of these reports were extracted, PHI was censored, and other elements such as visual elements (charts and graphics) and metadata were discarded. Echo reports found at Mayo Clinic generally consisted of three types of sections: unstructured text, numbered final impressions, and semi-structured tables of echocardiogram data. These reports were manually annotated with data elements extracted by the *EchoExtractor* by an experienced nurse at the Mayo Clinic which generated an annotation corpus as Mayo Clinic's gold standard.

Table 1. List of target concepts found in echo reports at Weill Cornell Medicine, The Mayo Clinic, Northwestern Medicine, and MIMIC sites. Each concept is shown as either present (P) or absent (A) at individual sites.

Concept	WCM	MAYO	NU	MIMIC
aortic valve mean gradient	P	A	P	A
aortic valve orifice area	P	P	P	P
aortic valve regurgitation	P	P	P	P
aortic valve regurgitation peak velocity	A	A	P	A
aortic valve stenosis	P	P	P	P
e/e prime ratio	P	P	P	P
inter-ventricular septum dimension at end diastole	P	P	P	P
left atrium size at end systole	P	P	P	P
left ventricular dimension at end diastole	P	P	P	P
left ventricular dimension at end systole	P	P	P	A
left ventricular size	P	P	P	P
left ventricular ejection fraction	P	P	P	P
left ventricular posterior wall thickness at end diastole	P	P	P	P
mitral valve mean gradient	A	P	P	A
mitral valve orifice area	P	A	P	A
mitral valve regurgitation	P	P	P	P
mitral valve regurgitation peak velocity	A	A	P	A
mitral valve stenosis	P	P	P	P
pulmonary artery pressure	P	P	P	P
right atrial pressure	P	P	P	P
tricuspid valve mean gradient	A	A	P	A
tricuspid valve orifice area	A	A	P	A
tricuspid valve regurgitation	P	P	P	P
tricuspid valve regurgitation peak velocity	P	P	P	P

At Northwestern Medicine, we retrieved 251,521 free-text echocardiogram reports for patients over 18 years old at the time of the echocardiogram between 2000 and 2017 from Northwestern Medicine® Enterprise Data Warehouse (NMEDW). The NMEDW hosts a comprehensive and integrated repository of the most used clinical and research data sources, at the time from 2 hospitals: Northwestern Memorial Hospital (NMH) and Lake Forest Hospital (LFH). Echo reports typically contained a mixture of semi-structured tables of echocardiogram data, along with unstructured impressions and findings. Overall structure of reports was primarily a metadata section containing patient and procedure information, semi-structured tabular results, followed by the impression (although many variations in this structure were observed). We randomly selected 200 echocardiogram reports and multiple study team members (YZ, YD, LY, JP, LR, YL) manually annotated the data elements extracted by the *EchoExtractor*. Across reviewers, we have discordant findings reconciled in a group setting to ensure consistent coding and interpretation, sometimes in

consultation with a cardiovascular physician colleague. The final consolidated list of annotations was used as the gold standard at Northwestern.

We also obtained echocardiograms from the Medical Information Mart for Intensive Care III (MIMIC-III) database[22]. The latest version, MIMIC-III, contains de-identified patient records for >40,000 critical care patients admitted to Beth Israel Deaconess Medical Center between 2001 and 2012. We extracted 44,559 echo reports from 21,645 adult patients who were 18 years or older. We then randomly selected 200 echo reports for 200 unique patients and a study team member (PA) manually annotated the data elements extracted by the *EchoExtractor* to create the gold standard for the MIMIC dataset. Unlike the Echo reports from three medical centers, MIMIC reports were in free-text format with no tabular or semi-structured sections. In general, these reports have an interpretation section comprising subsections for findings and conclusions. Findings section has detailed descriptions about left atrium, right atrium, left ventricle, right ventricle, aorta, aortic valve, mitral valve, tricuspid valve, and pulmonic valve/pulmonic artery. MIMIC reports are for patients admitted to the intensive care unit; hence, patient characteristics are different from those patients treated at various outpatient settings in WCM, Mayo Clinic and Northwestern. At each medical center we limited the number of reports for manual review to 200 mainly due to limited resource availability, and also due to the fact that, in general, echo reports tend to follow same format(s) within each clinical setting. Figure 1 shows typical examples of echo report from the four EHR systems investigated in this study.

Figure 1. Typical examples of echocardiogram originated from multiple EHR systems

Extraction methodology

Patterson et al. has described in detail the logic for concept extraction employed in the present study[21]. Various target concepts were extracted through a multi-stage process in a rule-based algorithm that involves identifying a concept mention in the text, identifying a value mention in the text, and linking appropriate concept and value mentions in a relationship. Term identification and term mapping was used to detect a concept mention in the text. Term identification was performed using regular expressions matching and term mapping was achieved by building a custom term lookup dictionary. Value identification was achieved through regular expressions that match both quantitative

and qualitative values for concepts. Finally, a relationship between a concept and value was established through pattern matching, from a set of patterns identified for the clinical subdomain of echocardiograms.

<u>Reference standard and performance evaluations</u>

For all three study sites and for the MIMIC dataset, we determined performance of the NLP pipeline individually by developing independent gold standard datasets (see above). Results of *EchoExtractor* output were tabulated and compared against the gold standard through manual reviews. For each concept, a reviewer compared the output generated by *EchoExtractor* to each of the 200 reports. The reviewer identified all mentions of a concept and the associated quantitative or qualitative values. The reviewer also confirmed that the concepts not found by *EchoExtractor* in a given report were truly not mentioned in the respective report. The reviewer identified both numeric values as well as qualitative descriptions such as normal, mild, moderate, dilated, and severe. If a numeric value is documented as a range, then we identified both the values as the minimum and maximum. In reports containing multiple instances of a concept, we identified all instances. In cases where concept value is expressed using a greater than or less than symbol (e.g. LVEF >55), the reviewer simply identified the value ignoring the symbol. Some echo reports express uncertainty about certain concepts such as LVEF value using a question mark (e.g. LVEF ?55-70). In such cases, the reviewer extracted the value, ignoring the question mark.

Each report was classified as one of four possible cases: true positive (report had the concept present and *EchoExtractor* identified the "concept-value pair" and matched with the value given in the reference standard); false positive (report had the concept absent, but *EchoExtractor* produced a concept-value pair); true negative (report had the concept absent, and *EchoExtractor* did not find any concept-value pair); and false negative (report had the concept present, but *EchoExtractor* did not identify a concept-value pair, or the value extracted did not match with the corresponding value in the text). A value for a concept can be either a quantitative measure as in "systolic function (LVEF>55%)" for ejection fraction, or a qualitative assessment as in "no aortic regurgitation". Note that we are not assessing if a given concept is present or absent in the report, rather whether the NLP method correctly identified the concept-value pairs documented in the report. For example, for a statement such as "There is no aortic valve stenosis", if the NLP system extracted 'no' as the assessment for the concept 'aortic valve stenosis', we will consider this as a true positive case. For a given target concept, when there were multiple instances of concept-value pair extracted by *EchoExtractor*, we used the following heuristic measure to select a given instance of the concept in order to compare directly with the reference standard: we selected the last concept-value pair, normally in the conclusion part of the report (as the order of results and text in the reports was usually chronological). The total outcomes of the four possible cases were then used to calculate various statistical performance measures, namely, precision (positive predictive value), recall (sensitivity or true positive rate), and the F-score (the harmonic mean of recall and precision).

Results

Results of *EchoExtractor* validation performed on echo reports from WCM, Mayo, Northwestern and MIMIC datasets are presented in Table 2. Among all datasets aortic valve regurgitation, aortic valve stenosis, left ventricular size, LVEF, mitral valve regurgitation, pulmonary artery pressure, and tricuspid valve regurgitation were the most frequently mentioned concepts. Other concepts were mentioned infrequently or in some cases absent all together as indicated by the letter 'A' in Table 2. While target concepts such as aortic valve regurgitation peak velocity, mitral valve regurgitation peak velocity, tricuspid valve mean gradient, and tricuspid valve orifice area were absent from the echo reports from WCM, Mayo Clinic and MIMIC reports, only aortic valve regurgitation peak velocity was absent from the Northwestern dataset. In addition, mitral valve mean gradient was absent from the WCM echo reports and mitral valve orifice area was absent from the MIMIC and Mayo echo reports. At WCM, aortic valve mean gradient, aortic valve orifice area, inter-ventricular septum dimension at end diastole, left atrium size at end systole, left ventricular dimension at end diastole and left ventricular dimension at end systole were also found in high frequency. Among Mayo, Northwestern and MIMIC echo reports, left atrium size at end systole was found frequently. Both WCM, Northwestern and MIMIC reports showed high recall and precision in extracting LVEF, with an F-score of 0.99, 0.99 and 0.95 respectively. We observed that *EchoExtractor* showed high precision (99%), while low recall (55%) for extracting LVEF from Mayo Clinic echo reports. Across all four datasets a F-score >0.90 was found on only two target concepts: aortic valve regurgitation and left atrium size at end systole. At WCM, Mayo Clinic and Northwestern, an F-score >0.90 was observed for aortic valve regurgitation, left atrium size at end systole, mitral valve regurgitation, and tricuspid valve regurgitation. Between WCM, Northwestern and MIMIC reports, a F-score >0.90 was observed for the concepts aortic valve regurgitation, aortic valve stenosis, left atrium size at end systole, left ventricular ejection fraction, and tricuspid valve regurgitation. Between WCM and Northwestern sites, mitral valve regurgitation showed an F-score >0.90.

Table 2. Performance evaluation of EchoExtractor NLP pipeline on echocardiograms originated at three medical centers (WCM, Mayo Clinic, Northwestern) and MIMIC

Concept	WCM Recall(%)	WCM Precision (%)	WCM F-score	MAYO Recall(%)	MAYO Precision (%)	MAYO F-score	NU Recall(%)	NU Precision (%)	NU F-score	MIMIC Recall(%)	MIMIC Precision (%)	MIMIC F-score
aortic valve mean gradient	100	100	100	A	A	A	39	100	56	A	A	A
aortic valve orifice area	100	77	87	25	75	38	8	64	13	0	0	0
aortic valve regurgitation	96	100	98	87	94	90	90	100	95	87	100	93
aortic valve regurgitation peak velocity	A	A	A	A	A	A	0		0	A	A	A
aortic valve stenosis	88	100	94	32	100	49	92	100	96	87	94	90
e/e prime ratio	0	0	0	25	50	33	94	100	97	100	100	100
inter-ventricular septum dimension at end diastole	99	100	100	38	25	30	28	95	43	100	38	56
left atrium size at end systole	98	100	99	98	96	97	92	99	96	85	98	91
left ventricular dimension at end diastole	100	100	100	0	0	0	28	100	44	0	0	0
left ventricular dimension at end systole	95	100	97	0	0	0	28	100	44	A	A	A
left ventricular size	67	97	79	73	94	82	39	100	56	38	99	54
left ventricular ejection fraction	99	99	99	55	99	71	96	100	99	91	100	95
left ventricular posterior wall thickness at end diastole	6	100	12	67	86	75	43	99	60	0	0	0
mitral valve mean gradient	A	A	A	6	100	11	39	100	56	A	A	A
mitral valve orifice area	100	6	11	A	A	A	31	100	48	A	A	A
mitral valve regurgitation	96	100	98	91	96	94	85	100	92	66	100	79
mitral valve regurgitation peak velocity	A	A	A	A	A	A	9	100	16	A	A	A
mitral valve stenosis	60	100	75	17	100	29	83	100	91	40	100	57
pulmonary artery pressure	86	100	93	32	35	33	49	95	65	87	100	93
right atrial pressure	0	0	0	0	0	0	33	96	49	67	50	57
tricuspid valve mean gradient	A	A	A	A	A	A	84	70	76	A	A	A
tricuspid valve orifice area	A	A	A	A	A	A	67	100	80	A	A	A
tricuspid valve regurgitation	90	100	95	92	97	95	98	100	99	71	100	83
tricuspid valve regurgitation peak velocity	0	0	0	0	0	0	96	98	97	0	0	0

Error analysis of the echo reports from the study datasets revealed that false positive cases are mainly due to term mismapping or the reference range for a concept being wrongly mapped as the measured value. For instance, in WCM echo reports several target concepts are documented in semi-structured tabular format where a given concept is documented by the measured value followed by the reference value (e.g. "Aortic Valve Area   1.1 > 2.4 cm2").

When there is no measured value documented as in several instances of aortic valve orifice area (e.g. "Aortic Valve Area > 2.40 cm2") or left ventricular ejection fraction (e.g. "Ejection Fraction 0.55 - 0.75") or mitral valve orifice area (e.g. "Mitral Valve Area > 3 cm2"), then the NLP system extracted the value given for the reference range as the measured value. For aortic valve orifice area and mitral valve orifice area, 23% and 94%, respectively, of reports were found to be false positive this way. Term mismapping also leads to a higher proportion of false positives. For example, the term "e:a" ratio being wrongly mapped to e/e prime ratio as in "Mitral E:A Rt 2" accounting 18% of false positive cases for the target concept e/e prime ratio. Although rare, the system wrongly mapped to terms "dilated aorta arch" to target concept left ventricular size with no apparent reason. In MIMIC echo reports, majority of false positive cases are due to term mismapping. Examples are "interatrial septum" being mapped to the concept interventricular septum dimension at end diastole, left ventricular cavity being mapped to the concept left ventricular dimension at end diastole, dilated descending aorta being mapped to left ventricular size, or mitral valve jet being mapped to tricuspid valve regurgitation peak velocity. Also, when a short representation of a target concept is part of a word preceded by a space character; the system wrongly mapped that concept. An example is "ava" for aortic valve orifice area being mapped to the word "available". Also, some false positive cases were due to wrong mapping of the qualitative assessment of one concept to another. For example, the assessment "mildly thickened" of aortic valve leaflets was mapped to aortic stenosis as in "The aortic valve leaflets (3) are mildly thickened but aortic stenosis is not present."

Error analysis on false negative cases revealed several instances where the system missed the target concept due to variations in the terms used to represent that concept in different clinical settings. For example, in WCM echo reports, the concept aortic valve max pressure gradient is consistently documented under peak aortic gradient, and since this is not one of the lookup terms for that concept and missed entirely. A second cause for false negative cases was when a line break (\n) or special characters (;, ~ , &), or parentheses appear between a concept and the corresponding value or qualitative assessment. Yet another cause for false negative cases was when a term for a target concept appears in two lines (e.g. right "\n" atrial pressure), and the system failed to extract the concept. Similarly, when two target concepts appeared too close to each other, the system failed to map those concepts with the correct value or assessment. For example, in MIMIC echocardiograms, there were several instances where aortic stenosis was described along with aortic valve orifice area (e.g. Mild AS (AoVA 1.2-1.9cm2)), and the system failed to correctly extract aortic valve area.

For the Mayo Clinic's echo reports, we found that semi-structured table format had negative impact on the performance of the concept extraction. For example, "e/e prime ratio" listed in the table entries could not be detected consistently and accurately. Therefore, the semi-structured tables were excluded from the performance evaluation. Otherwise, the Mayo Clinic's echo reports appeared to have relatively consistent syntax. We also found that semicolon (;) was used in place of colon (:) for the representation of some concept values such as "Calculated left ventricular ejection fraction; 65 %", which resulted in roughly half of the false negatives being recorded for this concept. In addition, some phrase patterns appeared to be more commonly used at Mayo Clinic (e.g. Mitral valve sclerosis without stenosis), which resulted in 58% of the false negatives for the detection of the mitral valve stenosis concept and 47% of the false negatives for the aortic valve stenosis concept. A similar issue occurred for "Aortic valve systolic mean Doppler gradient 12 mmHg:", where the insertion of "Doppler" term caused a failure to detect the gradient concept. The high occurrence rate of these specific phrase patterns appeared to be institution-specific and needed to be handled in the future.

At Northwestern, we also found multiple semi-structured table formats in the echocardiogram reports, which also led to poor NLP performance for some of the target concepts. Also, we found many of the other sources of errors as in other datasets, including the interjection of spacing and special characters in the reports, omissions of spacing between concepts and values (e.g., Ao valve open2.2 cm), different concept terms used, and multiple concepts in the same sentence. Of the few false positives we found, the NLP did not recognize some less common abbreviations such as those for aortic insufficiency (AI) that were not listed in most reports. Many of the false negatives were due to the above reasons; in particular, multiple concepts were frequently missed as they were in a separate table and the description of the measures were partly abbreviated (vs. fully spelled out or completely abbreviated as in other reports). For example, left ventricular size measures such as left ventricular dimension at end systole were listed in the table as "LV Size-end systole" instead of "LVESD" or "left ventricular size-end systole." The remaining few false negatives were due to missing values in sentences where other concepts were mentioned. For example, some reports started their summary section at the end of the report with summarizing multiple normal measures in 1 statement, such as "Left ventricular size, systolic function, wall thickness, and wall motion are all normal." There were very few instances of

FNs where the value extracted by the NLP was the wrong value for the concept-value pair; thus, the NLP algorithm was good at finding the correct value when it found a concept.

Discussion

While many modern EHR systems record various concepts from echocardiograms in a more structured manner, semi-structured or unstructured echocardiogram reports are still the primary source in most clinical settings. In the past a number of NLP methods have been developed to extract various cardiac measures documented in echocardiography reports. While majority of these studies focused on the extraction of LVEF[12,23-25], some have attempted to extract a broader range of cardiac concepts[21,26]. However, to the best of our knowledge these methods have not been tested for their performance on data originated from clinical sites other than the original NLP algorithm and method development site. As a result, "true portability" of the developed NLP systems has never been rigorously assessed. The present study fill this knowledge gap by adopting a system developed initially by the VA, and then subsequently implemented at three non-VA medical centers without any modifications, to test its performance in processing echocardiogram reports.

In particular, the NLP system under investigation - *EchoExtractor* - was designed to extract 27 concepts from narrative text, echocardiograms, and radiology reports, and the original developers reported an average precision of concept-value extraction of 0.982 in echo documents, 0.969 in Radiology, and 0.936 in narrative text, with F-score of 0.844, 0.877, and 0.872 respectively. Specifically, the system reported precision and recall values for 20 data elements in echocardiograms. The remaining 7 target concepts were either absent or too low in total mentions to report performance concepts. Out of the 20 target concepts in echocardiograms, 17 elements have been reported with a precision of 0.90 or higher[21]. When compared to the reported data, the overall performance of the NLP system was found to be mixed. For an NLP system to be portable, a low variability in the recall and precision values across sites is needed. As shown in Table 2, this is true for some target concepts but not for others. For those concepts where we found low inter-column variability, *EchoExtractor* can be deemed as an effective NLP tool. In WCM dataset, target concepts such as inter-ventricular septum dimension at end diastole, left ventricular dimension at end diastole, and left ventricular dimension at end systole are documented as part of a tabular format within echo reports and appeared consistently across all reports, and hence resulted high performance for these concepts. On the other hand, concepts such as aortic valve orifice area and mitral valve orifice area, although appear in a tabular format and are typically documented along with a reference range in the echo reports. For these concepts we observed that the NLP system often failed to correctly identify the measured values from the reference range, thereby resulting in lower precision.

The Mayo Clinic dataset comprised of semi-structured tables to represent the echocardiography concepts. While the table formats had modest variations over the years, the NLP system was generally unable to detect concepts located in these tables successfully. Our evaluations suggest that the representation of target concepts using a table format is likely to decrease the NLP system performance, and alternative approaches to capturing data in these sections should be evaluated. Specifically, we observed 16 target concepts in the Mayo Clinic dataset, out of which the NLP system achieve a high precision and F-score >0.90 for only 4 concepts. A notable exception in Mayo Clinic dataset compared to other datasets is the low recall observed on LVEF. While WCM, Northwestern and MIMIC echo reports showed high recall (99%, 97%, and 91%, respectively), Mayo Clinic echo reports showed only 55 percent. On further analysis, we concluded that the high rate of failure to detect LVEF is due to a semicolon ';' character (e.g. ejection fraction; 67) that appeared between the concept and value, which was causing system error. Common institutional phrasings such as "sclerosis without stenosis" that do not fit into the predefined patterns can be a significant source for concern and poor performance.

We observed 23 target measurements on Northwestern echo reports, out of which the NLP system showed a precision and F-score values >0.90 for 10 elements. Unlike echo reports from other sites, the concept tricuspid valve regurgitation peak velocity was found with high recall and precisions. Although this concept was mentioned infrequently in reports from other datasets, they were mentioned in ~50% reports at Northwestern using system recognized concept names. Similarly, the measurement tricuspid valve orifice area is found only in Northwestern echo reports, and extracted with a high recall and precision.

Unlike the echocardiograms from WCM, Northwestern and Mayo, MIMIC reports are entirely in a free-text format. Of the 16 target concepts found in MIMIC dataset, NLP system achieved a high precision and F-score values for 6 concepts. Majority of false positive and false negative cases in datasets that are more semi-structured arise out of differences in local document formats. Thus, the differences in performance between MIMIC and other datasets indicates a possible correlation between structure and tool usefulness. This is particularly important, since an ideal NLP system should be able to perform well when analyzing unstructured narratives. For those concepts observed with

low performance, it is also possible that physicians entering data in MIMIC are constrained by internal guidelines to use peculiar phrasing or punctuation and the NLP system missed to correctly identify those concepts.

Conclusion

The work described in this paper details a case study to adopt a specific instance (*EchoExtractor*) of a NLP system (Leo) previously developed by the VA. The adopters (WCM, Mayo Clinic and Northwestern) individually led the NLP system implementation without any modification, drawing on existing resources, and employing conventional software skills. The study demonstrates that concept-value extraction from echocardiograms can vary depending on local text formats and variations in lexical terms used to document various concept across different clinical settings. The NLP system used a custom lookup dictionary as part of its extraction algorithm[21], and our analysis showed that the performance of the system for several of the target concept can be improved significantly by extending this lookup dictionary with additional terms specific to individual sites. Many of the false negative cases resulted because of the absence of terms in the custom dictionary. Despite the fact that this system was developed based on documents originated at thousands of points of care and authored by a multitude of clinical professionals within the VA, it also suffers from some of the same drawbacks as other NLP systems have reported including the need for a local customization of extraction logic and extending the lookup dictionary based on local data source.

Acknowledgements

References

1. McAlister NH, McAlister NK, Buttoo K. Understanding cardiac "echo" reports. Practical guide for referring physicians. Can Fam Physician. 2006;52(7):869-74. PMID: 16893150, PMCID: PMC1781094
2. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, Liu S, Zeng Y, Mehrabi S, Sohn S, Liu H, Clinical information extraction applications: A literature review, Journal of Biomedical Informatics, 2018;77:34-49, 1532-0464. PMID: 29162496 PMCID: PMC5771858
3. Yim WW, Yetisgen M, Harris WP, Kwan SW. Natural Language Processing in Oncology: A Review. JAMA Oncol. 2016 Jun 1;2(6):797-804. doi: 10.1001/jamaoncol.2016.0213. Review. PubMed PMID: 27124593.
4. Pons E, Braun LM, Hunink MG, Kors JA. Natural Language Processing in Radiology: A Systematic Review. Radiology. 2016 May;279(2):329-43. doi: 10.1148/radiol.16142770. Review. PubMed PMID: 27089187.
5. Névéol A, Zweigenbaum P. Clinical Natural Language Processing in 2014: Foundational Methods Supporting Efficient Healthcare. Yearb Med Inform. 2015 Aug 13;10(1):194-8. doi: 10.15265/IY-2015-035. Review. PubMed PMID: 26293868; PubMed Central PMCID: PMC4587052.
6. Griffon N, Charlet J, Darmoni SJ. Managing free text for secondary use of health data. Yearb Med Inform. 2014 Aug 15;9:167-9. doi: 10.15265/IY-2014-0037. Review. PubMed PMID: 25123738; PubMed Central PMCID: PMC4287081
7. Demner-Fushman D, Elhadad N. Aspiring to Unintended Consequences of Natural Language Processing: A Review of Recent Developments in Clinical and Consumer-Generated Text Processing. Yearb Med Inform. 2016 Nov 10;(1):224-233. Review. PubMed PMID: 27830255
8. Mehrabi S, Krishnan A, Roch AM, Schmidt H, Li D, Kesterson J, Beesley C, Dexter P, Schmidt M, Palakal M, Liu H. Identification of Patients with Family History of Pancreatic Cancer--Investigation of an NLP System Portability. Stud Health Technol Inform. 2015;216:604-8. PubMed PMID: 26262122.
9. Martinez D, Pitson G, MacKinlay A, Cavedon L. Cross-hospital portability of information extraction of cancer staging information. Artif Intell Med. 2014 Sep;62(1):11-21. doi: 10.1016/j.artmed.2014.06.002. PubMed PMID: 25001545.
10. Carroll RJ, Thompson WK, Eyler AE, Mandelin AM, Cai T, Zink RM, Pacheco JA, Boomershine CS, Lasko TA, Xu H, Karlson EW, Perez RG, Gainer VS, Murphy SN, Ruderman EM, Pope RM, Plenge RM, Kho AN, Liao KP, Denny JC. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. J Am Med Inform Assoc. 2012 Jun;19(e1):e162-9. PubMed PMID: 22374935; PubMed Central PMCID: PMC3392871.
11. Zheng K, Vydiswaran VG, Liu Y, Wang Y, Stubbs A, Uzuner Ö, Gururaj AE, Bayer S, Aberdeen J, Rumshisky A, Pakhomov S, Liu H, Xu H. Ease of adoption of clinical natural language processing software: An evaluation of five systems. J Biomed Inform. 2015 Dec;58 Suppl:S189-96. doi: 10.1016/j.jbi.2015.07.008. PubMed PMID: 26210361; PubMed Central PMCID: PMC4974203.

12. Johnson SB, Adekkanattu P, Campion Jr. TR, Flory J, Pathak J, Patterson OV, et al. From Sour Grapes to Low-Hanging Fruit: A Case StudyDemonstrating a Practical Strategy for Natural Language Processing Portability. AMIA Annu Symp Proc. 2017 Nov. PMID: 29888051 PMCID: PMC5961788

13. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc. 2017 Sep 1;24(5):986-991. doi: 10.1093/jamia/ocx039. PubMed PMID: 28419261

14. Z. Zeng, Y. Deng, X. Li, T. Naumann and Y. Luo, "Natural Language Processing for EHR-Based Computational Phenotyping," in IEEE/ACM Transactions on Computational Biology and Bioinformatics, vol. 16, no. 1, pp. 139-153, 1 Jan.-Feb. 2019. doi: 10.1109/TCBB.2018.2849968

15. Mo H, Thompson WK, Rasmussen LV, et al. Desiderata for computable representations of electronic health records-driven phenotype algorithms. J Am Med Inform Assoc. 2015;22(6):1220-30. PMID: 26342218 PMCID: PMC4639716

16. Sharma H, Mao C, Zhang Y, Vatani H, Yao L, Zhong Y, Rasmussen LV, Jiang G, Pathak J, Luo Y Developing a Portable Natural Language Processing Based Phenotyping System. CoRR, (2018). abs/1807.06638.

17. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;10(3-4):327-348

18. Apache. Unstructured Information Management Architecture (UIMA). https://uima.apache.org. Accessed 11 March 2019

19. The Department of Veterans Affairs : EchoExtractor https://github.com/department-of-veterans-affairs/EchoExtractor Accessed 11 March 2019

20. The Department of Veterans Affairs : Leo https://department-of-veterans-affairs.github.io/Leo/userguide.html Accessed 11 March 2019

21. Patterson OV, Freiberg MS, Skanderson M, J Fodeh S, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC Cardiovasc Disord. 2017;17(1):151. Published 2017 Jun 12. doi:10.1186/s12872-017-0580-8. PMID: 28606104 PMCID: PMC5469017

22. Johnson, Alistair EW, David J. Stone, Leo A. Celi, and Tom J. Pollard. "The MIMIC Code Repository: enabling reproducibility in critical care research." Journal of the American Medical Informatics Association (2017): ocx084

23. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, Pickard S, Heidenreich P, Shen S, Weir C, Samore M, Goldstein MK. 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):859–66. doi: 10.1136/amiajnl-2011-000535.

24. Kim Y, Garvin JH, Goldstein MK, Hwang TS, Redd A, Bolton D, Heidenreich PA, Meystre SM. Extraction of left ventricular ejection fraction information from various types of clinical reports. J Biomed Inform. 2017;67:42–8. doi: 10.1016/j.jbi.2017.01.017.

25. Xie F, Zheng C, Yuh-Jer Shen A, Chen W. Extracting and analyzing ejection fraction values from electronic echocardiography reports in a large health maintenance organization. Health Inform J. 2016. doi:10.1177/1460458216651917.

26. Nath C, Albaghdadi MS, Jonnalagadda SR. A Natural Language Processing Tool for Large-Scale Data Extraction from Echocardiography Reports. PLoS ONE. 2016;11(4):0153749. doi: 10.1371/journal.pone.0153749