From Sour Grapes to Low-Hanging Fruit:  
A Case Study Demonstrating a Practical Strategy for Natural Language Processing Portability

Stephen B. Johnson, Ph.D. 1, Prakash Adekkanattu, Ph.D. 2, Thomas R. Campion, Jr., Ph.D. 12, James Flory, M.D., M.S. 1, Jyotishman Pathak, Ph.D. 1, Olga V. Patterson, Ph.D. 3,4, Scott L. DuVall, Ph.D. 3,4, Vincent Major, M.S. 5, Yindalon Aphinyanaphongs, M.D., Ph.D. 5

1Healthcare Policy and Research, Weill Cornell Medicine, New York, New York; 2Information Technologies & Services, Weill Cornell Medicine, New York, New York; 3VA Salt Lake City Health Care System; 4University of Utah, Salt Lake City, UT; 5Center for Health Informatics and Bioinformatics, NYU Langone Medical Center, New York, New York

Abstract
Natural Language Processing (NLP) holds potential for patient care and clinical research, but a gap exists between promise and reality. While some studies have demonstrated portability of NLP systems across multiple sites, challenges remain. Strategies to mitigate these challenges can strive for complex NLP problems using advanced methods (hard-to-reach fruit), or focus on simple NLP problems using practical methods (low-hanging fruit). This paper investigates a practical strategy for NLP portability using extraction of left ventricular ejection fraction (LVEF) as a use case. We used a tool developed at the Department of Veterans Affairs (VA) to extract the LVEF values from free-text echocardiograms in the MIMIC-III database. The approach showed an accuracy of 98.4%, sensitivity of 99.4%, a positive predictive value of 98.7%, and F-score of 99.0%. This experience, in which a simple NLP solution proved highly portable with excellent performance, illustrates the point that simple NLP applications may be easier to disseminate and adapt, and in the short term may prove more useful, than complex applications.

Introduction
Natural Language Processing (NLP) holds tremendous potential for patient care and clinical research. 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. The NLP community has engaged in numerous challenge tasks in recent years, which have been beneficial in improving technical methods and research collaboration. But, due to the artificial nature of tasks suitable for such competitions, these efforts have had limited impact on real-world problems. Several studies have demonstrated success in portability of NLP technologies across institutions. However, a recent paper by Carrell et al. argues that there remain serious challenges in adapting NLP systems across multiple sites, which include assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions.

Carrell et al. suggest a variety of mitigation strategies, such as heuristic record linkage, acquisition of local knowledge, active learning, and tailoring with machine learning. In contrast, Demner-Fushman and Elhadad suggest sharing patterns for simple tasks and “more work on porting pipelines with easy domain adaptation”. These two strategies may be broadly contrasted as seeking hard-to-reach fruit (which may turn out to be sour grapes for some institutions) or low-hanging fruit, respectively. This paper investigates which factors might allow one to pursue the latter approach as a practical strategy for NLP portability.

Our project was motivated by the New York City Clinical Data Research Network (CDRN), a collaboration among six academic medical centers in the metropolitan area, seeking to collect and integrate clinical data to support patient-centered clinical research. The CDRN needed an approach that would leverage existing NLP resources at specific sites, while enabling sharing of resources across sites. The first main consideration was to select a system architecture for NLP based on standards, which has become a crucial strategy to facilitate portability and scalability. The second main consideration was to select a task with potential to be replicated across all sites. We chose left ventricular ejection fraction (LVEF), a primary diagnostic measurement of heart failure. LVEF is the ratio of the
volume of blood ejected during systole to blood volume in the ventricle at the end of diastole. LVEF is typically measured by echocardiography and recorded in narrative text. A number of previous studies have shown success in extracting LVEF from clinical documents\textsuperscript{16-19}.

Based on these factors, we chose to work with a system architecture called Leo, which was developed by the Department of Veterans Affairs (VA) Informatics and Computing Infrastructure (VINCI)\textsuperscript{20}. Leo is a set of libraries that facilitate rapid development and scalable deployment of NLP systems, and builds upon the Apache Unstructured Information Management Architecture Asynchronous Scaleout (UIMA AS)\textsuperscript{21}. In particular, this study focuses on a specific instance of Leo named Ejection Fraction Extractor (EFEx)\textsuperscript{22}.

VINCI developed EFEx to extract LVEF values from clinical documents that originate at various centers within the VA\textsuperscript{23}. These studies were conducted entirely on VA documents, which raises a question about generalizability outside the VA system. However, with over 1,700 points of care and thousands of clinical authors, the VA system provides an exceptional data source for system training. Therefore, we expected that EFEx would be a better candidate for portability than a tool developed using data from a single medical center. This report details the initiative that we undertook to install and configure EFEx at Weill Cornell Medicine, and to extract LVEF from echocardiogram reports available in the MIMIC-III database.

Methods

Data source
We obtained echocardiograms from the Medical Information Mart for Intensive Care III (MIMIC-III) database\textsuperscript{24}. MIMIC is an openly available database developed by the MIT Lab for Computational Physiology. The latest version, MIMIC-III contains de-identified patient records for >40,000 critical care patients between 2001 and 2012. Researchers wishing to use the data must accept the data use agreement and provide evidence of completion of appropriate human subject research training. The MIT research team de-identified the data according to Health Insurance Portability and Accountability Act Privacy Rules, which included random date shifting, which preserves temporal relationships within a given patient but not across patients.

We extracted 8707 echocardiogram reports from the NOTEEVENTS table by selecting CATEGORY field for ‘Echo’. The table was filtered to only the first echocardiogram report, in chronological order, for each unique hospital admission (coded in MIMIC by HADM_ID). In this study we restricted our data corpus to a single document type of echocardiograms (low-hanging fruit) and originated at an independent source, which in this case is Beth Israel Deaconess Medical Center. One consideration behind such a selection was to investigate the effectiveness of EFEx on documents originated at an independent source. This is important if we want to eventually deploy EFEx at other CDRN centers while maintaining same level of performance. MIMIC is an open source de-identified dataset not subjected to institutional review board approval. The performance of EFEx on echocardiograms originated within Weill Cornell Medicine is an ongoing study and will be the subject of a future report.

System description
Leo follows the model of UIMA AS with client and service components, along with an additional core library. The client defines inputs and outputs for processing and sends requests to the services. Setting up a client consists of selecting the required collection reader and listener, which could be a database or a local file system. The core contains tools that have been developed in conjunction with Leo to facilitate various NLP and annotation needs. The service component contains the server functionality for launching UIMA AS services. The service component also defines the type system and annotators as a pipeline architecture that implements all the logic necessary to extract a target information from unstructured documents. The basic architecture of Leo is shown in Figure 1, with the flow beginning at the reader.
Leo is built using the Java language and requires the Java runtime environment and the Apache package manager Maven, and can be installed on Windows, Linux, or Mac. We installed instances of EFEx running on Linux and Mac environments, and the setup procedure was essentially identical. The following steps were performed to create a fully functional EFEx instance.

1. We installed Java SDK 8 on our machines and set up an environment variable JAVA_HOME pointing to the JDK bin location, and added this to the PATH environment variable.
2. We installed Maven 3.3.9 and setup an environment variable MAVEN_HOME pointing to maven bin location, and added this to the PATH variable.
3. We downloaded UIMA-AS (http://uima.apache.org/downloads.cgi) and extracted the content to a suitable folder. We installed UIMA version 2.6.0 (uima-as-2.6.0-source-release.zip), and followed the instructions to compile and package the UIMA-AS. We set the UIMA_HOME environmental variable pointing to UIMA-AS root folder and added the bin folder to PATH variable.

The distribution package for EFEx was made available through a VA github repository\(^{25}\). Installation of EFEx mainly involved downloading and extracting the content to a folder location on the machine. As part of the configuration setup, we created a folder called amq-broker under the uima-as folder and provided write permission to this folder. This folder is required for the broker service to copy all its configuration settings. The entire installation and basic configuration was completed in one day at WCM. However, the overall installation time may vary depending on technical skills available at individual centers.

Reference standard
The reference standard was developed at NYU Langone Medical Center. At WCM, all values were further confirmed through manual review of the entire document collection. Two reviewers examined each document on Excel spreadsheet. They were given training based on previously defined guidelines. These guidelines included identifying all mentions of LVEF and the associated quantitative values. If there were differences between the two reviewers’ findings, a third reviewer serving as adjudicator resolved the discrepancy. The reviewers also confirmed all documents that did not have any mention of LVEF information. We identified two values, EFmin and EFmax, corresponding to the lowest and the highest values of LVEF for each document in the dataset. The reviewer identified numeric values and ranges of LVEF (e.g. 55, 50-70), as well as severity-based descriptors such as normal, mild, moderate, and severe. The majority of reports had either a numerical value or a range of values. In documents that contain multiple instances of LVEF, we employed the following logic for determining the reference values:

1. A LVEF instance in the conclusion part, normally at the end of the document, takes precedence over one in the finding section.
2. A LVEF mention in the postoperative section takes precedence over one in finding or conclusion sections. (The postoperative section always follows the conclusion section in the document.)
In some reports, the LVEF value is expressed using a greater than or less than symbol (e.g. \(LVEF > 55\)). In these cases the reviewer extracted the value ignoring the symbol. Some echo reports express uncertainty about the LVEF value using a question mark (e.g. \(LVEF ? 55-70\)). In such cases, the reviewer extracted the value, provided that there was no other instance mentioned elsewhere in the document.

In reports where there was no quantitative value for LVEF available, we assigned a numerical value or a range of values using other information. LVEF concept synonyms were identified, including ‘lvef’, ‘left ventricular’, ‘LV’, ‘ejection fraction’, and modifiers were defined, such as ‘depressed’, ‘impaired’, ‘systolic dysfunction’ etc. When a concept was preceded or followed by modifiers to the severity level, such as mildly depressed, moderately depressed, or severely depressed, a quantitative value was assigned. Table 1 shows examples of modifiers and the corresponding values assigned. Despite using this mapping scheme, there were still documents with no concept-value pair identified in the reference standard. In general, these documents did not mention LVEF, or it was not possible to assign any meaningful value from the available information.

**Table 1.** Values assigned for concepts of ejection fraction with qualitative modifiers in developing the reference standard.

| Modifiers               | Values |
|-------------------------|--------|
| Severely depressed      | 5-29   |
| Moderately depressed    | 30-44  |
| Mildly depressed        | 45-54  |
| Grossly preserved       | 50-55  |
| Normal                  | 70     |

While developing the reference standard, the context as well as the overall content was taken into consideration in assigning a value or range of values to a concept. For example, there could be instances of LVEF expressed as quantitative values as well as qualitative descriptors, such as when the phrase ‘normal global systolic function’ was mentioned along with ‘severe regional left ventricular systolic function’ and ‘EF 20-25%’. In such cases, the numerical value took precedence over qualitative descriptors. Another example, when a document contained the phrases ‘moderately depressed LVEF’ and ‘(LVEF=30%)’ as well as ‘LVEF 70% previously, now 30%.’ In this case, the 30% is taken as the value for EFmin.

**Extraction methodology**

Patterson, et al. has described the logic for concept extraction employed in the present study in detail. EFEx is a rule-based system that identifies the set of core concepts for LVEF using regular expressions, pattern matching, and filters. Because of the ambiguous nature of some of the concepts (such as ‘function’), the preceding text to each mention of the concept was used as a filter. Quantitative values were found using number patterns, but allowed for with or without modifiers such as ‘\(=\)’, ‘\(\leq\)’, ‘\(>\)’, ‘\(<\)’, ‘\(\geq\)’, ‘\(\leq\)’, and ranges of values.

Figure 2 shows the overall logic that was implemented in finding the concept-value pairs of LVEF. Steps A through K are used to extract concept-value pairs, if there is one found in the document. For those cases when no concept-value pair is identified through steps A to K, functionality was added to the original EFEx to look for qualitative modifiers used to describe the LVEF concept. This extended logic was implemented through steps M through O and effectively simulates the mapping scheme adopted in creating the reference standard. We identified 100 reports that were previously shown to have no output when processed by EFEx and used these as a training set for developing the extended logic. The output from the training set was manually reviewed to adjust the regular expression patterns through an iterative process.

**Data analysis**

We analyzed the current results on MIMIC-III data in two ways. In the first case, we analyzed data using the extraction logic implemented in the original EFEx for LVEF. This instance that was ported from VA has extraction logic implemented only through steps A to K as described in Figure 2. We refer this version as **Original EFEx**. Upon analyzing results on MIMIC data, we observed that the **Original EFEx** missed a significant number of documents where EF concept is described in a qualitative manner without any numerical value assigned. So at
WCM we further extended the extraction methodology by implementing additional logics to discover EF concept-value pair based on qualitative assessment through a mapping scheme. The extended logic implemented as steps M through O in Figure 2 improved the performance of EFEx significantly. The algorithm searched for both numeric values, ranges of LVEF (e.g., 55, 50-70) and severity-based descriptors based around the clinically relevant normal, mild, moderate, and severe labels. We refer this version as Extended EFEx. Performance measures were then calculated in both Original EFEx and Extended EFEx instances on the entire documents.

Figure 2. Extraction logic for LVEF implemented in EFEx system.

The results of EFEx output were tabulated and compared against the reference standard. Each document was classified as one of four possible cases: true positive (document had an LVEF mention and EFEx identified the concept-value pair and matched with the value given in the reference standard); false positive (document had no LVEF mention as given by a null value in reference standard, but EFEx produced a non-null concept-value pair); true negative (document had no LVEF concept-value mention as given by a null value in the reference standard, and EFEx did not find any concept-value pair); and false negative (document had an LVEF concept-value mention as given by a non-null value in the reference standard, but EFEx did not identify a concept-value pair, or the value extracted did not match with the corresponding reference standard value).

When there were multiple instances of concept-value pair extracted by EFEx, we used the following heuristic measures to select a given instance of LVEF in order to compare directly with the reference standard. We either selected the last one, normally in the conclusion part of the report (time usually moves forward in the report), or the lowest value (the disease typically worsens). The total outcomes of the four cases were then used to calculate various statistical performance measures. These included precision (positive predictive value), recall (sensitivity or
Results

There were 8707 documents for analysis. Using the Original EFEx, we classified each document as one of four cases, for the purpose of calculating performance measures: true positive (6568), true negative (1124), false positive (0), and false negative (1015). These values resulted an overall accuracy of 88.3% (95% CI 87.7% - 88.9%), sensitivity of 86.6% (95% CI 85.8% - 87.4%), specificity of 100% (95% CI 99.6% - 100%), positive predictive value 100% (95% CI 99.9% - 100%), and an F-score of 92.8%. Percentage of severely and moderately depressed ejection fraction (LVEF < 45) cases is calculated to be 10.7%.

Using the Extended EFEx, we classified documents as true positive (7541), true negative (1026), false positive (98), and false negative (42). These values resulted an overall accuracy of 98.4% (95% CI 97.8% - 98.8%), sensitivity of 99.4% (95% CI 99.2% - 99.6%), specificity of 91.3% (95% CI 89.4% – 92.8%), positive predictive value 98.7% (95% CI 98.4% - 99.0%), and an F-score of 99.0. We observed an increased percentage of severely and moderately depressed cases for ejection fraction (18.1%).

Discussion

This experience illustrates how ejection fraction is an excellent example of ‘low hanging fruit’: a simple potential application for NLP that is relatively easily portable to new clinical settings. One limitation of this study is that the task of identifying LVEF measurements is relatively simple, with low variability of expressions and values to extract. In addition, the study examined only one document type. However, the relative simplicity of the task does not mean it is not important: ready availability of this important quantitative parameter has important implications for research, quality improvement, and clinical care.

The EFEx development team reported that the system achieved 98% positive predictive value and 93% sensitivity at the instance level across all medical centers across all VA

22. Garvin, et al., has developed an NLP system based on the UIMA architecture for extracting LVEF values from echocardiograms that are generated at four centers within the VA

23. They have reported for document-level classification of EF of <40% had a sensitivity of 98.41%, a specificity of 100%, a positive predictive value of 100%, and an F-score of 99.2%. Also system test results at a concept level it was reported a sensitivity of 88.9%, a positive predictive value of 95%, and an F-score of 91.9%. It should be noted that the discovery logic that was developed in that study is not the one implemented in the present EFEx system, although they both share some common features. The present results on the MIMIC-III dataset show a comparable overall performance when analyzed with EFEx without the qualitative concept mapping. The results on the Extended EFEx showed an improved performance matching the document level values reported above.

Recently, Nath et al. has reported an NLP tool named EchoInfer for large-scale data extraction from echocardiography reports at a single medical center

18. They have reported a recall 95.99% and a precision > 96% for LVEF. When compared to the performance of EchoInfer the result obtained from the EFEx system shows a slightly lower performance, when using EFEx without any concept-mapping scheme in its discovery logic on MIMIC data. However, with the concept-mapping scheme, the performance of EFEx improved significantly and the values are slightly better than the EchoInfer reported above. For the entire dataset, the Original EFEx classified 1015 documents as false negative. However, with Extended EFEx, we observed only 42 false negative cases. Some of the documents where Extended EFEx failed to identify a value for LVEF are one in which both left and right ventricle is mentioned together in one statement. Similarly, for documents in which the target concept value immediately followed by a different numeral (e.g. a list index number), our extraction logic failed to identify the correct value for EF. The adoption of the mapping scheme significantly improved the identification of severely and moderately depressed cases in the dataset. With Extended EFEx, we observed a 40% increase in the number of cases with LVEF < 45, which were confirmed by the manual review. This substantial increase further supports the effectiveness of the extended logic that was implemented, as these additional cases would not have been discovered using the original logic alone.
On the flip side, implementation of the extended logic introduced several false positive cases for EF. While no false positive case was observed with Original EFEx, 98 false positive cases were observed on Extended EFEx. The mapping scheme we implemented does not assign values for cases such as mild to moderate depressed, moderate to severely depressed, borderline depressed, or more depressed. For documents with these statements, EFEx assigns incorrect values for EF. Similarly, a statement such as Preserved LVEF (effective forward LVEF may be depressed given the severity of valvular regurgitation) (HADM_ID = 182611) is subject to interpretation and no value is given in the reference standard. Our extended logic assigned a value of 45. Typical of most NLP systems, there is room for further improvement in the extraction logic as evident from some of the false positive and false negative cases observed with EFEx.

Recent papers have identified a number of challenges facing NLP portability\textsuperscript{5-11}, such as assembling clinical corpora, managing diverse document structures and handling idiosyncratic linguistic expressions. An additional challenge arises when using standardized NLP architectures such as UIMA, especially when integrating multiple NLP modules\textsuperscript{13-15}. A final challenge not identified by these papers involves leadership of dissemination project. In general, the vast amount of dissemination of informatics technology has been a push from a small number of innovators (“benchmark institutions”) to adopters rather than a pull from the adopter\textsuperscript{26}.

The five challenges for NLP portability are summarized in Table 2, along with strategies for mitigating the challenges described in the cited literature. The strategies can be roughly partitioned into technologically advanced methods addressing complex NLP tasks (striving for the hard-to-reach fruit), and more practical methods addressing simpler NLP tasks (settling for the low-hanging fruit). Focusing on target concepts with low sensitivity to document location is found to be a good practical strategy for the portability of NLP tools. Our own experience showed that simple concepts where the associated values follow a general convention or prescribed format are good candidates for Leo. At WCM, our ongoing development effort resulted in other instances of Leo were we used this strategy effectively. We had achieved high performance in extracting PHQ-9 score from encounter notes. Similarly, we achieved high performance in extracting TNM stages, Gleason score and ICD-9/10 diagnosis codes from surgical pathology reports. In these cases, the precision and recall of Leo instances were sufficiently high enough, and we are currently in the process of making these data available in the i2b2 instance at WCM.

Table 2. NLP portability challenges, and mitigation strategies that require advanced methods (hard-to-reach), and more practical methods (low-hanging).

| Challenge                                      | Strategy: Hard-to-Reach                                      | Strategy: Low-Hanging                                    |
|------------------------------------------------|-------------------------------------------------------------|----------------------------------------------------------|
| Assemble corpora with heterogeneous document types | Use heuristic linkage methods; develop document classifiers | Exploit metadata; focus on single document type          |
| Navigate diverse report structures             | Customize document segmentation algorithms; employ active learning | Select pattern with low sensitivity to document location |
| Analyze idiosyncratic linguistic expressions   | Use machine learning to tailor complex patterns              | Re-use or adapt simple patterns developed previously     |
| Integrate multiple NLP modules                 | Employ large number of modules; adapt to meet architecture standards | Employ small number of modules; re-use use or adapt modules previously standardized |
| Lead the dissemination project                 | Acquire funding to support the innovator site; supply expertise in NLP methods | Draw on existing resources at the adopter site; use conventional software skills |

Conclusion

We extracted LVEF information from echocardiogram reports from the MIMIC-III database using the EFEx NLP system. We compared the results to a reference standard developed manually by human reviewers. EFEx in its original version showed lower performance compared to the performance reported on VA documents that are different in document formats and content. However, when the extraction logic was modified to include a concept-
value mapping scheme similar to the mapping scheme used in developing the reference standard, EFEx had an accuracy of 98.4%, sensitivity of 99.4%, a positive predictive value of 98.7%, and an F-score of 99.0%. These values match reasonably well with that reported earlier on VA generated echocardiograms. The extended extraction logic also improved the discovery of cases having severely or moderately depressed LVEF by 40%. The current study on the LVEF extraction from the MIMIC dataset suggests that the EFEx performance varies depending on documents that are originated at different clinical settings.

The project described in this paper pursues a practical strategy to pursue a relatively simple NLP task (low-hanging fruit). We exploited database metadata to focus on single document type (cardiology reports). We chose a pattern with low sensitivity to document location (we used the last occurrence of LVEF). We adapted simple rule based extraction logic, and a specific instance (EFEx) of a NLP system (Leo) previously developed by the VA. The adopter (WCM) led the dissemination project, drawing on existing resources, and employing conventional software skills. This case study provides evidence that an NLP system can be ported successfully from one institution to another, enable customization to a new data source, and achieve comparable performance. The identification of practical strategies for NLP portability has paved the way for sharing NLP tools among the multiple institutions in the NYC CDRN, and may provide useful guidance for other institutions interested in pursuing a similar approach.

References

1. 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.
2. 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.
3. 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.
4. 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.
5. 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.
6. Huang CC, Lu Z. Community challenges in biomedical text mining over 10 years: success, failure and the future. Brief Bioinform. 2016 Jan;17(1):132-44. doi: 10.1093/bib/bbv024. Epub 2015 May 1. Review. PubMed PMID: 25935162; PubMed Central PMCID: PMC4719069.
7. 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.
8. 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.
9. 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.
10. Zheng K, Vydisswaran 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.
11. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, Crockett SD, Gourevitch RA, Dean KM, Mehrotra A. Challenges in adapting 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/oxx039. PubMed PMID: 28419261.

12. Kaushal R, Hripcsak G, Ascheid DM, Bloom T, Campion TR Jr, Caplan AL, Currie BP, Check T, Deland EL, Gourevitch MN, Hart R, Horowitz CR, Kastenbaum I, Levin AA, Low AF, Meissner P, Mirhaji P, Pincus HA, Scaglione C, Shelley D, Tobin JN; NYC-CDRN. Changing the research landscape: the New York City Clinical Data Research Network. J Am Med Inform Assoc. 2014 Jul-Aug;21(4):587-90. doi: 10.1136/amiajn-2014-002764. PubMed PMID: 24821739; PubMed Central PMCID: PMC4078297.

13. Doan S, Conway M, Phuong TM, Ohno-Machado L. Natural language processing in biomedicine: a unified system architecture overview. Methods Mol Biol. 2014;1168:275-94. doi: 10.1007/978-1-4939-0847-9_16. Review. PubMed PMID: 24870142.

14. Divita G, Carter M, Redd A, Zeng Q, Gupta K, Trautner B, Samore M, Gundlapalli A. Scaling-up NLP Pipelines to Process Large Corpora of Clinical Notes. Methods Inf Med. 2015;54(6):548-52. doi: 10.3414/ME14-02-0018. Epub 2015 Nov 4. PubMed PMID: 26534722.

15. Chute CG, Pathak J, Savova GK, Bailey KR, Schor MI, Hart LA, Beebe CE, Huff SM. The SHARPn project on secondary use of Electronic Medical Record data: progress, plans, and possibilities. AMIA Annu Symp Proc. 2011;2011:248-56. Epub 2011 Oct 22. PubMed PMID: 22195076; PubMed Central PMCID: PMC3243296.

16. Kim Y, Garvin JH, Goldstein MK, Hwang TS, Redd A, Bolton D, et al. Extraction of Left Ventricular Ejection Fraction Information from Various Types of Clinical Reports. J Biomed Inform. 2017;

17. Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc [Internet]. 2016 Jul 12; Available from: http://www.ncbi.nlm.nih.gov/pubmed/27413122

18. Nath C, Albaghdadi MS, Jonnalagadda SR (2016) A natural language processing tool for large-scale data extraction from echocardiography reports. PLoS ONE 11(4): e0153749. doi:10.1371/journal.pone.0153749

19. Chung J, Murphy S. Concept-value pair extraction from semi-structured clinical narrative: a case study using echocardiogram reports. AMIA Annu Symp Proc 2005:131e5

20. Cornia R, Patterson O V, Ginter T, DuVall SL. Rapid NLP Development with Leo. In: AMIA Annu Symp Proc. 2014. p. 1356.

21. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Natural Language Engineering (2004);10(3-4):327–348. doi: 10.1017/S1351324904003523.

22. Patterson OV, Freiberg MS, Brandt C, DuVall SL. Unlocking echocardiogram report measures for heart disease research through natural language processing. BMC Cardiovasc Disord. 2016. Submitted.

23. Garvin JH, DuVall SL, South BR, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in information management architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012; 19: 859–866.

24. Johnson AEW, Pollard TJ, Shen L, Lehman L-WH, Feng M, Ghassemi M, et al. MIMIC-III, a freely accessible critical care database. Sci data [Internet]. 2016 May 24;3:160035. Available from: http://www.nature.com/articles/sdata201635%5Cnpapers3://publication/doi/10.1038/sdata.2016.35

25. The Department of Veterans Affairs. EFEx https://github.com/department-of-veterans-affairs/efex (Accessed 6 October 2016)

26. Chaudhry B, Wang J, Wu S, Maglione M, Mojica W, Roth E, Morton SC, Shekelle PG. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. Ann Intern Med. 2006 May 16;144(10):742-52. Epub 2006 Apr 11. Review. PubMed PMID: 16702590.