Published in final edited form as:

J Biomed Inform. 2017 September ; 73: 14–29. doi:10.1016/j.jbi.2017.07.012.

Natural language processing systems for capturing and standardizing unstructured clinical information: A systematic review

Kory Kreimeyer[a,*], **Matthew Foster**[a], **Abhishek Pandey**[a], **Nina Arya**[a], **Gwendolyn Halford**[b], **Sandra F Jones**[c], **Richard Forshee**[a], **Mark Walderhaug**[a], **Taxiarchis Botsis**[a]

[a]Office of Biostatistics and Epidemiology, Center for Biologics Evaluation and Research, US Food and Drug Administration, Silver Spring, MD, United States

[b]FDA Library, US Food and Drug Administration, Silver Spring, MD, United States

[c]Cancer Surveillance Branch, Division of Cancer Prevention and Control, National Center for Chronic Disease Prevention and Health Promotion, Centers for Disease Control and Prevention, Atlanta, GA, United States

Abstract

We followed a systematic approach based on the Preferred Reporting Items for Systematic Reviews and Meta-Analyses to identify existing clinical natural language processing (NLP) systems that generate structured information from unstructured free text. Seven literature databases were searched with a query combining the concepts of natural language processing and structured data capture. Two reviewers screened all records for relevance during two screening phases, and information about clinical NLP systems was collected from the final set of papers. A total of 7149 records (after removing duplicates) were retrieved and screened, and 86 were determined to fit the review criteria. These papers contained information about 71 different clinical NLP systems, which were then analyzed. The NLP systems address a wide variety of important clinical and research tasks. Certain tasks are well addressed by the existing systems, while others remain as open challenges that only a small number of systems attempt, such as extraction of temporal information or normalization of concepts to standard terminologies. This review has identified many NLP systems capable of processing clinical free text and generating structured output, and the information collected and evaluated here will be important for prioritizing development of new approaches for clinical NLP.

[*]Corresponding author at: Office of Biostatistics & Epidemiology | Center for, Biologics Evaluation and Research | FDA, 10903 New Hampshire Ave, Bldg 71 Rm, 1309A, Silver Spring, MD 20993-0002, United States. Kory.Kreimeyer@fda.hhs.gov (K. Kreimeyer).

5. Competing interests

None.

Appendix A

See Table A1.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/jjbi.2017.07.012.

Keywords

Review; Systematic; Natural language processing; Common data elements

1. Introduction

Important clinical information is often recorded in unstructured free text, and converting it to a structured format can be a time-consuming task that may not successfully capture all facets of the information. However, there are at least two large incentives for translating unstructured data into structured data: i) the reduction of time required for manual expert review and ii) the secondary use of these data for large scale automated processing. The former goal is an obvious benefit for anyone involved in clinical practice, where physicians and other experts examine patients' electronic health records (EHRs) on a regular basis and spend considerable time reading free text. Safety reviewers at the US Food and Drug Administration (FDA) who read large numbers of narratives from adverse event reports for the medical products they _regulate and any practitioner who tries to keep up to date on the medical literature in their field are two more examples of potential beneficiaries of a robust free text structuring process. The second important gain from the creation of structured data is in the ability to manage and mine clinical data in large volumes or across large time scales. This is vital for implementing algorithms to define patients at risk of certain diseases, eligible for certain clinical trials, or who fit the case definitions of certain diseases, to name a few.

Given the rate at which unstructured clinical information is created, it is clear that automated solutions utilizing Natural Language Processing (NLP) are needed to analyze this text and generate structured representations. However, clinical text possesses several properties (e.g. poor structure, abundant shorthand, domain-specific vocabularies) that make the application of NLP challenging. Current NLP systems have proven to be useful for certain activities and have, for example, reduced the time required for screening candidates for clinical trial eligibility [1] and identified potential adverse drug reactions [2]. There are, however, other challenges in the field, such as identification of temporal associations, evaluation of context-dependent text, and concept normalization to particular terminologies, that remain open [3–7].

The FDA and the Centers for Disease Control and Prevention (CDC) recently launched a collaborative effort for the "Development of a Natural Language Processing (NLP) Web Service for Structuring and Standardizing Unstructured Clinical Information". This project aims to create a NLP Platform for clinical text (initially cancer data and safety data) that will be extensible for many different subdomains [8]. The overall plan is to perform the necessary development for maximizing the use of existing tools and filling certain gaps, e.g. when there are no efficient solutions for real-life tasks. This project will initially focus on improving the efficiency of clinically relevant use cases involving the structuring and coding of unstructured information for the two domains.

As the first step in this NLP development project, we have conducted a systematic literature review to identify the existing NLP solutions that may support our project objectives. This

review is intended to compile a list of currently-in-use, complete NLP solutions for clinical text that are capable of encoding free text into standardized clinical terminologies and of capturing common data elements to fill specified forms or templates, a process known as structured data capture (SDC) [9]. This is a fairly specific topic, but it forms the core of some of the major needs in the biomedical field, especially as they relate to utilization of free-text information. We will first describe the methodology of the review and then present the information gathered about existing NLP solutions. We will close with a discussion of the remaining open challenges in the field and the next steps for the development of the clinical NLP Platform.

2. Methods

We based our review procedure on meeting the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) recommendations for reporting in systematic reviews, which provide a description of the components involved in a review and best practices for performing the work and publishing the results [10]. After reviewing the PRISMA guidelines, we structured our review into 4 phases:

1. Publication Retrieval
2. Review of Titles and Abstracts
3. Review of Full Text
4. System Information Collection

In the initial publication retrieval phase, we looked for publications related to NLP and SDC and meeting several filtering criteria:

- Published in English
- Published from 2006–01-01 to the time of the search(2016–06–15)
- Peer-reviewed

During the two review phases, we examined each publication to find the ones that were relevant to our exploration while attempting to minimize information loss. In particular, we defined the following inclusion criteria:

- Published and existing NLP algorithms, methods, or tools for the processing of clinical data—including EHRs, clinical trial summaries, post-market reports, medical product labels, and medical literature—used by federal agencies, public health agencies, academic centers, commercial vendors, or the National Patient-Centered Clinical Research Network (PCORnet) participants.[1]
- Activities related to SDC and Common Data Elements (CDEs) that are in use by federal agencies, public health agencies, academic centers, commercial vendors, or PCORnet participants and involve some NLP or text mining process.

[1]The identification of the PCORnet efforts was a requirement included in the FDA and the CDC proposal.

- Complete package solutions that convert clinical texts to structured and/or coded data.

We also used the following exclusion criteria:

- Text mining approaches that perform limited tasks, such as keyword extraction for topic modeling or document indexing.
- SDC systems that utilize existing humanly assigned codes only and do not process actual text using either a NLP or text mining process.
- Systems that require manual processing or preprocessing.

In the final phase, we collected information about the NLP tools described in the relevant papers and compiled the results.

We used the EPPI-Reviewer (version 4) software [11] produced by the EPPI-Centre at University College London to manage many of the screening and coding processes of our systematic review process. Its features broadly matched our desired workflow. In particular, EPPI-Reviewer allows comparison screening with two or more reviewers and an adjudicator, contains many sorting, filtering, and reporting functionalities for use after screening, and supports information collection by allowing codes to be applied to specific sections of text. We also made use of the EndNote X7 reference management software [12] from Thomson Reuters to store and organize citations at several stages throughout the review.

For the publication retrieval phase, we coordinated with a reference librarian at the FDA Library experienced in systematic reviews (GH) to expand and implement our search query. In June 2016, a structured search was conducted on NLP systems in several databases: PubMed, Excerpta Medica dataBASE (Embase), Cumulative Index to Nursing and Allied Health Literature (CINAHL), International Pharmaceutical Abstracts, Web of Science, and Science Direct. The query was constructed using keywords from the Medical Subject Headings (PubMed) and Emtree (Embase) thesauri related to the concepts of NLP and SDC. Each concept was searched separately, and the two concepts were combined using the "AND" operator to form the following query:

> natural language processing ***OR*** text mining ***OR*** data mining ***OR*** datamining ***OR*** information storage and retrieval ***OR*** information retrieval
>
> AND
>
> common data element ***OR*** common data elements ***OR*** common data item ***OR*** common data items ***OR*** common data format ***OR*** common data formats ***OR*** common data variable ***OR*** common data variables ***OR*** common data model ***OR*** common data models ***OR*** structured data capture ***OR*** structured data collection ***OR*** structured data collections

The citations were first compiled in EndNote, and the "Find Duplicates" function, which allows comparison of all fields for suspected duplicates, was used to review and delete duplicates. Some manual deletion was also performed by the reference librarian (GH).

In addition, the Department of Defense's Defense Technical Information Center online resource, "Research & Engineering Gateway," was searched for technical reports available for public release using the same query. These records were subsequently imported into the same EndNote library file containing the citations from the other databases and deduplicated.

The entire list of records was then uploaded to EPPI-Reviewer 4. All citations from all search databases were run through the duplicate search in EPPI-Reviewer 4, and suspected duplicates were manually reviewed by one of the authors (KK) based on title, authors, date, and journal, but not abstract.

ln the second phase, records were screened for the inclusion and exclusion criteria based on their titles and abstracts. Two reviewers (MF, AP) independently reviewed the full set of records, and an adjudicator (KK) resolved the disagreements.

An initial set of 1% of the entries was screened to test the design and applicability of the inclusion and exclusion criteria in EPPI-reviewer before moving on with the remainder of the screening. The records were then randomly segmented into 12 screening sets for ease of management. The two reviewers performed the initial screening and assigned codes to the records from the following list:

- Exclude as not text mining
- Exclude as not clinical
- Exclude for limited text mining
- Exclude for human codes
- Exclude for manual steps
- Exclude as not in use
- Include for now – Uncertain
- Include on title & abstract

The *Include for now – Uncertain* code was applied to records when the reviewer could not make a clear determination of whether the publication matched the inclusion or exclusion criteria based on the limited information available in the title and abstract. This code allowed us to reduce information loss by retaining potentially interesting records even if they did not contain pertinent details in their abstracts. The two reviewers generally selected only a single exclusion reason—the one that was most applicable—however, for some records they applied two or more reasons. All records for which the two reviewers disagreed with an *Exclude* and an *Include* code, or where both reviewers assigned the *Include for now – Uncertain* code, were adjudicated. For disagreements on the particular code but not on the include vs exclude judgment, the final code was chosen from one of the reviewers' selected codes in an alternating fashion. The agreement between the reviewers was measured using the direct percentage of the number of agreements divided by the total number of screened records. This was done for both exact agreement, in which both reviewers had to assign the exact same code(s), and for agreement only on the main decision of include or exclude.

Following the first screening, we used the EndNote "Find Full Text" feature to obtain full text PDF files for as many of the included records as possible. To obtain the rest, we used EPPI-Reviewer's "Find" function to perform a Google search for the title and author list. Additional manual searching was required for some records. The full texts that could not be found or could not be obtained because of access restrictions were then requested through the FDA Library.

For any records of conference proceedings, which represented a large collection of conference papers, we did not examine the full text of every paper. Instead, we briefly scanned the titles to determine if any were relevant to our review.

The third phase of the review was screening based on review of the full text of the publications. Each paper's *Methodology* section was reviewed to judge if it fit the review criteria. Additional sections of the paper were also read when necessary to make a clear decision. We selected a very similar methodology for this phase as in the previous screening phase: the same two reviewers independently reviewed the full text for all records, and the same adjudicator resolved disagreements between them. The overall inclusion and exclusion criteria were the same for this phase. The screening codes in this phase were nearly the same as the previous screening phase: the exclusion reasons were unchanged, but the *Include for now – Uncertain* code was removed. *Include* and *Exclude* disagreements were adjudicated. The records to be screened were separated into four bins alphabetically by author last name for ease of processing.

The fourth and final phase involved collecting information about the NLP system(s) described in each paper in the final set of records that satisfied the inclusion criteria. We defined seven broad categories of information about NLP systems and multiple items within each category to support the generation of a complete system description. The categories were:

- System Characteristics: General information about the system itself, the system functionalities, and the NLP approach used.
- System Development Cycle: Information about the development process (leaders, timeline, versions) of the system.
- System Input: Information about text sources that can be processed by the system.
- System Output: Information about the results generated by the system after processing clinical text.
- System Evaluation: Information about how well the system has performed while being tested and validated, especially in any external validations.
- System Use: Information about projects/tasks that the system has been used for.
- System Availability: Information about obtaining access to the system and its source code.

We decided to review the full text PDF files and annotate sections of text using seven different codes for these broad information categories. Then, the individual information points were collected from these selected pieces of text. EPPI-Reviewer broadly supports this workflow by allowing for codes to be applied to specific text sections within PDF documents.

The two reviewers who performed both rounds of screening also performed the initial annotation. They highlighted specific sentences or paragraphs from the full text of the papers, and marked them with codes for one or more of the seven categories of clinical NLP system information. Both reviewers independently annotated the full set of papers that were included in this collection phase. Then, the adjudicator read the annotated portions of text using the report generated by EPPI-Reviewer containing all the highlighted text and recorded all the information that could be gathered about the system(s) in a Microsoft Access database.

Additionally, citation counts for the final set of papers were searched in Web of Science or, if not found, in Google Scholar. The number of citing papers was totaled by year from 2006 to 2016. Citation counts were obtained on 2016–11–03.

3. Results

In the case of Embase, the original query generated a set of results with a high irrelevancy rate and failed to capture many pertinent papers. To remedy these problems, keywords were enclosed in quotes to generate a more precise result set. In addition, the list of citations from the Research & Engineering Gateway could only be obtained in the BibTeX file format, which is not natively supported in EndNote. Instead, the JabRef reference manager [13] (version 3.4) was used to convert the BibTeX citations to the RIS format that is supported by EndNote.

The results of the searches in Phase 1 (Publication Retrieval) across all databases showed that the largest number of records was obtained from PubMed (2949), followed by Web of Science (2893), the Research & Engineering Gateway (1737), Embase (37), and CINAHL (16). Neither Science Direct nor International Pharmaceutical Abstracts returned any results for the applied query. Following deduplication, the remaining 7149 records were moved to Phase 2 (Review of Titles and Abstracts) for *Include/Exclude* coding. The results of coding following adjudication are shown in Fig. 1. The most frequently assigned code was *Exclude as not clinical* (4356 records), followed by *Exclude as not text mining* (2394), *Include on title & abstract* (132), *Exclude for limited text-mining* (118), *Include for now – Uncertain* (81), *Exclude for human codes* (76), and *Exclude for manual steps* (13). The sum of these categories exceeds 7149 because some records were given more than one exclude code.

During Phase 2, exact agreement of reviewers, based on the specific include or exclude code, was 71.8%. General agreement on *Include* versus *Exclude* was 93.7%. A total of 533 records required adjudication (of which 447 were adjudicated for *Include/Exclude* disagreements, 60 were adjudicated because both reviewers chose *Include For Now – Uncertain,* and 26 were

adjudicated due to a disagreement on exclusion codes in the initial criteria testing set of 1% of records).

Following Phase 2, the 213 records with an adjudicated code *of Include for now – Uncertain* or *Include on title & abstract* were moved to Phase 3 (Review of Full Text). Full text PDFs were retrieved for 136 of these records using the EndNote "Find Full Text" feature. Google searches returned 56 additional PDFs, and 21 records were requested through the FDA Library's Inter-Library Loan program. A total of 6 records were not found or could not be used in the following step. Four were conference proceedings covering dozens of papers, and 2 were conference posters that could not be obtained. We did not find any relevant paper titles from the conference proceedings listings after quickly reviewing the titles.

During the screening of Phase 3 (Review of Full Text), the 207 full text records that were retrieved were screened and assigned an *Include/Exclude* code. The results of coding after adjudication are shown in Fig. 2. The most frequently assigned code was *Include on full text* (86 records), followed by *Exclude for limited text mining* (68), *Exclude as not text mining* (29), *Exclude as not clinical* (14), *Exclude for manual steps* (8), and *Exclude as not in use* (2). Exact agreement of reviewers, based on the specific include or exclude code, was 47.8%. General agreement on *Include* versus *Exclude* was 63.3%. ln Phase 3 the *Include for now – Uncertain* code was removed, which contributed to lower agreement as compared to Phase 2 screening. A total of 76 records required adjudication. The review flowchart for Phases 1 through 3 is shown in Fig. 3.

The citations for the records can be found as Supplementary Material in three separate files in RIS format. These represent snapshots of the record review process at three different stages: i) at the start of Phase 2 (Review of Titles and Abstracts); ii) at the end of Phase 2 (Review of Titles and Abstracts); and iii) at the end of Phase 3 (Review of Full Text).

ln Phase 4 (System Information Collection), the 86 full text records that met the inclusion criteria were annotated and reviewed using the seven categories of system information. The information collected from the final 86 papers was grouped per system rather than per paper, yielding a list of 71 systems. A system name was identified for 49 out of the 71 systems.

During the system information collection, it was determined that a total of five papers did not actually include information about a system matching our criteria. Two of the papers focused on extracting biological entities like genes and proteins from text [14,15]. Two papers described systems that only performed indexing or word-counting of documents [16,17]. The final paper noted that several institutions in a large consortium were using NLP, but did not provide details about any specific system [18].

Since the information collection process was based only on the identified papers, there were gaps in the collected information about many systems. Certain data were much more difficult to gather than others, and some statements in papers were ambiguous. We focused only on clear and certain facts that could be gathered directly from the text, leading to high missingness for certain systems or for certain fields. In fact, for 6 systems (8.5%), we did not gather any information about the NLP approach used or whether it used rules or machine learning techniques. For 4 systems (5.6%), we found no information about the types of input

texts that the system had been applied to, and for 7 systems (9.86%), we did not identify the specific type or format of the output generated by the system. For 15 systems (21.1%), we found no clear information about how the system's performance had been evaluated. There were 49 systems (69.0%) lacking information about their licensing model or availability of their source code.

Fig. 4 shows the distribution of papers from the original queried set by publication year, as well as the percentage from each year that were included after each screening phase. There was an increase over time in both the total number of papers matching the query and the share of those papers that were judged to fit the review criteria.

Multiple papers in the final set had zero citations (N = 14), or between one and five citations (N = 30). A heatmap of the citations over time for the top ten most highly cited papers is shown in Fig. 5. Another heatmap showing all 81 papers (excluding five that did not contain a system of interest) is included as Supplementary Data 1.

The 71 systems identified within the review are presented in Table 1, along with the records they were located in. For 12 of the 14 systems with only brief mentions within the final set of papers, we have included an external citation that describes the system. We also provide a short description of each system and the results of any evaluations found for the system in Table A1 in Appendix A. Commonly reported metrics for performance include recall (or sensitivity), precision (or positive predictive value), F-Measure (harmonic mean of recall and precision), accuracy, and specificity. A few of the entries in the table are for NLP frameworks that can support different NLP pipelines by selecting or creating different components. Some evaluation results were identified for specific implementations using these frameworks.

We noted a number of trends based on the (sometimes limited) information collected from the final set of papers. Rule-based NLP approaches were the most common (N = 33), with hybrid systems also being strongly represented (N = 19); few systems used a purely machine learning approach to NLP (N = 4). The most common form of input text processed was clinical notes (N = 35), followed by radiology reports (N = 11), pathology reports (N = 9), biomedical literature (N = 7), and clinical trial documents (N = 5). The programming language for system development was frequently not found, but Java (N = 8) was the most common among those reported, with Perl (N = 3) and Python (N = 2) being the only other languages found more than once. Also, more systems were available in an open-source model (N = 12) than were unavailable (N = 7), although this information was not determined for a large number of systems. We also identified a number of systems using the Unstructured Information Management Architecture [19] (N = 7) and the General Architecture for Text Engineering [20,21] (N = 6) frameworks.

Finally, we recorded several additional observations about features and commonalities of the reviewed systems. A large number of systems in our review have leveraged prior work by incorporating existing components or by expanding existing systems for new domains or with new techniques. We have found multiple systems that utilize solutions like the NegEx algorithm [22] or the Weka machine learning workbench [23]. Other systems have adapted

some or the full suite of components from Stanford CoreNLP [24], NLTK [25], or OpenNLP [26]. The goals of the systems cover a very broad range of clinical tasks across multiple domains, but there are a few areas that have been the focus of more than one system. For example, multiple systems have been developed to identify medication information [27–34] and to extract tumor and cancer characteristics from pathology reports [35–39]. There are also systems attempting to process clinical trial eligibility criteria for easier cohort matching [40,41] and to determine the smoking history of patients [42,43]. Further descriptions of each system can be found in Table A1 in Appendix A.

4. Discussion

In this systematic review of clinical NLP solutions, we reviewed over 7000 publication records and narrowed our focus to a set of 86 papers. These papers provided information about 71 different NLP systems that could be used to process unstructured clinical text and generate structured output. This output varied from extraction of specific numeric values to complete normalization of multiple types of clinical data into standardized terminologies like the Unified Medical Language System or Systematized Nomenclature of Medicine Clinical Terms. Many of the systems seemed to be developed to address a specific need and had a fairly narrow focus, whether it was extracting medication dosage information or classifying cancer staging from pathology reports. For many tasks, especially involving extraction of numbers, acceptable performance was often achieved with relatively simple rule-based approaches (e.g. regular expression patterns) [41,108,109,113,116]. These rules were required to be highly tuned to the intricacies of data at a specific institution, however, and most practitioners agree that they are not scalable. Machine learning approaches have been rarer overall, but may grow in popularity as more public data sets are made available, helping to overcome the initial hurdle of obtaining training data. We were also heart-ened to see that a number of system developers were willing to provide their tools to the community.

Our systematic review has several limitations. First, our initial query involved the combination of two concepts (NLP and SDC). This was required because the query for the NLP concept by itself returned over 90,000 records from PubMed alone, and we did not have the resources to perform a proper review of this scale. It also means that the results for this particular query may not have contained certain papers about relevant NLP systems, although these systems might still have been captured if they were mentioned in other papers. Second, the information that could be gathered from the final set of papers was incomplete for many of the NLP systems. It is possible some of this missing information could have been obtained from other publications that were excluded during the screening phases, but this data source was too large to search in a reasonable time for this information. Third, we did not robustly gather citation counts for the final set of papers, but instead relied on only two sources which do not necessarily provide a complete or comparable set of citation information. This was only a brief exploration of the citation rates for these systems and was not a major goal of the analysis.

This review has highlighted the importance and, in many ways, the difficulty of performing comprehensive extraction and standardization of clinical information. Many of the most difficult tasks in this area have also been the target of open community challenges at

ongoing conferences like the Informatics for Integrating Biology & the Bedside (i2b2), the Sharing Annotated Resources/Conference and Labs of the Evaluation Forum (ShARe/CLEF) eHealth challenges, and the International Workshop on Semantic Evaluation (SemEval). For example, the 2009 i2b2 Challenge on Medication Information focused on identifying medication names, dosages, modes, and frequencies [119], which was also the goal of several systems found in the review. Task 1 of the 2013 ShARe/CLEF eHealth challenge and the Analysis of Clinical Text tasks at the 2014 and 2015 SemEval conferences have focused on identifying disorder mentions in text and then normalizing them to ontology terms [5–7]. This is a vital task for general purpose NLP and has still been shown to be somewhat challenging. Only a limited set of systems in this review attempt to do such a complete task. Some community challenges have also focused on important clinical NLP topics that have not been strongly addressed by the systems from the review, such as extraction of temporal information. The organizers of the 2012 i2b2 challenge on temporal relations concluded that even state of the art techniques did not provide good performance for temporal relation identification [3]. Indeed, we found few systems in this review that appeared to utilize any temporal information in text at all.

Addressing these few remaining challenges will help move clinical NLP systems toward the full-scale, mainstream acceptance and daily use that they have so far struggled to achieve. The promise of improving health outcomes with advanced methodologies like NLP is only possible when the systems can reliably satisfy unmet needs or support real-life use cases on a routine basis.

The rate of publication in the clinical NLP field is increasing, and literature-based reviews like this one may have difficulty in keeping up with new developments. It is therefore worthwhile to point out some other avenues that may prove useful in identifying and characterizing new systems. The ongoing NLP challenges mentioned above are excellent sources for new insights. A great deal of information can also be obtained by searching for clinical NLP software in online code repositories, as these listings are updated on a much faster basis than publications. For example, the amount of current activity around a software tool can be inferred by looking at the frequency of updates. Sources like these should be explored for the most up-to-date information.

Although we found evidence for a few focused specialty areas and identified a handful of tools that attempt to cover a broad set of needs, this review has demonstrated that there cannot be a single one-size-fits-all solution for the full set of clinical NLP subdomains. Nearly every system is focused on addressing a single clinical need (or at least started out that way), and true improvements in the field will come from making quality NLP applications available for specific use cases. Our joint project with CDC aims to accomplish exactly this, by creating a versatile platform with available pipelines for many specific tasks. Where needed, we will supplement the review information with additional searches to support our decision-making processes. Although we cannot guarantee we have identified every relevant system with this review, the compiled information provides an excellent base for evaluating the next steps for both our specific clinical NLP Platform and for the field as a whole. Certain areas of clinical NLP remain as open problems and will require development

of new advancements and approaches, but for the well-studied and "solved" aspects, we must ensure we are leveraging existing knowledge and systems in moving forward.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

The findings and conclusions in this report are those of the authors and do not necessarily represent the official position of the Food and Drug Administration or the Centers for Disease Control and Prevention.

This work was supported in part by the appointment of Nina Arya, Matthew Foster, Kory Kreimeyer, and Abhishek Pandey to the Research Participation Program administered by ORISE through an interagency agreement between the US Department of Energy and the US FDA.

The authors are grateful to Christine Baker of the FDA Library for an informative discussion regarding citations.

6. Funding

This work was supported by the Office of the Secretary Patient-Centered Outcomes Research Trust Fund under Interagency Agreement # 750116PE060014.

Appendix

Table A.1:

Detailed information about the Natural Language Processing systems.

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
Amalga	a commercial system for extracting data from clinical event messages	Unknown	Unknown	Unknown	[44]
ASLForm	an adaptive learning system with some basic rules for finding candidate text and a machine learning implementation that continuously updates as a user selects appropriate output	adaptively evaluated on subsets of the i2b2 2006 Smoking Status and 2008 Obesity challenge discharge summary corpora	precision 0.91–0.94 for single-value extractions and 0.81–0.89 for multi-value extractions (with recall 0.84–0.86)	Unknown	[45]
BaselineM	extracts frequent semantic tags from clinical trial eligibility criteria	Unknown	Unknown	Unknown	[40]
Bio-LarK Concept Recognizer	a semantic concept retrieval system that can process literature abstracts for common diseases and generate Human Phenotype	run on over 5 million PubMed records, but results were only evaluated after applying additional steps	Unknown	Unknown	[46]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	Ontology annotations				
BioMedICUS	system capable of finding family history statements in clinical notes by applying rules within likely sentences identified by a stochastic gradient descent classifier	tested on a set of a few hundred History and Physical notes from MTSamples.com	F-Measure was 0.92 for family member words and 0.65 and lower for relationships	Unknown	[47]
BioMedLEE	an extension of the MedLEE system to extract phenotypic information from MEDLINE abstracts	Unknown	Unknown	Unknown	[48];[49] *
CaRE	developed for processing discharge summaries	evaluated on hospital discharge summaries	F-Measure above 0.9 for retrieval of relevant medical concepts	Clinical notes, speech-to-text translated patient consultations	[50]
caTIES	a GATE-based system from University of Pittsburgh for coding pathology reports with terms from NCI Thesaurus	Unknown	Unknown	Unknown	[51];[52] *
ChartIndex	uses Stanford Parser and MetaMap to find noun phrases in radiology or pathology reports and map them to SNOMED-CT or UMLS terms	evaluated on 400 surgical pathology reports for extracting anatomic site and findings/ diagnosis	Precision was 0.88	Radiology reports, pathology reports	[53]
ClearForest	processes breast cancer pathology reports and uses rules to identify certain diagnoses along with dates and laterality	run on a sample of over 75,000 pathology reports from three hospitals	sensitivity was 0.99, specificity was 0.965, precision was 0.986	Unknown	[38]
ClinREAD	a rule-based system for clinical notes that has been used to categorize pain status	was evaluated on 24,000 pages for 33 cancer patients	F-Measure of 0.95 for pain detection and 0.81 for pain severity management	Unknown	[54];[55] *
COAT	a framework for clinical note processing allowing rule-based and machine learning (through WEKA) components and offering integration with	an implementation was tested on pathology reports to extract Gleason score, tumor stage, and surgical margin status	accuracy was 0.99 for Gleason score and tumor stage; 0.97 for surgical margin status	Unknown	[56]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	MetaMap Transfer				
CR	processes PubMed abstracts to find text chunks that can be mapped to UMLS concepts by considering the inverse document frequencies of terms	evaluated on 3663–7669 PubMed abstracts for 3 different diseases; also run on over 800,000 PubMed abstracts in the two Collaborative Annotation of a Large-scale Biomedical Corpus challenges, but results not identified	compared to assigned MeSH terms, precision was < 0.41, recall was < 0.52	Unknown	[57]
CRIS-IE-Smoking	rule-based system to extract patient smoking status from open-text fields of mental health case records	tested on 100 random mental health records from South London and Maudsley case register	precision was 0.93, recall was 0.58	Unknown	[43]
cTAKES	a large and well-used open-source system using the UIMA framework that extracts clinical data with contextual attributes like polarity and certainty and generates structured output using SNOMED-CT, UMLS, and RxNorm	many components have been evaluated separately; evaluated on 150 chest CT reports; medication extraction tested on 1507 breast cancer patients' notes for two drugs	accuracy for sentence boundary detector was 0.949, for tokenizer 0.949, for POS tagger 0.936, for shallow parser 0.924, and F-score for named entity recognizer was 0.715 for exact matching and 0.824 for overlapping span; for chest CT reports, precision was 0.72, recall was 0.48; medication accuracy was 0.925	Clinical notes, radiology reports	[30,40,45,51,58–68]
EpiDEA	an extension of cTAKES focused on epilepsy that extracts and structures data using the Epilepsy and Seizure Ontology and has been incorporated into the MEDCIS federated query platform	tested against 104 manually annotated discharge summaries	precision was 0.936, recall was 0.84	Unknown	[51,69,70]
FreePharma	a commercial system for structuring	Unknown	Unknown	Unknown	[31,34]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	dosage instructions				
GATE	an extensive open-source framework for text processing with a basic information extraction pipeline and a wide variety of additional components	an implementation was evaluated to extract mini mental state exam scores and dates from mental health records	precision was at least 0.85, recall was at least 0.85, depending on note type	Drug patent documents, mental health notes	[20,50]
HITEx	a system utilizing many GATE resources to process clinical notes and discharge summaries and extract normalized diagnostic and family history terms	evaluated on 150 discharge summaries for principal diagnosis, comorbidity, and smoking status; also evaluated separately for extraction of family history information from discharge summaries	accuracy was 0.82 for principal diagnosis, 0.87 for comorbidity, and 0.9 for smoking status; for family history, precision was 0.96, recall was 0.93	Unknown	[59,71,72]
i2b2 Workbench	an optional NLP component is included in the i2b2 Workbench, but limited to specific purposes	Unknown	Unknown	Unknown	[73]; [74] *
I2E	an indexing and standardizing tool for clinical trial documents that finds concepts from NCI Thesaurus, MeSH, and PubChem	tested retrieval of documents for queries based on certain compounds	“relative” recall and precision were 0.94–0.99	Unknown	[75]
LEXIMER	a machine learning system for CT and MR imaging reports that locates important clinical findings or recommendations in text; has been incorporated into the Render medical imaging platform	evaluated on a set of over 1000 radiology reports	sensitivity and specificity were 0.989 and 0.949 for marking important findings and 0.982 and 0.999 for marking recommendations for more imaging	Unknown	[58,64,76]
LifeCode	a commercial system for extracting billing codes and dosage information from medical records	has also been extended and tested for cancer findings in 500 reports	identified 4347 of 5139 findings	Unknown	[34,42,58]; [77] *
LINNAEUS	a NER tool with customizable dictionary	tested on 50 and 25 PubMed documents, respectively, using custom	precision and recall were 1.0 and 0.896 for pain terms and	Unknown	[78,79]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
		dictionaries of pain and disease terms	0.96 and 0.96 for disease terms		
LSP-MLP	one of the earliest clinical NLP systems for parsing clinical notes and converting to SNOMED codes	had been evaluated for information retrieval tasks	precision was 0.986, recall was 0.925	X-ray reports	[59]; [80] *
MedEx	an UIMA-based system for medication extraction	Unknown	Unknown	Unknown	[59]; [81] *
MedKATp	a pathology extraction system that uses rules to map text to elements of the Cancer Disease Knowledge Representation Model	Unknown	Unknown	Unknown	[35,39,59]
MedLEE	an early rule-based system for structuring radiography reports that was expanded for nearly all types of clinical notes and was later commercialized	evaluated on 150 random sentences from clinical documents	precision was 0.83, recall was 0.77	radiology reports, pathology reports, clinical notes, discharge summaries	[48,51,58,59,64,82–92]
MedTagger	an adaptation of BioTagger-GM for extracting UMLS concept mentions from clinical notes	participated in the concept mention task in the 2010 i2b2 NLP challenge	F-Measure was 0.84	Unknown	[93]
Medtex	extracts terms and SNOMED-CT concepts from death certificate text	Unknown	Unknown	Unknown	[94]
MedXN	extracts medication information from clinical notes into a form based on the RxNorm dictionary	evaluated on an unknown clinical note corpus	F-Measure was 0.92 for dosage and 0.84 for frequency	Unknown	[28]; [95] *
MERKI	a rule-based system to extract medication information, including context of drugs (history, hospital, discharge, or not administered), from discharge summaries	tested on 26 manually annotated discharge summaries	precision was 0.94, recall was 0.825 for drug names; context was correct for 66% of found drugs	Unknown	[34,96]
MetaMap	a long-running system originally designed for	evaluated on 42 publications related to sleep	precision was 0.7, recall was 0.77 for sleep	biomedical literature, clinical notes	[35,53,58,59,78,93,96–98]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	literature abstracts that assigns the best candidate UMLS terms to segments of text and can map output to any constituent terminology in the UMLS	disorder; previously evaluated on mining ICD codes from records of pneumonia and influenza death; also tested for mining concepts for respiratory syndromes from emergency department reports	disorder publications; precision was 0.98, recall was 0.998 for ICD codes from death reports; precision was 0.56, recall was 0.72 for emergency department reports		
MTERMS	a rule-based system for extracting terms, concepts, and drug information from standard terminologies including UMLS, RxNorm, and SNOMED-CT	evaluated for medication extraction on 30 charts (1108 terms); later evaluated on 200 notes for family history info	precision and recall were 0.9 for drugs; precision was 1.0, recall was 0.97 for family history	Unknown	[71,99]
Multithreaded Clinical Vocabulary Server	extracts SNOMED-CT concepts from radiology reports and other clinical notes	evaluated on chest radiography and CT reports to identify pneumonia	sensitivity was 1.0, specificity was 0.98	radiography and CT reports, clinical notes	[58,100]
Natural Language Patient Record	a commercial system for identifying drugs and dosage information in medical records	Unknown	Unknown	Unknown	[34]
NCBO Annotator	uses drug and disease ontologies to identify terms and map them to medical concepts from UMLS and RxNorm, including flags for things like negation	evaluated for recognizing 16 disease events on the i2b2 2008 Obesity challenge discharge summary corpus; also separately evaluated for drug exposure recognition	sensitivity was 0.74, specificity was 0.96 for events; precision was 0.84, recall was 0.84 for drugs	clinical notes, pathology reports,radiology reports	[101,102]
NILE	identifies location of pulmonary embolism from radiology reports	Unknown	Unknown	Unknown	[58]; [103]*
OntoFusion	takes structured and unstructured clinical and genomic data and uses known and inferred relationships to generate a logical data schema	Unknown	Unknown	clinical trial documents, biomedical literature	[104]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
ONYX	processes chest radiography reports to determine if they are consistent or inconsistent with pneumonia	run on a set of 5000 chest radiography reports (where it decided 12% of reports needed manual review for decision)	sensitivity was 0.75, specificity was 0.95	Unknown	[45,58]; [105]*
OpenNLP	an Apache project for NLP that includes components like a sentence boundary detector, tokenizer, symbol remover, and POS tagger, as well as MaxEnt and Perceptron named entity recognizers	Unknown	Unknown	Unknown	[29,45,60,106]
PEP	an extension of MedKATp with additional annotators still focusing on pathology reports	measured extraction results for 22 fields from 400 pathology reports in the Strategic Partnering to Evaluate Cancer Signatures database	F-Measure above 0.9 for half of fields, above 0.8 for all fields	prostate, breast, and lung pathology reports	[39]
RADA	uses a specialized domain glossary and predefined grammar rules to extract key medical concepts and their attributes from radiology reports	Unknown	Unknown	Unknown	[64]; [107]*
REDEx	a system for extracting numeric values like body weight by algorithmically creating regular expression patterns from annotated instances	tested on 568 notes annotated for body weight in 10-fold cross-validation	precision was 0.98, recall was 0.98	Unknown	[108]
Regextractor	a rule-based system for extracting numeric values from pulmonary function test text data	tested on pulmonary function test result charts for 100 subjects (1100 data points)	99.5% congruency with manual chart abstraction	cardiac catheterization and echocardiograph data	[109]
SymText / MPLUS	a long-running and much-updated system with Bayesian network-based	evaluated on 292 chest radiography reports to detect pneumonia as	sensitivity was 0.94, specificity was 0.91	Unknown	[58,59,64,83]; [110]*

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	semantic grammar that can extract and normalize findings from radiography reports	determined by consensus of 4 physicians			
TagLine	processes 80 character lines from VA health records and first classifies each line with a decision tree classifier and then uses rules to extract info from them	evaluated on 47 notes containing over 5000 lines	accuracy of line classifier was 0.985; precision was above 0.95, recall was 0.99 for single values; lower for lists	Unknown	[111]
TextMiner	a rule-based system for identifying possible drug side effect mentions in notes by parsing sentences containing a drug mention into labeled phrase structures	tested on 242 notes for statin-related adverse events	precision was 0.99, recall was 0.81 for sentence level information	Unknown	[112]
Valx	extracts and structures numeric value ranges or inequalities by using synonyms in the UMLS and manually defined heuristics	evaluated on an unknown number of Type 1 and Type 2 diabetes trial eligibility requirements	F-Measure was >=0.97 for “HbA1c” extraction and >=0.92 for “Glucose” extraction	Unknown	[41]
Abhyankar & Demner-Fushman 2013 (unnamed)	uses regular expressions to extract maternal information including lab test results from NICU admission notes and discharge summaries	manually evaluated results for about 500 data-rich NICU notes from MIMIC II	recall above 0.9 and precision above 0.95 for most features	Unknown	[113]
Barrett et al. 2013 (unnamed)	a machine learning system with rules for number extraction to identify 17 serious sentinel events (e.g. dyspnea, delirium, and sepsis) in palliative care consult letters	tested on datasets of15 and 215 labeled consult letters	average accuracy of 0.7–0.8 for different sentinel events	Unknown	[114]
Cameron et al. 2012 (unnamed)	extracts smoker semantic types from progress notes	tested on a set of notes for 100 patients	precision was 0.87	Unknown	[42]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
Chen et al. 2015 (unnamed)	uses Stanford recursive neural network parser and some regular expressions and heuristics to extract noun phrases and lab result values from sleep reports	tested noun phrase extraction and regular expressions on 100 documents	precision was 0.8, recall was 0.75 for noun phrases; accuracy was 0.98 for regular expressions	Unknown	[73]
Fang et al. 2014 (unnamed)	uses stemming and grouping algorithms (and NegEx) to do NER on clinical notes for diseases, symptoms, mental behaviors, and medications and converts to UMLS terms	Unknown	Unknown	Unknown	[115]
Hao & Weng 2015 (unnamed)	an extension of the BaselineM system to extract frequent semantic tags using some heuristic rules and some algorithms from NLTK	run on 500 clinical trial summaries, 500 paragraphs from 12 clinical trial protocols, and 500 clinical data warehouse requests	found 10%−20% more frequent semantic tags than BaselineM	Clinical Data Warehouse requests; Clinical Trial summaries and protocols	[40]
Hassanpour & Langlotz 2016 (unnamed)	applies many pieces of the Stanford NER toolkit, including CMM and CRF methods, to extract radiology terms from chest CT reports	evaluated on 150 manually annotated chest CT radiology reports from RadCore	precision was 0.87, recall was 0.84 for both methods	Unknown	[64]
He et al. 2014 (unnamed)	a system for processing clinical trial eligibility criteria that checks substrings for UMLS matches and uses the Valx system to extract numeric values like lab results	Unknown	Unknown	Unknown	[41]
Kadra et al. 2015 (unnamed)	a GATE-based system to extract prescription information (especially anti-psychotic medications) from health records	tested on records for 120 patients over 6 months for 6 frequent anti-psychotic medications	precision was 0.94–0.97, recall was 0.57–0.77 across 6 drugs	Unknown	[27]
Karystianis et al. 2016 (unnamed)	a rule-based system to capture dosage instructions	evaluated on 220 free text prescription instructions from	accuracy was 0.94–1.0 for individual attributes and	Unknown	[28]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	(including min and max variability) from special instruction notes to patients	Clinical Practice Research Datalink	0.91 for full prescriptions		
Korkontzelos et al. 2015 (unnamed)	a system with several different approaches for performing drug NER on biomedical literature with minimal or no gold standard corpus	evaluated on 120 MEDLINE abstracts	precision was 0.973, recall was 0.93 when using genetically evolved patterns and MaxEnt method	Unknown	[29]
Li et al. 2015 (unnamed)	uses a CRF model to identify medications and their attributes in clinical notes and compares the list to the structured medication list for discrepancies	tested on 300 discharge summaries with corresponding structured prescription lists	precision and recall were 0.88–0.95 for all features except duration	Unknown	[30]
Martinez et al. 2014 (unnamed)	classifies pathology reports for cancer staging information with Genia Tagger, MetaMap, and NegEx as input	tested several different machine learning methods and data subsets from about 400 pathology reports from 2 hospitals	F-Measures from 0.7–0.9 for different methods, but lower when training and testing used different hospitals	Unknown	[35]
Mork et al. 2010 (unnamed)	a rule-based system using MetaMap that was built for a medication extraction challenge and that informed further development of MetaMap	participated in the 2009 i2b2 Medication Extraction Challenge	precision was 0.78, recall was 0.82	Unknown	[96]
Otal et al. 2013 (unnamed)	a machine learning system using WEKA algorithms to detect T cancer staging classification	tested for 68 non-metastatic breast cancer patients	agreement with expert was 93% for best portable algorithm	Unknown	[36]
Shah & Martinez 2006 (unnamed)	extracts medication dosage instruction information from free text instructions to patient	tested on 1000 prescription records from General Practice Research Database; then further training and testing on prescription records from the Adverse Drug Reactions On-line Information	99.4% correct daily dose extraction in GPRD; accuracy was 0.93 for ADROIT	Unknown	[31,34]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
		Tracking database			
Turchin et al. 2014 (unnamed)	a rule-based system that identifies discrepancies between free text prescription information and structured medication information	evaluated on 1000 electronic prescriptions manually reviewed for discrepancies by two experts	precision was 0.84, recall was 0.76	Unknown	[32]
Voorham & Denig 2007 (unnamed)	a rule-based system for processing clinical notes and extracting 13 numeric measurements for evaluating the quality of diabetes care (e.g. blood pressure, weight, height, and serum glucose)	tested on 60 annotated patient records	Precision was >=0.91 for 10 of 13 values, recall was >=0.94 for 11 of 13 values	Unknown	[116]
Wieneke et al. 2015 (unnamed)	a machine learning system for breast pathology reports that extracts procedure, result, and laterality; when high PPV and high NPV classifiers disagreed, reports were marked for manual review	evaluated on 324 breast pathology reports after training on nearly 3000 reports	12.7% were completely correctly coded; half of reports sent to manual review	Unknown	[37]
Xu et al. 2012 (unnamed)	a voting-based system built (after the fact) for the i2b2 2010 three-part challenge	evaluated on the i2b2 2010 corpus of 477 discharge summaries for the concept extraction, assertion classification, and relation identification tasks	micro-averaged F-Measure for the 3 tasks: 0.85, 0.94, 0.73	Unknown	[117]
Yli-Hietanen et al. 2009 (unnamed)	an expansion to a chief complaint normalization system to add approximate matching for misspelling	tested on chief complaints from 5 hospital emergency departments; checked results for 500 of them	99% correctly normalized, 98.4% when allowing misspellings	Unknown	[118]
Zheng et al. 2015 (unnamed)	uses some proprietary NLP components/ software and a custom dictionary of aspirin terms to extract low dose aspirin	evaluated on 5339 manually annotated notes from patients with atrial fibrillation	precision was 0.93, recall was 0.955	Unknown	[33]

System Name	Brief Description	Evaluation	Performance	Usage On Other Texts	Citations
	medication information				

*These citations did not originate through the systematic review process, but are provided for the sake of convenience.

NLP: Natural Language Processing; **POS:** Part-of-Speech; **NER:** Named Entity Recognition; **CMM:** Conditional Markov Model; **CRF:** Conditional Random Fields; **WEKA:** Waikato Environment for Knowledge Analysis; **NLTK:** Natural Language Toolkit; **i2b2:** Informatics for Integrating Biology & the Bedside; **NCI:** National Cancer Institute; **MeSH:** Medical Subject Headings; **UMLS:** Unified Medical Language System; **SNOMED-CT:** Systematized Nomenclature of Medicine – Clinical Terms; **UIMA:** Unstructured Information Management Architecture; **GATE:** General Architecture for Text Engineering; **CT:** Computerized Tomography; **PPV:** Positive Predictive Value; **NPV:** Negative Predictive Value

References

[1]. Ni Y, Kennebeck S, Dexheimer JW, McAneney CM, Tang H, Lingren T, et al., Automated clinical trial eligibility prescreening: increasing the efficiency of patient identification for clinical trials in the emergency department, J. Am. Med. Inform. Assoc. 22 (2015) 166–178. [PubMed: 25030032]

[2]. Wang G, Jung K, Winnenburg R, Shah NH, A method for systematic discovery of adverse drug events from clinical notes, J. Am. Med. Inform. Assoc. 22 (2015) 1196–1204. [PubMed: 26232442]

[3]. Sun W, Rumshisky A, Uzuner O, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, J. Am. Med. Inform. Assoc. 20 (2013) 806–813. [PubMed: 23564629]

[4]. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (2012) 786–791. [PubMed: 22366294]

[5]. Pradhan S, Elhadad N, South BR, Martinez D, Vogel A, Suominen H, et al., Task 1: ShARe/CLEF eHealth Evaluation Lab, 2013.

[6]. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G, SemEval-2014 Task 7: Analysis of Clinical Text. Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014): Association for Computational Linguistics and Dublin City University; 2014, pp. 54–62.

[7]. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G, SemEval-2015 Task 14: Analysis of Clinical Text. Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015): Association for Computational Linguistics; 2015, pp. 303–310.

[8]. Jones S, Development of a Natural Language Processing (NLP) Web Service for Structuring and Standardizing Unstructured Clinical Information. NAACCR 2016 Annual Conference St. Louis, MO, 2016.

[9]. Structured Data Capture Charter and Members. Standards & Interoperability Framework.

[10]. Moher D, Liberati A, Tetzlaff J, Altman DG, the PG, Preferred reporting items for systematic reviews and meta-analyses: The prisma statement, Ann. Intern. Med. 151 (2009) 264–269. [PubMed: 19622511]

[11]. Thomas J, Brunton J, Graziosa S, EPPI-Reviewer 4: Software for Research Synthesis, Social Science Research Unit, UCL Institute of Education, EPPI-Centre Software London, 2010.

[12]. EndNote. <http://endnote.com/>. [Last accessed 2017 Jun 7].

[13]. JabRef. <http://www.jabref.org/>. [Last accessed 2017 Jun 7].

[14]. Szostak J, Ansari S, Madan S, Fluck J, Talikka M, Iskandar A, et al. Construction of biological networks from unstructured information based on a semiautomated curation workflow. Database (Oxford). 2015;2015:bav057.

[15]. Miwa M, Saetre R, Kim JD, Tsujii J, Event extraction with complex event classification using rich features, J. Bioinform. Comput. Biol. 8 (2010) 131–146. [PubMed: 20183879]

[16]. Hoehndorf R, Schofield PN, Gkoutos GV, Analysis of the human diseasome using phenotype similarity between common, genetic, and infectious diseases, Sci. Rep. 5 (2015) 10888.

[17]. He Q, Veldkamp BP, de V, Screening for posttraumatic stress disorder using verbal features in self narratives: a text mining approach, Psychiatry Res. 198 (2012) 441–447. [PubMed: 22464046]

[18]. Kho AN, Pacheco JA, Peissig PL, Rasmussen L, Newton KM, Weston N, Electronic medical records for genetic research: results of the eMERGE consortium, Sci. Transl. Med. 3 (2011). 79re1.

[19]. Ferrucci D, Lally A, UIMA: an architectural approach to unstructured information processing in the corporate research environment, Nat. Lang. Eng. 10 (2004) 327–348.

[20]. Cunningham H, Tablan V, Roberts A, Bontcheva K, Getting more out of biomedical documents with GATE's full lifecycle open source text analytics, Plos Comput. Biol. (2013) 9.

[21]. Cunningham H, Maynard D, Bontcheva K, Text Processing with GATE (Version 6): Gateway Press CA, 2011.

[22]. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG, A simple algorithm for identifying negated findings and diseases in discharge summaries, J. Biomed. Inform. 34 (2001) 301–310. [PubMed: 12123149]

[23]. Frank E, Hall MA, Witten IH, The WEKA Workbench Data Mining: Practical Machine Learning Tools and Techniques. Fourth ed: Morgan Kaugmann, 2016.

[24]. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D, The Stanford CoreNLP Natural Language Processing Toolkit, in: Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations Baltimore, Maryland, 2014, pp. 55–60.

[25]. Bird S, Klein E, Loper E, Natural Language Processing with Python: O'Reilly Media, Inc, 2009.

[26]. Apache OpenNLP. <http://opennlp.apache.org/>. [Last accessed on 2017 May 30].

[27]. Kadra G, Stewart R, Shetty H, Jackson RG, Greenwood MA, Roberts A, et al., Extracting antipsychotic polypharmacy data from electronic health records: developing and evaluating a novel process, BMC Psychiatry 15 (2015) 166. [PubMed: 26198696]

[28]. Karystianis G, Sheppard T, Dixon WG, Nenadic G, Modelling and extraction of variability in free-text medication prescriptions from an anonymised primary care electronic medical record research database, BMC Med. Inform. Decis. Mak. 16 (2016) 18. [PubMed: 26860263]

[29]. Korkontzelos I, Piliouras D, Dowsey AW, Ananiadou S, Boosting drug named entity recognition using an aggregate classifier, Artif. Intell. Med. 65 (2015)145–153. [PubMed: 26116947]

[30]. Li Q, Spooner SA, Kaiser M, Lingren N, Robbins J, Lingren T, An end-to-end hybrid algorithm for automated medication discrepancy detection, BMC Med. Inform. Decision Making 15 (2015).

[31]. Shah AD, Martinez C, An algorithm to derive a numerical daily dose from unstructured text dosage instructions, Pharmacoepidemiol. Drug Saf. 15 (2006)161–166. [PubMed: 16170830]

[32]. Turchin A, Sawarkar A, Dementieva YA, Breydo E, Ramelson H, Effect of EHR user interface changes on internal prescription discrepancies, Appl Clin Inform. 5 (2014) 708–720. [PubMed: 25298811]

[33]. Zheng C, Rashid N, Koblick R, An J, Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation, Clin. Ther. 37 (2048–58) (2015) e2.

[34]. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G, Extracting structured medication event information from discharge summaries, AMIA Annu. Symp. Proc. 237–41 (2008). [PubMed: 18999147]

[35]. Martinez D, Pitson G, MacKinlay A, Cavedon L, Cross-hospital portability of information extraction of cancer staging information, Artif. Intell. Med. 62 (2014) 11–21. [PubMed: 25001545]

[36]. Otal RG, Guerra JLL, Calderon CLP, Garcia AM, Gironzini VS, Serrano JP, et al., Application of artificial intelligence in tumors sizing classification for, Breast Cancer (2013).

[37]. Wieneke AE, Bowles EJ, Cronkite D, Wernli KJ, Gao H, Carrell D, et al., Validation of natural language processing to extract breast cancer pathology procedures and results, J. Pathol. Inform. 6 (2015) 38. [PubMed: 26167382]

[38]. Buckley JM, Coopey SB, Sharko J, Polubriaginof F, Drohan B, Belli AK, et al., The feasibility of using natural language processing to extract clinical information from breast pathology reports, J. Pathol. Inform. 3 (2012) 23. [PubMed: 22934236]

[39]. Ashish N, Dahm L, Boicey C, University of California, Irvine-pathology extraction pipeline: the pathology extraction pipeline for information extraction from pathology reports, Health Inform. J. 20 (2014) 288–305.

[40]. Hao T, Weng C, Adaptive semantic tag mining from heterogeneous clinical research texts, Methods Inf. Med. 54 (2015) 164–170. [PubMed: 25327613]

[41]. He Z, Carini S, Hao T, Sim I, Weng C, A method for analyzing commonalities in clinical trial target populations, AMIA Annu Symp Proc. 2014 (2014) 1777–1786. [PubMed: 25954450]

[42]. Cameron D, Bhagwan V, Sheth AP, Towards comprehensive longitudinal healthcare data capture, in: Gao J, Dubitzky W, Wu C, Liebman M, Alhaij R, Ungar L, et al. (Eds.), 2012 leee International Conference on Bioinformatics and Biomedicine Workshops, 2012.

[43]. Wu CY, Chang CK, Robson D, Jackson R, Chen SJ, Hayes RD, et al., Evaluation of smoking status identification using electronic health records and open-text information in a large mental health case register, PLoS One 8 (2013) e74262.

[44]. Devine EB, Capurro D, van Eaton E, Alfonso-Cristancho R, Devlin A, Yanez ND, et al., Preparing Electronic Clinical Data for Quality Improvement and Comparative Effectiveness Research: The SCOAP CERTAIN Automation and Validation Project. EGEMS (Wash DC), vol. 1, 2013, pp. 1025. [PubMed: 25848565]

[45]. Zheng S, Wang F, Lu JJ, ASLForm: an adaptive self learning medical form generating system, AMIA Annu. Symp. Proc. 2013 (2013) 1590–1599. [PubMed: 24551429]

[46]. Groza T, Kohler S, Moldenhauer D, Vasilevsky N, Baynam G, Zemojtel T, et al., The human phenotype ontology: semantic unification of common and rare disease, Am. J. Hum. Genet. 97 (2015) 111–124. [PubMed: 26119816]

[47]. Bill R, Pakhomov S, Chen ES, Winden TJ, Carter EW, Melton GB, Automated extraction of family history information from clinical notes, AMIA Annu. Symp. Proc. 2014 (2014) 1709–1717. [PubMed: 25954443]

[48]. Friedman C, Borlawsky T, Shagina L, Xing HR, Lussier YA, Bio-ontology and text: bridging the modeling gap, Bioinformatics 22 (2006) 2421–2429. [PubMed: 16870928]

[49]. Chen L, Friedman C, Extracting phenotypic information from the literature via natural language processing, Stud. Health Technol. Inform. 107 (2004) 758–762. [PubMed: 15360914]

[50]. Klann JG, Szolovits P, An intelligent listening framework for capturing encounter notes from a doctor-patient dialog, BMC Med. Inform. Decis. Mak. 9 (Suppl 1) (2009) S3. [PubMed: 19891797]

[51]. Cui L, Bozorgi A, Lhatoo SD, Zhang GQ, Sahoo SS, EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification, AMIA Annu. Symp. Proc. 2012 (2012) 1191–1200. [PubMed: 23304396]

[52]. Crowley RS, Castine M, Mitchell K, Chavan G, McSherry T, Feldman M, caTIES: a grid based system for coding and retrieval of surgical pathology reports and tissue specimens in support of translational research, J. Am. Med. Inform. Assoc.: JAMIA 17 (2010) 253–264. [PubMed: 20442142]

[53]. Lowe HJ, Huang Y, Regula DP, Using a statistical natural language Parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports, AMIA Annu. Symp. Proc. 2009 (2009) 386–390. [PubMed: 20351885]

[54]. Kreuzthaler M, Schulz S, Berghold A, Secondary use of electronic health records for building cohort studies through top-down information extraction, J. Biomed. Inform. 53 (2015) 188–195. [PubMed: 25451102]

[55]. Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ, Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data, J. Am. Med. Inform. Assoc. 16 (2009) 571–575. [PubMed: 19390103]

[56]. D'Avolio LW, Bui AA, The clinical outcomes assessment toolkit: a framework to support automated clinical records-based outcomes assessment and performance measurement research, J. Am. Med. Inform. Assoc. 15 (2008) 333–340. [PubMed: 18308990]

[57]. Berlanga R, Jimenez-Ruiz E, Nebot V, Exploring and linking biomedical resources through multidimensional semantic spaces, BMC Bioinform. 13 (Suppl 1) (2012) S6.

[58]. Cai T, Giannopoulos AA, Yu S, Kelil T, Ripley B, Kumamaru KK, et al., Natural language processing technologies in radiology research and clinical applications, Radiographics 36 (2016) 176–191. [PubMed: 26761536]

[59]. Doan S, Conway M, Phuong TM, Ohno-Machado L, Natural language processing in biomedicine: a unified system architecture overview, Methods Mol. Biol. 1168 (2014) 275–294. [PubMed: 24870142]

[60]. Piliouras D, Korkontzelos I, Dowsey A, Ananiadou S, Ieee, Dealing with data sparsity in Drug Named Entity Recognition, 2013 Ieee International Conference on Healthcare Informatics (Ichi 2013), 2013, pp. 14–21.

[61]. Xu H, Fu Z, Shah A, Chen Y, Peterson NB, Chen Q, et al., Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases, AMIA Annu. Symp. Proc. 2011 (2011) 1564–1572. [PubMed: 22195222]

[62]. Savova GK, Olson JE, Murphy SP, Cafourek VL, Couch FJ, Goetz MP, et al., Automated discovery of drug treatment patterns for endocrine therapy of breast cancer within an electronic medical record, J. Am. Med. Inform. Assoc. 19 (2012) e83–e89. [PubMed: 22140207]

[63]. Wu ST, Kaggal VC, Dligach D, Masanz JJ, Chen P, Becker L, et al., A common type system for clinical natural language processing, J. Biomed. Semantics 4 (2013) 1. [PubMed: 23286462]

[64]. Hassanpour S, Langlotz CP, Information extraction from multi-institutional radiology reports, Artif. Intell. Med. 66 (2016) 29–39. [PubMed: 26481140]

[65]. Lin C, Karlson EW, Dligach D, Ramirez MP, Miller TA, Mo H, et al., Automatic identification of methotrexate-induced liver toxicity in patients with rheumatoid arthritis from the electronic medical record, J. Am. Med. Inform. Assoc. 22 (2015) e151–e161. [PubMed: 25344930]

[66]. Pathak J, Murphy SP, Willaert BN, Kremers HM, Yawn BP, Rocca WA, et al., Using RxNorm and NDF-RT to classify medication data extracted from electronic health records: experiences from the Rochester Epidemiology Project, AMIA Annu. Symp. Proc. 2011 (2011) 1089–1098. [PubMed: 22195170]

[67]. Pathak J, Bailey KR, Beebe CE, Bethard S, Carrell DC, Chen PJ, et al., Normalization and standardization of electronic health records for high-throughput phenotyping: the SHARPn consortium, J. Am. Med. Inform. Assoc. 20 (2013) e341–e348. [PubMed: 24190931]

[68]. Rea S, Pathak J, Savova G, Oniki TA, Westberg L, Beebe CE, et al., Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project, J. Biomed. Inform. 45 (2012) 763–771. [PubMed: 22326800]

[69]. Sahoo SS, Lhatoo SD, Gupta DK, Cui L, Zhao M, Jayapandian C, et al., Epilepsy and seizure ontology: towards an epilepsy informatics infrastructure for clinical research and patient care, J. Am. Med. Inform. Assoc. 21 (2014) 82–89. [PubMed: 23686934]

[70]. Zhang GQ, Cui L, Lhatoo S, Schuele SU, Sahoo SS, MEDCIS: multi-modality epilepsy data capture and integration system, AMIA Annu. Symp. Proc. 2014 (2014) 1248–1257. [PubMed: 25954436]

[71]. Zhou L, Lu Y, Vitale CJ, Mar PL, Chang F, Dhopeshwarkar N, et al., Representation of information about family relatives as structured data in electronic health records, Appl. Clin. Inform. 5 (2014) 349–367. [PubMed: 25024754]

[72]. Liao KP, Ananthakrishnan AN, Kumar V, Xia Z, Cagan A, Gainer VS, et al., Methods to develop an electronic medical record phenotype algorithm to compare the risk of coronary artery disease across 3 chronic disease cohorts, PLoS One 10 (2015) e0136651.

[73]. Chen W, Kowatch R, Lin S, Splaingard M, Huang Y, Interactive cohort identification of sleep disorder patients using natural language processing and i2b2, Appl. Clin. Inform. 6 (2015) 345–363. [PubMed: 26171080]

[74]. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al., Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2), J. Am. Med. Inform. Assoc. 17 (2010) 124–130. [PubMed: 20190053]

[75]. Chang MP, Chang M, Reed JZ, Milward D, Xu JJ, Cornell WD, Developing timely insights into comparative effectiveness research with a text-mining pipeline, Drug Discovery Today 21 (2016) 473–480. [PubMed: 26854423]

[76]. Dang PA, Kalra MK, Schultz TJ, Graham SA, Dreyer KJ, Informatics in radiology: render: an online searchable radiology study repository, Radiographics 29 (2009) 1233–1246. [PubMed: 19564253]

[77]. Heinze DT, Morsch M, Sheffer R, Jimmink M, Jennings M, Morris W, et al., LifeCode: a deployed application for automated medical coding, Ai Magazine 22 (2001) 76.

[78]. Lam C, Lai FC, Wang CH, Lai MH, Hsu N, Chung MH, Text mining of journal articles for sleep disorder terminologies, PLoS One 11 (2016) e0156031.

[79]. Jamieson DG, Roberts PM, Robertson DL, Sidders B, Nenadic G, Cataloging the biomedical world of pain through semi-automated curation of molecular interactions, Database (Oxford), 2013;2013, bat033.

[80]. Sager N, Lyman M, Bucknall C, Nhan N, Tick LJ, Natural language processing and the representation of clinical data, J. Am. Med. Inform. Assoc. 1 (1994) 142–160. [PubMed: 7719796]

[81]. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inform. Assoc. 17 (2010) 19–24. [PubMed: 20064797]

[82]. Johnson SB, Bakken S, Dine D, Hyun S, Mendonca E, Morrison F, et al., An electronic health record based on structured narrative, J. Am. Med. Inform. Assoc. 15 (2008) 54–64. [PubMed: 17947628]

[83]. Chen ES, Hripcsak G, Friedman C, Disseminating natural language processed clinical narratives, AMIA Annu. Symp. Proc. 126–30 (2006). [PubMed: 17238316]

[84]. Hripcsak G, Knirsch C, Zhou L, Wilcox A, Melton GB, Using discordance to improve classification in narrative clinical databases: an application to community-acquired pneumonia, Comput. Biol. Med. 37 (2007) 296–304. [PubMed: 16620802]

[85]. Hripcsak G, Soulakis ND, Li L, Morrison FP, Lai AM, Friedman C, et al., Syndromic surveillance using ambulatory electronic health records, J. Am. Med. Inform. Assoc. 16 (2009) 354–361. [PubMed: 19261941]

[86]. Hyun S, Johnson SB, Bakken S, Exploring the ability of natural language processing to extract data from nursing narratives, Cin-Comput. Inform. Nurs. 27 (2009) 215–223.

[87]. Li L, Chase HS, Patel CO, Friedman C, Weng C, Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials prescreening: a case study, AMIA Annu. Symp. Proc. 404–8 (2008). [PubMed: 18999285]

[88]. Morrison FP, Li L, Lai AM, Hripcsak G, Repurposing the clinical record: can an existing natural language processing system de-identify clinical notes?, J Am. Med. Inform. Assoc. 16 (2009) 37–39. [PubMed: 18952938]

[89]. Peissig PL, Rasmussen LV, Berg RL, Linneman JG, McCarty CA, Waudby C, et al., Importance of multi-modal approaches to effectively identify cataract cases from electronic health records, J. Am. Med. Inform. Assoc. 19 (2012) 225–234. [PubMed: 22319176]

[90]. Salmasian H, Freedberg DE, Friedman C, Deriving comorbidities from medical records using natural language processing, J. Am. Med. Inform. Assoc. 20 (2013) e239–e242. [PubMed: 24177145]

[91]. Yadav K, Sarioglu E, Smith M, Choi HA, Automated outcome classification of emergency department computed tomography imaging reports, Acad. Emerg. Med. 20 (2013) 848–854. [PubMed: 24033628]

[92]. Yadav K, Sarioglu E, Choi HA, Cartwright W.B.t., Hinds PS, Chamberlain JM, Automated outcome classification of computed tomography imaging reports for pediatric traumatic brain injury, Acad. Emerg. Med. 23 (2016) 171–178. [PubMed: 26766600]

[93]. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al., Towards a semantic lexicon for clinical natural language processing, AMIA Annu. Symp. Proc. 2012 (2012) 568–576. [PubMed: 23304329]

[94]. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N, Automatic ICD-10 classification of cancers from free-text death certificates, Int. J. Med. Inform. 84 (2015) 956–965. [PubMed: 26323193]

[95]. Sohn S, Clark C, Halgrim SR, Murphy SP, Chute CG, Liu H, MedXN: an open source medication extraction and normalization tool for clinical text, J. Am. Med. Inform. Assoc. 21 (2014) 858–865. [PubMed: 24637954]

[96]. Mork JG, Bodenreider O, Demner-Fushman D, Dogan RI, Lang FM, Lu Z, et al., Extracting Rx information from clinical narrative, J. Am. Med. Inform. Assoc. 17 (2010) 536–539. [PubMed: 20819859]

[97]. Jiang L, Edwards SM, Thomsen B, Workman CT, Guldbrandtsen B, Sorensen P, A random set scoring model for prioritization of disease candidate genes using protein complexes and data-mining of GeneRIF, OMIM and PubMed records, BMC Bioinform. 15 (2014) 315.

[98]. Yin SM, Y Li C, Zhou YG, Huang J, Detecting hotspots in insulin-like growth factors 1 research through metamap and data mining technologies, in: Huang Z, Liu C, He J, Huang G (Eds.), Web Information Systems Engineering - Wise 2013 Workshops, 2014, pp. 359–372.

[99]. Zhou L, Plasek JM, Mahoney LM, Karipineni N, Chang F, Yan X, et al., Using medical text extraction, reasoning and mapping system (MTERMS) to process medication information in outpatient clinical notes, AmIa Annu. Symp. Proc. 2011 (2011) 1639–1648. [PubMed: 22195230]

[100]. FitzHenry F, Murff HJ, Matheny ME, Gentry N, Fielstein EM, Brown SH, et al., Exploring the frontier of electronic health record surveillance: the case of postoperative complications, Med. Care 51 (2013) 509–516. [PubMed: 23673394]

[101]. Huang SH, LePendu P, Iyer SV, Tai-Seale M, Carrell D, Shah NH, Toward personalizing treatment for depression: predicting diagnosis and severity, J. Am. Med. Inform. Assoc. 21 (2014) 1069–1075. [PubMed: 24988898]

[102]. Cole TS, Frankovich J, Iyer S, LePendu P, Bauer-Mehren A, Shah NH, Profiling risk factors for chronic uveitis in juvenile idiopathic arthritis: a new model for EHR-based research, Pediatr. Rheumatol. 11 (2013).

[103]. Yu S, Cai T, A Short Introduction to NILE. arXiv:13116063 2013.

[104]. Garcia-Remesal M, Maojo V, Billhardt H, Crespo J, Integration of relational and textual biomedical sources. A pilot experiment using a semi-automated method for logical schema acquisition, Methods Inf. Med. 49 (2010) 337–348. [PubMed: 19936436]

[105]. Christensen L, Harkema H, Haug P, Irwin J, Chapman W, ONYX: a system for the semantic analysis of clinical text, in: Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing. Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 19–27.

[106]. Lin CH, Wu NY, Liou DM, A multi-technique approach to bridge electronic case report form design and data standard adoption, J. Biomed. Inform. 53 (2015) 49–57. [PubMed: 25200473]

[107]. Johnson DB, Taira RK, Cardenas AF, Aberle DR, Extracting information from free text radiology reports, Int. J. Digit. Libr. 1 (1997) 297–308.

[108]. Murtaugh MA, Gibson BS, Redd D, Zeng-Treitler Q, Regular expression-based learning to extract bodyweight values from clinical notes, J. Biomed. Inform. 54 (2015) 186–190. [PubMed: 25746391]

[109]. Hinchcliff M, Just E, Podlusky S, Varga J, Chang RW, Kibbe WA, Text data extraction for a prospective, research-focused data mart: implementation and validation, BMC Med. Inform. Decis. Mak. 12 (2012) 106. [PubMed: 22970696]

[110]. Christensen L, Haug P, Fiszman M, MPLUS: a probabilistic medical language understanding system, in: Proceedings of the ACL-02 workshop on Natural language processing in the

biomedical domain - Volume 3. Phildadelphia, Pennsylvania: Association for Computational Linguistics, 2002, pp. 29–36.

[111]. Finch DK, McCart JA, Luther SL, TagLine: information extraction for semi-structured text in medical progress notes, AMIA Annu. Symp. Proc. 2014 (2014) 534–543. [PubMed: 25954358]

[112]. Skentzos S, Shubina M, Plutzky J, Turchin A, Structured vs. unstructured: factors affecting adverse drug reaction documentation in an EMR repository, AMIA Annu. Symp. Proc. 2011 (2011) 1270–1279. [PubMed: 22195188]

[113]. Abhyankar S, Demner-Fushman D, A simple method to extract key maternal data from neonatal clinical notes, AMIA Annu. Symp. Proc. 2013 (2013) 2–9. [PubMed: 24551317]

[114]. Barrett N, Weber-Jahnke JH, Thai V, Engineering natural language processing solutions for structured information from clinical text: extracting sentinel events from palliative care consult letters, Stud. Health Technol. Inform. 192 (2013) 594–598. [PubMed: 23920625]

[115]. Fang S, Palakal M, Xia Y, Grannis Shaun J, Williams Jennifer L, Health-Terrain: Visualizing Large Scale Health Data. INDIANA UNIV INDIANAPOLIS, 2014, pp. 79.

[116]. Voorham J, Denig P, Computerized extraction of information on the quality of diabetes care from free text in electronic patient records of general practitioners, J. Am. Med. Inform. Assoc. 14 (2007) 349–354. [PubMed: 17329733]

[117]. Xu Y, Hong K, Tsujii J, Chang EI, Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries, J. Am. Med. Inform. Assoc. 19 (2012) 824–832. [PubMed: 22586067]

[118]. Yli-Hietanen J, Niiranen S, Aswell M, Nathanson L, Domain-specific analytical language modeling–the chief complaint as a case study, Int. J. Med. Inform. 78 (2009) e27–e30. [PubMed: 19307149]

[119]. Uzuner Ö, Solti I, Cadag E, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (2010) 514–518. [PubMed: 20819854]

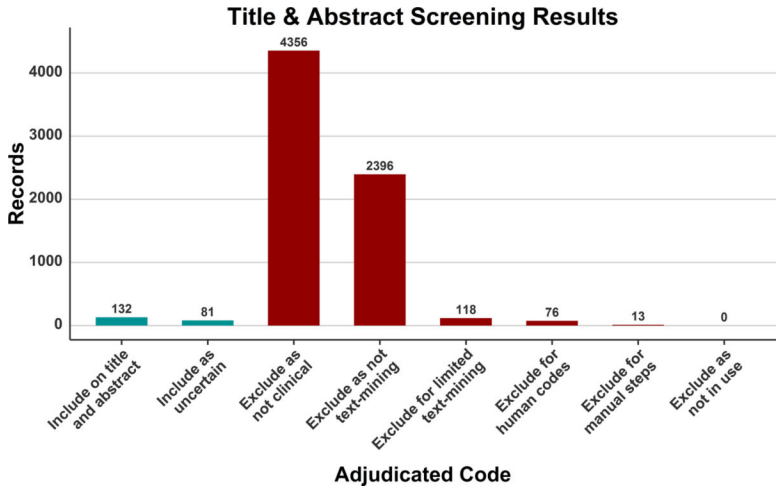

Fig. 1.

The coding results of the title and abstract screening following adjudication.

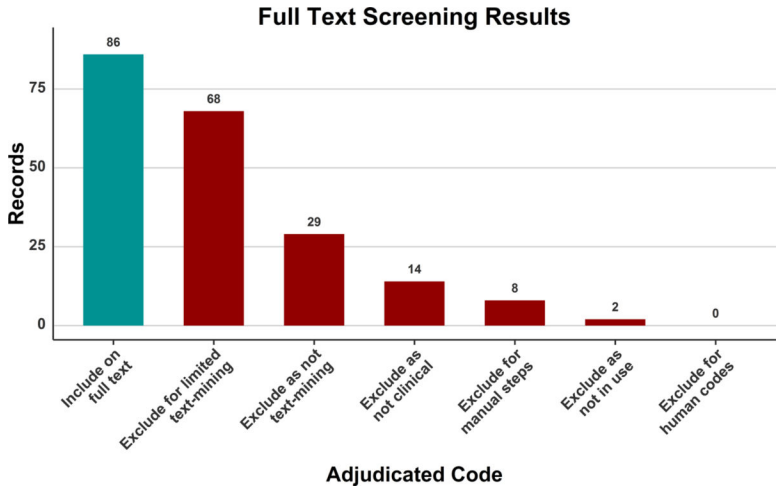

Fig. 2.
The coding results of the full text screening following adjudication.

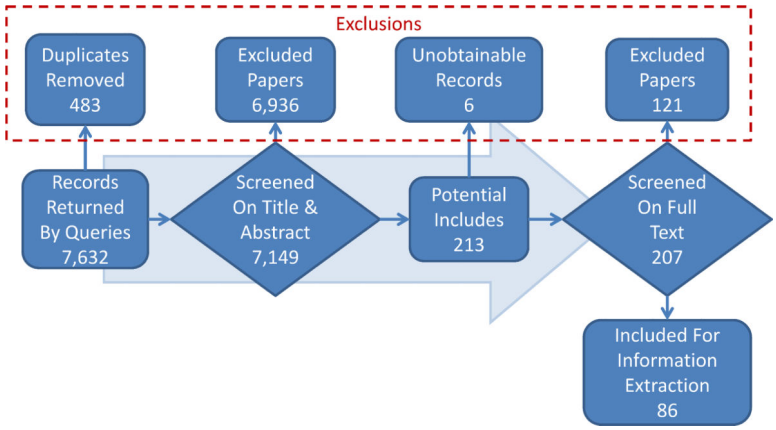

Fig. 3.
The review process and the number of records in Phases 1–3.

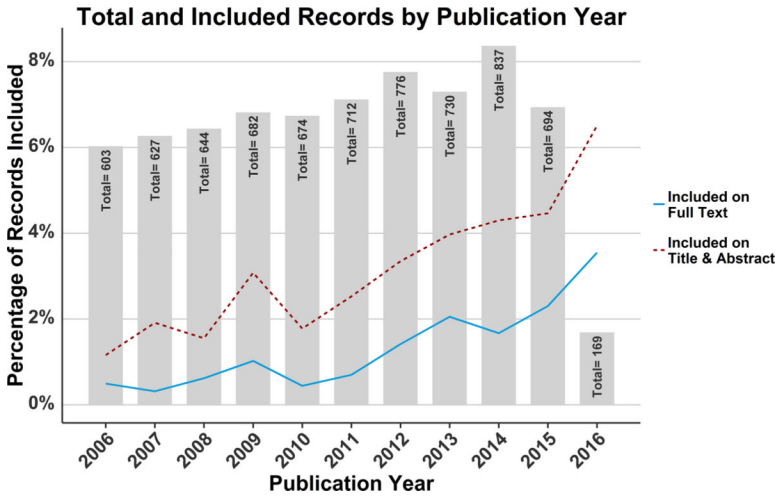

Fig. 4.

The queried records and the percentage included after each screening phase by publication year. The number in each bar represents the total records found by the initial query that were published during that year. Note that 2016 is an incomplete year because the query was run on June 15 of that year. The two lines show the percentage of records published in each year that were included after review of their title and abstract (red dashed line) and full text (blue solid line). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

Paper	Citation Year											Total Citations
	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	
Rea et al. 2012 [68]							1	12	17	18	7	55
Voorham & Denig 2007 [116]		0	5	4	10	2	6	7	4	6	6	50
Peissig et al. 2012 [89]							2	6	18	7	4	37
Cunningham et al. 2013 [20]								3	10	13	7	33
Pathak et al. 2013 [67]								0	2	12	9	23
Johnson et al. 2008 [82]			0	0	1	7	8	2	2	0	1	21
Groza et al. 2015 [46]										3	17	20
Hripcsak et al. 2009 [85]				0	5	5	2	4	0	1	3	20
FitzHenry et al. 2013 [100]								0	3	9	2	14
Dang et al. 2009 [76]				1	1	2	3	3	1	0	3	14

Fig. 5.

The citation counts per year for the 10 publications with the most citations. Retrieved from Web of Science on 2016–11–03.

Table 1:

Basic information about the Natural Language Processing systems.

Entries marked with 'Not Found' indicate that no relevant information was obtained from the final set of papers, and entries marked with 'Not Determined' indicate that some information was found but there was not a clear enough statement to unambiguously label the system.

System Name	Full Name*	NLP Type	Framework	Open-source	Mentioned In**	Described/Used In
Amalga	Amalga Unified Intelligence System	Not Found	Not Found	No	N/A	[44]
ASLForm	N/A	Hybrid	Not Found	Not Determined	N/A	[45]
BaselineM	N/A	Not Determined	Not Found	Not Found	N/A	[40]
Bio-LarK Concept Recognizer	Biomedical Large Knowledge Collider Concept Recognizer	Not Determined	Not Found	Not Found	N/A	[46]
BioMedlCUS	Biomedical Information Collection and Understanding System	Hybrid	UIMA	Yes	N/A	[47]
BioMedLEE	BioMedical Language Extraction and Encoding	Rule-based	Not Found	Not Found	[48]	[49]***
CaRE	Category and Relationship Extractor	Not Determined	Not Found	Not Found	N/A	[50]
caTIES	Cancer Text Information Extraction System	Not Found	GATE	Not Found	[51]	[52]***
ChartIndex	N/A	Hybrid	Not Found	Not Found	N/A	[53]
ClearForest	N/A	Rule-based	Not Found	Not Found	N/A	[38]
ClinREAD	N/A	Rule-based	Not Found	Not Found	[54]	[55]***
COAT	Clinical Outcomes Assessment Toolkit	Hybrid	Not Found	Not Found	N/A	[56]
CR	Concept Retrieval	Rule-based	Not Found	Not Found	N/A	[57]
CRIS-IE-Smoking	Clinical Record Interactive Search-IE-Smoking	Rule-based	GATE	Yes	N/A	[43]
cTAKES	clinical Text Analysis Knowledge Extraction System	Hybrid	UIMA	Yes	[40,45,58–61]	[30,51,62–68]
EpiDEA	Epilepsy Data Extraction and Annotation	Hybrid	UIMA	Not Found	[69]	[51,70]
FreePharma	N/A	Not Found	Not Found	No	[31,34]	N/A
GATE	General Architecture for Text Engineering	Hybrid	GATE	Yes	N/A	[20,50]
HITEx	Health Information Text Extraction	Not Determined	GATE	Yes	[59,71]	[72]
i2b2 Workbench	Informatics for Integrating Biology & the Bedside	Not Determined	Not Found	Not Found	[73]	[74]***
I2E	N/A	Not Determined	Not Found	Not Found	N/A	[75]
LEXIMER	Lexicon Mediated Entropy Reduction	Machine Learning	Not Found	Not Found	[58,64]	[76]
LifeCode	N/A	Rule-based	Not Found	No	[34,42,58]	[77]***

System Name	Full Name*	NLP Type	Framework	Open-source	Mentioned In**	Described/Used In
LINNAEUS	N/A	Rule-based	Not Found	Not Found	[78]	[79]
LSP-MLP	Linguistic String Project - Medical Language Processor	Rule-based	Not Found	Not Found	[59]	[80]***
MedEx	N/A	Not Found	UIMA	Not Found	[59]	[81]***
MedKATp	Medical Knowledge Analysis Tool pipeline	Rule-based	UIMA	Yes	[59]	[35,39]
MedLEE	Medical Language Extraction and Encoding System	Rule-based	Not Found	No	[48,51,58,59,64,82]	[61,83–92]
MedTagger	N/A	Hybrid	Not Found	Not Found	N/A	[93]
Medtex	N/A	Not Found	Not Found	Not Found	N/A	[94]
MedXN	Medication Extraction and Normalization	Not Determined	Not Found	Not Found	[28]	[95]***
MERKI	Medication Extraction and Reconciliation Knowledge Instrument	Rule-based	Not Found	Yes	[96]	[34]
MetaMap	N/A	Rule-based	Not Found	No	[58,59,93]	[35,53,78,96–98]
MTERMS	Medical Text Extraction, Reasoning and Mapping System	Rule-based	Not Found	Not Found	N/A	[71,99]
Multithreaded Clinical Vocabulary Server	N/A	Rule-based	Not Found	Not Found	[58]	[100]
Natural Language Patient Record	N/A	Not Found	Not Found	No	[34]	N/A
NCBO Annotator	National Center for Biomedical Ontology Annotator	Not Determined	Not Found	Not Found	N/A	[101,102]
NILE	Narrative Information Linear Extraction	Hybrid	Not Found	Not Found	[58]	[103]***
OntoFusion	N/A	Rule-based	Not Found	Not Found	N/A	[104]
ONYX	N/A	Rule-based	Not Found	Not Found	[45,58]	[105]***
OpenNLP	N/A	Hybrid	Not Found	Yes	N/A	[29,45,60,106]
PEP	Pathology Extraction Pipeline	Rule-based	UIMA	Yes	N/A	[39]
RADA	Radiology Analysis	Rule-based	Not Found	Not Found	[64]	[107]***
REDEx	Regular Expression Discovery Extractor	Rule-based	Not Found	Not Found	N/A	[108]
Regextractor	N/A	Rule-based	Not Found	Yes	N/A	[109]
SymText / MPLUS	Symbolic Text Processor	Hybrid	Not Found	No	[58,59,64,83]	[110]***
TagLine	N/A	Hybrid	GATE	Not Found	N/A	[111]
TextMiner	N/A	Rule-based	Not Found	Not Found	N/A	[112]
Valx	N/A	Rule-based	Not Found	Not Found	N/A	[41]

System Name	Full Name*	NLP Type	Framework	Open-source	Mentioned In**	Described/Used In
Abhyankar & Demner-Fushman 2013 (unnamed)	N/A	Rule-based	Not Found	Yes	N/A	[113]
Barrett et al. 2013 (unnamed)	N/A	Hybrid	Not Found	Not Determined	N/A	[114]
Cameron et al. 2012 (unnamed)	N/A	Rule-based	UIMA	Not Found	N/A	[42]
Chen et al. 2015 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[73]
Fang et al. 2014 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[115]
Hao & Weng 2015 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[40]
Hassanpour & Langlotz 2016 (unnamed)	N/A	Machine Learning	Not Found	Not Found	N/A	[64]
He et al. 2014 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[41]
Kadra et al. 2015 (unnamed)	N/A	Not Determined	GATE	Not Found	N/A	[27]
Karystianis et al. 2016 (unnamed)	N/A	Rule-based	Not Found	Yes	N/A	[28]
Korkontzelos et al. 2015 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[29]
Li et al. 2015 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[30]
Martinez et al. 2014 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[35]
Mork et al. 2010 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[96]
Otal et al. 2013 (unnamed)	N/A	MachineLearning	Not Found	Not Found	N/A	[36]
Shah & Martinez 2006 (unnamed)	N/A	Rule-based	Not Found	Not Found	[34]	[31]
Turchin et al. 2014 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[32]
Voorham & Denig 2007 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[116]
Wieneke et al. 2015 (unnamed)	N/A	MachineLearning	Not Found	Not Found	N/A	[37]
Xu et al. 2012 (unnamed)	N/A	Hybrid	Not Found	Not Found	N/A	[117]
Yli-Hietanen et al. 2009 (unnamed)	N/A	Rule-based	Not Found	Not Found	N/A	[118]
Zheng et al. 2015 (unnamed)	N/A	Rule-based	Not Found	No	N/A	[33]

*This information was gathered from sources outside of the final set of papers.

**Only substantive mentions that included some detail about the system were counted.

***These citations did not originate through the systematic review process, but are provided for the sake of convenience.

NLP: Natural Language Processing; **UIMA:** Unstructured Information Management Architecture; **GATE:** General Architecture for Text Engineering.