Journal of Biomedical Informatics 77 (2018) 34–49

Contents lists available at ScienceDirect

Journal of Biomedical Informatics

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

Methodological Review

Clinical information extraction applications: A literature review

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng[1], Saeed Mehrabi[2], Sunghwan Sohn, Hongfang Liu*

Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo Clinic, Rochester, MN, United States

A R T I C L E I N F O

Keywords:
Information extraction
Natural language processing
Application
Clinical notes
Electronic health records

A B S T R A C T

Background: With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research. One critical component used to facilitate the secondary use of EHR data is the information extraction (IE) task, which automatically extracts and encodes clinical information from text.
Objectives: In this literature review, we present a review of recent published research on clinical information extraction (IE) applications.
Methods: A literature search was conducted for articles published from January 2009 to September 2016 based on Ovid MEDLINE In-Process & Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and ACM Digital Library.
Results: A total of 1917 publications were identified for title and abstract screening. Of these publications, 263 articles were selected and discussed in this review in terms of publication venues and data sources, clinical IE tools, methods, and applications in the areas of disease- and drug-related studies, and clinical workflow optimizations.
Conclusions: Clinical IE has been used for a wide range of applications, however, there is a considerable gap between clinical studies using EHR data and studies using clinical IE. This study enabled us to gain a more concrete understanding of the gap and to provide potential solutions to bridge this gap.

1. Introduction

With the rapid adoption of electronic health records (EHRs), it is desirable to harvest information and knowledge from EHRs to support automated systems at the point of care and to enable secondary use of EHRs for clinical and translational research. Following the Health Information Technology for Economic and Clinical Health Act (HITECH Act) legislation in 2009, many health care institutions adopted EHRs, and the number of studies using EHRs has increased dramatically [1]. For example, Ellsworth et al. [2] conducted a review to evaluate methodological and reporting trends in the usability of EHRs; Goldstein et al. [3] evaluated the state of EHR-based risk prediction modeling through a systematic review of clinical prediction studies using EHR data.

However, much of the EHR data is in free-text form [4]. Compared to structured data, free text is a more natural and expressive method to document clinical events and facilitate communication among the care team in the health care environment. One critical component to facilitate the use of EHR data for clinical decision support, quality improvement, or clinical and translation research is the information extraction (IE) task, which automatically extracts and encodes clinical information from text. In the general domain, IE is commonly recognized as a specialized area in empirical natural language processing (NLP) and refers to the automatic extraction of concepts, entities, and events, as well as their relations and associated attributes from free text [5–7]. Most IE systems are expert-based systems that consist of patterns defining lexical, syntactic, and semantic constraints. An IE application generally involves one or more of the following subtasks: concept or named entity recognition that identifies concept mentions or entity names from text (e.g., person names or locations) [8], coreference resolution that associates mentions or names referring to the same entity [9], and relation extraction that identifies relations between concepts, entities, and attributes (e.g., person-affiliation and organization-location) [10].

* Corresponding author at: Division of Biomedical Statistics and Informatics, Mayo Clinic, 200 First St SW, Rochester, MN 55905, United States.
E-mail address: liu.hongfang@mayo.edu (H. Liu).
1 Current address: Seventh Affiliated Hospital of Sun Yat-sen University, Shenzhen, China.
2 Current address: Siemens Healthineers, Mountain View, California, United States.

https://doi.org/10.1016/j.jbi.2017.11.011
Received 30 June 2017; Received in revised form 1 November 2017; Accepted 17 November 2017
Available online 21 November 2017
1532-0464/

NLP focuses on "developing computational models for understanding natural language" [11]. An NLP system can include syntactic processing modules (e.g., tokenization, sentence detection, Part-of-Speech tagging) and/or semantic processing modules (e.g., named entity recognition, concept identification, relation extraction, anaphoric resolution). An IE application is an NLP system with semantic processing modules for extracting predefined types of information from text. In the clinical domain, researchers have used NLP systems to identify clinical syndromes and common biomedical concepts from radiology reports [12], discharge summaries [13], problem lists [14], nursing documentation [15], and medical education documents [16]. Different NLP systems have been developed and utilized to extract events and clinical concepts from text, including MedLEE [17], MetaMap [18], KnowledgeMap [19], cTAKES [20], HiTEX [21], and MedTagger [22]. Success stories in applying these tools have been reported widely [23–34].

A review done by Spyns [35] looked at NLP research in the clinical domain in 1996 and Meystre et al. [11] conducted a review of studies published from 1995 to 2008. Other reviews focus on NLP in a specific clinical area. For example, Yim et al. [36] provided the potential applications of NLP in cancer-case identification, staging, and outcomes quantification; Pons et al. [37] took a close look at NLP methods and tools that support practical applications in radiology. This review focuses on research published after 2009 regarding clinical IE applications.

Another motivation for our review is to gain a concrete understanding of the under-utilization of NLP in EHR-based clinical research. Fig. 1 shows the number of publications retrieved from PubMed using the keywords "electronic health records" in comparison with "natural language processing" from the year 2002 through 2015. We can observe that (1) there were fewer NLP-related publications than EHR-related publications and (2) EHR-related publications increased exponentially from 2009 to 2015, while NLP-related publications increased only moderately. One possible reason is federal incentives for EHR adoption (e.g., HITECH Act), which accelerated the progression of publications about EHR. Having said that, we consider that clinical IE has not been widely utilized in the clinical research community despite the growing availability of open-source IE tools. The under-utilization of IE in clinical studies is in part due to the fact that traditional statistical programmers or study coordinators may not have the NLP competency to extract information from text. Through this literature review, we hope to gain some insights and develop strategies to improve the utilization of NLP in the clinical domain.

2. Methods

We followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) [38] guidelines to perform our review.

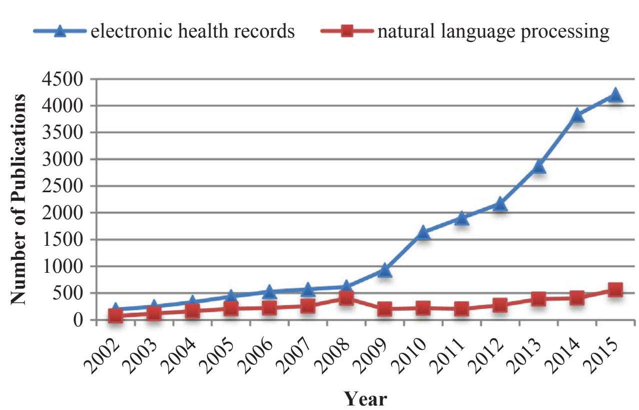

Fig. 1. The number of natural language processing (NLP)-related articles compared to the number of electronic health record (EHR) articles from 2002 through 2015.

2.1. Data sources and search strategies

We conducted a comprehensive search of several databases for articles from January 1, 2009, to September 6, 2016. The databases included Ovid MEDLINE In-Process & Other Non-Indexed Citations, Ovid MEDLINE, Ovid EMBASE, Scopus, Web of Science, and ACM Digital Library. We included articles written in English and excluded those in the form of editorial, review, erratum, letter, note, or comment. The search strategy was designed and conducted by an experienced librarian. The selected keywords and the associations between these keywords were identical for searches in each database: (clinical OR clinic OR electronic health record OR electronic health records) AND (information extraction OR named entity extraction OR named entity recognition OR coreference resolution OR relation extraction OR text mining OR natural language processing) AND (NOT information retrieval). The search strings were carefully designed to be exhaustive and effective for each database and are provided in the Appendix.

2.2. Article selection

The search strategy retrieved 1917 articles after removing duplicates. Nine reviewers (Y.W., L.W., M.R.M., S.M., F.S., N.A., S.L., Y.Z., S.M.) independently screened the titles and abstracts of these articles (each reviewer was given around 210 articles). Articles were excluded based on two criteria: (1) if they were overall unrelated to IE or (2) if they did not use clinical narratives written in English. After this screening process, 415 studies were considered for subsequent categorization. According to the main focus of those studies, one reviewer (Y.W.) categorized each article into one of three categories: (1) application, (2) methodology, or (3) software tool. Eventually, 263 articles were identified as IE application studies, 125 articles focused on proposing new IE methodologies, and 27 articles were about releasing new software tools. In this review, we focus on the 263 articles about clinical IE applications. Thus, those 263 studies underwent full-text review, performed by the same nine reviewers. A flow chart of this article selection process is shown in Fig. 2.

3. Results

In the first analysis, we analyzed the publication venues of the 263 included studies and their data sources. Since clinical IE is an interdisciplinary field of medicine and computer science, publication venues indicate the research communities that have NLP competency to leverage IE techniques. Since developing clinical NLP talent is difficult in large part due to the limited availability of clinical data needed, we provided analysis of data sources used in clinical IE research and the accessibility of these data sources. We hope to provide insight into addressing the data challenge in this domain. Next, we summarized the clinical IE tools and prevalent methods. We provided a list of clinical IE tools used in the 263 articles, an overview of their characteristics (what tools were used for what specific task), and their licenses (are they publically available or not). In addition, the methodologies prevalently adopted in clinical IE were demonstrated. Finally, we described the practical IE applications in the clinical domain, including disease areas that have been studied, drug-related studies, and utility of IE for optimizing clinical workflow. In the statistics presented below, each individual topic is reported. As a result, a single paper, for example, can be counted multiple times if it contains a discussion of multiple IE tools. The details of the included publications and review summaries are provided in the supplementary material.

3.1. Publication venues and data sources

3.1.1. Publication venues

The 263 articles were published in 117 unique venues, comprising 94 journals and 23 conferences. We manually categorized the

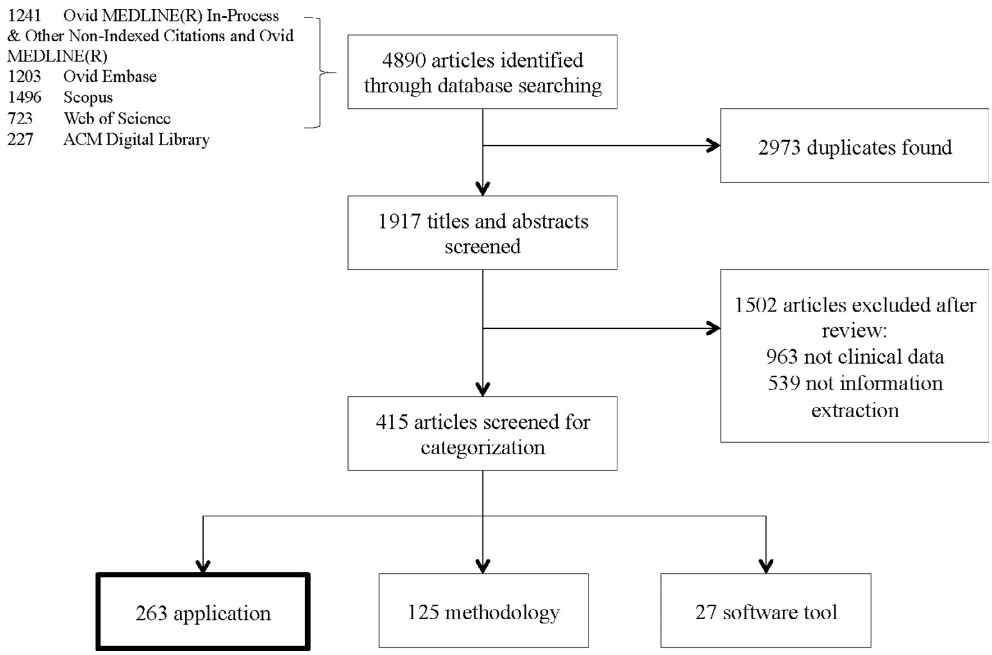

Fig. 2. Article selection flow chart.

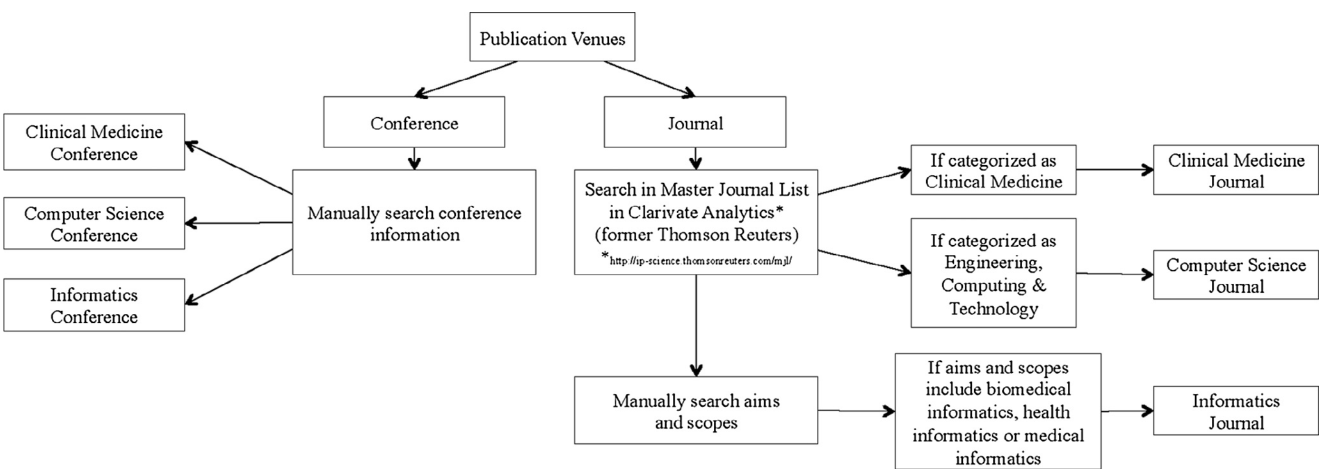

Fig. 3. Categorization of publication venues.

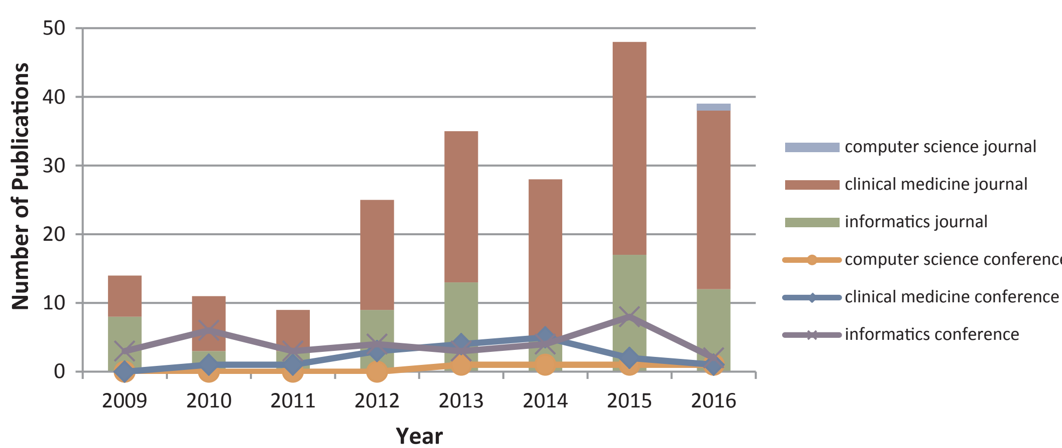

Fig. 4. Distribution of included studies, stratified by category and year (from January 1, 2009, to September 6, 2016).

publication venues into three categories: (1) clinical medicine, (2) informatics, and (3) computer science. The categorization process is summarized in Figs. 3 and 4 shows the number of included studies in each category.

We observed that the number of journal articles in the categories of clinical medicine and informatics are much larger than the number of conference articles in these categories; those findings were shown to be inversed in the category of computer science. Though the number of publications from informatics journals is smaller compared to clinical medicine journals, it shows that there are more informatics conference publications than other conference publications. The reason might be that informatics conferences, e.g., the American Medical Informatics Association (AMIA) Annual Symposium, recruit more regular papers than clinical medicine conferences. Overall, clinical medicine journals are the most popular venues for IE application publications.

Papers in the clinical medicine category are published in a variety of clinical-specific journals, such as *Arthritis & Rheumatism*. Publications in informatics are mostly published in two venues: (1) *Journal of the American Medical Informatics Association* (n = 26, n denotes the number of publications hereafter) and (2) *AMIA Annual Symposium Proceedings/AMIA Symposium* (n = 24). In Fig. 4, we observe a generally increasing trend of IE publications, except for the years 2014 and 2016 (due to the partial-year retrieval). This might be due to the World Congress on Medical and Health Informatics occurring bi-annually (MedInfo, odd year only, n = 13). We note that the MedInfo proceedings are published as special issues in Studies in Health Technology and Informatics, which is categorized as clinical medicine journal. Fig. 4 also shows an increasing attention and demand in the application of IE techniques in both the clinical research and informatics communities. Interestingly, although IE is a traditional research topic in computer science, only one computer science journal and a few computer science conferences (e.g., International Conference of the Italian Association for Artificial Intelligence, International Conference on System Sciences) are found. Overall, the top five publication venues having the largest number of publications are: (1) *Journal of the American Medical Informatics Association* (n = 26), (2) *AMIA Annual Symposium Proceedings/AMIA Symposium* (n = 24), (3) *Pharmacoepidemiology and Drug Safety* (n = 16), (4) *Studies in Health Technology and Informatics* (n = 13), and (5) *Journal of Biomedical Informatics* (n = 10). The results suggest that only a small portion of papers in JAMIA and AMIA focus on the use of NLP tools for clinical applications. This may be partially due to the tendency of the academic informatics community to prefer innovations in methodology rather than research reporting the use of informatics tools. It may also be due to the dependency and the lack of clear distinction of NLP with relevant fields, such as data mining and knowledge management on text data.

3.1.2. Data sources

The majority of the 263 studies were conducted in the United States (n = 236), while others were conducted in Canada (n = 9), United Kingdom (n = 5), Australia (n = 3), and other countries. Among the 236 US studies, 163 used only clinical documents and 56 used both clinical documents and structured EHR data, such as *International Statistical Classification of Diseases, Ninth Revision* (ICD-9) codes (n = 25). We found that other resources were also used in conjunction with clinical data, such as biomedical literature (n = 3) and health-related websites (n = 2).

Table 1 shows the number of papers with diverse types of clinical documents being used. Here, we classify clinical documents into two main categories, clinical notes and diagnostic reports. Clinical notes refer to documentation of a patient's visit with a health care provider, which may include the patient's medical/social history and physical examination, clinical observations, summaries of diagnostic and therapeutic procedures, plan of treatment, and instructions to the patients which can be telephonic or electronic interactions with the patient. Diagnostic reports refer to the reports provided by diagnostic services, such as laboratory reports, radiology reports, and pathology reports. We counted the number of publications according to their mentions of note types in the papers and listed the most frequently used note types with brief descriptions for clinical notes and diagnostic reports in Table 1. Most of the studies were conducted by the following institutions: US Department of Veterans Affairs (VA) (n = 34), Mayo Clinic (n = 12), Vanderbilt University (n = 8), Humedica (n = 7), and Kaiser Permanente (n = 7), either within individual institutions or through collaboration across multiple institutions.

We summarized the time range of clinical data utilized in those studies and found that the time period ranged from 1987 through 2015. We counted the number of studies using the data in each specific year and these results are shown in Fig. 5. The average time span of the clinical data used in the selected papers was 6.77 years. A rapid growth of data can be observed since 1995, and the amount of data utilized in those studies reached a peak in 2009. A large quantity of EHR data became available after 2009. However, Fig. 5 implies that these data have not been adequately utilized by clinical IE studies.

Note that clinical documents in individual institutions are not accessible to external researchers without collaborative projects, and only a few EHR data sets are accessible to external researchers. Here, we introduce four important clinical text corpora. The first is the i2b2 NLP Challenges data (n = 14), where fully de-identified notes from the Research Patient Data Repository at Partners HealthCare were created for a series of NLP challenges, 1500 notes of which have been released. In order to access these notes, one needs to register at the i2b2 website (https://www.i2b2.org/NLP/DataSets/) and submit a proposal which is then reviewed by the i2b2 organizers. The second is MIMIC II (n = 2) [39], a data set consisting of EHR data for over 40,000 de-identified intensive care unit stays at the Beth Israel Deaconess Medical Center, including clinical notes, discharge summaries, radiology reports, laboratory results, and structured clinical data. Physiologic time series are accessible publicly (https://physionet.org/physiobank/database/mimic2db/), and clinical data are accessible with a data use agreement (see http://physionet.org/mimic2/mimic2_access.shtml). The third corpus is MTsamples, which is a large collection of publicly available transcribed medical reports (http://www.mtsamples.com/). It contains sample transcription reports, provided by various transcriptionists for many specialties and different work types, and thus the accuracy and quality of the notes is not guaranteed [40]. Finally, the THYME corpus [41] contains de-identified clinical, pathology, and radiology records for a large number of patients, focusing on brain and

Table 1
The most frequently used note types for clinical notes (top 5) and diagnostic reports (top 3) and the corresponding brief descriptions and number of papers in the included publications.

	Note type	Brief description	No. of papers
Clinical notes	Discharge summaries	A document that describes the outcome of a patient's hospitalization, disposition, and provisions for follow-up care.	26
	Progress notes	A document that describes a patient's clinical status or achievements during the course of a hospitalization or over the course of outpatient care.	15
	Admission notes	A document that describes a patient's status (including history and physical examination findings), reasons why the patient is being admitted for inpatient care to a hospital or other facility, and the initial instructions for that patient's care.	9
	Operative notes	A document that describes the details of a surgery.	5
	Primary care notes	A document that describes the details of an outpatient during a primary care.	3
Diagnostic reports	Radiology reports	Results of radiological scans and X-ray images of various parts of the patient's body and specific organs.	43
	Pathology reports	Results of pathological examinations of tissue samples and tissues of organs removed during surgical procedures.	22
	Colonoscopy reports	Results of a colonoscopy.	4

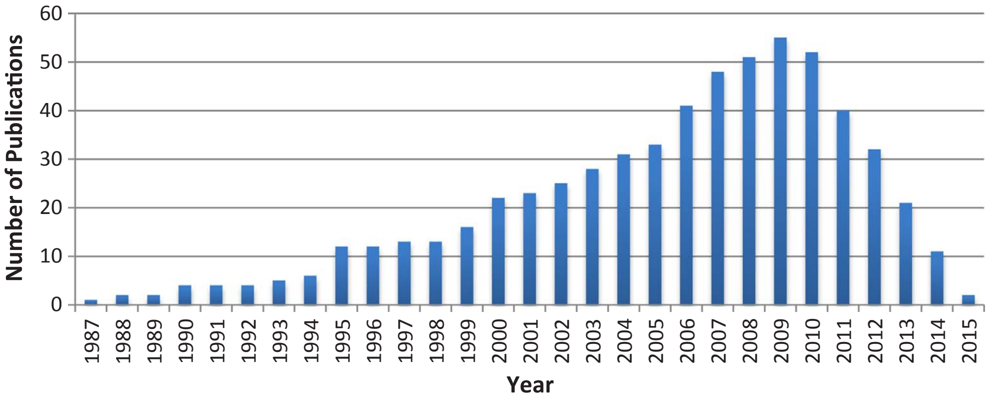

Fig. 5. The distribution of studies in terms of clinical narrative data utilized per year.

colon cancer from a large healthcare practice (Mayo Clinic). It also provides NLP annotations, created by annotators and adjudicators at the University of Colorado at Boulder and Boston Harvard Children's Medical Center, including temporal entity and relation, coreference, and UMLS named entity. It is available to researchers involved in NLP research under a data use agreement with Mayo Clinic (see https://github.com/stylerw/thymedata and https://clear.colorado.edu/TemporalWiki/index.php/Main_Page).

3.2. Implementations

In the next section, we briefly report the frameworks, tools, and toolkits being utilized in the selected publication. The second part summarizes two main categories of methods being used for clinical IE: rule-based and machine learning. These two areas were analyzed separately so readers can explore them based on their interests. Finally, we introduce the efforts of clinical IE-related NLP shared tasks in the community.

3.2.1. Clinical information extraction tools

The clinical IE tools used in the 263 studies included are summarized in Table 2. The most frequently used tools for IE in the clinical domain are cTAKES [20] (n = 26), MetaMap [18] (n = 12), and MedLEE [17] (n = 10). cTAKES, developed by Mayo Clinic and later transitioned to an Apache project, is the most commonly used tool. It is built upon multiple Apache open-source projects, the Apache Unstructured Information Management Architecture (UIMA) framework [42] and the Apache OpenNLP toolkit [43]. It contains several analysis engines for various linguistics and clinical tasks, such as sentence detection, tokenization, part-of-speech tagging, concept detection, and normalization. cTAKES has been adopted for identification of patient phenotype cohorts [28,44–54], smoking status extraction [55–58], genome-wide association studies [30], extraction of adverse drug events [59], detection of medication discrepancies [60], temporal relation discovery [61], risk stratification [25], and risk factor identification [62] from EHRs. MetaMap was developed by the National Library of Medicine (NLM) with the goal of mapping biomedical text to the Unified Medical Language System (UMLS) Metathesaurus, or vice versa. It was originally developed to improve biomedical text retrieval of MEDLINE/PubMed citations. Later, MetaMap's ability was improved to process clinical text [63], which is reflected by the large number of studies using MetaMap for clinical IE tasks. In the included studies, MetaMap has been used for phenotype extraction [31,64–69], assessment of emergency department use [27,70], drug-disease treatment relationships [71], fragment recognition in clinical documents [72], and extraction of patient-related attributes [73]. MedLEE is one of the earliest clinical NLP systems developed and is mostly used for pharmacovigilance [26,74,75] and pharmacoepidemiology [76,77].

Other tools focus more on one specific task. For example, GATE [78,79], NLTK [80], and OpenNLP [81] are typically used for various NLP preprocessing tasks, such as sentence boundary detection, tokenization, and part-of-speech (POS) tagging; MedEx [7] focuses on extracting drug names and doses; MALLET [82] and WEKA [83] are used for IE tasks that leverage machine learning algorithms, such as classification, clustering, and topic modeling; and Protégé [84] is a tool that has been frequently used for ontology building. Note that the tools summarized in this review are from the 263 application articles and that many IE tools, such as TextHunter [85], Patrick et al's cascaded IE tool [86], KneeTex [87], Textractor [88], and NOBLE [89], in the 27 tool articles and the 125 methodology articles (many of them are participant systems in shared tasks) are not included in this review and subject to a future study.

3.2.2. Methods for clinical information extraction

Approaches to clinical IE can be roughly divided into two main categories: rule-based and machine learning. Rule-based IE systems primarily consist of rules and an interpreter to apply the rules. A rule is usually a pattern of properties that need to be fulfilled by a position in the document. A common form of the rule is a regular expression that uses a sequence of characters to define a search pattern. Among the included 263 articles, 171 (65%) used rule-based IE systems. For example, Savova et al. [51] used regular expressions to identify peripheral arterial disease (PAD). A positive PAD was extracted if the pre-defined patterns were matched (e.g., "severe atherosclerosis" where "severe" was from a list of modifiers associated with positive PAD evidence and "atherosclerosis" was from a dictionary tailored to the specific task of PAD discovery). Another form of the rule is logic. Sohn and Savova [57] developed a set of logic rules to improve smoking status classification. In their approach, they first extracted smoking status for each sentence and then utilized precedence logic rules to determine a document-level smoking status. Current smoker has the highest precedence, followed by past smoker, smoker, non-smoker, and unknown (e.g., if current smoker was extracted from any sentence in a document, then the document was labeled as current smoker). The final patient-level smoking status was based on similar logic rules (e.g., if there is a current smoker document but no past smoker document belonging to a patient, then the patient was assigned as a current smoker). A clinical IE system is often composed of many rules that are written by a human-knowledge engineer. The rule could be developed through two means, manual knowledge engineering (78 studies) and leveraging knowledge bases (53 studies), or a hybrid system (40 studies). Manual knowledge engineering can be time consuming and requires collaboration with physicians. It is usually very accurate, since it is based on physicians' knowledge and experience. Sohn, Savova, and colleagues [51] provide examples of successful applications. A knowledge base is a computerized database system that stores complex structured information, such as UMLS (medical concepts), phenome-wide association studies (PheWAS) [90] (disease-gene relations), and DrugBank [91] (drug-gene

Table 2
IE frameworks, tools and toolkits used in the included publications.

Name	Description	License	Website	No. of Papers
Frameworks				
UIMA	Software framework for the analysis of unstructured contents like: text, video and audio data	Apache	https://uima.apache.org/	31
GATE	Java-based open-source software for various NLP tasks such as information extraction and semantic annotation	GNU Lesser General Public License	https://gate.ac.uk/	5
Protégé	Open-source ontology editor and framework for building intelligent systems	MIT License	http://protege.stanford.edu/	1
Tools				
cTAKES	Open-source NLP system based on UIMA framework for extraction of information from electronic health records unstructured clinical text	Apache	http://ctakes.apache.org/	26
MetaMap	National Institutes of Health (NIH)-developed NLP tool that maps biomedical text to UMLS concepts	UMLS Metathesaurus	https://metamap.nlm.nih.gov/	12
MedLEE	NLP system that extracts, structures, and encodes clinical information from narrative clinical notes	NLP International for commercial use	http://zellig.cpmc.columbia.edu/medlee/	10
KnowledgeMap Concept Indexer (KMCI)	NLP system that identifies biomedical concepts and maps them to UMLS concepts	Vanderbilt License	https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/kmci-knowledgemap-concept-indexer	4
HITEx	Open-source NLP tool built on top of the GATE framework for various tasks such as principal diagnoses extraction and smoking status extraction	i2b2 Software License Agreement	https://www.i2b2.org/software/projects/hitex/hitex_manual.html	4
MedEx	NLP tool used to recognize drug names, dose, route, and frequency from free-text clinical records	Apache	https://medschool.vanderbilt.edu/cpm/center-precision-medicine-blog/medex-tool-finding-medication-information	4
MedTagger	Open-source NLP pipeline based on UIMA framework for indexing based on dictionaries, information extraction, and machine learning–based named entity recognition from clinical text	Apache	http://ohnlp.org/index.php/MedTagger	3
ARC	Automated retrieval console (ARC) is an open-source NLP pipeline that converts unstructured text to structured data such as Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) or UMLS codes	Apache	http://blulab.chpc.utah.edu/content/arc-automated-retrieval-console	2
Medtex	Clinical NLP software that extracts meaningful information from narrative text to facilitate clinical staff in decision-making process	No license information available	https://aehrc.com/research/projects/medical-free-text-retrieval-and-analytics/#medtex	2
CLAMP	NLP software system based on UIMA framework for clinical language annotation, modeling, processing and machine learning	Software Research License	https://sbmi.uth.edu/ccb/resources/clamp.htm	1
MedXN	A tool to extract comprehensive medication information from clinical narratives and normalize it to RxNorm	Apache	http://ohnlp.org/index.php/MedXN	1
MedTime	A tool to extract temporal information from clinical narratives and normalize it to the TIMEX3 standard	GNU General Public License	http://ohnlp.org/index.php/MedTime	1
PredMED	NLP application developed by IBM to extract full prescriptions from narrative clinical notes	Commercial		1
SAS Text Miner	A plug-in for SAS Enterprise Miner environment provides tools that enable you to extract information from a collection of text documents and uncover the themes and concepts that are concealed in them.	Commercial		1
Toolkits				
WEKA	Open source toolkit that contains various machine learning algorithms for data-mining tasks	GNU General Public License	http://www.cs.waikato.ac.nz/ml/weka/	5
MALLET	Java-based package for various NLP tasks such as document classification, information extraction, and topic modeling	Common Public License	http://mallet.cs.umass.edu/	4
OpenNLP	Open-source machine learning toolkit for processing of natural language text	Apache	https://opennlp.apache.org/	1
NLTK	Python-based NLP toolkit for natural language text	Apache	http://www.nltk.org/	1
SPLAT	Statistical parsing and linguistic analysis toolkit (SPLAT) is a linguistic analysis toolkit for natural language developed by Microsoft research	Commercial	https://www.microsoft.com/en-us/research/project/msr-splat/	1

Table 3
The most frequently used machine learning methods (top 6) and the corresponding number of papers in the included publications.

Method	No. of Papers
Support Vector Machine (SVM)	26
Logistic regression (LR)	11
Conditional random field (CRF)	9
Decision Tree (DT)	8
Naïve Bayes (NB)	6
Random Forest (RF)	4

relations). For example, Martinez et al. [69] mapped phrases into UMLS medical concepts by MetaMap; Hassanpour and Langlotz [53] used RadLex, a controlled lexicon for radiology terminology, to identify semantic classes for terms in radiology reports; and Elkin et al. [92] coded signs, symptoms, diseases, and other findings of influenza from encounter notes into Systematized Nomenclature of Medicine – Clinical Terms (SNOMED CT) medical terminology.

Machine learning–based IE approaches have gained much more interest due to their efficiency and effectiveness [93–95], particularly their success in many shared tasks [96]. Among the 263 included studies, 61 articles have illustrations on using machine learning algorithms. Some articles included different machine learning approaches for evaluation purposes. We took all of those approaches into consideration and counted their frequency of appearance and listed the six most frequently used methods in Table 3. Support Vector Machine (SVM) is the most frequently employed method by researchers. Barrett et al. [97] integrated feature-based classification (SVM) and template-based extraction for IE from clinical text. Roberts et al. [94] proposed an approach to use SVM with various features to extract anatomic sites of appendicitis-related findings. Sarker et al. [98] proposed an automatic text classification approach for detecting adverse drug reaction using SVM. Himes et al. [99] conducted a study to classify chronic obstructive pulmonary disease with SVM among asthma patients recorded in the electronic medical record. Logistic regression (LR) is mostly used for entity and relation detections. For example, Chen et al. [100] applied LR to detect geriatric competency exposures from students' clinical notes; and Rochefort et al. [101] used multivariate LR to detect events with adverse relations from EHRs. Conditional random field (CRF) is another widely used method in many papers for the purpose of entity detection. For example, Deleger et al. [23] used CRF to extract Pediatric Appendicitis Score (PAS) elements from clinical notes; and Li et al. [60] used it to detect medication names and attributes from clinical notes for automated medication discrepancy detection. Based on our observation, many machine learning algorithms leveraged outputs from IE as features. For example, Yadav et al. [102] used IE tools to extract medical word features and then utilized those features as input for a decision tree to classify emergency department computed tomography imaging reports. Some researchers compared different machine learning approaches in one paper for the purpose of performance comparison. For example, to better identify patients with depression in free-text clinical documents, Zhou et al. [86] compared SVM, Generalized nearest neighbor (NNge), Repeated Incremental Pruning to Produce Error Propositional Rule (RIPPER), and DT for performance evaluation, and found that DT and NNge yielded the best F-measure with high confidence, while RIPPER outperformed other approaches with intermediate confidence.

3.2.3. Clinical IE-related NLP shared tasks

Multiple clinical NLP shared tasks have leveraged community efforts for methodology advancement. Though we have categorized most studies resulting from those shared tasks as methodology publications, we would like to briefly describe those shared tasks due to their significant impact on the clinical NLP research. Table 4 summarizes the most recognizable clinical IE-related NLP shared tasks in the community.

3.3. Applications of clinical information extraction

In this section, we summarize the application of clinical IE in terms of disease study areas, drug-related study areas, and clinical workflow optimization.

3.3.1. Disease study areas

IE for phenotyping accounted for a large portion of the studies. Among 263 papers, 135 focused on IE of 88 unique diseases or conditions from clinical notes, pathology reports, or radiology reports. For further analysis, we used ICD-9 to categorize diseases, as shown in Table 5. Our findings showed that the neoplasms category was the most studied disease area (e.g., hepatocellular cancer [120] and colorectal cancer [121]), followed by diseases of the circulatory system (e.g., heart failure [122] and peripheral arterial disease [51]), diseases of the digestive system (e.g., pancreatic cyst [123] and celiac disease [124]), diseases of the nervous system (e.g., headache [125], endocrine, nutritional, and metabolic diseases), and immunity disorders (e.g., diabetes mellitus [126]).

The included IE studies involved 14 disease categories among a total of 19 ICD-9 categories. Five disease areas were not covered in these studies (i.e., diseases of the sense organs; complications of pregnancy, childbirth, and the puerperium; congenital anomalies; certain conditions originating in the perinatal period; and external causes of injury and supplemental classification). Recent studies showed a research trend to look further into refined diseases with specific features (e.g., drug-resistant pediatric epilepsy [127], severe early-onset childhood obesity [49], non-severe hypoglycemic events [128], and neuropsychiatric disorder [129]). This research trend reflects the fact that IE techniques could play an important role when exact ICD-9 codes are not available for data extraction. IE has been used to identify patients having rare diseases with no specific ICD-9 diagnosis codes, such as acquired hemophilia [130]. The most frequently studied individual diseases (focused by more than 5 papers) were cancer, venous thromboembolism, PAD, and diabetes mellitus.

Various aspects of malignancy have been extensively focused, including identifying specific cancer type [131] or molecular testing data in a specific cancer type [132], cancer recurrence [44], diagnosis, primary site, laterality, histological type/grade, metastasis site/status [133], cancer metastases [134], and cancer stage [135]. Mehrabi et al. [131] developed a rule-based NLP system to identify patients with a family history of pancreatic cancer. This study showed consistent precision across the institutions ranging from 0.889 in the Indiana University (IU) dataset to 0.878 in the Mayo Clinic dataset. Customizing the algorithm to Mayo Clinic data, the precision increased to 0.881. Carrell et al. [44] developed an NLP system using cTAKES to process clinical notes for women with early-stage breast cancer to identify whether recurrences were diagnosed and if so, the timing of these diagnoses. The NLP system correctly identified 0.92 of recurrences with 0.96 specificity. Farrugia et al. proposed an NLP solution for which preliminary results of correctly identifying primary tumor stream, metastases, and recurrence are up to 0.973 [134]. Nguyen et al. [133] used Medtex to automatically extract cancer data and achieved an overall recall of 0.78, precision of 0.83, and F-measure of 0.80 over seven categories, namely, basis of diagnosis, primary site, laterality, histological, histological grade, metastasis site, and metastatic status. Warner et al. [135] developed an NLP algorithm to extract cancer staging information from narrative clinical notes. The study looked at the four stages of lung cancer patients and showed that the algorithm was able to calculate the exact stage of 0.72 of patients.

To extract venous thromboembolism, Tian et al. [136] used unigrams, bigrams, and list of negation modifiers to develop rules for identifying if a sentence from clinical reports refers to positive case of

Table 4
Clinical IE-related NLP shared tasks.

Shared Task	Year	Brief Description	No. of Participants	Best Participant Performance (F-measure)	Website
i2b2 de-identification and smoking challenge [103,104]	2006	Automatic de-identification of personal health information and identification of patient smoking status.	15	De-identification: 0.98; Smoking identification: 0.90.	https://www.i2b2.org/NLP/DataSets/
i2b2 obesity challenge [105]	2008	Identification of obesity and its co-morbidities.	30	0.9773	
2b2 medication challenge [106]	2009	Identification of medications, their dosages, modes (routes) of administration, frequencies, durations, and reasons for administration in discharge summaries.	20	Durations identification: 0.525; Reason identification: 0.459.	
i2b2 relations challenge [107]	2010	Concept extraction, assertion classification and relation classification.	30	Concept extraction: 0.852; Assertion and relation classification: 0.936.	
i2b2 coreference challenge [108]	2011	Coreference resolution.	20	0.827	
i2b2 temporal relations challenge [109]	2012	Extraction of temporal relations in clinical records including three specific tasks: clinically significant events, temporal expressions and temporal relations.	18	Event: 0.92; Temporal expression: 0.90; Temporal relation: 0.69.	
i2b2 de-identification and heart disease risk factors challenge [110,111]	2014	Automatic de-identification and identification of medical risk factors related to coronary artery disease in the narratives of longitudinal medical records of diabetic patients.	10	De-identification: 0.9586; Risk factor: 0.9276.	
CLEF eHealth shared task 1 [112]	2013	Named entity recognition in clinical notes.	22	0.75	https://sites.google.com/site/shareclefehealth/
CLEF eHealth shared task 2 [113]	2014	Normalization of acronyms or abbreviations.	10	Task 2a: 0.868 (accuracy);Task 2b: 0.576 (F-measure) .	https://sites.google.com/site/clefehealth2014/task-2
CLEF eHealth shared task 1b [114]	2015	Clinical named entity recognition from French medical text.	7	Plain entity recognition: 0.756; Normalized entity recognition: 0.711; Entity normalization: 0.872.	https://sites.google.com/site/clefehealth2015/task-1/task-1b
CLEF eHealth shared task 2 [115]	2016	Clinical named entity recognition from French medical text.	7	Plain entity recognition: 0.702; Normalized entity recognition: 0.529; Entity normalization: 0.524.	https://sites.google.com/site/clefehealth2016/task-2
SemEval task 9 [116]	2013	Extraction of drug-drug interactions from biomedical texts.	14	Recognition of drugs: 0.715; Extraction of drug-drug interactions: 0.651.	https://www.cs.york.ac.uk/semeval-2013/task9.html
SemEval task 7 [117]	2014	Identification and normalization of diseases and disorders in clinical reports.	21	Identification: 0.813;Normalization: 0.741 (accuracy) .	http://alt.qcri.org/semeval2014/task7/index.php?id=task-description
SemEval task 14 [118]	2015	Named entity recognition and template slot filling for clinical texts.	16	Named entity recognition: 0.757;Template slot filling: 0.886 (accuracy);Disorder recognition and template slot filling: 0.808 (accuracy) .	http://alt.qcri.org/semeval2015/task14/
SemEval task 12 [119]	2016	Temporal information extraction from clinical texts including time expression identification, event expression identification and temporal relation identification.	14	Time expression identification: 0.795; Event expression identification: 0.903; Temporal relation identification: 0.573.	http://alt.qcri.org/semeval2016/task12/

Table 5
Application areas of clinical IE and the corresponding number of publications.

Application Areas	No. of Papers
Disease study areas	
Neoplasms	27
Diseases of the circulatory system	23
Diseases of the digestive system	12
Diseases of the nervous system	12
Endocrine, nutritional and metabolic diseases, and immunity disorders	12
Mental disorders	12
Diseases of the respiratory system	11
Injury and poisoning	8
Diseases of the musculoskeletal system and connective tissue	6
Symptoms, signs, and ill-defined conditions	5
Infectious and parasitic diseases	3
Diseases of the genitourinary system	2
Diseases of the blood and blood-forming organs	1
External causes of injury and supplemental classification	1
Drug-related studies	
Adverse drug reaction	3
Medication extraction	9
Drug exposure	2
Drug-treatment classification	1
Dosage extraction	3
Clinical workflow optimization	
Adverse events	5
Quality control	8
Patient management	6
Measurement value extraction	8

deep vein thrombosis (DVT) or Pulmonary embolism, and NLP achieved 0.94 sensitivity, 0.96 specificity and 0.73 PPV for DVT. McPeek Hinz et al. [137] tried to capture both acute and historical cases of thromboembolic disease using a general purpose NLP algorithm, and obtained a positive predictive value of 0.847 and sensitivity of 0.953 for an F-measure of 0.897.

For PAD, Savova et al. [51] used cTAKES to identify four groups of PAD patients, positive, negative, probable and unknown based on radiology reports, and the positive predictive value was in the high 90s. Duke et al. [138] implemented an NLP system to improve identification of PAD patients from EHR. The results showed that using unstructured data is able to identify more PAD patients compared to structured data. The NLP system was able to identify 98% of PAD patients in their dataset but when only structured data was used only 22% of PAD patients were captured. The NLP system developed by Afzal et al. [139] ascertained PAD status from clinical notes with sensitivity (0.96), positive predictive value (0.92), negative predictive value (0.99), and specificity (0.98).

Currently extraction of diabetes from clinical text can achieve a performance score of over 0.95. For example, Wei et al. [140] combined NLP, a machine learning algorithm (e.g., SVM), and ontology (SNOMED-CT) for the automatic identification of patients with Type 2 Diabetes Mellitus, achieving an F-measure of above 0.950.

3.3.2. Drug-related studies

Out of 263 papers in our collection, 17 used IE for drug-related studies. Table 5 shows our categorization of drug-related studies and the number of papers in each category. In this section, we review papers in each category and highlight their novelties.

3.3.2.1. Drug-named entity recognition. One of the main components in drug-related studies is identifying drug names in clinical notes. Most of these studies used a rule-based keyword search approach. MedEx, developed by Xu et al. [141], has been applied in several studies, such as the application in [142]. MedEx is a rule-based system that extracts medication name, strength, route, and frequency. The system was evaluated on 50 discharge summaries, and an F-measure of 0.93 was reported. Sohn et al. [143] studied semantic and context patterns for describing medication information in clinical notes. They analyzed two different corpora: 159 clinical notes from Mayo Clinic and 253 discharge summaries from the i2b2 shared task. They illustrated that 12 semantic patterns cover 95% of medication mentions. Zheng et al. [144] developed an NLP system to extract mentions of aspirin use and dosage information from clinical notes. The system had several components, including sentence splitting, tokenization, part-of-speech tagging, etc. To identify the mentions, the system used a keyword search plus a word-sense disambiguation component. The authors trained the systems on 2949 notes and evaluated it on 5339 notes. The system achieved 0.955 sensitivity and 0.989 specificity.

3.3.2.2. Dosage information extraction. A few drug-related studies focused on extracting dosage information from clinical notes. Xu et al. [145] extended MedEx to extract dosage information from clinical notes and then calculated daily doses of medications. They tested the system for tacrolimus medication on four data sets and reported precision in the range of 0.90–1.0 and a recall rate of 0.81–1.0. In another study, Xu et al. [24] evaluated MedEx in an automating data-extraction process for pharmacogenetic studies. The study used a cohort of patients with a stable warfarin dose. They evaluated the system on 500 physician-annotated sentences and achieved 0.997 recall and 0.908 precision. The extracted information was used to study the association between the dose of warfarin and genetic variants.

3.3.2.3. Adverse drug reaction detection. We identified three research studies on extracting adverse drug reactions (ADRs) from clinical notes. Wang et al. [75] conducted the first study to use unstructured data in EHR for identifying an ADR. In this study, the authors used MedLEE to identify medication entities and events. They considered co-occurrences of entities and events as indications of ADR. The system evaluated for seven drug classes and their known ADRs; the authors reported 0.75 recall and 0.31 precision. Sohn et al. [59] developed two systems, a rule-based system to discover individual adverse effects and causative drug relationships, and a hybrid system of machine learning (C4.5-based decision tree) and a rule-based system to tag sentences containing adverse effects. They evaluated the system in the domain of psychiatry and psychology and reported 0.80 F-measure for the rule-based system and 0.75 for the hybrid system. Haerian et al. [26] studied ADRs from another perspective, confounders. They designed and implemented an NLP system to identify cases in which the event is due to a patient's disease rather than a drug. They evaluated the system for two ADRs, rhabdomyolysis and agranulocytosis, and reported 0.938 sensitivity and 0.918 specificity.

Conclusions from these studies show that ADR identification is a complex task and needs more sophisticated systems. Nevertheless, the mentioned systems could assist experts in the process of manual review of ADR identification.

3.3.2.4. Drug exposure extraction. Liu et al. [146] and Feng et al. [147] developed NLP systems to determine patient drug exposure histories. The former system, which is a hybrid system of NLP and machine learning, first identifies drug names and then drug events. While detecting drug events, the system labels drug mentions with an "on" or "stop" label. Finally, the system models drug exposure for a patient based on temporal information for each drug. The authors evaluated the system for warfarin exposure and reported 0.87 precision and 0.79 recall. The latter system used NLP to identify drug exposure histories for patients exposed to multiple statin dosages.

3.3.3. Clinical workflow optimization

Many studies leveraged clinical IE to improve and optimize clinical workflow. Table 5 lists four categories of clinical workflow and the number of papers in each category. In this section, we review papers in each category and highlight their novelties.

3.3.3.1. Adverse Event detection. Adverse events (AEs) are injuries caused by medical management rather than the underlying condition of the patient. Automated IE tools have been developed to detect AEs. Rochefort et al. [101] utilized rules to detect AEs of (1) hospital-acquired pneumonias, (2) central venous catheter–associated bloodstream infections, and (3) in-hospital falls. Receiver operating characteristic (ROC) was used to find the optimal threshold for detection of AEs based on values of blood cell counts, abnormal ventilator settings, or elevated body temperature. In another of their studies [148], Rochefort and colleagues used similar techniques to detect three highly prevalent AEs in elderly patients: (1) DVT, (2) pulmonary embolism (PE), and (3) pneumonia. Zhang et al. [149] extracted information on adverse reactions to statins from a combination of structured EHR entries. Hazlehurst et al. [150] used an NLP software, MediClass, to detect vaccine AEs based on concepts, terms, and rules. Baer et al. [151] developed Vaccine Adverse Event Text Mining (VaeTM) to extract features about AEs, including diagnosis and cause of death, from clinical notes. They found that the clinical conclusion from VaeTM agreed with the full text in 93% of cases, even though 74% of words were reduced.

3.3.3.2. Quality control. Inappropriate emergency department (ED) usage increases the workload of emergency care services due to the fact that patients with non-urgent problems make up a substantial proportion of ED visits. Using IE to automatically identify inappropriate ED caseloads could accurately predict inappropriate use. In two studies [27,70], researchers used GATE- and MetaMap-extracted biopsychosocial concepts from the primary care records of patients and studied their relationship to inappropriate use of ED visits. The study [27] extracted over 38 thousand distinct UMLS codes from 13,836 patients' primary records; and the codes of mental health and pain were associated with inappropriate ED room use with statistical significance ($p < .001$). It showed the feasibility of using IE to reduce inappropriate ED usage. Tamang et al. [152] utilized rules to detect unplanned care in EHRs, such as emergency care, unplanned inpatient care, and a trip to an outpatient urgent care center, in order to reduce these unplanned care episodes.

Researchers from UCLA conducted quality assessment of radiologic interpretations using, as a reference, other clinical information, such as pathology reports [153]. They developed a rule-based system to automatically extract patient medical data and characterize concordance between clinical sources, and showed the application of IE tools to facilitate health care quality improvement.

The increased use of imaging has resulted in repeated imaging examinations [154]. Ip et al. [155] utilized GATE [78] to extract imaging recommendations from radiology reports and quantify repeat imaging rates in patients. Since ADR is an important quality metric for colonoscopy performance, a few studies showed the application of IE tools in automatically extracting components to calculate ADR. Mehrotra and Harkema [156] developed an IE tool to measure published colonoscopy quality indicators from major gastroenterology societies, including documentation of cecal landmarks and bowel preparation quality. Raju et al. [157,158] developed an NLP program to identify adenomas and sessile serrated adenomas from pathology reports for reporting ADR. Gawron et al. [159] developed a flexible, portable IE tool—QUINCE—to accurately extract pathology results associated with colonoscopies, which is useful for reporting ADRs across institutions and health care systems.

3.3.3.3. Patient management. Popejoy et al. [15] described a care coordination ontology that was built to identify and extract care coordination activities from nursing notes and show how these activities can be quantified. Activities include communication and/or management of elderly patient needs. The study by Gundlapalli et al. [160] aimed to detect homeless status using free-text Veterans Affairs (VA) EHRs. In this study, a total of 356 concepts about risk factors among the homeless population were categorized into eight categories, including direct evidence, "doubling up," mentions of mental health diagnoses, etc.

Arranging and documenting follow-up appointments prior to patient dismissal is important in patient care. Information contained in the dismissal record is beneficial for performance measurement to support quality improvement activities and quality-related research. Ruud et al. [161] used the SAS text mining tool (SAS Text Miner) [162] to extract date, time, physician, and location information of follow-up appointment arrangements from 6481 free-text dismissal records at Mayo Clinic. The SAS Text Miner tool automatically extracts words and phrases and labels them as "terms." This is used to facilitate the IE process of dismissal records. The total annotation time can be reduced from 43 h to 14 h. Were et al. [163] evaluated the Regenstrief EXtracion (REX) tool to extract follow-up provider information from free-text discharge summaries at two hospitals. Comparing three physician reviewers showed that the tool was beneficial at extracting follow-up provider information.

3.3.3.4. Measurement values extraction. Rubin et al. [164] used GATE framework to identify device mentions in portable chest radiography reports and to extract the information, indicating whether the device was removed or remained present. The aim was to study complications, such as infections that could be related to the presence and length of time that devices were present. Hao et al. [165] developed a tool called Valx to extract and normalize numeric laboratory test expressions from clinical texts and evaluated them using clinical trial eligibility criteria text. Garvin et al. [166,167] used regular expressions in UIMA to extract left ventricular ejection fraction value, which is a key clinical component of heart failure quality measure, from echocardiogram reports, and achieved accurate results. Meystre et al. [168] developed a system called CHIEF, which was also based on the UMIA framework, to extract congestive heart failure (CHF) treatment performance measures, such as left ventricular function mentions and values, CHF medications, and documented reasons for a patient not receiving these medications, from clinical notes in a Veterans Health Administration project, and achieved high recall (> 0.990) and good precision (0.960–0.978).

4. Discussion

Observing that clinical IE has been underutilized for clinical and translational research, we have systematically reviewed the literature published between 2009 and 2016 in this study. Our review indicates that clinical IE has been used for a wide range of applications, but there is a considerable gap between clinical studies using EHR data and studies using clinical IE. This study enabled us to gain a more concrete understanding of underlying reasons for this gap.

First, NLP experts trained in the general domain have limited exposure to EHR data as well as limited experience in collaborating with clinicians. Few clinical data sets are available in the public domain due to the Health Insurance Portability and Accountability Act (HIPAA) privacy rule and institutional concerns [169]. Our review showed that the majority of clinical IE publications are from a handful of health care institutions, usually with a strong informatics team (including NLP experts). The development of clinical IE solutions often requires NLP experts to work closely with clinicians who can provide the necessary domain knowledge. However, even with the availability of some EHR data sets to the general community accessible with a data-use agreement (e.g., i2b2 and MIMIC II), they are still underutilized.

Second, as an applied domain, clinical NLP has been dominated by rule-based approaches, which is considerably different from the general

NLP community. We demonstrated that more than 60% of the studies in this review used only rule-based IE systems. However, in the academic NLP research domain (as opposed to the applied or commercial NLP domain), rule-based IE is widely considered obsolete, and statistical machine learning models dominate the research. For example, Chiticariu et al. [170] examined 177 research papers in four best NLP conference proceedings (NLP, EMNLP, ACL, and NAACL) from 2003 through 2012 and found that only 6 papers relied solely on rules. The skew of clinical IE toward rule-based approaches is very similar to the situation of commercial IE products in the general NLP application domain (as opposed to the specialized clinical NLP domain). Chiticariu and colleagues [170] also conducted an industry survey on 54 different IE products in the general domain and found that only one-third of the vendors relied entirely on machine learning. The systems developed by large vendors, such as IBM, SAP, and Microsoft, are completely rule-based. Like these commercial products in the general domain, clinical IE systems greatly value rule-based approaches due to their interpretability to clinicians. In addition, rule-based IE can incorporate domain knowledge from knowledge bases or experts, which is essential for clinical applications. We found that seven machine learning algorithms were applied on four NLP subtasks in 15 studies, and 16 machine learning algorithms were adopted on classification and regression tasks in 64 studies. Most machine learning methods were used for data prediction (e.g., chronic obstructive pulmonary disease prediction [99]), estimation (e.g., lesion malignancy estimation [171]), and association mining (e.g., association between deep vein thrombosis and pulmonary embolism [172]), while only a small group of them were applied directly to NLP tasks (e.g., tumor information extraction [67] and smoking status extraction [55]). Deep learning [173], the prevalent representation-learning method, has not been utilized in the 263 included studies. Nevertheless, there are over 2800 deep-learning publications in the Scopus database in the year 2015 alone. This is again partially due to the limited availability of clinical data sets to researchers. Other reasons include the challenge of interpretability of machine learning methods [174] and the difficulty of correcting specific errors reported by end users (compared to rule-based approaches, which can trivially modify rules correct specific errors). Efforts, such as organizing shared tasks to release clinical text data, are needed to encourage more NLP researchers to contribute to clinical NLP research.

Additionally, the portability and generalizability of clinical IE systems are still limited, partially due to the lack of access to EHRs across institutions to train the systems, and partially due to the lack of standardization. Rule-based IE systems require handcrafted IE rules, while machine learning–based IE systems require a set of manually annotated examples. The resultant IE systems may lack portability, primarily due to the sublanguage difference across heterogeneous sources. One potential solution to this lack of portability is to adopt advanced IE techniques, such as bootstrapping or distant supervision, to build portable and generalizable IE systems [175–179]. These techniques take advantage of a large amount of raw corpus, information redundancy across multiple sources, and existing knowledge bases to automatically or semi-automatically acquire IE knowledge. For example, we can generate raw annotated examples by utilizing an information redundancy across multiple sources and known relationships recorded in knowledge bases. Additionally, most IE tasks are defined without standard information models (a model defining a representation of concepts and the relationships, constraints, rules, and operations to specify data semantics) or value sets (typically used to represent the possible values of a coded data element in an information model), which also limit their portability and generalizability.

We believe the above issues could be alleviated through the training of NLP experts with cross-disciplinary experience, the adoption of standard information models and value sets to improve the interoperability of NLP systems and downstream applications, and collaboration among multiple institutions to advance privacy-preserving data analysis models. Training NLP experts with cross-disciplinary experience is critical to the biomedical informatics community, amplified by the area's interdisciplinary nature. Most NLP courses in informatics training focus on state-of-the-art NLP techniques, while our review demonstrates the widespread use of rule-based NLP systems for real-world practice and clinical research. It may imply an opportunity in informatics training to distinguish academic informatics from applied informatics. Even machine learning-based NLP systems achieve the state-of-the-art performance, however, it is difficult for clinicians and clinical researchers to participate in the system development process.

Standardizing semantics involves two components: (1) information models and (2) value sets. Information models generally specify data semantics and define the representation of entities or concepts, relationships, constraints, rules, and operations, while value sets specify permissible values. The adoption of standards will improve the interoperability of NLP systems and, therefore, facilitate the use of NLP for EHR-based studies. A potential solution is to leverage an international consensus information model, such as the Clinical Information Modeling Initiative (CIMI), and use the compositional grammar for SNOMED-CT concepts in Health Level Seven International (HL7) as standard representations. There are a few existing efforts focusing on sharing clinical data of a group of patients. For example, the clinical e-science framework (CLEF) [180], a UK MRC–sponsored project, aims to establish policies and infrastructure for clinical data sharing of cancer patients to enable the next generation of integrated clinical and bioscience research. However, no prior effort exists for privacy-preserving computing (PPC) on NLP artifacts with distributional information [181,182]. PPC strategies could combine different forms provided by different data resources within the topic of privacy restrictions. A primary issue of leveraging this technique is building a PPC infrastructure. Advanced PPC infrastructure, such as integrating Data for Analysis, Anonymization, and SHaring (iDASH) [183], may be a viable option. Through existing collaborating efforts or building and leveraging this privacy-preserving computing infrastructure, it will become more prevalent to use EHR data for structuring of clinical narratives and supporting the extraction of clinical information for downstream applications.

This review has examined the last 8 years of clinical information extraction applications literature. There are a few limitations in this review. First, this study may have missed relevant articles published after September 7, 2016. Second, the review is limited to articles written in the English language. Articles written in other languages would also provide valuable information. Third, the search strings and databases selected in this review might not be sufficient and might have introduced bias into the review. Fourth, the articles utilizing clinical narratives from non-EHR systems, such as clinical trials [184], are not considered in this review. Finally, the 27 articles about releasing new IE tools and 125 methodology articles are not included in this literature review and will be the focus of future work.

Funding

This work was made possible by NIGMS R01GM102282, NCATS U01TR002062, NLM R01LM11934, and NIBIB R01EB19403.

Competing Interests

None.

Contributors

Y.W.: conceptualized, designed, and wrote the study; designed the analysis of clinical workflow optimization. L.W.: analyzed the data; designed disease areas; edited the manuscript. M.R.M.: analyzed the data; designed drug-related studies; edited the manuscript. S.M.: analyzed the data; designed data sources; edited the manuscript. F.S.: analyzed the data; designed machine learning methods; edited the manuscript. N.A.: analyzed the data; designed clinical IE tools; edited the manuscript. S.L.: analyzed the data; designed publication venues; edited the manuscript. Y.Z.: analyzed the data. S.M.: analyzed the data. S.S.: edited the manuscript. H.L: conceptualized, designed, and edited the manuscript.

Conflict statement

We have nothing to disclose.

Appendix A. Search strategy

A.1. Ovid

Database(s): Embase 1988 to 2016 Week 36, Ovid MEDLINE(R) In-Process & Other Non-Indexed Citations and Ovid MEDLINE(R) 1946 to Present Search Strategy:

No.	Searches	Results
1	(clinic or clinical or "electronic health record" or "electronic health records").mp.	10,297,015
2	("coreference resolution" or "co-reference resolution" or "information extraction" or "named entity extraction" or "named entity recognition" or "natural language processing" or "relation extraction" or "text mining").mp.	10,981
3	"information retrieval".mp.	29,773
4	(1 and 2) not 3	3245
5	limit 4 to English language	3204
6	limit 5 to yr = "2009 -Current"	2480
7	limit 6 to (editorial or erratum or letter or note or comment) [Limit not valid in Embase, Ovid MEDLINE(R), Ovid MEDLINE(R) In-Process; records were retained]	36
8	6 not 7	2444
9	remove duplicates from 8	1651

A.2. Scopus

1 TITLE-ABS-KEY(clinic OR clinical OR "electronic health record" OR "electronic health records")
2 TITLE-ABS-KEY("coreference resolution" OR "co-reference resolution" OR "information extraction" OR "named entity extraction" OR "named entity recognition" OR "natural language processing" OR "relation extraction" OR "text mining")
3 TITLE-ABS-KEY("information retrieval")
4 PUBYEAR AFT 2008 AND LANGUAGE(english)
5 (1 and 2 and 4) and not 3
6 DOCTYPE(le) OR DOCTYPE(ed) OR DOCTYPE(bk) OR DOCTYPE(er) OR DOCTYPE(no) OR DOCTYPE(sh)
7 5 and not 6
8 PMID(0*) OR PMID(1*) OR PMID(2*) OR PMID(3*) OR PMID(4*) OR PMID(5*) OR PMID(6*) OR PMID(7*) OR PMID(8*) OR PMID(9*)
9 7 and not 8

A.3. Web of science

1 TOPIC: (clinic OR clinical OR "electronic health record" OR "electronic health records") AND TOPIC: ("coreference resolution" OR "co-reference resolution" OR "information extraction" OR "named entity extraction" OR "named entity recognition" OR "natural language processing" OR "relation extraction" OR "text mining") AND LANGUAGE: (English) AND DOCUMENT TYPES: (Article OR Abstract of Published Item OR Book OR Book Chapter OR Meeting Abstract OR Proceedings Paper OR Review) Indexes = SCI-EXPANDED Timespan = 2009–2016

2	TS = ("information retrieval")
3	1 NOT 2
4	PMID = (0* or 1* or 2* or 3* or 4* or 5* or 6* or 7* or 8* or 9*)
5	3 NOT 4

A.4. ACM Digital Library

+clinic +"information extraction" −"information retrieval"
+clinical +"information extraction" −"information retrieval"
+"electronic health record" +"information extraction" −"information retrieval"
+"electronic health records" +"information extraction" −"information retrieval"
+clinic +"coreference resolution" −"information retrieval"
+clinical +" coreference resolution" −"information retrieval"

+“electronic health record” +“ coreference resolution” −“information retrieval”
+“electronic health records” +“ coreference resolution” −“information retrieval”
+clinic +“co−reference resolution” −“information retrieval”
+clinical +“co−reference resolution” −“information retrieval”
+“electronic health record” +“co−reference resolution” −“information retrieval”
+“electronic health records” +“co−reference resolution” −“information retrieval”
+clinic +“named entity extraction” −“information retrieval”
+clinical +“named entity extraction” −“information retrieval”
+“electronic health record” +“named entity extraction” −“information retrieval”
+“electronic health records” +“named entity extraction” −“information retrieval”
+clinic +“named entity recognition” −“information retrieval”
+clinical +“named entity recognition” −“information retrieval”
+“electronic health record” +“named entity recognition” −“information retrieval”
+“electronic health records” +“named entity recognition” −“information retrieval”
+clinic +“natural language processing” −“information retrieval”
+clinical +“natural language processing” −“information retrieval ”
+“electronic health record” +“natural language processing” −“information retrieval”
+“electronic health records” +“natural language processing” −“information retrieval”
+clinic +“relation extraction” −“information retrieval”
+clinical +“relation extraction” −“information retrieval ”
+“electronic health record” +“relation extraction” −“information retrieval”
+“electronic health records” +“relation extraction” −“information retrieval”
+clinic +“text mining” −“information retrieval”
+clinical +“text mining” −“information retrieval ”
+“electronic health record” +“text mining” −“information retrieval”
+“electronic health records” +“text mining” −“information retrieval”
All limited from January 1, 2009 to September 6, 2016.

Appendix B. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.jbi.2017.11.011.

References

[1] D. Blumenthal, Launching hitech, N. Engl. J. Med. 362 (2010) 382–385.
[2] M.A. Ellsworth, M. Dziadzko, J.C. O'Horo, A.M. Farrell, J. Zhang, V. Herasevich, An appraisal of published usability evaluations of electronic health records via systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 218–226.
[3] B.A. Goldstein, A.M. Navar, M.J. Pencina, J. Ioannidis, Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review, J. Am. Med. Inform. Assoc. 24 (2017) 198–208.
[4] K. Jensen, C. Soguero-Ruiz, K.O. Mikalsen, R.-O. Lindsetmo, I. Kouskoumvekaki, M. Girolami, et al., Analysis of free text in electronic health records for identification of cancer patient trajectories, Sci. Rep. 7 (2017).
[5] Sarawagi S. Information extraction. Foundations and Trends® in Databases. 2008;1:261–377.
[6] S.G. Small, L. Medsker, Review of information extraction technologies and applications, Neural Comput. Appl.. 25 (2014) 533–548.
[7] J. Cowie, W. Lehnert, Information extraction, Commun. ACM 39 (1996) 80–91.
[8] D. Nadeau, S. Sekine, A survey of named entity recognition and classification, Lingvisticae Investigationes. 30 (2007) 3–26.
[9] H. Lee, Y. Peirsman, A. Chang, N. Chambers, M. Surdeanu, D. Jurafsky, Stanford's multi-pass sieve coreference resolution system at the CoNLL-2011 shared task, in: Proceedings of the Fifteenth Conference on Computational Natural Language Learning: Shared Task: Association for Computational Linguistics, 2011, pp. 28–34.
[10] N. Bach, S. Badaskar, A review of relation extraction, Lit. Rev. Lang. Stat. II (2007).
[11] S.M. Meystre, G.K. Savova, K.C. Kipper-Schuler, J.F. Hurdle, Extracting information from textual documents in the electronic health record: a review of recent research, Yearb Med. Inform. 35 (2008) 44.
[12] R.W.V. Flynn, T.M. Macdonald, N. Schembri, G.D. Murray, A.S.F. Doney, Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes, Pharmacoepidemiol. Drug Saf. 19 (2010) 843–847.
[13] H. Yang, I. Spasic, J.A. Keane, G. Nenadic, A text mining approach to the prediction of disease status from clinical discharge summaries, J. Am. Med. Inform. Assoc. 16 (2009) 596–600.
[14] R. Kung, A. Ma, J.B. Dever, J. Vadivelu, E. Cherk, J.D. Koola, et al., A natural language processing alogrithm for identification of patients with cirrhosis from electronic medical records, Gastroenterology 1 (2015) S1071–S1072.
[15] L.L. Popejoy, M.A. Khalilia, M. Popescu, C. Galambos, V. Lyons, M. Rantz, et al., Quantifying care coordination using natural language processing and domain-specific ontology, J. Am. Med. Inform. Assoc. 22 (2015) e93–e103.
[16] C. Di Marco, P. Bray, H.D. Covvey, D.D. Cowan, V. Di Ciccio, E. Hovy, et al., Authoring and generation of individualized patient education materials, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2006, p. 195.
[17] C. Friedman, P.O. Alderson, J.H. Austin, J.J. Cimino, S.B. Johnson, A general natural-language text processor for clinical radiology, J. Am. Med. Inform. Assoc. 1 (1994) 161–174.
[18] A.R. Aronson, F.-M. Lang, An overview of MetaMap: historical perspective and recent advances, J. Am. Med. Inform. Assoc. 17 (2010) 229–236.
[19] J.C. Denny, P.R. Irani, F.H. Wehbe, J.D. Smithers, A. Spickard III, The KnowledgeMap Project: Development of a Concept-based Medical School Curriculum Database, Citeseer, AMIA, 2003.
[20] G.K. Savova, J.J. Masanz, P.V. Ogren, J. Zheng, S. Sohn, K.C. Kipper-Schuler, et al., Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications, J. Am. Med. Inform. Assoc. 17 (2010) 507–513.
[21] S. Goryachev, M. Sordo, Q.T. Zeng, A suite of natural language processing tools developed for the I2B2 project, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2006, p. 931.
[22] H. Liu, S.J. Bielinski, S. Sohn, S. Murphy, K.B. Wagholikar, S.R. Jonnalagadda, et al., An information extraction framework for cohort identification using electronic health records, AMIA Summits Transl. Sci. Proc. 2013 (2013) 149–153.
[23] J.C. Denny, M.D. Ritchie, M.A. Basford, J.M. Pulley, L. Bastarache, K. Brown-Gentry, et al., PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations, Bioinformatics 26 (2010) 1205–1210.
[24] H. Xu, M. Jiang, M. Oetjens, E.A. Bowton, A.H. Ramirez, J.M. Jeff, et al., Facilitating pharmacogenetic studies using electronic health records and natural-language processing: a case study of warfarin, J. Am. Med. Inform. Assoc. 18 (2011) 387–391.
[25] L. Deleger, H. Brodzinski, H. Zhai, Q. Li, T. Lingren, E.S. Kirkendall, et al., Developing and evaluating an automated appendicitis risk stratification algorithm for pediatric patients in the emergency department, J. Am. Med. Inform. Assoc. 20 (2013) e212–e220.
[26] K. Haerian, D. Varn, S. Vaidya, L. Ena, H.S. Chase, C. Friedman, Detection of pharmacovigilance-related adverse events using electronic health records and automated methods, Clin. Pharmacol. Ther. 92 (2012) 228–234.
[27] J. St-Maurice, M.H. Kuo, Analyzing primary care data to characterize inappropriate emergency room use, Stud. Health Technol. Inform. 180 (2012) 990–994.
[28] V. Kumar, K. Liao, S.C. Cheng, S. Yu, U. Kartoun, A. Brettman, et al., Natural language processing improves phenotypic accuracy in an electronic medical record cohort of type 2 diabetes and cardiovascular disease, J. Am. Coll. Cardiol. 1

(2014) A1359.
[29] R. Patel, T. Lloyd, R. Jackson, M. Ball, H. Shetty, M. Broadbent, et al., Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes, BMJ Open. 5 (2015) e007504.
[30] I.J. Kullo, J. Fan, J. Pathak, G.K. Savova, Z. Ali, C.G. Chute, Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease, J. Am. Med. Inform. Assoc. 17 (2010) 568–574.
[31] K. Davis, C. Staes, J. Duncan, S. Igo, J.C. Facelli, Identification of pneumonia and influenza deaths using the Death Certificate Pipeline, BMC Med. Inf. Decis. Mak. 12 (2012) 37.
[32] C.-I. Wi, S. Sohn, M.C. Rolfes, A. Seabright, E. Ryu, G. Voge, et al., Application of a natural language processing algorithm to asthma ascertainment: an automated chart review, Am. J. Resp. Crit. Care Med. (2017).
[33] N. Afzal, S. Sohn, S. Abram, C.G. Scott, R. Chaudhry, H. Liu, et al., Mining peripheral arterial disease cases from narrative clinical notes using natural language processing, J. Vasc. Surg. 65 (2017) 1753–1761.
[34] S. Sohn, Z. Ye, H. Liu, C.G. Chute, I.J. Kullo, Identifying abdominal aortic aneurysm cases and controls using natural language processing of radiology reports, AMIA Summits Transl. Sci. Proc. 2013 (2013) 249.
[35] P. Spyns, Natural language processing, MethodsInf Med. 35 (1996) 285–301.
[36] Yim W-w, Yetisgen M, Harris WP, Kwan SW. Natural language processing in oncology: a review, JAMA Oncol. 2 (2016) 797–804.
[37] E. Pons, L.M. Braun, M.M. Hunink, J.A. Kors, Natural language processing in radiology: a systematic review, Radiology 279 (2016) 329–343.
[38] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, P. Group, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, PLoS Med. 6 (2009) e1000097.
[39] M. Saeed, M. Villarroel, A.T. Reisner, G. Clifford, L.-W. Lehman, G. Moody, et al., Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II): a public-access intensive care unit database, Crit. Care Med. 39 (2011) 952.
[40] K.R. Gøeg, P.B. Elberg, Global applicability of a local physical examination template. Scandinavian Conference on Health Informatics 2012; October 2–3; Linköping; Sverige: Linköping University Electronic Press, 2012, pp. 1–7.
[41] W.F. Styler IV, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P.C. de Groen, et al., Temporal annotation in the clinical domain, Trans. Assoc. Comput. Linguist. 2 (2014) 143–154.
[42] D. Ferrucci, A. Lally, UIMA: an architectural approach to unstructured information processing in the corporate research environment, Nat. Lang. Eng. 10 (2004) 327–348.
[43] J. Baldridge, The opennlp project, 2005. URL: < http://opennlpapacheorg/indexhtml > .
[44] D.S. Carrell, S. Halgrim, D.-T. Tran, D.S.M. Buist, J. Chubak, W.W. Chapman, et al., Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence, Am. J. Epidemiol. 179 (2014) 749–758.
[45] W.-Q. Wei, C. Tao, G. Jiang, C.G. Chute, A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes, AMIA Annu. Symp. Proc. 2010 (2010) 857–861.
[46] C. Lin, E.W. Karlson, D. Dligach, M.P. Ramirez, T.A. Miller, H. Mo, 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.
[47] H. Hamid, S. Fodeh, G.A. Lizama, R. Czlapinski, M.J. Pugh, W. LaFrance, et al., Validating a natural language processing tool to exclude psychogenic non-epileptic seizures in electronic medical record based epilepsy research, Epilepsy Curr. 14 (2014) 279.
[48] Z. Xia, E. Secor, L.B. Chibnik, R.M. Bove, S. Cheng, T. Chitnis, et al., Modeling disease severity in multiple sclerosis using electronic health records, PLoS ONE 8 (2013) e78927.
[49] T. Lingren, V. Thaker, C. Brady, B. Namjou, S. Kennebeck, J. Bickel, et al., Developing an algorithm to detect early childhood obesity in two tertiary pediatric medical centers, Appl. Clin. Inform. 7 (2016) 693–706.
[50] S. Mehrabi, C.M. Schmidt, J.A. Waters, C. Beesley, A. Krishnan, J. Kesterson, et al., An efficient pancreatic cyst identification methodology using natural language processing, Stud. Health Technol. Inform. 192 (2013) 822–826.
[51] G.K. Savova, J. Fan, Z. Ye, S.P. Murphy, J. Zheng, C.G. Chute, et al., Discovering peripheral arterial disease cases from radiology notes using natural language processing, AMIA Annu Symp Proc. 2010 (2010) 722–726.
[52] L. Cui, A. Bozorgi, S.D. Lhatoo, G.-Q. Zhang, S.S. Sahoo, EpiDEA: extracting structured epilepsy and seizure information from patient discharge summaries for cohort identification, AMIA Annu. Symp. Proc. 2012 (2012) 1191–1200.
[53] S. Hassanpour, C.P. Langlotz, Information extraction from multi-institutional radiology reports, Artif. Intell. Med. 66 (2016) 29–39.
[54] J. Pathak, D.K. Hall-Flavin, J.M. Biernacka, G.D. Jenkins, K.T. Bruce, S.P. Murphy, et al., Using electronic health records driven phenotyping for major depressive disorder, Biol. Psychiat. 1 (2014) 343S.
[55] M. Liu, A. Shah, M. Jiang, N.B. Peterson, Q. Dai, M.C. Aldrich, et al., A study of transportability of an existing smoking status detection module across institutions, AMIA Annu. Symp. Proc. 2012 (2012) 577–586.
[56] R. Khor, W. Yip, M. Bressel, W. Rose, G. Duchesne, F. Foroudi, Automated smoking status extraction from free text: adapting a system for use in the Australian context, J. Med. Imag. Radiat. Oncol. 57 (2013) 148.
[57] S. Sohn, G.K. Savova, Mayo clinic smoking status classification system: extensions and improvements, AMIA Annu. Symp. Proc. 2009 (2009) 619–623.
[58] R. Khor, W.-K. Yip, M. Bressel, W. Rose, G. Duchesne, F. Foroudi, Practical implementation of an existing smoking detection pipeline and reduced support vector machine training corpus requirements, J. Am. Med. Inform. Assoc. 21 (2014) 27–30.
[59] S. Sohn, Kocher J-PA, Chute CG, Savova GK. Drug side effect extraction from clinical narratives of psychiatry and psychology patients, J. Am. Med. Inform. Assoc. 18 (Suppl 1) (2011) i144–i149.
[60] Q. Li, S.A. Spooner, M. Kaiser, N. Lingren, J. Robbins, T. Lingren, et al., An end-to-end hybrid algorithm for automated medication discrepancy detection, BMC Med. Inf. Decis. Mak. 15 (2015) 37.
[61] C. Lin, D. Dligach, T.A. Miller, S. Bethard, G.K. Savova, Multilayered temporal modeling for the clinical domain, J. Am. Med. Inform. Assoc. 23 (2016) 387–395.
[62] A. Khalifa, S. Meystre, Adapting existing natural language processing resources for cardiovascular risk factors identification in clinical notes, J. Biomed. Inform. 58 (Suppl) (2015) S128–S132.
[63] A.R. Aronson, J.G. Mork, A. Névéol, S.E. Shooshan, D. Demner-Fushman, Methodology for creating UMLS content views appropriate for biomedical natural language processing, AMIA (2008).
[64] M. Yetisgen-Yildiz, C.A. Bejan, L. Vanderwende, F. Xia, H.L. Evans, M.M. Wurfel, Automated tools for phenotype extraction from medical records, AMIA Summits Transl. Sci. Proc. 2013 (2013) 283.
[65] C.A. Bejan, F. Xia, L. Vanderwende, M.M. Wurfel, M. Yetisgen-Yildiz, Pneumonia identification using statistical feature selection, J. Am. Med. Inform. Assoc. 19 (2012) 817–823.
[66] A.V. Gundlapalli, A. Redd, M. Carter, G. Divita, S. Shen, M. Palmer, et al., Validating a strategy for psychosocial phenotyping using a large corpus of clinical text, J. Am. Med. Inform. Assoc. 20 (2013) e355–e364.
[67] W.-W. Yim, T. Denman, S.W. Kwan, M. Yetisgen, Tumor information extraction in radiology reports for hepatocellular carcinoma patients, AMIA Summits Transl Sci Proc. 2016 (2016) 455–464.
[68] M. Sevenster, J. Buurman, P. Liu, J.F. Peters, P.J. Chang, Natural language processing techniques for extracting and categorizing finding measurements in narrative radiology reports, Appl. Clin. Inform. 6 (2015) 600–1110.
[69] D. Martinez, M.R. Ananda-Rajah, H. Suominen, M.A. Slavin, K.A. Thursky, L. Cavedon, Automatic detection of patients with invasive fungal disease from free-text computed tomography (CT) scans, J. Biomed. Inform. 53 (2015) 251–260.
[70] J. St-Maurice, M.H. Kuo, P. Gooch, A proof of concept for assessing emergency room use with primary care data and natural language processing, Methods Inf. Med. 52 (2013) 33–42.
[71] R. Khare, J. Li, Z. Lu, LabeledIn: cataloging labeled indications for human drugs, J. Biomed. Inform. 52 (2014) 448–456.
[72] C. Thorne, E. Cardillo, C. Eccher, M. Montali, D. Calvanese, Process fragment recognition in clinical documents, in: 13th International Conference of the Italian Association for Artificial Intelligence, AI*IA 2013, Turin 2013, pp. 227–238.
[73] H. Zhu, Y. Ni, P. Cai, Z. Qiu, F. Cao, Automatic extracting of patient-related attributes: disease, age, gender and race, Stud. Health Technol. Inform. 180 (2012) 589–593.
[74] X. Wang, G. Hripcsak, C. Friedman, Characterizing environmental and phenotypic associations using information theory and electronic health records, BMC Bioinf. 10 (Suppl 9) (2009) S13.
[75] X. Wang, G. Hripcsak, M. Markatou, C. Friedman, Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study, J. Am. Med. Inform. Assoc. 16 (2009) 328–337.
[76] H. Salmasian, D.E. Freedberg, J.A. Abrams, C. Friedman, An automated tool for detecting medication overuse based on the electronic health records, Pharmacoepidemiol. Drug Saf. 22 (2013) 183–189.
[77] A.W.C. Kamauu, G. Petratos, A. Amey, P. Bechtel, D. Dine, Extracting meaningful, searchable and discrete data from unstructured medical text, Pharmacoepidemiol. Drug Saf. 19 (2010) S75.
[78] H. Cunningham, D. Maynard, K. Bontcheva, V. Tablan, GATE: an architecture for development of robust HLT applications, in: Proceedings of the 40th annual meeting on association for computational linguistics: Association for Computational Linguistics, 2002, pp. 168–175.
[79] H. Cunningham, V. Tablan, A. Roberts, K. Bontcheva, Getting more out of biomedical documents with GATE's full lifecycle open source text analytics, PLoS Comput. Biol. 9 (2013) e1002854.
[80] S. Bird, NLTK: the natural language toolkit, Proceedings of the COLING/ACL on Interactive presentation sessions: Association for Computational Linguistics, 2006, pp. 69–72.
[81] T. Morton, J. Kottmann, J. Baldridge, G. Bierner, Opennlp: A java-based nlp toolkit, 2005.
[82] A.K. McCallum, Mallet: A machine learning for language toolkit, 2002.
[83] G. Holmes, A. Donkin, I.H. Witten, Weka: A machine learning workbench. Intelligent Information Systems, in: 1994 Proceedings of the 1994 Second Australian and New Zealand Conference on: IEEE, 1994, pp. 357–361.
[84] M.A. Musen, The Protégé project: a look back and a look forward, AI Matters 1 (2015) 4–12.
[85] M. Ball, R. Patel, R.D. Hayes, R.J. Dobson, R. Stewart, TextHunter–A User Friendly Tool for Extracting Generic Concepts from Free Text in Clinical Research, in: AMIA Annual Symposium Proceedings: American Medical Informatics Association, 2014, p. 729.
[86] J.D. Patrick, D.H. Nguyen, Y. Wang, M. Li, A knowledge discovery and reuse pipeline for information extraction in clinical notes, J. Am. Med. Inform. Assoc. 18 (2011) 574–579.
[87] I. Spasić, B. Zhao, C.B. Jones, K. Button, KneeTex: an ontology–driven system for information extraction from MRI reports, J. Biomed. Semantics 6 (2015) 34.

[88] S.M. Meystre, J. Thibault, S. Shen, J.F. Hurdle, B.R. South, Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents, J. Am. Med. Inform. Assoc. 17 (2010) 559–562.
[89] E. Tseytlin, K. Mitchell, E. Legowski, J. Corrigan, G. Chavan, R.S. Jacobson, NOBLE–Flexible concept recognition for large-scale biomedical natural language processing, BMC Bioinf. 17 (2016) 32.
[90] S.J. Hebbring, The challenges, advantages and future of phenome-wide association studies, Immunology 141 (2014) 157–165.
[91] V. Law, C. Knox, Y. Djoumbou, T. Jewison, A.C. Guo, Y. Liu, et al., DrugBank 4.0: shedding new light on drug metabolism, Nucl. Acids Res. 42 (2014) D1091–D1097.
[92] P.L. Elkin, D.A. Froehling, D.L. Wahner-Roedler, S.H. Brown, K.R. Bailey, Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes, Ann. Intern. Med. 156 (2012) 11–18.
[93] S. Horng, D.A. Sontag, N.I. Shapiro, L.A. Nathanson, Machine learning algorithms can identify patients who will benefit from targeted sepsis decision support, Ann. Emerg. Med. 1 (2012) S121.
[94] K. Roberts, B. Rink, S.M. Harabagiu, R.H. Scheuermann, S. Toomay, T. Browning, et al., A machine learning approach for identifying anatomical locations of actionable findings in radiology reports, AMIA Annu. Symp. Proc. 2012 (2012) 779–788.
[95] C. Zheng, N. Rashid, T.C. Cheetham, Y.L. Wu, G.D. Levy, Using natural language processing and machine learning to identify gout flares from electronic clinical notes, Arthritis Rheum. 65 (2013) S856–S857.
[96] P. Kluegl, M. Toepfer, P.-D. Beck, G. Fette, F. Puppe, UIMA Ruta: Rapid development of rule-based information extraction applications, Nat. Lang. Eng. 22 (2016) 1–40.
[97] N. Barrett, J.H. Weber-Jahnke, V. Thai, 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.
[98] A. Sarker, G. Gonzalez, Portable automatic text classification for adverse drug reaction detection via multi-corpus training, J. Biomed. Inform. 53 (2015) 196–207.
[99] B.E. Himes, Y. Dai, I.S. Kohane, S.T. Weiss, M.F. Ramoni, Prediction of chronic obstructive pulmonary disease (COPD) in asthma patients using electronic medical records, J. Am. Med. Inform. Assoc. 16 (2009) 371–379.
[100] Y. Chen, J. Wrenn, H. Xu, A. Spickard 3rd, R. Habermann, J. Powers, et al., Automated assessment of medical students' clinical exposures according to AAMC geriatric competencies, AMIA Annu. Symp. Proc. 2014 (2014) 375–384.
[101] C.M. Rochefort, D.L. Buckeridge, A.J. Forster, Accuracy of using automated methods for detecting adverse events from electronic health record data: a research protocol, Implement Sci. 10 (2015) 5.
[102] K. Yadav, E. Sarioglu, M. Smith, H.-A. Choi, Automated outcome classification of emergency department computed tomography imaging reports, Acad. Emerg. Med. 20 (2013) 848–854.
[103] Ö. Uzuner, Y. Luo, P. Szolovits, Evaluating the state-of-the-art in automatic de-identification, J. Am. Med. Inform. Assoc. 14 (2007) 550–563.
[104] Ö. Uzuner, I. Goldstein, Y. Luo, I. Kohane, Identifying patient smoking status from medical discharge records, J. Am. Med. Inform. Assoc. 15 (2008) 14–24.
[105] O. Uzuner, Recognizing obesity and comorbidities in sparse data, J. Am. Med. Inform. Assoc. 16 (2009) 561–570.
[106] O. Uzuner, I. Solti, E. Cadag, Extracting medication information from clinical text, J. Am. Med. Inform. Assoc. 17 (2010) 514–518.
[107] O. Uzuner, B.R. South, S. Shen, S.L. DuVall, 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text, J. Am. Med. Inform. Assoc. 18 (2011) 552–556.
[108] O. Uzuner, A. Bodnari, S. Shen, T. Forbush, J. Pestian, B.R. South, Evaluating the state of the art in coreference resolution for electronic medical records, J. Am. Med. Inform. Assoc. 19 (2012) 786–791.
[109] W. Sun, A. Rumshisky, O. Uzuner, Evaluating temporal relations in clinical text: 2012 i2b2 Challenge, J. Am. Med. Inform. Assoc. 20 (2013) 806–813.
[110] A. Stubbs, C. Kotfila, H. Xu, O. Uzuner, Identifying risk factors for heart disease over time: Overview of 2014 i2b2/UTHealth shared task Track 2, J. Biomed. Inform. 58 (Suppl) (2015) S67–S77.
[111] A. Stubbs, C. Kotfila, Ö. Uzuner, Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 i2b2/UTHealth shared task Track 1, J. Biomed. Inform. 58 (2015) S11–S19.
[112] S. Pradhan, N. Elhadad, B.R. South, D. Martinez, L.M. Christensen, A. Vogel, et al., Task 1: ShARe/CLEF eHealth Evaluation Lab 2013. CLEF (Working Notes), 2013.
[113] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D.L. Mowery, et al., Overview of the share/clef ehealth evaluation lab 2014. International Conference of the Cross-Language Evaluation Forum for European Languages: Springer, 2014, pp. 172–191.
[114] L. Goeuriot, L. Kelly, H. Suominen, L. Hanlen, A. Névéol, C. Grouin, et al., Overview of the CLEF eHealth evaluation lab 2015, in: International Conference of the Cross-Language Evaluation Forum for European Languages: Springer, 2015, pp. 429–443.
[115] L. Kelly, L. Goeuriot, H. Suominen, A. Névéol, J. Palotti, G. Zuccon, Overview of the CLEF eHealth evaluation lab 2016, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2016, pp. 255–266.
[116] I. Segura-Bedmar, P. Martínez, M.H. Zazo, Semeval-2013 task 9: Extraction of drug-drug interactions from biomedical texts (ddiextraction 2013). Second Joint Conference on Lexical and Computational Semantics (* SEM), in: Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013), 2013, pp. 341–350.
[117] S. Pradhan, N. Elhadad, W.W. Chapman, S. Manandhar, G. Savova, SemEval-2014 Task 7: Analysis of Clinical Text. SemEval@ COLING2014, pp. 54–62.
[118] N. Elhadad, S. Pradhan, S.L. Gorman, S. Manandhar, W.W. Chapman, G.K. Savova, SemEval-2015 Task 14: Analysis of Clinical Text. SemEval@ NAACL-HLT2015, pp. 303–310.
[119] S. Bethard, G. Savova, W.-T. Chen, L. Derczynski, J. Pustejovsky, M. Verhagen, Semeval-2016 task 12: Clinical tempeval, Proc. SemEval. (2016) 1052–1062.
[120] Y. Sada, J. Hou, P. Richardson, H. El-Serag, J. Davila, Validation of case finding algorithms for hepatocellular cancer from administrative data and electronic health records using natural language processing, Med. Care. 54 (2016) e9–e14.
[121] H. Xu, Z. Fu, A. Shah, Y. Chen, N.B. Peterson, Q. Chen, et al., Extracting and integrating data from entire electronic health records for detecting colorectal cancer cases, AMIA Annu. Symp. Proc. 2011 (2011) 1564–1572.
[122] Y. Kim, J. Garvin, J. Heavirland, S.M. Meystre, Improving heart failure information extraction by domain adaptation, Stud. Health Technol. Inform. 192 (2013) 185–189.
[123] A.M. Roch, S. Mehrabi, A. Krishnan, H.E. Schmidt, J. Kesterson, C. Beesley, et al., Automated pancreatic cyst screening using natural language processing: a new tool in the early detection of pancreatic cancer, Hpb 17 (2015) 447–453.
[124] J.F. Ludvigsson, J. Pathak, S. Murphy, M. Durski, P.S. Kirsch, C.G. Chute, et al., Use of computerized algorithm to identify individuals in need of testing for celiac disease, J. Am. Med. Inform. Assoc. 20 (2013) e306–e310.
[125] G. Sances, C. Larizza, M. Gabetta, M. Bucalo, E. Guaschino, G. Milani, et al., Application of bioinformatics in headache: The I2B2-pavia project, J. Headache Pain 11 (2010) S134–S135.
[126] A. Graystone, R. Bhatia, R. Davies, S. McClinton, Validation of the DM reporter text mining application for evaluating the management of high risk populations with diabetes, in: Diabetes Conference: 70th Scientific Sessions of the American Diabetes Association Orlando, FL United States Conference Start, 2010.
[127] K.B. Cohen, B. Glass, H.M. Greiner, K. Holland-Bouley, S. Standridge, R. Arya, et al., Methodological issues in predicting pediatric epilepsy surgery candidates through natural language processing and machine learning, Biomedical 8 (2016) 11–18.
[128] A.P. Nunes, J. Yang, K. Tunceli, K. Kurtyka, L. Radican, S.S. Engel, et al., Interim results on the relationship between mild-moderate and severe hypoglycaemia and cardiovascular disease in a cohort of sulfonylurea users, Diabetologia 1 (2015) S62.
[129] S. Lyalina, B. Percha, P. LePendu, S.V. Iyer, R.B. Altman, N.H. Shah, Identifying phenotypic signatures of neuropsychiatric disorders from electronic medical records, J. Am. Med. Inform. Assoc. 20 (2013) e297–e305.
[130] M. Wang, A. Cyhaniuk, D.L. Cooper, N.N. Iyer, Identification of persons with acquired hemophilia in a large electronic health record database, Blood 126 (23) (2015) 3271.
[131] S. Mehrabi, A. Krishnan, A.M. Roch, H. Schmidt, D. Li, J. Kesterson, et al., Identification of Patients with Family History of Pancreatic Cancer-Investigation of an NLP System Portability, Stud Health Technol Inform. 216 (2015) 604–608.
[132] C. Hirst, J. Hill, S. Khosla, K. Schweikert, C. Senerchia, K. Kitzmann, et al., The application of natural language processing (NLP) technology to enrich electronic medical records (EMRS) for outcomes research in oncology, Value Health 17 (3) (2014) A6.
[133] A.N. Nguyen, J. Moore, J. O'Dwyer, S. Philpot, Assessing the utility of automatic cancer registry notifications data extraction from free-text pathology reports, AMIA Annu. Symp. Proc. 2015 (2015) 953–962.
[134] H. Farrugia, G. Marr, G. Giles, Implementing a natural language processing solution to capture cancer stage and recurrence, J. Med. Imag. Radiat. Oncol. 56 (2012) 5.
[135] J.L. Warner, M.A. Levy, M.N. Neuss, J.L. Warner, M.A. Levy, M.N. Neuss, ReCAP: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data, J. Oncol. Pract. 12 (157–8) (2016) e69–e77.
[136] Z. Tian, S. Sun, T. Eguale, C. Rochefort, Automated extraction of VTE events from narrative radiology reports in electronic health records: a validation study, Pharmacoepidemiol. Drug Saf. 24 (2015) 166.
[137] E.R. McPeek Hinz, L. Bastarache, J.C. Denny, A natural language processing algorithm to define a venous thromboembolism phenotype, AMIA Annu. Symp. Proc. 2013 (2013) 975–983.
[138] J. Duke, M. Chase, N. Poznanski-Ring, J. Martin, R. Fuhr, A. Chatterjee, et al., Natural language processing to improve identification of peripheral arterial disease in electronic health data, J. Am. Coll. Cardiol. 1 (2016) 2280.
[139] N. Afzal, S. Sohn, S. Sohn, H. Liu, I.J. Kullo, A.M. Arruda-Olson, Identifying peripheral arterial disease cases using natural language processing of clinical notes, in: 3rd IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2016: Institute of Electrical and Electronics Engineers Inc., 2016, pp. 126–131.
[140] A. Rao, J.M. Ehrenfeld, R. Peterfreund, M. Zalis, M. Harris, Automated analysis of free text electronic medical records to identify patients with specific medical diagnoses, in: Anesthesia and Analgesia Conference, 2011, p. 112.
[141] H. Xu, S.P. Stenner, S. Doan, K.B. Johnson, L.R. Waitman, J.C. Denny, MedEx: a medication information extraction system for clinical narratives, J. Am. Med. Inform. Assoc. 17 (2010) 19–24.
[142] H. Xu, M.C. Aldrich, Q. Chen, H. Liu, N.B. Peterson, Q. Dai, et al., Validating drug repurposing signals using electronic health records: a case study of metformin associated with reduced cancer mortality, J. Am. Med. Inform. Assoc. 22 (2015) 179–191.
[143] S. Sohn, C. Clark, S.R. Halgrim, S.P. Murphy, S.R. Jonnalagadda, K.B. Wagholikar, et al., Analysis of cross-institutional medication description patterns in clinical

narratives, Biomedical 6 (2013) 7–16.
[144] C. Zheng, N. Rashid, R. Koblick, J. An, 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.
[145] H. Xu, S. Doan, K.A. Birdwell, J.D. Cowan, A.J. Vincz, D.W. Haas, et al., An automated approach to calculating the daily dose of tacrolimus in electronic health records, AMIA Summits Transl. Sci. Proc. 2010 (2010) 71–75.
[146] M. Liu, M. Jiang, V.K. Kawai, C.M. Stein, D.M. Roden, J.C. Denny, et al., Modeling drug exposure data in electronic medical records: an application to warfarin, AMIA Annu. Symp. Proc. 2011 (2011) 815–823.
[147] Q. Feng, M.S. Waitara, L. Jiang, H. Xu, M. Jiang, C.A. McCarty, et al., Dose-response curves extracted from electronic medical records identify sort-1 as a novel genetic predictor of statin potency (ED50), Clin. Pharmacol. Ther. 91 (2012) S48–S49.
[148] C. Rochefort, A. Verma, T. Eguale, D. Buckeridge, Surveillance of adverse events in elderly patients: a study on the accuracy of applying natural language processing techniques to electronic health record data, Eur. Geriat. Med. 6 (2015) S15.
[149] H. Zhang, J. Plutzky, S. Skentzos, F. Morrison, P. Mar, M. Shubina, et al., Epidemiology of adverse reaction to statins in routine care settings, in: Endocrine Reviews Conference: 94th Annual Meeting and Expo of the Endocrine Society, ENDO, 2012, p. 33.
[150] B. Hazlehurst, A. Naleway, J. Mullooly, Detecting possible vaccine adverse events in clinical notes of the electronic medical record, Vaccine 27 (2009) 2077–2083.
[151] B. Baer, M. Nguyen, E.J. Woo, S. Winiecki, J. Scott, D. Martin, et al., Can Natural language processing improve the efficiency of vaccine adverse event report review? MethodsInf. Med. 55 (2016) 144–150.
[152] S. Tamang, M.I. Patel, D.W. Blayney, J. Kuznetsov, S.G. Finlayson, Y. Vetteth, et al., Detecting unplanned care from clinician notes in electronic health records, J Oncol Pract. 11 (2015) e313–e319.
[153] W. Hsu, S.X. Han, C.W. Arnold, A.A. Bui, D.R. Enzmann, A data-driven approach for quality assessment of radiologic interpretations, J. Am. Med. Inform. Assoc. 23 (2016) e152–e156.
[154] R. Smith-Bindman, D.L. Miglioretti, E.B. Larson, Rising use of diagnostic medical imaging in a large integrated health system, Health Aff. 27 (2008) 1491–1502.
[155] I.K. Ip, K.J. Mortele, L.M. Prevedello, R. Khorasani, Repeat abdominal imaging examinations in a tertiary care hospital, Am. J. Med. 125 (2012) 155–161.
[156] A. Mehrotra, H. Harkema, Development and validation of a natural language processing computer program to measure the quality of colonoscopy, J. Gen. Intern. Med. 26 (2011) S339–S340.
[157] G.S. Raju, W.A. Ross, P. Lum, P.M. Lynch, R.S. Slack, E. Miller, et al., Natural language processing (NLP) as an alternative to manual reporting of colonoscopy quality metrics, Gastrointest Endosc. 1 (2014) AB116–AB117.
[158] G.S. Raju, P.J. Lum, R.S. Slack, S. Thirumurthi, P.M. Lynch, E. Miller, et al., Natural language processing as an alternative to manual reporting of colonoscopy quality metrics, Gastrointest Endosc. 82 (2015) 512–519.
[159] A.J. Gawron, J.A. Pacheco, B. Scuba, W. Chapman, T. Kaltenbach, W.K. Thompson, Quality improvement natural language processing colonoscopy evaluation tool (QUINCE): a flexible, portable tool to extract pathology results for colonoscopy quality reporting, Gastroenterology 1 (2016) S637.
[160] A.V. Gundlapalli, M.E. Carter, G. Divita, S. Shen, M. Palmer, B. South, et al., Extracting concepts related to homelessness from the free text of VA electronic medical records, AMIA Annu. Symp. Proc. 2014 (2014) 589–598.
[161] K.L. Ruud, M.G. Johnson, J.T. Liesinger, C.A. Grafft, J.M. Naessens, Automated detection of follow-up appointments using text mining of discharge records, Int. J. Qual. Health Care 22 (2010) 229–235.
[162] M. Abell, SAS Text Miner: CreateSpace Independent Publishing Platform, 2014.
[163] M.C. Were, S. Gorbachev, J. Cadwallader, J. Kesterson, X. Li, J.M. Overhage, et al., Natural language processing to extract follow-up provider information from hospital discharge summaries, AMIA Annu. Symp. Proc. 2010 (2010) 872–876.
[164] D. Rubin, D. Wang, D.A. Chambers, J.G. Chambers, B.R. South, M.K. Goldstein, Natural language processing for lines and devices in portable chest x-rays, AMIA Annu Symp Proc. 2010 (2010) 692–696.
[165] T. Hao, H. Liu, C. Weng, Valx: a system for extracting and structuring numeric lab test comparison statements from text, MethodsInf. Med. 55 (2016) 266–275.
[166] J.H. Garvin, P.L. Elkin, S. Shen, S. Brown, B. Trusko, E. Wang, et al., Automated quality measurement in department of the veterans affairs discharge instructions for patients with congestive heart failure, J. Healthc. Qual. 35 (2013) 16–24.
[167] J.H. Garvin, S.L. DuVall, B.R. South, B.E. Bray, D. Bolton, J. Heavirland, et al., Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure, J. Am. Med. Inform. Assoc. 19 (2012) 859–866.
[168] S.M. Meystre, Y. Kim, G.T. Gobbel, M.E. Matheny, A. Redd, B.E. Bray, et al., Congestive heart failure information extraction framework for automated treatment performance measures assessment, J. Am. Med. Inform. Assoc. 24 (2016) e40–e46.
[169] C. Friedman, T.C. Rindflesch, M. Corn, Natural language processing: state of the art and prospects for significant progress, a workshop sponsored by the National Library of Medicine, J. Biomed. Inform. 46 (2013) 765–773.
[170] L. Chiticariu, Y. Li, F.R. Reiss, Rule-based information extraction is dead! long live rule-based information extraction systems! EMNLP2013, pp. 827–832.
[171] S. Bozkurt, F. Gimenez, E.S. Burnside, K.H. Gulkesen, D.L. Rubin, Using automatically extracted information from mammography reports for decision-support, J. Biomed. Inform. 62 (2016) 224–231.
[172] C.M. Rochefort, A.D. Verma, T. Eguale, T.C. Lee, D.L. Buckeridge, A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data, J. Am. Med. Inform. Assoc. 22 (2015) 155–165.
[173] Y. LeCun, Y. Bengio, G. Hinton, Deep learning, Nature 521 (2015) 436–444.
[174] A. Vellido, J.D. Martín-Guerrero, P.J. Lisboa, Making machine learning models interpretable, ESANN2012, pp. 163–172.
[175] E. Riloff, J. Wiebe, T. Wilson, Learning subjective nouns using extraction pattern bootstrapping, in: Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4: Association for Computational Linguistics, 2003, pp. 25–32.
[176] E. Riloff, R. Jones, Learning dictionaries for information extraction by multi-level bootstrapping, AAAI/IAAI1999, pp. 474–479.
[177] M. Mintz, S. Bills, R. Snow, D. Jurafsky, Distant supervision for relation extraction without labeled data, Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2: Association for Computational Linguistics, 2009, pp. 1003–1011.
[178] S. Takamatsu, I. Sato, H. Nakagawa, Reducing wrong labels in distant supervision for relation extraction, in: Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1: Association for Computational Linguistics, 2012, pp. 721–729.
[179] D.S. Weld, R. Hoffmann, F. Wu, Using wikipedia to bootstrap open information extraction, ACM SIGMOD Record. 37 (2009) 62–68.
[180] D. Kalra, P. Singleton, J. Milan, J. Mackay, D. Detmer, A. Rector, et al., Security and confidentiality approach for the Clinical E-Science Framework (CLEF), Methods Inf. Med. 44 (2004) 193–197.
[181] M.B. Malik, M.A. Ghazi, R. Ali, Privacy preserving data mining techniques: current scenario and future prospects, in: Computer and Communication Technology (ICCCT), 2012 Third International Conference on: IEEE, 2012, pp. 26–32.
[182] J. Gardner, L. Xiong, An integrated framework for de-identifying unstructured medical data, Data Knowl. Eng. 68 (2009) 1441–1451.
[183] L. Ohno-Machado, V. Bafna, A.A. Boxwala, B.E. Chapman, W.W. Chapman, K. Chaudhuri, et al., iDASH: integrating data for analysis, anonymization, and sharing, J. Am. Med. Inform. Assoc. 19 (2012) 196–201.
[184] C. Weng, X. Wu, Z. Luo, M.R. Boland, D. Theodoratos, S.B. Johnson, EliXR: an approach to eligibility criteria extraction and representation, J. Am. Med. Inform. Assoc. 18 (2011) i116–i124.