frontiers | Frontiers in Artificial Intelligence

TYPE Original Research
PUBLISHED 13 January 2023
DOI 10.3389/frai.2022.1051724

OPEN ACCESS

EDITED BY
Roland Roller,
German Research Center for Artificial Intelligence (DFKI), Germany

REVIEWED BY
Ernestina Menasalvas,
Polytechnic University of Madrid, Spain
Shuntaro Yada,
Nara Institute of Science and Technology (NAIST), Japan

*CORRESPONDENCE
Juan Antonio Lossio-Ventura
✉ juan.lossio@nih.gov

SPECIALTY SECTION
This article was submitted to Natural Language Processing, a section of the journal Frontiers in Artificial Intelligence

RECEIVED 23 September 2022
ACCEPTED 15 December 2022
PUBLISHED 13 January 2023

CITATION
Lossio-Ventura JA, Sun R, Boussard S and Hernandez-Boussard T (2023) Clinical concept recognition: Evaluation of existing systems on EHRs. *Front. Artif. Intell.* 5:1051724. doi: 10.3389/frai.2022.1051724

Clinical concept recognition: Evaluation of existing systems on EHRs

Juan Antonio Lossio-Ventura[1,2]*, Ran Sun[1], Sebastien Boussard[3] and Tina Hernandez-Boussard[1,4,5]

[1]Biomedical Informatics Research, Stanford University, Stanford, CA, United States, [2]National Institute of Mental Health, National Institutes of Health, Bethesda, MD, United States, [3]College of Engineering, Boston University, Boston, MA, United States, [4]Department of Biomedical Data Sciences, Stanford University, Stanford, CA, United States, [5]Department of Surgery, Stanford University, Stanford, CA, United States

Objective: The adoption of electronic health records (EHRs) has produced enormous amounts of data, creating research opportunities in clinical data sciences. Several concept recognition systems have been developed to facilitate clinical information extraction from these data. While studies exist that compare the performance of many concept recognition systems, they are typically developed internally and may be biased due to different internal implementations, parameters used, and limited number of systems included in the evaluations. The goal of this research is to evaluate the performance of existing systems to retrieve relevant clinical concepts from EHRs.

Methods: We investigated six concept recognition systems, including CLAMP, cTAKES, MetaMap, NCBO Annotator, QuickUMLS, and ScispaCy. Clinical concepts extracted included procedures, disorders, medications, and anatomical location. The system performance was evaluated on two datasets: the 2010 i2b2 and the MIMIC-III. Additionally, we assessed the performance of these systems in five challenging situations, including negation, severity, abbreviation, ambiguity, and misspelling.

Results: For clinical concept extraction, CLAMP achieved the best performance on exact and inexact matching, with an F-score of 0.70 and 0.94, respectively, on i2b2; and 0.39 and 0.50, respectively, on MIMIC-III. Across the five challenging situations, ScispaCy excelled in extracting abbreviation information (F-score: 0.86) followed by NCBO Annotator (F-score: 0.79). CLAMP outperformed in extracting severity terms (F-score 0.73) followed by NCBO Annotator (F-score: 0.68). CLAMP outperformed other systems in extracting negated concepts (F-score 0.63).

Conclusions: Several concept recognition systems exist to extract clinical information from unstructured data. This study provides an external evaluation by end-users of six commonly used systems across different extraction tasks. Our findings suggest that CLAMP provides the most comprehensive set of

annotations for clinical concept extraction tasks and associated challenges. Comparing standard extraction tasks across systems provides guidance to other clinical researchers when selecting a concept recognition system relevant to their clinical information extraction task.

KEYWORDS

clinical concept recognition, electronic health records, natural language processing, clinical information extraction, UMLS, named-entity recognition

1. Introduction

The ubiquity of electronic health records (EHRs) has created an excessive amount of digital clinical data for research (Evans, 2016). EHRs store structured health information in various formats and unstructured patient data such as progress notes and discharge summaries, account for more than 80% of the data (Murdoch and Detsky, 2013; Assale et al., 2019). These data include critical information about clinical decisions made on patients' diagnosis, prescribed medications, clinical procedures, and its related anatomical locations. Information from these unstructured data is sparse and conversion of these unstructured data to structured data is labor-intensive and expensive (Hersh et al., 2013). Tools have been developed to make use of these data and solve many of the biomedical text mining problems.

Natural language processing (NLP) techniques have been successful in advancing biomedical and clinical research by decreasing the time and effort to obtain critical information from clinical notes (Yim et al., 2016; Wang Y. et al., 2018). Clinical concept recognition, also known as named entity recognition, is a fundamental NLP task that aims to automatically recognize and classify concepts from clinical narratives such as disease diagnosis and medications. Over the past several years, concept recognition for the general domain has attracted considerable attention, and studies applying it to the clinical domain have also emerged (Uzuner et al., 2010, 2011; Pradhan et al., 2014, 2015). Concept recognition systems in clinical settings is crucial because it reduces the manual effort to review patients historical medical record, promotes information exchange across different EHR systems, efficiently summarizes the patient medical history, and help providers as well as patient quickly grasp patient information about their disease conditions, procedures performed, and medications used.

Several concept recognition systems have been proposed to extract clinical information from text to facilitate patient care and clinical research, such as MedLEE (Friedman et al., 1994; Friedman, 2000), MetaMap (Aronson and Lang, 2010), MetaMap Lite (Demner-Fushman et al., 2017), KnowledgeMap (Denny et al., 2003), Apache cTAKES (Savova et al., 2010; Kovačević et al., 2013), HiTEX (Zeng et al., 2006), NCBO Annotator (Jonquet et al., 2009), NOBLE (Tseytlin et al., 2016), ScispaCy (Neumann et al., 2019), MedTagger (Liu et al., 2013), CLAMP (Soysal et al., 2017), QuickUMLS (Soldaini and Goharian, 2016), Doc2Hpo (Liu et al., 2019), medspaCy (Eyre et al., 2021), EHRKit (Li et al., 2022), biomedical and clinical models of Stanza (Zhang et al., 2021), UmlsBERT (Michalopoulos et al., 2021), CancerBERT (Zhou et al., 2022), among others (Doan et al., 2014; Ford et al., 2016; Kreimeyer et al., 2017; Cho et al., 2020). Most of these clinical annotation systems rely on existing health and biomedical vocabularies such as Unified Medical Language System (UMLS) (Bodenreider, 2004) to perform a pattern matching to determine what information to extract and how to encode the extracted information. Documents in the EHRs often contain information that are challenging to extract such as negated sentences, abbreviations and acronym (Kaufman et al., 2016; Assale et al., 2019), and symptoms along with the terms describing their severity (Meystre et al., 2008); extracting such information is important to guide treatment and make informed decisions. For instance, prostate cancer patients treated by surgery, can report mild, moderate or severe urinary incontinence, quantifying this symptom can help determine the best treatment option either use protective pads or surgery (Bozkurt et al., 2020).

Several studies have compared the performance of concept recognition systems; however, they are typically developed internally and may be biased due to different internal implementations, parameters used, and limited number of systems included in the evaluations (Hassanzadeh et al., 2016; Gehrmann et al., 2018; Reátegui and Ratté, 2018; Wang X. et al., 2018). Thus, there is a lack of evidence on these systems' performances used by external scenarios (end-users) and for different clinical concept extraction tasks to support the most appropriate and suitable system for a particular clinical task. This study presents a comprehensive comparison of six concept recognition systems commonly used in the clinical and biomedical domain. We hypothesize that there is not a single system with the best performance over all clinical concept recognition tasks and challenges. We evaluate them using two datasets: the 2010 i2b2/VA challenge dataset for test, treatment, and problem concept extraction (Uzuner et al., 2011), and a sample drawn from the MIMIC-III (Medical Information Mart for Intensive Care III) (Johnson et al., 2016, 2018) clinical care

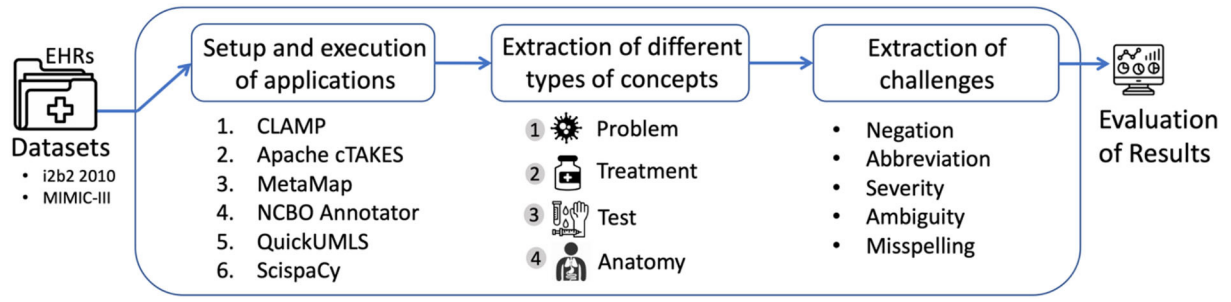

FIGURE 1
Workflow of the comparison of concept recognition systems across different tasks.

TABLE 1 Details of the i2b2 and MIMIC-III datasets.

	i2b2	MIMIC-III
Clinical records	Training: 170 Test: 256 Total: 426	Total: 27
Number of concepts	31,161	5,503
Concept types	Problem, treatment, and test	Problem, treatment, test, and anatomy
Total number of tokens	267,249	55,225
Average of number of tokens per note	1,043.9	2,045.4

database for problem, treatment, test, and anatomy concept extraction. We also evaluate how these systems handle known extraction challenges. This work fills a gap in the literature providing an external evaluation comparing concept recognition systems at extracting clinical concepts and known challenges on two clinical datasets.

2. Materials and methods

We investigated six biomedical/clinical concept recognition systems: (1) CLAMP, (2) Apache cTAKES, (3) MetaMap, (4) NCBO Annotator, (5) QuickUMLS, and (6) ScispaCy for the extraction of clinical concepts from unstructured EHR data using two datasets: i2b2 and MIMIC-III sample. We evaluated the performance on extracting clinical concepts, including problem (disease or disorder), treatment (procedure and drug), test, and anatomy. Additionally, we used MIMIC-III to examine the performance on how these systems handle known challenges, including abbreviations, negations, severity, ambiguity, and misspellings. Figure 1 outlines the workflow with the basic steps for this evaluation.

2.1. Datasets

Table 1 presents the detailed information of the two datasets.

2.1.1. i2b2

The 2010 i2b2 de-identified annotated dataset is composed of discharge summaries and progress reports from Beth Israel Deaconess Medical Center and University of Pittsburgh Medical Center (Uzuner et al., 2011). i2b2 is publicly-available and is used to evaluate several tasks based on clinical NLP methods such as assertion classification, and relation classification. In our study, we used the test partition of the i2b2 dataset for concept extraction which focused on the extraction of medical concepts such as problems, tests, and treatments from patient reports.

2.1.2. MIMIC-III

We used a sample drawn from MIMIC-III (Johnson et al., 2016, 2018) dataset that included clinical notes of patients in the ICU. We were interested in evaluating the different concept patterns from the health outcomes in patients receiving two chemotherapy agents, docetaxel and cisplatin. Docetaxel-cisplatin combination is a treatment option for specific types of aggressive cancer (Fan et al., 2013). Patients treated with the docetaxel-cisplatin were more likely to experience side effects such as anemia, nausea/vomiting, thrombocytopenia, etc. (Li et al., 2017), thus, their EHRs contained more symptoms and problems. A total of 27 clinical notes were included from MIMIC-III and were manually annotated by a clinical expert for four clinical concepts (Table 1), as well as for abbreviations, negations, severity, ambiguity, and misspellings (Table 2).

2.2. Concept recognition tools

We set up six clinical concept systems as described below.

TABLE 2 Details and examples of sentences annotated in the MIMIC-III dataset.

	Number of sentences	Number of concepts	Example sentences	Annotation
Abbreviation	123	169	He had a CXR that demonstrated possible left base consolidation	CXR: chest x-ray
Negation	169	228	She did not have fevers or chills until the day prior to admission when she noted chills	Fever, chills.
Severity	48	53	At least moderate pulmonary hypertension	Moderate
Ambiguity	26	26	He was then brought to the [**Hospital1 18**] ED for further management	ED: emergency department
Misspelling	43	43	Metastatic osteogenic sarcoma	Metastatic: metastatic

2.2.1. CLAMP

CLAMP is a Java-based clinical language annotation, modeling, and processing toolkit (CLAMP, 2021). CLAMP provides NLP modules, such as entity recognition, entity linking, normalization. It presents three different types of concept recognition methods: (1) a deep learning-based model that uses a recurrent neural network (RNN) within the bidirectional LSTM-CRF architecture; (2) a dictionary-based approach with comprehensive lexicon such as the UMLS; and (3) a regular expression-based algorithm to extract concept with common patterns. CLAMP includes NegEx (Chapman et al., 2001), a regular expression algorithm to identify negations. Moreover, additional negation lexicons and rules can be added.

2.2.2. cTAKES

Clinical Text Analysis and Knowledge Extraction System (cTAKES) is an open-source NLP system that combines rule-based and machine learning techniques to extract clinical information from EHR unstructured text (Savova et al., 2010; Kovačević et al., 2013). cTAKES executes some components in sequence to process clinical texts and mainly uses SNOMED-CT (Apache cTAKESTM, 2021). cTAKES also offers the extraction of negated concepts integrating NegEx.

2.2.3. MetaMap

MetaMap is a program providing access to the concepts in the unified medical language system (UMLS) Metathesaurus from biomedical text. It provides a link between the text of biomedical literature and the knowledge, including synonymy relationships, embedded in the Metathesaurus (Aronson and Lang, 2010; MetaMap, 2021). MetaMap includes the NegEx algorithm to extract negated concepts and allows the addition of new rules to identify negations.

2.2.4. NCBO annotator

The National Center for Biomedical Ontology (NCBO) Annotator is a publicly available Web service to process biomedical text and identify ontology concepts from over 1,100 ontologies (Jonquet et al., 2009; NCBO Annotator, 2021). The annotation is based on a syntactic concept recognition tool which uses concept names and synonyms. Moreover, new annotation features are provided through NCBO Annotator+, such as annotation scoring, negation detection (with NegEx/ConText algorithm), and temporality recognition (Tchechmedjiev et al., 2018).

2.2.5. ScispaCy

ScispaCy is a specialized Python NLP library for processing biomedical, scientific, and clinical texts (ScispaCy, 2021) which leverages the spaCy library (spaCy, 2021). ScispaCy is based on word embeddings and deep learning that uses a convolutional neural network (CNN) architecture. It contains three core released packages trained on biomedical text: (1) "en_core_sci_sm" with 100 k terms approximatively as vocabulary and no word vectors; (2) "en_core_sci_md" with 360 k terms as vocabulary and 50 k word vectors; and (3) "en_core_sci_lg" with 785 k terms approximatively as vocabulary and 600 k word vectors (Neumann et al., 2019; spaCy, 2021). ScispaCy does not include negation extraction and matches only 3-g terms to UMLS.

2.2.6. QuickUMLS

QuickUMLS is an unsupervised method for biomedical concept extraction (Soldaini and Goharian, 2016; QuickUMLS, 2021). QuickUMLS uses a simple and efficient algorithm for approximate dictionary matching designed for similarity measures such as cosine, Dice, Jaccard, and overlap coefficients (Okazaki and Tsujii, 2010). QuickUMLS does not provide the functionality to extract

negations and uses a subset of over 6 million concepts from UMLS.

2.3. Evaluation of clinical concept recognition systems

Using the two datasets, we evaluated the performance of the six concept recognition systems on extracting concepts, including problem, treatment, test, and anatomy. Additionally, we used MIMIC-III to examine the performance of the six systems in five challenging situations, including abbreviations, negations, severity, ambiguity, and misspellings. Results were compared among systems capable of addressing corresponding concepts. Sentences that contained the five challenges were annotated by the six clinical concept recognition systems, and the results were compared with the gold standard.

Evaluation of performance used the exact and inexact match of concepts (Uzuner et al., 2011). Exact means we only consider it correct when phrase boundaries and concept names matched exactly. Inexact matching represents a match over the surface string. The micro-averaged precision, recall, and F-score were compared across all systems for all types of concepts in the two datasets. Of note, we performed evaluations with different parameters and options that systems provide.

3. Results

3.1. Extraction of four clinical concepts

Table 3 shows the performance of the six systems on the two datasets for exact matching. Of note, CLAMP based on deep leaning outperformed the dictionary-based and regular expression-based methods, as well as ScispaCy with "en_core_sci_sm" outperformed the "en_core_sci_md" and "en_core_sci_lg" models. Thus, the following tables show the results of CLAMP based on deep leaning and ScispaCy with "en_core_sci_sm". By dataset, the performance varied across different systems in the MIMIC-III dataset with an F-score range between 0.03 and 0.39. While in the i2b2 dataset, we observed similar F-scores, ranging between 0.06 and 0.33 except for CLAMP (F-score 0.70). By clinical recognition systems, CLAMP achieved the best performance in both datasets with an F-score of 0.70 and 0.39 followed by ScispaCy with an F-score of 0.33 and 0.29.

Table 4 shows the performance of the systems for inexact matching. In general, all concept extraction systems performed better in inexact matching than exact matching evaluation. By dataset, the performance varied across different systems in the MIMIC-III dataset with an F-score range between 0.07 and 0.50. While in the i2b2 dataset, we observed similar F-scores, ranging between 0.08 and 0.56 except for CLAMP (F-score 0.94). By clinical recognition systems, CLAMP also obtained the best performance in both datasets with an F-score of 0.94 and 0.50 followed by ScispaCy with an F-score of 0.56 and 0.39.

3.2. Extraction of five challenges

We executed the six systems on the sentences that contained negations, abbreviations, severity, ambiguity, and misspellings in the MIMIC-III dataset and compared to the manually annotated results. Table 5 presents the six systems' performance on exact match at extracting the five challenging situations. We used exact matching since most of the entities are composed of a single word. Overall, there is no single system excelled in all tasks. Instead, each system performed differently in particular tasks. By clinical task, ScispaCy performed best in extracting abbreviation information with an F-score of 0.86, followed by NCBO annotator (F-score: 0.79).

In terms of extracting severity terms and negated concepts, CLAMP achieved the best performance with an F-score of 0.73 and 0.63, respectively. Figure 2 provides two sentences as examples to illustrate how different concept recognition systems extract negated concepts.

When evaluating severity terms, we observed that CLAMP outperformed the other systems. CLAMP was able to identify and categorize severity terms, while the other systems, such as MetaMap, usually identify such terms and categorize them into an UMLS semantic type named "qualitative measure", which is associated with "T080" as sematic type code (TUI). QuickUMLS identified and assigned these severity modifiers to another semantic type such as "finding" and "intellectual product" associated with TUIs "T033" and "T170". For example, in the sentence "moderate to severe tricuspid regurgitation", QuickUMLS identified "moderate" and assigned it "finding" and "intellectual product" as semantic types.

All systems obtained low F-scores for the extraction of ambiguity and misspelling. The list of concepts was composed of ambiguous abbreviations, i.e., abbreviations that were linked to more than one concept. Therefore, the systems were evaluated in terms of the extraction of the abbreviation and their full expansion. The best system for ambiguity abbreviations was MetaMap, which provided the CUI and the expansion of the abbreviation. While CLAMP was able to extract abbreviations, it did not provide the text expansion, only the CUI, therefore, its performance could not be evaluated. ScispaCy performed best at the identification of misspellings and the assignment to the correctly spelled terms. For instance, from the sentence "dilated and severely hypokinetic right ventricle", the systems

TABLE 3 Clinical concept recognition system performance on exact match at extracting clinical concepts from the i2b2 and MIMIC-III datasets.

	i2b2			MIMIC-III		
	Precision	Recall	*F*-score	Precision	Recall	*F*-score
CLAMP	**0.73**	**0.68**	**0.70**	**0.28**	**0.67**	**0.39**
cTAKES	0.16	0.24	0.19	0.18	0.37	0.24
MetaMap	0.13	0.46	0.20	0.12	0.58	0.20
NCBO Annotator	0.23	0.11	0.13	0.25	0.14	0.18
QuickUMLS	0.05	0.09	0.06	0.02	0.05	0.03
ScispaCy	0.25	0.54	0.33	0.19	0.64	0.29

TABLE 4 Clinical concept recognition system performance on inexact match at extracting clinical concepts from the i2b2 and MIMIC-III datasets.

	i2b2			MIMIC-III		
	Precision	Recall	*F*-score	Precision	Recall	*F*-score
CLAMP	**0.98**	**0.92**	**0.94**	**0.36**	**0.87**	**0.50**
cTAKES	0.30	0.45	0.35	0.25	0.51	0.34
MetaMap	0.24	0.83	0.36	0.17	0.81	0.28
NCBO Annotator	0.42	0.21	0.25	**0.36**	0.20	0.25
QuickUMLS	0.06	0.18	0.08	0.05	0.12	0.07
ScispaCy	0.41	0.91	0.56	0.26	**0.87**	0.39

TABLE 5 Clinical concept recognition system performance on exact match at extracting five challenges from MIMIC-III.

	Abbreviation			Negation			Severity			Ambiguity			Misspelling		
	R	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*	*R*	*P*	*F*
CLAMP	0.61	0.55	0.57	**0.61**	**0.66**	**0.63**	**0.73**	0.73	**0.73**						
cTAKES	0.62	0.53	0.56	0.41	0.49	0.43	0.39	0.56	0.44	0.04	0.04	0.04	0.02	0.02	0.02
MetaMap	0.66	0.59	0.61	0.52	0.52	0.51	0.47	0.62	0.52	**0.16**	**0.16**	**0.16**	0.02	0.02	0.02
NCBO Annotator	0.83	0.78	0.79	0.15	0.32	0.18	0.61	**0.85**	0.68	0.04	0.04	0.04	0.02	0.02	0.02
QuickUMLS	0.38	0.31	0.33				0.38	0.38	0.38	0.00	0.00	0.00	0.24	0.24	0.24
ScispaCy	**0.87**	**0.86**	**0.86**				0.33	0.33	0.33	0.04	0.04	0.04	**0.41**	**0.41**	**0.41**

are evaluated if they identify the incorrect word "severerly" and the assignment to the correct spelled word "severely".

4. Discussion

Clinical concept recognition is a common NLP task used to extract important concepts from clinical narrative text. There are many systems available to perform this task, yet limited external evaluation exists to guide end-users' selection. This study provides external evaluation of different clinical concept recognition tasks among six well-known systems, including CLAMP, cTAKES, MetaMap, NCBO Annotator, QuickUMLS, and ScispaCy. Our results indicate that CLAMP followed by ScispaCy outperformed the other systems when extracting clinical concepts from clinical notes. Similarly, CLAMP outperformed other systems regarding challenging concept recognition tasks, such as negation and ambiguous abbreviations. We observed that both CLAMP and ScispaCy systems integrate deep learning models (e.g., RNN and CNN architectures, respectively) that were trained using biomedical text, and this may explain the better performance of the systems. Moreover, NCBO Annotator, based on rule-based method, was the second best in clinical concept extraction on the MIMIC dataset as well as in abbreviation and severity recognition.

In our study, the concept recognition systems show better performance in the i2b2 compared to the MIMIC-III dataset. This is likely due to most of the systems used the i2b2 for

	Sentence 1 No other significant mediastinal, hilar or axillary lymphadenopathy is seen	***Sentence 2*** No Known Allergies / Adverse Drug Reactions
CLAMP	• other significant mediastinal • hilar • axillary lymphadenopathy	• Known Allergies / Adverse Drug Reactions
cTAKES	• axillary lymphadenopathy • lymphadenopathy	• Allergies • Drug Reactions
MetaMap	• axillary lymphadenopathy	• Allergies Adverse Drug Reactions
NCBO Annotator	-	• known

FIGURE 2
Example of negated concept extraction with four concept recognition systems.

their training processes and the different annotation process of clinical concepts on the two datasets. For instance, there are concepts composed of determiners in the i2b2 dataset. Determiners refer to those definite or indefinite articles (e.g., "the drop in hematocrit"), possessive pronouns (e.g., "her home medications"), and determinants (e.g., "a broken arm") before the concept name. The manually annotated MIMIC-III dataset did not include these determiners. Therefore, this indicates that system performance is directly related to the dataset's annotation process.

The systems evaluated in this paper performed differently on handling common challenges, including negated sentences, ambiguous terms, severity descriptions, acronyms/abbreviations, and misspellings. Challenges exist that limits the system performance on certain tasks such as extraction of negated concepts. Some terminologies already included negated concepts as part of their terminology. For instance, the negated concept "no thrombus" exists in SNOMED-CT. Systems that work with up-to-date terminologies first extract concepts that exist already as negated concepts in a terminology, and then they identify when a concept was negated (e.g., no presence of thrombus). In addition, some sentences contain more than one negated concept, and we found systems often fail to extract all negated concepts. If systems were able to extract the set of concept terms, they were often extracted and merged into one single term (see Figure 2, CLAMP extraction on Sentence 2). These issues compromised the system performance, thus a postprocessing step is recommended after concept extraction for an appropriate evaluation of the system performance. Also, CLAMP achieved the best performance in identifying and categorizing severity terms, while other systems, such as MetaMap and QuickUMLS, usually identify these terms and categorize them into a different UMLS semantic type. Thus, a careful evaluation of results of certain systems might be necessary to consider additional semantic types for a complete and correct extraction of severity terms.

All the concept recognition systems we evaluated showed relatively good results at identifying abbreviations. Most of the sentences in our study contained abbreviations that were linked to only one extension, such as "BRBPR" is associated with the extension "bright red blood per rectum", or "CXR" is associated with "chest x-ray". Still, we note that ScispaCy and NCBO annotator were more performant in extracting abbreviations than the other systems. ScispaCy depends on deep learning models with its own vocabularies trained on biomedical text, and NCBO annotator extracts more abbreviations since it uses more public terminologies than UMLS (Lossio-Ventura et al., 2019). The extraction of ambiguous abbreviations represents a harder challenge and all applications failed to extract such information, similar to a previous study (Wu et al., 2012). Our ambiguous terms list was composed of ambiguous abbreviations/acronyms, that can be associated with multiple meanings—which is common in clinical and biomedical text (Liu et al., 2015; Lossio-Ventura et al., 2018). For instance, in the sentence "He was then brought to the [**Hospital1 18**] ED for further management", all systems correctly identified "ED" as term, however, assigned the expansion/concept "erectile dysfunction" instead of "emergency department". In addition, many systems rely on the terminology from ULMS, which may not include all clinical abbreviations. Thus, future work in improving extraction of clinical ambiguous abbreviations is needed to ensure the correct interpretation of patient information from clinical notes.

Moreover, there are five important challenges related to concept recognition from clinical text, including negation, severity, abbreviation, ambiguity, and misspellings. These

tasks are important for clinical research, and particularly for electronic phenotyping and cohort selection (Banda et al., 2018; Hanauer et al., 2020). Eligibility criteria in clinical cohort may include patients that: did not have an arrhythmia, were diagnosed with coronary artery disease, and are taking some statin. The incorrect identification may lead to an incorrect patient cohort that include patients with the wrong eligibility or exclusion criteria. Such phenotyping is becoming critical, as federal initiative, such as the 21st Century Cures Act, are demanding the use of EHR text data to augment randomized control trials for clinical assertions (Hernandez-Boussard et al., 2019). Therefore, it may be reasonable that each system may be the most appropriate for different research tasks based on its performance.

On the other hand, in overall results were slightly different on both datasets i2b2 and MIMIC-III for exact and inexact match at extracting clinical concepts. As part of future work, other EHR-related datasets should be collected and annotated to allow the performance comparison of diverse clinical concept recognition systems on different datasets. In addition, a common preprocessing task could be added for all the datasets to reduce the noise and improve the recognition of concepts. Finally, named entity recognition tools recently proposed based on new deep learning techniques, such as medspaCy (Eyre et al., 2021), EHRKit (Li et al., 2022), biomedical and clinical models of Stanza (Zhang et al., 2021), and UmlsBERT (Michalopoulos et al., 2021), might be also added for comparison.

5. Conclusion

In conclusion, we found that each clinical concept recognition systems perform differently across various clinical tasks. Common challenges exist for all clinical concept recognition systems at extracting ambiguity and misspelling terms. Our work provides a benchmark for different clinical concept extraction systems in an external scenario by end-users that may be useful to other researchers when selecting a concept recognition system relevant to their clinical information extraction task. Our study suggests that CLAMP followed by ScispaCy are more consistent at extracting clinical information on clinical notes related to cancer patients receiving chemotherapy treatment. However, many challenges continue to underscore the performance of such systems, such as medical ambiguity and severity terms.

Data availability statement

The MIMIC-III dataset is freely available at https://mimic.physionet.org/ whose acquisition involves a required training, data use agreement, and corresponding credentials. The i2b2 (now n2c2) dataset is deidentified and freely available at https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/ upon completion of a data user agreement.

Author contributions

JALV contributed to conceiving the study idea and design, collected the data, set up the applications, and performed the evaluation. SB contributed to configure and evaluate the applications. RS led and performed two rounds of annotation on the MIMIC-III dataset. RS and THB contributed to the study design and provided significant feedback. JALV, RS, and THB wrote the initial draft and revised subsequent versions. All authors read, revised, and approved the final manuscript.

Funding

Research reported in this publication was supported by the National Cancer Institute of the National Institutes of Health under Award Number: R01CA183962.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Author disclaimer

The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Apache cTAKESTM (2021). *Clinical Text Analysis Knowledge Extraction System*. Available online at: https://ctakes.apache.org/ (accessed January 15, 2021).

Aronson, R., and Lang, F. M. (2010). An overview of MetaMap: historical perspective and recent advances. *J. Am. Med. Inform. Assoc.* 17, 229–236. doi: 10.1136/jamia.2009.002733

Assale, M., Dui, L. G., Cina, A., Seveso, A., and Cabitza, F. (2019). The revival of the notes field: leveraging the unstructured content in electronic health records. *Front. Med.* 6, 66. doi: 10.3389/fmed.2019.00066

Banda, J. M., Seneviratne, M., Hernandez-Boussard, T., and Shah, N. H. (2018). Advances in electronic phenotyping: from rule-based definitions to machine learning models. *Annu. Rev. Biomed. Data Sci.* 1, 53–68. doi: 10.1146/annurev-biodatasci-080917-013315

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* 32, D267–270. doi: 10.1093/nar/gkh061

Bozkurt, S., Paul, R., Coquet, J., Sun, R., Banerjee, I., Brooks, J. D., et al. (2020). Phenotyping severity of patient-centered outcomes using clinical notes: a prostate cancer use case. *Learn. Health Syst.* 4, e10237. doi: 10.1002/lrh2.10237

Chapman, W. W., Bridewell, W., Hanbury, P., Cooper, G. F., and Buchanan, G. B. (2001). A simple algorithm for identifying negated findings and diseases in discharge summaries. *J. Biomed. Inform.* 34, 301–310. doi: 10.1006/jbin.2001.1029

Cho, M., Ha, J., Park, C., and Park, S. (2020). Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition. *J. Biomed. Inform.* 103, 103381. doi: 10.1016/j.jbi.2020.103381

CLAMP (2021). *Natural Language Processing (NLP) Software*. Available online at: https://clamp.uth.edu/ (accessed January 15, 2021).

Demner-Fushman, D., Rogers, W. J., and Aronson, A. R. (2017). MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *J. Am. Med. Inform. Assoc.* 24, 841–844. doi: 10.1093/jamia/ocw177

Denny, J. C., Irani, P. R., Wehbe, F. H., Smithers, J. D., and Spickard, A. (2003). "The KnowledgeMap project: development of a concept-based medical school curriculum database," in *AMIA Annu. Symp. Proc. AMIA Symp.*, 195–199.

Doan, S., Conway, M., Phuong, T. M., and Ohno-Machado, L. (2014). Natural language processing in biomedicine: a unified system architecture overview. *Methods Mol. Biol.* 1168, 275–294. doi: 10.1007/978-1-4939-0847-9_16

Evans, R. S. (2016). Electronic health records: then, now, and in the future. *Yearb. Med. Inform.* 25 (Suppl. 1), S48–61. doi: 10.15265/IYS-2016-s006

Eyre, H., Chapman, A. B., Peterson, K. S., Shi, J., Alba, P. R., Jones, M. M., et al. (2021). Launching into clinical space with medspaCy: a new clinical text processing toolkit in Python. *AMIA Annu. Symp. Proc.* 2021, 438–447.

Fan, Y., Xu, B. H., Yuan, P., Ma, F., Wang, J. Y., Ding, X. Y., et al. (2013). Docetaxel-cisplatin might be superior to docetaxel-capecitabine in the first-line treatment of metastatic triple-negative breast cancer. *Ann. Oncol.* 24, 1219–1225. doi: 10.1093/annonc/mds603

Ford, E., Carroll, J. A., Smith, H. E., Scott, D., and Cassell, A. J. (2016). Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J. Am. Med. Inform. Assoc.* 23, 1007–1015. doi: 10.1093/jamia/ocv180

Friedman, C. (2000). "A broad-coverage natural language processing system," in *Proc. AMIA Symp*, 270–274.

Friedman, C., Alderson, P. O., Austin, J. H., Cimino, J. J., and Johnson, S. B. (1994). A general natural-language text processor for clinical radiology. *J. Am. Med. Inform. Assoc.* 1, 161–174. doi: 10.1136/jamia.1994.95236146

Gehrmann, S., Dernoncourt, F., Li, Y., Carlson, E. T., Wu, J. T., Welt, J., et al. (2018). Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PLoS ONE* 13, e0192360. doi: 10.1371/journal.pone.0192360

Hanauer, D. A., Barnholtz-Sloan, J. S., Beno, M. F., Fiol, G. D., Durbin, E. B., Gologorskaya, O., et al. (2020). Electronic medical record search engine (EMERSE): an information retrieval tool for supporting cancer research. *JCO Clin. Cancer Inform.* 4, 454–463. doi: 10.1200/CCI.19.00134

Hassanzadeh, H., Nguyen, A., and Koopman, B. (2016). "Evaluation of medical concept annotation systems on clinical records," in *Proceedings of the Australasian Language Technology Association Workshop 2016, Melbourne, Australia*, 15–24.

Hernandez-Boussard, T., Monda, K. L., Crespo, B. C., and Riskin, D. (2019). Real world evidence in cardiovascular medicine: ensuring data validity in electronic health record-based studies. *J. Am. Med. Inform. Assoc.* 26, 1189–1194. doi: 10.1093/jamia/ocz119

Hersh, W. R., Weiner, M. G., Embi, P. J., Logan, J. R., Payne, P. R. O., Bernstam, E. V., et al. (2013). Caveats for the use of operational electronic health record data in comparative effectiveness research. *Med. Care* 51 (Suppl. 3), S30–37. doi: 10.1097/MLR.0b013e31829b1dbd

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-W. H., Feng, M., Ghassemi, M., et al. (2016). MIMIC-III, a freely accessible critical care database. *Sci. Data* 3, 160035. doi: 10.1038/sdata.2016.35

Johnson, E., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The MIMIC Code Repository: enabling reproducibility in critical care research. *J. Am. Med. Inform. Assoc.* 25, 32–39. doi: 10.1093/jamia/ocx084

Jonquet, C., Shah, N. H., and Musen, A. M. (2009). The open biomedical annotator. *Summit Transl. Bioinforma.* 2009, 56–60.

Kaufman, D. R., Sheehan, B., Stetson, P., Bhatt, A. R., Field, A. I., Patel, C., et al. (2016). Natural language processing-enabled and conventional data capture methods for input to electronic health records: a comparative usability study. *JMIR Med. Inform.* 4, e35. doi: 10.2196/medinform.5544

Kovačević, A., Dehghan, A., Filannino, M., and Keane, J. A. (2013). Combining rules and machine learning for extraction of temporal expressions and events from clinical narratives. *J. Am. Med. Inform. Assoc.* 20, 859–866. doi: 10.1136/amiajnl-2013-001625

Kreimeyer, K., Foster, M., Pandey, A., Arya, N., Halford, G., Jones, S. F., et al. (2017). Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J. Biomed. Inform.* 73, 14–29. doi: 10.1016/j.jbi.2017.07.012

Li, A., Wei, Z.-J., Ding, H., Tang, H.-S., Zhou, H.-X., Yao, X., et al. (2017). Docetaxel versus docetaxel plus cisplatin for non-small-cell lung cancer: a meta-analysis of randomized clinical trials. *Oncotarget* 8, 57365–57378. doi: 10.18632/oncotarget.17071

Li, I., You, K., Tang, X., Qiao, Y., Huang, L., Hsieh, C. C., et al. (2022). *EHRKit: A Python Natural Language Processing Toolkit for Electronic Health Record Texts. arXiv*. Available online at: http://arxiv.org/abs/2204.06604 (accessed December 13, 2022).

Liu, C., Peres Kury, F. S., Li, Z., Ta, C., Wang, K., Weng, C., et al. (2019). Doc2Hpo: a web application for efficient and accurate HPO concept curation. *Nucleic Acids Res.* 47, W566–W570. doi: 10.1093/nar/gkz386

Liu, H., Bielinski, S. J., Sohn, S., Murphy, S., Wagholikar, K. B., Jonnalagadda, S. R., et al. (2013). An information extraction framework for cohort identification using electronic health records. *AMIA Jt. Summits Transl. Sci. Proc.* 2013, 149–153.

Liu, Y., Ge, T., Mathews, K. S., Ji, H., and McGuinness, D. L. (2015). Exploiting task-oriented resources to learn word embeddings for clinical abbreviation expansion. *Proc. BioNLP* 15, 92–97. doi: 10.18653/v1/W15-3810

Lossio-Ventura, J. A., Bian, J., Jonquet, C., Roche, M., and Teisseire, M. (2018). A novel framework for biomedical entity sense induction. *J. Biomed. Inform.* 84, 31–41. doi: 10.1016/j.jbi.2018.06.007

Lossio-Ventura, J. A., Boussard, S., Morzan, J., and Hernandez-Boussard, T. (2019). "Clinical named-entity recognition: a short comparison," in *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), San Diego, CA, USA*, 1548–1550.

MetaMap (2021). *A Tool For Recognizing UMLS Concepts in Text*. Available online at: https://metamap.nlm.nih.gov/ (accessed January 15, 2021).

Meystre, S. M., Savova, G. K., Kipper-Schuler, K. C., and Hurdle, J. F. (2008). Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb. Med. Inform.* 17, 128–144. doi: 10.1055/s-0038-1638592

Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., and Wong, A. (2021). "UmlsBERT: clinical domain knowledge augmentation of contextual embeddings using the unified medical language system metathesaurus," in *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 2021*, 1744–1753.

Murdoch, T. B., and Detsky, A. S. (2013). The inevitable application of big data to health care. *JAMA* 309, 1351–1352. doi: 10.1001/jama.2013.393

NCBO Annotator (2021). Available online at: https://bioportal.bioontology.org/annotator (accessed January 15, 2021).

Neumann, M., King, D., Beltagy, I., and Ammar, W. (2019). "ScispaCy: fast and robust models for biomedical natural language processing," in *Proc. 18th BioNLP Workshop Shar. Task*, 319–327.

Okazaki, N., and Tsujii, J. (2010). "Simple and efficient algorithm for approximate dictionary matching," in *Proceedings of the 23rd International Conference on Computational Linguistics, Beijing, China*, 1851–859. Available online at: https://www.aclweb.org/anthology/C10-1096/ (accessed January 15, 2021).

Pradhan, S., Elhadad, N., Chapman, W., Manandhar, S., and Savova, G. (2014). "SemEval-2014 task 7: analysis of clinical text," in *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), Dublin, Ireland*, 54–62.

Pradhan, S., Elhadad, N., South, B. R., Martinez, D., Christensen, L., Vogel, A., et al. (2015). Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *J. Am. Med. Inform. Assoc.* 22, 143–154. doi: 10.1136/amiajnl-2013-002544

QuickUMLS (2021). *System for Medical Concept Extraction and Linking*. https://github.com/Georgetown-IR-Lab/QuickUMLS (accessed January 15, 2021).

Reátegui, R., and Ratté, S. (2018). Comparison of MetaMap and cTAKES for entity extraction in clinical notes. *BMC Med. Inform. Decis. Mak.* 18, 74. doi: 10.1186/s12911-018-0654-2

Savova, G. K., Masanz, J. J., Ogren, P. V., Zheng, J., Sohn, S., Kipper-Schuler, K. C., et al. (2010). Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J. Am. Med. Inform. Assoc.* 17, 507–513. doi: 10.1136/jamia.2009.001560

ScispaCy (2021). *SpaCy Models for Biomedical Text Processing*. Available online at: https://github.com/allenai/scispacy (accessed January 15, 2021).

Soldaini, L., and Goharian, N. (2016). "QuickUMLS : a fast, unsupervised approach for medical concept extraction," in *Medical Information Retrieval (MedIR) Workshop, Pisa, Italy*, 4.

Soysal, E., Wang, J., Jiang, M., Wu, Y., Pakhomov, S., Liu, H., et al. (2017). CLAMP - a toolkit for efficiently building customized clinical natural language processing pipelines. *J. Am. Med. Inform. Assoc.* 25, 331–336. doi: 10.1093/jamia/ocx132

spaCy (2021). *Natural Language Processing in Python*. Available online at: https://spacy.io/. (accessed January 15, 2021).

Tchechmedjiev, A., Abdaoui, A., Emonet, V., Melzi, S., Jonnagaddala, J., Jonquet, C., et al. (2018). Enhanced functionalities for annotating and indexing clinical text with the NCBO Annotator. *Bioinforma. Oxf. Engl.* 34, 1962–1965. doi: 10.1093/bioinformatics/bty009

Tseytlin, E., Mitchell, K., Legowski, E., Corrigan, J., Chavan, G., Jacobson, R. S., et al. (2016). NOBLE - Flexible concept recognition for large-scale biomedical natural language processing. *BMC Bioinformatics* 17, 32. doi: 10.1186/s12859-015-0871-y

Uzuner, O., Solti, I., and Cadag, E. (2010). Extracting medication information from clinical text. *J. Am. Med. Inform. Assoc.* 17, 514–518. doi: 10.1136/jamia.2010.003947

Uzuner, Ö., South, B. R., Shen, S., and DuVall, S. L. (2011). 2010 i2b.2/VA challenge on concepts, assertions, and relations in clinical text. *J. Am. Med. Inform. Assoc.* 18, 552–556. doi: 10.1136/amiajnl-2011-000203

Wang, X., Yang, C., and Guan, R. (2018). A comparative study for biomedical named entity recognition. *Int. J. Mach. Learn. Cybern.* 9, 373–382. doi: 10.1007/s13042-015-0426-6

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., et al. (2018). Clinical information extraction applications: a literature review. *J. Biomed. Inform.* 77, 34–49. doi: 10.1016/j.jbi.2017.11.011

Wu, Y., Denny, J. C., Rosenbloom, S. T., Miller, R. A., Giuse, D. A., Xu, H., et al. (2012). A comparative study of current Clinical Natural Language Processing systems on handling abbreviations in discharge summaries. *AMIA Annu. Symp. Proc. AMIA Symp.* 2012, 997–1003.

Yim, W.-W., Yetisgen, M., Harris, W. P., and Kwan, W. S. (2016). Natural language processing in oncology: a review. *JAMA Oncol.* 2, 797–804. doi: 10.1001/jamaoncol.2016.0213

Zeng, Q. T., Goryachev, S., Weiss, S., Sordo, M., Murphy, S. N., Lazarus, R., et al. (2006). Extracting principal diagnosis, co-morbidity and smoking status for asthma research: evaluation of a natural language processing system. *BMC Med. Inform. Decis. Mak.* 6, 30. doi: 10.1186/1472-6947-6-30

Zhang, Y., Zhang, Y., Qi, P., Manning, C. D., and Langlotz, C. P. (2021). Biomedical and clinical English model packages for the Stanza Python NLP library. *J. Am. Med. Inform. Assoc.* 28, 1892–1899. doi: 10.1093/jamia/ocab090

Zhou, S., Wang, N., Wang, L., Liu, H., and Zhang, R. (2022). CancerBERT: a cancer domain-specific language model for extracting breast cancer phenotypes from electronic health records. *J. Am. Med. Inform. Assoc.* 29, 1208–1216. doi: 10.1093/jamia/ocac040