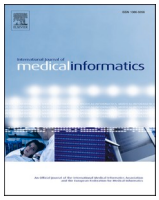
International Journal of Medical Informatics 177 (2023) 105122

Contents lists available at ScienceDirect

International Journal of Medical Informatics

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



Review article

Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review

David Fraile Navarro [a,*], Kiran Ijaz [a], Dana Rezazadegan [b], Hania Rahimi-Ardabili [a], Mark Dras [c], Enrico Coiera [a], Shlomo Berkovsky [a]

[a] Centre for Health Informatics, Australian Institute of Health Innovation, Macquarie University, Sydney, Australia
[b] Department of Computer Science and Software Engineering. School of Software and Electrical Engineering, Swinburne University of Technology, Melbourne, Australia
[c] Department of Computing, Macquarie University, Sydney, Australia

A B S T R A C T

Background: Natural Language Processing (NLP) applications have developed over the past years in various fields including its application to clinical free text for named entity recognition and relation extraction. However, there has been rapid developments the last few years that there's currently no overview of it. Moreover, it is unclear how these models and tools have been translated into clinical practice. We aim to synthesize and review these developments.

Methods: We reviewed literature from 2010 to date, searching PubMed, Scopus, the Association of Computational Linguistics (ACL), and Association of Computer Machinery (ACM) libraries for studies of NLP systems performing general-purpose (i.e., not disease- or treatment-specific) information extraction and relation extraction tasks in unstructured clinical text (e.g., discharge summaries).

Results: We included in the review 94 studies with 30 studies published in the last three years. Machine learning methods were used in 68 studies, rule-based in 5 studies, and both in 22 studies. 63 studies focused on Named Entity Recognition, 13 on Relation Extraction and 18 performed both. The most frequently extracted entities were "problem", "test" and "treatment". 72 studies used public datasets and 22 studies used proprietary datasets alone. Only 14 studies defined clearly a clinical or information task to be addressed by the system and just three studies reported its use outside the experimental setting. Only 7 studies shared a pre-trained model and only 8 an available software tool.

Discussion: Machine learning-based methods have dominated the NLP field on information extraction tasks. More recently, Transformer-based language models are taking the lead and showing the strongest performance. However, these developments are mostly based on a few datasets and generic annotations, with very few real-world use cases. This may raise questions about the generalizability of findings, translation into practice and highlights the need for robust clinical evaluation.

1. Introduction

With the advent of the information age, clinical documentation has moved from paper records into a digital format, commonly known as *electronic health records* (EHRs) [1]. Nowadays, EHRs have taken a central role in modern medicine and clinicians spend considerable time dealing with them [2], including an ever-increasing number of administrative tasks, contributing to the clinician's burnout [3] and using an increasing part of clinician's time [2].

Over the recent years, we have seen a proliferation of Artificial Intelligence (AI) applications in a range of fields, including medicine [4], in areas such as image classification in radiology [5] or dermatology [6]. Owing to novel AI techniques, natural language processing (NLP) has also experienced an increased interest in processing clinical free text. Among the most promising NLP tasks in clinical medicine [7], Information Extraction (IE) and its subcomponents of Named Entity Recognition (NER) and Relation Extraction (RE) are critical elements for developing learning systems and unleashing the potential of AI. Although no single definition of NER exists [8], NER could be defined as a *sub-task of IE consisting of extracting lexical units referring to a real-world entity in the specific domain of medicine*. Hence, RE is defined as the *extraction of semantic relations between two or more named entities*. Fig. 1 shows an example of NER extraction using the Stanford Stanza's system [9]. In this system, an example discharge letter is processed, and clinical entities are highlighted as "PROBLEM" (any form of symptom, sign, or disease) and "TREATMENT" (for any drug or therapy).

Among others, potential NER tasks in medicine include extraction of symptoms, signs, diseases, and treatments. Medically relevant RE includes medication-indication relations, drug-drug interactions, symptom-diagnose relations, and more. Harnessing NER and RE and building intelligent systems on top of them could allow reducing the time spent manually coding diagnoses, creating intelligent retrieval systems, or building data-intensive prediction models, all contributing to higher standards of care and potentially decreasing EHR associated

* Corresponding author at: Level 6, 75 Talavera Rd, North Ryde 2109, NSW, Australia.
E-mail address: david.frailenavarro@hdr.mq.edu.au (D. Fraile Navarro).

https://doi.org/10.1016/j.ijmedinf.2023.105122
Received 9 October 2022; Received in revised form 14 April 2023; Accepted 3 June 2023
Available online 5 June 2023

```
doc = nlp('The patient had a sore throat and fever that  was treated
with panadol and ibuprofen.')
for ent in doc.entities:
    print(f'{ent.text}\t{ent.type}')

a sore throat   PROBLEM
fever   PROBLEM
panadol TREATMENT
ibuprofen   TREATMENT
```

Fig. 1. Example of multi-entity Clinical NER using Stanza [9].

burnout. One of the most challenging tasks is to extract and establish clinical relations among the different terms of clinical semiology (e.g., symptoms and signs with diseases and syndromes). Establishing these relations into a coherent "ontology" has been one of the major tasks for which NLP approaches have been considered and researched extensively in the past [10].

Specific challenges limiting the development and application of NLP in the health domain include the need for both clinical datasets [11] and high-quality annotations to develop NLP models. To address these, a few activities, such as releasing public deidentified datasets such as MIMIC [12] or creating specific clinical NLP challenges such as i2b2 / n2c2 [13] have been introduced. While they produced a wealth of algorithmic techniques [13-15], it remains unclear how these solutions can be implemented in the healthcare context and there is a gap between the experimental development and its real-world implementation in the clinical practice [16].

Previous reviews have examined clinical applications of NLP [17,18] but are currently outdated in respect of the recent developments in the field, especially as they do not include Transformer-based models, that dramatically transfigured the NLP landscape, with models like BIOBERT [19] particularly relevant here (which was explicitly excluded in Wu et al.[18]). Moreover, previous reviews [17,18] provide a good general overview of all clinical NLP tasks but do not detail specifically how NER and RE tasks were conceptualized nor provide a comparison between systems, their performance and availability so it can guide its application by current NLP practitioners.

In this work, we aim to review NLP methods that extract clinical entities and/or their relations from unstructured clinical text and appraise their implementation potential.

2. Methods

2.1. Protocol and registration

The protocol of this systematic review has been registered in PROSPERO (International Prospective Register of Systematic Reviews) under the identifier CRD4202017803. We followed the PRISMA statement [20] for reporting and structuring this review.

2.2. Eligibility criteria

We included studies using any computational methods to extract diagnoses, treatments, and clinical semiology entities, defined as any form of free-text information related to the pathophysiological manifestations of diseases including symptoms and signs of disease. We included studies describing computational methods that perform either NER or RE. Only the studies that reported the above methods and performed these tasks in a multi-entity manner, meaning that they are designed to extract all clinical entities in a given text and not tailored to a specific disease or clinical task (e.g., smoking information, diabetes medication, cardiovascular semiology, etc.) were included. Studies that only extracted a specific type of clinical information (e.g., only medications, not in conjunction with symptoms or diagnoses) were excluded. We also excluded the studies that did not either conduct any development and model validation (e.g., to test performance) or report their deployment and related accuracy in clinical free-text. We defined *clinical free-text* as an unstructured, written clinical textual documentation produced during the care, management or follow-up of a patient. Examples of clinical free-texts include: discharge summaries, progress notes and referral letters. Studies where all the data came exclusively from biomedical literature, online forums, or social media, were excluded. We also excluded studies focusing solely on pathology or radiology reports as these diagnostic reports use a highly and structured language, where findings relate to imaging descriptions, either radiology or microscopy, and utilize a more constrained language than free-text clinical notes. Lastly, we excluded studies focusing on extracting temporal relations exclusively, rather than semantic relations, as this is usually considered a separate NLP task.

We have not limited our review to certain types of studies, including either clinical trials or synthetic experiments. Both publications of journals and conference proceedings are included as they are standard research outlets in the Computer Science field. However, secondary research such as opinion pieces or methodological surveys of the target technologies were excluded.

2.3. Information sources and search

The search was conducted on 15th June 2022 in PubMed, Scopus, the Association for Computer Machinery (ACM) Digital Library, and the Association for Computational Linguistics (ACL) library. It was restricted to articles published since 2010 in the English language. We did not search in pre-print repositories, however, after screening references of the included studies, additional highly relevant pre-print studies (cited more than 500 times in included peer-reviewed publications and not published subsequently as a peer reviewed publication) were included. The complete search strategy is available in Appendix 1.

2.4. Study selection

Three authors (DFN, KI, HR-A) performed the study selection and screening process. After initial calibration, with a 10% sample screened in duplicate, each reviewer screened the studies independently. Disagreements were resolved through discussion and with the help of a third reviewer (SB). The screening and study selection process was conducted using the RAYYAN platform [21].

2.5. Data extraction

We piloted data extraction with 10% of the selected studies extracted in duplicate by reviewers (DFN, DR, KI). After the initial calibration, these reviewers performed the complete data extraction independently. Disagreements were resolved through discussion and with the help of a third reviewer (DFN, DR, or KI).

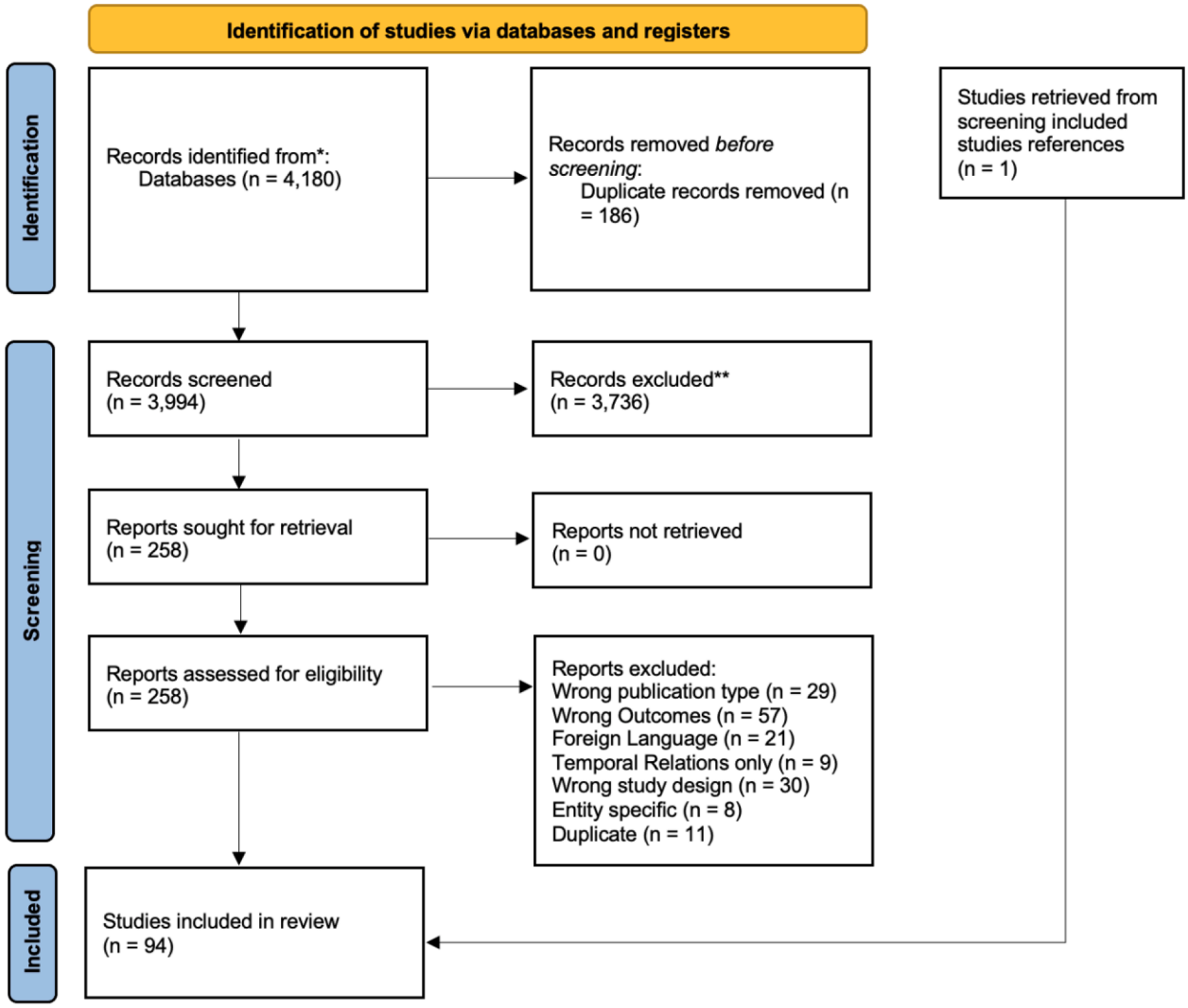


Fig. 2. PRISMA Flow-chart.

2.6. Data items, summary measures and synthesis of results

For each study, we extracted the following information: year of publication, NLP task(s) addressed (NER and/or RE), methods used (rule-based or machine learning (ML)), dataset(s) used, clinical entities of interest extracted, whether the study was part of a challenge, accuracy metrics(s) and additional performance metrics, and extraction of specific semantic properties: assertion (presence or absence of entities) and intensity (e.g., degree of severity for pain). We also extracted whether the system has been deployed in a real-world setting (defined as the system being used in piloting, trialing or routinely in any clinical setting) and the clinical or information task(s) to be addressed by the NLP system, as defined by the authors. There is currently no appropriate tool to measure the risk of bias in NLP clinical studies; however, we extracted from each study the description of the methods (either using algorithmic formulas, pseudocode, or a description of system architecture), whether the source code was provided, and whether the trained model or a software package was made available. When a study presented several systems and compared their performance for a given task, we extracted the accuracy metrics from the system that was proposed in the paper and/or achieved the highest score overall. For example, if a BERT model was developed and compared to a previous BiLSTM model, BERT's performance was extracted and reported. Additionally, for papers that proposed more than one method, we also extracted all the additional methods reported in the studies.

For each study, we compiled the reported accuracy metrics for the best performing NER or RE method. We calculated descriptive statistics for categorical variables and performed a descriptive analysis of the accuracy measurements and other variables collected (frequencies). Due to the nature of this review, and the high variety of methods, entities, and datasets; we did not perform a *meta*-analysis or use regression models across studies.

3. Results

Our search yielded 4,180 results. After removing duplicates, titles and abstracts of 3,613 articles were screened. 186 articles were selected for full-text screening and after a reference search of the included studies, given its contribution to the field (over 500 citations in Google Scholar) one additional study by Alsentzer et al.[22] published in a preprint server was included. In total, 94 articles were included in this review. Fig. 2 (PRISMA Flow-chart) provides detailed information on the inclusion and exclusion of studies. Considering the publication year, 15 studies (16%) were published in 2019 and additional 15 studies (16%) in 2020, 10 studies (11%) in 2021 followed by 8 studies (9%) in 2011 and 2018, 3 to 6 studies in 2011–2017, and only 2 studies in 2010. In the current year, 2022, 5 studies were published, bearing in mind that the search only included the first half of the year (see Fig. 3).

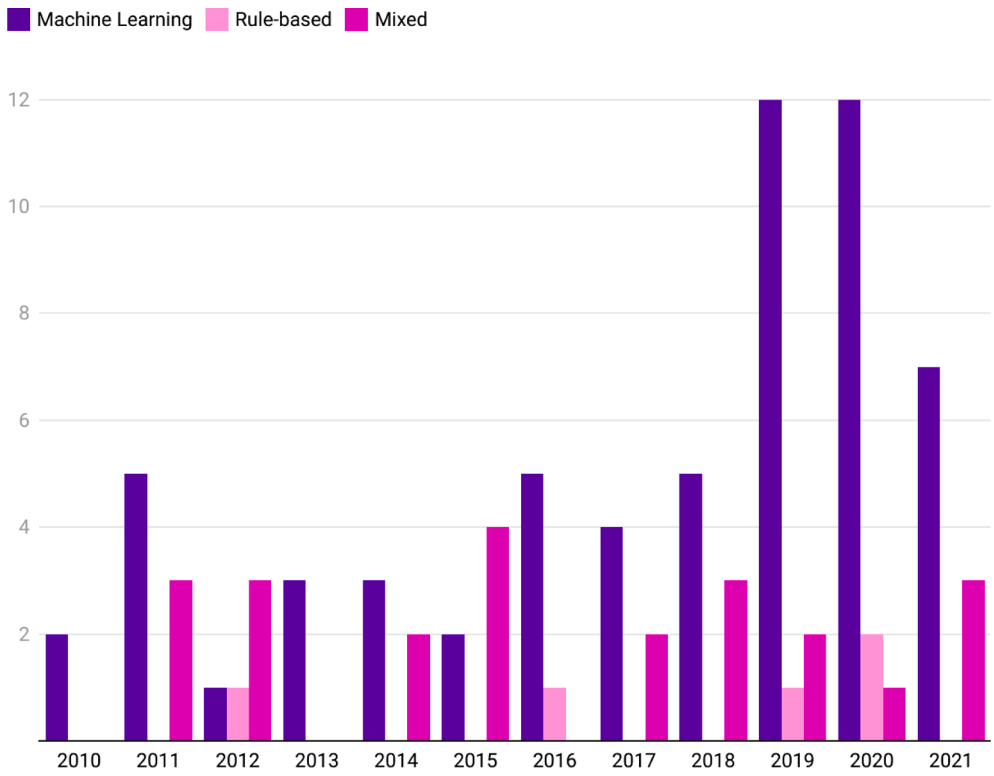


Fig. 3. Number of articles and methods included by the year of publication.

3.1. NLP methodology

Regarding NLP methodologies used, 68 (72%) of the included studies used ML methods compared to 22 studies (24%) combining ML and rule-based methods and 5 studies (5%) using rule-based methods alone. Considering the IE tasks explored, 63 studies (67%) focused on NER, 13 studies (14%) on RE and 18 studies (19%) performed both NER and RE (see Table 1**)**.

Considering the specific methods used to perform these tasks, we found a wide variety of ML and rule-based approaches. Across all the studies, 154 ML methods, pipelines, and algorithm implementations were reported in total. In 42 studies (45%) only one method was reported, whereas 35 studies (37%) reported two methods, and 14 studies (15%) reported more than two. The most frequent method was Conditional Random Fields (CRF) [23] which was used alone or in combination with other methods in 28 studies (30%) and Long Short-Term Memory (LSTM) [24] and Bidirectional Long Short-Term Memory (Bi-LSTM) [25] which were also used in 27 studies (29%), Bidirectional Encoder Representations from Transformers (BERT) [26] in 20 studies (21%), and Support Vector Machines (SVM) [27] were used in 16 studies (17%). The rest used different neural network (NN) architectures or did not specify the deployed model, referring generally to NNs, recurrent NNs, convolutional NNs, or deep NNs. Other specific NLP models included other word embeddings methods in 10 studies (11%), ELMO [28] architecture in 2 studies, global vectors (GloVe) [29] in two studies, and bag-of-words in another one. Studies not using ML methods included rule-based approaches in 19 studies (20%) and dictionaries and/or ontologies in 9 studies (10%). Table 2 describes the most common methods deployed in these studies.

3.2. Datasets and challenges

72 studies (77%) relied exclusively on public datasets, whereas 22 studies (23%) used proprietary datasets (e.g., hospital EHRs). 73 studies (78%) used the i2b2 challenge datasets and/or annotations, and among those, 55 studies (59%) used specifically the i2b2 2010 challenge dataset [15]. 13 studies (14%) used the MIMIC-II or MIMIC-III datasets [12], 10 studies (11%) used the SHARE/CLEF challenge [30] datasets and 8 studies (9%) used data from the SemEval challenges [31]. Lastly 4 studies (4%) used the MADE 1.0 challenge [32] dataset and 4 other used MedNLI (which is a subset of the MIMIC dataset) [33] (see Fig. 4**)**. Considering the challenges, 45 studies (48%) directly responded to a computational challenge and additional 17 studies (18%) addressed parts of those challenges or used them for benchmarking purposes. The remaining 32 studies (34%) did not respond to challenges, although some of them still used the datasets or annotations released by them.

3.3. Method description, source code, model, and software availability

Among the included studies, 32 (34%) provided the formal mathematical basis for their models, 48 studies (51%) included the pseudocode or a model architecture diagram, and 14 studies (15%) shared the source code utilized for model development. Only 7 studies (7%) supplied the trained models, and 8 studies (9%) provided a software solution available for download (see Table 3). Appendix 2 provides links to source materials, when these were made available by the authors.

3.4. Clinical named entities

Considering the entities being extracted, “treatments”, “medications” or “drugs” were the most common extracted type of clinical entity present in 66 studies (70%). 48 studies (51%) extracted the triad of entities (“problem”, “test”, “treatment”), as annotated in the i2b2 2010 challenge [15]. In the remaining works, the following entities were extracted: disorders or diagnoses (24 studies), symptoms and signs (14 studies), adverse drug effects or side effects (13 studies), indication or reason (10 studies), procedures (8 studies, drug-related entities (e.g., dosage, frequency, duration) (6 studies), body location or anatomical site (5 studies), laboratory tests (4 studies) and International Classification of Diseases (ICD) codes (3 studies). Other extracted entities

Table 1
Methods, NLP tasks, and named entities extracted.

Author	Year	NLP Method(s)	NLP task	Problem	Test	Treatment – Medication – Drug	Symptoms	Indication – Reason	Disorder-Disease-Diagnose	Relations	Adverse drug events – side effects	Procedure	Body Part	Other
Abadeer[34]	2020	ML	NER	X	X	X								
Alsentzer[22]	2019	ML	NER	X	X	X								
Bejan[35]	2014	ML	RE			X		X						
Ben Abacha[36]	2011	Rule-based + ML	NER	X	X	X								
Bhatia[37]	2019	ML	NER	X	X	X								
Chalapathy[38]	2016	ML	NER	X	X	X								
Chatzimina[39]	2014	ML	NER						X					
Chen[40]	2015	ML	NER	X	X	X								
Chodey[41]	2016	ML	NER						X					
D' Souza[42]	2014	ML + Rule-based	RE	X	X	X				X				
D'Avolio[43]	2011	ML	NER	X	X	X								
Dai[44]	2020	ML	NER	X	X	X								
de Bruijn[15]	2011	ML + Rule-based*	NER + RE	X	X	X								
Dirkson[45]	2021	ML	NER			X					X			
Divita[46]	2017	ML + Rule-based	NER				X							
Divita[47]	2014	ML	NER	X			X							
Dligach[48]	2013	ML	RE			X	X		X			X	X	
Doan[49]	2010	ML	NER			X								
Doğan[50]	2011	ML	NER + RE	X	X	X				X				
Ghiasvand[51]	2014	Rule-based + ML	NER						X					
Gligic[52]	2020	ML	NER + RE			X		X						Drug-related values
Hao[53]	2020	Rule-based + ML	NER + RE	X	X	X								temporal relation to past medical history
Hussain[54]	2020	Rule-based	NER						X					
Jagannatha[32]	2019	ML	NER + RE			X	X		X	X				
Jagannatha[55]	2016	ML	NER			X		X			X			
JianG[56]	2011	ML	NER	X	X	X								
Jiang[57]	2012	ML	NER	X	X	X								
Jiang[58]	2019	ML	NER	X	X	X								
Jonnalagadda[59]	2012	Rule-based + ML	NER	X	X	X								
Ju[60]	2020	ML	NER			X					X			
Kang[61]	2012	ML + Rule-based	NER	X	X	X								
Keretna[62]	2015	ML	NER	X	X	X								
Kim[63]	2019	ML + Rule-based	NER						X					cancer classification
Kim[64]	2015	ML + Rule-based	NER	X	X	X			X			X		

(*continued on next page*)

Table 1 (*continued*)

Author	Year	NLP Method(s)	NLP task	Problem	Test	Treatment – Medication – Drug	Symptoms	Indication – Reason	Disorder-Disease-Diagnose	Relations	Adverse drug events – side effects	Procedure	Body Part	Other
KraLjevic[65]	2021	Rule-based + ML	NER			x	x		X			x		Mental or behavioural dysfunction
Leaman[66]	2015	ML	NER						X					
Lee[67]	2019	ML	NER	X	X	X	X		X			X		lab test
Li F 1 [68]	2018	ML	NER + RE			X		X			X			SSLIF (any symptom sign or disease, not ADE)
Li F 2 [19]	2019	ML	NER			X		X			X			
Li L [69]	2020	ML	NER	X	X	X								
Li Z[70]	2019	ML	NER + RE	X	X	X				X				Relations previously described in other papers
Lin[71]	2017	ML	NER											TRNAP… etc ICD-10)
Liu[72]	2016	ML	NER	X	X	X								occurrence
Luo[73]	2017	ML	RE	X	X	X						X		
Manimaran[74]	2018	ML + Rule-based	NER	X	X	X						X		
Minard[75]	2011	ML	RE	X	X	X				X				
Narayanan[76]	2020	ML	NER								X			
Nath[77]	2021	Rule-based + ML	NER	X	X	X								
Nguyen[78]	2018	ML	NER											dx. codes
Patrick[79]	2011	ML + Rule-based	NER + RE	X	X	X								drug and dose
Peng[80]	2020	ML	NER + RE	X	X	X	X		X	X				
Pradhan[81]	2015	Rule-based + ML	NER						X					
Qin[82]	2018	ML	NER	X	X	X								
Raj[83]	2017	ML	RE	X	X	X				X				
Ramanan[84]	2016	Rule-based	NER				X		X				X	lab test, sex, age, outcome
Rea[85]	2012	Rule-based	NER			X			X					glucose
Rink[86]	2011	ML	RE							X				
Sahu[87]	2016	ML	RE							X				
Shi[88]	2019	ML	NER + RE						X	X			X	
Si [89]	2019	ML	NER	X	X	X								
Steinkamp[90]	2020	ML	NER				X							
Suster[91]	2018	ML + Rule-based	RE							X				
Tang[92]	2013	ML	NER	X	X	X								Frequency, duration
Tao[93]	2019	Rule-based + ML	NER + RE			X			X	X	X			
Tao[94]	2018	ML	NER + RE											
Tarcar[95]	2020	ML	NER			X	X		X					Dosage
Trivedi[96]	2020	Rule-based	NER		X		X		X			X		Family History, Situations affecting health

(*continued on next page*)

Table 1 (*continued*)

Author	Year	NLP Method(s)	NLP task	Problem	Test	Treatment – Medication – Drug	Symptoms	Indication – Reason	Disorder-Disease-Diagnose	Relations	Adverse drug events – side effects	Procedure	Body Part	Other
Wang[97]	2015	Rule-based + ML	NER						X					
Wang[98]	2018	ML	NER + RE											fractures, smoking, drug-drug interaction
Wei[99]	2020	ML	NER + RE			X		X			X			Drug-related values
Wu[100]	2017	ML	NER	X	X	X								
Wu[101]	2018	ML + Rule-based	NER	X	X	X								
Wu[102]	2015	ML	NER	X	X	X								
Xie[103]	2019	ML	NER											ICD-9
Xu J[104]	2019	ML	RE					X	X				X	lab test, drug-related values
Xu Y[105]	2012	ML + Rule-based	NER + RE	X	X	X				X				
Yang X 1[106]	2019	ML	NER + RE			X	X	X		X	X			
Yang X 2 [107]	2020	ML	NER + RE			X				X	X			
Yang X 3[108]	2020	ML	NER	X	X	X					X			
Yehia[109]	2019	Rule-based	NER + RE			X	X		X			X		demographics, vital signs, examination, lab test, RX
Zheng[110]	2017	Rule-based + ML	NER			X			X					
Zhu[111]	2013	ML	RE	X	X	X				X				
Roy[112]	2021	ML	RE							X				
Michalopoulos [113]	2021	Rule based + ML	NER	X	X	X								
Khetan[114]	2022	ML	RE							X				Causal Relations
Phan[115]	2022	ML	NER	X	X	X								
Khandelwal[116]	2022	ML	NER						X					
Narayanan & Mannam[117]	2022	ML	NER	X	X	X		X			X			
Mulyar[118]	2021	ML	NER	X	X	X		X			X			
Li & Zhou[119]	2021	ML	NER	X	X	X								
Zhang[9]	2021	ML	NER	X	X	X								
Tang & Yu[120]	2021	ML	NER	X	X	X								
Moqurrab[121]	2021	ML	NER	X	X	X								
Dave[122]	2022	ML	NER				X						X	Pain Symptoms
TOTAL	ML:66 ML + Rule-based: 23 Rule-based: 5 NER:63 RE: 13 NER + RE:18			48	48	66	14	10	24	18	13	8	5	–

*De Bruijn et al. reported multiple methods (both ML and rule-base.

Table 2
Methods deployed in the included studies.

Author	Dictionary / Ontology	Rule-based	Neural Network (not specified)	LSTM-BILSTM	CRF	SVM	Word embeddings (any type)	BERT / TRANSFORMERS	Other
Abadeer[34]								X	
Alsentzer[22]								X	ClinicalBERT*
Bejan[35]		X							
Ben Abacha[36]		X			X	X			
Bhatia[37]			X	X					
Chalapathy[38]					X				
Chatzimina[39]					X				PCFG-LA Probabilistic Context-Free Grammar with Latent Categories
Chen[40]					X				
Chodey[41]					X				
D' Souza[42]						X			
D'Avolio[43]					X				
Dai[44]				X				X	
de Bruijn[15]									List of different methods and approaches described.
Dirkson[45]								X	BERT (based on Distilbert) 2 implementations FuzzyBIO and BIOHD
Divita[46]	X								
Divita[47]		X							
Dligach[48]						X			
Doan[49]						X			
Doğan[50]					X	X			
Ghiasvand[51]		X			X				
Gligic[52]				X					seq2seq
Hao[53]	X							X	BERT pre-trained in MIMIC III plus UMLS Knowledge Base
Hussain[54]		X							
Jagannatha[32]				X	X				
Jagannatha[55]				X	X	X			Random Forest
JianG[56]		X							
Jiang[57]			X	X			X		ELMO, Flair
Jiang[58]	X								
Jonnalagadda[59]	X				X				
Ju[60]			X						
Kang[61]	X								
Keretna[62]			X		X				Naive Bayes, ME, Random Tree, C4.5 Ada boost, Random Forest
Kim[63]		X			X	X			
Kim[64]									Ensemble Methods
KraLjevic[65]		X					X	X	Word2vec and BERT
Leaman[66]					X				
Lee[67]					X				
Li F 1 [68]			X					X	
Li F 2 [19]				X	X				BILSTM-CRF and BILSTM ATTENTION
Li L [69]			X	X	X				Character level CNN-BLSTM-CRF
Li Z[70]			X	X					Attention layer, tensor-based representation layer
Lin[71]			X				X		
Liu[72]				X	X				
Luo[73]			X	X					
Manimaran[74]		X							
Minard[75]						X			
Narayanan[76]			X					X	ELMO and BERT
Nath[77]							X	X	Ontology-based (UMLS) and lexical vector augmentation
Nguyen[78]		X							

(*continued on next page*)

Table 2 (*continued*)

Author	Dictionary / Ontology	Rule-based	Neural Network (not specified)	LSTM-BILSTM	CRF	SVM	Word embeddings (any type)	BERT / TRANSFORMERS	Other
Patrick[79]		X			X	X			
Peng[80]								X	
Pradhan[81]	X	X			X	X			TF-IDF
Qin[82]				X					bag-of-words
Raj[83]			X						
Ramanan[84]		X							
Rea[85]		X							
Rink[86]									RDM (relation discovery model) based on LDA
Sahu[87]			X						
Shi[88]				X	X				
Si [89]				X	X			X	
Steinkamp[90]			X					X	GRUS, Transformer (BERT)
Suster[91]			X				X		
Tang[92]					X	X			
Tao[93]					X				
Tao[94]		X		X					ELMO-LSTM-CRF-HB
Tarcar[95]			X	X					
Trivedi[96]	X	X							
Wang[97]							X		Glove
Wang[98]	X	X				X			
Wei[99]				X	X	X			
Wu[100]			X		X				
Wu[101]			X				X		
Wu[102]				X			X		
Xie[103]			X				X		Attention Layer
Xu J[104]				X	X				
Xu Y[105]									Ensemble methods
Yang X 1[106]			X	X		X			
Yang X 2 [107]				X		X			Random forests and gradient boosting
Yang X 3[108]				X				X	BERT; ROBERTA, ALBERT, ELECTRA - RoBERTA MIMIC
Yehia[109]		X							
Zheng[110]	X	X							Hidden Markov Model
Zhu[111]						X			Markov Model, Logistic Regression, K-nearest neighbour
Roy[112]								X	
Michalopoulos[113]							X		
Khetan[114]								X	
Phan[115]								X	
Khandelwal[116]				X				X	Glove
Narayanan & Mannam [117]				X				X	
Mulyar[118]								X	
Li & Zhou[119]				X					GPT2, SegGANm, CTRL, CharRNN
Zhang[9]				X					
Tang & Yu[120]			X					X	
Moqurrab[121]			X	X					
Dave[122]					X				CLAMP Tool[123]
TOTAL	**9**	**19**	**21**	**27**	**28**	**16**	**10**	**20**	

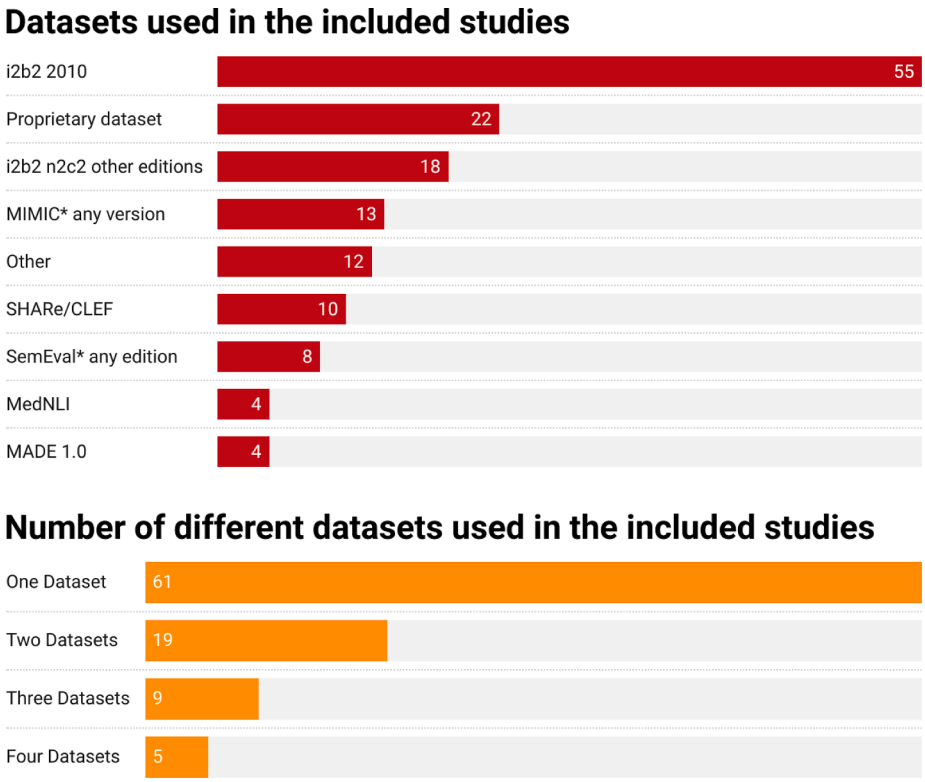


Fig. 4. Datasets used in the included studies.

included: demographics, history, vital signs, examination, fractures, smoking, sex, age, outcome, referral origin, encounters, cancer classification, department, and occurrence (see Table 1**)**.

3.5. Relation extraction

Among the 31 studies that performed RE tasks, the most common conceptualization followed the patterns presented in the 2010 i2b2 challenge [10]. This consisted of exploring relations between medical problems with treatments, tests, and relations between two medical problems. Other studies extracted different relational structures such as the 2009 i2b2 [124] and the 2018 n2c2 [13] challenges, which focused on medication relations. Shi et al.[88] extracted attributes related to disorders: negation, body, severity, change, assertion (uncertain, conditional), subject, and generic. Wang et al.[98] focused on drug-drug interaction and Yang et al.[106] identified the relations between medication that induced an adverse drug event. In contrast, Yehia et al. [109] used a rule-based method to extract the "*is-a*" relations between entity names and values, as well as entity values, attribute names and attribute values. Bejan et al.[35] extracted the relations such as "*A diagnoses B*", "*A causes B*", "*A is a location of B*", "*A is a B*", "*A treats B*" and "*A prevents B*". Dligach et al.[48] focused on extracting clinical entities and their relations to severity modifiers and body site. Tao et al. [93] extracted relations between prescriptions, and reasons for prescription; relations between positive diagnosis, the diagnosis being ruled out and concerns related to a given diagnostic, and drug-disease relations. Lastly, Xu et al.[104] extracted the semantic relations established in the SemEval 2015 [125] challenge, conceptualized as "disorder attributes" including the normalized disorder name, negation, subject, uncertainty, course, severity, conditional, generic mention, and body location.

3.6. Assertion and intensity

Assertion evaluation was described only in 16 studies (17%) [15,37,42,50,54,56,64,65,81,82,88,96,104,105,124,126], whereas in 78 (83%) this information was not available, and in 4 studies (4%) it was not sufficiently clear. The intensity of symptoms or severity of disease was reported in 11 studies (12%) [19,32,48,55,57,77,84,88,96,104,127] and it was mentioned in further 7 studies (7%); however, for the latter, it was unclear how it was measured. The remaining 76 studies (80%) did not report or measure intensity.

3.7. Evaluation metrics and benchmarking

Among the evaluation metrics reported, 48 studies (51%) reported Precision, Recall, and F1-Scores, whereas 88 studies (94%) reported only the F1 Scores. The remaining studies reported a combination of the above metrics and the Area Under the Curve (AUC). Three studies reported F1 scores and AUC. Two studies did not report precision but reported recall and F1. Three studies reported unaggregated metrics: true positives, true negatives, false positives, and false negatives [64,66,74]. Only one study did not report any accuracy metric [85]. The reported results, however, referred to different tasks and evaluation metrics. Thus, we neither summarize these results nor perform a *meta*-analysis.

Among the studies that used i2b2 2010 dataset (either for model development or benchmarking), 39 reported performances using the original test set. The highest reported F1 score (93.25%) was obtained by Moqurrab et al. [121] using a combination of CNN, Bi-LSTM, and CRF with non-complex embedding, followed by Si et al. [89] (F1 = 90.25%) using a Transformer model (BERT) [26]. With regards to the i2b2 2010 relation extraction task, the best F1 score (91.8%) was obtained by Roy et al. [112] using a BERT-based model. Table 4 contains all the studies reporting these benchmarks (average scores across the multiple

Table 3
Availability of source materials across studies:

Author	Mathematical formula	Pseudocode or system architecture provided	Source-code provided	Software available	Trained Model available
Abadeer[34]					X
Alsentzer[22]			X		X
Bejan[35]		X			
Ben Abacha[36]				X	
Bhatia[37]	X	X			
Chalapathy[38]	X		X		X
Chatzimina[39]		X			
Chen[40]					
Chodey[41]					
D' Souza[42]	X	X	X		
D'Avolio[43]				X	
Dai[44]	X	X			
de Bruijn[15]					
Dirkson[45]			X		
Divita[46]					
Divita[47]				X	
Dligach[48]		X		X	
Doan[49]					
Doğan[50]					
Ghiasvand[51]		X			
Gligic[52]					
Hao[53]		X			X
Hussain[54]					
Jagannatha[32]					
Jagannatha[55]	X				
JianG[56]					
Jiang & Denny[57]					
Jiang[58]	X	X			
Jonnalagadda	X				
Ju[60]		X			
Kang[61]					
Keretna[62]					
Kim[63]					
Kim[64]		X			
KraLjevic[65]		X	X	X	
Leaman[66]	X				
Lee[67]	X	X			
Li F 1 [68]	X	X		X	
Li F 2 [19]		X			
Li L [69]	X	X			
Li Z[70]	X	X			
Lin[71]		X			
Liu[72]	X	X			
Luo[73]	X				
Manimaran[74]		X			
Minard[75]					
Narayanan[76]		X	X		X
Nath[77]	X				
Nguyen[78]					
Patrick[79]					
Peng[80]	X	X			
Pradhan[81]			X		
Qin[82]	X	X	X		
Raj[83]					
Ramanan[84]					

(*continued on next page*)

Table 3 (*continued*)

Author	Mathematical formula	Pseudocode or system architecture provided	Source-code provided	Software available	Trained Model available
Rea[85]	X	X			
Rink[86]	X				
Sahu[87]	X	X			
Shi[88]					
Si [89]			X		
Steinkamp[90]			X		
Suster[91]	X				
Tang[92]					
Tao[93]					
Tao[94]		X			
Tarcar[95]					
Trivedi[96]	X				
Wang[97]	X				
Wang[98]		X			
Wei[99]					
Wu[100]	X				
Wu[101]		X			
Wu[102]	X	X			
Xie[103]					
Xu J[104]	X	X			
Xu Y[105]	X	X			
Yang X 1[106]		X			
Yang X 2 [107]			X		X
Yang X 3[108]		X			
Yehia[109]	X	X			
Zheng[110]		X			
Zhu[111]	X				
Roy[112]		X			
Michalopoulos[113]	X	X	X		X
Khetan[114]	X	X			
Phan[115]		X			
Khandelwal[116]		X			
Narayanan & Mannam[117]	X	X			
Mulyar[118]		X			
Li & Zhou[119]		X	X		
Zhang[9]		X		X	
Tang & Yu[120]	X	X			
Moqurrab[121]		X			
Dave[122]		X		X	
TOTAL	**32**	**48**	**14**	**8**	**7**

Table 4
I2b2-2010 concept extraction and relation classification benchmarks:

		Concept Extraction (NER)		
Author	Year	Precision	Recall	F1 Score
Abadeer[34]	2020	82	84	83
Nath[77]	2021			87
Yang X 2 [107]	2020			89.94
Divita[47]	2014	42.2	71.7	53.1
Lee[67]	2019			72.58
Ben Abacha[36]	2011	72.18	83.78	77.55
Chen[40]	2015			80
Patrick[79]	2011	84.88	78.92	81.79
Kang[61]	2012	83.3	81.2	82.2
Jiang[57]	2012	85.28	79.93	82.52
Wu[102]	2015	85.1	80.6	82.8
D'Avolio[43]	2011			83
Jonnalagadda[59]	2012	85.6	82	83.7
Chalapathy[38]	2016	84.36	83.41	83.88
Li L [69]	2020	83.83	85.41	84.61
Xu Y[105]	2012	86.53	83.19	84.82
de Bruijn*[15]	2011	86.88	83.64	85.23
Qin[82]	2018	84.24	86.53	85.37
Bhatia[37]	2019	85.4	85.8	85.5
JianG[56]	2011	88.28	82.98	85.55
Liu[72]	2016			85.81
Tang[92]	2013			85.82
Wu[100]	2017	85.33	86.56	85.94
Kim[63]	2019	88.6	83.5	86
Wu[101]	2018	87.37	85.09	86.21
Doğan[50]	2011	87.8	86.1	87
Jiang[58]	2019	88.03	86.91	87.44
Dai[44]	2020			87.8
Alsentzer[22]	2019	86	88	87.8
Hao[53]	2020			89.7
Si [89]	2019			90.25
Michalopoulos[113]	2021			87.7
Phan[115]	2022	86.92	88.55	87.73
Narayanan & Mannam[117]	2022			88.18
Mulyar[118]	2021			89.5
Li & Zhou[119]	2021			85.1
Tang & Yu[120]	2021			89.25
Zhang[9]	2021			88.13
Moqurrab[121]	2021	**94**	**94**	**93.57**
	Relation Extraction			
Author	Year	Precision	Recall	F1 Score
Xu Y[105]	2012	64	55.47	59.43
Raj[83]	2017	67.91	61.98	64.38
Doğan[50]	2011	64.6	74.6	69.2
D' Souza[42]	2014	66.7	72.9	69.6
Minard[75]	2011	62.8	80.3	70.5
Sahu[87]	2016	76.34	67.35	71.16
Patrick[79]	2011	73.07	67.51	72.63
Zhu[111]	2013	77.3	69.3	73.1
de Bruijn*[15]	2011	**77.38**	69.32	73.13
Peng[80]	2020			76
Li Z[70]	2019			74.3
Roy[112]	2021			**91.8**

*de Bruijn et al. reported best metric for the original i2b2/VA 2010 challenge.

subtasks). Over the years, we observed a modest improvement trend towards in the NER scores on i2b2/VA 2010 challenge and a steadier improvement in RE, compared to the initial best scores reported by De Bruijn et al.[15] in 2011 (Table 4 and Fig. 5). It is important to note that the exact performance is hard to compare due to the potential differences in the pre-processing steps, specific portions of the test data used, and other methodological discrepancies.".

3.8. Clinical task description

Considering the clinical or information task description, 33 studies (35%) did not describe any specific task or problem to be solved, 27 studies (29%) only described the potential use case that the system could be addressing, 20 studies (21%) provided motivational examples based on prior literature, and only 14 studies (15%) addressed one clinical or information task to solve. Lastly, 4 studies (4%) provided both hypothetical and literature examples. Specific tasks (either defined, potential or following on prior literature) included: adverse event detection and pharmacovigilance (21 studies), increasing clinician and patient understanding and helping patient management (21 studies), decision support (17 studies), drug efficacy and repurposing drug research (8 studies), coding and automating EHR tasks (15 studies), quality improvement (11 studies), public health and epidemiology research (6 studies), genotype and phenotype research (7 studies) and cohort identification and trial recruitment (3 studies) See Table 5.

3.9. Use in real-world settings

Only 3 studies (5%) showed any evidence of deployment in a real-world setting. Rea et al.[85] specify that the platform was deployed to "*(1) receive source EHR data in several formats, (2) generate structured data from EHR narrative text, and (3) normalize the EHR data using common detailed clinical models […], which were (4) accessed by a phenotyping service using normalized data specifications*". Divita et al.[46] mention an initial deployment for evaluation purposes, as "the model was folded into the NLP tool, scaled-up and run on a larger set of 964,105 records randomly chosen from the larger OEF/OEF cohort". Lastly, Kraljevic et al. [65] reported the development of MedCAT, a Multi-domain clinical annotation and extraction tool that was further validated in three London hospitals.

4. Discussion

In this review, we explored current literature on NLP systems that perform multi-entity NER and RE. We observed that most recent developments in this area leverage a limited number of datasets, with a scarce external validation (either on additional data or deployed in the real world). Most research is restricted to a limited number of annotations, shared tasks, and datasets. While about a quarter of the reviewed papers used proprietary datasets, these are not publicly available to the broader community, such that their value is limited. This may lead to generalization issues and the performance may be affected when applied in a different setting, dataset, or in the real world. Unfortunately, application in the clinical settings was almost non-existent as ready-to-use tools outside the experimental setting are scarce. Most of the studies focusing on the IE tasks used ML models and were mostly trained using a few datasets released for specific shared tasks (contests, challenges, competitions, etc) and often reported insufficiently the tools or methods utilized or lacked sufficient validation.

Although previous reviews have explored various aspects of NLP applications in medicine [17,18,128,129], in this review, we focused specifically on the combination of the NER and RE tasks and multi-entity systems and clinical semiology extraction. In contrast to previous reviews, our work produces a much-needed update, reflecting the proliferation of transfer learning and Transformer-based models that occurred in the last few years. Transformers have come to dominate the field taking the lead in several clinical benchmarks. Moreover, our review contains important findings for NLP researchers and ML practitioners. We highlight the current state of the art performance in concept extraction and relation extraction, and also identify the studies providing either code or pre-trained models ready for implementation. A strength of this review is in shedding light both on the ways these tools were developed from a technical point of view and exploring their potential for clinical implementation. As there is no validated tool to appraise the quality of the developed NLP models from an evidence-based medicine point of view, our approach may pave the way for establishing the criteria for this.

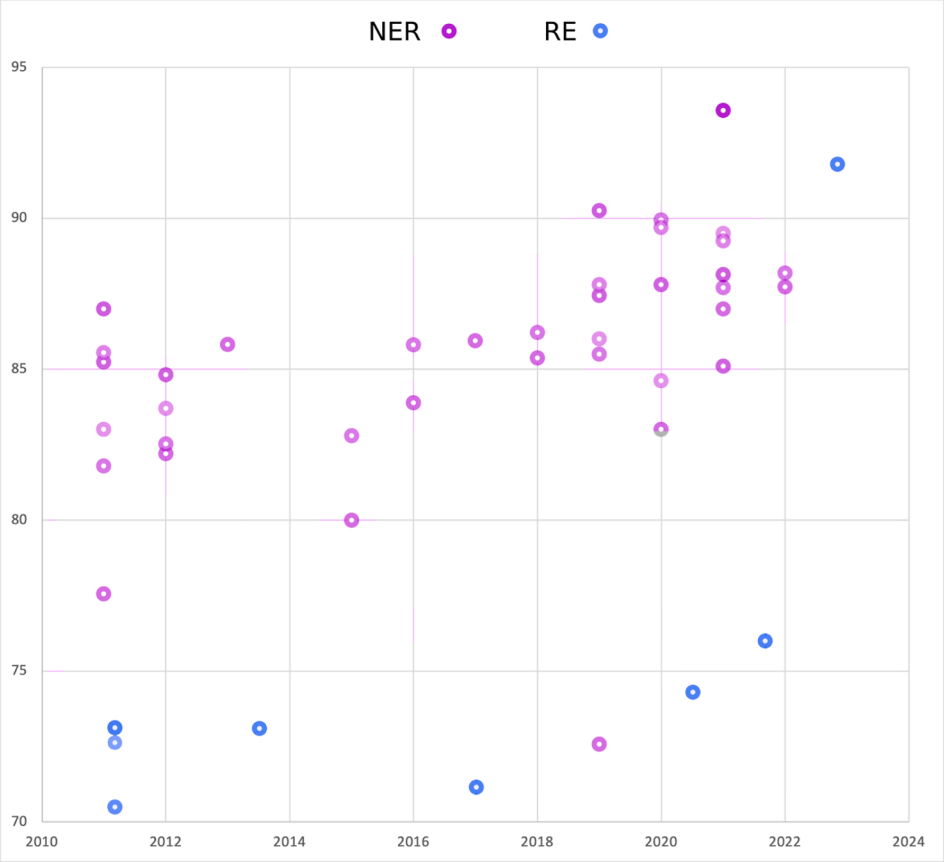


Fig. 5. F1 Scores by year for NER and RE in i2b2/VA 2010 dataset.

4.1. Limitations

The results of this review need to be interpreted considering some limitations. We have only analyzed systems developed in the English language, although only 10 publications in other languages were detected in our search (see Fig. 1). However, it is unlikely that we overlooked relevant literature, as the majority of tools detected were developed for English (e.g., MetaMap and cTAKES) as well as the pre-training of certain models such as Transformers and also the freely available datasets (such as i2b2/n2c2 or MIMIC) were all in the English language. Given this, we do not expect the works in other languages to invalidate our findings. While we did not calculate inter-observer agreement when extracting results, we minimize biases due to data extraction by performing a calibration with a 10% sub-sample and reconciling cases of discrepancy. Another potential limitation is that we analyzed systems that perform multi-entity information and relation extraction; that is systems that cover a broad range of clinical specialties, subjects and areas of medicine. Previous reviews [17,18] showed that single entity extraction may have dominated the space previously, especially considering older approaches such as regular expressions. It is possible then that most of the clinical applications of NLP are still limited to this single entity extraction, for instance for identification purposes. However, we believe that multi-entity systems, not linked to a particular condition or medical specialization, have a stronger potential for translation; thus, focusing on such systems. Another limitation is that given the current overwhelming progress of the NLP field, this systematic review could not include the most recent developments and papers. However, our review has captured the emergence of the transformer era in NLP, which has been the dominant NLP technology even in the most recent versions of Generative Pretrained Transformers (e.g., GPT-4 [130]). This review can also be useful as a departing point for future reviews, as we have shared our search strategy and several resources that may streamline the search task for future reviewers.

4.2. Datasets

Analyzing the sources of datasets used in many of the reviewed works, we noted a certain overlap among multiple datasets coming from the same source. For instance, the most widely used i2b2/VA 2010 shared task dataset [15] has approximately a third of its content coming from the MIMIC dataset.[12]. Likewise Share/CLEF [30] and some of the SemEval shared tasks [31] can be traced back to MIMIC. Although we cannot verify if the same parts of the dataset were re-used, this may raise potential validation issues, as the same data might have been used for both training and evaluation purposes. Moreover, although the above-mentioned shared tasks used to some extent different clinical sources, the MIMIC dataset refers exclusively to the ICU setting, and does not include other clinical scenarios, such as regular medical wards or primary care, potentially limiting the generalizability of the models developed with this dataset.

4.3. Transformer-era and newer developments

In this review, we noticed the strong uptake of pre-trained BERT-derived models, some of them tailored to the biomedical domain, such as those trained on medical literature, BLUEBERT [131] and PubMedBERT [132] and those specifically developed with clinical text, ClinicalBERT [22], and EHRBERT [19] that we analyzed in this review. However, few of these trained BERT models are openly available or ready to be used, due to various issues, mainly having used proprietary

Table 5
Clinical or information task description in included studies.

Author	HOW WAS THE CLINICAL OR INFORMATION TASK DEFINED (IF AT ALL)?	Which task(s) were defined, described or cited? CLINICIAN / PATIENT UNDERSTANDING AND MANAGEMENT	ADVERSE EVENTS DETECTION / PHARMACOVIGILANCE	HEALTHCARE QUALITY IMPROVEMENT	DECISION SUPPORT	DRUG EFFICACY / REPURPOSING DRUGS	COHORT IDENTIFICATION / TRIAL RECRUITMENT	GENOTYPE / PHENOTYPE	PUBLIC HEALTH / EPIDEMIOLOGY	CODING / AUTOMATING EHR TASKS	OTHER
Abadeer[34]	Follows on prior work										
Alsentzer[22]											
Bejan[35]	Follows on prior work	X	X	X							
Ben Abacha [36]											
Bhatia[37]	Potential use case and follows on prior work	X	X		X	X					
Chalapathy [38]	Potential use case	X		X							
Chatzimina [39]											
Chen[40]	Follows on prior work				X	X	X	X			
Chodey[41]											
D' Souza[42]											
D'Avolio[43]	Follows on prior work			X		X		X			"EBM"
Dai[44]											
de Bruijn[15]											
Dirkson[45]	Defined explicitly		X								
Divita[46]	Potential use case		X	X			X	X	X		
Divita[47]	Defined explicitly			X			X	X			
Dligach[48]	Follows on prior work									X	"Discovering tumour body sites for template filling"
Doan[49]											
Doğan[50]	Potential use case		X			X					
Ghiasvand [51]											
Gligic[52]	Follows on prior work		X	X							"Studying disease"
Hao[53]	Potential use case									X	
Hussain[54]	Defined explicitly									X	save human experts time and burden
Jagannatha [32]	Defined explicitly		X								Drug safety and post-marketing pharmaco-surveillance.
Jagannatha [55]	Potential use case		X			X					
JianG[56]											
Jiang[57]	Follows on prior work										"Body site identification for tumours"
Jiang[58]											
Jonnalagadda [59]	Potential use case	X			X						
Ju[60]	Follows on prior work	X									
Kang[61]	Follows on prior work		X		X					X	
Keretna[62]	Follows on prior work		X		X						
Kim[63]	Defined explicitly			X	X					X	

(*continued on next page*)

Table 5 (*continued*)

Author	HOW WAS THE CLINICAL OR INFORMATION TASK DEFINED (IF AT ALL)?	Which task(s) were defined, described or cited? CLINICIAN / PATIENT UNDERSTANDING AND MANAGEMENT	ADVERSE EVENTS DETECTION / PHARMACOVIGILANCE	HEALTHCARE QUALITY IMPROVEMENT	DECISION SUPPORT	DRUG EFFICACY / REPURPOSING DRUGS	COHORT IDENTIFICATION / TRIAL RECRUITMENT	GENOTYPE / PHENOTYPE	PUBLIC HEALTH / EPIDEMIOLOGY	CODING / AUTOMATING EHR TASKS	OTHER
Kim[64]	Potential use case and follows on prior work	X	X		X	X					
KraLjevic[65]	Potential use case									X	There is a need for a platform to accurately extract information from freeform health text in a scalable manner that is agnostic to underlying health informatics architectures
Leaman[66]	Follows on prior work	X	X		X		X		X		"Clinical associations"
Lee[67]											
Li F 1 [68]	Defined explicitly		X								
Li F 2 [19]	Potential use case	X		X	X					X	
Li L [69]	Follows on prior work	X			X						Knowledge graph
Li Z[70]											
Lin[71]	Potential use case								X		
Liu[72]											
Luo[73]											
Manimaran [74]	Follows on prior work		X								
Minard[75]	Potential use case										"Question Answering Systems"
Narayanan [76]	Defined explicitly		X		X						
Nath[77]	Potential use case	X			X					X	
Nguyen[78]	Defined explicitly									X	
Patrick[79]	Potential use case	X									
Peng[80]											
Pradhan[81]	Potential use case	X									
Qin[82]	Potential use case	X									
Raj[83]											
Ramanan[84]	Follows on prior work		X			X			X		
Rea[85]	Defined explicitly							X			
Rink[86]	Potential use case	X								X	
Sahu[87]	Potential use case					X					"Medical knowledge"
Shi[88]											
Si [89]											
Steinkamp [90]	Follows on prior work			X					X	X	Automatic EHR population
Suster[91]	Potential use case and follows on prior work									X	Automating healthcare systems and research
Tang[92]	Potential use case										Unspecified
Tao[93]	Potential use case			X							
Tao[94]	Potential use case	X			X					X	
Tarcar[95]											
Trivedi[96]	Follows on prior work										
Wang[97]											

(*continued on next page*)

Table 5 (*continued*)

Author	HOW WAS THE CLINICAL OR INFORMATION TASK DEFINED (IF AT ALL)?	Which task(s) were defined, described or cited? CLINICIAN / PATIENT UNDERSTANDING AND MANAGEMENT	ADVERSE EVENTS DETECTION / PHARMACOVIGILANCE	HEALTHCARE QUALITY IMPROVEMENT	DECISION SUPPORT	DRUG EFFICACY / REPURPOSING DRUGS	COHORT IDENTIFICATION / TRIAL RECRUITMENT	GENOTYPE / PHENOTYPE	PUBLIC HEALTH / EPIDEMIOLOGY	CODING / AUTOMATING EHR TASKS	OTHER
Wang[98]											
Wei[99]	Defined explicitly		X								
Wu[100]											
Wu[101]											
Wu[102]	Potential use case										Clinical and translational research
Xie[103]	Defined explicitly									X	
Xu J[104]	Potential use case				X						
Xu Y[105]	Potential use case	X									
Yang X 1 [106]											
Yang X 2 [107]	Follows on prior work										
Yang X 3 [108]	Defined explicitly		X								
Yehia[109]	Defined explicitly	X									
Zheng[110]											
Zhu[111]	Potential use case										Collect real-life, real-time, and large-sample-based knowledge from patients.
TOTAL	**Defined explicitly = 13, potential use case = 23, follows on prior work = 17, undefined = 29**	**17**	**18**	**10**	**14**	**8**	**3**	**5**	**5**	**14**	

datasets for their development [19,133]. This restricts their use and further validation or implementation and raises questions regarding the utility of their findings, reproducibility of the work, or its implementation potential. Although newer developments such as GPT-3 [134] have been popular, they seem to fail when evaluated in medical use cases [135]. Developers willing to implement or develop NLP clinical systems can reflect on the availability of these tools to avoid costly training and, thanks to novel transfer learning techniques and easy-to-use implementations [136].

4.4. Reproducibility and generalizability of clinical NER research

One of the problems detected in this review is the lack of metrics and appraisal tools to assess NLP systems development and applicability in the clinical context. Although the use of benchmarks such as GLUE [137] points towards an improved ability to compare across methods and strengthen validation and generalizability, it is uncertain how this should be interpreted from a clinical standpoint, where a different set of standards governs the evidence required for health interventions. Additionally, newer, larger ML models have demonstrated a non-negligible risk of training data to leak from the trained models, either accidentally or using reverse engineering [138,139]. Developing models using debiased datasets may help avoid artifacts or establish spurious relations, to prevent algorithmic bias, a common issue in medical AI applications [140].

4.5. The importance of computational shared tasks

Computational shared tasks, usually presented in form of contests, challenge or competitions, are a common methodology in the NLP field, with general recognition that shared tasks may lead to improvements in the state-of-the-art performance [141]. However, there is still a discussion about the ways to improve them [142,143]. When translating their findings into health practice, there is a potential tension between their usefulness as a means to advance the computational field and their fit to clinical reality, i.e., whether the generated knowledge can be transposed into clinical medicine. Considering this broad question, we highlight a clear need for interdisciplinary discourse and methods development, especially when those developments have the potential to affect something as delicate as patient's care.

4.6. Clinical translation and clinician interaction

Another issue that seems not to be sufficiently explored is the reasons for these models not being translated into clinical practice more frequently. Previous research has shown that clinicians are eager to use text automation [144] in practice. Yet there are several concerns around their use, including medico-legal issues, trust in system's extraction and data processing capabilities, and their compatibility with the existing workflows [145]. Although not a key focus of this review, an important and unresolved problem from the user perspective pertains to the meaningful implementation of clinical NLP tasks. We identified several studies that explored this parameter [146-149]. If the interaction with NLP systems is not designed properly [7], clinicians may not use them, notwithstanding their strong technical performance, as shown by previous cases in other fields [150].

5. Conclusion

In our review, we demonstrated that the information extraction field of NLP applications in medicine has developed steadily over the last 10 years. However, there remain several uncertainties surrounding its application to the clinical field, and so far, its translation into practice has been limited. More research is required to validate these systems in real-world scenarios and explore the ways clinicians can take advantage of these systems, to improve our future healthcare.

Summary points:

- Clinical information extraction tasks have developed steadily the last decade, with machine learning methods and recent transformer models gaining more power.
- There are few datasets and few annotations rules that have been used openly in most of the papers included.
- Very few studies provide trained models, datasets or ready to use tools. Even fewer report their use in clinical settings.
- There is a strong need for research translation and further validation of these models and tools in practice.

Funding

David Fraile Navarro is supported by an International Macquarie University Research Excellence Scholarship "iMQRES".

CRediT authorship contribution statement

David Fraile Navarro: Conceptualization, Investigation, Data curation, Formal analysis, Methodology, Project administration, Writing – original draft, Writing – review & editing. **Kiran Ijaz:** Investigation, Data curation, Formal analysis, Methodology, Writing – review & editing. **Dana Rezazadegan:** Conceptualization, Data curation, Formal analysis, Methodology, Investigation, Writing – review & editing. **Hania Rahimi-Ardabili:** Investigation, Data curation, Formal analysis, Methodology, Writing – review & editing. **Mark Dras:** Conceptualization, Investigation, Writing – review & editing. **Enrico Coiera:** Conceptualization, Investigation, Writing – review & editing. **Shlomo Berkovsky:** Supervision, Methodology, Conceptualization, Writing – review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ijmedinf.2023.105122.

References

[1] T. Heart, O. Ben-Assuli, I. Shabtai, A review of PHR, EMR and EHR integration: a more personalized healthcare and public health policy, Heal. Policy Technol. 6 (2017) 20–25, https://doi.org/10.1016/j.hlpt.2016.08.002.
[2] J. Marc Overhage, D. McCallie, Physician time spent using the electronic health record during outpatient encounters a descriptive study, Ann. Intern. Med. 172 (2020) 169–174, https://doi.org/10.7326/M18-3684.
[3] C. Dymek, B. Kim, G.B. Melton, T.H. Payne, H. Singh, C.-J. Hsiao, Building the evidence-base to reduce electronic health record–related clinician burden, J. Am. Med. Informatics Assoc. 00 (2020) 1–5, https://doi.org/10.1093/jamia/ocaa238.
[4] E.J. Topol, High-performance medicine: the convergence of human and artificial intelligence, Nat. Med. 25 (2019) 44–56, https://doi.org/10.1038/s41591-018-0300-7.
[5] A. Hosny, C. Parmar, J. Quackenbush, L.H. Schwartz, H.J.W.L. Aerts, Artificial intelligence in radiology, Nat. Rev. Cancer. 18 (2018) 500–510, https://doi.org/10.1038/s41568-018-0016-5.
[6] X. Du-Harpur, F.M. Watt, N.M. Luscombe, M.D. Lynch, What is AI? Applications of artificial intelligence to dermatology, Br. J. Dermatol. 183 (2020) 423–430, https://doi.org/10.1111/bjd.18880.
[7] D. Fraile Navarro, A.B. Kocaballi, M. Dras, S. Berkovsky, Understanding General Practitioners' attitudes towards natural language and text automation in clinical practice, Trans. Comput. Hum. Interact. (n.d.).
[8] D. Nouvel, M. Ehrmann, S. Rosset, Named entities for computational linguistics, Wiley Online, Library (2016).
[9] Y. Zhang, Y. Zhang, P. Qi, C.D. Manning, C.P. Langlotz, Biomedical and clinical English model packages for the Stanza Python NLP library, J. Am. Med. Informatics Assoc. 28 (2021) 1892–1899.

[10] 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, https://doi.org/10.1136/amiajnl-2011-000203.
[11] 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, https://doi.org/10.1016/j.jbi.2013.06.004.
[12] A.E.W. Johnson, T.J. Pollard, L. Shen, L.W.H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R.G. Mark, MIMIC-III, a freely accessible critical care database, Sci. Data. 3 (2016), https://doi.org/10.1038/sdata.2016.35.
[13] S. Henry, K. Buchan, M. Filannino, A. Stubbs, O. Uzuner, n2c2 shared task on adverse drug events and medication extraction in electronic health records, J. Am. Med. Informatics Assoc. 27 (2020) 3–12, https://doi.org/10.1093/jamia/ocz166.
[14] S. Henry, Y. Wang, F. Shen, O. Uzuner, The, National Natural language processing (NLP) Clinical Challenges (n2c2)/Open Health NLP (OHNLP) shared task on clinical concept normalization for clinical records, J. Am. Med. Informatics Assoc. 27 (2020) 1529–1537.
[15] B. de Bruijn, C. Cherry, S. Kiritchenko, J. Martin, X. Zhu, Machine-learned solutions for three stages of clinical information extraction: the state of the art at i2b2 2010, J. Am. Med. Informatics Assoc. 18 (2011) 557–562, https://doi.org/10.1136/amiajnl-2011-000150.
[16] M.G. Kersloot, F.J. P van Putten, A. Abu-Hanna, R. Cornet, D.L. Arts, Natural language processing algorithms for mapping clinical text fragments onto ontology concepts: a systematic review and recommendations for future studies, (n.d.). https://doi.org/10.1186/s13326-020-00231-z.
[17] Y. Wang, L. Wang, M. Rastegar-Mojarad, S. Moon, F. Shen, N. Afzal, S. Liu, Y. Zeng, S. Mehrabi, S. Sohn, Clinical information extraction applications: a literature review, J. Biomed. Inform. 77 (2018) 34–49.
[18] S. Wu, K. Roberts, S. Datta, J. Du, Z. Ji, Y. Si, S. Soni, Q. Wang, Q. Wei, Y. Xiang, Deep learning in clinical natural language processing: a methodical review, J. Am. Med. Informatics Assoc. 27 (2020) 457–470.
[19] F. Li, Y. Jin, W. Liu, B.P.S. Rawat, P. Cai, H. Yu, Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study, J. Med. Internet Res. 21 (2019), https://doi.org/10.2196/14830.
[20] D. Moher, A. Liberati, J. Tetzlaff, D.G. Altman, Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement, Ann. Intern. Med. 151 (2009) 264–269.
[21] M. Ouzzani, H. Hammady, Z. Fedorowicz, A. Elmagarmid, Rayyan—a web and mobile app for systematic reviews, Syst. Rev. 5 (2016) 1–10.
[22] E. Alsentzer, J. Murphy, W. Boag, W.-H. Weng, D. Jindi, T. Naumann, M. McDermott, Publicly Available Clinical, in: Proc. 2nd Clin. Nat. Lang. Process. Work., 2019: pp. 72–78. https://doi.org/10.18653/v1/w19-1909.
[23] H.M. Wallach, Conditional random fields: an introduction, Tech. Reports. (2004) 22.
[24] F.A. Gers, J. Schmidhuber, F. Cummins, Learning to forget, Continual prediction with LSTM (1999).
[25] A. Graves, J. Schmidhuber, Framewise phoneme classification with bidirectional LSTM and other neural network architectures, Neural Networks 18 (2005) 602–610.
[26] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, (2018). http://arxiv.org/abs/1810.04805 (accessed October 22, 2019).
[27] W.S. Noble, What is a support vector machine? Nat. Biotechnol. 24 (2006) 1565–1567.
[28] M.E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, L. Zettlemoyer, Deep Contextualized Word Representations, in: Proc. 2018 Conf. North {A} merican Chapter Assoc. Comput. Linguist. Hum. Lang. Technol. Vol. 1 (Long Pap., Association for Computational Linguistics, New Orleans, Louisiana, 2018: pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202.
[29] J. Pennington, R. Socher, C.D. Manning, Glove: Global vectors for word representation, in: Proc. 2014 Conf. Empir. Methods Nat. Lang. Process., 2014: pp. 1532–1543.
[30] H. Suominen, S. Salanterä, S. Velupillai, W.W. Chapman, G. Savova, N. Elhadad, S. Pradhan, B.R. South, D.L. Mowery, G.J.F. Jones, J. Leveling, L. Kelly, L. Goeuriot, D. Martinez, G. Zuccon, Overview of the ShARe/CLEF eHealth Evaluation Lab 2013 BT - information Access Evaluation. Multilinguality, Multimodality, and Visualization, in: P. Forner, H. Müller, R. Paredes, P. Rosso, B. Stein (Eds.), Springer, Berlin Heidelberg, Berlin, Heidelberg, 2013, pp. 212–231.
[31] SemEval - Wikipedia, Wikipedia. (n.d.).
[32] A. Jagannatha, F. Liu, W. Liu, H. Yu, Overview of the first natural language processing challenge for extracting medication, indication, and adverse drug events from electronic health record notes (MADE 1.0), Drug Saf. 42 (2019) 99–111, https://doi.org/10.1007/s40264-018-0762-z.
[33] C. Shivade, Mednli-a natural language inference dataset for the clinical domain, Publ. Online. (2017).
[34] M. Abadeer, Assessment of DistilBERT performance on Named Entity Recognition task for the detection of Protected Health Information and medical concepts, Proc. Ofthe 3rd Clin. Nat. Lang. Process. Work. (2020) 158–167, https://doi.org/10.18653/v1/2020.clinicalnlp-1.18.
[35] C.A. Bejan, W.-Q. Wei, J.C. Denny, Assessing the role of a medication-indication resource in the treatment relation extraction from clinical text, J. Am. Med. Informatics Assoc. 22 (2015) e162–e176, https://doi.org/10.1136/amiajnl-2014-002954.
[36] A. Ben Abacha, P. Zweigenbaum, Medical entity recognition: a comparison of semantic and statistical methods, 2011 Work, Biomed. Nat. Lang. Process. (2011) 56–64.
[37] P. Bhatia, B. Celikkaya, M. Khalilia, Joint entity extraction and assertion detection for clinical text, Association for Computational Linguistics, Florence, Italy, 2019 https://www.aclweb.org/anthology/P19-1091/.
[38] R. Chalapathy, E.Z. Borzeshi, M. Piccardi, Bidirectional LSTM-CRF for clinical concept extraction, Proc. Clin. Nat. Lang. Process. Work. (2016) 7–12. http://arxiv.org/abs/1611.08373.
[39] M.E. Chatzimina, C. Grouin, P. Zweigenbaum, Use of unsupervised word classes for entity recognition: application to the detection of disorders in clinical reports, European Language Resources Association (ELRA), Reykjavik, Iceland, 2014.
[40] Y. Chen, T.A. Lasko, Q. Mei, J.C. Denny, H. Xu, A study of active learning methods for named entity recognition in clinical text, J. Biomed. Inform. 58 (2015) 11–18, https://doi.org/10.1016/j.jbi.2015.09.010.
[41] K.P. Chodey, G. Hu, Clinical text analysis using machine learning methods, in: 2016 IEEE/ACIS 15th Int. Conf. Comput. Inf. Sci. ICIS 2016 - Proc., 2016. https://doi.org/10.1109/ICIS.2016.7550908.
[42] J. D'Souza, V. Ng, Ensemble-based medical relation classification, Dublin City University and Association for Computational Linguistics, Dublin, Ireland, 2014.
[43] L.W. D'Avolio, T.M. Nguyen, S. Goryachev, L.D. Fiore, Automated concept-level information extraction to reduce the need for custom software and rules development, J. Am. Med. Inform. Assoc. 18 (2011) 607–613, https://doi.org/10.1136/amiajnl-2011-000183.
[44] X. Dai, H. Adel, An Analysis of Simple Data Augmentation for Named Entity Recognition, Proc. 28th Int. Conf. Comput. Linguist. 2010 (2021) 3861–3867. https://doi.org/10.18653/v1/2020.coling-main.343.
[45] A. Dirkson, S. Verberne, W. Kraaij, FuzzyBIO, a proposal for fuzzy representation of discontinuous entities, Proc. 12th Int, Work. Heal. Text Min. Inf. Anal. (2021) 77–82. https://www.aclweb.org/anthology/2021.louhi-1.9.
[46] G. Divita, G. Luo, L.-T.-T. Tran, T.E. Workman, A.V. Gundlapalli, M.H. Samore, General symptom extraction from VA electronic medical notes, Stud. Health Technol. Inform. 245 (2017) 356–360.
[47] G. Divita, Q.T. Zeng, A.V. Gundlapalli, S. Duvall, J. Nebeker, M.H. Samore, Sophia: a Expedient UMLS concept extraction annotator, AMIA Annu. Symp. Proc. (2014) 467–476.
[48] D. Dligach, S. Bethard, L. Becker, T. Miller, G.K. Savova, Discovering body site and severity modifiers in clinical texts, J. Am. Med. Informatics Assoc. 21 (2014) 448–454, https://doi.org/10.1136/amiajnl-2013-001766.
[49] S. Doan, L. Bastarache, S. Klimkowski, J.C. Denny, H. Xu, Integrating existing natural language processing tools for medication extraction from discharge summaries, J. Am. Med. Inform. Assoc. 17 (2010) 528–531, https://doi.org/10.1136/jamia.2010.003855.
[50] R. Islamaj Doğan, A. Névéol, Z. Lu, A context-blocks model for identifying clinical relationships in patient records, BMC Bioinformatics. 12 (2011), https://doi.org/10.1186/1471-2105-12-S3-S3.
[51] O. Ghiasvand, R. Kate, UWM: Disorder Mention Extraction from Clinical Text Using CRFs and Normalization Using Learned Edit Distance Patterns, Proc. 8th Int. Work. Semant. Eval. (SemEval 2014). (2015) 828–832. https://doi.org/10.3115/v1/s14-2147.
[52] L. Gligic, A. Kormilitzin, P. Goldberg, A. Nevado-Holgado, Named entity recognition in electronic health records using transfer learning bootstrapped Neural Networks, Neural Networks 121 (2020) 132–139, https://doi.org/10.1016/j.neunet.2019.08.032.
[53] B. Hao, H. Zhu, I. Paschalidis, Enhancing Clinical BERT Embedding using a Biomedical Knowledge Base, Proc. 28th Int. Conf. Comput. Linguist. (2021) 657–661. https://doi.org/10.18653/v1/2020.coling-main.57.
[54] M. Hussain, D.J. Choi, S. Lee, Semantic based Clinical Notes Mining for Factual Information Extraction, Int. Conf. Inf. Netw. (2020-) 46–48, https://doi.org/10.1109/ICOIN48656.2020.9016559.
[55] A.N. Jagannatha, H. Yu, Structured prediction models for RNN based sequence labeling in clinical text, in: EMNLP 2016 - Conf. Empir. Methods Nat. Lang. Process. Proc., 2016: pp. 856–865.
[56] M. Jiang, Y. Chen, M. Liu, S.T. Rosenbloom, S. Mani, J.C. Denny, H. Xu, A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries, J. Am. Med. Informatics Assoc. 18 (2011) 601–606, https://doi.org/10.1136/amiajnl-2011-000163.
[57] M. Jiang, J.C. Denny, B. Tang, H. Cao, H. Xu, Extracting semantic lexicons from discharge summaries using machine learning and the C-Value method, AMIA Annu. Symp. Proceedings. AMIA Symp. (2012) 409–416.
[58] M. Jiang, T. Sanger, X. Liu, Combining contextualized embeddings and prior knowledge for clinical named entity recognition: Evaluation study, J. Med. Internet Res. 21 (2019), https://doi.org/10.2196/14850.
[59] J. S, C. T, W. S, G. G, Enhancing clinical concept extraction with distributional semantics, J. Biomed. Inform. 45 (2012) 129–140. internal-pdf://222.136.74.220/nihms337449.pdf.
[60] M. Ju, N.T.H. Nguyen, M. Miwa, S. Ananiadou, An ensemble of neural models for nested adverse drug events and medication extraction with subwords, J. Am. Med. Inform. Assoc. 27 (2020) 22–30, https://doi.org/10.1093/jamia/ocz075.
[61] N. Kang, Z. Afzal, B. Singh, E.M. van Mulligen, J.A. Kors, Using an ensemble system to improve concept extraction from clinical records, J. Biomed. Inform. 45 (2012) 423–428, https://doi.org/10.1016/j.jbi.2011.12.009.
[62] K. S, L. CP, C. D, S. KB, Enhancing medical named entity recognition with an extended segment representation technique, Comput. Methods Programs Biomed.

119 (2015) 88–100. internal-pdf://243.123.141.219/1-s2.0-S0169260715000425-main.pdf.
[63] Y. Kim, S.M. Meystre, A study of medical problem extraction for better disease management, Stud. Health Technol. Inform. 264 (2019) 193–197, https://doi.org/10.3233/SHTI190210.
[64] Y. Kim, E. Riloff, Stacked generalization for medical concept extraction from clinical notes, Proc. B}io{NLP 15 (2015).
[65] Z. Kraljevic, T. Searle, A. Shek, L. Roguski, K. Noor, D. Bean, A. Mascio, L. Zhu, A. A. Folarin, A. Roberts, R. Bendayan, M.P. Richardson, R. Stewart, A.D. Shah, W. K. Wong, Z. Ibrahim, J.T. Teo, R.J.B. Dobson, Multi-domain clinical natural language processing with MedCAT: the medical concept annotation toolkit, Artif. Intell. Med. 117 (2021). internal-pdf://0719885385/2010.01165.pdf.
[66] R. Leaman, R. Khare, Z. Lu, Challenges in clinical natural language processing for automated disorder normalization, J. Biomed. Inform. 57 (2015) 28–37, https://doi.org/10.1016/j.jbi.2015.07.010.
[67] W. Lee, J. Choi, Precursor-induced conditional random fields: Connecting separate entities by induction for improved clinical named entity recognition, BMC Med. Inform. Decis. Mak. 19 (2019), https://doi.org/10.1186/s12911-019-0865-1.
[68] F. Li, W. Liu, H. Yu, Extraction of information related to adverse drug events from electronic health record notes: Design of an end-to-end model based on deep learning, JMIR Med. Informatics. 6 (2018). internal-pdf://174.138.99.148/PDF (1).pdf.
[69] L. Li, W. Xu, H. Yu, Character-level neural network model based on Nadam optimization and its application in clinical concept extraction, Neurocomputing 414 (2020) 182–190. internal-pdf://81.95.0.124/1-s2.0-S0925231220311346-main.pdf.
[70] Z. Li, J. Yang, X. Gou, X. Qi, Recurrent neural networks with segment attention and entity description for relation extraction from clinical texts, Artif. Intell. Med. 97 (2019) 9–18. internal-pdf://178.123.81.126/10.1016@j.artmed.2019.04.003.pdf.
[71] C. Lin, C.-J. Hsu, Y.-S. Lou, S.-J. Yeh, C.-C. Lee, S.-L. Su, H.-C. Chen, Artificial intelligence learning semantics via external resources for classifying diagnosis codes in discharge notes, J. Med. Internet Res. 19 (2017) e380.
[72] Z. Liu, M. Yang, X. Wang, Q. Chen, B. Tang, Z. Wang, H. Xu, Entity recognition from clinical texts via recurrent neural network, BMC Med. Inform. Decis. Mak. 17 (2017), https://doi.org/10.1186/s12911-017-0468-7.
[73] S. Shah, X. Luo, Extracting modifiable risk factors from narrative preventive healthcare guidelines for EHR integration, in: Proc. - 2017 IEEE 17th Int. Conf. Bioinforma. Bioeng. BIBE 2017, 2017: pp. 514–519. https://doi.org/10.1109/BIBE.2017.000-2.
[74] J. Manimaran, T. Velmurugan, Evaluation of named entity recognition algorithms using clinical text data, Int. J. Eng. Technol. 7 (2018) 295–302, https://doi.org/10.14419/ijet.v7i4.5.20093.
[75] A.-L. Minard, A.-L. Ligozat, B. Grau, Multi-class SVM for Relation Extraction from Clinical Reports, Association for Computational Linguistics, Hissar, Bulgaria, 2011.
[76] S. Narayanan, K. Mannam, S.P. Rajan, P.V. Rangan, Evaluation of Transfer Learning for Adverse Drug Event ({ADE}) and Medication Entity Extraction, Association for Computational Linguistics, Online, n.d. internal-pdf://80.126.215.171/2020.clinicalnlp-1.6.pdf.
[77] N. Nath, S.H. Lee, M. McDonnell, I. Lee, The quest for better clinical word vectors: Ontology based and lexical vector augmentation versus clinical contextual embeddings, Comput. Biol. Med. 134 (2021). internal-pdf://107.104.7.154/1-s2.0-S0010482521002274-main.pdf.
[78] A.N. Nguyen, D. Truran, M. Kemp, B. Koopman, D. Conlan, J. O'Dwyer, M. Zhang, S. Karimi, H. Hassanzadeh, M.J. Lawley, D. Green, Computer-Assisted Diagnostic Coding: Effectiveness of an NLP-based approach using SNOMED CT to ICD-10 mappings, AMIA Annu. Symp. Proceedings. AMIA Symp. 2018 (2018) 807–816.
[79] J.D. Patrick, D.H.M. Nguyen, Y. Wang, M. Li, A knowledge discovery and reuse pipeline for information extraction in clinical notes, J. Am. Med. Informatics Assoc. 18 (2011) 574–579, https://doi.org/10.1136/amiajnl-2011-000302.
[80] Y. Peng, Q. Chen, Z. Lu, An Empirical Study of Multi-Task Learning on BERT for Biomedical Text Mining, Proc. BioNLP 2020 Work. (2020) 205–214. https://doi.org/10.18653/v1/2020.bionlp-1.22.
[81] S. Pradhan, N. Elhadad, B.R. South, D. Martinez, L. Christensen, A. Vogel, H. Suominen, W.W. Chapman, G. Savova, Evaluating the state of the art in disorder recognition and normalization of the clinical narrative, J. Am. Med. Informatics Assoc. 22 (2015) 143–154, https://doi.org/10.1136/amiajnl-2013-002544.
[82] Y. Qin, Y. Zeng, Research of Clinical Named Entity Recognition Based on Bi-LSTM-CRF, J. Shanghai Jiaotong Univ. 23 (2018) 392–397, https://doi.org/10.1007/s12204-018-1954-5.
[83] D. Raj, S.K. Sahu, A. Anand, Learning local and global contexts using a convolutional recurrent network model for relation classification in biomedical text, CoNLL 2017–21st, Conf. Comput. Nat. Lang. Learn. Proc. (2017) 311–321, https://doi.org/10.18653/v1/k17-1032.
[84] S.V. Ramanan, K. Radhakrishna, A. Waghmare, T. Raj, S.P. Nathan, S. M. Sreerama, S. Sampath, Dense Annotation of Free-Text Critical Care discharge summaries from an indian hospital and associated performance of a Clinical NLP Annotator, J. Med. Syst. 40 (2016) 187, https://doi.org/10.1007/s10916-016-0541-2.
[85] S. Rea, J. Pathak, G. Savova, T.A. Oniki, L. Westberg, C.E. Beebe, C. Tao, C. G. Parker, P.J. Haug, S.M. Huff, C.G. Chute, Building a robust, scalable and standards-driven infrastructure for secondary use of EHR data: the SHARPn project, J. Biomed. Inform. 45 (2012) 763–771, https://doi.org/10.1016/j.jbi.2012.01.009.
[86] B. Rink, S. Harabagiu, A generative model for unsupervised discovery of relations and argument classes from clinical texts, in: Proc. 2011 Conf. Empir. Methods Nat, Lang. Process., Association for Computational Linguistics, Edinburgh, Scotland, UK, 2011, pp. 519–528.
[87] S. Sahu, A. Anand, K. Oruganty, M. Gattu, Relation extraction from clinical texts using domain invariant convolutional neural network, Association for Computational Linguistics, Berlin, Germany, 2016 https://www.aclweb.org/anthology/W16-2928/.
[88] X. Shi, Y. Yi, Y. Xiong, B. Tang, Q. Chen, X. Wang, Z. Ji, Y. Zhang, H. Xu, Extracting entities with attributes in clinical text via joint deep learning, J. Am. Med. Informatics Assoc. 26 (2019) 1584–1591, https://doi.org/10.1093/jamia/ocz158.
[89] Y. Si, J. Wang, H. Xu, K. Roberts, Enhancing clinical concept extraction with contextual embeddings, J. Am. Med. Informatics Assoc. 26 (2019) 1297–1304, https://doi.org/10.1093/jamia/ocz096.
[90] J.M. Steinkamp, W. Bala, A. Sharma, J.J. Kantrowitz, Task definition, annotated dataset, and supervised natural language processing models for symptom extraction from unstructured clinical notes, J. Biomed. Inform. 102 (2020), https://doi.org/10.1016/j.jbi.2019.103354.
[91] S. Šuster, M. Sushil, W. Daelemans, Revisiting neural relation classification in clinical notes with external information, Association for Computational Linguistics, Brussels, Belgium, 2018.
[92] B. Tang, H. Cao, Y. Wu, M. Jiang, H. Xu, Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features., BMC Med, Inform. Decis. Mak. 13 (Suppl 1) (2013) S1, https://doi.org/10.1186/1472-6947-13-S1-S1.
[93] Y. Tao, B. Godefroy, G. Genthial, C. Potts, Effective Feature Representation for Clinical Text Concept Extraction, Proc. 2nd Clin. Nat. Lang. Process. Work. (2019) 1–14. https://doi.org/10.18653/v1/w19-1901.
[94] C. Tao, M. Filannino, Ö. Uzuner, FABLE: A Semi-Supervised Prescription Information Extraction System, AMIA Annu. Symp. Proceedings. AMIA Symp. (2018) 1534–1543.
[95] A.K. Tarcar, A. Tiwari, D. Rao, V.N. Dhaimodker, P. Rebelo, R. Desai, Healthcare NER models using language model pretraining, CEUR Workshop Proc. 2551 (2020) 12–18, https://doi.org/10.1145/3336191.3371879.
[96] S. Trivedi, R. Gildersleeve, S. Franco, A.S. Kanter, A. Chaudhry, Evaluation of a Concept Mapping Task Using Named Entity Recognition and Normalization in Unstructured Clinical Text, J. Healthc. Informatics Res. 4 (2020) 395–410. internal-pdf://141.237.114.99/Trivedi2020_Article_EvaluationOfAConceptMappin.pdf.
[97] C. Wang, R. Akella, A Hybrid Approach to Extracting Disorder Mentions from Clinical Notes, AMIA Jt, Summits Transl. Sci. Proceedings. AMIA Jt. Summits Transl. Sci. 2015 (2015) 183–187. internal-pdf://0149798315/2068895.pdf.
[98] Y. Wang, S. Liu, N. Afzal, M. Rastegar-Mojarad, L. Wang, F. Shen, P. Kingsbury, H. Liu, A comparison of word embeddings for the biomedical natural language processing, J. Biomed. Inform. 87 (2018) 12–20, https://doi.org/10.1016/j.jbi.2018.09.008.
[99] Q. Wei, Z. Ji, Z. Li, J. Du, J. Wang, J. Xu, Y. Xiang, F. Tiryaki, S. Wu, Y. Zhang, C. Tao, H. Xu, A study of deep learning approaches for medication and adverse drug event extraction from clinical text, J. Am. Med. Inform. Assoc. 27 (2020) 13–21, https://doi.org/10.1093/jamia/ocz063.
[100] Y. Wu, M. Jiang, J. Xu, D. Zhi, H. Xu, Clinical Named Entity Recognition Using Deep Learning Models, AMIA Annu. Symp. Proceedings. AMIA Symp. (2017) 1812–1819.
[101] Y. Wu, X. Yang, J. Bian, Y. Guo, H. Xu, W. Hogan, Combine factual medical knowledge and distributed word representation to improve clinical named entity recognition, AMIA Annu. Symp. Proceedings. AMIA Symp. (2018) 1110–1117.
[102] Y. Wu, J. Xu, M. Jiang, Y. Zhang, H. Xu, A study of neural word embeddings for named entity recognition in clinical text, AMIA Annu. Symp. Proceedings. AMIA Symp. (2015) 1326–1333.
[103] X. Xie, Y. Xiong, P.S. Yu, Y. Zhu, EHR Coding with Multi-Scale Feature Attention and Structured Knowledge Graph Propagation, in: Proc. 28th ACM Int. Conf. Inf. Knowl. Manag., Association for Computing Machinery, New York, NY, USA, 2019: pp. 649–658. https://doi.org/10.1145/3357384.3357897.
[104] J. Xu, Z. Li, Q. Wei, Y. Wu, Y. Xiang, H.-J. Lee, Y. Zhang, S. Wu, H. Xu, Applying a deep learning-based sequence labeling approach to detect attributes of medical concepts in clinical text, BMC Med. Inform. Decis. Mak. 19 (2019), https://doi.org/10.1186/s12911-019-0937-2.
[105] Y. Xu, K. Hong, J. Tsujii, E.-I.-C. Chang, Feature engineering combined with machine learning and rule-based methods for structured information extraction from narrative clinical discharge summaries, J. Am. Med. Informatics Assoc. 19 (2012) 824–832, https://doi.org/10.1136/amiajnl-2011-000776.
[106] X. Yang, J. Bian, Y. Gong, W.R. Hogan, Y. Wu, MADEx: a system for detecting medications, adverse drug events, and their relations from clinical notes, Drug Saf. 42 (2019) 123–133, https://doi.org/10.1007/s40264-018-0761-0.
[107] X. Yang, J. Bian, R. Fang, R.I. Bjarnadottir, W.R. Hogan, Y. Wu, Identifying relations of medications with adverse drug events using recurrent convolutional neural networks and gradient boosting, J. Am. Med. Informatics Assoc. 27 (2020) 65–72, internal-pdf://100.155.183.124/Yang-2020-Identifying relations of medications.pdf.
[108] X. Yang, J. Bian, W.R. Hogan, Y. Wu, Clinical concept extraction using transformers, J. Am. Med. Informatics Assoc. 27 (2020) 1935–1942, internal-pdf://152.1.21.190/Yang-2020-Clinical concept extraction using tr.pdf.

[109] E. Yehia, H. Boshnak, S. AbdelGaber, A. Abdo, D.S. Elzanfaly, Ontology-based clinical information extraction from physician's free-text notes, J. Biomed. Inform. 98 (2019).
[110] T. Zheng, W. Xie, L. Xu, X. He, Y. Zhang, M. You, G. Yang, Y. Chen, A machine learning-based framework to identify type 2 diabetes through electronic health records, Int. J. Med. Inform. 97 (2017) 120–127.
[111] X. Zhu, C. Cherry, S. Kiritchenko, J. Martin, B. de Bruijn, Detecting concept relations in clinical text: Insights from a state-of-the-art model, J. Biomed. Inform. 46 (2013) 275–285, https://doi.org/10.1016/j.jbi.2012.11.006.
[112] A. Roy, S. Pan, Incorporating medical knowledge in BERT for clinical relation extraction, in: Proc. 2021 Conf. Empir. Methods Nat. Lang. Process., Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 2021, pp. 5357–5366.
[113] G. Michalopoulos, Y. Wang, H. Kaka, H. Chen, A. Wong, {U}mls{BERT}: Clinical Domain Knowledge Augmentation of Contextual Embeddings Using the {U}nified {M}edical {L}anguage {S}ystem {M}etathesaurus, in: Proc. 2021 Conf. North Am. Chapter Assoc. Comput. Linguist. Hum. Lang. Technol., Association for Computational Linguistics, Online, 2021: pp. 1744–1753. https://doi.org/10.18653/v1/2021.naacl-main.139.
[114] V. Khetan, M.I.H. Rizvi, J. Huber, P. Bartusiak, B. Sacaleanu, A. Fano, MIMICause: Representation and automatic extraction of causal relation types from clinical notes, (2021) 764–773. https://doi.org/10.18653/v1/2022.findings-acl.63.
[115] U. Phan, N. Nguyen, Simple Semantic-based Data Augmentation for Named Entity Recognition in Biomedical Texts, in: Proc. 21st Work. Biomed, Lang. Process., Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 123–129.
[116] A. Khandelwal, A. Kar, V.R. Chikka, K. Karlapalem, Biomedical NER using Novel Schema and Distant Supervision, in: Proc. 21st Work. Biomed, Lang. Process., Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 155–160.
[117] S. Narayanan, K. Mannam, P. Achan, M.V. Ramesh, P.V. Rangan, S.P. Rajan, A contextual multi-task neural approach to medication and adverse events identification from clinical text, J. Biomed. Inform. 125 (2022). https://www.scopus.com/inward/record.uri?eid=2-s2.0-85120616157&doi=10.1016%2Fj.jbi.2021.103960&partnerID=40&md5=55a070400629d99923856765b7950031.
[118] A. Mulyar, O. Uzuner, B. McInnes, MT-clinical BERT: scaling clinical information extraction with multitask learning, J. Am. Med. Inform. Assoc. 28 (2021) 2108–2115. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116958963&doi=10.1093%2Fjamia%2Focab126&partnerID=40&md5=502b1e2c5ef305d87fa1e29e034713e0.
[119] J. Li, Y. Zhou, X. Jiang, K. Natarajan, S.V. Pakhomov, H. Liu, H. Xu, Are synthetic clinical notes useful for real natural language processing tasks: a case study on clinical entity recognition, J. Am. Med. Inform. Assoc. 28 (2021) 2193–2201. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85116958829&doi=10.1093%2Fjamia%2Focab112&partnerID=40&md5=312f6d8fce39ce2a1408fdc67394aa9d.
[120] Y. Tang, J. Yu, S. Li, B. Ji, Y. Tan, Q. Wu, Span classification based model for clinical concept extraction, Springer International Publishing (2021), https://doi.org/10.1007/978-3-030-70665-4_203.
[121] S.A. Moqurrab, U. Ayub, A. Anjum, S. Asghar, G. Srivastava, An accurate deep learning model for clinical entity recognition from clinical notes, IEEE J. Biomed. Heal. Informatics. 25 (2021) 3804–3811, https://doi.org/10.1109/JBHI.2021.3099755.
[122] A.D. Dave, G. Ruaño, J. Kost, X. Wang, Automated extraction of pain symptoms: a natural language approach using electronic health records, Pain Physician. 25 (2022) E245–E254.
[123] E. Soysal, J. Wang, M. Jiang, Y. Wu, S. Pakhomov, H. Liu, H. Xu, CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines, J. Am. Med. Informatics Assoc. 25 (2018) 331–336.
[124] J. Patrick, M. Li, High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge, (n.d.). https://doi.org/10.1136/jamia.2010.003939.
[125] N. Elhadad, S. Pradhan, S. Gorman, S. Manandhar, W. Chapman, G. Savova, SemEval-2015 task 14: Analysis of clinical text, in: Proc. 9th Int. Work. Semant. Eval. (SemEval 2015), 2015: pp. 303–310.
[126] K. Zheng, R.M. Ratwani, J. Adler-Milstein, Studying workflow and workarounds in electronic health record–supported work to improve health system performance, Ann. Intern. Med. 172 (2020) S116–S122.
[127] D. Ad, G. Ruano, J. Kost, X. Wang, Automated extraction of pain symptoms: a natural language approach using electronic health records, Pain Physician. 25 (2022) E245–E254. https://pubmed.ncbi.nlm.nih.gov/35322976/.
[128] C. Dreisbach, T.A. Koleck, P.E. Bourne, S. Bakken, A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data, Int. J. Med. Inform. 125 (2019) 37–46, https://doi.org/10.1016/j.ijmedinf.2019.02.008.
[129] T.A. Koleck, C. Dreisbach, P.E. Bourne, S. Bakken, Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review, J. Am. Med. Inform. Assoc. 26 (2019) 364–379, https://doi.org/10.1093/jamia/ocy173.
[130] OpenAI, GPT-4 Technical Report, (2023). https://arxiv.org/abs/2303.08774 (accessed April 13, 2023).
[131] Y. Peng, S. Yan, Z. Lu, Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets, ArXiv Prepr. ArXiv1906.05474 (2019).
[132] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ArXiv Prepr. ArXiv2007.15779 (2020).
[133] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, D. Zhi, Med-BERT: pre-trained contextualized embeddings on large-scale structured electronic health records for disease prediction, n.d.
[134] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D.M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language Models are Few-Shot Learners, (2020). https://arxiv.org/abs/2005.14165 (accessed July 23, 2020).
[135] A.-L. Rousseau, C. Baudelaire, K. Riera, Doctor GPT-3: hype or reality? - Nabla, (2020). https://www.nabla.com/blog/gpt-3/ (accessed March 2, 2021).
[136] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, Huggingface's transformers: State-of-the-art natural language processing, ArXiv Prepr. ArXiv1910.03771 (2019).
[137] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, S.R. Bowman, GLUE: A multi-task benchmark and analysis platform for natural language understanding, ArXiv Prepr. ArXiv1804.07461 (2018).
[138] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, C. Raffel, Extracting Training Data from Large Language Models (2020). http://arxiv.org/abs/2012.07805.
[139] X. Pan, M. Zhang, S. Ji, M. Yang, Privacy risks of general-purpose language models, in: 2020 IEEE Symp. Secur. Priv., IEEE, 2020: pp. 1314–1331.
[140] C.G. Walsh, B. Chaudhry, P. Dua, K.W. Goodman, B. Kaplan, R. Kavuluru, A. Solomonides, V. Subbian, Stigma, biomarkers, and algorithmic bias: recommendations for precision behavioral health with artificial intelligence, JAMIA Open. 3 (2020) 9–15.
[141] P. Paroubek, S. Chaudiron, L. Hirschman, Principles of evaluation in natural language processing, Rev. TAL. 48 (2007) 7–31.
[142] M. Nissim, L. Abzianidze, K. Evang, R. van der Goot, H. Haagsma, B. Plank, M. Wieling, Sharing is caring: The future of shared tasks, Comput. Linguist. 43 (2017) 897–904.
[143] C.P. Escartín, T. Lynn, J. Moorkens, J. Dunne, Towards transparency in NLP shared tasks, (2021). https://arxiv.org/abs/2105.05020 (accessed April 13, 2023).
[144] A.B. Kocaballi, K. Ijaz, L. Laranjo, J.C. Quiroz, D. Rezazadegan, H.L. Tong, S. Willcock, S. Berkovsky, E. Coiera, Envisioning an artificial intelligence documentation assistant for future primary care consultations: a co-design study with general practitioners, J. Am. Med. Informatics Assoc. 27 (2020) 1695–1704.
[145] D.F. Navarro, A.B. Kocaballi, M. Dras, S. Berkovsky, Collaboration, not Confrontation: Understanding General Practitioners' Attitudes Towards Natural Language and Text Automation in Clinical Practice, ACM Trans. Comput. Interact. (n.d.).
[146] K. Zheng, V.G.V. Vydiswaran, Y. Liu, Y. Wang, A. Stubbs, O. Uzuner, A.E. Gururaj, S. Bayer, J. Aberdeen, A. Rumshisky, S. Pakhomov, H. Liu, H. Xu, Ease of adoption of clinical natural language processing software: an evaluation of five systems, J. Biomed Inform. 58 (Suppl) (2015) S189–S196, https://doi.org/10.1016/j.jbi.2015.07.008.
[147] D. Sonntag, H.-J. Profitlich, An architecture of open-source tools to combine textual information extraction, faceted search and information visualisation, Artif. Intell. Med. 93 (2019) 13–28, https://doi.org/10.1016/j.artmed.2018.08.003.
[148] W. Hsu, R.K. Taira, S. El-Saden, H. Kangarloo, A.A.T. Bui, Context-based electronic health record: toward patient specific healthcare, IEEE Trans. Inf. Technol. Biomed. 16 (2012) 228–234.
[149] W. Hsu, R.K. Taira, F. Vinuela, A.A.T. Bui, A Case-Based Retrieval System Using Natural Language Processing and Population-Based Visualization, in: Proc. 2011 IEEE First Int. Conf. Healthc. Informatics, Imaging Syst. Biol., IEEE Computer Society, USA, 2011: pp. 221–228. https://doi.org/10.1109/HISB.2011.3.
[150] W. Xu, Toward human-centered AI: a perspective from human-computer interaction, Interactions 6 (2019) 42–46.