FAIR Data in Medical Research

Incorporating the FAIR Principles in the Research Data Life Cycle

Kersloot, M.G.

Publication date
2022

Chapter 4

Natural Language Processing algorithms for mapping clinical text fragments onto ontology concepts

A systematic review and recommendations for future studies

Martijn G. Kersloot
Florentien J. P. van Putten
Ameen Abu-Hanna
Ronald Cornet
Derk L. Arts

Journal of Biomedical Semantics 11, 14 (2020).
DOI: 10.1186/s13326-020-00231-z

Abstract

Background

Free-text descriptions in electronic health records (EHRs) can be of interest for clinical research and care optimization. However, free text cannot be readily interpreted by a computer and, therefore, has limited value. Natural Language Processing (NLP) algorithms can make free text machine-interpretable by attaching ontology concepts to it. However, implementations of NLP algorithms are not evaluated consistently. Therefore, the objective of this study was to review the current methods used for developing and evaluating NLP algorithms that map clinical text fragments onto ontology concepts. To standardize the evaluation of algorithms and reduce heterogeneity between studies, we propose a list of recommendations.

Methods

Two reviewers examined publications indexed by Scopus, IEEE, MEDLINE, EMBASE, the ACM Digital Library, and the ACL Anthology. Publications reporting on NLP for mapping clinical text from EHRs to ontology concepts were included. Year, country, setting, objective, evaluation and validation methods, NLP algorithms, terminology systems, dataset size and language, performance measures, reference standard, generalizability, operational use, and source code availability were extracted. The studies' objectives were categorized by way of induction. These results were used to define recommendations.

Results

Two thousand three hundred fifty five unique studies were identified. Two hundred fifty six studies reported on the development of NLP algorithms for mapping free text to ontology concepts. Seventy-seven described development and evaluation. Twenty-two studies did not perform a validation on unseen data and 68 studies did not perform external validation. Of 23 studies that claimed that their algorithm was generalizable, 5 tested this by external validation. A list of sixteen recommendations regarding the usage of NLP systems and algorithms, usage of data, evaluation and validation, presentation of results, and generalizability of results was developed.

Conclusion

We found many heterogeneous approaches to the reporting on the development and evaluation of NLP algorithms that map clinical text to ontology concepts. Over one-fourth of the identified publications did not perform an evaluation. In addition, over one-fourth of the included studies did not perform a validation, and 88% did not perform external validation. We believe that our recommendations, alongside an existing reporting standard, will increase the reproducibility and reusability of future studies and NLP algorithms in medicine.

Supplementary files referenced in this chapter can be accessed through
doi.org/10.1186/s13326-020-00231-z.

Background

One of the main activities of clinicians, besides providing direct patient care, is documenting care in the electronic health record (EHR). Currently, clinicians document clinical findings and symptoms primarily as free-text descriptions within clinical notes in the EHR since they are not able to fully express complex clinical findings and nuances of every patient in a structured format [40, 41]. These free-text descriptions are, amongst other purposes, of interest for clinical research [67, 68], as they cover more information about patients than structured EHR data [69]. However, free-text descriptions cannot be readily processed by a computer and, therefore, have limited value in research and care optimization.

One method to make free text machine-processable is entity linking, also known as annotation, i.e., mapping free-text phrases to ontology concepts that express the phrases' meaning. Ontologies are explicit formal specifications of the concepts in a domain and relations among them [22]. In the medical domain, SNOMED CT [70] and the Human Phenotype Ontology (HPO) [71] are examples of widely used ontologies to annotate clinical data. After the data has been annotated, it can be reused by clinicians to query EHRs [72,73], to classify patients into different risk groups [74,75], to detect a patient's eligibility for clinical trials [76], and for clinical research [77].

Natural Language Processing (NLP) can be used to (semi-)automatically process free text. The literature indicates that NLP algorithms have been broadly adopted and implemented in the field of medicine [78, 79], including algorithms that map clinical text to ontology concepts [80]. Unfortunately, implementations of these algorithms are not being evaluated consistently or according to a predefined framework and limited availability of data sets and tools hampers external validation [81].

To improve and standardize the development and evaluation of NLP algorithms, a good practice guideline for evaluating NLP implementations is desirable [82, 83]. Such a guideline would enable researchers to reduce the heterogeneity between the evaluation methodology and reporting of their studies. Generic reporting guidelines such as TRIPOD [84] for prediction models, STROBE [85] for observational studies, RECORD [86] for studies conducted using routinely-collected health data, and STARD [87] for diagnostic accuracy studies, are available, but are often not used in NLP research. This is presumably because some guideline elements do not apply to NLP and some NLP-related elements are missing or unclear. We, therefore, believe that a list of recommendations for the evaluation methods of and reporting on NLP studies, complementary to the generic reporting guidelines, will help to improve the quality of future studies.

In this study, we will systematically review the current state of the development and evaluation of NLP algorithms that map clinical text onto ontology concepts, in order to quantify the heterogeneity of methodologies used. We will propose a structured list of recommendations, which is harmonized from existing standards and based on the outcomes of the review, to support the systematic evaluation of the algorithms in future studies.

Methods

This study consists of two phases: a systematic review of the literature and the formation of recommendations based on the findings of the review.

Literature review

A systematic review of the literature was performed using the Preferred Reporting Items for Systematic reviews and Meta-Analyses (PRISMA) statement [88].

Search strategy and study selection

We searched Scopus, IEEE, MEDLINE, EMBASE, the Association for Computing Machinery (ACM) Digital Library, and the Association for Computational Linguistics (ACL) Anthology for the following keywords: Natural Language Processing, Medical Language Processing, Electronic Health Record, reports, charts, clinical notes, clinical text, medical notes, ontolog*, concept*, encod*, annotat*, code, and coding. We excluded the words 'reports' and 'charts' in the ACL and ACM databases since these databases also contain publications on non-medical subjects. The detailed search strategies for each database can be found in online supplementary file 2. We searched until December 19, 2019 and applied the filters "English" and "has abstract" for all databases. Moreover, we applied the filters "Medicine, Health Professions, and Nursing" for Scopus, the filters "Conferences", "Journals", and "Early Access Articles" for IEEE, and the filter "Article" for Scopus and EMBASE. EndNote X9 [89] and Rayyan [90] were used to review and delete duplicates.

The selection process consisted of three phases. In the first phase, two independent reviewers (MK, FP) individually assessed the resulting titles and abstracts and selected publications that fitted the criteria described in Box 4.1.

Some studies do not describe the application of NLP in their study by only listing NLP as the used method, instead of describing its specific implementation. Additionally, some studies create their own ontology to perform NLP tasks, instead of using an established, domain-accepted ontology. Both approaches limit the generalizability of the study's methods. Therefore, we defined exclusion criteria that are described in Box 4.2.

In the second phase, both reviewers excluded publications where the developed NLP algorithm was not evaluated by assessing the titles, abstracts, and, in case of uncertainty, the Method section of the publication. In the third phase, both reviewers independently

Box 4.1: Inclusion criteria

- Medical language processing as the main topic of the publication
- Use of EHR data, clinical reports, or clinical notes
- Algorithm performs annotation
- Publication is written in English

Box 4.2: Exclusion criteria

- Implementation was not described
- Implementation does not use an existing established ontology for encoding
- Not published in a peer-reviewed journal (except for ACL and ACM publications)

evaluated the resulting full-text articles for relevance. The reviewers used Rayyan [90] in the first phase and Covidence [91] in the second and third phases to store the information about the articles and their inclusion. In all phases, both reviewers independently reviewed all publications. After each phase the reviewers discussed any disagreement until consensus was reached.

Data extraction and categorization

Both reviewers categorized the implementations of the found algorithms and noted their characteristics in a structured form in Covidence. The objectives of the included studies and their associated NLP tasks were categorized by way of induction. The results were compared and merged into one result set.

We collected the following characteristics of the studies, based on a combination of TRIPOD [84], STROBE [85], RECORD [86], and STARD [87] statement elements (see online supplementary file 3): year, country, setting, objectives, evaluation methods, used NLP systems or algorithms, used terminology systems, size of datasets, performance measures, reference standard, language of the free-text data, validation methods, generalizability, operational use, and source code availability.

List of recommendations

Based on the findings of the systematic review and elements from the TRIPOD, STROBE, RECORD, and STARD statements, we formed a list of recommendations. The recommendations focus on the development and evaluation of NLP algorithms for mapping clinical text fragments onto ontology concepts and the reporting of evaluation results.

Results

The literature search generated a total of 2355 unique publications. After reviewing the titles and abstracts, we selected 256 publications for additional screening. Out of the 256 publications, we excluded 65 publications, as the described Natural Language Processing algorithms in those publications were not evaluated. The full text of the remaining 191 publications was assessed and 114 publications did not meet our criteria, of which 3 publications in which the algorithm was not evaluated, resulting in 77 included articles describing 77 studies. Reference checking did not provide any additional publications. The PRISMA flow diagram is presented in Figure 4.1.

The induction process resulted in eight categories and ten associated NLP tasks that describe the objectives of the papers: computer-assisted coding, information comparison, information enrichment, information extraction, prediction, software development and evaluation, and text processing. Our definitions of these NLP tasks and the associated categories are given in Table 4.1 and Table 4.2.

Table 4.3 lists the included publications with their first author, year, title, and country. Table 4.4 lists the included publications with their evaluation methodologies. The non-

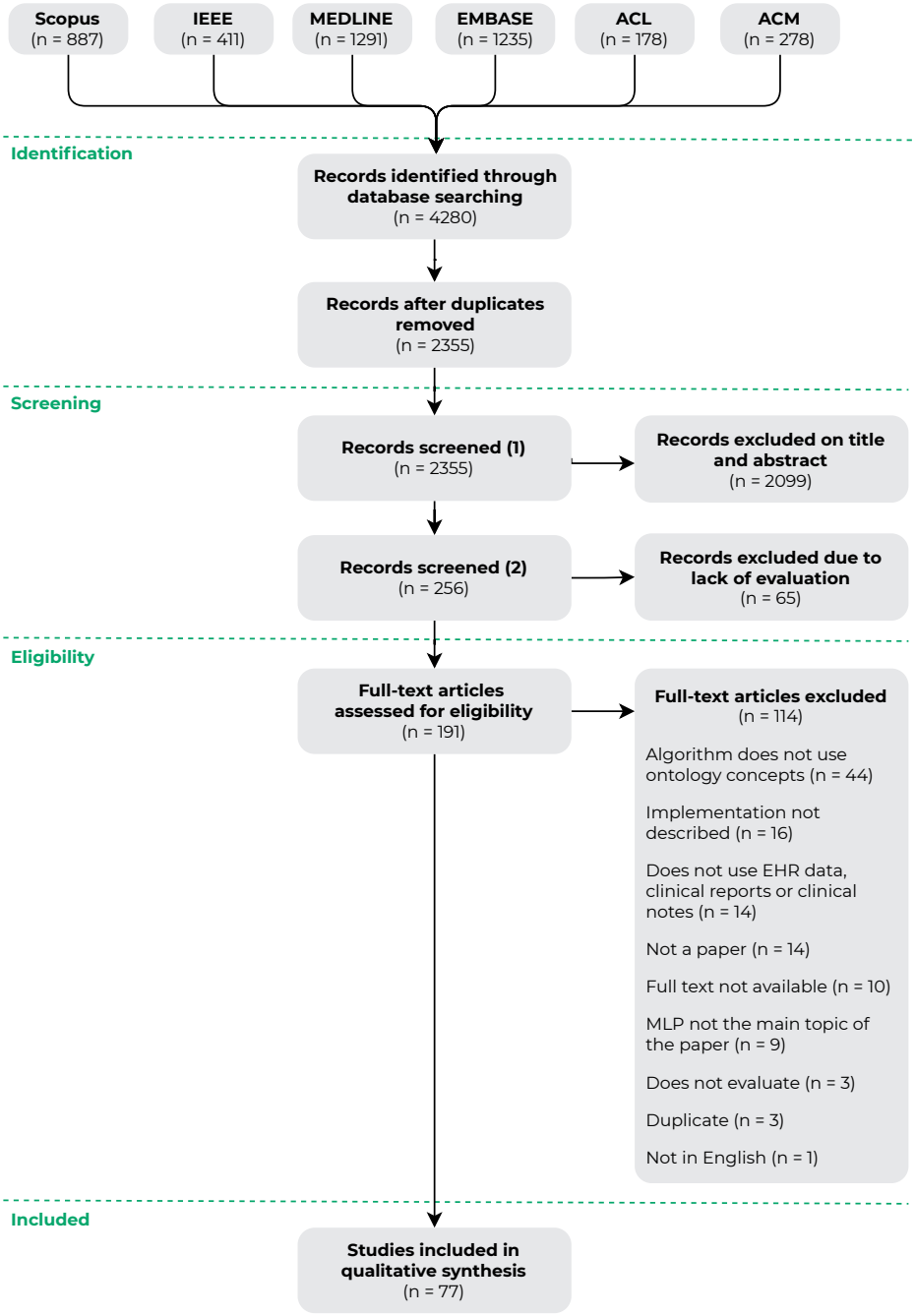

Figure 4.1: PRISMA flow diagram

induced data, including data regarding the sizes of the datasets used in the studies, can be found in online supplementary file 1.

Table 4.5 summarizes the general characteristics of the included studies and Table 4.6 summarizes the evaluation methods used in these studies. In all 77 papers, we found twenty different performance measures (Table 4.7).

Discussion

In this systematic review, we reviewed the current state of NLP algorithms that map clinical text fragments onto ontology concepts with regard to their development and evaluation, in order to propose recommendations for future studies.

Main findings and recommendations

We identified 256 studies that reported on the development of such algorithms, of which 68 did not evaluate the performance of the system. We included 77 studies. Many publications did not report their findings in a structured way, which made it challenging to extract all the data in a reliable manner. We discuss our findings and recommendations in the following five categories: Used NLP systems and algorithms, Used data, Evaluation and validation, Presentation of results, and Generalizability of results. A checklist for determining if the recommendations are followed in the reporting of an NLP study is added as supplementary material to this paper.

Used NLP systems and algorithms

A variety of NLP systems are used in the reviewed studies. Researchers use existing systems (n = 29, 38%), develop new systems with existing components (n = 25, 33%), or develop a completely new system (n = 23, 30%). Most studies, however, do not publish their (adapted) source code (n = 57, 74%), and a description of the algorithm in the final publication is often not detailed enough to replicate it. To ensure reproducibility, implementation details, including details on data processing, and preferably the source code should be published, allowing other researchers to compare their implementations or to reproduce the results. Based on these findings, we formulated three recommendations (Box 4.3).

Used data

Most authors evaluate their algorithms with manual annotations (n = 40, 52%) and use data present in their institutions (n = 55, 71%). However, it is not clear what these datasets consist of. Most studies describe the data as 'reports', 'notes', or 'summaries', but do not list the contents or example rows from the dataset. It is, therefore, not clear what types of patients and what specific types of data are included, making the study hard to reproduce. Finally, we found a wide range of dataset sizes. The training datasets, for example, ranged from 10 clinical notes to 636.439 discharge reports. The use of small datasets can result in an overfitted algorithm that either performs well on the dataset,

Box 4.3: Recommendations regarding the use of systems and algorithms

1. Describe the system or algorithm that is used or the system that is developed for the specific NLP task.
    1. When an existing NLP system or algorithm is used, describe how it is set up, how it is implemented in practice, and if and how the implementation differs from the original implementation.
    2. When a new system is developed, describe the components and features used in the system, and preferably include a flow chart that explains how these elements work together.
2. Include the source code of the developed algorithm as supplementary material to the publication or upload the source code to a repository such as GitHub.
3. Specify which ontologies are used in the encoding task, including the version of the ontology.
    1. If a new ontology is developed for the encoding task, report on the development and content of the ontology and rationale for the development of a new ontology instead of the use of an existing one. The MIRO guidelines could be used to structure the report [92].

but not on an external dataset, or performs poorly, for the algorithm was only trained on a specific type of data. More difficult recognition tasks require more data, and therefore sample size planning is recommended [93]. To improve the description and availability of datasets used in NLP studies, we formulated three recommendations (Box 4.4).

Box 4.4: Recommendations regarding the use of data

1. To ensure that new algorithms can be compared against your system, aim to publish the used training, development, and validation data in a data repository.
    1. In case the data cannot be published, determine if the data can be accessed on request or can be used in a federated learning approach (i.e., a learning process in which the data owners collaboratively train a model in which process any data owner does not expose the data to others [94]).
2. In case a reference standard is used, include information about the origin of the data (external dataset, subset of the dataset) and the characteristics of the data in the dataset. If possible, reference the dataset using a DOI or URL.
3. If an external dataset is used, give a short description of the data present in the dataset and reference the source of the dataset.

Evaluation and validation

Evaluation of the algorithm determines its performance on the dataset, and validation determines if the algorithm is not overfitted on that dataset and thus if the algorithm might work on other datasets as well. Over one-fourth of the studies (n = 68, 27%) that we identified did not evaluate their algorithms. In addition, 22 included studies (29%) did not validate the developed algorithm. A statement claiming that an algorithm can be used in clinical practice can be questioned if the algorithm has not been evaluated

and validated. Across all studies, 20 performance measures were used. To harmonize evaluation and validation efforts, we formulated three recommendations (Box 4.5).

Box 4.5: Recommendations regarding the evaluation and validation of Natural Language Processing algorithms

1. Perform an evaluation using generic (i.e., precision, recall, and F-score) performance measures and appropriate aspects of evaluation including discrimination, calibration, and preferably accuracies of predictions (e.g., AUC, calibration graphs, and the Brier score).
    1. Include a motivation for the choice of measures, with references to existing literature where appropriate (e.g., Sokolova and Lapalme's analysis of performance measures [95]).
2. Perform an error analysis and discuss the errors in the Discussion section of the paper. Include possible changes to the algorithm that could improve its performance for these specific errors.
3. When using a non-probabilistic NLP method: determine the cut-off value (a priori) for a 'good' test result before evaluating the algorithm. Elaborate why this cut-off value is chosen.

Presentation of results

Authors report the evaluation results in various formats. Only twelve articles (16%) included a confusion matrix which helps the reader understand the results and their impact. Not including the true positives, true negatives, false positives, and false negatives in the Results section of the publication, could lead to misinterpretation of the results of the publication's readers. For example, a high F-score in an evaluation study does not directly mean that the algorithm performs well. There is also a possibility that out of 100 included cases in the study, there was only one true positive case, and 99 true negative cases, indicating that the author should have used a different dataset. Results should be clearly presented to the user, preferably in a table, as results only described in the text do not provide a proper overview of the evaluation outcomes (Box 4.6). This also helps the reader interpret results, as opposed to having to scan a free text paragraph. Most publications did not perform an error analysis, while this will help to understand the limitations of the algorithm and implies topics for future research.

Box 4.6: Recommendations regarding the presentation of results

1. Report the outcomes of the evaluation in a clear manner, preferably in a table accompanied by a textual description of the outcomes.
    1. Aim to include a confusion matrix in the reporting of the outcomes.
2. Use figures if they contribute to the making the results more readable and understandable for the reader. If a figure is used, make sure that the data are also available in the text or in a table.

Generalizability of results

88% of the studies did not perform external validation (n = 68). Of the studies that claimed that their algorithm was generalizable, only 22% (n = 5) assessed this claim through external validation. However, one cannot claim generalizability without testing for it. Moreover, in 19% (n = 3) of the cases where external datasets were used, the datasets were not referenced and only listed in the text of the article, making it harder to find the used data and reproduce the results. Algorithm performance should be compared to that of other state-of-the-art algorithms, as this helps the reader decide whether the new algorithm could be considered useful for clinical practice. However, only 24 studies (31%) made this comparison, and four of those studies (17%) tested the performance difference for statistical significance. We also found that the authors' descriptions of generalizability are rather ambiguous and unclear. We formulated five recommendations regarding the generalizability of results (Box 4.7).

Box 4.7: Recommendations regarding the generalizability of results

1. Compare the results of the evaluated algorithm with other algorithms by using the same dataset as reported in the publication of the other algorithm or by processing the same dataset with another algorithm available through the literature. Report the outcomes of both experiments and test for statistical significance.
2. Describe in what setting the research is performed. Include if the research is part of a challenge (e.g., i2b2 challenge), or that the research is carried out in a specific institute or department.
3. Before claiming generalizability, perform external validation by testing the algorithm on a different, external dataset from other research projects or other publicly available datasets. Aim to use a dataset with a different case mix, different individuals, and different types of text.
4. Determine and describe if there are potential sources of bias in data selection, data use by the NLP algorithm or system, and evaluation.
5. When claiming generalizability, clearly describe the conditions under which the algorithm can be used in a different setting. Describe for which population, domain, and type and language of data the algorithm can be used.

Strengths

Our study has three main strengths: First, to our knowledge, this is the first systematic review that focuses on the evaluation of NLP algorithms in medicine. Second, we used a large number of databases for our search, resulting in publications from many different sources, such as medical journals and computer science conferences. Third, we used existing statements and guidelines and harmonized them to induce our findings and used these findings to propose a list of recommendations.

Limitations

Several limitations of our study should be noted as well. First, we only focused on algorithms that evaluated the outcomes of the developed algorithms. Second, the majority

of the studies found by our literature search used NLP methods that are not considered to be state of the art. We found that only a small part of the included studies was using state-of-the-art NLP methods, such as word and graph embeddings. This indicates that these methods are not broadly applied yet for algorithms that map clinical text to ontology concepts in medicine and that future research into these methods is needed. Lastly, we did not focus on the outcomes of the evaluation, nor did we exclude publications that were of low methodological quality. However, we feel that NLP publications are too heterogeneous to compare and that including all types of evaluations, including those of lesser quality, gives a good overview of the state of the art.

Conclusion

In this study, we found many heterogeneous approaches to the development and evaluation of NLP algorithms that map clinical text fragments to ontology concepts and the reporting of the evaluation results. Over one-fourth of the publications that report on the use of such NLP algorithms did not evaluate the developed or implemented algorithm. In addition, over one-fourth of the included studies did not perform a validation and nearly nine out of ten studies did not perform external validation. Of the studies that claimed that their algorithm was generalizable, only one-fifth tested this by external validation. Based on the assessment of the approaches and findings from the literature, we developed a list of sixteen recommendations for future studies. We believe that our recommendations, along with the use of a generic reporting standard, such as TRIPOD, STROBE, RECORD, or STARD, will increase the reproducibility and reusability of future studies and algorithms.

Table 4.1: Induced objective tasks with their definition and an example

Induced NLP task(s)	Description	Example
Concept detection [1]	Assign ontology concepts to phrases in free text (i.e., entity linking or annotation)	"Systolic blood pressure" can be represented as SNOMED-CT concept 271649006 | Systolic blood pressure (observable entity) |
Event detection	Detect events in free text	"Patient visited the outpatient clinic in January 2020" is an event of type Visit.
Relationship detection	Detect semantic relationships between concepts in free text	The concept Lung cancer in "This patient was diagnosed with recurrent lung cancer" is related to the concept Recurrence.
Text normalization	Transform free text into a single canonical form	"This patient was diagnosed with influenza last year." becomes "This patient be diagnose with influenza last year."
Text summarization	Create a short summary of free text and possible restructure the text based on this summary	"Last year, this patient visited the clinic and was diagnosed with diabetes mellitus type 2, and in addition to his diabetes, the patient was also diagnosed with hypertension" becomes "Last year, this patient was diagnosed with diabetes mellitus type 2 and hypertension".
Classification	Assign categories to free text	A report containing the text "This patient is not diagnosed yet" will be assigned to the category Undiagnosed.
Prediction	Create a predictive model based on free text	Predict the outcome of the APACHE score based on the (free-text) content in a patient chart.
Identification	Identify documents (e.g., reports or patient charts) that match a specific condition based on the contents of the document	Find all patient charts that describe patients with hypertension and a BMI above 30.
Software development	Develop new or build upon existing NLP software	A new algorithm was developed to map ontology concepts to free text in clinical reports.
Software evaluation	Evaluate the effectiveness of NLP software	The mapping algorithm has an F-score of 0.874.

Table 4.2: Induced objective categories with their definition and associated NLP task(s)

Induced category	Induced NLP task(s)	Definition
Computer-assisted coding	· Concept detection	Perform semi-automated annotation (i.e., with a human in the loop)
Information comparison	· Concept detection · Event detection · Relationship detection	Compare extracted structured information to information available in free-text form
Information enrichment	· Concept detection · Event detection · Relationship detection · Text normalization · Text summarization	Extract structured information from free text and attach this new information to the source
Information extraction	· Concept detection · Event detection · Relationship detection	Extract structured information from free text
Prediction	· Classification · Prediction · Identification	Use structured information to classify free-text reports, predict outcomes, or identify cases
Software development and evaluation	· Software development · Software evaluation	Develop new NLP software or evaluate new or existing NLP software
Text processing	· Text normalization · Text summarization	Transform free text into a new, more comprehensible form

Table 4.3: Included publications and their first author, year, title, and country

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Afshar	2019	USA	No	Inf. extr.	Clinical Data Warehouse Data	●	English	◧	UMLS (CPT, HCPCS, ICD-10, ICD10CM / ICD9CM, LOINC, MeSH, SNOMED-CT, RxNorm)	?	No, only links to cTAKES source code	[96]
Alnazzawi	2016	UK	No	Inf. enr.	PhenoCHF corpus [1]	○	English	■	UMLS	?	N/A	[97]
Atutxa	2018	Spain	No	Inf. enr.	EHR documents	●	Spanish	□	ICD (SNOMED-CT for normalization)	Not yet, aim to embed it in human-supervised loop	?	[98]
Barrett	2013	USA	No	Inf. extr.	Palliative care consult letters	●	English	□	SNOMED CT	?	No, but planned	[99]
Becker	2016	Germany	No	Inf. extr.	ShARe/CLEF corpus (2013) [2]	○	German	■	SNOMED CT (English), UMLS (German)	Not yet, still under development	N/A	[100]
Becker	2019	Germany	No	Inf. extr.	Clinical notes of patients with known colorectal cancer	●	German	◧	UMLS	Yes, led to improved quality of care for colorectal patients	?	[101]
Bejan	2015	USA	No	Inf. extr.	Discharge summaries and i2b2/VA challenge dataset (2010) [3]	◑	English	■	UMLS	No	N/A	[102]
Castro	2010	Spain	No	Inf. extr.	Clinical notes with 'most relevant information'	●	Spanish	■	SNOMED CT	?	N/A	[103]
Catling	2018	UK	No	Software	MIMIC-III dataset [4]	○	English	□	ICD-9-CM	?	?	[104]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Chapman	2004	USA	No	Inf. extr.	Emergency department reports	●	English	■	UMLS	?	N/A	[105]
Chen	2016	USA	No	Inf. enr.	Discharge summaries and progress notes	●	English	◧	UMLS	?	?	[106]
Chiaramello	2016	Italy	No	Inf. extr.	Clinical notes (cardiology, diabetology, hepatology, nephrology, and oncology)	●	Italian	■	UMLS	?	N/A	[107]
Chodey	2016	USA	SemEval (2014)	Inf. extr.	ICU Data: Discharge summaries, ECG, echo, and radiology	○	English	◧	UMLS	?	?	[108]
Chung	2005	USA	No	Inf. extr.	Echocardiogram reports	●	English	◧	UMLS	Not yet, it will be used to populate a registry	?	[109]
Combi	2018	Italy	No	Inf. extr.	VigiSegn (adverse drug reactions) reports	●	Italian + English	□	MedDRA	Yes, implemented in VigiFarmaco	Pseudo-code	[110]
De Bruijn	2011	Canada	i2b2/VA (2010)	Inf. extr.	Hospital discharge summaries and progress reports	○	English	◧	UMLS	?	?	[111]
Deisseroth	2019	USA	No	Inf. extr.	Six sets of real patient data from four different medical centers.	●	English	□	HPO	?	Yes	[112]
Demner-Fushman	2017	USA	No	Software	BioScope [5], NCBI disease corpus [6], i2b2/VA challenge corpus (2010) [3], ShARe corpus [7], LHC test collection (biological/clinical journal abstracts)	○	English	◧	UMLS	Yes, used in other papers identified in literature search	Yes	[113]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Divita	2014	USA	Parts: i2b2/VA (2010)	Software	Randomly selected clinical records from the most frequent document types	●	English	□	UMLS (level 0 + 9)	Yes, used by VA Informatics and Computing Infrastructure	Yes	[114]
Duarte	2018	Portugal	No	Inf. enr.	Death certificates, clinical bulletins, and autopsy reports	●	Portuguese	□	ICD-10	Yes, used by Portugese Ministry of Health for near real-time death cause surveillance	?	[115]
Falis	2019	UK	No	Inf. extr.	MIMIC-III dataset [4]	○	English	□	ICD-9	?	?	[116]
Ferrão	2013	Portugal	No	Inf. enr.	Inpatient adult episodes from the EHR	●	Portuguese	□	ICD-9-CM	?	?	[117]
Gerbier	2011	France	No	Inf. extr.	Computerized emergency department medical records	●	French	□	ICD-10, CCAM, SNOMED CT, ATC, MeSH, ICPC-2, DCR	Not yet, will be integrated into a CDSS	?	[118]
Goicoechea Salazar	2013	Spain	No	Inf. enr.	Diagnostic text from patient records	●	Spanish	□	ICD-9-CM	?	?	[119]
Hamid	2013	USA	No	Class.	Notes of Iraq and Afghanistan veterans from the VA national clinical database	●	English	■	UMLS	?	N/A	[120]
Hassanzadeh	2016	Australia	No	Inf. extr.	ShARe/CLEF corpus (2013) [2]	○	English	■	UMLS, SNOMED CT	N/A	N/A	[121]
Helwe	2017	Lebanon	No	Comp-ass. cod.	MIMIC-III dataset	○	English	□	UMLS, ICD	?	?	[122]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Hersh	2001	USA	No	Inf. enr.	Radiology image reports	●	English	■	UMLS	No, still in develop-ment/testing	Pseudo-code	[123]
Hoogendoorn	2015	Netherlands	No	Pred.	Consultation notes of patients in a primary care setting	●	Dutch	□	SNOMED-CT, UMLS, ICPC	⍰	⍰	[124]
Jindal	2013	USA	i2b2 (2012)	Inf. extr.	i2b2 challenge corpus (2012) [8]	○	English	◧	UMLS, SNOMED CT, MeSH	⍰	⍰	[125]
Kang	2009	Korea	No	Inf. extr.	Discharge summaries	●	Korean	□	KOMET, UMLS	⍰	⍰	[126]
Kersloot	2019	Netherlands	No	Inf. extr.	(Non-small cell) Lung cancer charts	●	English	◧	SNOMED CT	⍰	Yes	[127]
König	2019	Germany	No	Software	Discharge letters from BASE-II study	●	German	◧	Wingert-Nomenclature	No, still has to prove its value	⍰	[128]
Li	2015	USA	No	Inf. comp.	Clinical notes and discharge prescription lists	●	English	◧	UMLS, SNOMED CT, RxNorm	Not yet, plans to move to production	Pseudo-code	[129]
Li	2019	USA	No	Inf. extr.	EHR notes	●	English	◧	UMLS, SNOMED CT, MedDRA	⍰	⍰	[130]
Lingren	2016	USA	No	Class.	Structured and unstructured data from two EHR databases	●	English	◧	UMLS, ICD-9, RxNorm	⍰	⍰	[75]
Liu	2019	USA	No	Inf. extr.	Clinical notes from different institutions + PubMed Case report abstracts	◑	English	■	HPO	⍰	N/A	[131]
Lowe	2009	USA	No	Inf. extr.	Single-specimen pathology reports	●	English	■	UMLS, SNOMED CT	⍰	N/A	[132]
Luo	2014	USA	No	Inf. extr.	Pathology reports	●	English	◧	UMLS, SNOMED CT	Yes, currently working on project in multiple hospitals	⍰	[133]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ⍰ : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Meystre	2006	USA	No	Inf. enr.	Clinical documents form adult inpatients in a cardiovascular unit	●	English	◧	UMLS (level 0), SNOMED CT	Not yet, testing in practice	?	[134]
Meystre	2010	USA	i2b2 (2009)	Inf. extr.	i2b2 challenge dataset (2009) [9]	○	English	□	UMLS	Not yet, possible integration in research infrastructure	?	[135]
Minard	2011	France	i2b2/VA (2010)	Inf. extr.	i2b2/VA challenge corpus (2010) [3]	○	English	◧	UMLS	?	?	[136]
Mishra	2019	USA	No	Inf. extr.	Clinical notes from NIH Clinical Center data warehouse	●	English	■	UMLS, HPO	?	N/A	[137]
Nguyen	2018	Australia	No	Comp-ass. cod.	Hospital progress notes	●	English	◧	SNOMED CT, ICD-10-AM	?	?	[138]
Oellrich	2015	UK	No	Inf. extr.	PubMed abstracts, clinical trial information, i2b2/VA challenge corpus (2010) 3, SHARE/CLEF (2013) [2]	○	English	■	UMLS	?	N/A	[139]
Patrick	2011	Australia	i2b2/VA (2010)	Inf. extr.	i2b2/VA challenge corpus (2010) [3]	○	English	□	UMLS, SNOMED CT	?	?	[140]
Pérez	2018	Spain	No	Text pro-cessing	Spontaneous DTs randomly selected entries	●	Spanish	□	ICD	?	?	[141]
Reátegui	2018	Canada	No	Inf. extr.	i2b2 challenge corpus (2008) [10]	○	English	◧	UMLS, SNOMED CT, RxNorm	?	?	[142]
Roberts	2011	USA	i2b2/VA (2010)	Inf. extr.	i2b2/VA challenge corpus (2010) [3]	○	English	◧	UMLS, ICD-9	?	?	[143]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Rousseau	2019	USA	No	Inf. comp.	ED encounters for patients with headaches who received head CT	●	English	■	UMLS: SNOMED CT, RadLex	?	N/A	[144]
Savova	2010	USA	i2b2 (2006, 2008)	Inf. extr.	Subset of clinical notes from the EMR	●	English	◧	UMLS, SNOMED CT, RxNorm	Yes, used in other papers identified in literature search	Yes	[145]
Shivade	2015	USA	i2b2/UTHealth (2014)	Class.	i2b2 challenge corpus (2014) [11]	○	English	■	UMLS	?	N/A	[74]
Shoenbill	2019	USA	No	Inf. extr.	EHR notes from hypertension patients	●	English	■	UMLS, SNOMED CT	?	N/A	[146]
Sohn	2014	USA	No	Inf. extr.	Clinical notes with medication mentions	●	English	□	RxNorm	?	Yes	[147]
Solti	2008	USA	No	Inf. enr.	Cardiology ambulatory progress notes	●	English	■	UMLS	?	N/A	[148]
Soriano	2019	Spain	No	Inf. extr.	clinical emergency discharge reports	●	Spanish	□	SNOMED CT	Not yet	Yes	[149]
Soysal	2018	USA	Parts: i2b2 (2009 + 2010), ShARe/CLEF (2013), Sem-EVAL (2014)	Software	Discharge summaries from the i2b2/VA challenge corpus (2010) [3], outpatient clinic visit notes, mock clinical documents	◑	English	□	UMLS	Yes, used by various institutions and industrial entities	Yes	[150]
Spasić	2015	UK	No	Inf. extr.	MRI reports of patients	●	English	◧	TRAK, UMLS, MEDCIN, RadLex	?	Yes	[151]
Strauss	2013	USA	No	Inf. extr.	Pathology reports of breast and prostate cancer patients	●	English	□	SNOMED CT	?	Yes	[152]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Sung	2018	Taiwan	No	Inf. extr.	Cases of adult patients with AIS	●	English	■	UMLS	?	N/A	[153]
Tchechmedjiev	2018	France	No	Inf. extr.	Quaero (French MEDLINE abstract titles + EMEA drug labels) + CépiDC (ICD-10 coding of death certificates)	○	French	◧	UMLS terminologies (ICD-10)	Yes, available in SIFR BioPortal	Yes	[154]
Ternois	2018	France	No	Class.	Endoscopy reports written between 2015 and 2016	●	French	□	CCAM	?	?	[155]
Travers	2004	USA	No	Inf. extr.	Chief complaint text entries for all emergency department visits	●	English	□	UMLS	?	?	[156]
Tulkens	2019	Belgium	No	Inf. extr.	i2b2/VA challenge corpus (2010) [3]	○	English	◧	UMLS	?	Yes	[157]
Usui	2018	Japan	No	Pred.	Electronic medication history data from pharmacy	●	Japanese	□	ICD-10	Not yet, expect to use it	?	[158]
Valtchinov	2019	USA	No	Class.	Radiology reports, emergency department notes + other clinical reports	●	English	■	SNOMED CT, RadLex	?	N/A	[159]
Wadia	2018	USA	No	Class.	Chest CT reports	●	English	■	SNOMED CT, UMLS	?	N/A	[160]
Walker	2019	USA	No	Inf. extr.	Treatment sites from EMR	●	English	□	UMLS	?	?	[161]
Xie	2019	China	No	Inf. extr.	MIMIC-III dataset [4]	○	English	□	ICD-9-CM, ICD-10	?	?	[162]
Xu	2011	USA	No	Class.	CRC patient cases from the Synthetic Derivative database	●	English	■	UMLS	No, still under development	N/A	[163]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

Continued on next page

Author	Year	Country	Challenge	Objective	Data origin	Data	Language	System	Term. systems	In use	Code	Ref
Yadav	2013	USA	No	Pred.	Emergency department CT imaging reports	●	English	■	UMLS	?	Yes, command line	[164]
Yao	2019	USA	No	Pred.	i2b2 challenge corpus (2008) [10]	○	English	◧	UMLS	?	Part (Sorl)	[165]
Zeng	2018	USA	No	Class.	Progress notes and breast cancer surgical pathology reports	●	English	◧	UMLS	?	?	[166]
Zhang	2013	USA	No	Inf. extr.	i2b2/VA challenge corpus (2010) [3] and GENIA corpus (MEDLINE abstracts)	○	English	□	UMLS	?	?	[167]
Zhou	2006	USA	No	Inf. extr.	Records of patients with breast complaints	●	English	□	UMLS	No, still under development	?	[168]
Zhou	2011	USA	No	Software	COPD and CAD patients	●	English	□	SNOMED CT, RxNorm, UMLS, PPL, MDD, HL7 value sets	Yes, described in other paper ([169])	?	[170]
Zhou	2014	USA	No	Inf. extr.	Admission notes and discharge summaries	●	English	■	SNOMED CT, HL7 RoleCodes	?	N/A	[169]

● : Own dataset ○ : Existing dataset ◑ : Own and existing dataset □ : New system ■ : Existing system ◧ : New and existing system ? : Not listed N/A: Not applicable

1. PhenoCHF corpus: narrative reports from electronic health records (EHRs) and literature articles
2. ShARe/CLEF corpus (2013): narrative clinical reports
3. i2b2/VA challenge dataset (2010): discharge summaries and progress reports
4. MIMIC-III dataset: demographics, vital sign measurements, laboratory test results, procedures, medications, caregiver notes, imaging reports, and mortality
5. BioScope corpus: medical free texts, biological full papers and biological scientific abstracts
6. NCBI disease corpus: PubMed abstracts
7. ShARe corpus: deidentified clinical free-text notes from the MIMIC II database
8. i2b2 challenge corpus (2012): discharge summaries
9. i2b2 challenge dataset (2009): de-identified hospital discharge summaries
10. i2b2 challenge corpus (2008): discharge summaries of overweight and diabetic patients
11. i2b2 challenge corpus (2014): longitudinally ordered clinical notes from three cohorts of diabetic patients

Table 4.4: Included publications and their evaluation methodologies

Author	Year	Ref. std.	Validation	External	Generalizability *	Ref
Afshar	2019	Existing EHR data	Hold-out (train, test, dev.)	No	No, validation is needed	[96]
Alnazzawi	2016	Existing annotated corpus	External	ShARe/CLEF, NCBI disease, Heart failure and pulmonary embolism corpora	Yes, achieves competitive performance on other corpora	[97]
Atutxa	2018	Manual retrospective review	Hold-out (train, test, dev.)	No	Yes, easily portable to other languages	[98]
Barrett	2013	Manual annotations	10-fold cross validation	Multiple datasets (different provider)	Yes, expect that it is generalizable	[99]
Becker	2016	Existing annotated corpus	Not used	No	⍰	[100]
Becker	2019	Manual annotations	Hold-out (train, test, dev.)	No	⍰	[101]
Bejan	2015	Manual annotations	External	i2b2 data (2010)	Yes, good performance on the i2b2 dataset, even though not optimized on it	[102]
Castro	2010	Manual annotations	Not used	No	⍰	[103]
Catling	2018	Existing annotated corpus	Hold-out (train, test, dev.)	No	⍰	[104]
Chapman	2004	Manual annotations	Not used	No	Yes, generalizable to other domains within and outside of bio surveillance	[105]
Chen	2016	Manual annotations	10-fold cross validation	No	⍰	[106]
Chiaramello	2016	Manual annotations	Not used	No	⍰	[107]
Chodey	2016	Existing annotated corpus	Hold-out (train, test)	No	⍰	[108]
Chung	2005	Manual annotations	Hold-out (train, test)	Reports from a second hospital	⍰	[109]
Combi	2018	Manual annotations	Not used	No	⍰	[110]
De Bruijn	2011	Existing annotated corpus	15-fold cross validation	No	⍰	[111]
Deisseroth	2019	Manual annotations	Hold-out (train, test)	Data from a second hospital	Yes, it can be immediately incorporated into clinical practice	[112]
Demner-Fushman	2017	Existing annotated corpus	External	Multiple datasets	⍰	[113]
Divita	2014	Manual annotations	Not used	No	⍰	[114]
Duarte	2018	Manual annotations	Hold-out (train, test)	Second dataset	⍰	[115]

* As reported by authors, dev.: development ⍰ : Not listed

Continued on next page

Author	Year	Ref. std.	Validation	External	Generalizability *	Ref
Falis	2019	Existing annotated corpus	Hold-out (train, test, dev.)	No	Yes, method is not specific to an ontology, and could be used for a graph of any formation	[116]
Ferrão	2013	Existing EHR data	Hold-out (train, test)	No	?	[117]
Gerbier	2011	Manual annotations	Hold-out (train, test)	No	Yes, it could also serve other types of clinical decision support systems	[118]
Goicoechea Salazar	2013	Manual annotations	Hold-out (train, test)	No	?	[119]
Hamid	2013	Manual annotations	10-fold cross validation	No	Possible, the classifier may be applicable in academic hospital samples	[120]
Hassanzadeh	2016	Existing annotated corpus	Hold-out (train, test)	No	Not applicable	[121]
Helwe	2017	Existing annotated corpus	Hold-out (train, test, dev.)	No	?	[122]
Hersh	2001	Manual annotations	Hold-out (train, test)	No	?	[123]
Hoogendoorn	2015	Existing EHR data	5-fold cross validation	No	?	[124]
Jindal	2013	Existing annotated corpus	Hold-out (train, test)	No	Yes, broad applicability	[125]
Kang	2009	Manual annotations	Hold-out (train, test)	No	Yes, extensible to other languages	[126]
Kersloot	2019	Manual annotations	Hold-out (dev., test)	No	Possible, but external validation is needed	[127]
König	2019	Existing EHR data	Not used	No	Still to be tested	[128]
Li	2015	Manual annotations	10-fold cross validation	No	?	[129]
Li	2019	Existing annotated corpus	Hold-out (train, test, dev.)	No	?	[130]
Lingren	2016	Manual annotations	Hold-out (train, test, dev.)	No	?	[75]
Liu	2019	Manual annotations	Not used	No (but multiple datasets / non-trained)	No, limited because of NYP/CUIMC and Mayo notes.	[131]
Lowe	2009	Manual retrospective review	Hold-out (train, test)	No	Yes, has the potential to index other classes of clinical documents	[132]
Luo	2014	Existing EHR data	10-fold cross validation	No	No, challenging, not currently working on it	[133]
Meystre	2006	Manual retrospective review	Not used	No	?	[134]
Meystre	2010	Existing annotated corpus	Hold-out (train, test)	No	?	[135]
Minard	2011	Existing annotated corpus	Hold-out (train, test, dev.)	No	?	[136]
Mishra	2019	Manual annotations	Not used	No	?	[137]
Nguyen	2018	Existing EHR data	?	No	?	[138]

* As reported by authors, dev.: development ? : Not listed

Continued on next page

Author	Year	Ref. std.	Validation	External	Generalizability *	Ref
Oellrich	2015	Existing annotated corpus	External	Multiple datasets	?	[139]
Patrick	2011	Existing annotated corpus	10-fold cross validation	No	Yes, adaptable to different requirements in clinical information extraction and classification by choosing relevant feature sets	[140]
Pérez	2018	Existing annotated corpus	Hold-out (train, test, dev.)	No	Yes, extensible to different hospital-sections and hospitals	[141]
Reátegui	2018	Existing annotated corpus	Not used	No	?	[142]
Roberts	2011	Existing annotated corpus	Hold-out (train, test)	No	?	[143]
Rousseau	2019	Manual annotations	Not used	No	?	[144]
Savova	2010	Manual annotations	10-fold cross validation	No	Yes, implemented in several applications	[145]
Shivade	2015	Manual annotations	Hold-out (train, test)	No	?	[74]
Shoenbill	2019	Manual annotations	Hold-out (train, test)	No	Yes, can allow further evaluation and improvement in care delivery models and treatment approaches to multiple chronic illnesses	[146]
Sohn	2014	Manual annotations	Hold-out (train, test, dev.)	No	Yes, with adaptions: create flexible mechanism for adaptation process	[147]
Solti	2008	Manual annotations	Hold-out (train, test)	No	?	[148]
Soriano	2019	Manual annotations	?	No	?	[149]
Soysal	2018	Existing annotated corpus	Hold-out (train, test)	No	Yes, can be used to quickly develop customized clinical information extraction pipelines	[150]
Spasić	2015	Manual annotations	Hold-out (train, test)	No	?	[151]
Strauss	2013	Manual annotations	Not used	No	Yes, can be shared between institutions and used to support clinical + epidemiological research	[152]
Sung	2018	Manual annotations	?	No	?	[153]
Tchechmedjiev	2018	Existing annotated corpus	Hold-out (train, test, dev.)	No	Yes, but not universally	[154]
Ternois	2018	Existing EHR data	5-fold cross validation + Hold-out (train, test)	No	?	[155]
Travers	2004	Manual retrospective review	Not used	No	?	[156]
Tulkens	2019	Existing annotated corpus	Hold-out (train, test, dev.)	No	?	[157]
Usui	2018	Manual annotations	Not used	No	?	[158]
Valtchinov	2019	Manual annotations	Not used	No	No	[159]

* As reported by authors, dev.: development ? : Not listed

Continued on next page

Author	Year	Ref. std.	Validation	External	Generalizability *	Ref
Wadia	2018	Manual annotations	Not used	No	?	[160]
Walker	2019	Manual retrospective review	Hold-out (dev., test)	No	Yes, it can be incorporated in institutional data warehouse	[161]
Xie	2019	Existing annotated corpus	Hold-out (train, test, dev.)	No	?	[162]
Xu	2011	Manual annotations	Hold-out (train, test)	No	Yes, generable approach to combine information from heterogeneous data sources in EHRs	[163]
Yadav	2013	Manual annotations	Not used	No	Yes, should be broadly applicate to outcomes of clinical interest	[164]
Yao	2019	Existing annotated corpus	Hold-out (train, test)	No	?	[165]
Zeng	2018	Manual annotations	5-fold cross validation + Hold-out (train, test)	No	Yes, potential to be replicated	[166]
Zhang	2013	Existing annotated corpus	External	Two different sets with same settings	Yes, can be adapted to different semantic categories and text genres	[167]
Zhou	2006	Manual annotations	5-fold cross validation	No	?	[168]
Zhou	2011	Manual retrospective review	Hold-out (train, test)	No	?	[170]
Zhou	2014	Manual annotations	Not used	No	?	[169]

* As reported by authors, dev.: development ? : Not listed

Table 4.5: Characteristics of the included studies

	n	%	
Main objective			
Information extraction	45	58%	[96, 99–103, 105, 107–112, 116, 118, 121, 125–127, 130–133, 135–137, 139, 140, 142, 143, 145–147, 149, 151–154, 156, 157, 161, 162, 167–169]
Information enrichment	9	12%	[97, 98, 106, 115, 117, 119, 123, 134, 148]
Classification	8	10%	[74, 75, 120, 155, 159, 160, 163, 166]
Software development and evaluation	6	7.8%	[104, 113, 114, 128, 150, 170]
Prediction	4	5.2%	[124, 158, 164, 165]
Information comparison	2	2.6%	[129, 144]
Computer-assisted coding	2	2.6%	[122, 138]
Text processing	1	1.3%	[141]
Part of challenge			
i2b2 1	10	13%	[74, 111, 114, 125, 135, 136, 140, 143, 145, 150]
Entire system	8	10%	[74, 111, 125, 135, 136, 140, 143, 145]
Parts of the system	2	2.6%	[114, 150]
SemEval	2	2.6%	[108, 150]
Entire system	1	1.3%	[108]
Parts of the system	1	1.3%	[150]
ShARe/CLEF	1	1.3%	[150]
Parts of the system	1	1.3%	[150]
Dataset: language			
English	60	78%	[74, 75, 96, 97, 99, 102, 104–106, 108–114, 116, 120–123, 125, 127, 129–140, 142–148, 150–153, 156, 157, 159–170]
Spanish	5	6.5%	[98, 103, 119, 141, 149]
French	3	3.9%	[118, 154, 155]
German	3	3.9%	[100, 101, 128]
Italian	2	2.6%	[107, 110]
Portuguese	2	2.6%	[115, 117]
Dutch	1	1.3%	[124]
Japanese	1	1.3%	[158]
Korean	1	1.3%	[126]

1. Informatics for Integrating Biology and the Bedside, 2. Semantic Evaluation, 3. Shared Annotated Resources/Conference and Labs of the Evaluation Forum

Continued on next page

	n	%	
Dataset: Origin			
Data present in institute	55	71%	[75, 96, 98, 99, 101–103, 105–107, 109, 110, 112, 114, 115, 117–120, 123, 124, 126–134, 137, 138, 141, 144–153, 155, 156, 158–161, 163, 164, 166, 168–170]
Existing dataset	25	33%	[74, 97, 100, 102, 104, 108, 111, 113, 116, 121, 122, 125, 131, 135, 136, 139, 140, 142, 143, 150, 154, 157, 162, 165, 167]
Included reference to dataset	21	27%	[74, 97, 102, 104, 108, 111, 113, 116, 121, 122, 125, 131, 139, 142, 143, 150, 154, 157, 162, 165, 167]
Training of algorithm			
Trained	47	61%	[74, 75, 96, 98, 99, 101, 104, 106, 108, 109, 111, 112, 115–126, 129, 130, 132, 133, 135, 136, 140, 141, 143, 145–151, 154, 155, 157, 162, 163, 165, 166]
Not listed	3	3.9%	[97, 168, 170]
Development of algorithm			
Use of development set	16	21%	[75, 96, 98, 101, 104, 116, 122, 127, 130, 136, 141, 147, 154, 157, 161, 162]
Not listed	4	5.2%	[97, 149, 150, 168]
Used NLP system or algorithm			
New NLP system or algorithm	29	38%	[98, 99, 104, 110, 112, 114–119, 122, 124, 126, 135, 140, 141, 147, 149, 150, 152, 155, 156, 158, 161, 162, 167, 168, 170]
New NLP system or algorithm with existing components	25	33%	[75, 96, 101, 106, 108, 109, 111, 113, 125, 127–130, 133, 134, 136, 138, 142, 143, 145, 151, 154, 157, 165, 166]
Existing NLP system or algorithm	23	30%	[74, 97, 100, 102, 103, 105, 107, 120, 121, 123, 131, 132, 137, 139, 144, 146, 148, 153, 160, 163, 164, 169]
Use in practice			
Plans to implement / still under development and testing	12	16%	[98, 100, 118, 123, 129, 133–135, 149, 158, 163, 168]
Implemented in practice	10	13%	[101, 109, 110, 113–115, 145, 150, 154, 170]
Availability of code			
Published algorithm or source code	15	20%	[98, 112–114, 127, 145, 147, 149–152, 154, 157, 164, 165]
Pseudocode in manuscript	3	3.9%	[110, 123, 129]
Planning to publish algorithm or source code	1	1.3%	[99]
Not applicable, used an existing system	20	26%	[74, 97, 100, 102, 103, 105, 107, 120, 121, 131, 132, 137, 139, 144, 146, 148, 153, 160, 163, 169]

Table 4.6: Evaluation methods of the included studies

Description	n	%	
Evaluation: Reference standard			
Manual annotations	40	52%	[74,75,99,101–103,105–107,109,110,112,114,115,118–120,123,126,127,129,131,137,144–149,151–153,158–160,163,164,166,168,169]
Existing annotated corpus	24	31%	[97,100,104,108,111,113,116,121,122,125,130,135,136,139–143,150,154,157,162,165,167]
Existing EHR data	7	9.1%	[96,117,124,128,133,138,155]
Manual retrospective review	6	7.8%	[98,132,134,156,161,170]
Evaluation: Validation			
Hold-out validation	40	52%	[74,75,96,98,101,104,108,109,112,115–119,121–123,125–127,130,132,135,136,141,143,146–148,150,151,154,155,157,161–163,165,166,170]
Cross-validation	12	16%	[99,106,111,120,124,129,133,140,145,155,166,168]
External validation	9	12%	[97,99,102,109,112,113,115,139,167]
Solely external validation	5	6.5%	[97,102,113,139,167]
In addition to another type of validation	4	5.2%	[99,109,112,115]
Not performed or not listed	22	29%	[100,103,105,107,110,114,128,131,134,137,138,142,144,149,152,153,156,158–160,164,169]
Generalizability			
Claimed	23	30%	[97–99,102,105,112,116,118,125,126,132,140,141,145–147,150,152,154,161,163,164,167]
Externally validated	5	6.5%	[97,99,102,112,167]
Comparison			
Compared to other existing algorithms or models	24	31%	[97,102,106,112–114,116,121,125,127,130,131,139,142,147,150,154,157,161,162,165–168]
Tested difference in outcomes for statistical significance	4	5.2%	[102,106,127,130]

Table 4.7: Performance measures used in the included studies

Description	Formula	n	%	
Confusion Matrix	Lists True Positives (TP), True Negatives (TN), False Positives (FP), False Negatives (FN), and the Total (n) amount in a 2 × 2 contingency Table	12	16%	[101,111,114,118,123,125,127,128,151,154,158,160]
	TP: Text annotated with ontology concept when concept is present in reference standard			
	TN: Text not annotated with ontology concept when concept is absent in reference standard			
	FP: Text annotated with ontology concept when concept is absent in reference standard			
	FN: Text not annotated with ontology concept when concept is present in reference standard			
Performance measures				
Recall	$\frac{TP}{FN+TP}$	68	88%	[74,75,96–98,100–121,123–125,127–131,133–140,142–155,157–161,163,166–170]
Precision	$\frac{TP}{FP+TP}$	66	86%	[74,75,96–98,100–103,105–118,120,121,123–125,127–140,142–155,157,158,160,161,163,166–170]
F-score	$2 \bullet \frac{Precision \bullet Recall}{Precision+Recall}$	57	74%	[74,75,97,98,100–103,106–108,111,113–117,119–122,124–130,133–140,142–147,149–151,153–155,157,158,162,163,165–167,169,170]
Accuracy	$\frac{TP+TN}{n}$	11	14%	[97,99,101,108,115,119,134,141,145,159,163]
Specificity	$\frac{TN}{FP+TN}$	6	7.8%	[96,101,152,159,160,163]
AUC	Not applicable	5	6.5%	[96,106,124,162,166]
Kappa	$\frac{p_o - p_e}{1-p_e} = 1 - \frac{1-p_o}{1-p_e}$	3	3.9%	[152,156,164]
Processing time	Not applicable	3	3.9%	[99,114,150]
Negative Predictive Value	$\frac{TN}{FN+TN}$	3	3.9%	[96,152,160]
False Positive Rate	$\frac{FP}{FP+TN}$	1	1.3%	[101]
False Negative Rate	$\frac{FN}{TP+FN}$	1	1.3%	[101]
Information entropy	$-\sum_{i=1}^{n} P_i \log\left(P_i\right)$	1	1.3%	[131]
Mean Reciprocal Rank	$\frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}$	1	1.3%	[141]
Initial annotator agreement	Not applicable	1	1.3%	[146]
Match/no match (%)	Not applicable	1	1.3%	[156]
Overgeneration	$\frac{FP}{TP+FP}$	1	1.3%	[160]
Undergeneration	$\frac{FN}{TP+FN}$	1	1.3%	[135]
Error	$\frac{FN+FP}{TP+FN+FP}$	1	1.3%	[135]
Fallout	$\frac{FP}{TN+FP}$	1	1.3%	[135]
Mean Standard Error	$\frac{1}{n}\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2$	1	1.3%	[124]