21

Clinical Natural Language Processing in Secondary Use of EHR for Research

Sunyang Fu, Andrew Wen, and Hongfang Liu

Abstract

The rapid proliferation and implementation of electronic health record (EHR) systems have reshaped the documentation and management of patient data. This transformation has facilitated and accelerated the secondary use of EHRs for clinical research. A common approach to leveraging EHRs is via manual chart review, a process of reviewing or extracting information for clinical research investigations. As a significant portion of clinical information is represented in textual format, execution of such a human-operated approach is time-consuming, labor-intensive, and non-standardized (Kaur et al., BMC Pulm Med 18:34, 2018; Wang et al., J Biomed Inform 77:34–49, 2018; Gilbert et al., Ann Emerg Med 27:305–308, 1996; Fu et al., AMIA Summits Transl Sci Proc 2020:171, 2020). Clinical natural language processing (NLP) has therefore been adopted to computationally facilitate information retrieval and extraction for clinical research. This chapter describes the foundation of clinical NLP and explains different NLP techniques that can be employed in the context of extracting and transforming narrative information in EHR to support clinical research.

Keywords

Electronic health records · Clinical natural language processing · Symbolic approach Machine learning · Deep learning · Clinical research

Learning Objectives

1. Define the term natural language processing (NLP), and describe its relevance to clinical research.
2. List and describe four different approaches to developing NLP.
3. Describe the importance of a gold standard clinical corpus, and describe the five steps for developing a gold standard clinical corpus.
4. Discuss the benefits of using NLP to facilitate multisite clinical research and national research registries and describe challenges and strategies for deploying existing NLP solutions into different EHR environments.

S. Fu · A. Wen · H. Liu (✉)
Department of Artificial Intelligence and Informatics, Mayo Clinic, Rochester, MN, USA
e-mail: liu.hongfang@mayo.edu

R. L. Richesson et al. (eds.), *Clinical Research Informatics*, Health Informatics,
https://doi.org/10.1007/978-3-031-27173-1_21

The Role of Clinical Natural Language Processing in the Secondary Use of EHR

The Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 provides incentives for the rapid adoption and implementation of electronic health record (EHR) systems across the nation [1]. As a result, the availability of longitudinal and dense EHR data offers an unprecedented opportunity to conduct cost-effective clinical research (patient-oriented research, epidemiological and behavioral studies, or outcomes and health services research) [2]. Since then, there has been a rapid increase of studies reported using EHR data with applications including investigation of patient outcomes [3], disease comorbidities [4], risk stratifications [2], and drug interactions [5].

An EHR is a computerized health record for documenting patient information at care encounters [6]. EHRs can be represented through a variety of different formats such as (1) structured (e.g., demographic information, procedures), (2) semi-structured (e.g., patient provided information), (3) unstructured (e.g., clinical notes, radiology reports, pathology reports, operative reports), and (4) binary files (e.g., medical imaging files). A well-known challenge in EHR-based clinical research is that much of the detailed patient information is embedded within clinical narratives and represented in semi-structured or unstructured formats. A traditional method of screening or extracting information from EHRs for clinical research is manual chart review, a process of reviewing or abstracting information and assembling patient cohorts or data sets for research investigation [7]. As a significant amount of clinical information is represented in textual format, execution of such a human-assisted approach is time-consuming, labor-intensive, and non-standardized [7–10]. Natural language processing (NLP), a subfield of computer science and linguistics, has been leveraged to computationally process and analyze EHRs for clinical research [11, 12]. In the following sections, we exhibit two common use case examples of how clinical NLP techniques are leveraged to support research.

Use Case 1: Information Retrieval for Eligibility Screening or Cohort Identification

Information retrieval (IR) is the process of computationally ranking and acquiring information resources (e.g., patient phenotypic profiles and clinical documents) based on relevant information needs (i.e., queries) from a collection of resources (e.g., patient lists and clinical documents), where NLP techniques can be adopted [11]. Common IR applications in clinical research are eligibility screening (i.e., cohort identification or patient phenotype retrieval [13]), a process of determining a participant's eligibility for enrolling in a study based on pre-defined inclusion and exclusion criteria [14, 15]. In recent years, an increasing number of academic institutions and medical centers have applied the IR technology to their internal EHR data to electronically screen eligible patients for clinical studies. Advanced Text Explorer (ATE) is an example of such an IR system developed by Mayo Clinic. The system leverages Elasticsearch, a distributed full-text search engine that is built on Apache Lucene, to handle large-scale real-time document retrieval tasks [16]. EMERSE is a similar IR system developed by the University of Michigan that leverages Apache Solr, also an Apache Lucene-based search engine, for document indexing [17].

For illustrative purposes, the IR system allows users to input customized queries based on the pre-defined eligibility criteria to search clinical documents for selecting or removing prospective study candidates. Based on the example presented in Fig. 21.1, studies designed to investigate the effect of night shift work on cognitive functioning would need to identify participants with a history of working nightshift. Subsequent queries can be established to search EHRs and identify prospective candidates. Based on the search result, users can decide to continue to improve the search query or conduct a chart review for case validation.

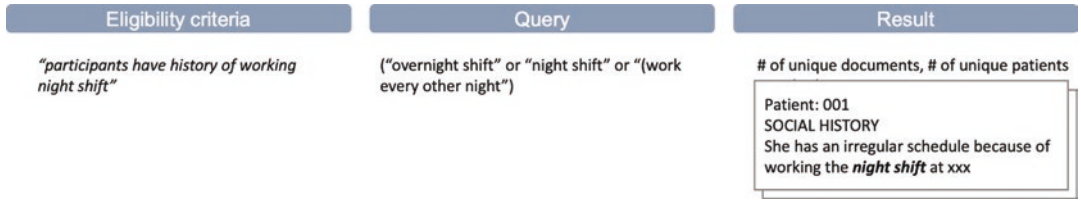

Fig. 21.1 An example of using information retrieval (IR) for cohort identification

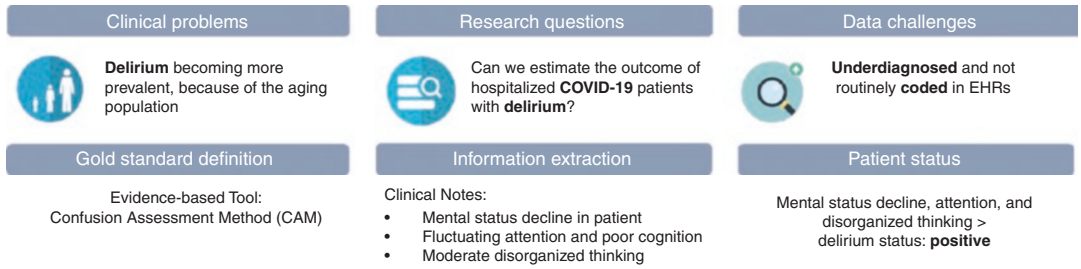

Fig. 21.2 An example of information extraction being used for delirium research

Use Case 2: Information Extraction for Assembling Clinical Research Data Sets

Information extraction (IE) is a sub-task of NLP aiming to automatically extract pre-defined clinical concepts from unstructured text through concept mention detection (i.e., named entity recognition [NER]) and concept normalization (i.e., map the mentions to concepts in standard or pre-defined terminologies) [9, 18–23]. IE can be utilized to assist clinical research by computationally extracting information from clinical documents and assembling a research data set for various research purposes. Common research tasks for clinical IE include case ascertainment [23, 24] and data abstraction [7, 25–28].

For illustrative purposes, a typical clinical IE task is presented in Fig. 21.2. Because delirium is underdiagnosed in clinical practice and is not routinely coded for billing, NLP can serve a distinct role to facilitate case ascertainment. In this particular use case, the goal is to extract cognitive and neuropsychological data elements based on the standard definition to identify patients with delirium from unstructured EHR text [29]. Based on the defined research objectives, the standard definition - confusion assessment method (CAM) is subsequently established by either adopting existing clinical criteria or developing new definitions by domain experts. Corresponding NLP algorithms are created based on these definitions and applied to relevant data sources such as clinical notes. We can then infer a positive status of delirium based on positive status of the extracted concepts "mental status decline," "fluctuating attention," and "disorganized thinking." The generated results can then be used in downstream analytics to help answer specific clinical questions (e.g., how is delirium associated with outcomes in hospitalized COVID-19 patients?).

Foundations of Clinical Natural Language Processing

The steps involved in the development of a gold standard clinical corpus can be divided into five key components: (1) task formulation, (2) corpus annotation (e.g., annotation guideline development, training, and production), (3) model development, (4) model evaluation, and (5) model application (Fig. 21.3) [30–33]. In the ensuing subsections, we will delve into each of these components in further detail.

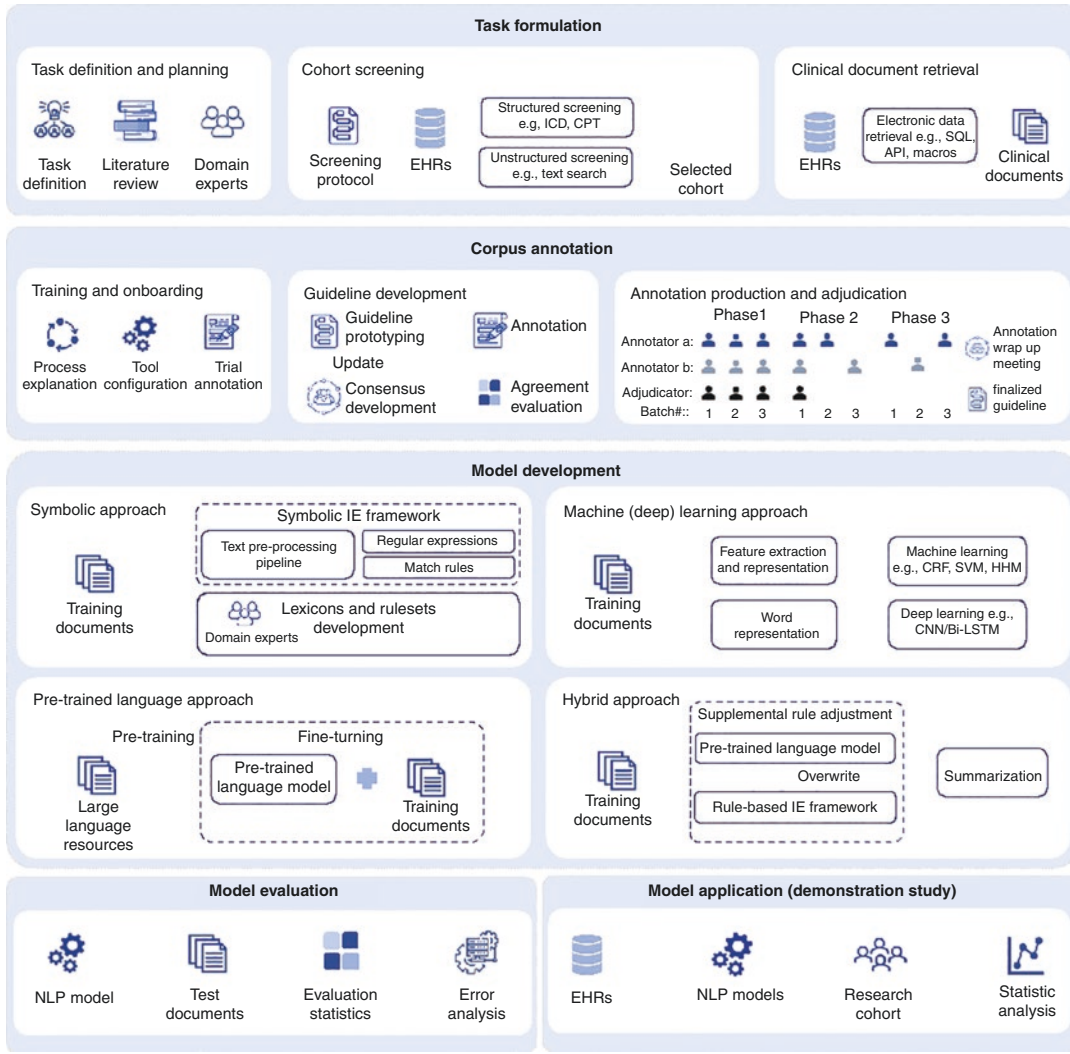

Fig. 21.3 An overview of NLP development and evaluation for clinical research

Task Formulation

Formulation of a clinical NLP task involves defining targets of interest to extract, conducting a literature review, consulting domain experts, and identifying study stakeholders such as annotators with specialized knowledge [34]. Cohort screening is the process of identifying study participants based on eligibility criteria. The initial step is to establish a screening protocol highlighting detailed inclusion and exclusion definitions. These definitions will then be operationalized using EHR data such as patient demographics, procedure codes, diagnosis codes, and problem lists to assemble study cohorts. Based on the established cohort, corresponding clinical documents (e.g., clinical notes) are retrospectively retrieved leveraging APIs (Application Programming Interface) or SQL (Structured Query Language) to query against enterprise data warehouses.

Corpus Annotation

Corpus annotation is the practice of marking predefined clinical or linguistic information to a given document [35]. In general, there are three

phases in the annotation process: (1) training and onboarding, (2) guideline development, and (3) annotation production and adjudication. The initial step starts by assembling an annotation team to identify key stakeholders such as annotators and adjudicators. This step is followed by organizing a preliminary meeting to discuss the overall goal of the study and to walk through the generic annotation process. Training sessions can be hosted to allow annotators to become familiar with the annotation tool and definitions of interest. In the guideline development phase, the process involves the development of a detailed annotation guideline specifying the common standards and definitions for the given task. The steps for developing guidelines can be iterative and commonly involves the following activities: prototyping a baseline guideline, performing annotation, calculating inter-annotator agreement (IAA), organizing consensus meetings, and updating guideline. IAA is often calculated through Cohen's kappa coefficient [36] or F1-score [37]. The process repeats until a satisfactory performance is reached (e.g., a kappa agreement greater or equal to 0.9). Annotation production and adjudication can be organized into a batch-based process for quality control. The production process is similar to guideline development except for allowing more documents to be annotated per batch. Adjudication is the process to resolve inconsistencies between different annotators. There are several ways to perform adjudication. The most common method is to have a third independent domain expert direct overwrite the result or apply majority votes. Team- or panel-based adjudication can be applied for resolving challenging cases. When an independent adjudicator is not available, the two original annotators may reach the final consensus through extensive discussion.

Model Development

Due to the high prevalence and usage of information extraction applications in clinical research, we will primarily focus on IE-related methodologies in this section. Methods for developing IE applications can typically be stratified into symbolic, traditional machine learning (non-deep learning variants), deep learning, or hybrid approaches. The Linguistic String Project-Medical Language Processing (LSP-MLP) project was an early effort aiming to develop clinical IE applications to extract medical concepts from clinical narratives leveraging semantic lexicons (terms) and rules [38, 39]. Since 1990, there has been an increasing number of statistical NLP studies published [12]. Recent advances in computational technologies such as graphics processing units (GPUs) have influenced the adoption of deep learning approaches for clinical IE [40–42]. Through combining both symbolic and machine learning approaches, hybrid approaches have also gained substantial popularity due to the benefits of both comprehensiveness. The following sections provide a methodological overview of each approach.

Symbolic Approach

Symbolic or rule-based approaches use a comprehensive set of lexicons and rules to identify pre-defined patterns in text [43, 44]. This approach has been adopted in many clinical applications due to interpretability and customizability, i.e., the effectiveness of implementing domain-specific knowledge [9] and/or controlled vocabularies [45]. For example, one advantage of the symbolic approach is the ability to leverage existing resources such as clinical criteria, guidelines, medical dictionaries, and knowledge bases. The strategy is to incorporate well-curated clinical knowledge resources such as Unified Medical Language System (UMLS) Metathesaurus [46], Medical Subject Headings (MeSH) [47], and MEDLINE® to facilitate the curation and normalization of lexicons [48]. Based on specific tasks, the combination of rules and well-curated dictionaries can result in promising performance. In addition, to strengthen the ability for capturing important contextual patterns such as family history, negated, possible, and hypothetical sentences, context algorithms are commonly utilized. As an example, NegEx, developed by Chapman et al., is one of the most popular context algorithms used in clinical NLP [49].

The development of lexicons and rules is a manual and iterative process that can be summa-

rized into the following steps: (1) adopting an existing symbolic NLP framework (see section "An Overview of Clinical NLP Systems and Toolkits"), (2) assessing existing knowledge resources, (3) crafting lexicons and rules based on clinical criteria and/or expert opinions, and (4) evaluating and refining lexicons and rules. The refinement of customized lexicons and rules is a recursive process involving multiple subject matter experts. At each iteration, the rules are applied to a reference standard corpus, and its results are evaluated. Based on the evaluation performance, domain experts review false classified mentions or sentences and determine the reasons for misclassification. This pattern was then repeated until it reached a reasonable performance (e.g., F1-score ≥ 0.95).

Traditional Machine Learning

"Traditional" machine learning (i.e., non-deep learning variants) can automatically learn patterns without explicit programming [50–53]. In contrast to deep learning methods, traditional machine learning approaches require more human intervention in the form of feature engineering, a process of selecting and converting raw text into features that can be used in machine learning models. Although feature engineering can be complex, the ability to process and learn from large document corpora greatly reduces the need to manually develop lexicons and rules.

The process of developing traditional machine learning models can be summarized into the following steps: task formulation, data pre-processing, word representation (feature engineering), model training, optimization, and evaluation. In clinical IE, there are two common tasks to be formulated: (1) classification: assign documents or sentences with pre-defined labels; and (2) structured prediction: sequence labeling and segmentation to recognize entities or other semantic units. Commonly reported clinical IE tasks include boundary detection-based classification and sequential labeling. Boundary detection is aimed at detecting the boundaries of the target type of information. For example, the BIO tags use B for beginning, I for inside, and O for outside of a concept. Sequential labeling-based extraction methods transform each sentence into a sequence of tokens with a corresponding property or label. One advantage of sequential labeling is the consideration of the dependencies of the target information. Existing pre-processing steps can be achieved by (1) segmenting documents into sentences, dividing a set of text into individual words (tokenization), and reducing a word to its word stem (stemming). Existing word representation methods for classification tasks include bag-of-words [54–59], continuous bag-of-words (CBOW) [60, 61], or word embedding [62–65] models. Traditional bag-of-words models convert words into a high-dimensional one-hot space, which potentially introduces sparsity, increases the size of data, and removes any sense of semantic similarity between words. Word embeddings can enhance the word semantic encoding by capturing latent syntactic and semantic similarities [66].

Frequently used traditional machine learning models for clinical IE include decision tree (DT) [67], logistic regression (LR) [68], Bayesian network [69], k nearest neighbor (k-NN) [70], random forests [71], hidden Markov model (HMM) [72], support vector machine (SVM) [73], structural support vector machines (SSVMs) [74], and conditional random fields (CRF) [75]. Among the aforementioned models, CRFs and the SVM are the two most popular models for clinical IE [76]. CRFs can be thought of as a generalization of LR for sequential data. SVMs use various kernels to transform data into a more easily discriminative hyperspace. In addition, structural support vector machines is an algorithm that combines the advantages of both CRFs and SVMs [76].

Deep Learning

Deep learning, a subfield of machine learning that focuses on learning patterns from dense representations of a large amount of data, has become an emerging trend in clinical NLP research [42, 77, 78]. In contrast to traditional machine learning approaches, deep learning approaches reduce the need to explicitly engineer data representations. In clinical NLP, the deep

learning algorithms are focused on neural networks or their variants such as convolutional neural networks (CNN) [79–82], recurrent neural networks (RNN) [83–85], gated recurrent unit (GRU) [86], long short-term memory (LSTM) networks [87], and transformers [88].

CNN is a type of artificial neural network (ANN) that relies on convolutional filters to capture spatial relationships in the inputs and pooling layers to minimize computational complexity. Although the models have been found to be exceptionally effective for computer vision tasks, CNN may have a difficult time capturing long-distance relationships in text [89]. RNNs are neural networks that explicitly model connections along a sequence, making RNNs uniquely suited for tasks that require long short-term dependencies to be captured [90, 91]. Conventional RNNs are, however, limited in modeling capability by the length of text due to problems with vanishing gradients. Variants such as LSTM [87] and GRU [86] have been developed to address this issue by separating the propagation of the gradient and control of the propagation through "gates." Meanwhile, many of the researchers have combined deep learning architectures with the CRF framework to further improve the model performance. This is to take advantage of their relative strengths: long-distance modeling of RNNs and CRF's ability to jointly connect output tags. Well-known architectures include CNN-CRF, Bi-LSTM-CRF, and Bi-LSTM-Attention-CRF. More recently, transformer architectures have been proposed to further improve the ability to capture complex dependencies and context. The architecture enables the segmentation of sentences, and adding subsequent layers is therefore needed to allow the model to accommodate long sequences of text without crippling memory constraints [88]. Thus, transformers can effectively model relationships with long word distance and are much more computationally efficient compared to RNN variants. Pre-trained representations based on this architecture such as BERT [40] and GPT [92] have yielded significant improvements in state-of-the-art performance in many NLP tasks [93].

Hybrid

Leveraging the advantages of both rule- and machine learning-based approaches, hybrid approaches combine them into one system potentially offering a comprehensive solution. There are two major hybrid architectures. The first architecture uses a symbolic system to extract features. These features are then will then be used as input for the machine learning system. This architecture may have the potential of achieving improved performance compared with purely symbolic or machine learning-based approach due to the informative features supplied by the symbolic system. As an example, Szarvas et al. applied pattern-based trigger words to improve their NER model for clinical de-identification tasks [94]. The second architecture uses machine learning approaches (or symbolic approaches) to rectify incorrect cases from symbolic approaches (or machine learning approaches). This architecture is also referred to as a "supplemental hybrid approach" or "post-hoc design" [23, 95] and has been leveraged to develop a generic IE framework [96] or to extract specific concept mentions [95, 97].

Model Evaluation

Rigorous model evaluation is crucial for developing valid and reliable clinical IE applications. Evaluation starts by defining the granularity of subjects to be assessed. Common levels of granularity include concept (or mention), sentence, document, and patient. The specific level selected with which evaluation was performed is typically determined based on the specific task or application. Most studies reported using the combination of concept and document-level evaluations [23]. Once the level is defined, the evaluation can then be performed by constructing a confusion matrix or a contingency table to derive error ratios including true positives, false positives, false negatives, and true negatives. From these measures, common evaluation metrics, including sensitivity or recall, specificity, precision or positive predictive value (PPV), negative predictive value (NPV), and F1-score or F-measure, can then be determined based on the error ratios. F1-score that measures the harmony of sensitivity and precision is a well-established metric in

the information retrieval community [37]. In addition, the area under the ROC curve (AUC) and the area under the precision-recall curve (PRAUC) are commonly used for evaluating machine learning models. The designs for evaluation include the hold-out method, where the model is trained on training sets and evaluated on the blinded test set or (nested-) cross-validation (CV), where the prediction error of a model is estimated by iteratively training part of the data and leaving the rest for testing [98, 99].

Model Application

After the evaluation process is finished, the model can be deployed and applied to assemble clinical cohorts or assist in data abstraction in the context of the problem that the model is designed for. The process can be achieved by treating the model as a standalone tool. Corresponding clinical data can be assembled by following the steps highlighted in the section "Task Formulation". A more integrated solution is to deploy the model into the existing data infrastructure or EHR environment. However, the implementation process varies and can be dependent on the maturity of each site's specific infrastructure and policy [100].

A Step-by-Step Case Demonstration

In this section, we present a step-by-step case demonstration for developing two different NLP approaches (symbolic and deep learning) under a case study of aging. Falls are a leading cause of unintentional injury. However, studies have found that the use of billing codes may underestimate true fall events [101]. The case study aims to fully leverage the EHR data and NLP to accurately identify fall events from clinical notes. We supplied additional supporting materials to assist the case demonstration (https://github.com/OHNLP/CRI_Chapter22).

Task Formulation

The task was defined to develop two NLP models (symbolic and pre-trained language approaches) to extract fall-related mentions and sentences from clinical notes at Mayo Clinic Rochester. A literature review was conducted to identify existing methods and dictionaries for adoption [102–107]. Domain experts included in the project are two geriatricians and one palliative care physician. A screening protocol was co-developed by the study team using diagnosis codes. The protocol defines the study participants as Mayo Clinic Biobank patients with age greater or equal to 65 at the time of enrollment. Cases were identified using fall-related ICD-9 and 10 codes: E804, E833–E835, E843, E880–E888, E917.5–E917.8, E929.3, E987, and W00.0XXA-W18.49XS. Controls were matched with age and sex. A total of 300 patients (150 cases and 150 controls) were assembled through an open-source clinical data warehousing research platform i2b2 (Informatics for Integrating Biology & the Bedside) [108] (Fig. 21.4). Clinical notes were subsequently retrieved for these 300 patients directly from the enterprise data warehouse (EDW) using customized SQL.

Corpus Annotation

In this example, the task was formulated as annotating mentions of fall-related expressions in clinical notes. The annotation team is assembled with two trained nurse abstractors as annotators and one geriatrician as the adjudicator. We choose MedTator as the annotation tool. MedTator is a free and serverless annotation tool released under the Apache Software License [109]. To develop an annotation guideline, we first adopt existing definitions from the ANA National database for nursing quality indicators [11]: "An unplanned descent to the floor (or extension of the floor, e.g., trash can or other equipment) with or without injury." Fall events that result from either physiological reasons or environmental reasons are included. Based on this definition, the annotation task can be specified as highlighting both fall-related mentions, indications, and the associated attributes as presented in Table 21.1. Based on the annotation definition, the corresponding annotation schema (.dtd file) is created (Textbox 21.1).

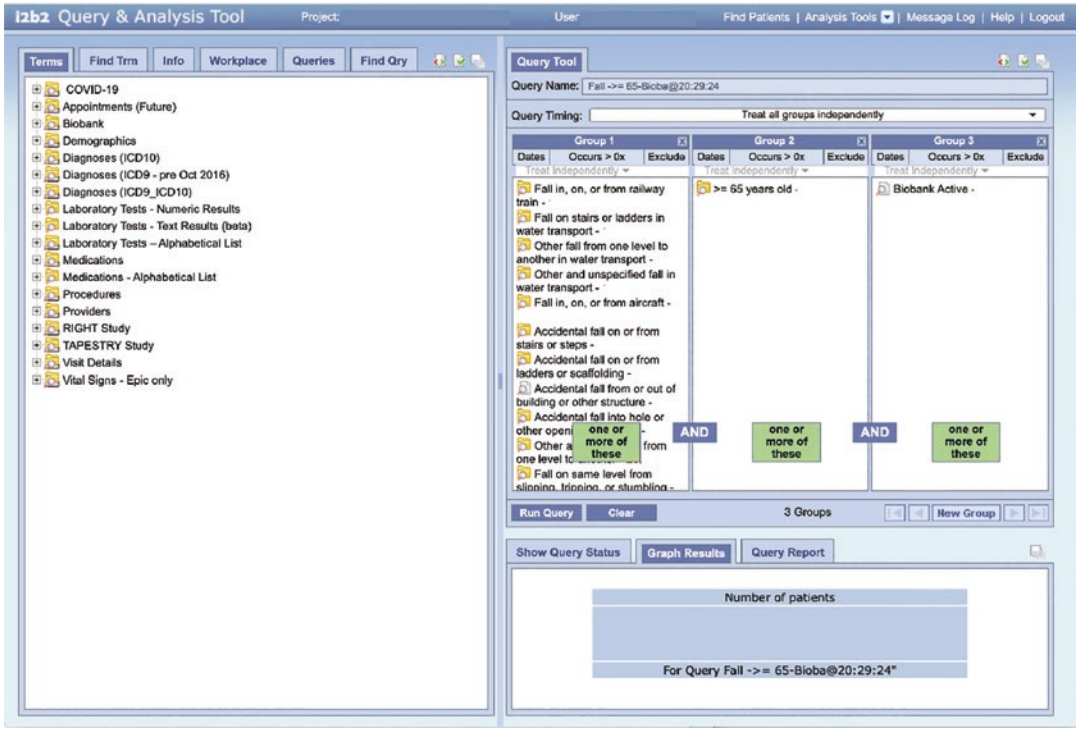

Fig. 21.4 Cohort screening interface based on i2b2

Table 21.1 Fall annotation definition

Concept	Examples	Attribute
Mention	Fall/fell, tripped, slipped, slid	Certainty (negated, possible, hypothetical, confirmed) Status (present, follow-up visit, history)
Indication	Seizure, syncope/fainting, narcolepsy	Experiencer (patient, other) Exclusion (yes, no) (e.g., fall asleep—exclusion: yes)

Textbox 21.1 Example of Fall Annotation Schema in .dtd Format

```
<!ENTITY name "fall_schema_1_1">
<!ELEMENT fall_mention ( #PCDATA ) >
<!ATTLIST fall_mention certainty ( confirmed | hypothetical | possible | negated ) #IMPLIED "confirmed" >
<!ATTLIST fall_mention status ( current | past ) #IMPLIED "" >
<!ATTLIST fall_mention experiencer ( patient | family_member | other ) #IMPLIED "patient" >
<!ATTLIST fall_mention exclusion ( yes | no ) #IMPLIED "no" >
<!ATTLIST fall_mention comment CDATA "" >
<!ELEMENT fall_indication ( #PCDATA ) >
<!ATTLIST fall_indication certainty ( confirmed | hypothetical | possible | negated ) #IMPLIED "confirmed" >
<!ATTLIST fall_indication status ( current | past ) #IMPLIED "" >
<!ATTLIST fall_indication experiencer ( patient | family_member | other ) #IMPLIED "patient" >
<!ATTLIST fall_indication exclusion ( yes | no ) #IMPLIED "no" >
<!ATTLIST fall_indication comment CDATA "" >
```

Once the schema is created, annotation can be performed using the MedTator tool. The tool can be accessed through the URL: https://ohnlp.github.io/MedTator/. After the web interface is opened, the first step is to load the annotation schema. This can be achieved by dragging the .dtd file to the top left (first) box. Similarly, raw clinical documents can be dragged into the second box for annotation. If you don't have a schema or text file yet, you could explore the online sample by clicking the "Sample" button in the top right location.

According to the example presented in Fig. 21.5, "risk of falling" is highlighted as "fall_mention" with certainty as "confirmed," status as "current," patient as "experiencer," and exclusion as "yes." "fall from ladder" is highlighted as a "fall_mention" with certainty as "confirmed," status as "past," patient as "experiencer," and exclusion as "no." During the annotation, the task is usually defined to treat each unique concept independently. It is recommended to choose the smallest possible span that semantically encloses the problem, condition, or diagnosis. Additional annotation best practices can be found at https://github.com/OHNLP/annotation-best-practices.

Model Development

Symbolic Approach

We use the open-source clinical NLP pipeline MedTagger (https://github.com/OHNLP/MedTagger) to develop the symbolic model. First, the initial keywords and regex search patterns based on existing studies [12, 102–107] and domain experts are compiled (Textbox 21.2). These patterns are then applied to the training data. False-positive and false-negative cases are manually reviewed for refinement. This process is repeated after an acceptable performance is reached (e.g., F1-score > 0.95).

Textbox 21.2 Example Keywords and Regex Patterns for Fall Identification

a fall; recurrent fall; time of fall; falls?; fell; fallen; collapsed; slipped; tripped; syncope; falling; syncopal (events?|episodes?|spells?); found (\S+\s+){0,3}on the ground; on (\S+\s+){0,3} way down

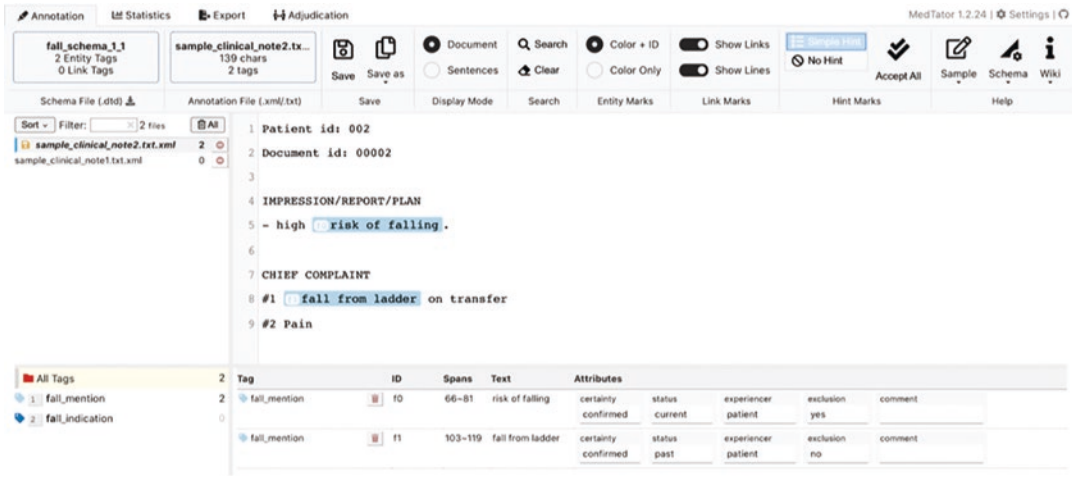

Fig. 21.5 MedTator interface for fall annotation

Deep Learning Approach

We use BERTbase, a pre-trained model with pre-trained sentences on unpublished books and Wikipedia, to perform the sequential sentence classification task. The pre-trained BERT model is adopted from the original Google BERT GitHub repository (https://github.com/google-research/bert). The model contains 768 hidden layers and 12 self-attention heads. For the model fine-tuning, the maximum sequence length (e.g., 512) and batch size (e.g., 32) need to be configured. The early stopping technique is applied to identify the epoch number and prevent overfitting. Sample codes for both approaches can be found at https://github.com/OHNLP/CRI_Chapter22.

Model Evaluation

The models are evaluated on an independent test set based on the mention or sentence level. The presented evaluation results in Fig. 21.6 indicated the model achieve 0.895, 0.9912, 0.770, 0.997, and 0.828 in sensitivity, specificity, PPV, NPV, and F1-score, respectively. The error analysis can be performed by manually reviewing incorrect cases. Through the error analysis, we are able to identify false-negative and false-positive samples for future improvement.

Clinical NLP Resources

An Overview of Clinical NLP Community Challenges

Clinical NLP-related challenges or shared tasks are community activities or competitions with the objective of developing task-specific NLP algorithms within a certain timeline. Solutions will be evaluated using standardized criteria across all participating teams. The top winning team will be awarded small prizes or be invited to disseminate their methods through conference or journal submissions. The challenge starts by calling for participation and releasing the task details. For example, in the 2019 National NLP Clinical Challenge (n2c2) Family History Extraction challenge, the task was to extract mentions of family members in clinical notes and observations (diseases) in the family history. Common timeline for the challenge includes participant registration (e.g., team formulation, data usage agreement), training data release, test data release, submission due, results release, and abstract or manuscript submission. Community challenges have been serving as a vital role in advancing NLP methodologies, disseminating NLP knowledge resources (e.g., annotation guidelines and corpora), engaging informatics

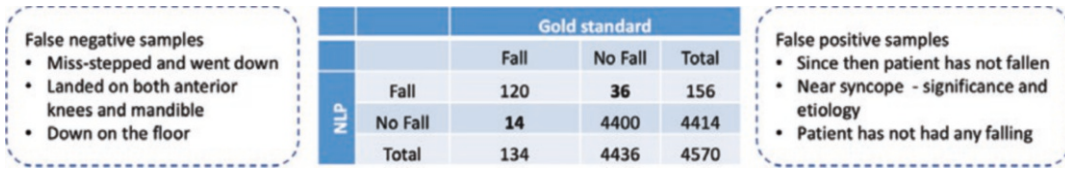

Fig. 21.6 Example of confusion matrix and error cases

researchers, and promoting interdisciplinary collaboration. Furthermore, since the tasks in each challenge are well-defined and standardized by the organizers, coupling with de-identified and made publicly accessible corpora, they are usually regarded as standard benchmarks for the state-of-the-art NLP performance evaluation. Well-known clinical NLP tasks include the Semantic Evaluation (SemEval) challenges [110–112], BioCreative/OHNLP [113–116], the Informatics for Integrating Biology and the Bedside (i2b2) challenges [117–121], the National NLP Clinical Challenge (n2c2) [122], and the Conference and Labs of the Evaluation Forum (CLEF) eHealth challenges [110, 111].

An Overview of Clinical NLP Systems and Toolkits

An Overview of Clinical NLP Systems

NLP systems (frameworks) are important resources for the development, standardization, and streamlined execution of symbolic methods. The key advantage of NLP systems is the built-in and modularized text (pre-)processing pipeline such as sentence detector, tokenizer, part-of-speech tagger, chunking annotator, section detector, information extractor, and context annotator [123, 124]. Different NLP systems have been developed at different institutions, including MedLEE [125], MetaMap [126], KnowledgeMap [127], cTAKES [123], HiTEX [128], CLAMP [129], and MedTagger [124]. MedLEE is one of the earliest clinical NLP systems developed and was originally developed for providing clinical decision support for radiographs. The system has been subsequently expanded for processing different clinical documents such as discharge summaries, pathology reports, and radiology reports [125, 130]. MetaMap, developed by the National Library of Medicine (NLM), is a highly configurable system for providing access and mapping from clinical text to the Unified Medical Language System (UMLS) Metathesaurus [126]. cTAKES is one of the most commonly used tools developed using the Unstructured Information Management Architecture framework (UIMA) [131] and OpenNLP natural language processing toolkit under the Apache project. MedTagger is a resource-driven open-source UIMA-based IE framework developed under the Open Health Natural Language Processing (OHNLP) Consortium aiming to create an interoperable, scalable, and usable NLP ecosystem [124]. Meanwhile, major technology companies have all embraced clinical NLP with commercial solutions available on the market (e.g., IBM Watson [132], Google Healthcare Natural Language API [133], or Amazon Comprehend Medical [134]).

An Overview of Clinical NLP Toolkits and Packages

NLP packages and toolkits are useful resources for developing clinical NLP solutions, especially for text-preprocessing and machine learning approaches. Well-known toolkits include WEKA [135], MALLET [136], OpenNLP [137], SPLAT [138], NLTK [139], and SpaCy [140]. Recently, there has been a rapid growth in the number of open-source deep learning packages (frameworks). Common examples of these packages are Torch [141], Theano [142], MxNet [143], TensorFlow [144], PyTorch [145], Keras [146], and CNTK [147]. Although studies have found variations in the GPU performance and memory management among these libraries [148, 149], most of the packages share similar core competencies, and the selection of appropriate packages can be based on the research environment and user preference.

Challenges, Opportunities, and Future Directions

Despite the notable benefits of leveraging NLP to facilitate clinical research, there remain several open challenges. In this section, we discussed three challenges that need to be investigated in the future including reproducibility and scientific rigor, multisite NLP collaboration, and federated learning and evaluation.

Reproducibility and Scientific Rigor

Considering that many NLP solutions could serve as middleware applications (i.e., supplying research data) for clinical research, the validity of research outcomes for such studies is dependent on the robustness and trustworthiness of the NLP models used as well as the quality of the data being fed into these models [150–152]. Existing clinical NLP applications face challenges in the form of various data quality issues caused by the heterogeneity of the EHR environment. Since EHR systems are primarily designed for patient care and billing, routinely generated and documented clinical information may suffer from potential data quality issues when being used for clinical research. Furthermore, the EHR system itself may have a strong impact on the syntactic and semantic meaning of patient narratives due to its built-in documentation functionality such as smart forms, templates, and macros. Therefore, it is important to have a good understanding of EHR data before the model development and deployment effort. In addition to data quality, reproducibility, which measures the ability to obtain the same (or similar enough) result following the same (or sufficient details) computational steps, is another important criterion for trusted NLP solutions. In the context of clinical NLP, the criterion emphasizes the need for information resource (e.g., corpus, system, and associated research metadata such as inclusion and exclusion criteria used) provenance and process transparency to ensure scientific rigor. Another quality dimension that is commonly referred to as a potential factor of "user trust" and safety is interpretability [153]. In clinical research, the explanations of NLP results may serve as important criteria for the evaluation of the model's capability to explain why a certain decision is made.

Multisite NLP Collaboration

Compared with manual chart review, NLP solutions are distinctive in their ability to systematically extract clinical concepts from clinical text, offering high-throughput solutions for automated data abstraction across multiple different institutions. Therefore, NLP has strong potential to be used to facilitate multisite clinical research collaborations and national-wide research registry development. However, successfully deploying an existing NLP solution to a different EHR environment is nontrivial. We highlight three important NLP dimensions to be considered including implementability, portability, and customizability. Implementability evaluates the feasibility of deploying NLP solutions to the clinical environment. The NLP implementation process is highly dependent on institutional infrastructure, system requirements, data usage agreements, and research and practice objectives. Besides, how NLP models are packaged can also affect the complexity of implementation. For example, whether the NLP solutions can be packaged into a standalone tool or need to be integrated into existing infrastructures would demand different implementation processes [100]. After the deployment, the performance of NLP needs to be re-evaluated in each local environment. Many studies have found that NLP algorithms developed in one institution for a study may not perform well when reused in the same institution or deployed to a different institution or for different studies [154]. The degradation of NLP performance at a different site is often referred to as an NLP portability issue. The differences in EHR systems, care practice, and data documentation standards across institutions may contribute to the variability in clinical documentation and non-optimal performance of NLP systems. To address that, a local evaluation and refinement process can potentially improve the system. The feasibility of system refinement is dependent on the customizability of each system, which measures how easily each model can be adapted, modified, and refined based on existing implementation when a concept definition is changed or there is an update to clinical guidelines. This quality dimension can affect the choices between different NLP approaches (e.g., symbolic vs. machine learning) for multisite studies.

Federated Learning and Evaluation

Another barrier of developing robust and portable NLP solutions is the lack of multisite data due to the regulations, privacy, and security requirements surrounding protected health information (PHI) and the high cost of creating well-annotated and curated clinical corpus [34, 155]. Federated learning, a machine learning approach to train statistical models on remote devices, can be potentially leveraged to address data sharing challenges [156, 157]. The learning can be achieved by allowing individual sites to collaboratively train a model and send incremental updates for immediate aggregation to achieve the shared learning objectives without the need to distribute data [156, 157]. Traditional federated learning is, however, limited only to machine learning approaches. To further enhance the process transparency and model interpretability, the OHNLP Consortium [158] adapt the federated learning approach and proposed a collaborative NLP development framework [159]. The framework contains a user-centric crowdsourcing interface for collaborative ruleset development and a transparent multisite participation workflow on corpus development and evaluation [159]. Site-specific knowledge and findings can therefore be effectively aggregated and synthesized. Another similar concept is federated evaluation, a process of deploying NLP solutions to local institutions, running models on local data, sharing performances to a centralized location (e.g., cloud server). For example, the NLP Sandbox, developed by the National Center for Data to Health (CD2H), is a federated evaluation platform that enables the continuous benchmarking of NLP models on data hosted at different sites through Docker containers. Through this approach, institutional-specific findings and knowledge can be learned and shared without transferring PHI information.

Conclusion

In conclusion, this chapter provided an overview of clinical NLP in the context of the secondary use of EHR for clinical research. A case study of aging was conducted to demonstrate an end-to-end process of NLP development and evaluation. We further discussed three open challenges and highlighted the importance of translational science and community engagement efforts for leveraging clinical NLP applications to support research.

References

1. Jha AK. Meaningful use of electronic health records: the road ahead. JAMA. 2010;304(15):1709–10.
2. McCoy TH Jr, Han L, Pellegrini AM, Tanzi RE, Berretta S, Perlis RH. Stratifying risk for dementia onset using large-scale electronic health record data: a retrospective cohort study. Alzheimers Dement. 2020;16:531.
3. Reis BY, Kohane IS, Mandl KD. Longitudinal histories as predictors of future diagnoses of domestic abuse: modelling study. BMJ. 2009;339:339.
4. Qeadan F, VanSant-Webb E, Tingey B, Rogers TN, Brooks E, Mensah NA, et al. Racial disparities in COVID-19 outcomes exist despite comparable Elixhauser comorbidity indices between blacks, Hispanics, native Americans, and whites. Sci Rep. 2021;11(1):1–11.
5. Zhou M, Zheng C, Xu R. Combining phenome-driven drug-target interaction prediction with patients' electronic health records-based clinical corroboration toward drug discovery. Bioinformatics. 2020;36(Suppl_1):i436–44.
6. Garets D, Davis M. Electronic medical records vs. electronic health records: yes, there is a difference. Policy white paper Chicago, HIMSS Analytics. 2006:1–14.
7. Gilbert EH, Lowenstein SR, Koziol-McLain J, Barta DC, Steiner J. Chart reviews in emergency medicine research: where are the methods? Ann Emerg Med. 1996;27(3):305–8.
8. Kaur H, Sohn S, Wi CI, Ryu E, Park MA, Bachman K, et al. Automated chart review utilizing natural language processing algorithm for asthma predictive index. BMC Pulm Med. 2018;18(1):34.
9. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, Afzal N, et al. Clinical information extraction applications: a literature review. J Biomed Inform. 2018;77:34–49.
10. Fu S, Carlson LA, Peterson KJ, Wang N, Zhou X, Peng S, Jiang J, Wang Y, St Sauver J, Liu H. Natural language processing for the evaluation of methodological standards and best practices of EHR-based clinical research. AMIA Summits Transl Sci Proc. 2020;2020:171–80.
11. Manning C, Raghavan P, Schütze H. Introduction to information retrieval. Nat Lang Eng. 2010;16(1):100–3.
12. Manning CD, Manning CD, Schütze H. Foundations of statistical natural language processing. MIT press; 1999.

13. Chute CG. The horizontal and vertical nature of patient phenotype retrieval: new directions for clinical text processing. In: Proceedings of the AMIA symposium. American Medical Informatics Association; 2002.
14. Weng C, Tu SW, Sim I, Richesson R. Formal representation of eligibility criteria: a literature review. J Biomed Inform. 2010;43(3):451–67.
15. Van Spall HG, Toren A, Kiss A, Fowler RA. Eligibility criteria of randomized controlled trials published in high-impact general medical journals: a systematic sampling review. JAMA. 2007;297(11):1233–40.
16. Kaggal VC, Elayavilli RK, Mehrabi S, Pankratz JJ, Sohn S, Wang Y, et al. Toward a learning health-care system–knowledge delivery at the point of care empowered by big data and NLP. Biomed Inform Insights. 2016;8:BII.S37977.
17. Hanauer DA, Mei Q, Law J, Khanna R, Zheng K. Supporting information retrieval from electronic health records: a report of University of Michigan's nine-year experience in developing and using the Electronic Medical Record Search Engine (EMERSE). J Biomed Inform. 2015;55:290–300.
18. Cowie J, Wilks Y. Information extraction. In: Handbook of natural language processing, vol. 56; 2000. p. 57.
19. Nadeau D, Sekine S. A survey of named entity recognition and classification. Lingvisticae Investigationes. 2007;30(1):3–26.
20. Marsh E, Perzanowski D. MUC-7 evaluation of IE technology: overview of results. In: Seventh message understanding conference (MUC-7): proceedings of a conference held in Fairfax, Virginia, Apr 29–May 1, 1998.
21. Torii M, Wagholikar K, Liu H. Using machine learning for concept extraction on clinical documents from multiple data sources. J Am Med Inform Assoc. 2011;18(5):580–7.
22. Si Y, Wang J, Xu H, Roberts K. Enhancing clinical concept extraction with contextual embeddings. J Am Med Inform Assoc. 2019;26(11):1297–304.
23. Fu S, Chen D, He H, Liu S, Moon S, Peterson KJ, et al. Clinical concept extraction: a methodology review. J Biomed Inform. 2020;109:103526.
24. Kent DM, Leung LY, Zhou Y, Luetmer PH, Kallmes DF, Nelson J, et al. Association of silent cerebrovascular disease identified using natural language processing and future ischemic stroke. Neurology. 2021;97(13):e1313–21.
25. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, et al. Use of natural language processing algorithms to identify common data elements in operative notes for total hip arthroplasty. J Bone Joint Surg Am. 2019;101(21):1931.
26. Fu S, Wyles CC, Osmon DR, Carvour ML, Sagheb E, Ramazanian T, et al. Automated detection of periprosthetic joint infections and data elements using natural language processing. J Arthroplast. 2021;36(2):688–92.
27. Lott JP, Boudreau DM, Barnhill RL, Weinstock MA, Knopp E, Piepkorn MW, et al. Population-based analysis of histologically confirmed melanocytic proliferations using natural language processing. JAMA Dermatol. 2018;154(1):24–9.
28. Hylan TR, Von Korff M, Saunders K, Masters E, Palmer RE, Carrell D, et al. Automated prediction of risk for problem opioid use in a primary care setting. J Pain. 2015;16(4):380–7.
29. Fu S, Lopes GS, Pagali SR, Thorsteinsdottir B, LeBrasseur NK, Wen A, et al. Ascertainment of delirium status using natural language processing from electronic health records. J Gerontol A. 2022;77(3):524–30.
30. Developing a framework for detecting asthma endotypes from electronic health records. Am J Respir Crit Care Med. In: 2014 Conference American Thoracic Society International Conference, ATS 2014, San Diego, CA, p 189
31. Fu S, Leung LY, Wang Y, Raulli A-O, Kallmes DF, Kinsman KA, et al. Natural language processing for the identification of silent brain infarcts from neuroimaging reports. JMIR Med Inform. 2019;7(2):e12109.
32. Chase HS, Mitrani LR, Lu GG, Fulgieri DJ. Early recognition of multiple sclerosis using natural language processing of the electronic health record. BMC Med Inform Decis Mak. 2017;17(1):24.
33. Wu ST, Wi CI, Sohn S, Liu H, Juhn YJ. Staggered NLP-assisted refinement for clinical annotations of chronic disease events. In: 10th International conference on language resources and evaluation, LREC 2016. European Language Resources Association (ELRA); 2016.
34. Fu S, Leung LY, Raulli A-O, Kallmes DF, Kinsman KA, Nelson KB, et al. Assessment of the impact of EHR heterogeneity for clinical research through a case study of silent brain infarction. BMC Med Inform Decis Mak. 2020;20:1–12.
35. Leech G. Corpus annotation schemes. Literary Linguist Comput. 1993;8(4):275–81.
36. Cohen J. A coefficient of agreement for nominal scales. Educ Psychol Meas. 1960;20(1):37–46.
37. Van Rijsbergen CJ. The geometry of information retrieval. Cambridge University Press; 2004.
38. Sager N. Natural language information processing: a computer grammmar of english and its applications. Addison-Wesley Longman Publishing Co., Inc.; 1981.
39. Sager N, Friedman C, Lyman MS. Medical language processing: computer management of narrative data. Addison-Wesley Longman Publishing Co., Inc.; 1987.
40. Devlin J, Chang M-W, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding. 2018. arXiv preprint arXiv:181004805.
41. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018;25(10):1419–28.
42. Wu S, Roberts K, Datta S, Du J, Ji Z, Si Y, et al. Deep learning in clinical natural language process-

ing: a methodical review. J Am Med Inform Assoc. 2019;27:457.
43. Childs LC, Enelow R, Simonsen L, Heintzelman NH, Kowalski KM, Taylor RJ. Description of a rule-based system for the i2b2 challenge in natural language processing for clinical data. J Am Med Inform Assoc. 2009;16(4):571–5.
44. Clancey WJ. The epistemology of a rule-based expert system—a framework for explanation. Artif Intell. 1983;20(3):215–51.
45. Cimino JJ. Desiderata for controlled medical vocabularies in the twenty-first century. Methods Inf Med. 1998;37(4/5):394–403.
46. Bodenreider O. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 2004;32(Suppl_1):D267–70.
47. Lipscomb CE. Medical subject headings (MeSH). Bull Med Libr Assoc. 2000;88(3):265.
48. Carrell DS, Schoen RE, Leffler DA, Morris M, Rose S, Baer A, et al. Challenges in adapting existing clinical natural language processing systems to multiple, diverse health care settings. J Am Med Inform Assoc. 2017;24(5):986–91.
49. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A simple algorithm for identifying negated findings and diseases in discharge summaries. J Biomed Inform. 2001;34(5):301–10.
50. Sebastiani F. Machine learning in automated text categorization. ACM Comput Surv (CSUR). 2002;34(1):1–47.
51. Freitag D. Machine learning for information extraction in informal domains. Mach Learn. 2000;39(2–3):169–202.
52. Alpaydin E. Introduction to machine learning. MIT Press; 2009.
53. Hastie T, Tibshirani R, Friedman J, Franklin J. The elements of statistical learning: data mining, inference and prediction. Math Intell. 2005;27(2):83–5.
54. Doan S, Xu H. Recognizing medication related entities in hospital discharge summaries using support vector machine. Proc Int Conf Comput Ling. 2010;2010:259–66.
55. Hoogendoorn M, Szolovits P, Moons LMG, Numans ME. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. Artif Intell Med. 2016;69:53–61.
56. Sarker A, Gonzalez G. Portable automatic text classification for adverse drug reaction detection via multi-corpus training. J Biomed Inform. 2015;53:196–207.
57. Sohn S, Larson DW, Habermann EB, Naessens JM, Alabbad JY, Liu H. Detection of clinically important colorectal surgical site infection using Bayesian network. J Surg Res. 2017;209:168–73.
58. Rochefort CM, Verma AD, Eguale T, Lee TC, Buckeridge DL. A novel method of adverse event detection can accurately identify venous thromboembolisms (VTEs) from narrative electronic health record data. J Am Med Inform Assoc. 2014;22(1):155–65.
59. Gaebel J, Kolter T, Arlt F, Denecke K. Extraction of adverse events from clinical documents to support decision making using semantic preprocessing. Stud Health Technol Inform. 2015;216:1030.
60. Pandey C, Ibrahim Z, Wu H, Iqbal E, Dobson R. Improving RNN with attention and embedding for adverse drug reactions. In: 7th International conference on digital health, DH 2017. Association for Computing Machinery; 2017.
61. Liu Z, Yang M, Wang X, Chen Q, Tang B, Wang Z, et al. Entity recognition from clinical texts via recurrent neural network. BMC Med Inform Decis Mak. 2017;17(Suppl 2):67.
62. Liu Z, Tang B, Wang X, Chen Q. De-identification of clinical notes via recurrent neural network and conditional random field. J Biomed Inform. 2017;75S:S34–42.
63. Luu TM, Phan R, Davey R, Chetty G. A multilevel NER framework for automatic clinical name entity recognition. In: 17th IEEE international conference on data mining workshops, ICDMW 2017. IEEE Computer Society; 2017.
64. Tran T, Kavuluru R. Predicting mental conditions based on "history of present illness" in psychiatric notes with deep neural networks. J Biomed Inform. 2017;75S:S138–S48.
65. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, Welt J, et al. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. PLoS One. 2018;13(2):e0192360.
66. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems; 2013.
67. Quinlan JR. Induction of decision trees. Mach Learn. 1986;1(1):81–106.
68. Kleinbaum DG, Dietz K, Gail M, Klein M, Klein M. Logistic regression. Springer; 2002.
69. Pearl J. Bayesian networks: a model cf self-activated memory for evidential reasoning. In: Proceedings of the 7th conference of the Cognitive Science Society. Irvine, CA: University of California; 1985.
70. Fix E, Hodges JL. Discriminatory analysis. Nonparametric discrimination: consistency properties. Int Stat Rev/Revue Internationale de Statistique. 1989;57(3):238–47.
71. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32.
72. Baum LE, Petrie T. Statistical inference for probabilistic functions of finite state Markov chains. Ann Math Stat. 1966;37(6):1554–63.
73. Cortes C, Vapnik V. Support-vector networks. Mach Learn. 1995;20(3):273–97.
74. Tsochantaridis I, Joachims T, Hofmann T, Altun Y. Large margin methods for structured and interdependent output variables. J Mach Learn Res. 2005;6:1453–84.
75. Lafferty J, McCallum A, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data. 2001.

76. Tang B, Cao H, Wu Y, Jiang M, Xu H. Recognizing clinical entities in hospital discharge summaries using Structural Support Vector Machines with word representation features. BMC Med Inform Decis Mak. 2013;13:S1.
77. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436.
78. Chen D, Liu S, Kingsbury P, Sohn S, Storlie CB, Habermann EB, et al. Deep learning and alternative learning strategies for retrospective real-world clinical data. NPJ Digit Med. 2019;2(1):43.
79. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998;86(11):2278–324.
80. Chen H, Lin Z, Ding G, Lou J, Zhang Y, Karlsson B. GRN: gated relation network to enhance convolutional neural network for named entity recognition. Proc AAAI. 2019;33:6236.
81. Tan LK, Liew YM, Lim E, McLaughlin RA. Convolutional neural network regression for short-axis left ventricle segmentation in cardiac cine MR sequences. Med Image Anal. 2017;39:78–86.
82. Rios A, Kavuluru R. Convolutional neural networks for biomedical text classification: application in indexing biomedical articles. In: Proceedings of the 6th ACM conference on bioinformatics, computational biology and health informatics. Atlanta, GA: ACM; 2015.
83. Rumelhart DE, Hinton GE, Williams R. Learning representations by back-propagating errors. Nature. 1986;323(6088):533–6.
84. Cocos A, Fiks AG, Masino AJ. Deep learning for pharmacovigilance: recurrent neural network architectures for labeling adverse drug reactions in twitter posts. J Am Med Inform Assoc. 2017;24(4):813–21.
85. Jauregi Unanue I, Zare Borzeshi E, Piccardi M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. J Biomed Inform. 2017;76:102–9.
86. Cho K, Van Merriënboer B, Gulcehre C, Bahdanau D, Bougares F, Schwenk H, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation. 2014.
87. Hochreiter S, Schmidhuber J. Long short-term memory. Neural Comput. 1997;9(8):1735–80.
88. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Advances in neural information processing systems; 2017.
89. Zhang D, Wang D. Relation classification via recurrent neural network. 2015.
90. Hochreiter S. The vanishing gradient problem during learning recurrent neural nets and problem solutions. Int J Uncertainty Fuzziness Knowl Based Syst. 1998;6(2):107–16.
91. Chung J, Gulcehre C, Cho K, Bengio Y. Gated feedback recurrent neural networks. International conference on machine learning. 2015.
92. Radford A, Narasimhan K, Salimans T, Sutskever I. Improving language understanding by generative pre-training. 2018. https://s3-us-west-2amazonaws.com/openai-assets/researchcovers/languageunsupervised/languageunderstandingpaper.pdf.
93. Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. 2019. arXiv preprint arXiv:190605474.
94. Szarvas G, Farkas R, Busa-Fekete R. State-of-the-art anonymization of medical records using an iterative machine learning framework. J Am Med Inform Assoc. 2007;14(5):574–80.
95. Fu S, Thorsteinsdottir B, Zhang X, Lopes GS, Pagali SR, LeBrasseur NK, et al. A hybrid model to identify fall occurrence from electronic health records. Int J Med Inform. 2022;162:104736.
96. Zheng S, Lu JJ, Ghasemzadeh N, Hayek SS, Quyyumi AA, Wang F. Effective information extraction framework for heterogeneous clinical reports using online machine learning and controlled vocabularies. JMIR Med Inform. 2017;5(2):e7235.
97. Meystre SM, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Congestive heart failure information extraction framework for automated treatment performance measures assessment. J Am Med Inform Assoc. 2017;24(e1):e40–e6.
98. Simon R, Radmacher MD, Dobbin K, McShane LM. Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification. J Natl Cancer Inst. 2003;95(1):14–8.
99. Varma S, Simon R. Bias in error estimation when using cross-validation for model selection. BMC Bioinformatics. 2006;7(1):91.
100. Wen A, Fu S, Moon S, El Wazir M, Rosenbaum A, Kaggal VC, et al. Desiderata for delivering NLP to accelerate healthcare AI advancement and a Mayo Clinic NLP-as-a-service implementation. NPJ Digit Med. 2019;2(1):130.
101. Luther SL, McCart JA, Berndt DJ, Hahm B, Finch D, Jarman J, et al. Improving identification of fall-related injuries in ambulatory care using statistical text mining. Am J Public Health. 2015;105(6):1168–73.
102. Tremblay MC, Berndt DJ, Luther SL, Foulis PR, French DD. Identifying fall-related injuries: text mining the electronic medical record. Inf Technol Manag. 2009;10(4):253.
103. Zhu VJ, Walker TD, Warren RW, Jenny PB, Meystre S, Lenert LA. Identifying falls risk screenings not documented with administrative codes using natural language processing. In: AMIA annual symposium proceedings. American Medical Informatics Association; 2017.
104. Patterson BW, Jacobsohn GC, Shah MN, Song Y, Maru A, Venkatesh AK, et al. Development and validation of a pragmatic natural language processing approach to identifying falls in older adults in the emergency department. BMC Med Inform Decis Mak. 2019;19(1):138.
105. McCart JA, Berndt DJ, Jarman J, Finch DK, Luther SL. Finding falls in ambulatory care clinical docu-

ments using statistical text mining. J Am Med Inform Assoc. 2013;20(5):906–14.
106. Toyabe S. Detecting inpatient falls by using natural language processing of electronic medical records. BMC Health Serv Res. 2012;12(448):448.
107. dos Santos HDP, Silva AP, Maciel MCO, Burin HMV, Urbanetto JS, Vieira R. Fall detection in EHR using word embeddings and deep learning. In: 2019 IEEE 19th international conference on bioinformatics and bioengineering (BIBE). IEEE; 2019.
108. Murphy SN, Weber G, Mendis M, Gainer V, Chueh HC, Churchill S, et al. Serving the enterprise and beyond with informatics for integrating biology and the bedside (i2b2). J Am Med Inform Assoc. 2010;17(2):124–30.
109. He H, Fu S, Wang L, Liu S, Wen A, Liu H. MedTator: a serverless annotation tool for corpus development. Bioinformatics. 2022;38:1776.
110. Pradhan S, Elhadad N, Chapman W, Manandhar S, Savova G. Semeval-2014 task 7: analysis of clinical text. In: Proceedings of the 8th international workshop on semantic evaluation (SemEval 2014); 2014.
111. Elhadad N, Pradhan S, Gorman S, Manandhar S, Chapman W, Savova G. SemEval-2015 task 14: analysis of clinical text. In: Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015); 2015.
112. Bethard S, Savova G, Chen W-T, Derczynski L, Pustejovsky J, Verhagen M. Semeval-2016 task 12: clinical tempeval. In: Proceedings of the 10th international workshop on semantic evaluation (SemEval-2016); 2016.
113. Liu S, Wang Y, Liu H. Selected articles from the BioCreative/OHNLP challenge 2018. Springer; 2019.
114. Rastegar-Mojarad M, Liu S, Wang Y, Afzal N, Wang L, Shen F, et al.. BioCreative/OHNLP challenge 2018. In: Proceedings of the 2018 ACM international conference on bioinformatics, computational biology, and health informatics. ACM; 2018.
115. Wang Y, Afzal N, Liu S, Rastegar-Mojarad M, Wang L, Shen F, et al. Overview of the BioCreative/OHNLP challenge 2018 task 2: clinical semantic textual similarity. 2018.
116. Liu S, Mojarad MR, Wang Y, Wang L, Shen F, Fu S, et al. Overview of the BioCreative/OHNLP 2018 family history extraction task. 2018
117. Uzuner Ö, Luo Y, Szolovits P. Evaluating the state-of-the-art in automatic de-identification. J Am Med Inform Assoc. 2007;14(5):550–63.
118. Uzuner Ö, Goldstein I, Luo Y, Kohane I. Identifying patient smoking status from medical discharge records. J Am Med Inform Assoc. 2008;15(1):14–24.
119. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009;16(4):561–70.
120. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514–8.
121. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. Evaluating the state of the art in coreference resolution for electronic medical records. J Am Med Inform Assoc. 2012;19(5):786–91.
122. Stubbs A, Filannino M, Soysal E, Henry S, Uzuner Ö. Cohort selection for clinical trials: n2c2 2018 shared task track 1. J Am Med Inform Assoc. 2019;26(11):1163–71.
123. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, Kipper-Schuler KC, et al. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J Am Med Inform Assoc. 2010;17(5):507–13.
124. Liu H, Bielinski SJ, Sohn S, Murphy S, Wagholikar KB, Jonnalagadda SR, et al. An information extraction framework for cohort identification using electronic health records. AMIA Summits Transl Sci Proc. 2013;2013:149.
125. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. A general natural-language text processor for clinical radiology. J Am Med Inform Assoc. 1994;1(2):161–74.
126. Aronson AR, Lang F-M. An overview of MetaMap: historical perspective and recent advances. J Am Med Inform Assoc. 2010;17(3):229–36.
127. Denny JC, Irani PR, Wehbe FH, Smithers JD, Spickard III A. The KnowledgeMap project: development of a concept-based medical school curriculum database. In: AMIA annual symposium proceedings. American Medical Informatics Association; 2003.
128. Goryachev S, Sordo M, Zeng QT. A suite of natural language processing tools developed for the I2B2 project. In: AMIA annual symposium proceedings. American Medical Informatics Association; 2006.
129. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, Liu H, et al. CLAMP–a toolkit for efficiently building customized clinical natural language processing pipelines. J Am Med Inform Assoc. 2018;25(3):331–6.
130. Bakken S, Hyun S, Friedman C, Johnson S. A comparison of semantic categories of the ISO reference terminology models for nursing and the MedLEE natural language processing system. In: MEDINFO 2004. IOS Press; 2004.
131. Ferrucci D, Lally A. UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat Lang Eng. 2004;10(3–4):327–48.
132. High R. The era of cognitive systems: an inside look at IBM Watson and how it works, vol. 1. IBM Corporation, Redbooks; 2012. p. 16.
133. Cloud G. Using the healthcare natural language API. 2022. Available from: https://cloud.google.com/healthcare-api/docs/how-tos/nlp.
134. Medical AC. Amazon Comprehend Medical—extract information from unstructured medical text accurately and quickly. 2022. Available from: https://aws.amazon.com/comprehend/medical/.

135. Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD Explor Newsl. 2009;11(1):10–8.
136. Mimno D. Machine learning with MALLET. 2004.
137. OpenNLP. Welcome to Apache OpenNLP. 2022. Available from: https://opennlp.apache.org/.
138. Quirk C, Choudhury P, Gao J, Suzuki H, Toutanova K, Gamon M, et al.. MSR SPLAT, a language analysis toolkit. In: Proceedings of NAACL-HLT 2012; 2012.
139. Loper E, Bird S. Nltk: the natural language toolkit. 2002. arXiv preprint cs/0205028.
140. Honnibal M, Montani I. spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. Sentometr Res. 2017;7(1):411–20.
141. Collobert R, Kavukcuoglu K, Farabet C. Torch7: a matlab-like environment for machine learning. In: BigLearn, NIPS workshop; 2011.
142. Bastien F, Lamblin P, Pascanu R, Bergstra J, Goodfellow I, Bergeron A, et al. Theano: new features and speed improvements. 2012. arXiv preprint arXiv:12115590.
143. Chen T, Li M, Li Y, Lin M, Wang N, Wang M, et al. Mxnet: a flexible and efficient machine learning library for heterogeneous distributed systems. 2015. arXiv preprint arXiv:151201274.
144. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: a system for large-scale machine learning. In: 12th USENIX symposium on operating systems design and implementation (OSDI 16); 2016.
145. Paszke A, Gross S, Chintala S, Chanan G. PyTorch: tensors and dynamic neural networks in Python with strong GPU acceleration. 2017;6(3).
146. Chollet F. Keras: the python deep learning library. Astrophysics source code library. 2018:ascl:1806.022.
147. Seide F, Agarwal A. CNTK: Microsoft's open-source deep-learning toolkit. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining; 2016.
148. Yapici MM, Topaloğlu N. Performance comparison of deep learning frameworks. Comput Inform. 2021;1(1):1–11.
149. Elshawi R, Wahab A, Barnawi A, Sakr S. DLBench: a comprehensive experimental evaluation of deep learning frameworks. Clust Comput. 2021;24(3):2017–38.
150. Fu S. TRUST: clinical text retrieval and use towards scientific rigor and transparent process. University of Minnesota; 2021.
151. Fu S, Wen A, Pagali S, Zong N. The implication of latent information quality to the reproducibility of secondary use of electronic health records. Stud Health Technol Inform. 2022;290:173.
152. Fu S, Wen A, Schaeferle GM, Wilson PM. Assessment of data quality variability across two ehr systems through a case study of post-surgical complications. AMIA Annu Symp Proc. 2022;2022:196.
153. Du M, Liu N, Hu XJ. Techniques for interpretable machine learning. Commun ACM. 2019;63(1):68–77.
154. Wagholikar K, Torii M, Jonnalagadda S, Liu H. Feasibility of pooling annotated corpora for clinical concept extraction. AMIA Summits Transl Sci Proc. 2012;2012:38.
155. Chapman WW, Nadkarni PM, Hirschman L, D'avolio LW, Savova GK, Uzuner O. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions. J Am Med Inform Assoc. 2011;18:540.
156. Li T, Sahu AK, Talwalkar A, Smith V. Federated learning: challenges, methods, and future directions. IEEE Signal Process Mag. 2020;37(3):50–60.
157. Li L, Fan Y, Tse M, Lin K-Y. A review of applications in federated learning. Comput Ind Eng. 2020;149:106854.
158. Consortium O. OHNLP Consortium 2022. Available from: http://ohnlp.org/.
159. Liu S, Wen A, Wang L, He H, Fu S, Miller R, et al. An open natural language processing development framework for ehr-based clinical research: a case demonstration using the National COVID Cohort Collaborative (N3C). 2021. arXiv preprint arXiv:211010780.