Modern Clinical Text Mining: A Guide and Review

Bethany Percha[1,2]

[1]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10025; email: bethany.percha@mssm.edu
[2]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA, 10025

Xxxx. Xxx. Xxx. Xxx. YYYYAA

https://doi.org/10.1146/((please add article doi))

Keywords

text mining, natural language processing, electronic health record, clinical text, machine learning

Abstract

Electronic health records (EHRs) are becoming a vital source of data for healthcare quality improvement, research, and operations. However, much of the most valuable information contained in EHRs remains buried in unstructured text. The field of clinical text mining has advanced rapidly in recent years, transitioning from rule-based approaches to machine learning and, more recently, deep learning. With new methods come new challenges, however, especially for those new to the field. This review provides an overview of clinical text mining for those who are encountering it for the first time (e.g. physician researchers, operational analytics teams, machine learning scientists from other domains). While not a comprehensive survey, it describes the state of the art, with a particular focus on new tasks and methods developed over the past few years. It also identifies key barriers between these remarkable technical advances and the practical realities of implementation at health systems and in industry.

 doi:10.20944/preprints202010.0649.v2

Contents

1. INTRODUCTION

Among the most significant barriers to large-scale deployment of electronic health records (EHRs) in quality improvement, operations, and research is the amount of EHR data stored as unstructured text (1). Structured, machine computable data, such as procedure and diagnosis codes, are in the minority. The bulk of information relating clinical findings to decisions, and communicating the logical and deductive processes of medicine, is buried within progress notes, radiology and pathology reports, and other free text documents (2, 3). Examples include:

- Treatment goals and outcomes (e.g. success or failure of treatments, criteria for success, decisions about subsequent treatments)
- Interpretations of radiology and pathology images and laboratory test results
- Social determinants of health (e.g. social connection/isolation, housing issues, mentions of financial resource strain) (4)
- Symptoms, symptom changes, and their interpretation (5)
- Past medical history and family history
- Patient's emotional disposition, mood, and interactions with health providers
- Detailed descriptions of procedures (e.g. labor and delivery, heart catheterization, imaging studies, surgeries)
- Adherence to treatment plans (e.g. medications, physical therapy, procedures)
- Allergies, side effects, and other adverse events
- Results of physical examination (e.g. review of systems and interpretation of findings)
- Patient's reasons for seeing a health provider; primary and secondary complaints
- Psychiatric evaluations and records of therapy sessions

 doi:10.20944/preprints202010.0649.v2

- Discharge summaries and follow-up plans

Some have speculated that modern machine learning algorithms, combined with EHR and other patient data, will enable the convergence of human and machine intelligence in healthcare (6, 7). From a practical standpoint, such a vision hinges on text mining. Without the ability to reliably process and interpret vast quantities of clinical text, all attempts to create high-performance predictive models, phenotyping algorithms, and data-driven treatment strategies ("precision medicine") will face substantial challenges.

For the past several decades, a community of researchers working at the intersection of computer science and medicine has developed strategies for information extraction and modeling of clinical text, using techniques somewhat distinct from those of the broader natural language processing (NLP) research community (8, 9). Their efforts have led to the development of new methods and the production of both commercial (10) and open-source (11) software systems for clinical text mining. In recent years, technology giants like Amazon and Google have also recognized the importance of clinical text mining and joined the fray; Amazon Comprehend Medical (12) now comes packaged as a software add-on to Amazon Web Services, incentivizing storage of EHR data on Amazon's HIPAA-compliant cloud platform by providing seamless clinical text processing. Dedicated clinical text processing companies such as (as of this writing) Clinithink (`clinithink.com`), Linguamatics (`linguamatics.com`), and Apixio (`apixio.com`) have built proprietary systems of their own, promising to improve clinical trial recruitment, disease registry creation, government reporting, and billing, all through improved mining of unstructured clinical text.

As a data scientist with a background in biomedical text mining, I am frequently approached by physician colleagues and academic and industry collaborators who, for various reasons, have found themselves needing to process clinical text. Many perceive clinical text mining to be a "solved" problem, believing that one can simply apply a packaged clinical NLP system to extract structured data for a variety of downstream applications. As a result, I often find myself explaining the limits of current NLP technology and the fact that clinical NLP encompasses many different goals, progress on some of which is further along than others. The purpose of this review, therefore, is to provide a starting point for those who are encountering clinical text mining for the first time. Far from a comprehensive survey, it focuses on a subset of methods and ideas that are particularly clear and generalizable and can serve as starting points for further explorations of the field. Importantly, nothing I discuss here requires access to institution-specific or proprietary software, rule sets, or training corpora. My goal is to provide "outsiders" with a realistic baseline for what it is possible to accomplish with clinical text mining today.

Information Extraction: Often considered a subdomain of NLP, a term referring to any method that extracts structured information, such as concepts, relations, or events, from unstructured text. Examples of information extraction tasks include named entity recognition, concept normalization, and relation extraction.

Electronic Phenotyping: Also called cohort identification, this is the task of identifying patients with certain characteristics of interest (e.g. exposures, diseases, or outcomes), usually from EHR, claims, or other administrative data.

2. A SHORT TAXONOMY OF TASKS AND APPROACHES

2.1. Information Extraction vs. Modeling

Clinical text can play multiple roles in a project, so it is important to start by defining one's overall goal and how the text fits in. For example, electronic phenotyping algorithms (13, 14, 15, 16) often combine clinical notes with structured data, such as diagnosis codes, medication orders, and procedures, to make a prediction about whether a patient has a disease or other phenotype. Here the primary goal of text mining is information extraction: converting the text into a set of structured features that can be combined with other types of features to produce an answer (17). EHR search indexing, knowledge base construction, and patient timeline building are similar in their focus on information extraction.

Equally important, however, are problems where the goal is to make a prediction or inference from the text itself – for example, to classify mammography reports by BI-RADS category (18) or to cluster clinical documents to uncover latent structure (19). This may or may not require a separate information extraction step. For example, methods such as end-to-end, deep learning-based text classification models (20, 21, 22), which produce answers directly from the raw text, often shine in such cases. One important consideration is whether human-interpretable features are necessary for the project or whether the algorithm can be allowed to learn its own representations of text automatically in the course of solving a downstream task (22).

2.2. Rule-Based vs. Statistical Approaches

Clinical NLP systems fall into two broad categories: rule-based and statistical. Rule-based systems codify expert knowledge into a set of structured rules, or templates, which produce structured information when applied to unstructured text. For example, a rule might specify patterns of words, phrases, or parts of speech that signal the presence of a particular type of entity; e.g. "if the word 'received' is followed by a noun followed by 'for' and then a disease name, assume the noun is a drug name". Many of the best-performing clinical NLP systems are rule-based: of 263 clinical text mining articles reviewed by Wang *et al* in 2018 (17), 171 (65%) used rule-based methods. However, rule-based systems have two important disadvantages. First, domain experts must often expend substantial time and effort to construct the rules. Second, because they are domain-specific, they do not generalize well to new problems; a rule-based system for identifying drug names in text will not be good at anything other than identifying drug names in text.

Rule-Based NLP System: Applies a set of expert-defined rules, or templates, to perform an NLP task. Downside is need for expert time and effort.

Statistical NLP System: Learns how to perform an NLP task by applying machine learning algorithms to training data. Downside is need for [often large amounts of] training data.

The alternative is a system built using a statistical learning ("machine learning") algorithm. If provided with some text in which all of the drug names are labeled, for example, the algorithm will try to identify patterns that indicate a particular span of text is a drug name (9, Ch. 8). Learning algorithms themselves are often task-independent, which is one of their key advantages. However, statistical learning algorithms require annotated training data, which in the clinical domain is often limited or nonexistent (23). In addition, privacy concerns often make it impossible to share training data across institutions. As a result, while the NLP community has increasingly turned toward machine learning and away from rules-based approaches, clinical text mining maintains a strong focus on rules (8).

3. SOFTWARE FOR CLINICAL INFORMATION EXTRACTION

The three most common information extraction tasks – named entity recognition (37, 38, 39), concept normalization (40, 41), and relation extraction (Section 7) – are still active areas of research. However, because of the broad need for basic information extraction in applied tasks like medical coding, search, and case finding, software systems have been developed to perform these tasks automatically. This section reviews current state-of-the-art methods and systems and provides examples of the type of output one can expect from each system.

 doi:10.20944/preprints202010.0649.v2

CLINICAL TEXT MINING SOFTWARE & RESOURCES

The following tools are popular choices for general and clinical text processing (e.g. word and sentence tokenization, part-of-speech tagging, chunking, parsing, NER, word and phrase embeddings). The first section contains general-purpose libraries, while the second contains resources specific to clinical text.

Resource	Language	URL	Reference
NLTK Toolkit	Python	`nltk.org`	(24)
Stanford CoreNLP	Java	`stanfordnlp.github.io/CoreNLP`	(25)
Stanza	Python	`stanfordnlp.github.io/stanza`	(26)
spaCy	Python, Cython	`spacy.io`	
scispaCy	Python	`allenai.github.io/scispacy`	(27)
Apache OpenNLP	Java	`opennlp.apache.org`	
CRFSuite	Python	`chokkan.org/software/crfsuite`	
Scikit-learn	Python	`scikit-learn.org` (text preprocessing: `sklearn.feature_extraction.text`)	
Gensim	Python	`radimrehurek.com/gensim/index.html`	(28)
BERT	Python	`github.com/google-research/bert`	(29)
MetaMap	Java	`metamap.nlm.nih.gov`	(30)
MetaMap Lite	Java	`metamap.nlm.nih.gov/MetaMapLite.shtml`	(31)
cTAKES	Java	`ctakes.apache.org`	(11)
Stanza clinical	Python	`stanza.run/bio`	(32)
DNorm	Java, REST API	`ncbi.nlm.nih.gov/research/bionlp/Tools/dnorm`	(33)
Clinical BERT	Python	`github.com/EmilyAlsentzer/clinicalBERT`	(34)
	Python	`github.com/kexinhuang12345/clinicalBERT`	(35)
UMLS	N/A (extraction software in Java)	`nlm.nih.gov/research/umls/index.html`	(36)

3.1. Named Entity Recognition

Named entity recognition (NER) is the task of identifying and locating mentions of conceptual categories, such as drug, symptom, or disease names, in text. It is perhaps the most widely-studied information extraction task, and existing clinical NER systems can already identify a variety of clinically-relevant entities, including problems, tests, and treatments (32, 37, 38), medication and adverse event names (42, 43), and protected health information (PHI) (44, 45). Figure 1 shows the raw output from Stanza (26, 32), a state-of-the-art NER system trained to tag clinical entities using *test*, *problem*, and *treatment* concept annotations from the 2010 i2b2/VA dataset (46).

Named Entity Recognition: The task of identifying and locating mentions of conceptual categories, such as drug, symptom, or disease names, in text.

The simplest NER systems are dictionary-based: they simply compare text strings to a list of terms from a specific category, such as disease names. While these approaches are common and frequently yield acceptable performance on clinical text (47, 48), modern clinical NER systems more commonly employ machine learning models adapted for sequence data, including conditional random fields (CRFs), recurrent neural networks (RNNs), and RNN variants such as long short-term memory networks (LSTMs). For example, Stanza (Figure 1) uses pretrained character-level language models (49) fed into Bi-LSTM-CRF sequence taggers (32, 50, 51). Trained using corpora hand-annotated with entity type(s)

 doi:10.20944/preprints202010.0649.v2

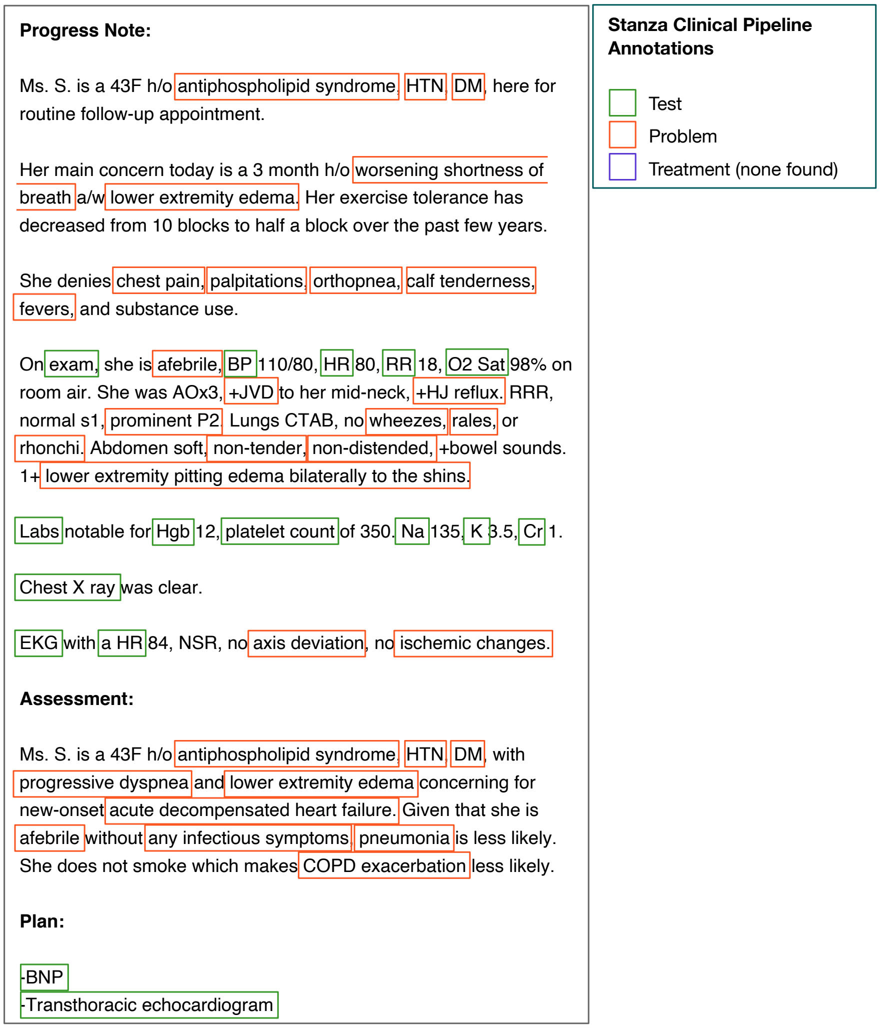

Figure 1

A sample clinical progress note (not a real patient) with named entity annotations provided by the Stanza clinical text processing pipeline, trained using data from the 2010 i2b2/VA challenge. The Stanza pipeline tags three types of named entities: *treatment*, *problem*, and *test*. For this particular note, no treatment entities were found. Medical terms, abbreviations and acronyms: HTN: hypertension; DM: diabetes mellitus; h/o: history of; a/w: along with; AOx3: alert and oriented to person, time, and place; JVD: jugular vein distention; HJ reflux: hepatojugular reflux, distention of jugular vein produced by applying manual pressure to the liver; RRR: regular rate and rhythm (of pulse); S1: heart sound produced by closure of atrioventricular (mitral and tricuspid) valves; P2: heart sound produced by closure of pulmonic valve; CTAB: "clear to auscultation bilaterally", an abbreviation used in lung examinations; pitting edema: if area of swelling pressed, a pit remains; Hgb: hemoglobin, measured in units of g/dL; Cr: creatinine, measured in units of mg/dL; NSR: normal sinus rhythm; BNP: brain natriuretic peptide test (indicative of heart failure).

of interest, these algorithms learn to identify features of a text string and its surrounding context that predict whether it is one of the desired types.

What constitutes a "feature" depends on the system. Traditionally, NER algorithms have selected features from predefined sets, including morphological (capitalization and punctuation patterns, presence/absence/location of numbers, etc.), syntactic (parts of speech, grammatical dependencies, etc.), semantic (membership in a lexicon, position in an ontology, etc.), and other specialized or hand-coded features (52). Some systems incorporate pre-trained word or phrase embeddings (Section 4), and modern neural network-based NER systems often learn higher-order embeddings directly from patterns in the text itself (50, 53).

A major advantage of machine learning-based NER systems is their flexibility. The same system can often learn to tag different types of entities simply by swapping training datasets. For example, while Figure 1 shows "test", "problem", and "treatment" annotations, the Stanza system has also been trained using a corpus of 150 chest CT radiology reports (54) to tag "anatomy", "anatomy modifier", "observation", "observation modifier", and "uncertainty" concepts (32). The reverse is also true: different machine learning algorithms can be trained using the same training data. In fact, when looking for an NER system or any other clinical NLP system, a useful strategy is to identify an annotated corpus for that task and look for papers that have cited the corpus. For example, like Stanza, the Clinical Named Entity Recognition (CliNER) system (38) was trained using concept annotations from the 2010 i2b2/VA NLP challenge (46). Other widely cited systems trained on the same dataset are the Bi-LSTM-CRF systems by Chalapathy *et al* (55) and Unanue *et al* (56) and Tang *et al*'s system combining support vector machines (SVMs) with CRFs (57).

3.2. Domain Specificity and Key Challenges

It is worth pausing here to consider the conceptual and practical challenges illustrated by Figure 1. First, it is clear that NER only makes sense when the entities involved are discrete and have well-defined locations in text. Many clinically important concepts, such as income, housing, and employment history, are unlikely to be described using simple and consistent terminology that can be picked up by NER algorithms. Second, not all of the NER annotations in Figure 1 are correct or meaningful without consideration of the surrounding context. Labeling the term "acute decompensated heart failure" as a *problem*, for example, means little without the qualification that it is a suspected, not definite, diagnosis.

There are also practical concerns regarding NER model training and maintenance. Although dozens of different clinical NER systems have been developed, many are now obsolete, and not all are released as "production-ready" code (i.e. easy to download and use). In addition, if one is interested in an entity class for which no pre-annotated corpus or pre-trained model is available, there is no alternative but to train one's own system; this means either defining a rules-based approach or creating a custom, annotated training set.

Finally, there is the issue of domain specificity. Clinical text is complex, incorporating specialized medical terms, numerical measures and scores, abbreviations (see Figure 1 caption), misspelled words, and poor grammar (20). Many high-quality general domain NER models exist, such as those from the Stanford CoreNLP (25) and spaCy libraries (`https://spacy.io`). However, these are trained using general domain text, such as telephone conversations, newswire, newsgroups, broadcast news, broadcast conversation, and

Table 1 Examples of cTAKES annotations associated with the note in Figure 2. The annotations in the top section are correct mappings, and those in the bottom section are incorrect mappings. There were 122 unique cTAKES annotations for this note.

Line Number	Annotation Type	Original String	Normalized Term	UMLS Concept ID
2	DiseaseDisorderMention	HTN	Hypertensive disease	C0020538
4-5	SignSymptomMention	shortness of breath	Dyspnea	C0013404
7	SignSymptomMention	chest pain (negated)	Chest pain (negated)	C0008031
10	SignSymptomMention	JVD	Jugular venous engorgement	C0425687
16	ProcedureMention	EKG	Electrocardiography	C1623258
24	ProcedureMention	Transthoracic echocardiogram	Transthoracic echocardiography	C0430462
10	DiseaseDisorderMention	reflux	Gastroesophageal reflux disease	C0017168
11	MedicationMention	CTAB	Cetrimonium bromide	C0951233
23	MedicationMention	BNP	Nesiritide	C0054015

blogs. As such, they tag somewhat generic entities like *person*, *number*, and *place* names, which may or may not be relevant in a clinical text mining context. A second class of systems are those that have been trained using biomedical text, usually from PubMed research articles and abstracts. Recent examples are the scispaCy library (27) and the transformer-based language model BioBERT, fine-tuned for NER (58). These systems often tag entity types that are relevant to clinical text, such as gene names; however, because they were trained using scientific writing, they may suffer reduced accuracy on clinical text. The issue of domain specificity and the need for domain-specific models extends beyond NER, affecting nearly all tasks in clinical text mining.

Concept Normalization: The task of assigning a unique identity to an entity name recognized in the text. In the biomedical domain, this typically involves mapping the name to a known concept from a structured terminology or ontology.

3.3. Concept Normalization

The output of a clinical NER system (Figure 1) is a set of named entities of one or more types. The obvious downside to such output is that it tells one nothing about the entities except their type(s); for example, there is no way of knowing that the strings "HTN" and "hypertension" – even if they are in the same note and both labeled as *problems* – refer to the same concept. Likewise, although an NER system may recognize multi-word phrases (e.g. "lower extremity pitting edema bilaterally to the shins", Line 13, Figure 1), it does not understand how the component words contribute to the meaning of each phrase, and it cannot easily connect a given phrase to coreferent phrases, even from the same passage (e.g. "lower extremity edema", Line 5, Figure 1).

Concept normalization, also known as *entity linking*, is the task of assigning a unique identity to each entity name mentioned in text. In the clinical domain, this typically involves mapping each entity name to a known concept from a structured terminology or ontology. The task is closely related to NER and indeed, systems often combine the two processes (59). Coreference resolution, in which strings referring to the same entity (e.g. a pronoun and its antecedent) are grouped, is a similar task; it is essentially normalization without the ontology mapping step (60, Ch. 22).

Clinical text is incredibly diverse (61), and practitioners from different medical specialties, or who have been trained at different institutions, will often choose different terms for

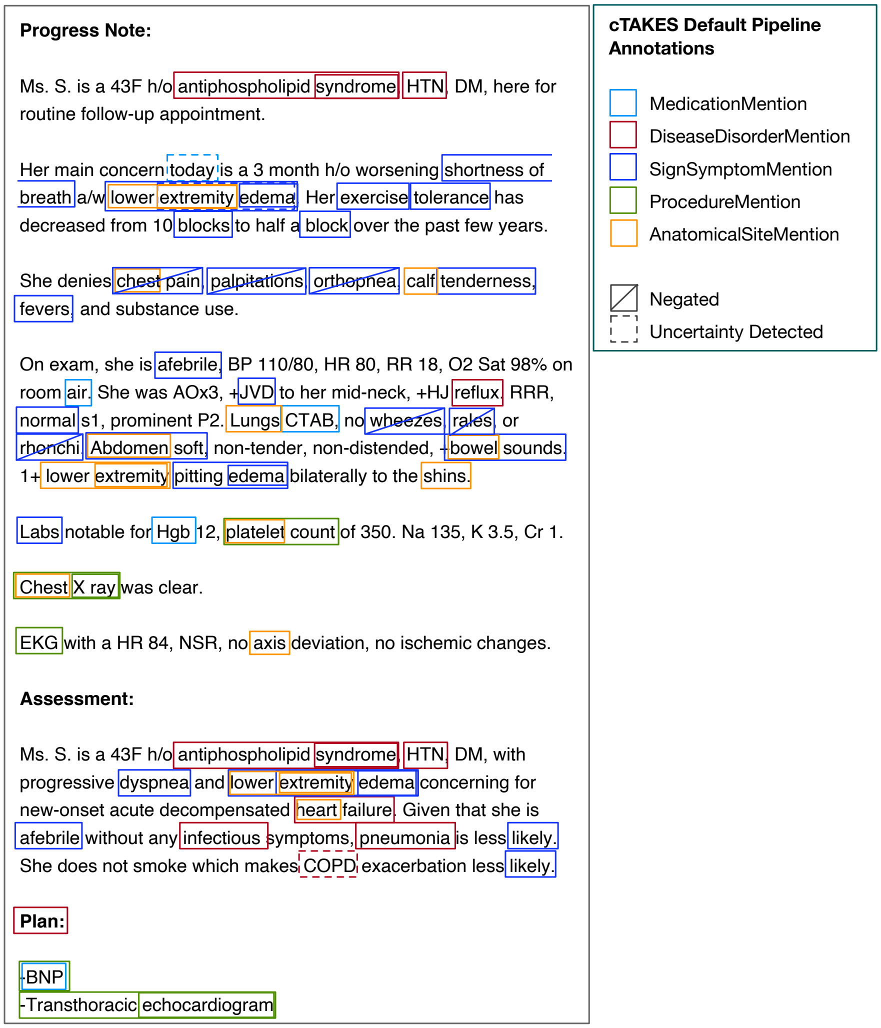

Figure 2

The same clinical progress note as in Figure 1, with annotations provided by the cTAKES (version 4.0) default pipeline. The abbreviations are the same as in Figure 1. The cTAKES pipeline detects negation and uncertainty and maps each entity to its corresponding concept in UMLS. A selection of the UMLS concepts found in this note is in Table 1.

the same concept. The Unified Medical Language System (UMLS), a project begun in 1986 at the National Library of Medicine, was designed to address this issue (36). UMLS is a compendium of biomedical ontologies and terminologies in which concepts occurring across multiple resources are mapped to a single unique identifer (a “Concept Unique Identifer”, or CUI). Today, the predominant strategy for clinical concept normalization is to map a given text string to one of these CUIs. End-to-end clinical text mining systems like MedLEE (now Health Fidelity; (10)), MetaMap (30), MetaMap Lite (31), and cTAKES (11), all have this functionality. The CLAMP system (62) provides an easy-to-use graphical user interface for

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2021 doi:10.20944/preprints202010.0649.v2

N-Gram: A contiguous sequence of N items in text. In NLP, the term "N-Gram" most often refers to a sequence of N words, but it can also refer to sequences of characters, syllables, etc.

Negation Detection: The task of identifying whether a term or concept is negated in the text. Simple pattern-based algorithms, such as NegEx (63), often suffice.

Coreference Resolution: The task of grouping strings from a passage that refer to the same entity, such as a pronoun and its antecedent.

building and deploying clinical NLP pipelines, including UMLS mapping.

The same note analyzed by Stanza in Figure 1 is shown in Figure 2, this time with annotations produced by cTAKES, a popular system developed at the Mayo Clinic (11). A selection of the 122 detailed UMLS mappings produced by cTAKES is in Table 1. In addition to UMLS-based concept normalization, cTAKES detects negation (63), uncertainty, and experiencer (whether the statement refers to the patient or, e.g. a family member). The results shown in Figure 2 are from the default cTAKES pipeline, i.e. what one could expect running cTAKES "out of the box". Most of the annotations are correct; for example, cTAKES correctly maps the string "HTN" to the normalized concept *Hypertensive disease* (CUI C0020538) and understands that "shortness of breath" is a synonym for *Dyspnea* (CUI C0013404). A key shortcoming, however, is cTAKES' reliance on dictionary-based lookups to identify and normalize named entities. This is apparent in Figure 2, where cTAKES labels the strings "CTAB", "BNP", and "Hgb" as medications because of spurious UMLS mappings (e.g. "CTAB" maps to "cetrimonium bromide" in UMLS). If the specificity of extracted medication terms were crucial for one's application, therefore, it might make sense to include a dedicated named entity recognition system for medication names in the cTAKES pipeline. In addition, depending on the application, full concept normalization may not be necessary; in one recent study (64), using cTAKES annotations as features in a 30-day readmission model yielded no better performance than N-grams.

Like named entity recognition, clinical concept normalization is still an active area of research. For those interested in this task, a good place to start are the disorder normalization systems built for the SHARe/CLEF eHealth 2013 Evaluation Lab, a community NLP challenge focusing on clinical named entity recognition and concept normalization (65, 41). DNorm (33, 61) was the top-performing system on the concept normalization task, deploying a pairwise learning-to-rank approach that was the first of its kind in the clinical concept normalization literature. More recent studies have applied deep learning models to the same task and dataset (66, 67).

3.4. Numbers, Ranges, and Sections

There are a few information extraction tasks of particular importance to clinical text for which dedicated systems have been developed. These systems are generally rule-based and rely on regular expressions (68). For example, extraction of lab values and vital signs is a distinct task from named entity recognition because it requires interpreting numeric values and ranges. The Valx system (69) extracts and structures lab test comparison statements, though so far it has only been applied to trial descriptions from `ClinicalTrials.gov`. The CNN-based system developed by Xie *et al* (70) identifies blood pressure readings, determines the exactness of the readings, and classifies the readings into three classes: general, treatment, and suggestion. Their machine learning-based workflow could be adapted to extract other types of numeric values.

Section identification is another task somewhat unique to the clinical text mining literature. It involves identifying the section labels associated with each span of text within a note (e.g. *Progress Note*, *Assessment*, and *Plan* in Figure 1), which informs the interpretation of whatever is found there. To date, the only section identification system used outside the institution in which it was developed is the SecTag system by Denny *et al* (71). A complete review of section identification methods and systems can be found in (72).

4. EMBEDDINGS AND PRETRAINING

The core idea behind concept normalization (Section 3.3) is semantic relatedness; two terms can look different, yet refer to the same concept. However, semantic relatedness extends beyond the dichotomy of same vs. different; terms can have degrees of similarity (e.g. "dog" vs. "cat" as opposed to "dog" vs. "volcano") and can be similar in different ways (e.g. "queen" vs. "king" as opposed to "queen" vs. "president"). Modern NLP systems represent this idea mathematically using a construct called an *embedding*.

Embedding: A mathematical representation of a word, phrase, or other piece of text, typically a vector of a fixed length, designed in such a way that pieces of text with similar meaning have similar vectors.

Transfer Learning: The technique of storing knowledge gained while solving one problem for later use on a different, but related, problem. In text mining, the strategy of learning embeddings of words and phrases on large, unlabeled corpora and later incorporating them in task-specific supervised models is a form of transfer learning.

4.1. Word, Phrase, and Character Embeddings

An embedding is a semantically-meaningful mathematical representation of a word, phrase, or other piece of text. Usually a vector, it is designed in such a way that words and phrases with similar meaning have similar vectors. Meaning is difficult to represent using numbers, so embedding methods replace "meaning" with "context" and build vectors to reflect usage patterns, typically within large, unlabeled corpora. The NLP subfield of distributional semantics, which originated with Latent Semantic Analysis in 1988 and reached a milestone with the development of word2vec (73) and GloVe (74) in 2013–2014, is a collection of methods all built around the central goal of creating vector-space embeddings of words and phrases that reflect how they are used in context. To compare the meaning of two words, one simply calculates the cosine similarity of their corresponding vectors.

From a clinical text mining standpoint, embeddings are useful in two ways. First, because they do not require annotated corpora for training, it is easy to create embeddings that are specific to clinical text, or that capture regularities of expression within a particular clinical subfield or institution. These will often outperform general-domain embeddings on clinical text mining tasks (53). Specialized clinical text embeddings have been used to improve clinical NER (75), resolve abbreviations in clinical text (76), expand a structured lexicon of radiology terms (77) and build specialized lexicons from scratch (78). Second, an embedding can incorporate structured information beyond what is found in the text (79), and embeddings have been created to represent CUIs (80), documents (81, 82), or entire patient records (83). Any task in which the notion of similarity is important, particularly when that similarity is based on patterns in text, can probably benefit from embeddings.

For more information about embeddings, readers are encouraged to consult Turney and Pantel (84) for a review of early methods and Kalyan *et al* (85) for a review of embedding methods currently in use in clinical text mining.

4.2. Contextual Embeddings and Pretraining

Until the last few years, embeddings consisted of one vector per entity; that is, one vector per word, phrase, or document. However, novel neural network architectures (22) have permitted the creation of embeddings that vary depending on the context; this has expanded the representational power of embedding methods and led to the creation of massive pretrained language models like BERT (Bidirectional Encoder Representations from Transformers) (29) and GPT-3 (`openai.com`). These models are generally too resource-intensive to be trained from scratch. Instead, a transfer learning approach (86) is used in which models trained on general-domain corpora are either further pre-trained or fine-tuned on clinical text for use in clinical text mining tasks (85). For example, Alsentzer *et al* recently trained BERT models on 2 million notes from the MIMIC-III (87) database. They produced two models, one for generic clinical text and another for discharge summaries, which they released publicly (34). They and others have demonstrated that BERT models fine-tuned on clinical corpora improve the state of the art on clinical NER, de-identification, inference, and concept normalization tasks (88, 89), though in at least one case, UMLS features still contributed valuable additional information (90).

The downside of these models is that they require some technical sophistication to adapt and apply. Whereas the original word2vec could be run on a plain text corpus using a single script and output vectors to a text file, to use BERT requires knowledge of how to "wire up" a pre-trained model to task-specific output layers for fine-tuning. However, it is likely that end-to-end clinical text processing systems, like cTAKES, will begin to incorporate BERT and related methods into different annotation modules as the technology develops.

5. TEXT CLASSIFICATION

Text Classification: The task of assigning a label, or category, to text based on its content. Examples include document classification (e.g. of radiology, pathology, or autopsy reports) and sentence classification.

Text classification is perhaps the most sought-after application of clinical text mining. A recent survey (22) found that of 212 clinical text mining papers employing deep learning methods, 88 (41.5%) focused on text classification; text classification and NER together encompassed 75.5% of articles. The goal of text classification is to classify documents (or sentences, phrases, etc.) into two or more discrete categories. Examples from the clinical domain include classifying primary care descriptions of back pain into acute vs. lower back pain (91), distinguishing normal vs. abnormal knee MRI reports (92), and assessing whether a patient is a current or former smoker vs. a non-smoker based on clinical notes (93). Text classification is a modeling task – typically, it is its own goal. Often it will incorporate features identified through information extraction (Section 3), like named entities or CUIs, or embeddings (Section 4).

A recent systematic review of clinical text classification describes standard text classification algorithms, as well as popular approaches to preprocessing, feature selection, and training set construction (20). An older but still relevant review surveys text classification methods for automated clinical coding (94). In general, text classification methods for clinical text are similar to those for other domains, with the exception that specialized medical resources, such as UMLS, often serve as additional sources of features.

5.1. Feature Construction and Selection

The use of individual words or N-grams as features, while common in text classification (95, Ch. 13), often results in undesirable levels of feature sparsity when applied to clinical text. As a result, feature selection and dimensionality reduction methods are of particular importance in clinical text classification. Feature selection based on TFIDF weighting (95, Ch. 6) is common, as are embeddings (Section 4), which turn a potentially unmanageable number of word and text features into dense representations of fixed dimensionality (96). Concept normalization (Section 3.3) also plays a particularly important role in clinical text classification; it is common to preprocess clinical text with a system like cTAKES or MetaMap to merge different term and phrase variants into the same structured concept, then use those concepts in a classification model (97, 98). It is also possible to exploit parent-child relationships from the UMLS hierarchy to create additional features, e.g. by including all parent terms for a given concept. Such ontology-guided feature engineering has been shown to improve performance on downstream clinical text classification tasks (99).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2021 doi:10.20944/preprints202010.0649.v2

Finally, one can choose a classification algorithm that provides implicit feature selection. In one study, elastic net (100) was used to classify ICU patients into risk strata based on the text of nursing notes. It reduced the number of text features by over a thousandfold while maintaining near-optimal performance (101).

5.2. Deep Learning for Clinical Text Classification

Aside from those that have employed task-specific rules (Section 2.2), the majority of clinical text classification studies to date have used standard supervised machine learning algorithms, including support vector machines, naive Bayes, random forests, and boosting (92, 102, 103). However, over the past five years, deep learning algorithms have begun to displace other classifiers. One of their key advantages is a reduced need for feature engineering; embeddings of words, phrases, and higher-order text structures can be learned as part of the overall training process or incorporated via transfer learning from other pre-trained models. Several studies have deployed convolutional neural networks (CNNs) with high success on a variety of clinical text classification tasks: assigning diagnosis codes (104, 105), classifying radiology reports (21, 106), subtyping diseases (91), and determining the presence or absence of comorbidities (107). Alternative neural network architectures, such as LSTMs and attention networks, are commonly used in text classification tasks in the general NLP domain, although as of this writing, CNNs have been the dominant architecture in clinical text classification (22, 108). One recent paper exemplifies the end-to-end deep learning approach to clinical text classification, tying rule-based features together with word and UMLS-based concept embeddings in a single CNN-based classifier (107).

Weak Supervision: Supervised learning using "weak", or noisy, labels. For example, simple heuristic rules, or "labeling functions", may be used to create large, weakly-annotated training sets.

Distant Supervision: Supervised learning using training signals that do not directly label the training examples. For example, the use of structured clinical data to train a supervised text mining algorithm is a form of distant supervision, since the structured data are generally associated with patients or encounters, not individual text records (i.e. sentences or documents).

6. WEAK AND DISTANT SUPERVISION

As discussed in Section 2.2, machine learning approaches to clinical NLP generally suffer from a lack of training data (23). In addition, existing clinical information extraction and text classification models have generally been trained using the same few annotated datasets (46, 109, 110, 65, 111), which restricts the range and quality of annotations they produce. Most applied clinical text mining projects will therefore confront, at some point, the problem of insufficient or inappropriate training data. Two practical solutions to this problem are weak and distant supervision. Weak supervision is the act of creating "silver standard" training data by applying a weak, or noisy, labeling function to large amounts of unlabeled data. Distant supervision is a related practice in which external data sources, such as knowledge bases, are used as training signals. One can, in fact, view distant supervision as a form of weak supervision, and in practice the terms are often used interchangeably.

The paradigmatic clinical text mining example of distant supervision is using structured information from the EHR, such as ICD codes, as a labeling mechanism for unstructured text documents. For example, outcomes such as in-hospital mortality (16), hospital readmission (112, 113), and reportable adverse events (114) are routinely captured in the course of health system operations. Although this information is typically attached to patients or encounters, not individual text documents, one can use it as a source of noisy training labels for discharge summaries or other narrative documents attached to the encounters. These noisy labels then serve as a source of supervision for text classification algorithms. Similar results have been achieved using structured ICD9/10 diagnosis (115, 116, 117) and procedure codes (118) as class labels. However, this technique is somewhat limited to the task of document classification; to obtain labels for specific words or text spans (i.e. for NER or relation extraction), one needs a labeling mechanism that works directly on the text.

An alternative is to apply simple heuristic rules to create noisy labels. For example, Wang *et al* used keyword-based weak labels for two separate tasks: smoking status classification and hip fracture classification (93). Importantly, they noted that their best-performing deep learning classifier, a CNN, was robust to the massive label noise created by the weak labeling. Their paper was, to my knowledge, the first to apply a combination of weak supervision and deep learning to clinical text classification; most earlier applications of weak supervision in the biomedical domain focused on images or text from biomedical research articles. Two earlier studies of note in the biomedical domain are Sabbir *et al*'s study of distant supervision for biomedical word sense disambiguation (119) and Fries *et al*'s description of the SwellShark system (120), a generative model for biomedical NER that uses lexicons and ontologies for weak labeling. The Snorkel system, on which SwellShark is based, was recently used to weakly label clinical notes for the purposes of extracting implant details and reports of complications and pain after hip replacement; the weakly labeled notes were then used to train deep learning models to recognize (pain, anatomy) and (complication, implant) relations (121). These methods improved classification performance by 12.8-53.9% over rule-based methods and detected over six times as many complication events compared to structured data alone.

Alternative approaches to the efficient annotation of training sets for clinical text mining include crowdsourcing and active learning. Crowdsourcing is not usually a viable option in the clinical domain because of privacy concerns. Active learning is a strategy for minimizing annotation effort by iteratively sampling subsets of data for human annotation based on the current performance of a supervised learning algorithm (122, 123). However, it still requires recruiting one or more experts to create the annotations.

7. RELATION EXTRACTION AND INFERENCE

Relation extraction is the task of assigning a structured form to a relationship between or among entities based on how it is described in text. Typically this form includes the categories of the involved entities and a label denoting the nature of their relationship, such as "symptom *sign_of* disease" or "test *reveals* problem". For example, the phrase "progressive dyspnea and lower extremity edema concerning for new-onset acute decompensated heart failure" from the last paragraph in Figure 1 contains two different "symptom *sign_of* disease" relations. Relation extraction is usually framed as a text classification problem in which sentences or dependency paths (see sidebar) are classified into groups corresponding to relational labels. It is related to the task of knowledge base creation, which represents text as a network of structured relations over which inference can be performed to generate new knowledge (124).

Although ordinarily discussed alongside other information extraction tasks, such as NER, relation extraction is arguably one step closer to true language understanding. NER and text classification simply label text; they do not address compositionality, the combining of individual facts to generate composite ideas. Compositionality presents a particularly important challenge for clinical text mining because clinical writing reflects a high level of assumed knowledge, as well as unstated implications about the temporal and causal ordering of events. Current clinical text mining systems possess no ability to reason, as a

Dependency Parsing: Representing the syntactic structure of a sentence as a set of directed, binary grammatical relations between pairs of words (or lemmas). For example, a sentence containing the phrase "green car" would contain a dependency of the form "amod(car, green)", where "amod" represents the relationship "adjectival modifier".

Dependency Path: Treating a dependency parse as a directed acyclic graph (DAG), a dependency path is a list of all the edges traversed when moving from one entity to another. It tends to capture only those parts of the sentence that are relevant to the relationship between the two entities.

Relation Extraction: The task of assigning a structured form to a relationship between or among entities. Typically this form includes the identities of the involved entities and a label denoting the nature of their relationship, such as "drug treats disease" or "event 1 precedes event 2".

 doi:10.20944/preprints202010.0649.v2

human would, about the relationships between laboratory and clinical findings and specific diagnoses or treatments (in Figure 1, the meaning of a clear chest x-ray or the implication of pitting edema for a diagnosis of heart failure). Such reasoning will require incorporation of external knowledge derived from, e.g., textbooks or research articles. Relation extraction is a first step in this direction.

7.1. Methods for Clinical Relation Extraction

Modern clinical relation extraction systems are generally based on deep learning models, such as CNNs with pre-trained word2vec embeddings (125), segment CNNs (Seg-CNNs) (126), and coupled Bi-LSTMs with CNNs incorporating dependency path features (127), or other machine learning methods like SVMs (128, 129). They are typically built and evaluated using annotated corpora, such as the relation extraction corpus from the 2010 i2b2/VA dataset (46), which we have seen earlier; indeed, the five studies just mentioned all used this dataset. The recent 2018 n2c2 shared task on adverse drug event relations (130) provides a recent snapshot of the field; of the top 10 systems, five used deep learning, three used SVMs, one used a random forest and one used a rule-based algorithm.

One particular relational class that has been the focus of considerable research in recent years are temporal relations, reviewed in detail in (131). A standard language has been developed for annotating temporal relations in text, including events (EVENTs), time expressions (TIMEXs), and relations between EVENTs and TIMEXs (TLINKs). This formalism has led to the creation of two major annotated corpora for clinical temporal relation extraction: the THYME corpus (132), and the 2012 i2b2 temporal relations corpus (110). Methods for temporal relation extraction have followed those developed for other clinical relation extraction tasks; earlier papers used models such as CRFs and SVMs (133), while later papers apply deep learning approaches such as CNNs (134), Bi-LSTMs (135), and BERT (136).

7.2. Inference and Entailment

Natural language inference (NLI) is a variant of relation extraction with a longstanding presence in NLP, the goal of which is to determine whether one statement (the hypothesis) can be inferred from another (the premise). As of 2018, the clinical NLP community lacked any annotated corpora for NLI, owing in part to the difficulty and expense of getting medical experts to produce annotations and the inability to share patient data with non-expert (e.g. crowd-worker) annotators. However, Romanov and Shivade (137) recently produced the MedNLI dataset to facilitate NLI research in the clinical domain. Starting with premises from the MIMIC-III (87) dataset, physicians were asked to write sentences that (1) were definitely implied by the premise, (2) were neither contradicted nor implied by the premise, and (3) were definitely contradicted by the premise. Although the task is still in its infancy, shared tasks built around the MedNLI dataset have led to multiple new approaches for NLI in this domain, including BERT-BiLSTM-Attention architectures (138), and state-of-the-art ESIM (Enhanced Sequential Inference Model) architectures coupled with knowledge-enhanced word representations based on UMLS (139, 140).

Compositionality: A principle from philosophy and mathematical logic, usually attributed to George Boole, stating that the meaning of a complex expression is determined by the meanings of its constituent expressions and the rules used to combine them.

Entailment: A particular type of relation between two segments of text in which one implies the other; that is, the truth of the second statement follows from the first.

Natural Language Inference: Formerly called “entailment recognition”, the task of determining whether a given hypothesis (statement two) can be inferred from a given premise (statement one).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 3 February 2021 doi:10.20944/preprints202010.0649.v2

A REVIEW OF REVIEWS

The field of clinical text mining has been extensively reviewed in prior articles. The reviews selected below are those I found to be particularly useful surveys of specific research areas or the field in general.

Year	Author(s)	Title	Reference
2011	Chapman *et al*	Overcoming Barriers to NLP for Clinical Text: The Role of Shared Tasks and the Need for Additional Creative Solutions	(141)
2016	Ford *et al*	Extracting Information from the Text of Electronic Medical Records to Improve Case Detection: A Systematic Review	(142)
2016	Koleck *et al*	Natural Language Processing of Symptoms Documented in Free-Text Narratives of Electronic Health Records: A Systematic Review	(5)
2017	Kreimeyer *et al*	Natural Language Processing Systems for Capturing and Standardizing Unstructured Clinical Information: A Systematic Review	(8)
2019	Khattak *et al*	A Survey of Word Embeddings for Clinical Text	(143)
2019	Mujtaba *et al*	Clinical Text Classification Research Trends: Systematic Literature Review and Open Issues	(20)
2020	Spasic *et al*	Clinical Text Data in Machine Learning: Systematic Review	(23)
2010	Stanfill *et al*	A Systematic Literature Review of Automated Clinical Coding and Classification Systems	(94)
2018	Velupillai *et al*	Using Clinical Natural Language Processing for Health Outcomes Research: Overview and Actionable Suggestions for Future Advances	(144)
2018	Wang *et al*	Clinical Information Extraction Applications: A Literature Review	(17)
2020	Wu *et al*	Deep Learning in Clinical Natural Language Processing: A Methodical Review	(22)

8. CONCLUSION

The volume of electronic health record (EHR) data in the United States has skyrocketed in recent years. In 2008, only 42% of office-based physicians reported access to an EHR; this number had climbed to 86% a decade later (145). A favorable policy environment, created by the HITECH Act of 2009 and fueled by the 21st Century Cures Act of 2016, has promoted the meaningful use of EHRs and other observational health data to inform patient care, improve health system operations, facilitate research, and provide "real world evidence" for FDA approval. Over this same time period, methodological advances in machine learning (22, 86), the creation of dedicated clinical text processing software (11, 30) and the seemingly continuous development of high-performing predictive and diagnostic algorithms (6) have fueled enthusiasm about the ability of data and data science to change the way we deliver healthcare.

Amidst such excitement, it would be easy to overlook the fact that most predictive models built on EHR data have focused on outcomes captured in structured data fields, such as mortality, readmissions, length of stay, and diagnosis codes (146, 83). In addition,

 doi:10.20944/preprints202010.0649.v2

a recent systematic review found no net performance benefit of more sophisticated machine learning methods over logistic regression in clinical prediction models (147). Both of these observations can be explained, in part, by the limitations of clinical text mining. To date, the vast quantities of text contained within EHRs have primarily been treated as a source of features for downstream learning algorithms, improving predictive performance over structured data alone (148, 149, 142), but not enabling the types of fundamentally new studies that would result if clinical text mining systems could reason about text and incorporate prior knowledge the way a human would. Assessing whether a treatment failed or succeeded for a given patient, for example, is still a nearly impossible task to accomplish using EHR data without manual chart review. Even the most cutting-edge healthcare data science companies still employ human curators to extract this type of information from text. This situation limits both the types of questions we can ask of EHR data and the potential of even the most sophisticated predictive algorithms and causal models to answer them.

Modern clinical text mining systems have accomplished a great deal. They can now reliably tag a wide variety of clinically-relevant entities in text, map them to standard concepts from lexicons and ontologies, detect negation and uncertainty, and understand the person or people to whom they refer. Given sufficient training data, there are now established system architectures for performing tasks like text classification and relation extraction in the clinical domain. Production-grade clinical text mining systems are in use throughout industry and academia, finding wide application in health outcomes research (144), case detection and phenotyping (142), and automated coding and classification (94). Modern deep learning methods, particularly massive language models like BERT and GPT-3, have recently entered the clinical domain, improving state-of-the-art performance on a variety of clinical NLP tasks and rightfully generating much excitement (86). There remain open questions about the fundamental limitations of these methods to process and understand language (150), and to date the rate of publications describing the application of text mining to EHR data has not kept pace with the field of EHR data mining as a whole (17). However, the field of clinical text mining is also at an exciting turning point, as it is beginning to pursue questions of inference and logic that cut to the heart of what it means to build intelligent machines.

DISCLOSURE STATEMENT

The author is not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

ACKNOWLEDGMENTS

My most sincere thanks go to Cindy Gao for reference checking and editing, Edwin Yoo for writing the fake outpatient progress note that is included in this review, and Guergana Savova, the creator of cTAKES, for her advice on implementing the cTAKES default pipeline. I would also like to acknowledge Tudor Achim, a contributor to Stack Overflow, who was the source of the clear distinction between weak and distant supervision used in the paper (`https://stackoverflow.com/questions/18944805`). Thanks also to Marshall Pierce and Ben Glicksberg, who informally reviewed the first draft of the paper and provided helpful feedback on both the writing and the content.

LITERATURE CITED

1. Roberts A. 2017. Language, structure, and reuse in the electronic health record. *AMA Journal of Ethics* 19:281–288
2. Hersh WR, Weiner MG, Embi PJ, Logan JR, Payne PR, et al. 2013. Caveats for the use of operational electronic health record data in comparative effectiveness research. *Medical Care* 51:S30
3. Rosenbloom ST, Denny JC, Xu H, Lorenzi N, Stead WW, Johnson KB. 2011. Data from clinical notes: a perspective on the tension between structure and flexible documentation. *Journal of the American Medical Informatics Association* 18:181–186
4. Hatef E, Rouhizadeh M, Tia I, Lasser E, Hill-Briggs F, et al. 2019. Assessing the availability of data on social and behavioral determinants in structured and unstructured electronic health records: a retrospective analysis of a multilevel health care system. *JMIR Medical Informatics* 7:e13802
5. Koleck TA, Dreisbach C, Bourne PE, Bakken S. 2019. Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review. *Journal of the American Medical Informatics Association* 26:364–379
6. Topol EJ. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine* 25:44–56
7. Rajkomar A, Dean J, Kohane I. 2019. Machine learning in medicine. *New England Journal of Medicine* 380:1347–1358
8. Kreimeyer K, Foster M, Pandey A, Arya N, Halford G, et al. 2017. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *Journal of Biomedical Informatics* 73:14–29
9. Dalianis H. 2018. *Clinical text mining: Secondary use of electronic patient records*. Springer Nature
10. Friedman C, Alderson PO, Austin JH, Cimino JJ, Johnson SB. 1994. A general natural-language text processor for clinical radiology. *Journal of the American Medical Informatics Association* 1:161–174
11. Savova GK, Masanz JJ, Ogren PV, Zheng J, Sohn S, et al. 2010. Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17:507–513
12. Guzman B, Metzger I, Aphinyanaphongs Y, Grover H, et al. 2020. Assessment of Amazon Comprehend Medical: Medication information extraction. *arXiv preprint arXiv:2002.00481*
13. Wei WQ, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. 2016. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association* 23:e20–e27
14. Liao KP, Cai T, Savova GK, Murphy SN, Karlson EW, et al. 2015. Development of phenotype algorithms using electronic medical records and incorporating natural language processing. *The BMJ* 350:h1885
15. Marafino BJ, Park M, Davies JM, Thombley R, Luft HS, et al. 2018. Validation of prediction models for critical care outcomes using natural language processing of electronic health record data. *JAMA Network Open* 1:e185097–e185097
16. Weissman GE, Hubbard RA, Ungar LH, Harhay MO, Greene CS, et al. 2018. Inclusion of unstructured clinical text improves early prediction of death or prolonged ICU stay. *Critical Care Medicine* 46:1125
17. Wang Y, Wang L, Rastegar-Mojarad M, Moon S, Shen F, et al. 2018. Clinical information extraction applications: a literature review. *Journal of Biomedical Informatics* 77:34–49
18. Castro SM, Tseytlin E, Medvedeva O, Mitchell K, Visweswaran S, et al. 2017. Automated annotation and classification of BI-RADS assessment from radiology reports. *Journal of Biomedical Informatics* 69:177–187
19. Patterson O, Hurdle JF. 2011. *Document clustering of clinical narratives: a systematic study*

of clinical sublanguages. In *AMIA Annual Symposium Proceedings*, vol. 2011, pp. 1099. American Medical Informatics Association
20. Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, et al. 2019. Clinical text classification research trends: Systematic literature review and open issues. *Expert Systems with Applications* 116:494–520
21. Shin B, Chokshi FH, Lee T, Choi JD. 2017. *Classification of radiology reports using neural attention models*. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pp. 4363–4370. IEEE
22. Wu S, Roberts K, Datta S, Du J, Ji Z, et al. 2020. Deep learning in clinical natural language processing: a methodical review. *Journal of the American Medical Informatics Association* 27:457–470
23. Spasic I, Nenadic G. 2020. Clinical text data in machine learning: Systematic review. *JMIR Medical Informatics* 8:e17984
24. Bird S, Klein E, Loper E. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.
25. Manning CD, Surdeanu M, Bauer J, Finkel J, Bethard SJ, McClosky D. 2014. *The Stanford CoreNLP Natural Language Processing Toolkit*. In *Association for Computational Linguistics (ACL) System Demonstrations*, pp. 55–60
26. Qi P, Zhang Y, Zhang Y, Bolton J, Manning CD. 2020. *Stanza: A Python Natural Language Processing Toolkit for Many Human Languages*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*
27. Neumann M, King D, Beltagy I, Ammar W. 2019. *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing*. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 319–327
28. Řehůřek R, Sojka P. 2010. *Software Framework for Topic Modelling with Large Corpora*. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50. Valletta, Malta: ELRA. http://is.muni.cz/publication/884893/en
29. Devlin J, Chang MW, Lee K, Toutanova K. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*
30. Aronson AR, Lang FM. 2010. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association* 17:229–236
31. Demner-Fushman D, Rogers WJ, Aronson AR. 2017. MetaMap Lite: an evaluation of a new Java implementation of MetaMap. *Journal of the American Medical Informatics Association* 24:841–844
32. Zhang Y, Zhang Y, Qi P, Manning CD, Langlotz CP. 2020. Biomedical and clinical english model packages in the Stanza Python NLP library. *arXiv preprint arXiv:2007.14640*
33. Leaman R, Islamaj Doğan R, Lu Z. 2013. DNorm: disease name normalization with pairwise learning to rank. *Bioinformatics* 29:2909–2917
34. Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, et al. 2019. Publicly available clinical BERT embeddings. *arXiv preprint arXiv:1904.03323*
35. Huang K, Altosaar J, Ranganath R. 2019. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*
36. Bodenreider O. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research* 32:D267–D270
37. Wu Y, Jiang M, Xu J, Zhi D, Xu H. 2017. *Clinical named entity recognition using deep learning models*. In *AMIA Annual Symposium Proceedings*, vol. 2017, pp. 1812. American Medical Informatics Association
38. Boag W, Sergeeva E, Kulshreshtha S, Szolovits P, Rumshisky A, Naumann T. 2018. CliNER 2.0: Accessible and accurate clinical concept extraction. *arXiv preprint arXiv:1803.02245*
39. Goeuriot L, Suominen H, Kelly L, Miranda-Escalada A, Krallinger M, et al. 2020. *Overview of the CLEF eHealth Evaluation Lab 2020*. In *International Conference of the Cross-Language*

Evaluation Forum for European Languages, pp. 255–271. Springer

40. Luo YF, Sun W, Rumshisky A. 2019. MCN: A comprehensive corpus for medical concept normalization. *Journal of Biomedical Informatics* 92:103132
41. Pradhan S, Elhadad N, South BR, Martinez D, Christensen L, et al. 2015. Evaluating the state of the art in disorder recognition and normalization of the clinical narrative. *Journal of the American Medical Informatics Association* 22:143–154
42. Yang X, Bian J, Wu Y. 2018. *Detecting medications and adverse drug events in clinical notes using recurrent neural networks*. In *International Workshop on Medication and Adverse Drug Event Detection*, pp. 1–6
43. Jagannatha AN, Yu H. 2016. *Structured prediction models for RNN based sequence labeling in clinical text*. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, vol. 2016, pp. 856
44. Liu Z, Yang M, Wang X, Chen Q, Tang B, et al. 2017. Entity recognition from clinical texts via recurrent neural network. *BMC Medical Informatics and Decision Making* 17:67
45. Dernoncourt F, Lee JY, Uzuner O, Szolovits P. 2017. De-identification of patient notes with recurrent neural networks. *Journal of the American Medical Informatics Association* 24:596–606
46. Uzuner Ö, South BR, Shen S, DuVall SL. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association* 18:552–556
47. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. 2015. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. *Journal of the American Medical Informatics Association* 22:121–131
48. Quimbaya AP, Múnera AS, Rivera RAG, Rodríguez JCD, Velandia OMM, et al. 2016. Named entity recognition over electronic health records through a combined dictionary-based approach. *Procedia Computer Science* 100:55–61
49. Akbik A, Blythe D, Vollgraf R. 2018. *Contextual string embeddings for sequence labeling*. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1638–1649
50. Lample G, Ballesteros M, Subramanian S, Kawakami K, Dyer C. 2016. *Neural Architectures for Named Entity Recognition*. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 260–270
51. Huang Z, Xu W, Yu K. 2015. Bidirectional LSTM-CRF models for sequence tagging. *arXiv preprint arXiv:1508.01991*
52. Settles B. 2004. *Biomedical named entity recognition using conditional random fields and rich feature sets*. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA/BioNLP)*, pp. 107–110
53. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, et al. 2018. A comparison of word embeddings for the biomedical natural language processing. *Journal of Biomedical Informatics* 87:12–20
54. Hassanpour S, Langlotz CP. 2016. Information extraction from multi-institutional radiology reports. *Artificial Intelligence in Medicine* 66:29–39
55. Chalapathy R, Borzeshi EZ, Piccardi M. 2016. Bidirectional LSTM-CRF for clinical concept extraction. *arXiv preprint arXiv:1611.08373*
56. Unanue IJ, Borzeshi EZ, Piccardi M. 2017. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition. *Journal of Biomedical Informatics* 76:102–109
57. Tang B, Cao H, Wu Y, Jiang M, Xu H. 2012. *Clinical entity recognition using structural support vector machines with rich features*. In *Proceedings of the ACM Sixth International Workshop on Data and Text Mining in Biomedical Informatics*, pp. 13–20

58. Lee J, Yoon W, Kim S, Kim D, Kim S, et al. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36:1234–1240
59. Luo G, Huang X, Lin CY, Nie Z. 2015. *Joint entity recognition and disambiguation*. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 879–888
60. Jurafsky D, Martin JH. 2019. Speech and Language Processing (draft). October 2019. *URL https://web. stanford. edu/~jurafsky/slp3*
61. Leaman R, Khare R, Lu Z. 2015. Challenges in clinical natural language processing for automated disorder normalization. *Journal of Biomedical Informatics* 57:28–37
62. Soysal E, Wang J, Jiang M, Wu Y, Pakhomov S, et al. 2018. CLAMP: a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association* 25:331–336
63. Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of Biomedical Informatics* 34:301–310
64. Afshar M, Dligach D, Sharma B, Cai X, Boyda J, et al. 2019. Development and application of a high throughput natural language processing architecture to convert all clinical documents in a clinical data warehouse into standardized medical vocabularies. *Journal of the American Medical Informatics Association* 26:1364–1369
65. Suominen H, Salanterä S, Velupillai S, Chapman WW, Savova G, et al. 2013. *Overview of the ShARe/CLEF eHealth Evaluation Lab 2013*. In *International Conference of the Cross-Language Evaluation Forum for European Languages*, pp. 212–231
66. Li H, Chen Q, Tang B, Wang X, Xu H, et al. 2017. CNN-based ranking for biomedical entity normalization. *BMC Bioinformatics* 18:79–86
67. Luo YF, Sun W, Rumshisky A. 2019. A hybrid normalization method for medical concepts in clinical narrative using semantic matching. *AMIA Summits on Translational Science Proceedings* 2019:732
68. Turchin A, Kolatkar NS, Grant RW, Makhni EC, Pendergrass ML, Einbinder JS. 2006. Using regular expressions to abstract blood pressure and treatment intensification information from the text of physician notes. *Journal of the American Medical Informatics Association* 13:691–695
69. Hao T, Liu H, Weng C. 2016. Valx: a system for extracting and structuring numeric lab test comparison statements from text. *Methods of Information in Medicine* 55:266
70. Xie T, Zhen Y, Tavakoli M, Hundley G, Ge Y. 2020. A deep-learning based system for accurate extraction of blood pressure data in clinical narratives. *AMIA Summits on Translational Science Proceedings* 2020:703
71. Denny JC, Spickard III A, Johnson KB, Peterson NB, Peterson JF, Miller RA. 2009. Evaluation of a method to identify and categorize section headers in clinical documents. *Journal of the American Medical Informatics Association* 16:806–815
72. Pomares-Quimbaya A, Kreuzthaler M, Schulz S. 2019. Current approaches to identify sections within clinical narratives from electronic health records: a systematic review. *BMC Medical Research Methodology* 19:155
73. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J. 2013. *Distributed representations of words and phrases and their compositionality*. In *Advances in Neural Information Processing Systems*, pp. 3111–3119
74. Pennington J, Socher R, Manning CD. 2014. *GloVe: Global vectors for word representation*. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1532–1543
75. Wu Y, Xu J, Jiang M, Zhang Y, Xu H. 2015. *A study of neural word embeddings for named entity recognition in clinical text*. In *AMIA Annual Symposium Proceedings*, vol. 2015, pp. 1326. American Medical Informatics Association

76. Wu Y, Xu J, Zhang Y, Xu H. 2015. *Clinical abbreviation disambiguation using neural word embeddings*. In *Proceedings of BioNLP 15*, pp. 171–176
77. Percha B, Zhang Y, Bozkurt S, Rubin D, Altman RB, Langlotz CP. 2018. Expanding a radiology lexicon using contextual patterns in radiology reports. *Journal of the American Medical Informatics Association* 25:679–685
78. Fan Y, Pakhomov S, McEwan R, Zhao W, Lindemann E, Zhang R. 2019. Using word embeddings to expand terminology of dietary supplements on clinical notes. *JAMIA Open* 2:246–253
79. Lastra-Díaz JJ, Goikoetxea J, Taieb MAH, García-Serrano A, Aouicha MB, Agirre E. 2019. A reproducible survey on word embeddings and ontology-based methods for word similarity: linear combinations outperform the state of the art. *Engineering Applications of Artificial Intelligence* 85:645–665
80. Beam A, Kompa B, Schmaltz A, Fried I, Weber G, et al. 2020. *Clinical Concept Embeddings Learned from Massive Sources of Multimodal Medical Data*. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, vol. 25, pp. 295–306
81. Baumel T, Nassour-Kassis J, Cohen R, Elhadad M, Elhadad N. 2017. Multi-label classification of patient notes a case study on ICD code assignment. *arXiv preprint arXiv:1709.09587*
82. Banerjee I, Chen MC, Lungren MP, Rubin DL. 2018. Radiology report annotation using intelligent word embeddings: applied to multi-institutional chest CT cohort. *Journal of Biomedical Informatics* 77:11–20
83. Miotto R, Li L, Kidd BA, Dudley JT. 2016. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. *Scientific Reports* 6:1–10
84. Turney PD, Pantel P. 2010. From frequency to meaning: vector space models of semantics. *Journal of Artificial Intelligence Research* 37:141–188
85. Kalyan KS, Sangeetha S. 2020. SECNLP: A survey of embeddings in clinical natural language processing. *Journal of Biomedical Informatics* 101:103323
86. Peng Y, Yan S, Lu Z. 2019. *Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets*. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 58–65
87. Johnson AE, Pollard TJ, Shen L, Li-Wei HL, Feng M, et al. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:1–9
88. Si Y, Wang J, Xu H, Roberts K. 2019. Enhancing clinical concept extraction with contextual embeddings. *Journal of the American Medical Informatics Association* 26:1297–1304
89. Li F, Jin Y, Liu W, Rawat BPS, Cai P, Yu H. 2019. Fine-tuning bidirectional encoder representations from transformers (BERT)–based models on large-scale electronic health record notes: an empirical study. *JMIR Medical Informatics* 7:e14830
90. Xu D, Gopale M, Zhang J, Brown K, Begoli E, Bethard S. 2020. Unified medical language system resources improve sieve-based generation and bidirectional encoder representations from transformers (BERT)–based ranking for concept normalization. *Journal of the American Medical Informatics Association*
91. Miotto R, Percha BL, Glicksberg BS, Lee HC, Cruz L, et al. 2020. Identifying acute low back pain episodes in primary care practice from clinical notes: Observational study. *JMIR Medical Informatics* 8:e16878
92. Hassanpour S, Langlotz CP, Amrhein TJ, Befera NT, Lungren MP. 2017. Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: a tool to estimate diagnostic yield. *American Journal of Roentgenology* 208:750–753
93. Wang Y, Sohn S, Liu S, Shen F, Wang L, et al. 2019. A clinical text classification paradigm using weak supervision and deep representation. *BMC Medical Informatics and Decision Making* 19:1
94. Stanfill MH, Williams M, Fenton SH, Jenders RA, Hersh WR. 2010. A systematic literature review of automated clinical coding and classification systems. *Journal of the American Medical Informatics Association* 17:646–651

95. Manning CD, Schütze H, Raghavan P. 2008. *Introduction to Information Retrieval*. Cambridge University Press
96. Shao Y, Taylor S, Marshall N, Morioka C, Zeng-Treitler Q. 2018. *Clinical text classification with word embedding features vs. bag-of-words features*. In *2018 IEEE International Conference on Big Data (Big Data)*, pp. 2874–2878
97. Buchan K, Filannino M, Uzuner Ö. 2017. Automatic prediction of coronary artery disease from clinical narratives. *Journal of Biomedical Informatics* 72:23–32
98. Kocbek S, Cavedon L, Martinez D, Bain C, Mac Manus C, et al. 2016. Text mining electronic hospital records to automatically classify admissions against disease: measuring the impact of linking data sources. *Journal of Biomedical Informatics* 64:158–167
99. Garla VN, Brandt C. 2012. Ontology-guided feature engineering for clinical text classification. *Journal of Biomedical Informatics* 45:992–998
100. Zou H, Hastie T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67:301–320
101. Marafino BJ, Boscardin WJ, Dudley RA. 2015. Efficient and sparse feature selection for biomedical text classification via the elastic net: Application to ICU risk stratification from nursing notes. *Journal of Biomedical Informatics* 54:114–120
102. Lucini FR, Fogliatto FS, da Silveira GJ, Neyeloff JL, Anzanello MJ, et al. 2017. Text mining approach to predict hospital admissions using early medical records from the emergency department. *International Journal of Medical Informatics* 100:1–8
103. Kavuluru R, Rios A, Lu Y. 2015. An empirical evaluation of supervised learning approaches in assigning diagnosis codes to electronic medical records. *Artificial Intelligence in Medicine* 65:155–166
104. Rios A, Kavuluru R. 2015. *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*, pp. 258–267
105. Mullenbach J, Wiegreffe S, Duke J, Sun J, Eisenstein J. 2018. *Explainable Prediction of Medical Codes from Clinical Text*. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pp. 1101–1111
106. Chen MC, Ball RL, Yang L, Moradzadeh N, Chapman BE, et al. 2018. Deep learning to classify radiology free-text reports. *Radiology* 286:845–852
107. Yao L, Mao C, Luo Y. 2019. Clinical text classification with rule-based features and knowledge-guided convolutional neural networks. *BMC Medical Informatics and Decision Making* 19:71
108. Gehrmann S, Dernoncourt F, Li Y, Carlson ET, Wu JT, et al. 2018. Comparing deep learning and concept extraction based methods for patient phenotyping from clinical narratives. *PloS One* 13:e0192360
109. Uzuner O, Bodnari A, Shen S, Forbush T, Pestian J, South BR. 2012. Evaluating the state of the art in coreference resolution for electronic medical records. *Journal of the American Medical Informatics Association* 19:786–791
110. Sun W, Rumshisky A, Uzuner O. 2013. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *Journal of the American Medical Informatics Association* 20:806–813
111. Bethard S, Derczynski L, Savova G, Pustejovsky J, Verhagen M. 2015. *Semeval-2015 task 6: Clinical tempeval*. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pp. 806–814
112. Rumshisky A, Ghassemi M, Naumann T, Szolovits P, Castro V, et al. 2016. Predicting early psychiatric readmission with natural language processing of narrative discharge summaries. *Translational Psychiatry* 6:e921–e921
113. Agarwal A, Baechle C, Behara R, Zhu X. 2017. A natural language processing framework for assessing hospital readmissions for patients with COPD. *IEEE Journal of Biomedical and Health Informatics* 22:588–596

114. Young IJB, Luz S, Lone N. 2019. A systematic review of natural language processing for classification tasks in the field of incident reporting and adverse event analysis. *International Journal of Medical Informatics* 132:103971
115. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. 2016. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. *Journal of the American Medical Informatics Association* 23:1077–1084
116. Chen W, Huang Y, Boyle B, Lin S. 2016. The utility of including pathology reports in improving the computational identification of patients. *Journal of Pathology Informatics* 7:46
117. Venkataraman GR, Pineda AL, Bear Don't Walk IV OJ, Zehnder AM, Ayyar S, et al. 2020. FasTag: Automatic text classification of unstructured medical narratives. *PloS One* 15:e0234647
118. Roysden N, Wright A. 2015. *Predicting health care utilization after behavioral health referral using natural language processing and machine learning*. In *AMIA Annual Symposium Proceedings*, vol. 2015, pp. 2063. American Medical Informatics Association
119. Sabbir A, Jimeno-Yepes A, Kavuluru R. 2017. *Knowledge-based biomedical word sense disambiguation with neural concept embeddings*. In *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, pp. 163–170. IEEE
120. Fries J, Wu S, Ratner A, Ré C. 2017. Swellshark: A generative model for biomedical named entity recognition without labeled data. *arXiv preprint arXiv:1704.06360*
121. Callahan A, Fries JA, Ré C, Huddleston JI, Giori NJ, et al. 2019. Medical device surveillance with electronic health records. *npj Digital Medicine* 2:94
122. Chen Y, Lasko TA, Mei Q, Denny JC, Xu H. 2015. A study of active learning methods for named entity recognition in clinical text. *Journal of Biomedical Informatics* 58:11–18
123. Kholghi M, Sitbon L, Zuccon G, Nguyen A. 2016. Active learning: a step towards automating medical concept extraction. *Journal of the American Medical Informatics Association* 23:289–296
124. Meystre SM, Thibault J, Shen S, Hurdle JF, South BR. 2010. Textractor: a hybrid system for medications and reason for their prescription extraction from clinical text documents. *Journal of the American Medical Informatics Association* 17:559–562
125. Sahu S, Anand A, Oruganty K, Gattu M. 2016. *Relation extraction from clinical texts using domain invariant convolutional neural network*. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pp. 206–215
126. Luo Y, Cheng Y, Uzuner Ö, Szolovits P, Starren J. 2018. Segment convolutional neural networks (Seg-CNNs) for classifying relations in clinical notes. *Journal of the American Medical Informatics Association* 25:93–98
127. Li Z, Yang Z, Shen C, Xu J, Zhang Y, Xu H. 2019. Integrating shortest dependency path and sentence sequence into a deep learning framework for relation extraction in clinical text. *BMC Medical Informatics and Decision Making* 19:22
128. Rink B, Harabagiu S, Roberts K. 2011. Automatic extraction of relations between medical concepts in clinical texts. *Journal of the American Medical Informatics Association* 18:594–600
129. Munkhdalai T, Liu F, Yu H. 2018. Clinical relation extraction toward drug safety surveillance using electronic health record narratives: classical learning versus deep learning. *JMIR Public Health and Surveillance* 4:e29
130. Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2020. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association* 27:3–12
131. Alfattni G, Peek N, Nenadic G. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics* :103488
132. Styler IV WF, Bethard S, Finan S, Palmer M, Pradhan S, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the Association for Computational Linguistics* 2:143–154

133. Tang B, Wu Y, Jiang M, Chen Y, Denny JC, Xu H. 2013. A hybrid system for temporal information extraction from clinical text. *Journal of the American Medical Informatics Association* 20:828–835
134. Dligach D, Miller T, Lin C, Bethard S, Savova G. 2017. *Neural temporal relation extraction*. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 746–751
135. Tourille J, Ferret O, Neveol A, Tannier X. 2017. *Neural architecture for temporal relation extraction: A Bi-LSTM approach for detecting narrative containers*. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 224–230
136. Lin C, Miller T, Dligach D, Bethard S, Savova G. 2019. *A BERT-based universal model for both within-and cross-sentence clinical temporal relation extraction*. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pp. 65–71
137. Romanov A, Shivade C. 2018. *Lessons from Natural Language Inference in the Clinical Domain*. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1586–1596
138. Lee LH, Lu Y, Chen PH, Lee PL, Shyu KK. 2019. *NCUEE at MEDIQA 2019: Medical Text Inference Using Ensemble BERT-BiLSTM-Attention Model*. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pp. 528–532
139. Lu M, Fang Y, Yan F, Li M. 2019. Incorporating domain knowledge into natural language inference on clinical texts. *IEEE Access* 7:57623–57632
140. Sharma S, Santra B, Jana A, Santosh T, Ganguly N, Goyal P. 2019. Incorporating domain knowledge into medical NLI using knowledge graphs. *arXiv preprint arXiv:1909.00160*
141. Chapman WW, Nadkarni PM, Hirschman L, D'avolio LW, Savova GK, Uzuner O. 2011. Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions
142. Ford E, Carroll JA, Smith HE, Scott D, Cassell JA. 2016. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *Journal of the American Medical Informatics Association* 23:1007–1015
143. Khattak FK, Jeblee S, Pou-Prom C, Abdalla M, Meaney C, Rudzicz F. 2019. A survey of word embeddings for clinical text. *Journal of Biomedical Informatics: X* 4:100057
144. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, et al. 2018. Using clinical natural language processing for health outcomes research: Overview and actionable suggestions for future advances. *Journal of Biomedical Informatics* 88:11–19
145. Office of the National Coordinator for Health Information Technology. 2019. 'Office-based Physician Electronic Health Record Adoption,' Health IT Quick-Stat #50. Tech. rep., Office of the National Coordinator for Health Information Technology
146. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, et al. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1:18
147. Christodoulou E, Ma J, Collins GS, Steyerberg EW, Verbakel JY, Van Calster B. 2019. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of clinical epidemiology* 110:12–22
148. Castro VM, Minnier J, Murphy SN, Kohane I, Churchill SE, et al. 2015. Validation of electronic health record phenotyping of bipolar disorder cases and controls. *American Journal of Psychiatry* 172:363–372
149. Hoogendoorn M, Szolovits P, Moons LM, Numans ME. 2016. Utilizing uncoded consultation notes from electronic medical records for predictive modeling of colorectal cancer. *Artificial Intelligence in Medicine* 69:53–61
150. Marcus G. 2018. Deep learning: a critical appraisal. *arXiv preprint arXiv:1801.00631*