Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review

Adrien Bazoge, Emmanuel Morin, Béatrice Daille, Pierre-Antoine Gourraud

▶ **To cite this version:**

Adrien Bazoge, Emmanuel Morin, Béatrice Daille, Pierre-Antoine Gourraud. Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review. JMIR Medical Informatics, 2023, 10.2196/42477 . hal-04347465

HAL Id: hal-04347465
https://hal.science/hal-04347465v1

Submitted on 15 Dec 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons CC0 - Public Domain Dedication 4.0 International License

Review

Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review

Adrien Bazoge[1,2], MSc; Emmanuel Morin[1], PhD; Béatrice Daille[1], PhD; Pierre-Antoine Gourraud[2,3], PhD

[1]Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, F-44000 Nantes, France

[2]Nantes Université, CHU de Nantes, Pôle Hospitalo-Universitaire 11: Santé Publique, Clinique des données, INSERM, CIC 1413, F-44000 Nantes, France

[3]Nantes Université, INSERM, CHU de Nantes, École Centrale Nantes, Centre de Recherche Translationnelle en Transplantation et Immunologie, CR2TI, F-44000 Nantes, France

Corresponding Author:
Pierre-Antoine Gourraud, PhD
Nantes Université, INSERM, CHU de Nantes, École Centrale Nantes, Centre de Recherche Translationnelle en Transplantation et Immunologie, CR2TI
30 bvd Jean Monnet - 2eme étage
F-44000 Nantes
France
Phone: 33 2 447 68 234
Email: Pierre-Antoine.Gourraud@univ-nantes.fr

Abstract

Background: In recent years, health data collected during the clinical care process have been often repurposed for secondary use through clinical data warehouses (CDWs), which interconnect disparate data from different sources. A large amount of information of high clinical value is stored in unstructured text format. Natural language processing (NLP), which implements algorithms that can operate on massive unstructured textual data, has the potential to structure the data and make clinical information more accessible.

Objective: The aim of this review was to provide an overview of studies applying NLP to textual data from CDWs. It focuses on identifying the (1) NLP tasks applied to data from CDWs and (2) NLP methods used to tackle these tasks.

Methods: This review was performed according to the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines. We searched for relevant articles in 3 bibliographic databases: PubMed, Google Scholar, and ACL Anthology. We reviewed the titles and abstracts and included articles according to the following inclusion criteria: (1) focus on NLP applied to textual data from CDWs, (2) articles published between 1995 and 2021, and (3) written in English.

Results: We identified 1353 articles, of which 194 (14.34%) met the inclusion criteria. Among all identified NLP tasks in the included papers, information extraction from clinical text (112/194, 57.7%) and the identification of patients (51/194, 26.3%) were the most frequent tasks. To address the various tasks, symbolic methods were the most common NLP methods (124/232, 53.4%), showing that some tasks can be partially achieved with classical NLP techniques, such as regular expressions or pattern matching that exploit specialized lexica, such as drug lists and terminologies. Machine learning (70/232, 30.2%) and deep learning (38/232, 16.4%) have been increasingly used in recent years, including the most recent approaches based on transformers. NLP methods were mostly applied to English language data (153/194, 78.9%).

Conclusions: CDWs are central to the secondary use of clinical texts for research purposes. Although the use of NLP on data from CDWs is growing, there remain challenges in this field, especially with regard to languages other than English. Clinical NLP is an effective strategy for accessing, extracting, and transforming data from CDWs. Information retrieved with NLP can assist in clinical research and have an impact on clinical practice.

(JMIR Med Inform 2023;11:e42477) doi: 10.2196/42477

KEYWORDS

natural language processing; data warehousing; clinical data warehouse; artificial intelligence; AI

XSL•FO
RenderX

Introduction

Background

For >20 years, health data from patient care have been systematically archived in the form of electronic health records (EHRs) [1,2]. Databases have been created to gather both structured data (eg, vital signs and clinical-biological characteristics and demographics) and unstructured data (eg, textual reports of hospitalizations or visits). These large amounts of data involve multiple contributors: patients, for whom data are collected during hospitalizations or visits; caregivers, who care for the patients and collect the data; and health care institutions, which organize all operational and financial logistics involving the care and related data [3]. The first purpose of collecting these data is to broadly deliver high-quality care to patients, even if the data may be repurposed for secondary use, such as reduction in health care costs, population health management, and clinical research [1]. Human data in clinical research are intended for research purposes and limited in terms of sample size, scope, and longitudinal follow-up (ie, clinical trials or disease registries). The secondary use of EHRs allows to increase patient recruitment in trials [4] and enables access to a larger variety of clinical information for clinical research [5,6].

The rapid increase in digital data production prompted the construction of clinical data warehouses (CDWs), also known as health data warehouses or biomedical data warehouses, for the secondary use of EHRs [2]. *CDW* refers to the interconnection of disparate data from different sources, which are restructured into a common format and indexed using standard vocabularies. CDWs collect data from millions of patients treated in hospitals and can be accessed by stakeholders to analyze care situations and make critical decisions [7]. Unlike in the fields of logistics, marketing, and sales, the health care field has been slow to fully integrate data warehouses. CDWs require managing security and privacy constraints related to medical data [7]. Depending on which country houses the CDW, medical data–related policies can vary and potentially slow the construction process [8]. Data warehouses have been part of the health care landscape for decades [9], especially in the United States, where the first CDWs appeared in the 1990s. In some countries, such as France, CDWs have only been constructed more recently owing to policy constraints. At the institutional level, the use of CDWs underscores that organizations recognize the transformative potential and value of the data generated by their activity. This secondary use of data is facilitated by technological advances in artificial intelligence [10]. Among many types of data, textual data reinforce the popularity of a subgroup of artificial intelligence methods, natural language processing (NLP), which implements algorithms that can operate on massive unstructured textual data [11]. The majority of clinical information is stored in unstructured text format, and NLP allows accessing this information [12,13].

Objectives

This review aims at providing an overview of studies applying clinical NLP to textual data from CDWs. The focus of this review is to identify the (1) NLP tasks applied to data from CDWs and (2) NLP methods used for each task.

Methods

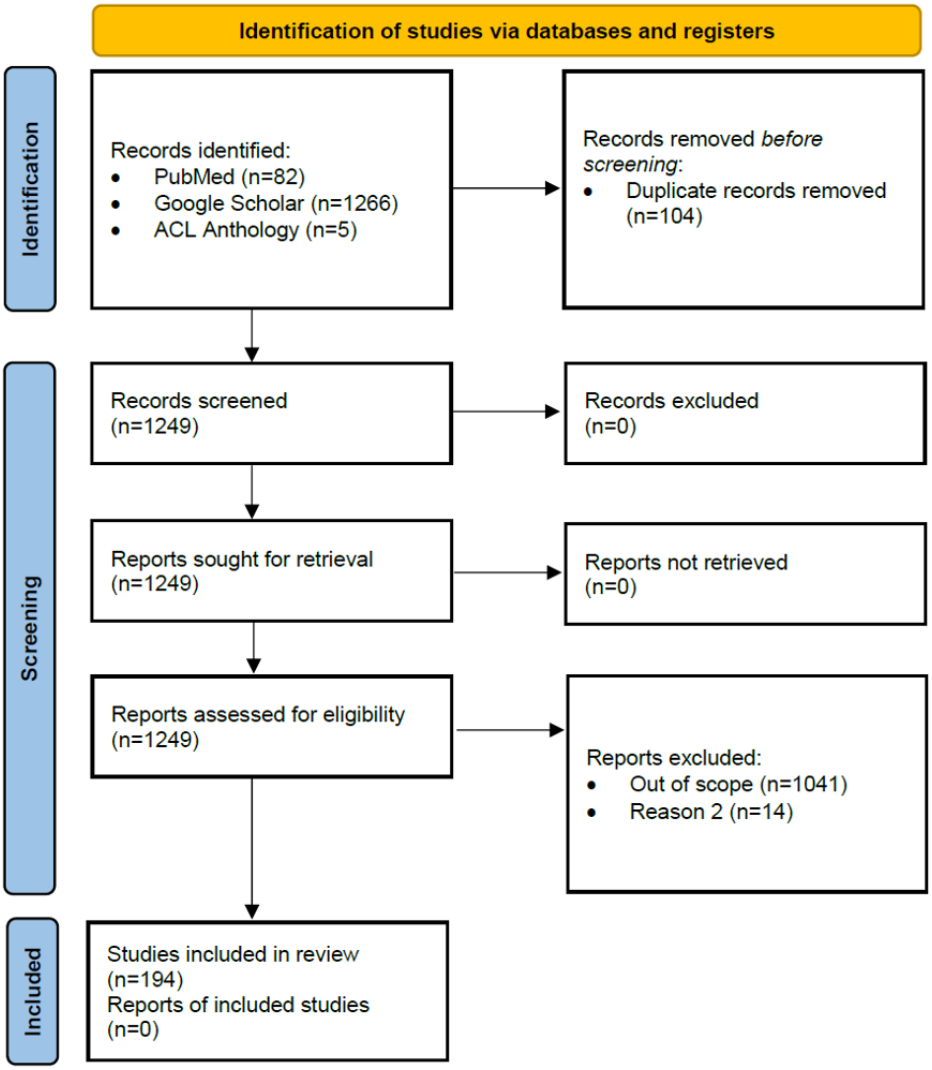
The PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) guidelines were followed for reporting this review (Multimedia Appendix 1).

Review Method and Selection Criteria

Articles identified from the queries were manually included on the basis of the following inclusion criteria: articles (1) mentioning the use of NLP on data from CDWs, (2) published between 1995 and 2021, and (3) written in English. The inclusion was carried out by reading titles and abstracts or by searching the article for the keywords used in the queries to determine whether it was relevant. Details of the article selection steps are described in Figure 1.

XSL•FO
RenderX

Figure 1. PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) article selection flowchart.



Bibliographic Databases

We searched for relevant articles in 3 bibliographic databases: PubMed, ACL Anthology, and Google Scholar. PubMed is specialized in biomedical literature; its query builder allows searchers to construct queries based on both Medical Subject Headings terms and natural language. ACL Anthology covers the literature published in conferences related to computational linguistics and NLP. Google Scholar does not have a dedicated area of specialty for the papers it references and covers a wide range of the literature.

Search Strategy

Identifying papers with NLP applied to data from CDWs involved combining multiple designations: the term *data warehouse* is sometimes referred to as a *database* or a *repository*. In addition, the source of the data used in clinical studies may only be listed in the main manuscript. Data collection requires using multiple queries to aim at both high specificity and high sensitivity.

To retrieve a representative selection of papers, we used queries based on specific keywords for each topic of interest, that is, (1) CDWs and (2) NLP:

- CDWs: “clinical data warehouse,” “biomedical data warehouse,” and “health data warehouse.” The selected keywords representing this topic correspond to the most commonly used terms for CDWs.
- NLP: “natural language processing,” “NLP,” and “text mining.” The keyword “text mining” complements the concept of the “natural language processing” keyword. Text mining stands out as the most frequently used NLP application in the medical field. As a result, the term “natural language processing” can sometimes be eclipsed by “text mining.”

Several queries were made using the selected keywords in each bibliographic database. The details of each query are available in Multimedia Appendix 2.

All queries were run on February 23, 2022. PubMed and ACL Anthology papers were retrieved by manually executing queries on the respective websites of these bibliographic databases. Google Scholar papers were collected using free software [14]. The results of the queries were merged, and duplicates were removed.

The queries are not exhaustive but rather aim to provide a limited and representative selection of papers covering the topics

XSL•FO
RenderX

of interest. Synonyms for *warehouse*, such as *database* or *repository*, were not used in the queries to avoid the collection of a significant number of irrelevant articles to review. Furthermore, some papers may also apply NLP to data from CDWs without mentioning the CDW and could be missed by the queries.

Data Collection

The following data were manually collected from the included articles: (1) NLP tasks addressed in the original paper (the NLP task classification is based on the one provided by Névéol et al [13]), (2) NLP methods used to address the tasks, (3) the CDW that is the source of the data, and (4) the language of the data used in the paper.

Results

Overview

A total of 1353 articles (PubMed: n=82, 6.06%; Google Scholar: n=1266, 93.57%; and ACL Anthology: n=5, 0.37%) were identified with the initial search strategy. After reviewing the title and abstract of each article, of the 1353 articles, 1159 (85.66%) were excluded owing to duplication (n=104, 8.97%), language issues (n=14, 1.21%), and for being out of the scope of this review (n=1041, 89.82%). Overall, of the initially identified 1353 articles, 194 (14.34%) met the inclusion criteria. These 194 articles were published between 2002 and 2021, which means that articles published between 1995 and 2001 did not meet the inclusion criteria.

This section gathers the topics covered in published research on NLP applied to data from CDWs. The results of the reviewed articles are presented by the NLP task mentioned in the articles. Although many articles address the same NLP task, we decided to not directly compare the performances of the methods used in the articles in this review. Methods have been evaluated with different data in different languages and with different metrics. Hence, we concluded that it was not relevant to perform this comparison.

Table 1 gives the count of studies based on the NLP task for 2 periods of time: 2002-2015 and 2016-2021. The 2 time periods were chosen owing to the transition in the NLP paradigm, shifting from knowledge-based to machine learning methods. This transition coincided with the emergence of new tasks, including language modeling.

XSL•FO
RenderX

Table 1. Natural language processing (NLP) tasks reported in the retrieved publications (n=194).

NLP tasks	NLP methods used, n (%)		References
	2002-2015	2016-2021	
Information extraction (n=112)			
Medical concepts (n=37)	S[a]: 14 (74); ML[b]: 5 (26)	S: 10 (40); ML: 11 (44); DL[c]: 4 (16)	[15-51]
Specific characteristics (n=40)	S: 4 (67); ML: 2 (33)	S: 22 (56); ML: 12 (31); DL: 5 (13)	[52-91]
Drugs and adverse events (n=26)	S: 10 (77); ML: 3 (23)	S: 8 (57); ML: 1 (7); DL: 5 (36)	[49,52,92-115]
Findings and symptoms (n=8)	S: 1 (50); ML: 1 (50)	S: 2 (25); ML: 2 (25); DL: 4 (50)	[49,52,116-121]
Relation extraction (n=1)	S: 1 (100)	N/A[d]	[50]
Classification (n=51)			
Phenotyping (n=38)	S: 7 (78); ML: 2 (22)	S: 17 (49); ML: 12 (34); DL: 6 (17)	[50,122-158]
Indexing and coding (n=7)	S: 3 (100)	S: 2 (50); ML: 1 (25); DL: 1 (25)	[159-165]
Topic modeling (n=3)	N/A	S: 1 (25); ML: 3 (75)	[166-168]
Patient identification (n=3)	N/A	S: 1 (25); ML: 2 (50); DL: 1 (25)	[169-171]
Context analysis (n=18)			
Similarity (n=6)	S: 2 (100)	S: 1 (25); DL: 3 (75)	[172-177]
Temporality (n=4)	S: 1 (100)	S: 2 (100)	[93,178-180]
Negation detection (n=3)	N/A	S: 2 (67); DL: 1 (33)	[178,181,182]
Abbreviation (n=2)	N/A	S: 2 (100)	[183,184]
Uncertainty (n=1)	N/A	S: 1 (100)	[180]
Experiencer (n=2)	N/A	S: 2 (100)	[178,182]
Language modeling (n=11)	N/A	ML: 6 (46); DL: 7 (54)	[171,185-194]
Resource development (n=6)			
Corpora and annotation (n=4)	N/A	ML: 1 (100)	[195-198]
Lexica (n=2)	N/A	S: 2 (67); ML: 1 (33)	[199,200]
Shared tasks (n=5)	S: 4 (57); ML: 3 (43)	S: 1 (100)	[201-205]
Deidentification (n=2)	S: 1 (50); ML: 1 (50)	DL: 1 (100)	[206,207]
Data cleaning (n=1)	N/A	ML: 1 (100)	[208]

[a]S: symbolic methods.

[b]ML: machine learning.

[c]DL: deep learning.

[d]N/A: not applicable.

Information Extraction

Information extraction is one of the most studied tasks in NLP within the clinical field. In the included articles, named entity recognition (NER) primarily focuses on identifying entities such as protected health information (PHI) to deidentify clinical documents [206,207], as well as various clinical concepts. These concepts encompass diseases [20,25,40,41,45,47,49]; findings and symptoms [49,52,116-119,121]; and medication names [49,52,93-95,99,100,102,106,107,112,113,115], along with their associated details such as dose, frequency, and duration [52,93-95,112,113,115] as well as potential adverse events [96-98,100,101,106-110,114]. These medical concepts can be mapped to terminologies or ontologies such as the Unified Medical Language System (UMLS) [23,24,30,37-39,41,46,97], Systematized Nomenclature of Medicine–Clinical Terms (SNOMED-CT) [27,28,30], or International Classification of Diseases, Ninth Revision (ICD-9) [21].

Several popular NLP systems have been extensively used for extracting, structuring, and encoding clinical information from narrative patient reports in English. Numerous studies detail the application of the Medical Language Extraction and Encoding System (MedLEE) for clinical concepts [24,27-29,32-36,50,51,121] or medication [103,104,111] extraction, as well as UMLS coding. The extraction and mapping of clinical information from clinical notes to UMLS has also been accomplished using the clinical Text Analysis and Knowledge Extraction System (cTAKES) [16,17,20,22,100,129,134,168], MetaMap [31,37,38,47], MedTagger [44,45,67,78,86,105], and the National Center for Biomedical Ontology (NCBO) Annotator

XSL•FO
RenderX

[97,99,106,107,109,114]. Extracted concepts can be mapped to other standard ontologies and terminologies, such as SNOMED-CT [27]. Caliskan et al [95] evaluated the Averbis Health Discovery NLP system on a medication extraction task on German clinical notes.

Other systems addressing NER or information extraction were customized to specific use cases. Rule-based methods encoded dictionaries and terminologies to match terms and concepts in clinical texts [40-42,49,102,108,112,113]. Machine learning methods take advantage of the clinical knowledge in the large amount of data in CDWs. According to the time period, methods that were used reflect the trend of using NLP state-of-the-art methods and language models. Conditional random fields (CRFs) were used to extract clinical concepts [23,46] or PHI for the deidentification of clinical documents [207]. Hierarchically supervised latent Dirichlet allocation was applied to hospital discharge summaries to predict ICD-9 codes [21]. Deep learning approaches such as bidirectional long short-term memory–CRF (BiLSTM-CRF) [93,113,115] and recurrent neural network grammars [93] performed medical entity extraction in French clinical texts. Chokshi et al [119] compared a bag-of-words model with support vector machine (SVM) and 2 neural network models: a convolutional neural network (CNN) and a neural attention model, both with Word2Vec embedding as input. The accuracies of the CNN and neural attention model models were relatively equal, but they were higher than the accuracy of the SVM model. Lerner et al [49] compared 3 systems for clinical NER: a terminology-based system built on UMLS and SNOMED-CT, a bidirectional gated recurrent unit–CRF system, and a hybrid system using the prediction of the terminology-based system as a feature for the bidirectional gated recurrent unit–CRF system. Yang et al [206] identified PHI from free text with a long short-term memory (LSTM)–CRF model.

Recent state-of-the-art models based on transformer neural architectures [209] were also applied to extract medical concepts. Neuraz et al [52] used a BiLSTM-CRF layer on top of a vector representation of tokens computed by Bidirectional Encoder Representations from Transformers (BERT) in French. BERT and Robustly Optimized BERT Pretraining Approach were examined to extract social and behavioral determinants of health concepts from clinical narratives [15]. Some of the studies paired a neural language model with simple pattern matching techniques; for example, Jouffroy et al [115] proposed a hybrid approach for the extraction of medication information from French clinical text that combined regular expressions to preannotate the text with contextual word embeddings (embeddings from language models [ELMo]) that are fed into a deep recurrent neural network (BiLSTM-CRF).

Some of the studies (31/194, 16%) addressed specific clinical information extracted from clinical texts. These included bone density [59], breast cancer gene 1 or 2 mentions [86], the predictors and timing of lifestyle modification for patients with hypertension [60], the determination of positivity at imaging presentation in radiology reports [66], Banff classification [69], surgical site infection [70], Breast Imaging Reporting and Database System category 3 [71,72], chemotherapy toxicities [76], vital signs [79], transurethral resection of bladder tumors [80], statin use [57], human leukocyte antigen genotypes [82], unplanned episodes of care [83], smoking status [65,84], monoclonal gammopathy [90], skeletal site-specific fractures [85], and social determinants of health [66]. Methods used to extract this information were rule based [67,69-72,76,79,80,82-85], statistical [59,60], or a combination of both [86,90].

Multiple pieces of information about patients were extracted from clinical texts for application in retrospective studies [56]. Ansoborlo et al [89] extracted 52 pieces of bioclinical information from French multidisciplinary team meeting reports concerning lung cancer by applying regular expressions and then compared this approach with a Bayesian classifier method.

Extracting information from clinical text was also carried out as a prediction task. Predicted data cover length of hospital stay [73], the likelihood of neuroscience intensive care unit admission [64], the risk of 30-day readmission in patients with heart failure [55], or quality metrics for the assessment of pretreatment digital rectal examination documentation [62]. Risk assessments of diseases or pathologies, including HIV [61,81], pancreatic cancer [75], pressure ulcer [91], chronic kidney disease [63], and breast cancer [54], have also been studied as prediction tasks. Predicting this clinical information can be achieved with rule-based methods [73,81], machine learning techniques such as latent Dirichlet allocation [63,73], or a combination of both [75,91].

Context Analysis

Linguistic occurrences are particularly relevant where medical information is concerned, such as negation, temporality, uncertainty, or experiencer (ie, determine whether the identified information is related to the patient or a third party, such as a family member). In the included studies, rule-based methods were often used to detect contextual information in clinical text [178,180,182]. Although these methods offer good results (with an approximate F_1-measure value of 0.90), they rely on handmade resources, such as terminologies and regular expressions, and customization is often needed for specific use cases. Temporality patterns have been studied by Liu et al [92] to discern adverse drug events from indications in clinical text. Zhou et al [179] describe a temporal constraint structure constructed from temporal expressions in discharge summaries to model these expressions. In the clinical domain, many temporal expressions have unique characteristics, and this structure provides comprehensive coverage in encoding these expressions. Abbreviations are widely used in medicine and have been studied in French [183] and English [184] clinical texts to better handle medical abbreviations. Recent embedding-based methods such as BERT have made it easier to study negation detection [181] and text similarity [173,174]. Text similarity has also been studied to identify semantically similar concepts [175], similar patients [177], or to detect redundancy in clinical texts [172,176].

Classification

Identifying patients is a key component in clinical research for constructing population studies. NLP can improve the querying and indexing of patients and their data in CDWs. Zhu et al [161]

XSL•FO
RenderX

addressed query expansion based on a large in-domain clinical corpus to solve problems of polysemy, synonymy, and hyponymy in clinical text to improve patient identification. Query expansion was also studied through 3 methods: synonym expansion strategy, topic modeling, and a predicate-based strategy derived from MEDLINE abstracts [165]. An automated electronic search algorithm for identifying postoperative complications was evaluated by Tien et al [162]. A semantic health data warehouse was designed to assist health professionals in prescreening eligible patients in clinical trials [163,164]. A combination of structured and unstructured German data was used by Scheurwegs et al [160] to assign clinical codes to patient stays.

Downstream of the query of CDWs, NLP can be applied to identify patients or documents of interest when the classification methods offered by CDWs are not precise enough. Patient identification can be carried out using methods such as rule-based approaches, which involve using terms related to eligible criteria [127,137,140-150,153,170], or learning-based approaches [126,131,133], or a combination of both [152,155-157,169]. Li et al [166] and Chen et al [167] applied latent Dirichlet allocation in clinical notes for topic modeling. Agarwal et al [154] detailed a logistic regression model of phenotypes learned on noisy labeled data. Some of the studies (4/194, 2.1%) relied on Dr Warehouse, a biomedical data warehouse oriented toward clinical narrative reports, developed at Necker Children's Hospital in Paris, France. This data warehouse was used to explore, using the frequency and term frequency–inverse document frequency (TF-IDF), the association between clinical phenotypes and rare diseases such as the potassium voltage-gated channel subfamily A member 2 variant in neurodevelopmental syndromes [138], Dravet syndrome [125], ciliopathy [139], and other rare diseases [136].

Language Modeling

Recent word embedding–based methods take advantage of the large amount of data stored in CDWs to learn effective semantic representations of clinical texts. In the included articles, these methods allowed to make calculations on words to find, for example, similar terms in the embedding space [88,130]. Among these methods, transformer-based models, such as BERT, were fine-tuned for multiple tasks, including text classification to map document titles to Logical Observation Identifiers Names and Codes Document Ontology [159] and sequence labeling to detect and estimate the location of abnormalities in whole-body scans [53]. Similarly, clinical text was structured with the classification of ICD-9 codes based on vectorization methods [190,191].

Some of the studies evaluated the effectiveness of word embedding models on multiple tasks. Lee et al [135] evaluated Node2Vec, singular value decomposition, Language Identification for Named Entities, Word2Vec, and global vectors for word representation (GloVe) in retrieving relevant medical features for phenotyping tasks. The authors demonstrated that GloVe, when trained on EHR data, outperforms other embedding methods. GloVe and Word2Vec were used in conjunction with LSTM and gated recurrent unit and evaluated across multiple tasks, with gated recurrent unit outperforming LSTM [192]. Similarly, Dynomant et al [193] compared on multiple tasks 3 embedding methods (Word2Vec, GloVe, and fastText) trained on a French corpus. The 3 methods were evaluated on 4 tasks, and Word2Vec with the skip-gram architecture had the highest score on 3 (75%) of the 4 tasks. Peng et al [185] evaluated 2 transformer-based models, BERT and ELMo, on 10 benchmark data sets and found that the BERT model achieved the best results. BERT was also evaluated on sentence similarity, relation extraction, inference, and NER tasks on data sets from clinical domains [186]. The study by Neuraz et al [188] comparing fastText and ELMo showed that models learned on clinical data performed better than models learned on data from the general domain. The study by Tawfik and Spruit [187] described a toolkit to evaluate the effectiveness of sentence representation learning models.

Text representation models are commonly used as embedding layers in neural network models developed for specific tasks. Word2Vec has been used in numerous studies for various purposes, including assessing bone scan use among patients with prostate cancer with a CNN [151], screening and diagnosing of breast cancer with a deep learning architecture [123], extracting features used for risk prediction of liver transplantation for hepatocellular cancer with a capsule neural network [124], and using a CNN to learn the clinical trial criteria eligibility status of patients for participation in cohort studies [171]. Lee et al [194] proposed a unified graph representation learning framework based on graph convolutional networks and LSTM to construct an EHR graph representation of medical entities. Dligach et al [189] developed a clinical text encoder for specific phenotypes. Experiments were conducted with a deep averaging network and a CNN to construct this text encoder.

Resource Development and Shared Tasks

Many NLP methods rely on clinically specific resources to be developed. In the included articles, data from CDWs, combined with clinical expert knowledge, allowed the development of resources such as annotation guidelines and schemes [195,196,198], lexica [200], ontologies [199], or frameworks to validate the outputs of NLP systems [197].

International community efforts have been demonstrated through shared tasks involving clinical notes from CDWs. In the included articles, the Informatics for Integrating Biology and the Bedside (i2b2) obesity challenge focused on obesity and its 15 most common comorbidities through a multiclass multilabel classification task [204,205]. Another i2b2 challenge held in 2009 concerned extracting medication information from clinical text [202,210]. Three tasks were proposed in the fourth i2b2 or Department of Veterans Affairs shared-task and workshop challenge: extraction of medical problems, tests, and treatments; classification of assertions made on medical problems; and classification of a relationship between a pair of concepts that appear in the same sentence where at least 1 concept is a medical problem [202]. These i2b2 shared tasks relied on deidentified discharge summaries from the Partners HealthCare research patient data repository. The 2018 National NLP Clinical Challenges (n2c2) shared-task workshop presented a cohort selection task for clinical trials [203].

XSL•FO
RenderX

Previously presented NLP tasks and methods were applied to medical data in different languages, with the majority being in English (153/194, 78.9%; Table 2).

Multimedia Appendix 3 presents the CDWs used in the publications presented in this review. Overall, the oldest CDWs, such as the Columbia University Irving Medical Center CDW, Mayo Clinic, and the Partners HealthCare research patient data repository, are the ones that reuse the most textual data and contribute the most to developing the application of NLP on EHR data.

Table 2. Language of the data used in the papers (n=194).

Data language	Publications, n (%)	References
English	153 (78.9)	[15-17,19-25,27-38,41-48,50,51,53-68,71,72,74,75,78-80,83-88,90-92,96,97,99-112,114,116,119-124,126-135,137,140,142-149,151-154,156-159,161,162,165-176,179,181,184-187,189-192,194-196,198,200-208]
French	27 (13.9)	[39,49,52,73,76,77,81,89,93,94,113,115,118,125,136,138,139,155,163,164,177,178,182,183,188,193,197]
German	9 (4.6)	[18,26,69,95,117,150,160,180,199]
Korean	3 (1.5)	[40,65,82]
Japanese	1 (0.5)	[98]
Not mentioned	2 (1)	[70,141]

Discussion

Principal Findings

As CDWs become more prevalent and are adopted in many countries, they open up opportunities for clinical NLP to flourish. This review shows that the use of NLP on data from CDWs is primarily focused on extracting information from clinical texts and identifying patients. Depending on the task, various methods can be used, from symbolic methods to machine learning and deep learning techniques. The oldest CDWs are associated with the most numerous publications. This shows that the use of NLP is not a 1-time event but is intended to be established in the long term. It contributes to the continuous quality improvement of data made available in CDWs.

Symbolic and linguistics methods have still been widely used in recent years, despite the preponderance of deep learning approaches that have shown excellent results across a majority of tasks. This shows that some tasks can be partially achieved with classical NLP techniques, such as regular expressions and pattern matching that exploit specialized lexica such as drug lists and terminologies. Existing information extraction tools such as cTAKES, MedLEE, and MetaMap offer easy handling and satisfactory results. As a result, they are often used for processing English language clinical text.

Interestingly, the number of data languages presented in our review is quite low—only 5 languages: English, French, German, Korean, and Japanese. This can be explained by three factors: (1) CDWs are not cited as data sources in articles, resulting in a bias related to queries; (2) CDWs are operational in another country, but NLP has not yet been used on these data; and (3) CDWs have not yet been adopted in every country.

Opportunities and Challenges

Although NLP methods are becoming increasingly popular, there remain challenges within the clinical field. This review demonstrates that the use of NLP in CDWs is becoming more frequent over time. However, CDWs still rarely provide open access for NLP research owing to medical data confidentiality. A first step to partially overcome the privacy constraints could involve working on deidentified or anonymized data from CDWs, as has been done in some recent shared tasks [202,204,205,210]. These shared tasks, crucial for making advances in medical NLP research, are too scarce, particularly for languages other than English [9]. Providing an appropriate measure to respect patient privacy should encourage collaboration among hospital and NLP research teams and facilitate access to clinical data.

The global movement is toward the structuring and interoperability of clinical data; yet, the finer points of medical reasoning are always expressed in textual reports, and such information cannot always be structured. The increase in NLP approaches applied to clinical data could lead to major advances in clinical research, both to identify the populations of interest and to retrieve relevant information of these patients for clinical research. NLP could also have a positive impact on the daily life of caregivers by speeding up access to information contained in patient EHRs using automated tools for the summarization of patient history. Indeed, caregivers invest a significant amount of time recording information gathered during care delivery in textual reports. Surprisingly, they also dedicate an equivalent amount of time sifting through numerous documents to retrieve this information when needed.

Structured or semistructured data stored in CDWs provide information about patient follow-up and can serve as a valuable resource for developing or enhancing NLP systems. Indeed, temporal data can offer guidance on where the information is most relevant in the text. In addition, other data such as PHI, including names, surnames, and addresses, can be used as a starting point in NLP systems.

Clinical data are a use case for NLP research. They possess the advantage of being accessible in multiple languages owing to the global nature of medical care. This accessibility enhances research efforts focused on multilingualism. Such data are available in abundance, facilitating the acquisition of effective clinical text representations that can be applied in deep neural networks to learn relevant concept models. Clinical data fall within the category of specialized domains or languages

XSL•FO
RenderX

designed for specific purposes. They share certain properties, such as specific knowledge, uses, and discourse. This also entails undertaking specific tasks such as deidentification or anonymization.

The analysis of the literature conducted here highlights the need for further development of CDWs, with a stronger integration of NLP applications throughout the entire data value chain.

Limitations

The NLP tasks identified in this review cover only a small part of all existing NLP tasks in the general domain. These tasks globally reflect the primary needs in clinical research, such as identifying the study population and extracting clinical information for a defined population. Other tasks, such as context analysis and language modeling, have been widely studied in the general domain NLP but are less prevalent in the clinical domain. In recent years, transformer-based approaches have emerged as the state-of-the-art methods for most NLP tasks. However, this review indicates that these methods have not fully spread to the clinical domain. This demonstrates a gap between methods that are well established in the general domain NLP and their adoption in specific domains such as the clinical domain.

This review focuses on 2 very specific subjects from different emerging domains: clinical NLP and CDWs. This combination of subjects implies the use of multiple bibliographic databases and the aggregation of multiple queries to ensure good coverage of the literature. Some bibliographic databases cover a wider range of articles and include articles already present in other more specialized sources. To avoid having a surfeit of duplicate articles, we prioritized the use of the most encompassing bibliographic databases: Google Scholar and PubMed. This introduces a bias of completeness because relevant articles could be missing from the selected bibliographic databases and be present in others we did not use in this review, such as Scopus, Web of Science, and Embase.

There is another bias of completeness related to the search by keywords in the bibliographic databases. A given concept can be expressed in various ways in natural language, using different keywords. The choice of keywords is crucial to aim at both high specificity and high sensitivity, even if the selected keywords are searched in the whole paper. In this review, we used very broad keywords to have the highest sensitivity but at the expense of specificity (n=194, 14.34% relevant articles among 1353 articles identified from the queries).

Conclusions

CDWs are central to the secondary use of clinical texts for research purposes. Our review highlights the growing interest in computerized health data, particularly in clinical texts, where NLP is used to address various clinical tasks. These tasks include patient identification and information extraction, as well as clinical NLP tasks such as language modeling, context analysis, and EHR deidentification. The broad spectrum of NLP approaches has been effectively leveraged, ranging from symbolic methods to machine learning and deep learning methods. Despite the prevalence of pretrained language models in the broader NLP domain, symbolic and linguistics methods have continued to be used in recent years. In the realm of clinical NLP for CDWs, the trends align with global NLP patterns, where resources and methods are predominantly developed for the English language. The development of NLP in the medical field will require cooperation between health care and NLP experts.

Acknowledgments

This work was supported by the French Agence Nationale de la Recherche (ANR; National Research Agency) AIBy4 project (ANR-20-THIA-0011).

Conflicts of Interest

None declared.

Multimedia Appendix 1

PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) checklist.
[PDF File (Adobe PDF File), 50 KB-Multimedia Appendix 1]

Multimedia Appendix 2

Search queries used in PubMed, Google Scholar, and ACL Anthology to retrieve publications for inclusion in this systematic review.
[DOCX File , 13 KB-Multimedia Appendix 2]

Multimedia Appendix 3

Clinical data warehouses from which data have been used in a publication.
[DOCX File , 21 KB-Multimedia Appendix 3]

References

XSL•FO
RenderX

1. Meystre SM, Lovis C, Bürkle T, Tognola G, Budrionis A, Lehmann CU. Clinical data reuse or secondary use: current status and potential future progress. Yearb Med Inform. 2017 Aug;26(1):38-52 [doi: 10.15265/IY-2017-007] [Medline: 28480475]
2. Adler-Milstein J, Holmgren AJ, Kralovec P, Worzala C, Searcy T, Patel V. Electronic health record adoption in US hospitals: the emergence of a digital "advanced use" divide. J Am Med Inform Assoc. 2017 Nov 01;24(6):1142-1148 [FREE Full text] [doi: 10.1093/jamia/ocx080] [Medline: 29016973]
3. Casto AB, Layman E. Principles of Healthcare Reimbursement. Springfield, IL. American Health Information Management Association; 2013:371
4. Köpcke F, Prokosch HU. Employing computers for the recruitment into clinical trials: a comprehensive systematic review. J Med Internet Res. 2014 Jul 01;16(7):e161 [FREE Full text] [doi: 10.2196/jmir.3446] [Medline: 24985568]
5. Shah SM, Khan RA. Secondary use of electronic health record: opportunities and challenges. IEEE Access. 2020;8:136947-136965 [doi: 10.1109/access.2020.3011099]
6. Sarwar T, Seifollahi S, Chan J, Zhang X, Aksakalli V, Hudson I, et al. The secondary use of electronic health records for data mining: data characteristics and challenges. ACM Comput Surv. 2022 Jan 18;55(2):1-40 [FREE Full text] [doi: 10.1145/3490234]
7. Hamoud A, Hashim A, Awadh W. Clinical data warehouse: a review. Iraqi J Comput Inform. 2018 Dec 31;44(2):16-26 [FREE Full text] [doi: 10.25195/ijci.v44i2.53]
8. Holmes JH, Elliott TE, Brown JS, Raebel MA, Davidson A, Nelson AF, et al. Clinical research data warehouse governance for distributed research networks in the USA: a systematic review of the literature. J Am Med Inform Assoc. 2014 Jul;21(4):730-736 [FREE Full text] [doi: 10.1136/amiajnl-2013-002370] [Medline: 24682495]
9. Gagalova KK, Leon Elizalde MA, Portales-Casamar E, Görges M. What you need to know before implementing a clinical research data warehouse: comparative review of integrated data repositories in health care institutions. JMIR Form Res. 2020 Aug 27;4(8):e17687 [FREE Full text] [doi: 10.2196/17687] [Medline: 32852280]
10. Lin WC, Chen JS, Chiang MF, Hribar MR. Applications of artificial intelligence to electronic health record data in ophthalmology. Transl Vis Sci Technol. 2020 Feb 27;9(2):13 [FREE Full text] [doi: 10.1167/tvst.9.2.13] [Medline: 32704419]
11. Juhn Y, Liu H. Artificial intelligence approaches using natural language processing to advance EHR-based clinical research. J Allergy Clin Immunol. 2020 Feb;145(2):463-469 [FREE Full text] [doi: 10.1016/j.jaci.2019.12.897] [Medline: 31883846]
12. Kim E, Rubinstein SM, Nead KT, Wojcieszynski AP, Gabriel PE, Warner JL. The evolving use of electronic health records (EHR) for research. Semin Radiat Oncol. 2019 Oct;29(4):354-361 [doi: 10.1016/j.semradonc.2019.05.010] [Medline: 31472738]
13. Névéol A, Dalianis H, Velupillai S, Savova G, Zweigenbaum P. Clinical natural language processing in languages other than English: opportunities and challenges. J Biomed Semantics. 2018 Mar 30;9(1):12 [FREE Full text] [doi: 10.1186/s13326-018-0179-8] [Medline: 29602312]
14. Publish or perish. Anne-Wil Harzing. URL: https://harzing.com/resources/publish-or-perish [accessed 2023-11-27]
15. Yu Z, Yang X, Dang C, Wu S, Adekkanattu P, Pathak J, et al. A study of social and behavioral determinants of health in lung cancer patients using transformers-based natural language processing models. AMIA Annu Symp Proc. 2021 Feb 21;2021:1225-1233 [FREE Full text] [Medline: 35309014]
16. Afshar M, Dligach D, Sharma B, Cai X, Boyda J, Birch S, et al. 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. J Am Med Inform Assoc. 2019 Nov 01;26(11):1364-1369 [FREE Full text] [doi: 10.1093/jamia/ocz068] [Medline: 31145455]
17. Raja AS, Pourjabbar S, Ip IK, Baugh CW, Sodickson AD, O'Leary M, et al. Impact of a health information technology-enabled appropriate use criterion on utilization of emergency department CT for renal colic. AJR Am J Roentgenol. 2019 Jan;212(1):142-145 [doi: 10.2214/AJR.18.19966] [Medline: 30403534]
18. Grön L, Bertels A, Heylen K. Leveraging sublanguage features for the semantic categorization of clinical terms. In: Proceedings of the 18th BioNLP Workshop and Shared Task. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 211-216 URL: https://aclanthology.org/W19-5022.pdf [doi: 10.18653/v1/w19-5022]
19. Wang L, Haug PJ, Del Fiol G. Using classification models for the generation of disease-specific medications from biomedical literature and clinical data repository. J Biomed Inform. 2017 May;69:259-266 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.014] [Medline: 28435015]
20. Walsh JA, Shao Y, Leng J, He T, Teng CC, Redd D, et al. Identifying axial spondyloarthritis in electronic medical records of US veterans. Arthritis Care Res (Hoboken). 2017 Sep;69(9):1414-1420 [doi: 10.1002/acr.23140] [Medline: 27813310]
21. Perotte A, Wood F, Elhadad N, Wood F. Hierarchically supervised Latent Dirichlet allocation. In: Proceedings of the 24th International Conference on Neural Information Processing Systems. 2011 Presented at: NIPS '11; December 12-15, 2011; Granada, Spain p. 2609-2617 URL: https://dl.acm.org/doi/10.5555/2986459.2986750
22. Zhong QY, Karlson EW, Gelaye B, Finan S, Avillach P, Smoller JW, et al. Screening pregnant women for suicidal behavior in electronic medical records: diagnostic codes vs. clinical notes processed by natural language processing. BMC Med Inform Decis Mak. 2018 May 29;18(1):30 [FREE Full text] [doi: 10.1186/s12911-018-0617-7] [Medline: 29843698]
23. Jonnalagadda S, Cohen T, Wu S, Gonzalez G. Enhancing clinical concept extraction with distributional semantics. J Biomed Inform. 2012 Feb;45(1):129-140 [FREE Full text] [doi: 10.1016/j.jbi.2011.10.007] [Medline: 22085698]

XSL•FO
RenderX

24. 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 Feb 28;17(1):24 [FREE Full text] [doi: 10.1186/s12911-017-0418-4] [Medline: 28241760]
25. Ashish N, Dahm L, Boicey C. University of California, Irvine-Pathology Extraction Pipeline: the pathology extraction pipeline for information extraction from pathology reports. Health Informatics J. 2014 Dec;20(4):288-305 [FREE Full text] [doi: 10.1177/1460458213494032] [Medline: 25155030]
26. Scheurwegs E, Luyckx K, Luyten L, Goethals B, Daelemans W. Assigning clinical codes with data-driven concept representation on Dutch clinical free text. J Biomed Inform. 2017 May;69:118-127 [FREE Full text] [doi: 10.1016/j.jbi.2017.04.007] [Medline: 28400312]
27. Melton GB, Parsons S, Morrison FP, Rothschild AS, Markatou M, Hripcsak G. Inter-patient distance metrics using SNOMED CT defining relationships. J Biomed Inform. 2006 Dec;39(6):697-705 [FREE Full text] [doi: 10.1016/j.jbi.2006.01.004] [Medline: 16554186]
28. Wang X, Chused A, Elhadad N, Friedman C, Markatou M. Automated knowledge acquisition from clinical narrative reports. AMIA Annu Symp Proc. 2008 Nov 06;2008:783-787 [FREE Full text] [Medline: 18999156]
29. Wang X, Hripcsak G, Markatou M, Friedman C. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. J Am Med Inform Assoc. 2009;16(3):328-337 [FREE Full text] [doi: 10.1197/jamia.M3028] [Medline: 19261932]
30. Lowe HJ, Huang Y, Regula DP. Using a statistical natural language parser augmented with the UMLS specialist lexicon to assign SNOMED CT codes to anatomic sites and pathologic diagnoses in full text pathology reports. AMIA Annu Symp Proc. 2009 Nov 14;2009:386-390 [FREE Full text] [Medline: 20351885]
31. Harris DR, Henderson DW, Corbeau A. sig2db: a workflow for processing natural language from prescription instructions for clinical data warehouses. AMIA Jt Summits Transl Sci Proc. 2020 May 30;2020:221-230 [FREE Full text] [Medline: 32477641]
32. Chuang JH, Friedman C, Hripcsak G. A comparison of the Charlson comorbidities derived from medical language processing and administrative data. Proc AMIA Symp. 2002:160-164 [FREE Full text] [Medline: 12463807]
33. Van Vleck TT, Wilcox A, Stetson PD, Johnson SB, Elhadad N. Content and structure of clinical problem lists: a corpus analysis. AMIA Annu Symp Proc. 2008 Nov 06;2008:753-757 [FREE Full text] [Medline: 18999284]
34. Li L, Chase HS, Patel CO, Friedman C, Weng C. Comparing ICD9-encoded diagnoses and NLP-processed discharge summaries for clinical trials pre-screening: a case study. AMIA Annu Symp Proc. 2008 Nov 06;2008:404-408 [FREE Full text] [Medline: 18999285]
35. Chen ES, Hripcsak G, Xu H, Markatou M, Friedman C. Automated acquisition of disease drug knowledge from biomedical and clinical documents: an initial study. J Am Med Inform Assoc. 2008;15(1):87-98 [FREE Full text] [doi: 10.1197/jamia.M2401] [Medline: 17947625]
36. Carlo L, Chase HS, Weng C. Aligning structured and unstructured medical problems using UMLS. AMIA Annu Symp Proc. 2010 Nov 13;2010:91-95 [FREE Full text] [Medline: 21346947]
37. Zhou X, Wang Y, Sohn S, Therneau TM, Liu H, Knopman DS. Automatic extraction and assessment of lifestyle exposures for Alzheimer's disease using natural language processing. Int J Med Inform. 2019 Oct;130:103943 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.08.003] [Medline: 31476655]
38. Singh K, Betensky RA, Wright A, Curhan GC, Bates DW, Waikar SS. A concept-wide association study of clinical notes to discover new predictors of kidney failure. Clin J Am Soc Nephrol. 2016 Dec 07;11(12):2150-2158 [FREE Full text] [doi: 10.2215/CJN.02420316] [Medline: 27927892]
39. Campillo-Gimenez B, Garcelon N, Jarno P, Chapplain JM, Cuggia M. Full-text automated detection of surgical site infections secondary to neurosurgery in Rennes, France. Stud Health Technol Inform. 2013;192:572-575 [Medline: 23920620]
40. Hong SN, Son HJ, Choi SK, Chang DK, Kim Y, Jung S, et al. A prediction model for advanced colorectal neoplasia in an asymptomatic screening population. PLoS One. 2017;12(8):e0181040 [FREE Full text] [doi: 10.1371/journal.pone.0181040] [Medline: 28841657]
41. Hunter-Zinck HS, Peck JS, Strout TD, Gaehde SA. Predicting emergency department orders with multilabel machine learning techniques and simulating effects on length of stay. J Am Med Inform Assoc. 2019 Dec 01;26(12):1427-1436 [FREE Full text] [doi: 10.1093/jamia/ocz171] [Medline: 31578568]
42. Kshatriya BS, Balls-Berry JE, Freeman WD, Zhang R, Wang Y. Completeness of Social and Behavioral Determinants of Health in Electronic Health Records: A case study on the Patient-Provided Information from a minority cohort with sexually transmitted diseases. Research Square. Preprint posted online December 10, 2020. 2020 [FREE Full text] [doi: 10.21203/rs.3.rs-123744/v1]
43. Baghal A, Al-Shukri S, Kumari A. Agile natural language processing model for pathology knowledge extraction and integration with clinical enterprise data warehouse. In: Proceedings of the 6th International Conference on Social Networks Analysis, Management and Security. 2019 Presented at: SNAMS '19; October 22-25, 2019; Granada, Spain p. 419-422 URL: https://ieeexplore.ieee.org/document/8931828 [doi: 10.1109/snams.2019.8931828]
44. Liu H, Wu ST, Li D, Jonnalagadda S, Sohn S, Wagholikar K, et al. Towards a semantic lexicon for clinical natural language processing. AMIA Annu Symp Proc. 2012;2012:568-576 [FREE Full text] [Medline: 23304329]

XSL•FO
RenderX

45. Afzal N, Sohn S, Abram S, Scott CG, Chaudhry R, Liu H, et al. Mining peripheral arterial disease cases from narrative clinical notes using natural language processing. J Vasc Surg. 2017 Jun;65(6):1753-1761 [FREE Full text] [doi: 10.1016/j.jvs.2016.11.031] [Medline: 28189359]
46. Jonnalagadda S, Cohen T, Wu S, Liu H, Gonzalez G. Using empirically constructed lexical resources for named entity recognition. Biomed Inform Insights. 2013 Jun 24;6(Suppl 1):17-27 [FREE Full text] [doi: 10.4137/BII.S11664] [Medline: 23847424]
47. Osborne JD, Wyatt M, Westfall AO, Willig J, Bethard S, Gordon G. Efficient identification of nationally mandated reportable cancer cases using natural language processing and machine learning. J Am Med Inform Assoc. 2016 Nov;23(6):1077-1084 [FREE Full text] [doi: 10.1093/jamia/ocw006] [Medline: 27026618]
48. Hernandez-Boussard T, Blayney DW, Brooks JD. Leveraging digital data to inform and improve quality cancer care. Cancer Epidemiol Biomarkers Prev. 2020 Apr;29(4):816-822 [FREE Full text] [doi: 10.1158/1055-9965.EPI-19-0873] [Medline: 32066619]
49. Lerner I, Paris N, Tannier X. Terminologies augmented recurrent neural network model for clinical named entity recognition. J Biomed Inform. 2020 Feb;102:103356 [FREE Full text] [doi: 10.1016/j.jbi.2019.103356] [Medline: 31837473]
50. Wang X, Chase H, Markatou M, Hripcsak G, Friedman C. Selecting information in electronic health records for knowledge acquisition. J Biomed Inform. 2010 Aug;43(4):595-601 [FREE Full text] [doi: 10.1016/j.jbi.2010.03.011] [Medline: 20362071]
51. Overby CL, Pathak J, Gottesman O, Haerian K, Perotte A, Murphy S, et al. A collaborative approach to developing an electronic health record phenotyping algorithm for drug-induced liver injury. J Am Med Inform Assoc. 2013 Dec;20(e2):e243-e252 [FREE Full text] [doi: 10.1136/amiajnl-2013-001930] [Medline: 23837993]
52. Neuraz A, Lerner I, Digan W, Paris N, Tsopra R, Rogier A, et al. AP-HP/Universities/INSERM COVID-19 Research Collaboration; AP-HP COVID CDR Initiative. Natural language processing for rapid response to emergent diseases: case study of calcium channel blockers and hypertension in the COVID-19 pandemic. J Med Internet Res. 2020 Aug 14;22(8):e20773 [FREE Full text] [doi: 10.2196/20773] [Medline: 32759101]
53. Eyuboglu S, Angus G, Patel BN, Pareek A, Davidzon G, Long J, et al. Multi-task weak supervision enables anatomically-resolved abnormality detection in whole-body FDG-PET/CT. Nat Commun. 2021 Mar 25;12(1):1880 [FREE Full text] [doi: 10.1038/s41467-021-22018-1] [Medline: 33767174]
54. He T, Puppala M, Ezeana CF, Huang Y, Chou P, Yu X, et al. A deep learning-based decision support tool for precision risk assessment of breast cancer. JCO Clin Cancer Inform. 2019 May;3:1-12 [FREE Full text] [doi: 10.1200/CCI.18.00121] [Medline: 31141423]
55. Golas SB, Shibahara T, Agboola S, Otaki H, Sato J, Nakae T, et al. A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: a retrospective analysis of electronic medical records data. BMC Med Inform Decis Mak. 2018 Jun 22;18(1):44 [FREE Full text] [doi: 10.1186/s12911-018-0620-z] [Medline: 29929496]
56. Sehdev A, Hayden R, Kuhar MJ, Cheng L, Warren SJ, Mark LA, et al. Prognostic role of BRAF mutation in malignant cutaneous melanoma. J Clin Oncol. 2018 May 20;36(15_suppl):e21599 [FREE Full text] [doi: 10.1200/jco.2018.36.15_suppl.e21599]
57. Riestenberg RA, Furman A, Cowen A, Pawlowksi A, Schneider D, Lewis AA, et al. Differences in statin utilization and lipid lowering by race, ethnicity, and HIV status in a real-world cohort of persons with human immunodeficiency virus and uninfected persons. Am Heart J. 2019 Mar;209:79-87 [FREE Full text] [doi: 10.1016/j.ahj.2018.11.012] [Medline: 30685678]
58. Abboud A, Nguonly A, Bean A, Brown KJ, Chen RF, Dudzinski D, et al. Rationale and design of the preserved versus reduced ejection fraction biomarker registry and precision medicine database for ambulatory patients with heart failure (PREFER-HF) study. Open Heart. 2021 Oct;8(2):e001704 [FREE Full text] [doi: 10.1136/openhrt-2021-001704] [Medline: 34663746]
59. Wang L, Xue Z, Ezeana CF, Puppala M, Chen S, Danforth RL, et al. Preventing inpatient falls with injuries using integrative machine learning prediction: a cohort study. NPJ Digit Med. 2019;2:127 [FREE Full text] [doi: 10.1038/s41746-019-0200-3] [Medline: 31872067]
60. Shoenbill K, Song Y, Craven M, Johnson H, Smith M, Mendonca EA. Identifying patterns and predictors of lifestyle modification in electronic health record documentation using statistical and machine learning methods. Prev Med. 2020 Jul;136:106061 [FREE Full text] [doi: 10.1016/j.ypmed.2020.106061] [Medline: 32179026]
61. Feller DJ, Zucker J, Yin MT, Gordon P, Elhadad N. Using clinical notes and natural language processing for automated HIV risk assessment. J Acquir Immune Defic Syndr. 2018 Feb 01;77(2):160-166 [FREE Full text] [doi: 10.1097/QAI.0000000000001580] [Medline: 29084046]
62. Bozkurt S, Kan KM, Ferrari MK, Rubin DL, Blayney DW, Hernandez-Boussard T, et al. Is it possible to automatically assess pretreatment digital rectal examination documentation using natural language processing? A single-centre retrospective study. BMJ Open. 2019 Jul 18;9(7):e027182 [FREE Full text] [doi: 10.1136/bmjopen-2018-027182] [Medline: 31324681]
63. Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis. J Am Med Inform Assoc. 2015 Jul;22(4):872-880 [FREE Full text] [doi: 10.1093/jamia/ocv024] [Medline: 25896647]

XSL•FO
RenderX

64. Klang E, Kummer BR, Dangayach NS, Zhong A, Kia MA, Timsina P, et al. Predicting adult neuroscience intensive care unit admission from emergency department triage using a retrospective, tabular-free text machine learning approach. Sci Rep. 2021 Jan 14;11(1):1381 [FREE Full text] [doi: 10.1038/s41598-021-80985-3] [Medline: 33446890]
65. Bae YS, Kim KH, Kim HK, Choi SW, Ko T, Seo HH, et al. Keyword extraction algorithm for classifying smoking status from unstructured bilingual electronic health records based on natural language processing. Appl Sci. 2021 Sep 22;11(19):8812 [FREE Full text] [doi: 10.3390/app11198812]
66. Stemerman R, Arguello J, Brice J, Krishnamurthy A, Houston M, Kitzmiller R. Identification of social determinants of health using multi-label classification of electronic health record clinical notes. JAMIA Open. 2021 Jul;4(3):ooaa069 [FREE Full text] [doi: 10.1093/jamiaopen/ooaa069] [Medline: 34514351]
67. Sharperson C, Hanna TN, Herr KD, Zygmont ME, Gerard RL, Johnson J. The effect of COVID-19 on emergency department imaging: what can we learn? Emerg Radiol. 2021 Apr;28(2):339-347 [FREE Full text] [doi: 10.1007/s10140-020-01889-9] [Medline: 33420529]
68. Moon S, Wen A, Scott C. An automated system for analysis of implantable cardioverter defibrillator reports in hypertrophic cardiomyopathy patients. Circulation. 2018;138(Suppl 1):A16215 [doi: 10.26226/morressier.5d19cfb257558b317a10dd93]
69. Zubke M, Katzensteiner M, Bott OJ. Stud Health Technol Inform. 2020 Jun 16;270:272-276 [doi: 10.3233/SHTI200165] [Medline: 32570389]
70. Ciofi Degli Atti ML, Pecoraro F, Piga S, Luzi D, Raponi M. Developing a surgical site infection surveillance system based on hospital unstructured clinical notes and text mining. Surg Infect (Larchmt). 2020 Oct;21(8):716-721 [doi: 10.1089/sur.2019.238] [Medline: 32105569]
71. Cochon LR, Giess CS, Khorasani R. Comparing diagnostic performance of digital breast tomosynthesis and full-field digital mammography. J Am Coll Radiol. 2020 Aug;17(8):999-1003 [doi: 10.1016/j.jacr.2020.01.010] [Medline: 32068009]
72. Lacson R, Wang A, Cochon L, Giess C, Desai S, Eappen S, et al. Factors associated with optimal follow-up in women with BI-RADS 3 breast findings. J Am Coll Radiol. 2020 Apr;17(4):469-474 [FREE Full text] [doi: 10.1016/j.jacr.2019.10.003] [Medline: 31669081]
73. Chrusciel J, Girardon F, Roquette L, Laplanche D, Duclos A, Sanchez S. The prediction of hospital length of stay using unstructured data. BMC Med Inform Decis Mak. 2021 Dec 18;21(1):351 [FREE Full text] [doi: 10.1186/s12911-021-01722-4] [Medline: 34922532]
74. Stein DM, Vawdrey DK, Stetson PD, Bakken S. An analysis of team checklists in physician signout notes. AMIA Annu Symp Proc. 2010 Nov 13;2010:767-771 [FREE Full text] [Medline: 21347082]
75. Chen W, Butler RK, Zhou Y, Parker RA, Jeon CY, Wu BU. Prediction of pancreatic cancer based on imaging features in patients with duct abnormalities. Pancreas. 2020 Mar;49(3):413-419 [FREE Full text] [doi: 10.1097/MPA.0000000000001499] [Medline: 32132511]
76. Rogier A, Coulet A, Rance B. Using an ontological representation of chemotherapy toxicities for guiding information extraction and integration from EHRs. Stud Health Technol Inform. 2022 Jun 06;290:91-95 [doi: 10.3233/SHTI220038] [Medline: 35672977]
77. Delespierre T, Denormandie P, Bar-Hen A, Josseran L. Empirical advances with text mining of electronic health records. BMC Med Inform Decis Mak. 2017 Aug 22;17(1):127 [FREE Full text] [doi: 10.1186/s12911-017-0519-0] [Medline: 28830417]
78. Wang L, Wampfler J, Dispenzieri A, Xu H, Yang P, Liu H. Achievability to extract specific date information for cancer research. AMIA Annu Symp Proc. 2019;2019:893-902 [FREE Full text] [Medline: 32308886]
79. Genes N, Chandra D, Ellis S, Baumlin K. Validating emergency department vital signs using a data quality engine for data warehouse. Open Med Inform J. 2013;7:34-39 [FREE Full text] [doi: 10.2174/1874431101307010034] [Medline: 24403981]
80. Glaser AP, Jordan BJ, Cohen J, Desai A, Silberman P, Meeks JJ. Automated extraction of grade, stage, and quality information from transurethral resection of bladder tumor pathology reports using natural language processing. JCO Clin Cancer Inform. 2018 Dec;2:1-8 [FREE Full text] [doi: 10.1200/CCI.17.00128] [Medline: 30652586]
81. Duthe JC, Bouzille G, Sylvestre E, Chazard E, Arvieux C, Cuggia M. How to identify potential candidates for HIV Pre-exposure prophylaxis: an AI algorithm reusing real-world hospital data. Stud Health Technol Inform. 2021 May 27;281:714-718 [doi: 10.3233/SHTI210265] [Medline: 34042669]
82. Lee KH, Kim HJ, Kim YJ, Kim JH, Song EY. Extracting structured genotype information from free-text HLA reports using a rule-based approach. J Korean Med Sci. 2020 Mar 30;35(12):e78 [FREE Full text] [doi: 10.3346/jkms.2020.35.e78] [Medline: 32233158]
83. Tamang S, Patel MI, Blayney DW, Kuznetsov J, Finlayson SG, Vetteth Y, et al. Detecting unplanned care from clinician notes in electronic health records. J Oncol Pract. 2015 May;11(3):e313-e319 [FREE Full text] [doi: 10.1200/JOP.2014.002741] [Medline: 25980019]
84. Yang X, Yang H, Lyu T, Yang S, Guo Y, Bian J, et al. A natural language processing tool to extract quantitative smoking status from clinical narratives. IEEE Int Conf Healthc Inform. 2020;2020:1109 [FREE Full text] [doi: 10.1109/ICHI48887.2020.9374369] [Medline: 33786419]

XSL•FO
RenderX

85. Wang Y, Mehrabi S, Sohn S, Atkinson EJ, Amin S, Liu H. Natural language processing of radiology reports for identification of skeletal site-specific fractures. BMC Med Inform Decis Mak. 2019 Apr 04;19(Suppl 3):73 [FREE Full text] [doi: 10.1186/s12911-019-0780-5] [Medline: 30943952]
86. Zhao Y, Weroha SJ, Goode EL, Liu H, Wang C. Generating real-world evidence from unstructured clinical notes to examine clinical utility of genetic tests: use case in BRCAness. BMC Med Inform Decis Mak. 2021 Jan 06;21(1):3 [FREE Full text] [doi: 10.1186/s12911-020-01364-y] [Medline: 33407429]
87. Haug PJ, Ferraro JP, Holmen J, Wu X, Mynam K, Ebert M, et al. An ontology-driven, diagnostic modeling system. J Am Med Inform Assoc. 2013 Jun;20(e1):e102-e110 [FREE Full text] [doi: 10.1136/amiajnl-2012-001376] [Medline: 23523876]
88. Magnani CJ, Bievre N, Baker LC, Brooks JD, Blayney DW, Hernandez-Boussard T. Real-world evidence to estimate prostate cancer costs for first-line treatment or active surveillance. Eur Urol Open Sci. 2021 Jan;23:20-29 [FREE Full text] [doi: 10.1016/j.euros.2020.11.004] [Medline: 33367287]
89. Ansoborlo M, Dhalluin T, Gaborit C, Cuggia M, Grammatico-Guilllon L. Prescreening in oncology using data sciences: the PreScIOUS study. Stud Health Technol Inform. 2021 May 27;281:123-127 [doi: 10.3233/SHTI210133] [Medline: 34042718]
90. Ryu JH, Zimolzak AJ. Natural language processing of serum protein electrophoresis reports in the veterans affairs health care system. JCO Clin Cancer Inform. 2020 Aug;4:749-756 [FREE Full text] [doi: 10.1200/CCI.19.00167] [Medline: 32813561]
91. Luther SL, Thomason SS, Sabharwal S, Finch DK, McCart J, Toyinbo P, et al. Leveraging electronic health care record information to measure pressure ulcer risk in veterans with spinal cord injury: a longitudinal study protocol. JMIR Res Protoc. 2017 Jan 19;6(1):e3 [FREE Full text] [doi: 10.2196/resprot.5948] [Medline: 28104580]
92. Liu Y, Lependu P, Iyer S, Shah NH. Using temporal patterns in medical records to discern adverse drug events from indications. AMIA Jt Summits Transl Sci Proc. 2012;2012:47-56 [FREE Full text] [Medline: 22779050]
93. Lerner I, Jouffroy J, Burgun A. Learning the grammar of drug prescription: recurrent neural network grammars for medication information extraction in clinical texts. arXiv. Preprint posted online April 24, 2020. 2020 [FREE Full text] [doi: 10.48550/arXiv.2004.11622]
94. Hoertel N, Sánchez-Rico M, Vernet R, Beeker N, Neuraz A, Alvarado JM, et al. AP-HP/Université de Paris/INSERM Covid-19 research collaborationAP-HP Covid CDR Initiative. Dexamethasone use and mortality in hospitalized patients with coronavirus disease 2019: a multicentre retrospective observational study. Br J Clin Pharmacol. 2021 Oct;87(10):3766-3775 [FREE Full text] [doi: 10.1111/bcp.14784] [Medline: 33608891]
95. Caliskan D, Zierk J, Kraska D, Schulz S, Daumke P, Prokosch HU, et al. First steps to evaluate an NLP tool's medication extraction accuracy from discharge letters. Stud Health Technol Inform. 2021 May 24;278:224-230 [doi: 10.3233/SHTI210073] [Medline: 34042898]
96. Rochefort CM, Buckeridge DL, Abrahamowicz M. Improving patient safety by optimizing the use of nursing human resources. Implement Sci. 2015 Jun 14;10(1):89 [FREE Full text] [doi: 10.1186/s13012-015-0278-1] [Medline: 26071752]
97. Wang G, Jung K, Winnenburg R, Shah NH. A method for systematic discovery of adverse drug events from clinical notes. J Am Med Inform Assoc. 2015 Nov;22(6):1196-1204 [FREE Full text] [doi: 10.1093/jamia/ocv102] [Medline: 26232442]
98. Shimai Y, Takeda T, Okada K, Manabe S, Teramoto K, Mihara N, et al. Screening of anticancer drugs to detect drug-induced interstitial pneumonia using the accumulated data in the electronic medical record. Pharmacol Res Perspect. 2018 Jul;6(4):e00421 [FREE Full text] [doi: 10.1002/prp2.421] [Medline: 30009034]
99. Jung K, Lependu P, Shah N. Automated detection of systematic off-label drug use in free text of electronic medical records. AMIA Jt Summits Transl Sci Proc. 2013;2013:94-98 [FREE Full text] [Medline: 24303308]
100. Geva A, Abman S, Manzi S, Ivy DD, Mullen MP, Griffin J, et al. Adverse drug event rates in pediatric pulmonary hypertension: a comparison of real-world data sources. J Am Med Inform Assoc. 2020 Feb 01;27(2):294-300 [FREE Full text] [doi: 10.1093/jamia/ocz194] [Medline: 31769835]
101. Rochefort CM, Buckeridge DL, Tanguay A, Biron A, D'Aragon F, Wang S, et al. Accuracy and generalizability of using automated methods for identifying adverse events from electronic health record data: a validation study protocol. BMC Health Serv Res. 2017 Feb 16;17(1):147 [FREE Full text] [doi: 10.1186/s12913-017-2069-7] [Medline: 28209197]
102. Gold S, Elhadad N, Zhu X, Cimino JJ, Hripcsak G. Extracting structured medication event information from discharge summaries. AMIA Annu Symp Proc. 2008 Nov 06;2008:237-241 [FREE Full text] [Medline: 18999147]
103. Li Y, Salmasian H, Harpaz R, Chase H, Friedman C. Determining the reasons for medication prescriptions in the EHR using knowledge and natural language processing. AMIA Annu Symp Proc. 2011;2011:768-776 [FREE Full text] [Medline: 22195134]
104. Harpaz R, Vilar S, Dumouchel W, Salmasian H, Haerian K, Shah NH, et al. Combing signals from spontaneous reports and electronic health records for detection of adverse drug reactions. J Am Med Inform Assoc. 2013 May 01;20(3):413-419 [FREE Full text] [doi: 10.1136/amiajnl-2012-000930] [Medline: 23118093]
105. Zhao Y, Dimou A, Shen F, Zong N, Davila JI, Liu H, et al. PO2RDF: representation of real-world data for precision oncology using resource description framework. BMC Med Genomics. 2022 Jul 30;15(1):167 [FREE Full text] [doi: 10.1186/s12920-022-01314-9] [Medline: 35907849]

XSL•FO
RenderX

106. Lependu P, Liu Y, Iyer S, Udell MR, Shah NH. Analyzing patterns of drug use in clinical notes for patient safety. AMIA Jt Summits Transl Sci Proc. 2012;2012:63-70 [FREE Full text] [Medline: 22779054]
107. Lependu P, Iyer SV, Fairon C, Shah NH. Annotation analysis for testing drug safety signals using unstructured clinical notes. J Biomed Semantics. 2012 Apr 24;3 Suppl 1(Suppl 1):S5 [FREE Full text] [doi: 10.1186/2041-1480-3-S1-S5] [Medline: 22541596]
108. LePendu P, Iyer SV, Bauer-Mehren A, Harpaz R, Mortensen JM, Podchiyska T, et al. Pharmacovigilance using clinical notes. Clin Pharmacol Ther. 2013 Jun;93(6):547-555 [FREE Full text] [doi: 10.1038/clpt.2013.47] [Medline: 23571773]
109. Jung K, LePendu P, Iyer S, Bauer-Mehren A, Percha B, Shah NH. Functional evaluation of out-of-the-box text-mining tools for data-mining tasks. J Am Med Inform Assoc. 2015 Jan;22(1):121-131 [FREE Full text] [doi: 10.1136/amiajnl-2014-002902] [Medline: 25336595]
110. Wright A, McCoy A, Henkin S, Flaherty M, Sittig D. Validation of an association rule mining-based method to infer associations between medications and problems. Appl Clin Inform. 2013;4(1):100-109 [FREE Full text] [doi: 10.4338/ACI-2012-12-RA-0051] [Medline: 23650491]
111. Malec SA, Wei P, Bernstam EV, Boyce RD, Cohen T. Using computable knowledge mined from the literature to elucidate confounders for EHR-based pharmacovigilance. J Biomed Inform. 2021 May;117:103719 [FREE Full text] [doi: 10.1016/j.jbi.2021.103719] [Medline: 33716168]
112. Weeks HL, Beck C, McNeer E, Williams ML, Bejan CA, Denny JC, et al. medExtractR: a targeted, customizable approach to medication extraction from electronic health records. J Am Med Inform Assoc. 2020 Mar 01;27(3):407-418 [FREE Full text] [doi: 10.1093/jamia/ocz207] [Medline: 31943012]
113. Chouchana L, Beeker N, Garcelon N, Rance B, Paris N, Salamanca E, et al. AP-HP/Universities/Inserm COVID-19 research collaboration, AP-HP Covid CDR Initiative,"Entrepôt de Données de Santé" AP-HP Consortium". Association of antihypertensive agents with the risk of in-hospital death in patients with COVID-19. Cardiovasc Drugs Ther. 2022 Jun;36(3):483-488 [FREE Full text] [doi: 10.1007/s10557-021-07155-5] [Medline: 33595761]
114. Leeper NJ, Bauer-Mehren A, Iyer SV, Lependu P, Olson C, Shah NH. Practice-based evidence: profiling the safety of cilostazol by text-mining of clinical notes. PLoS One. 2013;8(5):e63499 [FREE Full text] [doi: 10.1371/journal.pone.0063499] [Medline: 23717437]
115. Jouffroy J, Feldman SF, Lerner I, Rance B, Burgun A, Neuraz A. Hybrid deep learning for medication-related information extraction from clinical texts in french: MedExt algorithm development study. JMIR Med Inform. 2021 Mar 16;9(3):e17934 [FREE Full text] [doi: 10.2196/17934] [Medline: 33724196]
116. Min TL, Xu L, Choi JD, Hu R, Allen JW, Reeves C, et al. COVID-19 pandemic-associated changes in the acuity of brain MRI findings: a secondary analysis of reports using natural language processing. Curr Probl Diagn Radiol. 2022;51(4):529-533 [FREE Full text] [doi: 10.1067/j.cpradiol.2021.11.001] [Medline: 34955284]
117. Fiebeck J, Laser H, Winther HB, Gerbel S. Leaving no stone unturned: using machine learning based approaches for information extraction from full texts of a research data warehouse. In: Proceedings of the 13th International Conference on Data Integration in the Life Sciences. 2018 Presented at: DILS '18; November 20-21, 2018; Hannover, Germany p. 50-58 URL: https://link.springer.com/chapter/10.1007/978-3-030-06016-9_5 [doi: 10.1007/978-3-030-06016-9_5]
118. Pham AD, Névéol A, Lavergne T, Yasunaga D, Clément O, Meyer G, et al. Natural language processing of radiology reports for the detection of thromboembolic diseases and clinically relevant incidental findings. BMC Bioinformatics. 2014 Aug 07;15(1):266 [FREE Full text] [doi: 10.1186/1471-2105-15-266] [Medline: 25099227]
119. Chokshi FH, Shin B, Lee T. Natural language processing for classification of acute, communicable findings on unstructured head CT reports: comparison of neural network and non-neural machine learning techniques. bioRxiv. Preprint posted online August 10, 2017. 2017 [FREE Full text] [doi: 10.1101/173310]
120. Cao H, Markatou M, Melton GB, Chiang MF, Hripcsak G. Mining a clinical data warehouse to discover disease-finding associations using co-occurrence statistics. AMIA Annu Symp Proc. 2005;2005:106-110 [FREE Full text] [Medline: 16779011]
121. Patel TA, Puppala M, Ogunti RO, Ensor JE, He T, Shewale JB, et al. Correlating mammographic and pathologic findings in clinical decision support using natural language processing and data mining methods. Cancer. 2017 Jan 01;123(1):114-121 [FREE Full text] [doi: 10.1002/cncr.30245] [Medline: 27571243]
122. Olmsted ZT, Hadanny A, Marchese AM, DiMarzio M, Khazen O, Argoff C, et al. Recommendations for neuromodulation in diabetic neuropathic pain. Front Pain Res (Lausanne). 2021 Sep 07;2:726308 [FREE Full text] [doi: 10.3389/fpain.2021.726308] [Medline: 35295414]
123. He T, Puppala M, Ogunti R. Deep learning analytics for diagnostic support of breast cancer disease management. In: Proceedings of the 2017 IEEE EMBS International Conference on Biomedical & Health Informatic. 2017 Presented at: BHI '17; February 16-19, 2017; Orlando, FL p. 365-368 URL: https://ieeexplore.ieee.org/document/7897281 [doi: 10.1109/bhi.2017.7897281]
124. He T, Fong JN, Moore LW, Ezeana CF, Victor D, Divatia M, et al. An imageomics and multi-network based deep learning model for risk assessment of liver transplantation for hepatocellular cancer. Comput Med Imaging Graph. 2021 Apr;89:101894 [FREE Full text] [doi: 10.1016/j.compmedimag.2021.101894] [Medline: 33725579]

XSL•FO
RenderX

125. Lo Barco T, Kuchenbuch M, Garcelon N, Neuraz A, Nabbout R. Improving early diagnosis of rare diseases using natural language processing in unstructured medical records: an illustration from Dravet syndrome. Orphanet J Rare Dis. 2021 Jul 13;16(1):309 [FREE Full text] [doi: 10.1186/s13023-021-01936-9] [Medline: 34256808]
126. Alba PR, Gao A, Lee KM, Anglin-Foote T, Robison B, Katsoulakis E, et al. Ascertainment of veterans with metastatic prostate cancer in electronic health records: demonstrating the case for natural language processing. JCO Clin Cancer Inform. 2021 Sep;5:1005-1014 [FREE Full text] [doi: 10.1200/CCI.21.00030] [Medline: 34570630]
127. Zhu VJ, Lenert LA, Bunnell BE, Obeid JS, Jefferson M, Halbert CH. Automatically identifying social isolation from clinical narratives for patients with prostate Cancer. BMC Med Inform Decis Mak. 2019 Mar 14;19(1):43 [FREE Full text] [doi: 10.1186/s12911-019-0795-y] [Medline: 30871518]
128. To D, Sharma B, Karnik N, Joyce C, Dligach D, Afshar M. Validation of an alcohol misuse classifier in hospitalized patients. Alcohol. 2020 May;84:49-55 [doi: 10.1016/j.alcohol.2019.09.008] [Medline: 31574300]
129. Zong N, Ngo V, Stone DJ, Wen A, Zhao Y, Yu Y, et al. Leveraging genetic reports and electronic health records for the prediction of primary cancers: algorithm development and validation study. JMIR Med Inform. 2021 May 25;9(5):e23586 [FREE Full text] [doi: 10.2196/23586] [Medline: 34032581]
130. De Freitas JK, Johnson KW, Golden E, Nadkarni GN, Dudley JT, Bottinger EP, et al. Phe2vec: automated disease phenotyping based on unsupervised embeddings from electronic health records. Patterns (N Y). 2021 Sep 10;2(9):100337 [FREE Full text] [doi: 10.1016/j.patter.2021.100337] [Medline: 34553174]
131. Carter GC, Landsman-Blumberg PB, Johnson BH, Juneau P, Nicol SJ, Li L, et al. KRAS testing of patients with metastatic colorectal cancer in a community-based oncology setting: a retrospective database analysis. J Exp Clin Cancer Res. 2015 Mar 27;34(1):29 [FREE Full text] [doi: 10.1186/s13046-015-0146-5] [Medline: 25888436]
132. Banda JM, Halpern Y, Sontag D, Shah NH. Electronic phenotyping with APHRODITE and the Observational Health Sciences and Informatics (OHDSI) data network. AMIA Jt Summits Transl Sci Proc. 2017;2017:48-57 [FREE Full text] [Medline: 28815104]
133. Shao Y, Zeng QT, Chen KK, Shutes-David A, Thielke SM, Tsuang DW. Detection of probable dementia cases in undiagnosed patients using structured and unstructured electronic health records. BMC Med Inform Decis Mak. 2019 Jul 09;19(1):128 [FREE Full text] [doi: 10.1186/s12911-019-0846-4] [Medline: 31288818]
134. Sharma B, Dligach D, Swope K, Salisbury-Afshar E, Karnik NS, Joyce C, et al. Publicly available machine learning models for identifying opioid misuse from the clinical notes of hospitalized patients. BMC Med Inform Decis Mak. 2020 Apr 29;20(1):79 [FREE Full text] [doi: 10.1186/s12911-020-1099-y] [Medline: 32349766]
135. Lee J, Liu C, Kim JH, Butler A, Shang N, Pang C, et al. Comparative effectiveness of medical concept embedding for feature engineering in phenotyping. JAMIA Open. 2021 Apr;4(2):ooab028 [FREE Full text] [doi: 10.1093/jamiaopen/ooab028] [Medline: 34142015]
136. Garcelon N, Neuraz A, Salomon R, Bahi-Buisson N, Amiel J, Picard C, et al. Next generation phenotyping using narrative reports in a rare disease clinical data warehouse. Orphanet J Rare Dis. 2018 May 31;13(1):85 [FREE Full text] [doi: 10.1186/s13023-018-0830-6] [Medline: 29855327]
137. Afzal N, Mallipeddi VP, Sohn S, Liu H, Chaudhry R, Scott CG, et al. Natural language processing of clinical notes for identification of critical limb ischemia. Int J Med Inform. 2018 Mar;111:83-89 [FREE Full text] [doi: 10.1016/j.ijmedinf.2017.12.024] [Medline: 29425639]
138. Hully M, Lo Barco T, Kaminska A, Barcia G, Cances C, Mignot C, et al. Deep phenotyping unstructured data mining in an extensive pediatric database to unravel a common KCNA2 variant in neurodevelopmental syndromes. Genet Med. 2021 May;23(5):968-971 [FREE Full text] [doi: 10.1038/s41436-020-01039-z] [Medline: 33500571]
139. Chen X, Garcelon N, Neuraz A, Billot K, Lelarge M, Bonald T, et al. Phenotypic similarity for rare disease: ciliopathy diagnoses and subtyping. J Biomed Inform. 2019 Dec;100:103308 [FREE Full text] [doi: 10.1016/j.jbi.2019.103308] [Medline: 31622800]
140. Bastarache L, Hughey JJ, Goldstein JA, Bastraache JA, Das S, Zaki NC, et al. Improving the phenotype risk score as a scalable approach to identifying patients with Mendelian disease. J Am Med Inform Assoc. 2019 Dec 01;26(12):1437-1447 [FREE Full text] [doi: 10.1093/jamia/ocz179] [Medline: 31609419]
141. Stephen R, Boxwala A, Gertman P. Feasibility of using a large clinical data warehouse to automate the selection of diagnostic cohorts. AMIA Annu Symp Proc. 2003;2003:1019 [FREE Full text] [Medline: 14728522]
142. Yahi A, Tatonetti NP. A knowledge-based, automated method for phenotyping in the EHR using only clinical pathology reports. AMIA Jt Summits Transl Sci Proc. 2015;2015:64-68 [FREE Full text] [Medline: 26306239]
143. Hoffman SR, Vines AI, Halladay JR, Pfaff E, Schiff L, Westreich D, et al. Optimizing research in symptomatic uterine fibroids with development of a computable phenotype for use with electronic health records. Am J Obstet Gynecol. 2018 Jun;218(6):610.e1-610.e7 [FREE Full text] [doi: 10.1016/j.ajog.2018.02.002] [Medline: 29432754]
144. Haerian K, Salmasian H, Friedman C. Methods for identifying suicide or suicidal ideation in EHRs. AMIA Annu Symp Proc. 2012;2012:1244-1253 [FREE Full text] [Medline: 23304402]
145. Evans RS, Benuzillo J, Horne BD, Lloyd JF, Bradshaw A, Budge D, et al. Automated identification and predictive tools to help identify high-risk heart failure patients: pilot evaluation. J Am Med Inform Assoc. 2016 Sep;23(5):872-878 [doi: 10.1093/jamia/ocv197] [Medline: 26911827]

XSL•FO
RenderX

146. Upadhyaya SG, Murphree Jr DH, Ngufor CG, Knight AM, Cronk DJ, Cima RR, et al. Automated diabetes case identification using electronic health record data at a tertiary care facility. Mayo Clin Proc Innov Qual Outcomes. 2017 Jul;1(1):100-110 [FREE Full text] [doi: 10.1016/j.mayocpiqo.2017.04.005] [Medline: 30225406]
147. Ahmed A, Thongprayoon C, Pickering BW, Akhoundi A, Wilson G, Pieczkiewicz D, et al. Towards prevention of acute syndromes: electronic identification of at-risk patients during hospital admission. Appl Clin Inform. 2014;5(1):58-72 [FREE Full text] [doi: 10.4338/ACI-2013-07-RA-0045] [Medline: 24734124]
148. Redman JS, Natarajan Y, Hou JK, Wang J, Hanif M, Feng H, et al. Accurate identification of fatty liver disease in data warehouse utilizing natural language processing. Dig Dis Sci. 2017 Oct;62(10):2713-2718 [doi: 10.1007/s10620-017-4721-9] [Medline: 28861720]
149. Nigwekar SU, Solid CA, Ankers E, Malhotra R, Eggert W, Turchin A, et al. Quantifying a rare disease in administrative data: the example of calciphylaxis. J Gen Intern Med. 2014 Aug;29 Suppl 3(Suppl 3):S724-S731 [FREE Full text] [doi: 10.1007/s11606-014-2910-1] [Medline: 25029979]
150. Krebs J, Bittrich M, Dietrich G, Ertl M, Fette G, Kaspar M, et al. Finding needles in the haystack: identifying patients with rare subtype of multiple myeloma supported by a data warehouse and information extraction. Stud Health Technol Inform. 2018;253:160-164 [Medline: 30147064]
151. Coquet J, Bozkurt S, Kan KM, Ferrari MK, Blayney DW, Brooks JD, et al. Comparison of orthogonal NLP methods for clinical phenotyping and assessment of bone scan utilization among prostate cancer patients. J Biomed Inform. 2019 Jun;94:103184 [FREE Full text] [doi: 10.1016/j.jbi.2019.103184] [Medline: 31014980]
152. Bozkurt S, Paul R, Coquet J, Sun R, Banerjee I, Brooks JD, et al. Phenotyping severity of patient-centered outcomes using clinical notes: a prostate cancer use case. Learn Health Syst. 2020 Oct;4(4):e10237 [FREE Full text] [doi: 10.1002/lrh2.10237] [Medline: 33083539]
153. Meystre SM, Heider PM, Kim Y, Aruch DB, Britten CD. Automatic trial eligibility surveillance based on unstructured clinical data. Int J Med Inform. 2019 Sep;129:13-19 [FREE Full text] [doi: 10.1016/j.ijmedinf.2019.05.018] [Medline: 31445247]
154. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc. 2016 Dec;23(6):1166-1173 [FREE Full text] [doi: 10.1093/jamia/ocw028] [Medline: 27174893]
155. Ferté T, Cossin S, Schaeverbeke T, Barnetche T, Jouhet V, Hejblum BP. Automatic phenotyping of electronical health record: PheVis algorithm. J Biomed Inform. 2021 May;117:103746 [FREE Full text] [doi: 10.1016/j.jbi.2021.103746] [Medline: 33746080]
156. Chase HS, Radhakrishnan J, Shirazian S, Rao MK, Vawdrey DK. Under-documentation of chronic kidney disease in the electronic health record in outpatients. J Am Med Inform Assoc. 2010;17(5):588-594 [doi: 10.1136/jamia.2009.001396] [Medline: 20819869]
157. Kim C, Zhu V, Obeid J, Lenert L. Natural language processing and machine learning algorithm to identify brain MRI reports with acute ischemic stroke. PLoS One. 2019;14(2):e0212778 [FREE Full text] [doi: 10.1371/journal.pone.0212778] [Medline: 30818342]
158. Kim JH, Hua M, Whittington RA, Lee J, Liu C, Ta CN, et al. A machine learning approach to identifying delirium from electronic health records. JAMIA Open. 2022 Jul;5(2):ooac042 [FREE Full text] [doi: 10.1093/jamiaopen/ooac042] [Medline: 35663114]
159. Zuo X, Li J, Zhao B, Zhou Y, Dong X, Duke J, et al. Normalizing clinical document titles to LOINC document ontology: an initial study. AMIA Annu Symp Proc. 2021 Jan 25;2020:1441-1450 [FREE Full text] [Medline: 33936520]
160. Scheurwegs E, Luyckx K, Luyten L, Daelemans W, Van den Bulcke T. Data integration of structured and unstructured sources for assigning clinical codes to patient stays. J Am Med Inform Assoc. 2016 Apr;23(e1):e11-e19 [FREE Full text] [doi: 10.1093/jamia/ocv115] [Medline: 26316458]
161. Zhu D, Wu S, Carterette B, Liu H. Using large clinical corpora for query expansion in text-based cohort identification. J Biomed Inform. 2014 Jun;49:275-281 [FREE Full text] [doi: 10.1016/j.jbi.2014.03.010] [Medline: 24680983]
162. Tien M, Kashyap R, Wilson GA, Hernandez-Torres V, Jacob AK, Schroeder DR, et al. Retrospective derivation and validation of an automated electronic search algorithm to identify post operative cardiovascular and thromboembolic complications. Appl Clin Inform. 2015;6(3):565-576 [FREE Full text] [doi: 10.4338/ACI-2015-03-RA-0026] [Medline: 26448798]
163. Lelong R, Soualmia LF, Grosjean J, Taalba M, Darmoni SJ. Building a semantic health data warehouse in the context of clinical trials: development and usability study. JMIR Med Inform. 2019 Dec 20;7(4):e13917 [FREE Full text] [doi: 10.2196/13917] [Medline: 31859675]
164. Pressat-Laffouilhère T, Balayé P, Dahamna B, Lelong R, Billey K, Darmoni SJ, et al. Evaluation of Doc'EDS: a French semantic search tool to query health documents from a clinical data warehouse. BMC Med Inform Decis Mak. 2022 Feb 08;22(1):34 [FREE Full text] [doi: 10.1186/s12911-022-01762-4] [Medline: 35135538]
165. Zeng QT, Redd D, Rindflesch T, Nebeker J. Synonym, topic model and predicate-based query expansion for retrieving clinical documents. AMIA Annu Symp Proc. 2012;2012:1050-1059 [FREE Full text] [Medline: 23304381]

XSL•FO
RenderX

166. Li M, Lee K, Liu Z, Ma M, Pan Q, Chen R, et al. Applying Bayesian hyperparameter optimization towards accurate and efficient topic modeling in clinical notes. In: Proceedings of the IEEE 9th International Conference on Healthcare Informatics. 2021 Presented at: ICHI '21; August 9-12, 2021; Victoria, BC p. 493-494 URL: https://ieeexplore.ieee.org/document/9565781 [doi: 10.1109/ichi52183.2021.00086]
167. Chen JH, Goldstein MK, Asch SM, Mackey L, Altman RB. Predicting inpatient clinical order patterns with probabilistic topic models vs conventional order sets. J Am Med Inform Assoc. 2017 May 01;24(3):472-480 [FREE Full text] [doi: 10.1093/jamia/ocw136] [Medline: 27655861]
168. Afshar M, Joyce C, Dligach D, Sharma B, Kania R, Xie M, et al. Subtypes in patients with opioid misuse: a prognostic enrichment strategy using electronic health record data in hospitalized patients. PLoS One. 2019;14(7):e0219717 [FREE Full text] [doi: 10.1371/journal.pone.0219717] [Medline: 31310611]
169. Ling AY, Kurian AW, Caswell-Jin JL, Sledge GW, Shah NH, Tamang SR. Using natural language processing to construct a metastatic breast cancer cohort from linked cancer registry and electronic medical records data. JAMIA Open. 2019 Dec;2(4):528-537 [FREE Full text] [doi: 10.1093/jamiaopen/ooz040] [Medline: 32025650]
170. Wu DW, Bernstein JA, Bejerano G. Discovering monogenic patients with a confirmed molecular diagnosis in millions of clinical notes with MonoMiner. Genet Med. 2022 Oct;24(10):2091-2102 [FREE Full text] [doi: 10.1016/j.gim.2022.07.008] [Medline: 35976265]
171. Chen CJ, Warikoo N, Chang Y, Chen J, Hsu W. Medical knowledge infused convolutional neural networks for cohort selection in clinical trials. J Am Med Inform Assoc. 2019 Nov 01;26(11):1227-1236 [FREE Full text] [doi: 10.1093/jamia/ocz128] [Medline: 31390470]
172. Mutinda FW, Nigo S, Wakamiya S, Aramaki E. Detecting redundancy in electronic medical records using clinical BERT. The Association for Natural Language Processing. 2020. URL: https://www.anlp.jp/proceedings/annual_meeting/2020/pdf_dir/E3-3.pdf [accessed 2023-11-27]
173. Mahajan D, Poddar A, Liang JJ, Lin Y, Prager JM, Suryanarayanan P, et al. Identification of semantically similar sentences in clinical notes: iterative intermediate training using multi-task learning. JMIR Med Inform. 2020 Nov 27;8(11):e22508 [FREE Full text] [doi: 10.2196/22508] [Medline: 33245284]
174. Li J, Zhang X, Zhou X. ALBERT-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation: algorithm validation study. JMIR Med Inform. 2021 Jan 22;9(1):e23086 [FREE Full text] [doi: 10.2196/23086] [Medline: 33480858]
175. Pivovarov R, Elhadad N. A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts. J Biomed Inform. 2012 Jun;45(3):471-481 [FREE Full text] [doi: 10.1016/j.jbi.2012.01.002] [Medline: 22289420]
176. Cohen R, Elhadad M, Elhadad N. Redundancy in electronic health record corpora: analysis, impact on text mining performance and mitigation strategies. BMC Bioinformatics. 2013 Jan 16;14:10 [FREE Full text] [doi: 10.1186/1471-2105-14-10] [Medline: 23323800]
177. Garcelon N, Neuraz A, Benoit V, Salomon R, Kracker S, Suarez F, et al. Finding patients using similarity measures in a rare diseases-oriented clinical data warehouse: Dr. Warehouse and the needle in the needle stack. J Biomed Inform. 2017 Sep;73:51-61 [FREE Full text] [doi: 10.1016/j.jbi.2017.07.016] [Medline: 28754522]
178. Mirzapour M, Abdaoui A, Tchechmedjiev A, Digan W, Bringay S, Jonquet C. French FastContext: a publicly accessible system for detecting negation, temporality and experiencer in French clinical notes. J Biomed Inform. 2021 May;117:103733 [FREE Full text] [doi: 10.1016/j.jbi.2021.103733] [Medline: 33737205]
179. Zhou L, Melton GB, Parsons S, Hripcsak G. A temporal constraint structure for extracting temporal information from clinical narrative. J Biomed Inform. 2006 Aug;39(4):424-439 [FREE Full text] [doi: 10.1016/j.jbi.2005.07.002] [Medline: 16169282]
180. Klappe ES, van Putten FJ, de Keizer NF, Cornet R. Contextual property detection in Dutch diagnosis descriptions for uncertainty, laterality and temporality. BMC Med Inform Decis Mak. 2021 Apr 07;21(1):120 [FREE Full text] [doi: 10.1186/s12911-021-01477-y] [Medline: 33827555]
181. Lin C, Bethard S, Dligach D, Sadeque F, Savova G, Miller TA. Does BERT need domain adaptation for clinical negation detection? J Am Med Inform Assoc. 2020 Apr 01;27(4):584-591 [FREE Full text] [doi: 10.1093/jamia/ocaa001] [Medline: 32044989]
182. Garcelon N, Neuraz A, Benoit V, Salomon R, Burgun A. Improving a full-text search engine: the importance of negation detection and family history context to identify cases in a biomedical data warehouse. J Am Med Inform Assoc. 2017 May 01;24(3):607-613 [doi: 10.1093/jamia/ocw144] [Medline: 28339516]
183. Cossin S, Jolly M, Larrouture I, Griffier R, Jouhet V. Semi-automatic extraction of abbreviations and their senses from electronic health records. ResearchGate. Preprint posted online July 3, 2023. 2021 [FREE Full text]
184. Moon S, Ihrke D, Zeng Y, Liu H. Distinction between medical and non-medical usages of short forms in clinical narratives. AMIA Annu Symp Proc. 2017;2017:1302-1311 [FREE Full text] [Medline: 29854199]
185. Peng Y, Yan S, Lu Z. 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. 2019 Presented at: BioNLP '19; August 1, 2019; Florence, Italy p. 58-65 URL: https://aclanthology.org/W19-5006.pdf [doi: 10.18653/v1/w19-5006]

XSL•FO
RenderX

186. Peng Y, Yan S, Lu Z. An empirical study of multi-task learning on BERT for biomedical text mining. In: Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing. 2020 Presented at: BioNLP '20; July 9, 2020; Virtual Event p. 205-214 URL: https://aclanthology.org/2020.bionlp-1.22.pdf [doi: 10.18653/v1/2020.bionlp-1.22]
187. Tawfik NS, Spruit MR. Evaluating sentence representations for biomedical text: methods and experimental results. J Biomed Inform. 2020 Apr;104:103396 [doi: 10.1016/j.jbi.2020.103396] [Medline: 32147441]
188. Neuraz A, Looten V, Rance B, Daniel N, Garcelon N, Llanos LC, et al. Do you need embeddings trained on a massive specialized corpus for your clinical natural language processing task? Stud Health Technol Inform. 2019 Aug 21;264:1558-1559 [doi: 10.3233/SHTI190533] [Medline: 31438230]
189. Dligach D, Afshar M, Miller T. Toward a clinical text encoder: pretraining for clinical natural language processing with applications to substance misuse. J Am Med Inform Assoc. 2019 Nov 01;26(11):1272-1278 [FREE Full text] [doi: 10.1093/jamia/ocz072] [Medline: 31233140]
190. Lee YC, Jung S, Kumar A, Shim I, Song M, Kim MS, et al. ICD2Vec: Mathematical representation of diseases. J Biomed Inform. 2023 May;141:104361 [doi: 10.1016/j.jbi.2023.104361] [Medline: 37054960]
191. Zhan X, Humbert-Droz M, Mukherjee P, Gevaert O. Structuring clinical text with AI: old versus new natural language processing techniques evaluated on eight common cardiovascular diseases. Patterns (N Y). 2021 Jul 09;2(7):100289 [FREE Full text] [doi: 10.1016/j.patter.2021.100289] [Medline: 34286303]
192. Dubois S, Kale DC, Romano N, Shah N, Jung K. Learning effective representations from clinical Nnotes. arXiv. Preprint posted online May 19, 2017. 2017 [FREE Full text] [doi: 10.48550/arXiv.1705.07025]
193. Dynomant E, Lelong R, Dahamna B, Massonnaud C, Kerdelhué G, Grosjean J, et al. Word embedding for the French natural language in health care: comparative study. JMIR Med Inform. 2019 Jul 29;7(3):e12310 [FREE Full text] [doi: 10.2196/12310] [Medline: 31359873]
194. Lee D, Jiang X, Yu H. Harmonized representation learning on dynamic EHR graphs. J Biomed Inform. 2020 Jun;106:103426 [FREE Full text] [doi: 10.1016/j.jbi.2020.103426] [Medline: 32339747]
195. Roberts K, Si Y, Gandhi A, Bernstam E. A FrameNet for cancer information in clinical narratives: schema and annotation. In: Proceedings of the 11th International Conference on Language Resources and Evaluation. 2018 Presented at: LREC '18; July 15-18, 2018; Miyazaki, Japan p. 272-279 URL: https://aclanthology.org/L18-1041.pdf
196. Van Vleck TT, Stein DM, Stetson PD, Johnson SB. Assessing data relevance for automated generation of a clinical summary. AMIA Annu Symp Proc. 2007 Oct 11;2007:761-765 [FREE Full text] [Medline: 18693939]
197. Escudié JB, Jannot AS, Zapletal E, Cohen S, Malamut G, Burgun A, et al. Reviewing 741 patients records in two hours with FASTVISU. AMIA Annu Symp Proc. 2015;2015:553-559 [FREE Full text] [Medline: 26958189]
198. Feller DJ, Zucker J, Don't Walk OB, Srikishan B, Martinez R, Evans H, et al. Towards the inference of social and behavioral determinants of sexual health: development of a gold-standard corpus with semi-supervised learning. AMIA Annu Symp Proc. 2018;2018:422-429 [FREE Full text] [Medline: 30815082]
199. Loda S, Krebs J, Danhof S, Schreder M, Solimando AG, Strifler S, et al. Exploration of artificial intelligence use with ARIES in multiple myeloma research. J Clin Med. 2019 Jul 09;8(7):999 [FREE Full text] [doi: 10.3390/jcm8070999] [Medline: 31324026]
200. Song H, Gu Y, Leroy G, Donovan FM, Galgiani JN. Integrating automated biomedical lexicon creation for valley fever diagnosis. In: Proceedings of the 2021 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies. 2021 Presented at: CHASE '21; December 16-17, 2021; Washington, DC p. 111-112 URL: https://ieeexplore.ieee.org/document/9697921 [doi: 10.1109/chase52844.2021.00021]
201. Patrick J, Li M. High accuracy information extraction of medication information from clinical notes: 2009 i2b2 medication extraction challenge. J Am Med Inform Assoc. 2010 Oct;17(5):524-527 [FREE Full text] [doi: 10.1136/jamia.2010.003939] [Medline: 20819856]
202. Patrick JD, Nguyen DH, Wang Y, Li M. A knowledge discovery and reuse pipeline for information extraction in clinical notes. J Am Med Inform Assoc. 2011;18(5):574-579 [FREE Full text] [doi: 10.1136/amiajnl-2011-000302] [Medline: 21737844]
203. Chen L, Gu Y, Ji X, Lou C, Sun Z, Li H, et al. Clinical trial cohort selection based on multi-level rule-based natural language processing system. J Am Med Inform Assoc. 2019 Nov 01;26(11):1218-1226 [FREE Full text] [doi: 10.1093/jamia/ocz109] [Medline: 31300825]
204. Solt I, Tikk D, Gál V, Kardkovács ZT. Semantic classification of diseases in discharge summaries using a context-aware rule-based classifier. J Am Med Inform Assoc. 2009;16(4):580-584 [FREE Full text] [doi: 10.1197/jamia.M3087] [Medline: 19390101]
205. Uzuner Ö. Recognizing obesity and comorbidities in sparse data. J Am Med Inform Assoc. 2009;16(4):561-570 [FREE Full text] [doi: 10.1197/jamia.M3115] [Medline: 19390096]
206. Yang X, Lyu T, Li Q, Lee C, Bian J, Hogan WR, et al. A study of deep learning methods for de-identification of clinical notes in cross-institute settings. BMC Med Inform Decis Mak. 2019 Dec 05;19(Suppl 5):232 [FREE Full text] [doi: 10.1186/s12911-019-0935-4] [Medline: 31801524]

XSL•FO
RenderX

207. Ferrández O, South BR, Shen S, Friedlin FJ, Samore MH, Meystre SM. BoB, a best-of-breed automated text de-identification system for VHA clinical documents. J Am Med Inform Assoc. 2013 Jan 01;20(1):77-83 [FREE Full text] [doi: 10.1136/amiajnl-2012-001020] [Medline: 22947391]
208. Woo H, Kim K, Cha K, Lee J, Mun H, Cho SJ, et al. Application of efficient data cleaning using text clustering for semistructured medical reports to large-scale stool examination reports: methodology study. J Med Internet Res. 2019 Jan 08;21(1):e10013 [FREE Full text] [doi: 10.2196/10013] [Medline: 30622098]
209. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention is all you need. In: Proceedings of the 31st Conference on Neural Information Processing Systems. 2017 Presented at: NIPS '17; December 4-9, 2017; Long Beach, CA p. 1-11 URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
210. Uzuner Ö, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010;17(5):514-518 [FREE Full text] [doi: 10.1136/jamia.2010.003947] [Medline: 20819854]

Abbreviations

BERT: Bidirectional Encoder Representations from Transformers
BiLSTM-CRF: bidirectional long short-term memory–conditional random field
CDW: clinical data warehouse
CNN: convolutional neural network
CRF: conditional random field
cTAKES: clinical Text Analysis and Knowledge Extraction System
EHR: electronic health record
ELMo: embeddings from language models
GloVe: global vectors for word representation
i2b2: Informatics for Integrating Biology & the Bedside
ICD-9: International Classification of Diseases, Ninth Revision
LSTM: long short-term memory
MedLEE: Medical Language Extraction and Encoding System
n2c2: National NLP Clinical Challenges
NCBO: National Center for Biomedical Ontology
NER: named entity recognition
NLP: natural language processing
PHI: protected health information
PRISMA: Preferred Reporting Items for Systematic Reviews and Meta-Analyses
SNOMED-CT: Systematized Nomenclature of Medicine–Clinical Terms
SVM: support vector machine
TF-IDF: term frequency–inverse document frequency
UMLS: Unified Medical Language System

Edited by C Lovis; submitted 05.09.22; peer-reviewed by M Behzadifar, MF Kabir, B Hoyt; comments to author 17.11.22; revised version received 16.01.23; accepted 07.09.23; published 15.12.23

Please cite as:
Bazoge A, Morin E, Daille B, Gourraud PA
Applying Natural Language Processing to Textual Data From Clinical Data Warehouses: Systematic Review
JMIR Med Inform 2023;11:e42477
URL: https://medinform.jmir.org/2023/1/e42477
doi: 10.2196/42477
PMID:

©Adrien Bazoge, Emmanuel Morin, Béatrice Daille, Pierre-Antoine Gourraud. Originally published in JMIR Medical Informatics (https://medinform.jmir.org), 15.12.2023. This is an open-access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work, first published in JMIR Medical Informatics, is properly cited. The complete bibliographic information, a link to the original publication on https://medinform.jmir.org/, as well as this copyright and license information must be included.

XSL•FO
RenderX