Journal of the American Medical Informatics Association, 26(4), 2019, 364–379
doi: 10.1093/jamia/ocy173
Advance Access Publication Date: 6 February 2019
Review

Review

Natural language processing of symptoms documented in free-text narratives of electronic health records: a systematic review

Theresa A. Koleck,[1] Caitlin Dreisbach,[2,3] Philip E. Bourne,[3] and Suzanne Bakken[1,4,5]

[1]School of Nursing, Columbia University, New York, New York, USA, [2]School of Nursing, University of Virginia, Charlottesville, Virginia, USA, [3]Data Science Institute, University of Virginia, Charlottesville, Virginia, USA, [4]Department of Biomedical Informatics, Columbia University, New York, New York, USA, and [5]Data Science Institute, Columbia University, New York, New York, USA

Corresponding Author: Suzanne Bakken, RN, PhD, FAAN, FACMI, 560 West 168th Street, Mail Code 6, New York, NY 10032, USA (sbh22@cumc.columbia.edu)

Received 29 June 2018; Revised 20 November 2018; Editorial Decision 24 November 2018; Accepted 27 November 2018

ABSTRACT

Objective: Natural language processing (NLP) of symptoms from electronic health records (EHRs) could contribute to the advancement of symptom science. We aim to synthesize the literature on the use of NLP to process or analyze symptom information documented in EHR free-text narratives.

Materials and Methods: Our search of 1964 records from PubMed and EMBASE was narrowed to 27 eligible articles. Data related to the purpose, free-text corpus, patients, symptoms, NLP methodology, evaluation metrics, and quality indicators were extracted for each study.

Results: Symptom-related information was presented as a primary outcome in 14 studies. EHR narratives represented various inpatient and outpatient clinical specialties, with general, cardiology, and mental health occurring most frequently. Studies encompassed a wide variety of symptoms, including shortness of breath, pain, nausea, dizziness, disturbed sleep, constipation, and depressed mood. NLP approaches included previously developed NLP tools, classification methods, and manually curated rule-based processing. Only one-third (n = 9) of studies reported patient demographic characteristics.

Discussion: NLP is used to extract information from EHR free-text narratives written by a variety of healthcare providers on an expansive range of symptoms across diverse clinical specialties. The current focus of this field is on the development of methods to extract symptom information and the use of symptom information for disease classification tasks rather than the examination of symptoms themselves.

Conclusion: Future NLP studies should concentrate on the investigation of symptoms and symptom documentation in EHR free-text narratives. Efforts should be undertaken to examine patient characteristics and make symptom-related NLP algorithms or pipelines and vocabularies openly available.

Key words: natural language processing, signs and symptoms, electronic health records, review

BACKGROUND AND SIGNIFICANCE

Natural language processing (NLP) is currently the most widely used "big data" analytical technique in healthcare,[1] and is defined as "any computer-based algorithm that handles, augments, and transforms natural language so that it can be represented for computation."[2] NLP algorithms are used to perform syntactic processing (eg, tokenization, sentence detection), extract information (ie, convert unstructured text into a structured form), capture meaning (ie, assign a concept to a word or group of words), and detect relationships (ie, assign relationships between concepts) from natural language free text through the use of defined language rules and relevant domain knowledge.[2–4] While both the ambiguity and complexity of medical language makes the application of NLP challenging, NLP has been used for a variety of healthcare-related purposes, including identifying disease risk factors, evaluating efficiency of care and costs, and extracting information from free-text clinical narratives within electronic health records (EHRs).[1]

EHRs are longitudinal collections of electronic information related to the health of or healthcare provided to an individual.[5] EHRs are mainly comprised of 2 types of data, structured data (eg, billing diagnoses, medications, laboratory test results) and unstructured free-text narratives (eg, admission documents, discharge summaries, progress notes, nursing notes, and primary care clinic encounter notes).[6] Much of the rich, expressive clinical data captured in EHRs are documented and stored within these unstructured free-text narratives.[7] This is true for many patient-experienced or reported phenomena, especially symptoms. Consequently, such free-text narratives have been the data source for NLP "challenges" in the health NLP community.[8–12]

Symptoms are subjective indications of disease and include phenomena such as pain, fatigue, disturbed sleep, depressed mood, anxiety, nausea, dyspnea, and pruritus. Symptoms are challenging to manage and burden both the patient and healthcare system,[13] so much so that the National Institute of Nursing Research named "symptom science" as 1 of its key themes with the objective of "[providing] a better understanding of the symptoms of chronic illness and [improving] quality of life across diverse populations." The complexity and multidimensionality of symptoms pose a challenge for research. The volume of longitudinal symptom data available in free-text clinical narratives offers an unprecedented opportunity to study the biological and behavioral foundations of symptom occurrence as well as symptom documentation practices. Development of more effective symptom assessment and management strategies is essential for improving the health-related quality of life of patients.

To illustrate the importance of extracting symptom information from free-text clinical narratives and highlight the diversity of symptom descriptions, Forbush et al[14] manually reviewed and annotated 171 mental or social notes (ie, inpatient and outpatient psychiatry, psychology, social work, and case management) and 579 primary or specialty notes (ie, primary care clinic, specialty clinic, physical and occupational therapy, and inpatient) for symptom terms (eg, depressed mood; memory dysfunction) and subjective symptom expressions (eg, "I'm good for nothing anymore"; "Always forgetting where I put things"). They reported a mean average ($\bar{x}$) of 8.74 (range, 0-67) symptom terms per note for the mental or social notes and $\bar{x}$=6.14 (range, 0-69) for the primary or specialty notes, and $\bar{x}$=1.25 (range, 0-16) symptom expressions per note for the mental or social notes and $\bar{x}$=0.57 (range, 0-35) for the primary or specialty notes.[14] Importantly, they found that if International Classification of Diseases–Ninth Revision–Clinical Modification diagnosis codes were used alone to extract symptom information, only 36% of subjective symptom expressions would be captured.[14]

Symptom information has historically been extracted from patient records via manual review by clinical experts. This approach has clear limitations in scalability in addition to being time consuming, labor intensive, and expensive. The increased availability of EHRs for secondary data reuse has created an opportunity for NLP to be used to harness the potential of free-text narratives to study symptoms and symptom documentation. Systematic reviews related to the automated extraction of information from medical text using NLP and related methods have been published.[15–19] None of these previous reviews focused on symptoms. Due to the (1) prevalence of symptom-related patient and healthcare burden, (2) importance of accurate extraction of symptom information for other applications including disease classification and response to treatment, and (3) potential ability of NLP to facilitate the advancement of symptom science, we sought to review the body of literature and report the state of the science on the use of NLP to process or analyze symptom information from EHR free-text narratives.

OBJECTIVE

The purpose of the present study is to systematically review the literature on the use of NLP to process or analyze symptom information from free-text narratives of EHRs. In particular, we aim to describe and assess the following aspects of studies included in the review: (1) purpose and data source; (2) target clinical population and patient information; (3) symptom extraction and analysis; (4) NLP method, evaluation, and performance; and (5) indicators of quality. We further synthesize and discuss current trends and gaps related to this area and propose recommendations for future studies using NLP to investigate symptoms in the free-text narratives of EHRs.

MATERIALS AND METHODS

Our review procedures were based on the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) recommendations and carried out using Covidence (www.covidence.org), a web-based tool designed to facilitate screening and data extraction related to systematic reviews. The review consisted of 3 stages: (1) article retrieval, (2) study selection, and (3) data extraction and synthesis.

Article retrieval

We searched PubMed and EMBASE on February 5, 2018, to identify all potentially relevant abstracts related to NLP and symptoms. Search terms capturing the concepts of natural language processing and symptoms (Table 1) were derived from the Medical Subject Headings vocabulary (U.S. National Library of Medicine) for the database queries. The use of additional search terms for specific symptoms was guided by inclusion of the symptom in National Institute of Nursing Research common data element measures. Queries were limited to English language, but not by date constraints. Searches returned 811 records from PubMed and 1742 records from EMBASE, of which 589 were duplicates (Figure 1).

Study selection

To be eligible for inclusion in the review, the primary requirement was that the article needed to focus on the description, evaluation, or use of a NLP algorithm or pipeline to process or analyze patient symptom terms. We defined a symptom as a subjective indication of disease. Example symptom terms include *anxiety*, *depressed mood*, *fatigue*, *disturbed sleep*, *impaired cognition*, and *nausea*. Notably, symptoms are distinct from signs (eg, elevated blood pressure, fever, vomiting, rash, cough, hemoptysis, weight loss), which are objective findings that can be directly observed or measured by a healthcare provider. Due to the rigorous focus on symptoms, articles that used NLP to extract more general "problem" terms (which include disorders, procedures, signs, etc.) without specifically naming a symptom(s) were excluded. Review articles as well as articles not published in English or those without full text available were also excluded. While our initial intent was to survey NLP and patient symptoms across all types of free text, a corpus distinction between EHRs and electronic patient authored text (eg, online health

Table 1. Queries used to retrieve records

Database	Search Terms
PubMed	(natural language processing [mh] OR natural language processing [tw] OR NLP [tw] OR text mining [tw]) AND (signs and symptoms [mh] OR symptom [tw] OR nursing [mh] OR nurs* [tw] OR pain [mh] OR pain [tw] OR anxiety [mh] OR anxi* [tw] OR cognition [mh] OR cognit* [tw] OR cognitive function [tw] OR attention [tw] OR memory [tw] OR executive function [tw] OR sleep [mh] OR dyssomnias [mh] OR sleep* [tw] OR fatigue [mh] OR fatigue [tw] OR depression [mh] OR depress* [tw] OR affect [mh] OR affective symptoms [mh] OR affect* [tw] OR mood [tw] OR well being [tw] OR well-being [tw] OR nausea [mh] OR nausea [tw]) AND english [la]
EMBASE	('natural language processing'/exp OR 'natural language processing': ab, ti, kw OR 'nlp': ab, ti, kw OR 'text mining'/exp OR 'text mining': ab, ti, kw) AND ('symptom'/exp OR 'symptomatology'/exp OR 'symptom*': ab, ti, kw OR 'nursing'/exp OR 'nurs*': ab, ti, kw OR 'pain'/exp OR 'pain': ab, ti, kw OR 'anxiety'/exp OR 'anxi*': ab, ti, kw OR 'cognition'/exp OR 'cognit*': ab, ti, kw OR 'cognitive function': ab, ti, kw OR 'sleep'/exp OR 'sleep disorder'/exp OR 'sleep*': ab, ti, kw OR 'fatigue'/exp OR 'fatigue': ab, ti, kw OR 'depression'/exp OR 'depress*': ab, ti, kw OR 'mood disorder'/exp OR 'mood': ab, ti, kw OR 'affect*': ab, ti, kw OR 'wellbeing'/exp OR 'well being': ab, ti, kw OR 'well-being': ab, ti, kw OR 'nausea'/exp OR 'nausea': ab, ti, kw) AND [english]/lim

communities, Twitter) became apparent during the review process; thus, we pulled articles focused on electronic patient authored text for a separate systematic review. EHRs are the focus of the current review.

Two authors (CD, TAK) independently reviewed the title and abstract for each retrieved record. Articles were labeled by potential relevancy as "yes," "no," or "maybe" based on eligibility criteria. Disagreements and articles labeled as "maybe" were discussed to reach a consensus. The same 2 authors (CD, TAK) then independently reviewed the full text of 40 articles identified as potentially relevant during title and abstract screening. Articles were labeled as "include" in or "exclude" from the review. Disagreements were resolved through discussion. Thirteen articles were excluded during the full-text review. Nine of these articles were not symptom focused and 4 did not use NLP or a methodology of interest.

Data extraction and synthesis

Data were manually extracted by 1 of 2 authors (CD, TAK) from the remaining 27 articles included in the systematic review (Table 2).[20–46] A formal quality assessment was not conducted, as relevant reporting standards have not been established for NLP articles. Instead, we developed a data extraction spreadsheet guided by elements reported in previous NLP-focused systematic reviews.[15,18,19] We included information related to the study purpose, corpus (eg, data source, number of narratives, time period), patients (eg, target population, number of distinct patients, demographic information), symptoms (eg, symptoms studied), NLP (eg, methodology or tools used, evaluation measures and performance), and study outcomes (eg, reported symptom-related outcomes).

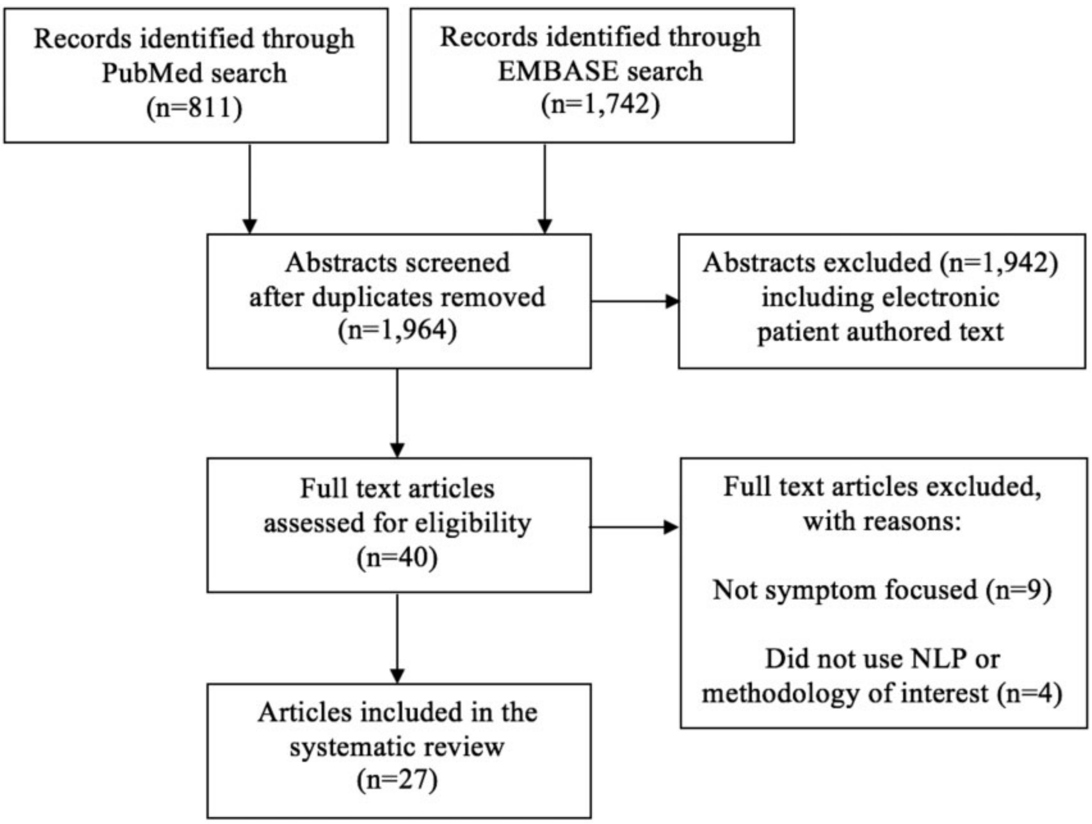

Figure 1. Flow diagram of included articles. NLP: natural language processing.

Table 2. Study purpose and EHR data information

Author	Purpose	Data Type and Source[a]	Number of Documents[b]	Relevant Outcomes	Symptom(s) as Primary Outcome[c]
Byrd et al, 2014[20]	To identify Framingham heart failure signs and symptom criteria	Notes from the EHR, primary care clinic	>3.3 million	System accurately identifies and labels affirmations and denials of Framingham diagnostic criteria in primary care clinical notes	✓
Chase et al, 2017[21]	To determine if patients with multiple sclerosis could be identified from clinical notes before the initial recognition by healthcare providers	Notes from a data repository, encounter notes	Not specified	Classifiers identified 40% of patients with multiple sclerosis before formal documentation by providers; symptom groups used as attributes for multiple sclerosis classification include cognition, dizziness and vertigo, eye and vision, fatigue, headache, mood, pain, motor, and sensory	
Dara et al, 2008[22]	To determine whether preprocessing chief complaints improves performance of syndromic classification	Notes from EHR, chief complaint text	Train: 28 990 Development: 20 293 Test: 10 161	Preprocessing with the chief complaint processor did not improve syndromic classification performance for a probabilistic or keyword-based classifier	
Divita et al, 2017[23]	To describe an NLP technique to identify symptoms from text	Notes from a data repository, encounter notes	948	59 412 symptom mentions were found; Distribution of organ system classes of the symptoms found in the cohort: general (10.03%), musculoskeletal (9.63%), immune (9.44%), respiratory (8.46%), nervous (8.38%), mental health (7.60%), cardiovascular (7.31%), lymphatic (6.74%), genitourinary (6.19%), digestive (5.82%), integumentary (5.63%), endocrine (5.48%), urinary (4.91%), and reproductive (4.38%)	✓
Elkin et al, 2012[24]	To evaluate biosurveillance using data from the encounter note compared with the chief complaint field alone	Notes and clinical data from EHR, chief complaint and encounter notes	Not specified	A biosurveillance model for influenza using the whole encounter note is more accurate than a model that uses only the chief complaint field; model included dyspnea and sore throat	
Friedman et al, 1999[25]	To automate determination of severity classes for patients with CAP	Notes and clinical encoded data from a data repository	Not specified	Feasible to automate determination of risk classes for patients with CAP by using NLP of patient reports; symptoms from discharge summaries were used	
Greenwald et al, 2017[26]	To build a model to identify hospitalized patients' risk for 30-day readmissions	Notes from EHR, admission and discharge documents	Test: 21 876 Train: 7289	Final logistic regression model for 30-day readmission risk included: mood problems (b=0.40±0.06, $P < .01$), suicidal or violent thoughts (b=0.11±0.05, P=0.03), and chronic or uncontrolled pain (b=0.10±0.06, P=0.09)	

(continued)

Table 2. continued

Author	Purpose	Data Type and Source[a]	Number of Documents[b]	Relevant Outcomes	Symptom(s) as Primary Outcome[c]
Gundlapalli et al, 2008[27]	To adapt MedLEE for identifying patients with symptoms suggestive of inflammatory bowel disease	Notes from a data repository, primary and specialty care encounters	76 500	Abdominal pain was identified as a specific symptom suggestive of inflammatory bowel disease and was included for 21% of patients with a reference standard diagnosis	✓
Gundlapalli et al, 2017[28]	To develop an NLP pipeline to extract concepts related to the presence of an indwelling urinary catheter	Notes from a data repository, medical and long- term care inpatient notes	Train: 1050 Test: 545	Performance of the NLP pipeline on extracting positively asserted and negated urinary symptoms was high; out of all the positively asserted symptoms (n = 219 total instances), 11.8% were for dysuria	
Hazlehurst et al, 2009[29]	To identify possible vaccine adverse events of patients who had a recent immunization	Notes from the EHR, ED visits and telephone contacts	13 414	Text classifier was able to identify many gastrointestinal adverse events that were not coded by clinicians in the EHR	
Heintzelman et al, 2013[30]	To test the feasibility of using text mining to depict experience of pain in patients with cancer	Paper records converted into electronic free text, oncology provider encounters	4409	The mean pain mention per record was 1.45; overall, pain increased markedly during the last 2 year of life; severe pain was associated with receipt of opioids (OR, 6.6; $P < .0001$) and palliative radiation (OR, 3.4; $P = .0002$)	✓
Hyun et al, 2009[31]	To explore the ability of NLP for capturing symptoms within nursing documentation	Nursing narratives from the EHR, oncology progress notes	553	The most frequently monitored and recorded symptoms in oncology nursing progress notes were related to chemotherapy care, such as adverse reactions, shortness of breath, nausea, and pain; additional nursing terms and abbreviations must be added to the lexicon to improve performance in the domain of nursing	✓
Iqbal et al, 2017[32]	To create a rule-based framework to identify adverse drug events	Notes from a data repository, clinical encounters and discharge summaries	Rule creation: 2310 Test: 6011	Pipeline achieves better performance in common and long-term adverse drug events than it does with rare and acute adverse drug events	
Jackson et al, 2017[33]	To develop a suite of models to identify key symptoms of severe mental illness	Notes from a data repository, routine mental health encounters	36 624	Symptomatology extracted from discharge summaries of 87% of patients with severe mental illness and 60% of patients with nonsevere mental illness; in the severe mental illness cohort, counts of patients exhibiting the various symptoms followed an approximately Poisson distribution and had prevalence ranging from common to very rare	

(continued)

Table 2. continued

Author	Purpose	Data Type and Source[a]	Number of Documents[b]	Relevant Outcomes	Symptom(s) as Primary Outcome[c]
Ling et al, 2015[34]	To build a system for extracting and clustering symptom/medication names from clinical notes	2009 and 2014 clinical notes datasets from the i2b2 workshop on NLP challenges	2009 data: 1239 2014 data: 1304	Using words, symptom names, and medication names together achieves the best performance for clinical document clustering	✓
Matheny et al, 2012[35]	To develop rule-based NLP algorithms for infectious symptom detection	Notes from EHR, clinical care notes	Train: 60 Test: 444	Among symptoms detected, 1223 (49.9%) had positive, 1215 (49.6%) had negative, and 13 (0.5%) had uncertain assertions; majority of symptoms with excellent performance are those most commonly documented (eg, chest pain or nausea) and those with poorest recall were uncommonly documented (eg, anorexia)	✓
Nunes et al, 2017[36]	To evaluate tolerability and drug effectiveness using EHR data	Notes from a data repository, clinical care notes	Not specified	In both white and African American patients, gastrointestinal symptoms tended to be higher in exenatide once weekly relative to basal insulin	✓
Pakhomov et al, 2007[37]	To test the hypothesis that NLP of the EHR improves chest pain detection over diagnostic codes	Notes from EHR, outpatient and inpatient clinical notes	Not specified	Method improved the detection of unspecified and exertional chest pain cases compared with diagnostic codes and consistently identified more patients with exertional chest pain over a 28-month follow-up	✓
Pakhomov et al, 2008[38]	To determine the agreement between patient-reported symptoms and physician documented symptoms	Notes from EHR, clinical care notes	Not specified	The positive agreement between clinical notes and patient provided forms was 74 for chest pain and 70 for dyspnea, while the negative agreement was 76 and 76; kappa statistics were 0.50 for chest pain and 0.46 for dyspnea	✓
Patel et al, 2015[39]	To assess the impact of mood instability on clinical outcomes of patients receiving secondary mental healthcare	Notes from a data repository, clinical care notes	Not specified	Mood instability was documented in 12.1% of patients presenting to mental healthcare and was associated with a greater number of days spent in the hospital ($b = 18.5$, $P < .001$) and greater frequency of hospitalization (incidence rate ratio, 1.95, $P < .001$)	✓
Tamang et al, 2015[40]	To detect unplanned clinical encounters documented in clinician notes using a clinical text-mining tool	Notes from a data repository, ED	308 096	Pain was the most prevalent symptom and was detected in 75% of ED visits; nausea (54%), anxiety (12%), and emotional distress (12%) were also detected	✓
Tang et al, 2017[41]	To determine whether the Food and Drug Administration's Adverse Event Reporting System data could serve as the basis of automated monitoring for adverse drug events	Notes from EHR, inpatient encounter notes, discharge summaries, ED	1 168 397	2475 adverse drug reaction-related drug-reaction pair sentences were identified	✓

(continued)

Table 2. continued

Author	Purpose	Data Type and Source[a]	Number of Documents[b]	Relevant Outcomes	Symptom(s) as Primary Outcome[c]
Vijayakrishnan et al, 2014[42]	To use NLP to determine the prevalence of the Framingham criteria symptoms	Notes from EHR, clinical care notes in primary care	>3.3 million	41.0% of heart failure cases and 28.1% of controls had paroxysmal nocturnal dyspnea and 87.4% of cases and 59.9% of controls had dyspnea on exertion documented at least once	✓
Wang et al, 2008[43]	To develop an automated approach to discover disease-symptom associations	Notes from a data repository, discharge documents	25 074	563 unique symptom entities and 31 249 unique disease–symptom co-occurring pairs were identified	
Wang et al, 2009[44]	To demonstrate the feasibility of NLP for pharmacovigilance purposes	Notes from a data repository, discharge documents	25 074	132 potential adverse drug events were found to be associated with 7 selected drugs: ibuprofen, morphine, warfarin, bupropion, paroxetine, rosiglitazone, and ACE inhibitors	
Weissman et al, 2016[45]	To characterize the discharge documents of patients diagnosed with acute respiratory distress syndrome	Notes from EHR, discharge documents	815	Symptoms or recommendations related to post–intensive care syndrome were included in 306 (38%) discharge documents; Percentage of reported symptom stem terms: weak/weakness (11.8%), depress* (9.9%), anxiety (5.8%), confus* (5.3%), and cognit* impair* (<0.5%)	
Zhou et al, 2015[46]	To identify patients with depression by applying an NLP system and machine learning classification algorithms	Notes from EHR, discharge documents	Train: 600 Test: 600	Automated approach identified ~20% additional depression cases compared with the structured problem list	

ACE: angiotensin-converting enzyme; CAP: community acquired pneumonia; ED: emergency department; EHR: electronic health record; i2b2: Informatics for Integrating Biology and the Bedside; NLP: natural language processing; OR, odds ratio.

[a]The term *clinical care notes* encompasses a range of notes from the care team including physician, nursing, pathology, social work, radiology, etc. whereas the term *encounter notes* specifies providers who can record clinical visits such as the physician;

[b]Total number of document used unless specified number among training, development, and testing;

[c]A checkmark indicates that the study presented symptom information as a primary outcome.

RESULTS

Twenty-seven articles were included in the review. Years of publication ranged from 1999 to 2017 with more than 90% (n = 25) of articles published in the last 10 years.

Study purpose and data sources

The main objectives of studies included in this review (Table 2) were to capture or detect symptoms (n = 10)[20,23,27,30,31,35,37–39,42]; identify, classify, or characterize disease (n = 8)[21,22,24,25,33,43,45,46]; study adverse drug (n = 5)[32,34,36,41,44] or vaccine (n = 1)[29] events; and identify or detect readmission (n = 1),[26] presence of a device (n = 1),[28] or unplanned clinical encounters (n = 1).[40] Approximately 52% (n = 14) of studies presented symptom-related information as a primary outcome.[20,23,27,30,31,34–42] Symptom-related outcomes relevant to this systematic review are described in Table 2. Free-text narratives were primarily from EHRs (n = 13)[20,22,24,26,29,31,35,37,38,41,42,45,46] and data repositories (n = 12).[21,23,25,27,28,32,33,36,39,40,43,44] Free-text narratives used in the 2 remaining studies were obtained from paper records converted into electronic free text[30] and Informatics for Integrating Biology & the Bedside Challenge datasets.[34] Narratives represented both inpatient (eg, admission documents, discharge summaries, emergency department documents, progress notes, nursing narratives) and outpatient (eg, primary care and specialty clinic documents, mental health encounters) settings and were written by various members of the clinical care team (eg, physicians, nurses). The number of documents parsed as part of each study ranged from 504 to more than 3.3 million. However, approximately 25% (n = 7) of studies did not specify the number of documents processed.[21,24,25,36–39]

Target clinical populations and patient information

Studies focused on 1 or more clinical specialties with general (n = 13),[21–23,25–28,34,35,37,41,43,44] cardiology (n = 5),[20,34,38,42,46] and mental health (n = 4)[32,33,39,46] occurring most frequently (Table 3). The number of distinct patients varied greatly, ranging from 22 to more than 50 000. Notably, the number of distinct patients from which clinical free text was obtained was not reported in approximately 25% (n = 7) of studies,[22,23,29,32,35,43,44] and only one-third (n = 9) of studies reported any patient demographic characteristics.[21,24,30,36–39,42,45] In addition, only 1 study featured a pediatric target population.[41]

Symptom extraction and analysis

All studies mentioned at least 1 specific symptom processed or evaluated using NLP in the study methods, results, or discussion sections. In approximately 37% of studies (n = 10), symptoms were referenced in general terms (eg, all signs and symptoms with concept unique identifiers in the Unified Medical Language System) rather than specifically naming symptoms of interest.[22,23,29,31,34,40,41,43,44,46] In these instances, we manually extracted all symptoms mentioned in the methods, results, or discussion sections of the article. The studies encompassed a wide range of emotional state (eg, mood instability, depressed mood, anxiety), circulatory and respiratory (eg, chest pain, shortness of breath), digestive and abdomen (eg, nausea, constipation, abdominal pain), cognition and perception (eg, cognitive impairment, memory dysfunction, paresthesia, blurred vision, tinnitus), pain (eg, pain, ache, discomfort, headache), fatigue and sleep disturbance (eg, fatigue, disturbed sleep, lethargy), nervous and musculoskeletal (eg, weakness, stiffness, myalgia), general (eg, chills), skin and subcutaneous tissue (eg, pruritus), and urinary (eg, dysuria, bladder discomfort) symptoms. Figure 2 displays the symptoms of interest for each study in this review. Symptoms featured in more than 5 studies included shortness of breath, dyspnea, or orthopnea (n = 13)[20,22,24,25,29,31,35,37,40–44]; pain, ache, or discomfort not specific to the chest or abdomen (n = 11)[21–23,26,30,31,34,35,40,41,44]; nausea (n = 11)[22,29,31,32,34–36,40,41,43,44]; chest pain, pressure, discomfort, or distress or angina (n = 9)[22,31,34,35,37,38,40,43,44]; dizziness or vertigo (n = 9)[21–23,29,31,32,41,43,44]; disturbed sleep, sleeplessness, sleepy, or insomnia (n = 8)[21,23,32,33,41,43,44,46]; abdominal or stomach pain (n = 7)[22,27,31,34,35,40,44]; constipation (n = 7)[21,31,32,34,36,41,44]; and depressed mood (n = 7).[21,23,34,41,43,45,46] With the exception of the study by Heintzelman et al,[30] which incorporated pain severity indicators into the NLP algorithm, documentation occurrence or frequency of occurrence was used to evaluate symptoms.

NLP approach, evaluation, and performance

A variety of different approaches were used to perform NLP and evaluate the NLP algorithms and pipelines (Table 4). Approaches included combinations of previously developed NLP tools, classification methods, and manually curated rule-based processing. Of the previously developed NLP tools, the Medical Language Extraction and Encoding system,[21,25,27,31,43,44] TextHunter,[33,39] Multithreaded Clinical Vocabulary Server,[24,35] and the v3NLP Framework[23,28] were used in more than 1 study. Almost half (n = 13) of studies incorporated manually curated rule-based processing.[23,26,28–30,32,33,35–37,40,45,46] The implementation of NLP was primarily (n = 23) for symptom extraction.[20,21,23–33,35,37–45] NLP algorithms or pipelines were also used for a combination of extraction and pre- or postprocessing[34,36,46] and preprocessing alone.[22]

With the exception of 2 studies that did not evaluate performance of the symptom-related NLP algorithm or pipeline,[36,40] all other studies reported 1 or more evaluation metrics such as sensitivity or recall, specificity, precision, accuracy, F-measure, kappa coefficient, area under the receiver-operating characteristic curve, and C-statistic. Of the 25 studies that reported evaluation metrics, 6 featured true comparative evaluation,[20,22,26,32,34,39] comparing the NLP algorithm or pipeline performance with that of other algorithms either developed as part of the study or previously. The remaining 19 studies compared the results of the NLP algorithm or pipeline with manual chart review or a manually created reference standard (n = 13),[25,27–30,35,37,41–46] cases and control subjects (n = 2),[21,24] clinical practice guidelines (n = 1),[31] International Classification of Diseases–Ninth Revision–Clinical Modification codes (n = 1),[38] "hold out" mentions (n = 1),[23] and with or without a negation algorithm (n = 1).[33] No trends in approach, evaluation, and performance over time were noted.

Indicators of quality across studies

Table 5 summarizes and compares indicators of quality across studies by year of publication. Quality indicators include the clarity of the study purpose statement, inclusion of symptoms as a primary outcome, adequacy of the description of the study approach, and presence of information related to the number of documents, number of patients, patient demographics, evaluation metrics, and comparative evaluation. All studies have at least 4 of the 8 quality indicators. Nine studies have at least 7 quality indicators,[20,27,30,31,34,38,39,41,42] with 1 study addressing all 8.[30] No trends among indicators of quality were identified over time.

Table 3. Clinical focus and patient information

Study	Clinical Specialty	Target Population	Number of Distinct Patients	Demographic Information Reported[a]
Byrd et al, 2014[20]	Cardiology	Primary care patients diagnosed with heart failure	32 407	
Chase et al, 2017[21]	General	Adult patients diagnosed with multiple sclerosis	2999	✓
Dara et al, 2008[22]	General	Patients presenting with a chief complaint	Not reported	
Divita et al, 2017[23]	General	Veterans receiving inpatient or outpatient care	Not reported	
Elkin et al, 2012[24]	Immunology	Patients diagnosed with influenza	2194	✓
Friedman et al, 1999[25]	General	Patients diagnosed with community acquired pneumonia	79	
Greenwald et al, 2017[26]	General	Hospitalized patients readmitted within 30 days of discharge	29 156	
Gundlapalli et al, 2008[27]	General, gastroenterology	Patients diagnosed with inflammatory bowel disease	15 377	
Gundlapalli et al, 2017[28]	General, genitourinary	Hospitalized patients with an indwelling urinary catheter	1222	
Hazlehurst et al, 2009[29]	Immunology, gastroenterology	Patients who had received an immunization	Not reported	
Heintzelman et al, 2012[30]	Oncology	Adult men diagnosed with metastatic prostate cancer	33	✓
Hyun et al, 2009[31]	Oncology	Patients receiving cancer-related inpatient care	22	
Iqbal et al, 2017[32]	Mental health	Patients prescribed antipsychotic or antidepressant medications	Not reported	
Jackson et al, 2017[33]	Mental health	Patients diagnosed with either severe or nonsevere mental illness	15 537	
Ling et al, 2015[34]	General, cardiology	General inpatient and patients diagnosed with coronary artery disease	296[b]	
Matheny et al, 2012[35]	General	General inpatient and outpatient with at least 1 surgical admission	Not reported	
Nunes et al, 2017[36]	Diabetes	Adult injectable-naïve patients diagnosed with type II diabetes mellitus who initiated either exenatide once weekly or basal insulin	5849	✓
Pakhomov et al, 2007[37]	Cardiology	Adult patients with angina pectoris	871	✓
Pakhomov et al, 2008[38]	General	Adult general ambulatory and hospitalized patients	1119	✓
Patel et al, 2015[39]	Mental health	Adult patients diagnosed with a psychotic, affective, or personality disorder	27 704	✓
Tamang et al, 2015[40]	Oncology	Patients with breast, gastrointestinal, or thoracic cancer who seek unplanned care	1263	
Tang et al, 2017[41]	General	Pediatric general inpatient and emergency	42 995	
Vijayakrishnan et al, 2014[42]	Cardiology	Adult primary care patients who have and have not developed heart failure	51 625	✓
Wang et al, 2008[43]	General	General inpatient	Not reported	
Wang et al, 2009[44]	General	General inpatient	Not reported	
Weissman et al, 2016[45]	Pulmonology	Patients diagnosed with acute respiratory distress syndrome	815	✓
Zhou et al, 2015[46]	Mental health, cardiology	Hospitalized patients with a history of ischemic heart disease	1200	

Note:
[a]A checkmark indicates that the study reported demographic information;
[b]Ling et al[34] used clinical note datasets from the i2b2 workshop on NLP challenges from 2009 and 2014. The number of patients is reported for the 2014 dataset only.

DISCUSSION

In this systematic review on the use of NLP to process or analyze symptom information from free-text narratives of patient EHRs, we reviewed and narrowed over 1900 records to a final set of 27 articles. Overall, we found that previously developed NLP tools, classification methods, and manually created rule-based algorithms have been used to primarily extract information on an extensive range of symptoms from EHR free-text narratives written by a variety of healthcare providers across a number of different clinical specialty settings.

One of the most revealing findings from this systematic review was related to the study objectives; only half of the studies presented symptom information as a primary outcome with approximately 30% of studies focusing on the use of symptoms to identify or classify disease. These results highlight how the state of the science on the study of symptoms from EHR free-text narratives is on the development of methods to extract symptom information and the use of symptom information for disease classification tasks rather than on the investigation of symptoms themselves. Considering the pervasiveness of symptom related patient and healthcare burden, there needs to be more investigations focused on symptoms and

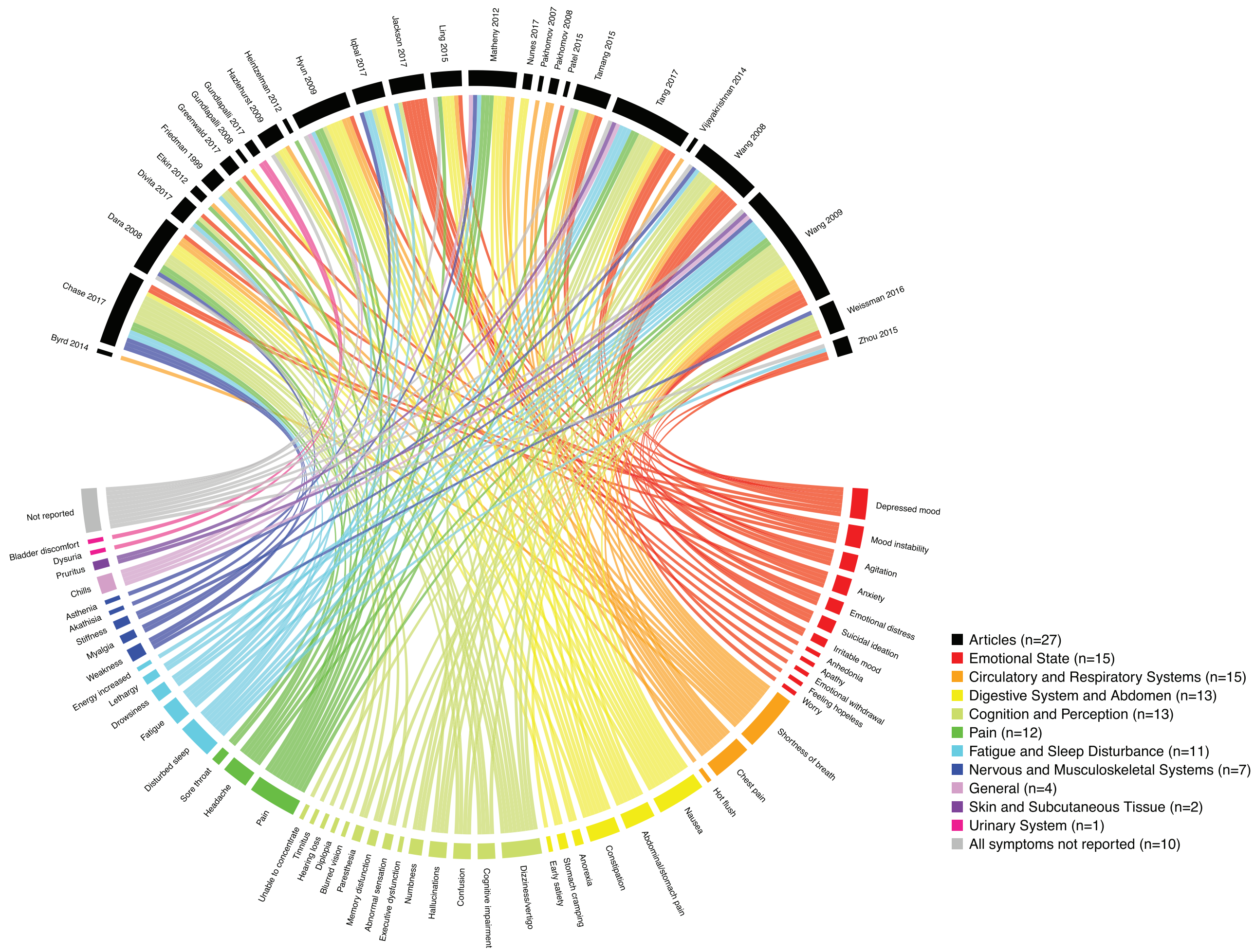

Figure 2. Chord diagram of symptoms by clinical category included in systematic review articles. Relationships between symptoms (color sectors and tracks) and articles (black sectors) included in the systematic review are displayed. Individual symptoms are arranged via color by clinical category. Symptom sector size is proportional to the number of unique articles that include a given symptom. Article sector size is proportional to the number of unique symptoms included in a given study. Sample sizes in the legend correspond to the number of unique articles overall and in each clinical category. Shortness of breath includes dyspnea and orthopnea. Pain includes pain, ache, or discomfort not specified as occurring in the chest or abdomen. The figure was generated using R statistical software (R Foundation for Statistical Computing (R version 3.3.1), Vienna, Austria).[47]

symptom documentation as well as symptom management as primary outcomes of interest from the free-text narratives of EHRs in addition to studies on the use of symptom information to characterize disease or predict response to treatment.

The study of symptoms and symptom documentation from the free-text narratives of EHRs could be facilitated through adherence to the tenets of open science, which aim to increase overall transparency in research and remove barriers for data and resource sharing.[48,49] A strength of a number of studies in this review was the inclusion of detailed information on the selection of symptoms or creation of rules for NLP symptom extraction by clinical experts. For instance, Matheny et al[35] provided the full set of detection rules for each symptom included in their study in appendices. Likewise, Iqbal et al[32] made their expert-developed dictionaries of adverse drug event–related terms available in a GitHub (github.com), which is commonly used to host open-source software projects, repository. However, this was not the case for all studies utilizing expert-developed rules and certainly not the case for the complete NLP pipelines or algorithms. Although open sharing of actual EHR free-text narratives may not be feasible due to the presence of patient protected health information (eg, name, birthdate), researchers can continue to develop and use generalized, open-source EHR-related NLP systems such as Apache cTAKES™ (ctakes.apache.org),[50] the clinical Text Analysis and Knowledge Extraction System, and make expert-developed rule-based NLP algorithms available on platforms such as GitHub to support transparency and replication of study findings and minimize duplicated efforts. Moreover, researchers can advance the symptom content in ontology-based vocabularies such as SNOMED–CT (snomed.org), which was used in multiple studies identified in this systematic review, and contribute to evolving symptom ontologies such as the Open Biological and Biomedical Foundry (obofoundry.org) adopted Symptom Ontology. In addition to symptom-related content, another future direction for NLP of symptom resource development is the normalization of extracted symptom terms to controlled vocabularies. Normalization is important, as many unique symptoms terms (eg, *discomfort*, *hurt*, *ache*, *tender*) are frequently used to represent a single symptom concept (ie, pain).

Table 4. Evaluation and performance metrics

	Approach[b]						
Author[a]	Text Processing	Vocabulary	Classification	Manually Curated Rule-Based Processing[c]	Implementation of NLP[d]	Primary Evaluation Metric	Comparative Evaluation[e]
Friedman et al, 1999[25]	MedLEE				Extraction	Accuracy = 0.93, sensitivity = 0.92, and specificity = 0.93 for processing discharge summaries	Comparison to reference standard
Pakhomov et al, 2007[38]	Text Analysis System (NLP)	SNOMED–CT			Extraction	Sensitivity = 0.62, specificity = 0.63 for any chest pain; sensitivity = 0.71 and specificity = 0.60 for exertional chest pain; and sensitivity = 0.88 and specificity = 0.58 for definitive Rose angina	Compared with ICD-9 codes
Dara et al, 2008[22]	CCP, EMT-P		Naïve Bayes classification, rule-based classification		Preprocessing	Sensitivity = 0.85 for the chief complaint processor preprocessing algorithm	✓
Gundlapalli et al, 2008[27]	MedLEE				Extraction	AUC ROC = 0.90, sensitivity = 0.86, and specificity = 0.95 for identifying concepts of inflammatory bowel disease	Comparison to reference standard
Pakhomov et al, 2008[37]	Unspecified NLP pipeline			✓	Extraction	Sensitivity = 0.91 for chest pain and sensitivity = 0.98 for dyspnea algorithms	Compared with manual extraction
Wang et al, 2008[43]	MedLEE				Extraction	Recall = 0.90 and precision = 0.92 for random sample of disease-symptom associations	Compared with manual extraction
Hazlehurst et al, 2009[29]	MediClass (NLP pipeline)			✓	Extraction	Precision = 0.89, NPV = 0.92, sensitivity = 0.75, and specificity = 0.97 for detection of vaccine reactions versus gold standard manual chart review	Comparison to manual review
Hyun et al, 2009[31]	Perl (text preprocessing), MedLEE				Extraction	18% and 43% of extracted terms matched with pain management and chemotherapy side effects, respectively	Compared with clinical practice guidelines
Wang et al, 2009[44]	MedLEE				Extraction	Recall = 0.75 and 0.31 for known adverse drug events	Compared with manual extraction

(continued)

Table 4. continued

Author[a]	Approach[b] Text Processing	Vocabulary	Classification	Manually Curated Rule-Based Processing[c]	Implementation of NLP[d]	Primary Evaluation Metric	Comparative Evaluation[e]
Elkin et al, 2012[24]	MCVS	SNOMED–CT			Extraction	AUC ROC = 0.929 for entire encounter note versus 0.703 for surveillance with the chief complaint field; kappa = 0.905 between automated method and human review	Case-control comparison and manual review
Matheny et al, 2012[35]	MCVS	SNOMED–CT		✓	Extraction	Precision = 0.91, recall = 0.84, and F-measure = 0.87 for overall symptom detection	Compared with manual review
Heintzelman et al, 2013[30]	ClinREAD (NLP pipeline)		Logistic regression analysis	✓	Extraction	F-measure = 0.95 for pain mention detection	Comparison to reference standard
Byrd et al, 2014[20]	IBM LanguageWare Resource Workbench (text processing), PredMed				Extraction	Precision = 0.925, recall = 0.896, and F-score = 0.910 for Framingham criteria extractions	✓
Vijayakrishnan et al, 2014[42]	Unspecified pipeline				Extraction	Precision = 0.925 and sensitivity = 0.896 for program for Framingham heart failure criteria	Compared with manual review
Ling et al, 2015[34]	Stanford CoreNLP (NLP toolkit), NegEx algorithm (text negation), MetaMap (tool for recognizing Unified Medical Language System concepts in text)		Non-negative matrix factorization		Preprocessing and extraction	Accuracy = 0.60 and normalized mutual information = 0.18 using words, symptom names, and medication names together for clinical document clustering	✓
Patel et al, 2015[39]	TextHunter (NLP tool)		Support vector machine		Extraction	Recall = 0.725, 0.456, and 0.608 and precision = 0.905, 0.911, and 0.980 for mood, affective, and emotional instability, respectively, after applying a probability threshold of precision $\geq$ 0.90	✓
Tamang et al, 2015[40]	ConText algorithm (text processing), unspecified text-mining pipeline			✓	Extraction	No evaluation of symptom text mining algorithm	
Zhou et al, 2015[46]	MTERMS	SNOMED–CT	Weka open-source toolkit	✓	Extraction and processing	F-measure = 0.896, precision = 0.869, recall = 0.924 for MTERMS algorithm	Compared with manual review
Weissman et al, 2016[45]	Unspecified text processing pipeline		Keyword-based document classifier in R	✓	Extraction	Accuracy = 0.95 for document classifier for symptoms of post–intensive care syndrome	Compared with manual review

(continued)

Table 4. continued

Author[a]	Approach[b]: Text Processing	Approach[b]: Vocabulary	Approach[b]: Classification	Approach[b]: Manually Curated Rule-Based Processing[c]	Implementation of NLP[d]	Primary Evaluation Metric	Comparative Evaluation[e]
Chase et al, 2017[21]	MedLEE		Naïve Bayes classification		Extraction	AUC ROC = 0.90, sensitivity = 0.75, and specificity = 0.91 for confirming multiple sclerosis in an enriched cohort	Case-control comparison
Divita et al, 2017[23]	v3NLP Framework (Apache Unstructured Information Management application framework for NLP)		Automated machine learning in Weka	✓	Extraction	Precision = 0.80, recall = 0.74 and F-score = 0.80 for symptom mentions	Held-out testing set
Greenwald et al, 2017[26]	Unspecified NLP pipeline			✓	Extraction	Validated C-statistic = 0.74 for final 30-day readmission risk model	✓
Gundlapalli et al, 2017[27]	v3NLP Framework (Apache Unstructured Information Management application framework for NLP)			✓	Extraction	Recall and precision >0.90 for extracting urinary symptoms	Comparison to reference standard
Iqbal et al, 2017[32]	GATE framework, ADEPt			✓	Extraction	Average F-measure = 0.83 and accuracy = 0.83 for the tool across all tested adverse drug events	✓
Jackson et al, 2017[33]	TextHunter (NLP tool), ConText algorithm (text processing)		Support vector machine	✓	Extraction	Median F1 score = 0.88, precision = 0.90, and recall = 0.85 for across all symptoms for the ConText plus machine learning model	Compared with and without ConText
Nunes et al, 2017[36]	Unspecified NLP pipeline			✓	Extraction and syntax processing	No evaluation of NLP algorithm	
Tang et al, 2017[41]	cTAKES, NegEx algorithm (text negation)				Extraction	Precision = 0.800, TP = 4 for ED notes; precision = 0.458, TP = 165 for progress notes; precision = 0.381, TP = 40 for discharge summaries; precision = 0.259, TP = 15 for H&P notes	Compared with manual annotation

ADEPt: Adverse Drug Event annotation Pipeline (preprocessing, NLP); AUC ROC: area under the receiver-operating characteristic curve; CCP: chief complaint processor (preprocessing); cTAKES: Clinical Text Analysis and Knowledge Extraction System (NLP); ED: emergency department; EMT-P: emergency medical text processor (preprocessing); GATE: General Architecture for Text Engineering (Java framework for NLP); H&P: history and physical; ICD-9: International Classification of Diseases–Ninth Revision; MCVS: Multithreaded Clinical Vocabulary Server (preprocessing and NLP); MedLEE: Medical Language Extraction and Encoding system (NLP pipeline); MTERMS: Medical Text Extraction, Reasoning and Mapping System (NLP pipeline); NLP: natural language processing; NPV: negative predictive value; PredMed: Predictive Modeling for Early Detection (NLP pipeline); SNOMED–CT: Systematized Nomenclature of Medicine–Clinical Terms (reference terminology); TP: true positive.

[a]Studies included in this table have been arranged in chronological order to assess trends of approach and analytic methods over time;

[b]Approach as outlined in the manuscript, includes text processing pipelines and tools, terminology vocabulary, classification method, and inclusion of manually curated rule-based processing;

[c]A checkmark indicates that the study used a rule-based methodology;

[d]Specific primary usage of NLP in the study;

[e]A checkmark indicates that the study compared their performance to another existing algorithm, otherwise text is added in this column about an available performance comparison group.

Table 5. Indicators of quality across articles

Author[a]	Clearly defined purpose[b]	Symptoms as primary outcome[c]	Approach adequately described[d]	Number of documents specified[e]	Number of patients specified[e]	Patient demographic information reported[e]	Evaluation metrics reported[e,f]	Inclusion of comparative evaluation[e,g]
Friedman et al, 1999[25]	✓		✓		✓		✓	✓
Pakhomov et al, 2007[38]	✓	✓	✓		✓	✓	✓	✓
Dara et al, 2008[22]	✓		✓	✓			✓	✓
Gundlapalli et al, 2008[27]	✓	✓	✓	✓	✓		✓	✓
Pakhomov et al, 2008[37]	✓	✓			✓	✓	✓	✓
Wang et al, 2008[43]	✓		✓	✓			✓	✓
Hazlehurst et al, 2009[29]	✓		✓	✓			✓	✓
Hyun et al, 2009[31]	✓	✓	✓	✓	✓		✓	✓
Wang et al, 2009[44]	✓		✓	✓			✓	✓
Elkin et al, 2012[24]	✓		✓		✓	✓	✓	✓
Matheny et al, 2012[35]	✓	✓	✓	✓			✓	✓
Heintzelman et al, 2013[30]	✓	✓	✓	✓	✓	✓	✓	✓
Byrd et al, 2014[20]	✓	✓	✓	✓	✓		✓	✓
Vijayakrishnan et al, 2014[42]	✓	✓		✓	✓	✓	✓	✓
Ling et al, 2015[34]	✓	✓	✓	✓	✓		✓	✓
Patel et al, 2015[39]	✓	✓	✓		✓	✓	✓	✓
Tamang et al, 2015[40]	✓	✓	✓	✓	✓			
Zhou et al, 2015[46]	✓		✓	✓	✓		✓	✓
Weissman et al, 2016[45]	✓			✓	✓	✓	✓	✓
Chase et al, 2017[21]	✓		✓		✓	✓	✓	✓
Divita et al, 2017[23]	✓	✓	✓	✓			✓	✓
Greenwald et al, 2017[26]	✓			✓	✓		✓	✓
Gundlapalli et al, 2017[28]	✓		✓	✓	✓		✓	✓
Iqbal et al, 2017[32]	✓		✓	✓			✓	✓
Jackson et al, 2017[33]	✓		✓	✓	✓		✓	✓
Nunes et al, 2017[36]	✓	✓			✓	✓		
Tang et al, 2017[41]	✓	✓	✓	✓	✓		✓	✓

[a]Studies included in this table have been arranged in chronological order to assess trend of quality indicators over time;
[b]A checkmark denotes reviewer judgement of clear statement of the study purpose;
[c]A checkmark denotes inclusion of symptoms as a primary outcome;
[d]A checkmark denotes reviewer judgement of adequate description of the study approach;
[e]A checkmark denotes the presence of information in the article;
[f]Evaluation metrics include accuracy, area under the curve, sensitivity, specificity, recall, or precision;
[g]Comparison includes another algorithm, held-out testing set, manual review or annotation, or a case-control design.

While our finding that almost half of the studies focused on general inpatient or outpatient populations was in line with expectations, we were surprised that only about 11% (n = 3) of studies featured oncology as the clinical specialty of interest. This lack of cancer- or cancer treatment–related symptoms being processed or analyzed using NLP from EHR free-text narratives is in contrast to what one would anticipate based on both the cancer symptom and cancer NLP literature. Providing evidence for the focus on oncology in the field of symptom science, Miaskowski et al[51] reported that approximately 83% of n = 158 articles surveyed for a review of co-occurring symptoms in chronic conditions studied patients with cancer. Moreover, a PubMed search of the MeSH (Medical Subject Headings) terms *signs and symptoms* and *neoplasm* returns almost 7000 articles from the past 10 years highlighting the clinical importance and, we would argue, the complexity of symptoms related to detection, diagnosis, treatment, and management of cancer or cancer treatment. Likewise, a recent review by Jiang et al[52] relayed that the major disease concentration area for artificial intelligence (including NLP as well as other computational techniques such as support vector machines and neural networks) in healthcare was cancer followed by neurology and cardiology. A clear opportunity exists to combine these fields and use NLP to study symptoms related to cancer or its treatments in the EHR.

Remarkably, <75% of articles reported the distinct number of patients from which clinical free-text was obtained and only 33% of articles reported any patient demographic characteristics. These findings appear to be related to the objective of the study, specifically, whether the purpose of study was to develop an algorithm for symptom identification versus to describe symptom related information for a defined clinical population. For example, the purposes of the articles by Iqbal et al[32] and Matheny et al,[35] which do not report the number of distinct patients or patient demographic information, were to develop rule-based algorithms for the identification of adverse drug events and infectious symptoms, respectively. In contrast, the articles by Patel et al[39] and Vijayakrishnan et al[42] aimed to study the impact of symptoms on clinical outcomes and prevalence of symptoms, respectively, in specific clinical populations; both of these articles report the distinct number of patients and patient demographic information, including age, gender, and race. The inclusion of information about the patients from whom clinical free-text was obtained is important because symptom experience is known to vary by common sociodemographic factors including age, sex or gender, race and ethnicity, and socioeconomic status.[53] It is essential for future NLP studies of symptoms documented in EHRs to analyze and report patient information for generalization of study findings, ascertainment of potential assessment or documentation biases, and development of tailored interventions.

While the studies in our review included a wide variety of symptoms, shortness of breath, dyspnea, or orthopnea; pain, ache, or discomfort not specific to the chest or abdomen; nausea; and chest pain, pressure, discomfort, or distress or angina were the most common symptoms mentioned in the methods, results, or discussion sections of included studies. These symptoms are consistent with the 10 leading principal reasons for emergency department visits, which include chest pain and related symptoms; shortness of breath; pain, site not referable to a specific body system; and vomiting (ie, the sign that typically accompanies nausea).[54]

However, we would like to point out that many studies investigated symptoms and signs concurrently, either not making the distinction between the 2 concepts or inaccurately classifying signs as symptoms. As mentioned earlier in this review, symptoms are subjective while signs are objective evidence of disease. The imprecision is not unexpected because symptoms (eg, pruritus or itchy skin) and signs (eg, rash) frequently occur simultaneously with signs often being termed "physical" symptoms. But this observation further highlights the focus of using symptom information from EHR free-text narratives to characterize or classify disease rather than study the symptoms themselves. Additionally, by and large, studies used documentation occurrence or frequency of occurrence to investigate symptoms. Though many studies included negation algorithms (eg, no shortness of breath) as part of NLP processing, only 1 study explicitly evaluated symptom severity.[30] Heintzelman et al[30] developed pain severity contextual rules to further categorize mentions of pain as no pain, some pain, controlled pain, and severe pain. Incorporation of accurate extraction of severity as well as other contextual factors such as symptom location or duration into EHR NLP algorithms is of great interest for future work.

Finally, we found it challenging to assess the quality of the studies within this systematic review as relevant formal standards have yet to be established for NLP articles. Instead, we focused on indicators of quality of the included articles. A number of the recurrent strengths and weaknesses of articles have already been discussed throughout this section. Additional strengths include the incorporation of concept modifiers into NLP algorithms or pipelines, control for covariates and confounders in analyses, and evaluation of NLP algorithm or pipeline performance. Additional weaknesses include small samples of patients or narratives, no incorporation of temporality, and lack of true comparative evaluation of the NLP algorithm or pipeline used in the study to other methods.

CONCLUSION

In this systematic review, we synthesized data from 27 articles on the use of NLP to process or analyze symptom information from free-text narratives of patient EHRs. In summary, we found that NLP tools, classification methods, and manually curated rule-based processing are being used to extract information from EHR free-text narratives written by a variety of healthcare providers on a wide range of symptoms across diverse clinical specialties. The current focus of this field is on the development of methods to extract symptom information and the use of symptom information for disease classification tasks rather than on the investigation of symptoms themselves. Considering the prevalence of symptom-related patient and healthcare burden, future work should concentrate on the study of specific symptoms and symptom documentation in free-text narratives of patient EHRs in addition to the use of symptoms to accomplish other tasks. The study of symptoms and symptom documentation from EHRs using NLP would greatly benefit from clear statement of the symptoms being evaluated as part of the study, a detailed description of the clinical population from which symptom information was extracted and analyzed, open sharing of user-developed symptom-related NLP algorithms or pipelines and vocabularies, and the establishment of formal reporting standards for investigations using NLP methodologies.

FUNDING

This work is supported by the Reducing Health Disparities Through Informatics training grant (T32NR007969 to SB and TAK), the Precision in Symptom Self-Management (PriSSM) Center (P30NR016587 to SB), Advancing Chronic Condition Symptom Cluster Science Through Use of Electronic Health Records and Data Science Techniques K99NR017651 to TAK), and Influence of Maternal Obesity on Microbial Function and Impaired Glucose Tolerance During Pregnancy (F31NR017821 (to CD).

CONTRIBUTORS

All authors contributed significantly to this work. TAK, CD, and SB conceptualized the study. TAK and CD searched for and retrieved relevant articles and analyzed data. TAK, CD, and SB interpreted the data. TAK drafted the manuscript, and CD, PEB, and SB made substantive revisions to the manuscript. All authors gave final approval of and accept accountability for the manuscript.

Conflict of interest statement. None declared.

REFERENCES

1. Mehta N, Pandit A. Concurrence of big data analytics and healthcare: a systematic review. *Int J Med Inform* 2018; 114: 57–65.
2. Yim W-W, Yetisgen M, Harris WP, *et al*. Natural language processing in oncology. *JAMA Oncol* 2016; 2 (6): 797–804.
3. Fleuren WWM, Alkema W. Application of text mining in the biomedical domain. *Methods* 2015; 74: 97–106.
4. Wang Y, Wang L, Rastegar-Mojarad M, *et al*. Clinical information extraction applications: a literature review. *J Biomed Inform* 2018; 77: 34–49.
5. Institute of Medicine (US) Committee on Data Standards for Patient Safety. *Key Capabilities of an Electronic Health Record System: Letter Report*. Washington, DC: National Academies Press. 2003.
6. Chen ES, Sarkar IN. Mining the electronic health record for disease knowledge. *Methods Mol Biol* 2014; 1159: 269–86.
7. Ross MK, Wei W, Ohno-Machado L. "Big data" and the electronic health record. *Yearb Med Inform* 2014; 9: 97–104.
8. Uzuner O, South BR, Shen S, *et al*. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *J Am Med Inform Assoc* 2011; 18 (5): 552–6.
9. Uzuner O, Stubbs A, Filannino M. A natural language processing challenge for clinical records: Research Domains Criteria (RDoC) for psychiatry. *J Biomed Inform* 2017; 75: S1–3.
10. Uzuner O. Recognizing obesity and comorbidities in sparse data. *J Am Med Inform Assoc* 2009; 16 (4): 561–70.
11. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. *J Am Med Inform Assoc* 2013; 20 (5): 806–13.
12. Stubbs A, Kotfila C, Xu H, *et al*. Identifying risk factors for heart disease over time: overview of 2014 i2b2/UTHealth shared task track 2. *J Biomed Inform* 2015; 58: S67–77.

13. Kwekkeboom KL. Cancer symptom cluster management. *Semin Oncol Nurs* 2016; 32 (4): 373–82.
14. Forbush TB, Gundlapalli AV, Palmer MN, *et al*. Sitting on pins and needles. Characterization of symptom descriptions in clinical notes. *AMIA Jt Summits Transl Sci Proc* 2013; 2013: 67–71.
15. Canan C, Polinski JM, Alexander GC, *et al*. Automatable algorithms to identify nonmedical opioid use using electronic data: a systematic review. *J Am Med Inform Assoc* 2017; 24 (6): 1204–10.
16. Ford E, Carroll JA, Smith HE, *et al*. Extracting information from the text of electronic medical records to improve case detection: a systematic review. *J Am Med Inform Assoc* 2016; 23 (5): 1007–15.
17. Kreimeyer K, Foster M, Pandey A, *et al*. Natural language processing systems for capturing and standardizing unstructured clinical information: a systematic review. *J Biomed Inform* 2017; 73: 14–29.
18. Pons E, Braun LMM, Hunink MGM, *et al*. Natural language processing in radiology: a systematic review. *Radiology* 2016; 279 (2): 329–43.
19. Mishra R, Bian J, Fiszman M, *et al*. Text summarization in the biomedical domain: a systematic review of recent research. *J Biomed Inform* 2014; 52: 457–67.
20. Byrd RJ, Steinhubl SR, Sun J, *et al*. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. *Int J Med Inform* 2014; 83 (12): 983–92.
21. Chase HS, Mitrani LR, Lu GG, *et al*. Early recognition of multiple sclerosis using natural language processing of the electronic health record. *BMC Med Inform Decis Mak* 2017; 17: 24.
22. Dara J, Dowling JN, Travers D, *et al*. Evaluation of preprocessing techniques for chief complaint classification. *J Biomed Inform* 2008; 41 (4): 613–23.
23. Divita G, Luo G, Tran L-TT, *et al*. General symptom extraction from VA electronic medical notes. *Stud Health Technol Inform* 2017; 245: 356–60.
24. Elkin PL, Froehling DA, Wahner-Roedler DL, *et al*. Comparison of natural language processing biosurveillance methods for identifying influenza from encounter notes. *Ann Intern Med* 2012; 156 (1_Part_1): 11–8.
25. Friedman C, Knirsch C, Shagina L, *et al*. Automating a severity score guideline for community-acquired pneumonia employing medical language processing of discharge summaries. *Proc AMIA Symp* 1999; 256–60.
26. Greenwald JL, Cronin PR, Carballo V, *et al*. A novel model for predicting rehospitalization risk incorporating physical function, cognitive status, and psychosocial support using natural language processing. *Med Care* 2017; 55 (3): 261–6.
27. Gundlapalli AV, South BR, Phansalkar S, *et al*. Application of natural language processing to VA electronic health records to identify phenotypic characteristics for clinical and research purposes. *Summit Transl Bioinform* 2008; 2008: 36–40.
28. Gundlapalli AV, Divita G, Redd A, *et al*. Detecting the presence of an indwelling urinary catheter and urinary symptoms in hospitalized patients using natural language processing. *J Biomed Inform* 2017; 71S: S39–45.
29. Hazlehurst B, Naleway A, Mullooly J. Detecting possible vaccine adverse events in clinical notes of the electronic medical record. *Vaccine* 2009; 27 (14): 2077–83.
30. Heintzelman NH, Taylor RJ, Simonsen L, *et al*. Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. *J Am Med Inform Assoc* 2013; 20 (5): 898–905.
31. Hyun S, Johnson SB, Bakken S. Exploring the ability of natural language processing to extract data from nursing narratives. *Comput Inform Nurs* 2009; 27: 215–23, quiz 224–5.
32. Iqbal E, Mallah R, Rhodes D, *et al*. ADEPt, a semantically-enriched pipeline for extracting adverse drug events from free-text electronic health records. *PLoS One* 2017; 12 (11): e0187121.
33. Jackson RG, Patel R, Jayatilleke N, *et al*. Natural language processing to extract symptoms of severe mental illness from clinical text: the Clinical Record Interactive Search Comprehensive Data Extraction (CRIS-CODE) project. *BMJ Open* 2017; 7 (1): e012012.
34. Ling Y, Pan X, Li G, *et al*. Clinical documents clustering based on medication/symptom names using multi-view nonnegative matrix factorization. *IEEE Trans Nanobioscience* 2015; 14 (5): 500–4.
35. Matheny ME, Fitzhenry F, Speroff T, *et al*. Detection of infectious symptoms from VA emergency department and primary care clinical documentation. *Int J Med Inform* 2012; 81 (3): 143–56.
36. Nunes AP, Loughlin AM, Qiao Q, *et al*. Tolerability and effectiveness of exenatide once weekly relative to basal insulin among type 2 diabetes patients of different races in routine care. *Diabetes Ther* 2017; 8 (6): 1349–64.
37. Pakhomov SV, Jacobsen SJ, Chute CG, *et al*. Agreement between patient-reported symptoms and their documentation in the medical record. *Am J Manag Care* 2008; 14: 530–9.
38. Pakhomov SSV, Hemingway H, Weston SA, *et al*. Epidemiology of angina pectoris: role of natural language processing of the medical record. *Am Heart J* 2007; 153 (4): 666–73.
39. Patel R, Lloyd T, Jackson R, *et al*. Mood instability is a common feature of mental health disorders and is associated with poor clinical outcomes. *BMJ Open* 2015; 5 (5): e007504.
40. Tamang S, Patel MI, Blayney DW, *et al*. Detecting unplanned care from clinician notes in electronic health records. *J Oncol Pract* 2015; 11 (3): e313–9.
41. Tang H, Solti I, Kirkendall E, *et al*. Leveraging Food and Drug Administration Adverse Event Reports for the automated monitoring of electronic health records in a pediatric hospital. *Biomed Inform Insights* 2017; 9: 1178222617713018.
42. Vijayakrishnan R, Steinhubl SR, Ng K, *et al*. Prevalence of heart failure signs and symptoms in a large primary care population identified through the use of text and data mining of the electronic health record. *J Card Fail* 2014; 20 (7): 459–64.
43. Wang X, Chused A, Elhadad N, *et al*. Automated knowledge acquisition from clinical narrative reports. *AMIA Annu Symp Proc* 2008; 2008: 783–7.
44. Wang X, Hripcsak G, Markatou M, *et al*. Active computerized pharmacovigilance using natural language processing, statistics, and electronic health records: a feasibility study. *J Am Med Inform Assoc* 2009; 16 (3): 328–37.
45. Weissman GE, Harhay MO, Lugo RM, *et al*. Natural kanguage processing to assess documentation of features of critical illness in discharge documents of acute respiratory distress syndrome survivors. *Ann Am Thorac Soc* 2016; 13 (9): 1538–45.
46. Zhou L, Baughman AW, Lei VJ, *et al*. Identifying patients with depression using free-text clinical documents. *Stud Health Technol Inform* 2015; 216: 629–33.
47. Gu Z, Gu L, Eils R, *et al*. circlize Implements and enhances circular visualization in R. *Bioinformatics* 2014; 30 (19): 2811–2.
48. Watson M. When will 'open science' become simply 'science'? *Genome Biol* 2015; 16: 101.
49. McKiernan EC, Bourne PE, Brown CT, *et al*. How open science helps researchers succeed. *Elife* 2016; 5: 372.
50. Savova GK, Masanz JJ, Ogren PV, *et al*. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *J Am Med Inform Assoc* 2010; 17 (5): 507–13.
51. Miaskowski C, Barsevick A, Berger A, *et al*. Advancing symptom science through symptom cluster research: expert panel proceedings and recommendations. *J Natl Cancer Inst* 2017; 109 (4): djw253.
52. Jiang F, Jiang Y, Zhi H, *et al*. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol* 2017; 2 (4): 230–43.
53. Corwin EJ, Berg JA, Armstrong TS, *et al*. Envisioning the future in symptom science. *Nurs Outlook* 2014; 62 (5): 346–51.
54. Rui P, Kang K. National Hospital Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables. http://www.cdc.gov/nchs/data/ahcd/nhamcs_emergency/2015_ed_web_tables.pdf. Accessed June 6, 2018.