Migration Letters

Volume: 19, No: S2 (2022), pp. 1298-1303

ISSN: 1741-8984 (Print) ISSN: 1741-8992 (Online)

www.migrationletters.com

The Use Of Natural Language Processing In Extracting Information From Medical Records

Mona Ali Ajeebi , Sultan Khudayr Muhamad Aleanzi , Khalid Ayidh Oudah , Budur Ali Alshehri , Sarah Ahmed Arif, Muath Mohammed Moraya , Abdullah Jaber Faqihi , Dhaifallah Mohammed Moraya , Nasser Mubarak Khalaf , Adel Mohammed Hakami , Dalal Funtul Alanazi , Majed Mislat Eid Albaqami

Abstract

Utilizing natural language processing (NLP) to analyze symptoms extracted from electronic health records (EHRs) has the potential to enhance the progress of symptom research. Our objective is to consolidate the existing research on the use of Natural Language Processing (NLP) in the processing and analysis of symptom information found in free-text narratives inside Electronic Health Records (EHR). For each research, information pertaining to the objective, free-text collection, patients, symptoms, NLP approach, assessment metrics, and quality indicators were retrieved. Fourteen studies reported symptom-related information as their main endpoint. The EHR narratives included a range of therapeutic disciplines, including general, cardiology, and mental health, with the latter being the most common. The studies examined a broad range of symptoms, such as dyspnea, pain, emesis, vertigo, disrupted sleep, constipation, and low mood. Natural Language Processing (NLP) methodologies included pre-existing NLP tools, categorization techniques, and meticulously maintained rule-based processing. Natural Language Processing (NLP) is used to retrieve data from Electronic Health Record (EHR) unstructured narratives authored by a multitude of healthcare practitioners, including a wide spectrum of symptoms across many clinical domains. The present emphasis in this research is on developing techniques to collect symptom data and use that data for illness categorization purposes, rather than studying the symptoms themselves. Future research in Natural Language Processing (NLP) should focus on examining symptoms and the documenting of symptoms in electronic health record (EHR) narratives that are written in free-text format. It is important to make an effort to analyze the features of patients and create publically accessible NLP algorithms or pipelines and vocabularies that are specifically designed for symptoms.

Keywords: *natural language processing, indications and manifestations, electronic health records, examination*

1. Introduction

Natural language processing (NLP) is now the predominant "big data" analytical method in healthcare.1 It refers to any computer-based program that manages, enhances, and converts natural language into a format suitable for computing.2 Natural Language Processing (NLP) algorithms are utilized to carry out syntactic processing tasks such as tokenization and sentence detection. They also extract information by converting unstructured text into a structured format. Furthermore, these algorithms assign concepts to words or groups of words to capture their meaning. Additionally, they identify relationships between concepts by utilizing defined language rules and domain knowledge. 2-4 Although the ambiguous and complex nature of medical language presents challenges for the implementation of NLP, NLP has been successfully employed in various healthcare contexts. These include identifying risk factors for diseases, assessing the effectiveness and costs of care, and extracting information from unstructured clinical narratives in electronic health records (EHRs).1

2. Electronic Health Records (EHRs)

Electronic Health Records (EHRs) are comprehensive and continuous repositories of digital data pertaining to an individual's medical condition and the medical services they have received.5 Electronic Health Records (EHRs) include primarily of two categories of data: structured data, which includes information such as billing diagnoses, prescriptions, and laboratory test results; and unstructured free-text narratives, which encompass documents like admission paperwork, discharge summaries, progress notes, nursing notes, and primary care clinic contact notes.6 A significant portion of the comprehensive and detailed clinical information recorded in electronic health records (EHRs) is written and kept as unstructured free-text narratives.7 This is true for several things that patients have encountered or reported, including symptoms. As a result, these unrestricted written accounts have served as the primary data for NLP "challenges" among the health NLP community.8-12

Symptoms are subjective manifestations of sickness and include phenomena such as pain, exhaustion, disrupted sleep, low mood, worry, nausea, difficulty breathing, and itching. The management of symptoms poses a significant challenge and places a burden on both the patient and the healthcare system. This issue is of such importance that the National Institute of Nursing Research has identified "symptom science" as one of its key themes. The objective is to gain a deeper understanding of the symptoms associated with chronic illness and to enhance the quality of life for diverse populations. The intricate and multifaceted nature of symptoms is a formidable obstacle for investigation. The abundance of longitudinal symptom data included in unstructured clinical narratives provides a unique chance to investigate the physiological and psychological basis of symptom occurrence, as well as the methods used to capture symptoms. Enhancing the health-related quality of life of patients requires the imperative development of more efficient ways for assessing and managing symptoms. 13

In order to emphasize the significance of extracting symptom information from clinical narratives and showcase the wide range of symptom descriptions, Forbush et al 14 conducted a manual review and annotation of 171 mental or social notes (such as inpatient and outpatient psychiatry, psychology, social work, and case management) and 579 primary or specialty notes (including primary care clinic, specialty clinic, physical and occupational therapy, and inpatient) for symptom terms (such as depressed mood; memory dysfunction) and subjective symptom expressions (such as "I'm good for nothing anymore"; "Always forgetting where I put things").

In the past, symptom information has been obtained from patient records by having clinical professionals manually evaluate them. This strategy has evident constraints in terms of scalability, as well as being characterized by a significant investment of time, labor, and financial resources. The expanded accessibility of electronic health records (EHRs) for the purpose of reusing secondary data has presented a chance for natural language processing (NLP) to be used in order to fully utilize the potential of unstructured text narratives for the examination of symptoms and the reporting of such symptoms. Published systematic studies have explored the automated extraction of information from medical text using NLP and similar technologies.15-19

None of the prior reviews addressed symptoms specifically. Given the widespread occurrence of symptoms and the resulting burden on patients and healthcare providers, the accurate extraction of symptom information is crucial for various purposes such as disease classification and treatment response. Additionally, natural language processing (NLP) has the potential to advance the field of symptom science. Therefore, our objective was to review existing literature and present the current state of using NLP to process and analyze symptom information from electronic health record (EHR) free-text narratives.

The objective of this work is to conduct a comprehensive assessment of existing literature about the use of Natural Language Processing (NLP) for the purpose of processing and analyzing symptom information extracted from unstructured text narratives found in Electronic Health Records (EHRs). Specifically, our objective is to provide a detailed analysis and evaluation of the following components of the research included in the review: The text covers five main areas: (1) the objective and data origin; (2) the specific clinical population and patient data; (3) the process of extracting and analyzing symptoms; (4) the natural language processing (NLP) technique, assessment, and performance; and (5) the criteria used to assess quality. In addition, we analyze and examine the present patterns and deficiencies associated with this field and provide suggestions for future research using NLP to examine symptoms in the unstructured narratives of EHRs.

3. Data Analysis

An important discovery from this comprehensive analysis was that only 50% of the research included symptom information as their main emphasis, whereas only 30% of the studies concentrated on utilizing symptoms to diagnose or categorize diseases. The findings emphasize that the current focus of research in the field of studying symptoms from EHR free-text narratives is mostly on developing techniques to extract symptom information and use this information for illness categorization purposes, rather than directly investigating the symptoms themselves. Given the widespread impact of symptoms on patients and healthcare systems, it is necessary to conduct further research specifically targeting symptoms and their documentation, as well as the management of symptoms as the main outcomes of interest. This research should include analyzing the free-text narratives in electronic health records (EHRs) to understand how symptoms are described and to explore their potential use in characterizing diseases or predicting treatment response.

Examining symptoms and recording them from the unstructured text narratives in electronic health records (EHRs) may be made easier by following the principles of open science. Open science seeks to enhance openness in research and eliminate obstacles to sharing data and resources. 20,21

One notable aspect of much research in this analysis was the incorporation of comprehensive data on the process of symptom selection or the development of rules for NLP symptom extraction by clinical specialists. Matheny et al. (22) published the whole set of detection criteria for each symptom in their investigation as appendices. Similarly, Iqbal et al 23 published their expert-created dictionaries of terminology linked to adverse medication events on GitHub, a popular platform for hosting open-source software projects. However, not all research that used expert-developed rules had the same outcome, and this was especially true for whole NLP pipelines or algorithms. While it may not be possible to openly share actual EHR free-text narratives due to the inclusion of protected patient health information, researchers can still create and utilize generalized, open-source EHR-related NLP systems like Apache cTAKES.

These systems, such as the clinical Text Analysis and Knowledge Extraction System, can provide expert-developed rule-based NLP algorithms that can be made accessible on platforms like GitHub. This promotes transparency and replication of study findings while reducing redundant work. In addition, researchers have the ability to enhance the symptom content in ontology-based vocabularies like SNOMED-CT (snomed.org), which has been utilized in several studies identified in this systematic review. They can also contribute to the development of evolving symptom ontologies such as the Symptom Ontology adopted by the Open Biological and Biomedical Foundry (obofoundry.org). Furthermore, an upcoming area of focus for NLP in the creation of symptom resources is the standardization of identified symptom words to controlled vocabularies. Normalization is crucial since several distinct symptom labels (such as discomfort, hurt, aching, sore) are often used to denote a same symptom idea (namely, pain). In support of the emphasis on oncology in the realm of symptom research, Miaskowski et al 24 found that over 83% of the n = 158 publications examined in a study of co-occurring symptoms in chronic illnesses focused on patients with cancer.

Surprisingly, less than 75% of papers included information about the specific number of patients from whom clinical free-text data was collected, and only 33% of studies included any details about the demographic features of the patients. The publications by Iqbal et al23 and Matheny et al22 aimed to create rule-based algorithms for identifying adverse medication events and infectious symptoms, respectively. These studies did not include information on the number of distinct patients or patient demographics. On the other hand, the articles authored by Patel et al25 and Vijayakrishnan et al26 sought to examine the influence of symptoms on clinical outcomes and the occurrence of symptoms in particular clinical populations. Both of these articles provide detailed information about the number of patients involved and their demographic characteristics, such as age, gender, and race. It is crucial to include information on the patients from whom clinical free-text was gathered since symptom experience is known to differ based on common sociodemographic characteristics such as age, sex or gender, race and ethnicity, and socioeconomic level.27 Analyzing and reporting patient information in electronic health records (EHRs) is crucial for future research in natural language processing (NLP) of symptoms. This allows for the generalization of study results, identification of possible biases in evaluation or documentation, and the creation of customized therapies.

The studies in our review encompassed a diverse range of symptoms. However, the most frequently mentioned symptoms in the methods, results, or discussion sections of these studies were 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, also known as angina. The presented symptoms align with the top 10 primary causes for emergency room visits, which include chest discomfort and associated symptoms, difficulty breathing, pain not localized to a single bodily system, and vomiting (i.e., the common indication accompanying nausea).28

Nevertheless, it is important to note that several research have examined symptoms and signs simultaneously, either by failing to differentiate between the two concepts or by incorrectly categorizing signals as symptoms. As previously stated in this overview, symptoms refer to individual experiences, whereas signs are tangible indications of an illness. The lack of clarity is unsurprising given symptoms (such as pruritus or skin itchiness) and signs (such as a rash) often occur together, with signs typically being referred to as "physical" symptoms. However, this discovery emphasizes the need of using symptom information from EHR free-text narratives to describe or diagnose illness, rather than only studying the symptoms themselves. In addition, research mostly used record of symptom incidence or frequency to evaluate symptoms. While many research used negation algorithms, such as the absence of shortness of breath, into their natural language processing (NLP) procedures, only one study specifically assessed the severity of symptoms.29 Heintzelman et al29 created contextual criteria to classify mentions of pain into categories such as no pain, some pain, managed pain, and severe pain, based on the degree of the discomfort. Future research is very interested in integrating precise identification of severity, along with other contextual criteria like symptom location or duration, into NLP algorithms for Electronic Health Records (EHR).

4. Summary

This review included the synthesis of data from 27 papers that examined the use of Natural Language Processing (NLP) to process or analyze symptom information obtained from the free-text narratives of patient Electronic Health Records (EHRs). To summarize, our findings indicate that NLP tools, classification techniques, and manually curated rule-based processing are used to extract information from EHR free-text narratives authored by various healthcare professionals. These narratives cover a broad spectrum of symptoms across different clinical specialties. The present emphasis in this research is on devising techniques to collect symptom data and use that data for illness categorization purposes, rather than studying the symptoms themselves.

To address the high occurrence of symptoms and the resulting load on patients and healthcare providers, future research should focus on examining individual symptoms and documenting them in the free-text narratives of patients' electronic health records (EHRs), as well as using symptoms for other purposes. For the study of symptoms and symptom documentation from electronic health records (EHRs) using natural language processing (NLP), it is crucial to have a clear statement of the symptoms being evaluated, a detailed description of the clinical population from which symptom information was extracted and analyzed, open sharing of user-developed NLP algorithms or pipelines and vocabularies related to symptoms, and the establishment of formal reporting standards for investigations using NLP methodologies.

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. Watson M. When will ‘open science’ become simply ‘science’? Genome Biol 2015; 16: 101.
21. McKiernan EC, Bourne PE, Brown CT, et al. How open science helps researchers succeed. Elife 2016; 5: 372.
22. 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
23. 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.
24. 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.
25. 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.
26. 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.
27. Corwin EJ, Berg JA, Armstrong TS, et al. Envisioning the future in symptom science. Nurs Outlook 2014; 62 (5): 346–51
28. Rui P, Kang K. National Hospital Ambulatory Medical Care Survey: 2015 Emergency Department Summary Tables.
29. 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.