How to Cite:
Alonayzan, H. H. M., Alenezi, Talal S. S., Alshammari, Khalaf S. F., Albakr, Mohammed S. B., Alshammari, Sanad H. S., Alghadeer, Saleh O. A., ... Alanzi, A. S. B. (2020). The use of natural language processing in medical record analysis: Review. *Tennessee Research International of Social Sciences*, *2*(2), 12–25. Retrieved from https://triss.org/index.php/journal/article/view/43

The use of natural language processing in medical record analysis: Review

Hamad Hassan Mohammed Alonayzan
KSA, National Guard Health Affairs
Email: Alonizanha@ngha.med.sa

Talal Sanian Salem Alenezi
KSA, National Guard Health Affairs
Email: Alenezyta@ngha.med.sa

Khalaf Saud Faryhan Alshammari
KSA, National Guard Health Affairs
Email: alshammarykh@ngha.med.sa

Mohammed Saad Bakr Albakr
KSA, National Guard Health Affairs
Email: al-bakrmo@ngha.med.sa

Sanad Hamdan Sanad Alshammari
KSA, National Guard Health Affairs
Email: aIshammarisa5@ngha.med.sa

Saleh Obaid Abdullah Alghadeer
KSA, National Guard Health Affairs
Email: Alghadersa@ngha.med.sa

Nezar Mohammad Mutlaq Alshammari
KSA, National Guard Health Affairs
Email: Alshammaryne@ngha.med.sa

Fahad Khalifah Salem Almughamis
KSA, National Guard Health Affairs
Email: almgamasfa@ngha.med.sa

Nuri Rawafa Alanzi
KSA, National Guard Health Affairs
Email: alenazinu@ngha.med.sa

Tennessee research international of social sciences © 2020.
ISSN: 2766-7464 (Online)
Publisher: Smoky Mountain Publishing
Manuscript submitted: 18 Sept 2020, Manuscript revised: 09 Oct 2020, Accepted for publication: 27 Nov 2020

Abdullah Ibrahim Hamran
KSA, National Guard Health Affairs
Email: Hamranab@ngha.med.sa

Fawaz Ayed Al-Sharari
KSA, National Guard Health Affairs

Ahmed Turki Alotaibi
KSA, National Guard Health Affairs
Email: A7mdoh85@gmail.com

Awad Shehab B Alanzi
KSA, National Guard Health Affairs

***Abstract*---**Background: In order to address the increasing prevalence of chronic illnesses globally, it is necessary to develop innovative methods that supplement and surpass evidence-based treatment in this field. An optimistic approach is the use of electronic health records (EHRs) for the purpose of conducting clinical and translational research by analyzing patient data. Machine learning methods used to electronic health records (EHRs) are leading to enhanced comprehension of patient clinical paths and the ability to forecast the risk of chronic diseases. This presents a distinct chance to uncover previously unidentified clinical knowledge. Nevertheless, a substantial amount of clinical histories are still inaccessible due to being stored as unstructured free-form text. Unlocking the whole potential of EHR data relies on the advancement of natural language processing (NLP) techniques to automatically convert clinical text into structured clinical data. This structured data may then be used to inform clinical choices and perhaps postpone or prevent the start of diseases. Aim of Work: The aim of the study was to provide a thorough examination of the progress and adoption of NLP techniques used in analyzing clinical notes about chronic illnesses. This included exploring the difficulties encountered by NLP methodology in comprehending clinical narratives. Methods: The study followed the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. Searches were performed in 5 databases using keywords such as "clinical notes," "natural language processing," and "chronic disease" to ensure comprehensive coverage of the articles. Results: Natural language processing (NLP) has been more heavily applied to medical records for diseases of the circulatory system. The analysis has shown a notable rise in the use of machine learning techniques in contrast to rule-based approaches. Nevertheless, deep learning methods are still in the early stages of development, as indicated by a sample size of three studies. As a result, most studies prioritize the categorization of illness characteristics, whereas only a few articles tackle the extraction of comorbidities from unstructured text or the merging of clinical notes with organized data. The use of relatively basic approaches, such as shallow classifiers (or their combination

with rule-based methods), is remarkable owing to the interpretability of predictions. However, this still poses a substantial challenge for more complicated methods. Insufficient development of more complex technologies, including extracting word embeddings from clinical notes, may have been caused by a lack of publicly accessible data. Conclusion: Further improvements are needed in clinical NLP methods to advance from extraction to understanding, including recognizing relationships between entities rather than considering them in isolation. Additionally, temporal extraction is necessary to comprehend past, present, and future clinical events. It is also important to utilize alternative sources of clinical knowledge and have access to large-scale, de-identified clinical corpora.

***Keywords*---**Clinical Notes, Electronic Health Records, Chronic Diseases, Natural Language.

Introduction

The prevalence of chronic illnesses, such as malignancies, diabetes, and hypertension, is generally acknowledged as a major obstacle in the field of healthcare. Despite major advancements in the identification of novel therapies and methods of prevention, this problem continues to exist and is actually increasing in frequency [1]. This has a substantial effect on the quality of life for patients and the expenses associated with their care. Hence, there is a need for innovative methodologies to supplement and surpass existing evidence-based therapy, which may mitigate the consequences of chronic ailments on contemporary society.

An auspicious avenue involves the secondary utilization of electronic health records (EHRs) for the purpose of scrutinizing patient data, propelling medical research forward, and enhancing clinical decision making via improved information. Analysis of electronic health records (EHRs) is leading to enhanced comprehension of patient clinical trajectories, as well as facilitating more accurate patient categorization and risk prediction. Specifically, the utilization of machine learning, particularly deep learning, for the analysis of electronic health records (EHRs) is presenting a distinct chance to uncover previously unidentified clinical knowledge [2]. This is particularly applicable to chronic illnesses because to their long-term nature, which generates a substantial and uninterrupted flow of data. From this data, significant patterns with clinical relevance may be identified and used to inform therapeutic choices, such as delaying or avoiding the beginning of the disease [3-5].

Nevertheless, Electronic Health Records (EHRs) pose difficulties in terms of their complex representation and modeling, primarily because of their extensive dimensions, presence of noise, heterogeneity, sparsity, incompleteness, random mistakes, and systematic biases. Furthermore, a vast amount of data on a patient's health history is often inaccessible due to the use of unstructured clinical narratives. This is because writing text is still the most instinctive and comprehensive way to record clinical occurrences [6-8]. The advancement of

natural language processing (NLP) techniques is crucial for the automated conversion of clinical text into organized clinical data that can be readily analyzed using machine learning algorithms. The clinical sector is increasingly adopting the use of NLP, which has different applications. These applications include the identification of biological concepts from radiology reports [9], nursing documentation [10], and discharge summaries [11]. Frameworks using Natural Language Processing (NLP) in clinical narratives have not been extensively used in clinical contexts for aiding decision support systems or processes.

Traditionally, important information from clinical notes has been obtained by professionals manually reviewing them, which has resulted in difficulties in expanding the process and increased expenses. This is especially important for chronic illnesses since clinical notes are more prevalent than organized data. For instance, Wei et al [12] quantitatively demonstrate the higher volume of clinical notes compared to structured data for chronic diseases such rheumatoid arthritis, Parkinson's disease, and Alzheimer's disease. The presence of this data presents a significant opportunity for NLP to extract clinically relevant information that might potentially postpone or prevent the beginning of diseases. However, this also gives rise to many obstacles. This study seeks to find strategies that might accelerate the use of Natural Language Processing (NLP) for analyzing clinical notes related to chronic illnesses. Additionally, it aims to give insights into the present obstacles and advancements in this field.

Aim of Work

Past publications have included systematic studies on the processing of clinical notes [13-18]. However, none of these reviews have particularly examined chronic illnesses, which make it challenging to draw definitive findings and suggestions in this particular and varied field. This work specifically examines the NLP difficulties associated with 43 distinct chronic illnesses found via our comprehensive assessment, and explores the patterns of using different NLP techniques for clinical translational research. The review's findings led us to propose several recommendations for future research. These include: advancing clinical NLP methods to focus on comprehension rather than just extraction; considering the relationships between entities rather than treating them in isolation; incorporating temporal extraction to gain insight into past, present, and future clinical events; utilizing additional sources of clinical knowledge; and making large-scale deidentified and annotated clinical corpora more accessible.

Methods

We conducted a comprehensive search across many databases, including Scopus, Web of Science (which includes MEDLINE) and PubMed, as well as the Association for Computing Machinery (ACM) Digital Library, to identify all publications published between January 1, 2007, and February 6, 2018, that may possibly be relevant. The search has been restricted to journal articles published only in the English language. The searches were conducted using the following groups of keywords: (1) "clinical notes," "medical notes," or "clinical narratives"; (2) "natural language processing," "medical language processing," "text mining," or "information extraction"; and (3) "chronic disease," "heart

disease," "stroke," "cancer," "diabetes," or "lung disease" (representing the top five chronic diseases). The search phrases were chosen to be comprehensive in order to maximize the extent of coverage of the articles.

Circulatory System Disorders
Cardiovascular diseases

The predominant research in this field has been on using Natural Language Processing (NLP) to assess the likelihood of developing cardiac disease. For instance, Chen et al [13] created a hybrid pipeline that combines machine learning and rules to detect medically significant information about the risk of heart disease and monitor the progression of the disease using longitudinal patient records, including clinical notes (Torri et al [14]). Karystianis et al [15] and Yang et al [16] assessed the detection of risk variables for heart disease using clinical notes of diabetes patients. Roberts et al [17] used a somewhat different strategy by concentrating on the estimation of heart disease risk via the categorization of 8 risk triggers, such as aspirin. Previous research in this field has concentrated on assessing the use of aspirin as a potential risk factor [18,19], deriving heart function values from echocardiograms [20], investigating deep vein thrombosis and pulmonary embolism [21], and examining the relationship between low-density lipoprotein levels and the usage of statins [22]. The likelihood of stroke and significant bleeding in individuals with atrial fibrillation has been forecasted by using structured data and clinical notes [23], however people with heart failure have been recognized only based on clinical notes [24]. Furthermore, Italian medical data have been used to detect instances of arrhythmia [25].

Peripheral and Coronary Arterial Disease

Multiple studies used natural language processing (NLP) to detect instances of peripheral arterial disease (PAD) and critical limb ischemia from clinical notes [26, 27]. This includes a genome-wide association research that specifically focused on PAD to identify medicines, illnesses, signs/symptoms, anatomical locations, and treatments [28]. Leeper et al [29] used natural language processing (NLP) to detect patients with peripheral artery disease (PAD) for the purpose of conducting a safety surveillance research on the effects of Cilostazol. The investigation revealed the occurrence of serious consequences such as malignant arrhythmia and sudden death, which were not previously detected in connection with the use of this medication. Moreover, the Clinical Text Analysis Knowledge Extraction System (cTAKES) has been used to analyze the clinical records of diabetic patients in order to forecast the occurrence of peripheral artery disease (PAD) [30].

Hypertension

The primary emphasis of research on hypertension has been on Natural Language Processing (NLP) techniques to extract important indicators, comorbidities, and pharmacological treatments. The Bulgarian language was analyzed in 100 million outpatient notes to extract numerical blood pressure readings with a high sensitivity and recall **[**31]. The phrase "hypertension" was recovered from free-text notes using a rule-based, open-source technology [32]. The identification of hypertensive people was facilitated by using clinical notes and other medical

records, employing the open-source pharmaceutical information extraction (IE) system MedEx [33].

Right-sided, left-sided, and congestive heart failure

Byrd et al [34] introduced a hybrid NLP model that aims to detect signs and symptoms of Framingham heart failure from clinical notes and EHRs. Specifically, the model classifies whether the Framingham criteria are claimed or not. The left ventricular ejection fraction was obtained from unstructured echocardiography data [35], whereas unorganized, longitudinal electronic health records (EHRs) of diabetic patients were used to extract pertinent information on heart disease, using naïve Bayes and conditional random field (CRF) classifiers [36].

Wang et al [37] introduced a technique for prospectively validating the detection of congestive heart failure (CHF) using electronic health records (EHRs). In addition, the left ventricular ejection fraction, together with its corresponding qualitative and quantitative measurements, was used to identify individuals who were at risk of congestive heart failure (CHF) [38]. Meanwhile, free-text notes were employed to differentiate between left and right heart failure [39].

Heart Failure

Topaz et al [40] devised an algorithm to detect heart failure (HF) patients who exhibit inadequate self-management of food, physical activity, medication adherence, and clinical visits via the analysis of discharge summary notes. In contrast, Garvin et al [41] concentrated on assessing the quality of treatment provided to HF patients. Vijayakrishnan et al [42] investigated the use of a text and data-mining technique that has been previously validated to detect the presence of criteria for signs and symptoms of heart failure in the electronic health records (EHRs) of a large primary care population. The researchers discovered that indications and symptoms of HF were recorded far more often among individuals who eventually had HF, even years before their first diagnosis. This suggests that early identification of HF may play a significant role in the future. Lastly, regular expressions were used to discover pre-established psychosocial characteristics that acted as predictors of the probability of hospital readmission after a case of HF [43].

Tumors

Kasthurirathne et al [44] assessed the efficacy of conventional classification algorithms in identifying cancer cases from unstructured pathology records using non-dictionary methods. Yim et al [45] investigated the use of a machine learning method to extract tumor features from radiology data by implementing reference resolution. Jensen et al [46] devised a model that enables the estimation of disease trajectories in cancer patients based on clinical literature. Napolitano et al [47] improved the process of extracting cancer staging information by introducing a model for semistructured reports, which demonstrated superior performance compared to relying only on unstructured data.

Several studies have examined various uses of NLP in the fields of pathology, histopathology, and radiology reports. These include extracting specific domain entities from cancer pathology reports, detecting negations of medical entities in pathology reports, translating sentences from pathology reports into graph representations, extracting information from pathology reports and classifications, and recognizing named entities from histopathology notes. The three predominant forms of malignancies detected are breast cancer (n=8), colorectal cancer (n=7), and prostate cancer (n=4).

Carrell et al [48] developed a natural language processing (NLP) system for analyzing clinical text and detecting instances of breast cancer recurrence. Miller et al [49] introduced a technique for coreference resolution in clinical literature, which was assessed both within the domain (colon cancer) and across domains (breast cancer). Mykowiecka et al [50] provide a rule-based information extraction (IE) system that was assessed using mammography records. Bozkurt et al [51] used natural language processing (NLP) techniques to identify abnormalities in unstructured mammography reports and extract their associated connections, resulting in a comprehensive information framework for each abnormality.

Electronic health records (EHRs) and natural language processing (NLP) were used to identify individuals who need colorectal cancer screening and to identify ideas linked to colonoscopy and temporal information. In addition, electronic health records (EHRs) and natural language processing (NLP) were used to identify individuals who had positive prostate biopsies for prostatic cancer [52]. Ping et al [53] collected textual data related to certain clinical themes from various clinical reports of patients diagnosed with liver cancer, while Al-Haddad et al [54] selected patients with verified surgical pathology diagnosis of intraductal papillary mucinous neoplasms.

Endocrine, Nutritional, and Metabolic Diseases

Some applications of Natural Language Processing (NLP) in the field of endocrine, nutritional, and metabolic diseases involve detecting negation and identifying mentions of family history in unstructured text notes. NLP is also used to assign temporal tags to medical concepts [35]. Furthermore, NLP techniques are employed in the identification of obesity [25] and diabetes [47-50]. NLP can also be used to analyze diabetes complications, such as foot examination findings, vision loss, and quantifying the occurrence of hypoglycemia.

The study used two support vector machines (SVMs) to automatically classify obesity kinds by extracting obesity and diabetes-related ideas from clinical literature, together with patient identification [55]. A Support Vector Machine (SVM) system was created and verified to detect Electronic Health Record (EHR) progress notes related to diabetes, while foot examination results from clinical reports were used to forecast quality of life. In addition, a comprehensive examination of a vast electronic health record (EHR) database was conducted to measure the frequency of hypoglycemia [24].

Additional Disease Classifications

Various studies focus on diseases of the musculoskeletal system and connective tissue. These studies specifically involve the classification of text snippets related to axial spondyloarthritis in the electronic medical records (EMRs) of US military veterans using natural language processing (NLP) and support vector machines (SVM), the characterization of systemic lupus erythematosus, and the identification of rheumatoid arthritis patients through ontology-based NLP and logistic regression. In the field of digestive system diseases, Chen et al [29] utilized natural language characteristics from pathology reports to identify patients with celiac disease. Soguero-Ruiz et al [56] employed feature selection and support vector machines (SVMs) to detect early complications following colorectal cancer. Chang et al [57] combined rule-based natural language processing (NLP) on medical notes with International Classification of Diseases, Ninth Revision (ICD-9) codes and laboratory values in an algorithm to more accurately define and assess the risk of patients with cirrhosis.

Two research publications assessed the effectiveness of deep learning in a domain that encompasses many diseases. Miotto et al [3] developed a patient representation using neural networks that combines structured clinical data and clinical notes from aggregated electronic health records (EHRs). This patient representation is designed to support clinical predictive modeling by providing a comprehensive view of the patient's state. The clinical notes were analyzed using the National Center for Biomedical Ontology's Open Biomedical Annotator to identify medical terminology. These phrases were then subjected to topic modeling using latent Dirichlet allocation. Shi et al [58] suggested evaluating the risk of illness based on patient clinical notes by using word embeddings and convolutional neural networks with a fully connected layer.

Neural networks were used to analyze clinical notes in order to classify mental diagnoses. Specifically, this model consisted of two neural networks, one proficient in accurately declining patients but lacking in its ability to select appropriate ones, and the other possessing the inverse qualities. Comorbidity networks were created using patient records from the biggest psychiatric hospital in Denmark to identify relationships between different mental and behavioral illnesses [59].

IE, or information extraction, based on NLP, or natural language processing, was utilized for various purposes. These include screening computed tomography reports for invasive pulmonary mold, discovering the co-occurrences of chronic obstructive pulmonary disease with other medical terms, quantifying the relationship between aggregated preoperative risk factors and cataract surgery complications, detecting patients with multiple sclerosis from clinical notes before their health care providers recognized it, and identifying patients on dialysis in the publicly available Multiparameter Intelligent Monitoring in Intensive Care II (MIMIC-II) dataset.

In their study, Pivovarov and Elhadad [60] utilized clinical notes from patients with chronic kidney disease to verify a new model for calculating the similarity between two medical concepts. This model combines additional information

obtained from the patterns of clinical documentation usage, accepted definitions, and the position of the concepts in ontology.

Methods for extracting information

To comprehend the patterns in NLP techniques for chronic illnesses, we conducted a thorough analysis of articles, focusing on the utilized approaches, namely machine learning vs rule-based learning. Although machine learning approaches are being used more often compared to rule-based methods, the difference is not as significant as anticipated, despite the higher performance of machine learning algorithms shown in the NLP literature [50-57]. This outcome may indicate the ongoing shift from rule-based techniques to machine learning algorithms, where rule-based methods are used as a benchmark for evaluating the effectiveness of machine learning approaches.

Natural Language Processing

The NLP methodologies discussed in the examined publications and related techniques indicate that the most often mentioned tasks are text categorization and entity recognition. The bulk of the publications discuss text classification problems using conventional techniques in natural language processing (NLP), such as Support Vector Machines (SVM) (n=12) and naïve Bayes (n=4). Entity recognition methods rely on both humanly created resources such as dictionaries, regular expressions, and handwritten rules, as well as machine learning techniques. Regarding the first option, there are ways that depend on dictionaries with a size of 5, while the second option involves approaches that use regular expressions with a size of 12. Regarding the latter, the methods mostly rely on conventional machine learning techniques like CRF and deep learning. Several studies outline methods for coreference resolution (n=2) and negation detection (n=3). Coreference resolution is tackled using Support Vector Machines (SVM), whereas negation detection is performed using SVM (n=2) or human rules (n=1).

Most of the publications discuss experiments conducted on datasets that are not accessible to the public. These datasets are usually comprised of clinical data acquired at healthcare facilities and used by internal natural language processing (NLP) teams. However, of the 16 articles that use publicly accessible corpora, 12 of them make use of the Informatics for Integrating Biology and the Bedside (i2b2) datasets. The other four public datasets used are MIMIC-II, PhenoCHF, Temporal Histories of Your Medical Event (THYME), and Cancer Deep Phenotype Extraction (DeepPhe) [49,61,62].

Conclusion

Our analysis has shown that Support Vector Machines (SVM) and naïve Bayes algorithms were often used for tasks involving machine learning, either alone or in conjunction with rule-based approaches. This might be attributed to the widespread use of these algorithms and the fact that naïve Bayes, being a very simple method, needs a comparatively minimal quantity of training data, especially when compared to more complex classifiers such as deep learning models. While it is not possible to directly compare the algorithmic performance of

the studies we examined due to the diversity of data and challenges they addressed, we have observed that the most frequently reported performance measures were sensitivity (recall), positive predictive value (precision), and F score.

Ultimately, our analysis has shown that the availability of public datasets continues to be limited. The anticipated result was largely predictable due to the delicate nature of clinical data, as well as the various legal and regulatory concerns, such as the Health Insurance Portability and Accountability Act and the Data Protection Directive (Directive 95/46/EC) of the European Law (replaced by the General Data Protection Regulation 2016/679). The papers analyzed in this work mostly originated from healthcare facilities that do research and have their own Natural Language Processing (NLP) teams. These teams have the ability to access clinical data. Thus, there is still a need for collaborative projects like i2b2 and access to data that would enhance involvement in clinical NLP and aid in the development of NLP techniques and algorithms specifically designed for clinical applications.

References

1. World Health Organization. WHO Global status report on noncommunicable diseases 2014 URL: https://www.who.int/nmh/publications/ncd-status-report-2014/en/
2. Kruse CS, Kothman K, Anerobi K, Abanaka L. Adoption factors of the electronic health record: a systematic review. JMIR Med Inform 2016 Jun 01;4(2):e19.
3. Miotto R, Li L, Kidd BA, Dudley JT. Deep patient: an unsupervised representation to predict the future of patients from the electronic health records. Sci Rep 2016 Dec 17;6:26094
4. Jensen P, Jensen L, Brunak S. Mining electronic health records: towards better research applications and clinical care. Nat Rev Genet 2012 May 02;13(6):395-405.
5. Goldstein BA, Navar AM, Pencina MJ, Ioannidis JPA. Opportunities and challenges in developing risk prediction models with electronic health records data: a systematic review. J Am Med Inform Assoc 2017 Jan;24(1):198-208.
6. Ye C, Fu T, Hao S, Zhang Y, Wang O, Jin B, et al. Prediction of incident hypertension within the next year: prospective study using statewide electronic health records and machine learning. J Med Internet Res 2018 Jan 30;20(1):e22
7. Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform 2017 May 06.
8. Jensen K, Soguero-Ruiz C, Oyvind MK, Lindsetmo R, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep 2017 Dec 07;7:46226
9. Flynn R, Macdonald TM, Schembri N, Murray GD, Doney ASF. Automated data capture from free-text radiology reports to enhance accuracy of hospital inpatient stroke codes. Pharmacoepidemiol Drug Saf 2010 Aug;19(8):843-847.

10. Popejoy LL, Khalilia MA, Popescu M, Galambos C, Lyons V, Rantz M, et al. Quantifying care coordination using natural language processing and domain-specific ontology. J Am Med Inform Assoc 2015 Apr;22(e1):e93-e103
11. Yang H, Spasic I, Keane JA, Nenadic G. A text mining approach to the prediction of disease status from clinical discharge summaries. J Am Med Inform Assoc 2009;16(4):596-600
12. Wei W, Teixeira PL, Mo H, Cronin RM, Warner JL, Denny JC. Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. J Am Med Inform Assoc 2016 Apr;23(e1):e20-e27
13. Chen Q, Li H, Tang B, Wang X, Liu X, Liu Z, et al. An automatic system to identify heart disease risk factors in clinical texts over time. J Biomed Inform 2015 Dec;58 Suppl:S158-S163
14. Torii M, Fan J, Yang W, Lee T, Wiley MT, Zisook DS, et al. Risk factor detection for heart disease by applying text analytics in electronic medical records. J Biomed Inform 2015 Dec;58 Suppl:S164-S170
15. Karystianis G, Dehghan A, Kovacevic A, Keane JA, Nenadic G. Using local lexicalized rules to identify heart disease risk factors in clinical notes. J Biomed Inform 2015 Dec;58 Suppl:S183-S188
16. Yang H, Garibaldi JM. A hybrid model for automatic identification of risk factors for heart disease. J Biomed Inform 2015 Dec;58 Suppl:S171-S182 [FREE Full text] [CrossRef] [Medline]
17. Roberts K, Shooshan SE, Rodriguez L, Abhyankar S, Kilicoglu H, Demner-Fushman D. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. J Biomed Inform 2015 Dec;58 Suppl:S111-S119
18. Pakhomov S, Shah N, Hanson P, Balasubramaniam S, Smith S. Automated processing of electronic medical records is a reliable method of determining aspirin use in populations at risk for cardiovascular events. Inform Prim Care 2010;18(2):125-133
19. Zheng C, Rashid N, Koblick R, An J. Medication extraction from electronic clinical notes in an integrated health system: a study on aspirin use in patients with nonvalvular atrial fibrillation. Clin Ther 2015 Sep;37(9):2048-2052.
20. Patterson OV, Freiberg MS, Skanderson M, Brandt CA, DuVall SL. Unlocking echocardiogram measurements for heart disease research through natural language processing. BMC Cardiovasc Disord 2017 Dec 12;17(1):151
21. Tian Z, Sun S, Eguale T, Rochefort CM. Automated extraction of VTE events from narrative radiology reports in electronic health records: a validation study. Med Care 2017 Dec;55(10):e73-e80
22. Ross EG, Shah N, Leeper N. Statin intensity or achieved LDL? Practice-based evidence for the evaluation of new cholesterol treatment guidelines. PLoS One 2016;11(5):e0154952
23. Wang SV, Rogers JR, Jin Y, Bates DW, Fischer MA. Use of electronic healthcare records to identify complex patients with atrial fibrillation for targeted intervention. J Am Med Inform Assoc 2017 Mar 01;24(2):339-344.
24. Pakhomov S, Weston S, Jacobsen S, Chute C, Meverden R, Roger V. Electronic medical records for clinical research: application to the identification of heart failure. Am J Manag Care 2007 Jun;13(6 Part 1):281-288

25. Viani N, Larizza C, Tibollo V, Napolitano C, Priori SG, Bellazzi R, et al. Information extraction from Italian medical reports: an ontology-driven approach. Int J Med Inform 2018 Mar;111:140-148.
26. 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
27. 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 Dec;65(6):1753-1761
28. Kullo IJ, Fan J, Pathak J, Savova GK, Ali Z, Chute CG. Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc 2010;17(5):568-574
29. 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
30. Buchan K, Filannino M, Uzuner O. Automatic prediction of coronary artery disease from clinical narratives. J Biomed Inform 2017 Dec;72:23-32
31. Boytcheva S, Angelova G, Angelov Z, Tcharaktchiev D. Text mining and big data analytics for retrospective analysis of clinical texts from outpatient care. Cybern Inf Technol 2015;1(4):55-77
32. Jonnagaddala J, Liaw S, Ray P, Kumar M. HTNSystem: hypertension information extraction system for unstructured clinical notes. Lect Notes Comput Sci 2014:219-227
33. Teixeira PL, Wei W, Cronin RM, Mo H, VanHouten JP, Carroll RJ, et al. Evaluating electronic health record data sources and algorithmic approaches to identify hypertensive individuals. J Am Med Inform Assoc 2017 Jan;24(1):162-171
34. Byrd RJ, Steinhubl SR, Sun J, Ebadollahi S, Stewart WF. Automatic identification of heart failure diagnostic criteria, using text analysis of clinical notes from electronic health records. Int J Med Inform 2014 Dec;83(12):983-992
35. Garvin JH, DuVall SL, South BR, Bray BE, Bolton D, Heavirland J, et al. Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure. J Am Med Inform Assoc 2012;19(5):859-866
36. Jonnagaddala J, Liaw S, Ray P, Kumar M, Dai H, Hsu C. Identification and progression of heart disease risk factors in diabetic patients from longitudinal electronic health records. Biomed Res Int 2015;2015:636371
37. Wang Y, Luo J, Hao S, Xu H, Shin AY, Jin B, et al. NLP based congestive heart failure case finding: a prospective analysis on statewide electronic medical records. Int J Med Inform 2015 Dec;84(12):1039-1047.
38. Kim Y, Garvin JH, Goldstein MK, Hwang TS, Redd A, Bolton D, et al. Extraction of left ventricular ejection fraction information from various types of clinical reports. J Biomed Inform 2017 Dec;67:42-48
39. American Heart Association. Types of Heart Failure URL: https://www.heart.org/en/health-topics/heart-failure/what-is-heart-failure/types-of-heart-failure [

40. Topaz M, Radhakrishnan K, Blackley S, Lei V, Lai K, Zhou L. Studying associations between heart failure self-management and rehospitalizations using natural language processing. West J Nurs Res 2017 Jan;39(1):147-165.
41. Garvin JH, Kim Y, Gobbel GT, Matheny ME, Redd A, Bray BE, et al. Automating quality measures for heart failure using natural language processing: a descriptive study in the Department of Veterans Affairs. JMIR Med Inform 2018 Jan 15;6(1):e5
42. Vijayakrishnan R, Steinhubl SR, Ng K, Sun J, Byrd RJ, Daar Z, 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 Jul;20(7):459-464
43. Watson AJ, O'Rourke J, Jethwani K, Cami A, Stern TA, Kvedar JC, et al. Linking electronic health record-extracted psychosocial data in real-time to risk of readmission for heart failure. Psychosomatics 2011;52(4):319-327
44. Kasthurirathne SN, Dixon BE, Gichoya J, Xu H, Xia Y, Mamlin B, et al. Toward better public health reporting using existing off the shelf approaches: the value of medical dictionaries in automated cancer detection using plaintext medical data. J Biomed Inform 2017 Dec;69:160-176
45. Yim W, Kwan SW, Yetisgen M. Tumor reference resolution and characteristic extraction in radiology reports for liver cancer stage prediction. J Biomed Inform 2016 Dec;64:179-191
46. Jensen K, Soguero-Ruiz C, Oyvind MK, Lindsetmo R, Kouskoumvekaki I, Girolami M, et al. Analysis of free text in electronic health records for identification of cancer patient trajectories. Sci Rep 2017 Dec 07;7:46226.
47. Napolitano G, Marshall A, Hamilton P, Gavin AT. Machine learning classification of surgical pathology reports and chunk recognition for information extraction noise reduction. Artif Intell Med 2016 Dec;70:77-83.
48. Carrell DS, Halgrim S, Tran D, Buist DSM, Chubak J, Chapman WW, et al. Using natural language processing to improve efficiency of manual chart abstraction in research: the case of breast cancer recurrence. Am J Epidemiol 2014 Mar 15;179(6):749-758
49. Miller T, Dligach D, Bethard S, Lin C, Savova G. Towards generalizable entity-centric clinical coreference resolution. J Biomed Inform 2017 Dec;69:251-258
50. Mykowiecka A, Marciniak M, Kupść A. Rule-based information extraction from patients' clinical data. J Biomed Inform 2009 Oct;42(5):923-936
51. Bozkurt S, Lipson JA, Senol U, Rubin DL. Automatic abstraction of imaging observations with their characteristics from mammography reports. J Am Med Inform Assoc 2015 Apr;22(e1):e81-e92.
52. Thomas AA, Zheng C, Jung H, Chang A, Kim B, Gelfond J, et al. Extracting data from electronic medical records: validation of a natural language processing program to assess prostate biopsy results. World J Urol 2014 Feb;32(1):99-103.
53. Ping X, Tseng Y, Chung Y, Wu Y, Hsu C, Yang P, et al. Information extraction for tracking liver cancer patients' statuses: from mixture of clinical narrative report types. Telemed J E Health 2013 Sep;19(9):704-710.
54. Al-Haddad MA, Friedlin J, Kesterson J, Waters JA, Aguilar-Saavedra JR, Schmidt CM. Natural language processing for the development of a clinical registry: a validation study in intraductal papillary mucinous neoplasms. HPB (Oxford) 2010 Dec;12(10):688-695

55. Wei W, Tao C, Jiang G, Chute CG. A high throughput semantic concept frequency based approach for patient identification: a case study using type 2 diabetes mellitus clinical notes. AMIA Annu Symp Proc 2010 Nov 13;2010:857-861
56. Soguero-Ruiz C, Hindberg K, Rojo-Alvarez JL, Skrovseth SO, Godtliebsen F, Mortensen K, et al. Support vector feature selection for early detection of anastomosis leakage from bag-of-words in electronic health records. IEEE J Biomed Health Inform 2016 Dec;20(5):1404-1415.
57. Chang EK, Yu CY, Clarke R, Hackbarth A, Sanders T, Esrailian E, et al. Defining a patient population with cirrhosis: an automated algorithm with natural language processing. J Clin Gastroenterol 2016;50(10):889-894
58. Shi X, Hu Y, Zhang Y, Li W, Hao Y, Alelaiwi A, et al. Multiple disease risk assessment with uniform model based on medical clinical notes. IEEE Access 2016;4:7074-7083.
59. Roque F, Jensen P, Schmock H, Dalgaard M, Andreatta M, Hansen T, et al. Using electronic patient records to discover disease correlations and stratify patient cohorts. PLoS Comput Biol 2011 Aug;7(8):e1002141
60. 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
61. Abhyankar S, Demner-Fushman D, Callaghan FM, McDonald CJ. Combining structured and unstructured data to identify a cohort of ICU patients who received dialysis. J Am Med Inform Assoc 2014;21(5):801-807
62. Alnazzawi N, Thompson P, Ananiadou S. Mapping phenotypic information in heterogeneous textual sources to a domain-specific terminological resource. PLoS One 2016;11(9):e0162287