Natural language processing systems for diagnosing and determining level of lung cancer: A systematic review

Mahdieh Montazeri1, Ali Afraz1, Raheleh Mahboob Farimani2, Fahimeh Ghasemian3*

1Medical Informatics Research Center, Institute for Futures Studies in Health, Kerman University of Medical Sciences, Kerman, Iran
2Department of Health Information Sciences, Faculty of Management and Medical Information Sciences, Kerman University of Medical Sciences, Kerman, Iran
3Department of Computer Engineering, Faculty of Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ABSTRACT

Introduction: Lung cancer is the second most common cancer for men and women. Using natural language processing to automatically extract information from text, lead to decrease labor of manual extraction from large volume of text material and save time. The aim of this study is to systematically review of studies which reviewed NLP methods in diagnosing and staging lung cancer.

Material and Methods: PubMed, Scopus, Web of science, Embase was searched for English language articles that reported diagnosing and staging methods in lung cancer Using NLP until DEC 2019. Two reviewers independently assessed original papers to determine eligibility for inclusion in the review.

Results: Of 231 studies, 7 studies were included. Three studies developed a NLP algorithm to scan radiology notes and determine the presence or absence of nodules to identify patients with incident lung nodules for treatment or follow-up. Two studies used NLP to transform the report text, including identification of UMLS terms and detection of negated findings to classifying reports, also one of them used an SVM-based text classification system for staging lung cancer patients. All studies reported various performance measures based on the difference between combinations of methods. Most of studies have reported sensitivity and specificity of the NLP algorithm for identifying the presence of lung nodules.

Conclusion: Evaluation of studies in diagnosing and staging methods in lung cancer using NLP shows there is a number of studies on diagnosing lung cancer but there are a few works on staging that. In some studies, combination of methods was considered and NLP isolated was not sufficient for capturing satisfying results. There are potentials to improve studies by adding other data sources, further refinement and subsequent validation.

INTRODUCTION

Today cancer is one of the main health issues all over the world [1-4]. The lung cancer is one of the most common cancers worldwide and is the major reason for mortality from cancer in the world [5-9]. Lung cancer is assessed to present 228,150 new cases and 142,670 deaths in 2019 in the United States [10]. Late diagnosis and treatment in lung cancer can result from a failure to act upon abnormal radiologic findings in a timely way [11]. The most common cause of lung cancer is the lengthy exposure to tobacco smoke, which is the explanation of 90% of lung cancers [12-14]. The Percent of lung cancer in individuals who do not smoke is 15% and the reason is because of a combination of factors including genetic factors, radon gas, asbestos, and air contamination such as cigarette smoke of another person [15]. Stage and treatment modality have been the main factors for lung cancer prognosis [16]. The stage of a cancer groups its progression, in terms of the size and location of the primary tumor, just as any contamination such as ci.
and to define outcomes as a basis for population-level analysis of health programs [17]. For many reasons, in any case, formal staging data is not regularly collected for all cancer patients. Staging of patients is suggested as a standard of care by national cancer bodies, e.g. [18], and provides the base for international benchmarking of outcomes. At the present time because of increasing costs and constrained resources from one viewpoint and significant costs of avoidance, screening and treatment of chronic diseases, particularly cancer on the other hand, Health care providers are searching for the most effective and cost-effective care [19]. While most lung nodules are benign, some can characterize malignancy. Whenever detected early and suitable care rendered, there is the possibility to save many lives and meaningfully decrease costs [20]. Significant clinical information is generally recorded in unstructured free text, and converting it to a structured format can be a time consuming task that may not effectively capture all aspects of the information. Nevertheless, there are at least two large motivations for converting unstructured data into structured data. First, the decrease of time required for manual expert review and second the secondary use of these information for large scale automated processing [20]. Text mining includes a range of approaches used for describing and transforming text [21]. NLP is a collection of syntactic and/or semantic rule- or statistical-based processing algorithms that can be used to parse, segment, extract, or analyze text data in text mining [22]. The use of text analysis differentiates these methods; text mining uses the words as a unit of analysis (e.g., frequencies, the presence or absence of specific words of interest), while NLP techniques use the hidden metadata including content and phrase patterns. Both NLP and text mining have been used in health-related online domains such as mental health [23], oncology [24], and infectious disease [25]. During the previous decade, automatic text processing has been used effectively to extract data from electronic radiology and other reports for a variety of purposes.

Clinical text has some properties such as poor structure, abundant shorthand and domain specific vocabularies that make the application of NLP challenging. Other challenges in the field, like identification of temporal associations, assessment of context-dependent text, and concept normalization to particular terminologies, that remain open [26-28]. We have conducted a systematic literature review to detect the used NLP solutions. However, to the best of our knowledge, no previous systematic literature review has summarized these studies. The aim of this study is to systematically review and assess the reporting and methodological quality of all development which were aimed to diagnosing or staging lung cancer using NLP solutions.

MATERIAL AND METHODS

Eligibility criteria

We followed international Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines in conducting this review [29]. Study characteristics for eligibility were: published research of experimental or methodological studies, conference papers, written in English, using NLP techniques on free text for this review. Editorials, commentaries, letters, books, and presentations were excluded. Publications were excluded if no full text could be retrieved. The application of NLP techniques should be in the field of human healthcare. The source of free text should radiology reports, the NLP technique should relate to diagnose of lung cancer or determining the level of this cancer.

Information sources and searches

Articles were identified for this review through a search of PubMed, Scopus, Web of Science, Embase databases up to Dec 2019. The keywords used to retrieve articles from PubMed data base was as following:

("NLP"[Title/Abstract] OR "natural language process"[Title/Abstract]) AND ("Cancer"[Title/Abstract] OR "neoplasm"[Title/Abstract])

Study selection

After removing duplicate studies, two authors independently reviewed titles and abstracts of all identified studies. Fig 1 illustrates the inclusion process. In the second stage, two authors independently assessed the full-text articles. The disagreements were resolved through discussion and if required with referral to a third researcher. During the screening of the studies, the reviewers documented the reasons for the exclusion of each study. We used a free web and mobile application platform for paper screening (Rayyan QCRI systematic review software) [30]. Included studies were those that: (a) used NLP (b) used radiology reports (c) relate to diagnose of lung cancer or determining level of this cancer.

Data collection process

We developed a customized data-extraction form to be used by 2 reviewers to extract specific details about each study. This form consisted of study location, NLP technique type, sample size, performance, library/platform, and evaluation method (Table 1).

Risk of bias assessment

As various databases were searched to extract the
relevant papers, the evidence selection bias may not occur. To decrease the risk of bias in data extraction a reliable data extraction form was developed and before data extraction the reviewers agreed and discussed the concept of each data items in the form.

![Diagram](image)

**Fig 1:** Study identification and selection process.

**RESULTS**

Literature searches identified 231 potentially relevant citations. Of those, 18 studies were deemed eligible for full-text review, and 7 studies fulfilled the inclusion criteria for this systematic review. Table 1 illustrates the characteristics of included studies.

**Performance Measures**

Performance measures that have been used in the studies in this review include sensitivity (also called recall in the field of NLP), specificity, positive predictive value (PPV) (also called precision), F score (harmonized average of recall and precision) [31] and accuracy. When describing the performance of a method, we focus on sensitivity and specificity; PPV is referred to if specificity is unavailable. Some studies provide the F score, which is frequently used in the field of NLP as a single, overall measure of system performance. All studies reported various performance measures based on the difference between combinations of methods. Most (71%) of studies have reported sensitivity and specificity of the NLP algorithm for identifying the presence of lung nodule(s).

**Data Representation and Data Analysis**

Three studies developed a NLP algorithm to scan radiology notes and determine the presence or absence of nodules to identify (all) patients with incident lung nodules for treatment or follow-up. Two studies [17, 32] used natural language processing to transform the report text, including identification of UMLS terms and detection of negated findings to classifying reports, also one of them [17] used an SVM-based text classification system for staging lung cancer patients.

Due to the heterogeneity and multidisciplinary nature of the included studies, a formal meta-analysis was not possible. If a study tested multiple methods for a particular application, we report the performance of the best performing method.

Only two latest studies [32, 33] use deep learning-based approach for extracting information from lung cancer patients’ lung CT radiology report

NLP technique used in included studies grouped in 3 categories: rule base (n=2) [32, 33], machine learning (n=3) [17, 32, 33], and combinations of rule base and machine learning (n=2) [34, 35].

Most studies (71%) reported outcomes for only one NLP technique. The most frequent models developed used variants of Machine Learning Models, including SVM and deep learning.

The majority of studies [32, 35-37] used a manually annotated corpus of free-text documents as the comparator. Internal validations in the studies were done using stratified random sampling and 10-fold or n-fold cross-validation.

**DISCUSSION**

**Principal finding**

Our systematic review has shown that despite the abundance of studies [32, 34-37] using image processing techniques for lung cancer-related applications, there are only seven studies were done about the automatic extraction of useful information from lung cancer patients’ lung CT radiology reports. Although the recent trend of deep learning libraries, toolkits (e.g. TensorFlow, Keras, Lasagne library, deeplearning4j), and studies for natural language processing applications, we found only two latest studies use deep learning-based approach for extracting information from lung cancer patients’ lung CT radiology report [33, 38-40]. Probably because deep learning for natural language processing is a developing area of study, most of the reports published in different type of publications like conferences, and electronic preprints repository like arXiv instead of journals.

None of the reviewed studies used public data which shows the scarcity of public data for this particular subject. This is due to the importance of privacy of clinical data, legal and regulatory compliance such as federal law of Health Insurance Portability and
Accountability Act of 1996 (HIPAA), and the Directive 95/46/EC of the European Parliament for data protection. There is a need for shared public data with compliance with regulation to facilitate the contribution of researchers around the world for improvement of NLP algorithms for clinical and healthcare purposes.

Table 1: Characteristics of included studies

<table>
<thead>
<tr>
<th>Reference</th>
<th>Authors</th>
<th>Publication Year</th>
<th>Study location</th>
<th>Sample size</th>
<th>NLP type</th>
<th>Library/platform</th>
<th>Evaluation</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>[16]</td>
<td>Danforth et al</td>
<td>2012</td>
<td>Kaiser Permanente Southern California</td>
<td>12,389</td>
<td>Rule Based</td>
<td>-</td>
<td>Internal validation: A stratified random sample of 116 patients were used for evaluation.</td>
<td>Sensitivity: 96%(95% CI: 88-99%), specificity: 86%(95% CI: 75-93%), positive predictive value: 87%(95% CI: 77-93%), negative predictive value: 96%(95% CI: 87-99%)</td>
</tr>
<tr>
<td>[17]</td>
<td>Beyer et al</td>
<td>2017</td>
<td>USA-a Lung Cancer Screening Center</td>
<td>1501</td>
<td>Rule Based</td>
<td>UIMA</td>
<td>Internal validation: 496 reports were used for evaluation.</td>
<td>Sensitivity (suspicions case): 74.1%, Specificity (suspicions case): 98.6%, Sensitivity (positive): 75% Specificity (positive): 98.8%</td>
</tr>
<tr>
<td>[17]</td>
<td>Lain McCowan et al</td>
<td>2006</td>
<td>Queensland Cancer Control and Analysis Team (QCCAT) in Queensland Health</td>
<td>718</td>
<td>Machine Learning</td>
<td>SVM light</td>
<td>Internal validation: Reports from more than 700 lung cancer patients were analyzed.</td>
<td>T staging (Primary Tumor) sensitivity: 0.61 specificity: 0.88 PPV: 0.62: N staging(Regional Lymph Nodes) sensitivity: 0.78: specificity: 0.90 PPV: 0.78: F1: 0.78</td>
</tr>
<tr>
<td>[14]</td>
<td>Bipin Kurnakar an et al</td>
<td>2017</td>
<td>• Mayo Clinic part-of-speech corpus • GENIA (literature mining project in molecular biology from the University of Tokyo) • Penn Treebank</td>
<td>30,000 Validation notes:1,096</td>
<td>Rule Based and Machine Learning</td>
<td>cTAKES</td>
<td>Internal validation: A total of 1,096 notes were validated</td>
<td>Precision: 0.873 Recall: 0.947 Accuracy: 0.912 F1: 0.908</td>
</tr>
<tr>
<td>[15]</td>
<td>Roxanne Wadia et al</td>
<td>2018</td>
<td></td>
<td>326</td>
<td>Rule Based and Machine Learning</td>
<td>Yale cTAKES</td>
<td>Internal validation: A total of 450 reports representing 428 patients were analyzed</td>
<td>NLP sensitivity: 77.3% manual coding sensitivity: 51.5% NLP specificity: 72.5% manual coding specificity: 82.5% NLP PPV: 88.4% manual coding PPV: 88.9% NLP NPV: 54% manual coding NPV: 38.5%</td>
</tr>
<tr>
<td>[12]</td>
<td>Liwei Wang et al</td>
<td>2019</td>
<td>Mayo Clinic EHR at Rochester site</td>
<td>2311</td>
<td>Machine Learning</td>
<td>-</td>
<td>Internal validation: precision range from 0.704 to 1. Recall from 0.779 to 1 for 30 days to 365 days</td>
<td>recall stage:99 recall histology: 98 recall tumor grade: 78 recall therapies achieving: 100 precision stage:70 precision histology: 88 precision tumor grade: 90 precision therapies achieving: 100</td>
</tr>
<tr>
<td>[22]</td>
<td>Er. Khushbu Gupta et al</td>
<td>2019</td>
<td>The Department of Radiology, Ramathibodi Hospital Mahidol University</td>
<td>2000 CT-scan reports of lung cancer patients</td>
<td>Machine Learning</td>
<td>-</td>
<td>Internal validation: accuracy 59% for T-staging, 36% for N-staging and 41% for M-staging</td>
<td>accuracy up to 85%</td>
</tr>
</tbody>
</table>

Due to difficulty in the acquisition of radiology reports from other institutions most authors would be forced to use data from the same institution for train and validation. This will result in a lack of external validations or scarcity of external validation in NLP studies in medicine, as such external validation was absent from all of the included studies of this review. Implementation of NLP methods in a real environment is also necessary. Despite commercially available NLP tools for radiology
reports, none of the included studies has implemented in healthcare institutions. Lack of generalizability due to paucity of data for external validation and clinicians’ distrust due to the unclear nature of the way of generating output evidently are important contributing factors to being not popular in practice.

In this review, we have only included application of NLP methods for lung cancer radiology report, but NLP techniques have been adapted in other kinds of radiology reports [41-43]. There are many other NLP medical applications such as EHR information extraction, and identification of diseases [44-46]. The use of NLP in radiology report applications may benefit from NLP applications that operate on EHR data from other fields. For instance, the NLP task of identifying a clinical diagnosis from broader EHR content can help provide a more definitive reference standard to diagnoses.

Limitations
To get quality studies, we searched four prestigious academic research databases (PubMed, Scopus, Web of Science, and Embase). Because of it, we may miss some new studies. Conference and electronic preprint manuscripts are more likely to be biased, which will lead to misrepresenting performance parameters to excluding studies that are more probable to be biased justify our decision.

In this review we found studies for identification of cancer patients, finding most urgent suspected lung cancer cases for follow up, determine the probability of having lung cancer, and determine the stage of lung cancer, in which various NLP methods were applied. The diversity of the different situations prevents us to do a meta-analysis of the studies. Also, there was only one study that compared two different NLP methods, and hence it is difficult to conclude on which NLP techniques is the best.

Recommendations
Consider the great potential of deep learning approaches for natural language processing, and the benefits of high accuracy automated information extraction of lung cancer patients’ chest CT radiology reports, there is a need for more research in this area.

Most of NLP applications in cancer diagnosis and treatment remain in proof of concept stage and never merge in clinical practice routine. Identify the existing challenges and finding their solutions help clinicians to save more lives at the end of days.

CONCLUSION
Evaluation of studies in diagnosing and staging methods in lung cancer using natural language processing (NLP) shows there is a number of studies on diagnosing lung cancer but there are a few works on staging that. In some studies, combination of methods was considered and NLP isolated was not sufficient for capturing satisfying results. There are potentials to improve studies by adding other data sources, further refinement and subsequent validation.

AUTHOR’S CONTRIBUTION
The authors agree on this final form of the manuscript, and attested that all authors contributed in the final draft of the manuscript.

CONFLICTS OF INTEREST
The authors declare no conflicts of interest regarding the publication of this study.

FINANCIAL DISCLOSURE
No financial interests related to the material of this manuscript have been declared.

REFERENCES
Natural Language Processing Systems for Diagnosing and Determining Level of Lung Cancer


