Published in final edited form as:

IEEE Int Conf Healthc Inform. 2019 June ; 2019: .

Information Extraction for Populating Lung Cancer Clinical Research Data

Liwei Wang,

Department of health Sciences Research Mayo Clinic, Rochester, MN, U.S.

Lei Luo,

Department of Good Clinical Practice Guizhou Province People's Hospital Guiyang, Guizhou, China.

Yanshan Wang,

Department of health Sciences Research Mayo Clinic, Rochester, MN, U.S.

Jason A. Wampfler,

Department of health Sciences Research Mayo Clinic, Rochester, MN, U.S. Wampfler.

Ping Yang,

Department of health Sciences Research Mayo Clinic, Scottsdale, AZ, U.S.

Hongfang Liu

Department of health Sciences Research Mayo Clinic, Rochester, MN, U.S.

Abstract

Lung cancer is the second most common cancer and the wide adoption of electronic health records (EHRs) offers a potential of accelerating cohort-related epidemiological studies using informatics approaches. In this study, we developed and evaluated a natural language processing (NLP) system to extract information on stage, histology, grade and therapies (chemotherapy, radiotherapy and surgery) automatically for lung cancer patients from clinical narratives including clinical notes, pathology reports and surgery reports. Evaluation showed promising results with the recalls for stage, histology, grade, and therapies achieving 89%, 98%, 80%, and 100% respectively and the precisions were 71%, 89%, 90%, and 100% respectively. This study demonstrated the feasibility and accuracy of extracting related information from clinical narratives for lung cancer research.

Keywords

natural language processing; lung cancer; stage; histology; grade; treatments

I. INTRODUCTION

Lung cancer is the second most common cancer and by far the leading cause of cancer-related death in both males and females, accounting for 1 in 4 cancer deaths in U.S. [1]. Accurate identification of lung cancer related information is very important for

epidemiological studies, especially in terms of prognosis [2], therefore is critical for improving cancer outcomes. Epidemiologists use electronic health records (EHR) with rich longitudinal data on large populations for epidemiologic research [3]. Since manual extraction from large volumes of text materials is time consuming and labor intensive, some efforts have emerged to automatically extract information from text for lung cancer patients using natural language processing (NLP), an artificial intelligence technique. In this study, we developed and evaluated an information extraction system to capture information on stage, histology, grade and therapy in lung cancer patients using various clinical narrative documents including clinical notes, pathology reports and surgery reports.

II. METHODS

Figure 1 shows the overall study design. Based on an existing lung cancer cohort including 2,311 lung cancer patients with information about stage, histology, grade, and therapies manually ascertained, an information extraction system was developed using the open source clinical NLP pipeline MedTagger as the platform [4]. Specifically we utilized the sentence detection and tokenization parts in MedTagger. Then the system integrated rules and algorithms to generate final normalized concept names for each data element. We evaluated the result of NLP system against the human abstracted results of the existing dataset.

In error analysis, we used deep learning, i.e., Convolutional Neural Networks (CNN) [5], to predict values for histological cell types using sentences labeled by NLP system as input. Then we analyzed NLP results, deep learning prediction results and the reference standard for error analysis in terms of histology extraction. we randomly selected 100 patients as the testing set and the remaining patients as training set for deep learning. Using the classifiers trained from deep learning, we predicted the histological types of the 100 patients using clinical notes, pathology reports and surgery notes.

In our implementation, we utilized the typical CNN framework that contains embedding layer, convolution layer and fully connected layer with a softmax function. For the embedding layer, we used a pre-trained word embedding matrix to represent each word in a clinical document in the embedding space. The word embedding matrix was pre-trained by word2vec [6] on a corpus of clinical notes of 113k patients who received their primary care at Mayo Clinic [7]. Then the sequences for a document from the embedding layer is input to a convolution layer where rectify linear unit (ReLU) is used as the convolutional function. 1-max pooling is then performed to choose the most useful feature from each sequence. Finally, to classify the document, we utilize a fully connected layer over global features and a softmax function with the dimension of the number of categories.

III. RESULTS

Results showed that the recalls for stage, histology, grade, and therapies achieved 89%, 98%, 80%, and 100% respectively and the precisions were 71%, 89%, 90%, and 100% respectively

In error analysis, we conducted combined analysis following the approach in the evaluation part. Then we analyzed 5 error results that weren't in the reference standard but in both deep learning prediction and NLP extracted results. The findings discovered that 4 resulted from the failure of identifying subtypes such as Adenocarcinoma but have been identified as up-level type non-small cell. Further investigation found that 2 patients have no subtype information in EHR including clinical notes, pathology reports or surgery reports. Another 2 patients have related subtype information only in clinical notes, but not in pathology reports, while our algorithm chose the result from pathology reports, i.e., non-small cell, therefore missed the subtype. The additional 1 patient was labeled as other NSCLC in reference standard, but was identified by both NLP and deep learning as "small cell". We looked into the EHR, all related data sources including pathology reports, clinical notes and surgery notes indicated the patient had small cell.

IV. DISCUSSION

Due to the historic reasons, patient report error or health providers' writing error, discordance of recording data elements was very common. In preparing the reference standards, human abstractors often met the same situation where a pathologist will be involved to make the final judgement. In our NLP system, we developed an algorithm to resolve the data discordance issue, where concepts with highest frequency or more advanced concepts have been used. Compared to the previous study on histological type and grade extraction [8], our NLP system obtained the similar precisions (0.88, 0.90). Our study focused on more specific stages such as Ia, Ib, IIa or IIb, not only stage I and stage II. Compared to the previous study on stage extraction [8], our NLP system obtained similar performance in distinguishing more specific stages. The previous study was able to distinguish stage IIIA from stage IIIB with the accuracy in the 64% to 79% range [8]. Our system yielded the precision around 70%. The reason why the performance for stage was not very high maybe because three staging systems have been used in the past 20 years at Mayo Clinic. In addition, all performances for therapies were 100%. These findings demonstrated that our NLP rules and algorithms were effective in identifying data elements.

V. CONCLUSION

This study demonstrated the feasibility and accuracy of extracting cancer related information from narrative EMR data for clinical research of lung cancer, as well as the feasibility of improving the efficiency of human abstractors through NLP techniques.

ACKNOWLEDGMENT

This work was supported by National Cancer Institute 1U24CA194215-01A1.

REFERENCES

1. Cancer Facts & Figures 2017 [<https://www.cancer.org/research/cancer-facts-statistics/all-cancer-facts-figures/cancer-facts-figures-2017.html>]
2. Yang P: Epidemiology of lung cancer prognosis: quantity and quality of life In: Cancer Epidemiology. Springer; 2009: 469–486.

3. Casey JA, Schwartz BS, Stewart WF, Adler NE: Using electronic health records for population health research: a review of methods and applications. *Annual review of public health* 2016, 37:61–81.
4. Liu H, Bielinski SJ, Sohn S, Murphy S, Waghlikar KB, Jonnalagadda SR, Ravikumar K, Wu ST, Kullo IJ, Chute CG: An information extraction framework for cohort identification using electronic health records. *AMIA Summits on Translational Science Proceedings* 2013, 2013:149.
5. LeCun Y, Bottou L, Bengio Y, Haffner P: Gradient-based learning applied to document recognition. *Proceedings of the IEEE* 1998, 86(11):2278–2324.
6. Mikolov T, Yih W-t, Zweig G: Linguistic regularities in continuous space word representations. In: *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 2013*; 2013: 746–751.
7. Wang Y, Liu S, Afzal N, Rastegar-Mojarad M, Wang L, Shen F, Liu H: A Comparison of Word Embeddings for the Biomedical Natural Language Processing. *arXiv preprint arXiv:180200400* 2018.
8. Soysal E, Warner JL, Denny JC, Xu H: Identifying Metastases-related Information from Pathology Reports of Lung Cancer Patients. *AMIA Summits on Translational Science Proceedings* 2017, 2017:268.

Figure 1.

Study design. EHR: Electronic Health Record, RS: related sentences, DL: deep learning