Artificial Intelligence in Oncology

Available online at: www.aioncology.org

Perspective

Mining Cancer-related Information in Electronic Healthcare Records with Natural Language Processing

Man Liu[a] *

National University of Singapore, Sinapore, 579599

ARTICLE INFO

Article history:
Received 05 May 2018
Accepted 05 May 2018

Keywords:
Artificial Intelligence, Healthcare Record, Data Mining

Perspectives

In 2018, around 1,735,350 new cases of cancer were estimated and 609,640 people will die from cancer in the United States in 2018. The number of new cases of cancer is 439.2 per 100,000 person per year. Specially, more than half of the world's new cancer incidents arise in Africa, Asia and Central and South America. And 70% of the people dying because of cancer are from these districts. With constant advancements in cancer research and healthcare provision, the overall cancer death rate has declined since the early 1990s in the United States. The cancer death rate has dropped 1.8% among men and 1.4% among women per year from 2004 to 2013. However, cancer still remains in the midst of leading causes of death followed by circulatory diseases such as heart disease and strokes .

A wealth of cancer-relevant information is conserved in a variety of types of healthcare records, for example, the electronic health record (EHR). Over the past 10 years, the adoption of EHR system has burgeoned, because of the Health Information Technology for Economic and Clinical Health (HITECH) Act of 2009 which brought 30 million dollars to encourage physician practices and hospitals to endorse EHR systems. Around 84% of hospitals have used at least a basic EHR system from the report of the National Coordinator of Health In-formation Technology office. EHR systems can be typically categorized into basic EHRs without clinic notes, basic EHRs with clinical notes and comprehensive systems [1]. Despite the shortage of advanced functionality of basic EHR systems, it can provide prolific information of patients' medical history, complications, and medication usage. While primarily de-signed for improving the healthcare efficiency from the operational applications (e.g. fortifying patient care by decreasing errors, increasing efficiency as well as improving coordination), EHRs were exploited for the alternative usage. For example, EHR systems store a wealth of patient information, including diagnoses, demographics, physical exams, sensor measurement, clinical notes, prescribed or administered medications etc. By mining the data in EHRs, it has the potential for various tasks such as patient trajectory modeling, medical concept extraction, disease inference, clinical decision support and so on [2].

However, part of the critical information is organized in the free narrative text which hampers machine to interpret the information underlying the text. The development of artificial intelligence provides a variety of solutions to this plight. For example, the technology of natural language processing (NLP) has emerged bridging the gap between free text and structured representation of cancer information [3]. In addition, it provides techniques for capturing cancer information in the free text. NLP can excavate the critical prognostic and predictive fac-tors which cannot be covered in the predefined categories in EHRs. For example, NLP is beneficial to process document collections for the purpose of information

* *Corresponding author.* Tel.: + 65 86879576;
E-mail address: lium@i2r.a-star.edu.sg

retrieval which can filter the relevant document, text classification which can utilize the content in the text and predict the appropriate tags for each document, information extraction which can select specific facts of predefined types of entities and relationships of interest, named entity recognition which can extract the pre-specified types of named entities, etc.

Recently, several researchers have published their work on unearthing cancer-related information in EHRs based on the NLP technology. Warner and Neuss [4] attempted to extract lung cancer stage information from narrative electronic healthcare record data. They developed a computerized algorithm based on NLP involving automated extracted rules to select the most likely stage when discordance was detected in EHRs. Their system achieved a correlation Cohen's κ = 0.996 compared with the tumor registry. For further analysis, they con-structed a network diagram to interpret the discordance causing the potential stage ambiguity. The node of the network represented different cancer stages and each edge between nodes was the particular co-occurrence observed across all patients. From this network, they reached a conclusion that the most common occurrence was stage IV and metastatic. Joshua and Neesha [5] reported their work on identification of colorectal cancer (CRC) tests in EHRs through NLP. This job was to find out whether patients had received appropriate screening of CRC. Four colorectal cancer tests including colonoscopy, flexible sigmoidoscopy, fecal occult blood test, and double contrast barium enema were identified. For identification of all CRC tests, this system achieved recall 93% and precision 94% and for finding out patients with a demand for screening, recall is 95% and precision is 88%. These are typical examples which reveal that NLP technique is important in the extraction of EHRs information. Besides the above cases, NLP has the application to discover additional imaging recommendation practice in radiology reports [6]. And NLP has also been applied to build a predictive model for further cancer clinical strategy making (i.e. risk of pancreatic cancer [7], survival in metastatic prostate cancer [8], prediction of invasive cancer across different age cohorts [9]).

Apart from the traditional NLP methods, the development of deep learning helps EHRs mining go further. Compared with the traditional NLP techniques, deep learning yields better performance and requires less labor-consuming feature engineering and preprocessing. Deep learning has been successfully exploited for applications in various tasks, such as temporal event extraction [10], relation extraction [11], abbreviation expansion [12], patient representation [13], static outcome prediction (e.g. prediction of the heart failure [14]), temporal out-come prediction (e.g. prediction of the future diagnoses [15-17]). In general, benefiting from the development of technologies, more potential application of EHRs on cancer will be dis-covered. However, challenges are also obvious. For example, EHR data is really heterogeneous which will lead to problems of data representations and the gap between cancer clinical knowledge and techniques will impede the development of real applications. Despite all of the challenges, we make a plea for our profession to make the area of cancer information mining in EHRs an ongoing area of future research.

References

1. Henry J, Pylypchuk Y, Searcy T, Patel V. (2016). Adoption of electronic health record systems among US Non-Federal Acute Care Hospitals: 2008-2015 (ONC Data Brief No. 35). Office of the National Coordinator for Health Information Technology, Washington, DC. doi:
2. Shickel B, Tighe PJ, Bihorac A, Rashidi P. (2017). Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. IEEE Journal of Biomedical and Health Informatics. 1-1. doi: 10.1109/JBHI.2017.2767063.
3. Spasić I, Livsey J, Keane JA, Nenadić G. (2014). Text mining of cancer-related information: review of current status and future directions. International journal of medical informatics. 83: 605-623. doi:
4. Warner JL, Levy MA, Neuss MN, Warner JL, Levy MA, Neuss MN. (2015). ReCAP: Feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. Journal of oncology practice. 12: 157-158. doi:
5. Denny JC, Choma NN, Peterson JF, Miller RA, Bastarache L, Li M, Peterson NB. (2012). Natural language processing improves identification of colorectal cancer testing in the electronic medical record. Medical Decision Making. 32: 188-197. doi:
6. Mamlin BW, Heinze DT, McDonald CJ. (2003). Automated extraction and normalization of findings from cancer-related free-text radiology reports. AMIA Annual Symposium Proceedings: American Medical Informatics Association), pp. 420.
7. Zhao D, Weng C. (2011). Combining PubMed knowledge and EHR data to develop a weighted bayesian network for pancreatic cancer prediction. Journal of biomedical informatics. 44: 859-868. doi:
8. Heintzelman NH, Taylor RJ, Simonsen L, Lustig R, Anderko D, Haythornthwaite JA, Childs LC, Bova GS. (2012). Longitudinal analysis of pain in patients with metastatic prostate cancer using natural language processing of medical record text. Journal of the American Medical Informatics Association. 20: 898-905. doi:
9. Nassif H, Page D, Ayvaci M, Shavlik J, Burnside ES. (2010). Uncovering age-specific invasive and DCIS breast cancer rules using Inductive Logic Programming. Proceedings of the 1st ACM International Health Informatics Symposium: ACM), pp. 76-82.
10. Fries JA, Center M. (2016). Brundlefly at SemEval-2016 Task 12: Recurrent Neural Networks vs. Joint Inference for Clinical Temporal Information Extraction. Proceedings of SemEval. 1274-1279. doi:
11. Lv X, Guan Y, Yang J, Wu J. (2016). Clinical relation extraction with deep learning. International Journal of Hybrid Information Technology. 9: 237-248. doi:
12. Liu Y, Ge T, Mathews KS, Ji H, McGuinness DL. (2015). Exploiting Task-Oriented Resources to Learn Word Embeddings for Clinical Abbreviation Expansion. ACL-IJCNLP 2015. 92. doi:
13. Choi E, Bahadori MT, Schuetz A, Stewart WF, Sun J. (2016). Doctor ai: Predicting clinical events via recurrent neural networks. Machine Learning for Healthcare Conference, pp. 301-318.
14. Choi E, Schuetz A, Stewart W, Sun J. Medical Concept Representation Learning from Electronic Health Records and its Application on Heart Failure Prediction. doi:
15. Lipton ZC, Kale DC, Elkan C, Wetzel R. LEARNING TO DIAGNOSE WITH LSTM RECURRENT NEURAL NETWORKS. doi:
16. Nickerson P, Tighe P, Shickel B, Rashidi P. (2016). Deep neural network architectures for forecasting analgesic response. Engineering in Medicine and Biology Society (EMBC), 2016 IEEE 38th Annual International Conference of the: IEEE), pp. 2966-2969.
17. Esteban C, Staeck O, Baier S, Yang Y, Tresp V. (2016). Predicting clinical events by combining static and dynamic information using recurrent neural networks. Healthcare Informatics (ICHI), 2016 IEEE International Conference on: IEEE), pp. 93-101.