Identifying Bladder Cancer Stage And Use Of Chemotherapy In The Electronic Medical Record: How Reliable Is Natural Language Processing?

Nirupama Devanathan, David Haggstrom, Clint Cary

Division of General Internal Medicine & Geriatrics, Indiana University School of Medicine; Center for Health Services and Outcomes Research, Regenstrief Institute; VA HSR&D Center for Health Information and Communication Indiana, Indiana University Department of Urology

Background:
Large automated electronic medical record (EMR) databases, together with natural language processing (NLP) algorithms, have the potential to be valuable tools in studying the patterns and effectiveness of treatment. Therefore, the current study sought to develop novel tools to identify bladder cancer cases, their clinical stage, and the chemotherapy they receive in electronic medical records.

Methods:
EMR data were obtained from Indiana University Health hospitals from 2008 to 2015. We developed 2 novel algorithms using natural language processing (NLP) on unstructured data to identify (a) bladder cancer cases and clinical stage, and (b) chemotherapy names and line of chemotherapy. The sensitivity, specificity, PPV, and NPV for the clinical staging and treatment algorithm were calculated against the gold standard of manual chart review

Results:
A total of 2,559 unique bladder cancer patients were identified and stratified using the clinical staging algorithm, defined as metastatic, muscle invasive, or non-muscle invasive. We identified 657 metastatic cases, 567 muscle invasive cases, and 604 non-muscle invasive cases. Further, we calculated the PPV for metastatic cases as 69.9%, muscle invasive as 80.4%, and non-muscle invasive as 79.1%. Next, the treatment algorithm was applied to metastatic patients to identify the type of chemotherapy received and 1st or 2nd line of therapy. The PPV for identifying the 1st and 2nd lines were 70.5% and 55.6%, respectively. The PPV for gemcitabine/carboplatin or cisplatin was 57.5%, but for methotrexate, vinblastine, doxorubicin, cisplatin, was 37.5%.

Conclusion and Potential Impact:
The performance of the algorithm demonstrates the potential for NLP to identify cancer cases, stage, and presence of treatment. While providing meaningful information, the accuracy of the approach suggests that a hybrid method using both NLP algorithms and manual chart review remains the most robust approach. The low performance of the algorithm to identify line of therapy further highlights the need for further NLP development in this area and emphasizes the ongoing need for either human entry or review of structured data.