Develop a Corpus for Immunotherapy-related Adverse Events from Clinical Notes in the Electronic Medical Records

Md Muntasir Zitu\textsuperscript{1}, Shijun Zhang, MS\textsuperscript{1}, Dwight Owen, MD\textsuperscript{2}, ChienWei Chiang, PhD\textsuperscript{1}, Lang Li, PhD\textsuperscript{1}

\textsuperscript{1}Department of Biomedical Informatics, College of Medicine, The Ohio State University, Columbus, Ohio, USA; \textsuperscript{2}Department of Internal Medicine, College of Medicine, The Ohio State University, Columbus, Ohio, USA

Introduction
Immunotherapy is an effective treatment for cancer patients; however, it is associated with significant adverse drug events (ADEs). Unstructured Electronic health records (EHRs) provide useful information to detect ADEs related to immunotherapy. Before developing and evaluating text mining methods, it is critical to construct corpus that have well annotated AEs related to immunotherapies.

Methodology
We collected total 195,401 clinical notes from 1097 individual patients treated with immunotherapy in The Ohio State University James Cancer Hospital. The first step was to identify the notes having immunotherapy drugs and ADE symptoms. We considered the drugs, Pembrolizumab, Nivolumab, Atezolizumab, Ipilimumab and Tremelimumab. We used the vocabularies based on Common Terminology Criteria for Adverse Events (CTCAE) version 5.0 for some of the common immune-related AEs, such as colitis, pneumonitis, rash, myalgia, hepatitis and thyroid abnormalities. After the screening, there were 35,331 notes having drug names and ADE symptoms among 1084 immunotherapy patients. We selected 600 notes from 17 patients. The annotations were further performed using the brat rapid annotation tool\textsuperscript{2}. Before annotation, all the selected notes were segmented into sentences using cTAKES\textsuperscript{3}. The immunotherapy drugs, ADE symptoms and dates mentioned in the clinical notes were tagged automatically. The other entities, such as grades and causal terms, were annotated manually. We further manually annotate the relations between entities, including drug-date, ADE-date, ADE-grade, drug-causal term, ADE-causal term, and drug-ADE. We annotated relations in the same sentence as well as in different sentences. We annotate both positive and negative relations between drugs and ADEs.

Results
In our corpus, there were totally 3822 drugs, 8106 ADE symptoms, 282 causal terms, 1331 grades and 30,716 dates found in those 600 notes. We were also able to extract 381 single sentence positive drug-ADE relations, 118 cross sentences positive drug-ADE relations, 42 single sentence negative drug-ADE relations, 7 cross sentences negative drug-ADE relations, 551 single sentence causal term-drug/ADE relations, 65 cross sentences causal term-drug/ADE, 343 grade-ADE relations, 1023 drug-date relations and 1321 ADE-date relations.

Conclusion
In this study, we develop a corpus for the immunotherapy induced ADEs reported in the clinical notes. After manual annotations, we recognize that drug-ADE relations were reported in the notes in a highly diverse format.

References