Cancer Treatment Classification with Electronic Medical Health Records (Student Abstract)

Jiaming Zeng, Imon Banerjee, Michael Gensheimer, Daniel Rubin
1Department of Management Science and Engineering, Stanford University
2Department of Computer Science, Emory University
3Department of Biomedical Data Science, Stanford University
4Department of Radiation Oncology, Stanford University

Abstract
We built a natural language processing (NLP) language model that can be used to extract cancer treatment information using structured and unstructured electronic medical records (EMR). Our work appears to be the first that combines EMR and NLP for treatment identification.

Introduction
Knowing the sequence of treatments administered to a cancer patient is important for personalized medicine and sequential treatment planning. Our final goal is to leverage the full EMR, including the information available in the clinical notes, to build causal models for treatment effectiveness. For that purpose, we need a sufficiently large dataset with labeled treatment information. However, cancer registries only record the initial line of treatment, even that requires hours of expensive manual labour.

We aim to build a NLP language model that can extract longitudinal treatment information using a combination of structured and unstructured EMR data. Starting with a model trained only on the initial line of treatment, we explore if we can extend the model to extract unrecorded or longitudinal treatment information. The extracted sequential treatments can then be used for future analysis and treatment planning.

Related Works
Automated clinical text classification is an area of growing interest in NLP applications (Wang et al. 2019). Existing literature has shown success in classifying cancer stage information (Warner et al. 2015), disease characteristics (Zhu et al. 2012), and medical subdomains (Weng et al. 2017). Our work appears to be the first work that utilizes NLP for cancer treatment classification from EMRs.

Dataset
We built our dataset from the Stanford Cancer Institute Research Database (SCIRDB), where we selected for patients with localized prostate, esophageal, and oropharynx cancer and diagnosis date after 2008. From SCIRDB, we pulled a total of 4,420 patients with 483,782 clinical notes among the three cancer types. From the California Cancer Registry (CCR), we pulled the initial treatment information available for these selected patients.

From the 4,420 patients, we filtered for patients with at least one recorded treatment in the CCR and at least one associated note/encounter in SCIRDB. We classify the initial line of treatment as all treatments performed within 6 months of the first administered treatment and performed classification for the 6 most common treatments found in the CCR. We randomly reserved 10% of the patients from the compiled dataset for testing. Notes associated with these patients were not used for training the embedding models.

After filtering, we compiled a dataset of 2,389 patients to train a treatment classification model. The treatment types selected for classification prediction are summarized in Table 1. For surgery, only patients with surgery at the primary site were selected. For the clinical notes associated with each treatment entry, we compiled the notes into one document by filtering for sentences that contained at least one treatment term. The treatment term dictionary was built from clinical terms in NCI metathesaurus vocabulary and medications curated by experts familiar with our treatments.

In addition to the clinical notes, we also extracted structured data from the information available in the SCIRDB. The features extracted are summarized in Table 2.

Methodology
We trained a machine learning model for treatment classification using combinations of structure and unstructured data and compared the performance. The methodology can be described in two major parts: 1) build a NLP language model using the entire vocabulary, 2) train a machine learn-
Table 2: Summary of the structured data.

Table 3: Testing accuracy from logistic regression.

Future Work

Our preliminary exploration for a treatment classification model shows promising results. In order to create a large treatment dataset, we plan to boost the model performance by evaluating non-linear ML methods, e.g. elastic net, random forest, and SGD boosting.

We plan to do more experimentation with structured and unstructured data separately, and explore more methods for combining. For the structured data, we will expand the features, e.g. ICD9, CPT codes. For the unstructured data, we will explore different ways of collapsing the documents. We also plan to implement a simple bag-of-words classifier as a baseline. For combining the structured and unstructured data, we will also explore other methods such as 1D convolution and other ensemble methods.

Furthermore, our current model aims to produce one model for prostate, oropharynx, and esophagus. We will build separate language models for each cancer type and explore model transfer effects.

References

Le, Q., and Mikolov, T. 2014. Distributed representations of sentences and documents. In International conference on machine learning, 1188–1196.

Wang, Y.; Sohn, S.; Liu, S.; Shen, F.; Wang, L.; Atkinson, E. J.; Amin, S.; and Liu, H. 2019. A clinical text classification paradigm using weak supervision and deep representation. BMC medical informatics and decision making 19(1):1.

Warner, J. L.; Levy, M. A.; Neuss, M. N.; Warner, J. L.; Levy, M. A.; and Neuss, M. N. 2015. Recap: feasibility and accuracy of extracting cancer stage information from narrative electronic health record data. Journal of oncology practice 12(2):157–158.

Weng, W.-H.; Waghholikar, K. B.; McCray, A. T.; Szolovits, P.; and Chueh, H. C. 2017. Medical subdomain classification of clinical notes using a machine learning-based natural language processing approach. BMC medical informatics and decision making 17(1):155.

Zhu, H.; Ni, Y.; Cai, P.; Qiu, Z.; and Cao, F. 2012. Automatic extracting of patient-related attributes: disease, age, gender and race. Studies in health technology and informatics 180:589–593.