2023 Multidisciplinary Research Program in Medicine Project: Finding breast cancer relapses in pathology reports using natural language processing

Hypothesis or Research Question(s): Metastatic disease is detected in about 12% of patients with breast cancer. We hypothesise that an NLP model can identify metastatic disease in pathology reports with 95% sensitivity and 90% specificity, so <5% of relapses are missed and <20% of cases have positive predictions that require review by human coders.

PROJECT BACKGROUND & SUMMARY

Through the Provincial Health Services Authority, BC Cancer and the Data Science Institute at the University of British Columbia have formed a research partnership to develop NLP tools for extracting staging information from the electronic medical records of breast cancer patients.

We have successfully developed and evaluated a customizable automated data extraction NLP pipeline for breast cancer outcomes data in a small cohort of patients as a pilot project. Using a combination of rule-based (pattern-matching, auto-correction) and neural-based (pre-trained biomedical word embeddings) methods, our NLP algorithm is robust, transparent, and adaptable. The NLP pipeline achieved near-human-level accuracy across 49 targeted outcomes relevant to breast cancer research using operative and pathology reports. For structured data in breast cancer patient reports, we have developed a fully customized NLP pipeline to automate extraction of clinically relevant outcomes based on 2607 data points. The NLP pipeline achieves an overall 91.9% accuracy for operative reports and 95.4% accuracy for pathology reports.

For breast cancer survivors, cancer relapse is an important long-term concern. Cancer relapse can happen years or decades after completing treatment. The BC Cancer Breast Cancer Outcomes Unit was formed in 1994 to collect outcomes data for retrospective breast cancer research. Over the years, its membership has produced over 200 reviewed publications, many of which were co-authored with UBC medical students. Relapse data is collected from many sources, such as imaging reports, surgical procedures, pathology reports, chemotherapy records and radiotherapy records.

Provincial cancer registries do not have adequate resources to follow patients over the course of their lifetime to detect if and when they develop a cancer relapse. Outside of British Columbia, the current understanding of breast cancer relapse is based on fewer than 5% of patients who have been enrolled in clinical trials. However, if NLP tools could detect breast cancer relapses from electronic health records, relapse data collection could become routine in all provinces, aiding with quality improvement of care delivery and resource planning for care of breast cancer survivors who have experienced relapse.

Breast cancer relapse necessitates additional treatment and often presages death by years. NLP uses artificial intelligence methodology to extract discrete data elements from free-text electronic medical records. Tracking of treatment outcomes and predicting therapy needs will be enhanced if an NLP model can be developed to extract relapse data from multiple data sources. A suitable project for an 8-week Summer Student Research Program would be to focus on one data source. Most patients with breast cancer relapses have a biopsy or cytology to diagnose their relapse. For this project, we plan to develop NLP models for extracting information about local, regional and distant relapses from pathology.

BENEFIT TO THE STUDENTS

TS will learn to critically evaluate the literature about the use of NLP in oncology and interact with the programmers in the UBC Data Science Institute (DSI) to learn about NLP. TS will be required to complete
2023 Multidisciplinary Research Program in Medicine Project: Finding breast cancer relapses in pathology reports using natural language processing

their TCPS2 (https://tcps2core.ca/welcome) training and they will be able to observe ethical research practices at BC Cancer and the DSI. In addition, they will have the opportunity to complete Good Clinical Practice and Responsible Conduct of Research Courses for Investigators through the CITI training program (citiprogram.org). By focusing on pathology reports, TS will interrogate a rich data source with a manageable number of reports for the 8-week duration of their SSRP project. They will have the opportunity to learn about pathology by interacting with the specialists who dictate the reports, the oncologists who interpret the reports, and the health record analysts who code information from the reports. They will also learn about artificial intelligence methodology from computer scientists and statisticians. TS will be expected to write and submit a research abstract to a conference. If they are keen, they could also be involved in writing a manuscript, but this would need to be completed after the end of their SSRP term. Members of the DSI will train the student to use report annotation software to collect the data for building NLP models. TS will code about 1000 pathology reports from patients with breast cancer, identifying phrases associated with local, regional and distant relapses. TS will learn about data dictionaries and research data management. The pathology reports will be divided into training, validation and test sets. NLP models will be trained using a collaborative process involving computer scientists, clinicians and TS. TS's familiarity with the text of the reports will be vital for designing the queries that are used to probe reports for relapse data.

The process of creating NLP models is interactive between programmers and clinical content experts. During model building, TS will gain familiarity with statistical methods for assessing model quality, such as receiver operating curves. It is expected that TS will need to troubleshoot the models by iteratively generating hypotheses and testing them by adjusting multiple programming parameters in conjunction with the DSI team. TS will have the opportunity to attend the weekly virtual research meetings of the Breast Cancer Outcomes Unit and the DSI.