Cancer BERT Network (caBERTnet): A Question-and-Answer System to Extract Data from Free-Text Pathology Reports

Moffitt researchers developed a BERT (Bidirectional Encoder Representations and Transformers)-based deep learning natural language processing (NLP) algorithm to automatically extract detailed tumor site and histology information from free text pathology reports. caBERTnet's (pronounced "Cabernet") accuracies for predicting group-level site and histology codes were 93.5% and 97.7%, respectively. The top-5 accuracies for predicting fine-grained ICD-O-3 site and histology codes with 5 or more samples each in the training dataset were 93.6% and 95.4%, respectively. This is the first time an NLP system has achieved expert-level performance predicting ICD-O-3 codes across a broad range of tumor sites and histologies. This new system could help reduce treatment delays, increase enrollment in clinical trials of new therapies, and improve patient outcomes.

COMMERCIAL OPPORTUNITY

Certified tumor registrars (CTRs) undergo an extensive training and internship program to become proficient at extracting quantitative and categorical data from unstructured pathology reports. They are widely employed by cancer centers and other organizations to extract data for clinical and research applications and for reporting to state and national agencies. The Moffitt Cancer Registry deploys state-of-the art quality assurance procedures: its benchmark for quality is 90% and its target accuracy is 95%. caBERTnet may help simplify and accelerate CTR workflows. For example, caBERTnet could preprocess pathology reports to identify the top-5 site and histology ICD-O-3 codes and their corresponding phrases. The phrases could then be highlighted within the report body. Two pull-down menus could be pre-populated with top-5 code predictions: one for site and the other for histology. The CTR could then quickly choose a code from either pull-down menu. If the correct code was not among the top-5, then the CTR would resort to their current workflow—entering this information by hand.

TECHNOLOGY

We pursued three specific aims: 1) extract accurate tumor site and histology descriptions from free-text pathology reports; 2) accommodate the diverse terminology used to indicate the same pathology; and 3) provide accurate standardized tumor site and histology codes for use by downstream applications. We first trained a base language-model to comprehend the technical language in pathology reports. This involved unsupervised learning on a training corpus of 275,605 electronic pathology reports from 164,531 unique patients that included 121 million words. Next, we trained a Q&A "head" that would connect to, and work with, the pathology language model to answer pathology questions. Our Q&A system was designed to search for the answers to two predefined questions in each pathology report: 1) "What organ contains the tumor?"; and 2) "What is the kind of tumor or carcinoma?" This involved supervised training on 8,197 pathology reports, each with ground truth answers to these two questions determined by Certified Tumor Registrars. The dataset included 214 tumor sites and 193 histologies. The tumor site and histology phrases extracted by the Q&A model were used to predict ICD-O-3 site and histology codes. This involved fine-tuning two additional BERT models: one to predict site codes, and the second to predict histology codes. Our final system includes a network of 3 BERT-based models. We call this caBERTnet. We evaluated caBERnet using a sequestered test dataset of 2,050 pathology reports with ground truth answers determined by Certified Tumor Registrars.

PUBLICATION/PATENT

- To be published in the Journal of Medicinal Chemistry 2021
- Provisional patent application to be filed for Dr. Mitchell

CONTACT

Haskell Adler PhD MBA
Senior Licensing Manager
Haskell.Adler@Moffitt.org
(813) 745-6596

LICENSING OPPORTUNITY

MOFFITT
CANCER CENTER