OncoNote: Enabling the Discovery of a Novel Digital Biomarker for Early Pan-Cancer Survival and Recurrence Prediction via Free-Text Clinical Narratives

Chandrasekhar, Vrishank (School: Lynbrook High School)

Cancer is the second leading cause of death worldwide, affecting 20 million people and being responsible for >9.7 million deaths annually. Early prediction of clinical outcomes is crucial for patient risk stratification and treatment management, yet current staging systems fail to stratify across cancers, with concordances as low as 52%. Thus, my project presents OncoNote: the first low-cost, non-invasive computational framework to predict the survival and recurrence rates of pan-cancer patients at initial diagnosis using unstructured free-text clinical narratives. First, I train an unsupervised FastText model to process the ~2.3 million clinical narratives from 56,339 patients by removing noise, segmenting sentences, and synthesizing clinical terminology. By learning robust, medically relevant word representations across 686 institutions, the model generates a dataset of document-level embeddings, which is used to train a Cox Proportional Hazards neural network. The model adaptively evaluates prognostically relevant clinical text and predicts a risk score for each patient. After extensive experimentation, the survival and recurrence prediction models achieved impressive concordances of 0.72 and 0.77 across 32 cancer types, surpassing all existing clinical techniques. OncoNote offers clinicians the first pan-cancer, text-based digital biomarker that can be generalized to any electronic health record and will lead to better clinical outcomes by sparing patients from overaggressive therapies. Ultimately, this platform is a scalable, affordable, and comprehensive solution that will enable clinicians to determine robust, targeted treatment plans earlier, prevent adverse clinical events from occurring in the first place, and save millions of lives.

Awards Won: