Brain Cancer Survival

Erdos Institute Data Science Boot Camp May 2024

Executive Summary

Project team: Daniel Cerkoney, Anthony Young

Overview

Cancer is the second leading cause of death in the US, and brain cancer in particular is one of the most difficult varieties to treat. In order to aid medical professionals in the treatment of cancer, the Surveillance, Epidemiology, and End Results (SEER) Program has compiled data about cancer patients since 1975 from several registries. The SEER dataset covers 42% of the US population and tracks a very large number of covariates. To understand the prognoses of patients, we perform a survival analysis on all brain tumor cases using a Cox proportional hazards regression model and Kaplan-Meier estimators.

Stakeholders

- Doctors
- Cancer patients
- Pharmaceutical companies
- Health insurance companies

Data

For our dataset, we considered the SEER 17 database (November 2022 submission), which includes US cancer registry data documented between 2000 and 2020. The dataset included over 9 million tumor records in total, from which we extracted case information and survival data for all 76,327 brain tumor entries. We then restricted the study cohort to only those patients with a single tumor, leaving 74,332 cases in total.

Methodology

Data cleaning and pre-processing and the train-test split were performed using scikit-learn, while the lifelines library was used to fit the survival models. We first selected a relevant feature subset, and then pre-processed the remaining data by implementing standard scaling & imputing, one-hot encoding of categoricals, and removal of collinear features. This resulted in 111 features after preprocessing. We then stratified the dataset by age based on established categorical age standards for survival analysis. We fit a Cox proportional hazards model with added LASSO and ridge regression penalty terms and compared this to baseline results for the dataset obtained via the Kaplan-Meier estimator, which does not adjust survival for multiple covariates. Goodness-of-fit for the final model was assessed via the concordance index (c-index).

Results

The hazard ratios predicted by our model reinforce many well-known aspects of cancer treatment, including the poor outcomes of glioblastoma cases and the Chromosome 1q/19q co-deletion associated with improved survival of oligodendroglioma patients. We show improvements in predicted survival curves using Cox regression over the Kaplan-Meier estimator.

Future Work

In the future, we would like to extend this investigation in a few key ways:

1. Quantify the validity of the proportional hazards assumption for our dataset. If important features are found to violate this assumption, we could further stratify on these features, or add interaction terms to the model for the relevant features.
2. Further analyze covariates relevant for improvement of survival outcomes for patients with severe diagnoses (e.g., glioblastoma).
3. Compare our results from Cox regression to one or more of the following:
    a. Random survival forests
    b. Survival support vector machines
    c. Deep learning methods for survival analysis