Data 601 Project Proposal

Romith Bondada, Gautham Nagaraj Chandra Shekariah, Lamis

January 2025

1 Introduction

Cancer remains one of the most pressing global health challenges, with early detection and effective treatment being critical to improving patient outcomes and survival rates. This project investigates patterns and trends in cancer patient data to identify key factors influencing survival, tumor progression, and treatment responses. By employing advanced data visualization techniques, we aim to uncover insights that can inform clinical decision-making and contribute to the broader field of medical data analysis. Our study leverages a comprehensive cancer dataset that includes demographic, clinical, and pathological information. This data set offers an opportunity to explore complex relationships among variables such as tumor characteristics, survival outcomes, and hormone receptor statuses. Through a systematic analysis, we seek to gain insights from the data.

2 Guiding Questions

- What are demographic factors (e.g. age, race, and marital status) associated with higher survival rates?
- How do tumor characteristics such as size, grade, and hormone receptor status correlate with survival outcomes?

The guiding questions will help provide a blueprint as to how the data will be wrangled/visualized and see if what factors affect the survival of the patient the most.

3 Dataset

The data set of breast cancer patients [1] was obtained from the 2017 November update of the NCI SEER Program, which provides information on population-based cancer statistics. The data set involved women with breast cancer diagnosed in 2006-2010. Patients with unknown tumor size examined regional lymph nodes, regional positive lymph nodes, and patients whose survival months were less than 1 month were excluded; thus, 4024 patients were ultimately included.

4 Tasks

Key tasks include removing duplicates, standardizing categorical variables, creating an age group column by binning, and filtering data. Visualizations include histograms for age distribution, bar charts for demographic survival comparisons, scatter plots for tumor size vs. survival outcomes, heatmaps for correlations, and boxplots for variability in survival across groups. The team will use Pandas, Matplotlib, Seaborn, and Plotly to clean data, identify trends, create visualizations, and compile a final report.

- Data cleaning, age group transformations, and filtering data will be done by Gautham Nagaraj Chandra Shekariah - Handle missing values, remove duplicates, and standardize categorical variables. - Create an age-group column by dividing ages into ranges. - Focus on breast cancer cases and filter for relevant demographic and tumor data.
- The demographic data aggregation and visualizations will be done by Lamis. Calculate average survival rates by age, race, and marital status. Create histograms, bar charts, and box plots for demographic factors.
- Correlation analysis and visualization of tumor characteristics will be performed by Romith Bondada

References

[1] JING TENG. Seer breast cancer data, 2019.