The determination of biological trends in Bladder Urothelial Carcinoma

Anjali Menon, Hayden Scott, Rhea Kaul

Introduction

Bladder Urothelial Carcinoma is one of the most common cancer types, representing the most common bladder cancer histotype (1), and further study of its trends would allow for better treatment. This study aims to isolate subgroups using genetic mutation data from the TCGA PanCancer Atlas and perform survival analysis on each subgroup to determine if diagnosis age plays an important role. This will be done using data pre-processing and clustering. Expected results are descriptions of a small or zero number of subgroups, given that our data is already within a single histotype. This study hopes to aid in assessing survival chances and help better inform treatment plans.

Project Goal

Research Question: The goal of this study is to determine the biological trend between mutations in the TCGA, PanCancer Atlas data set and the average age of patients diagnosed with Bladder Urothelial Carcinoma (UCC).

The correlation of biological mutations and average patient age is important in order to understand whether the risk of developing cancer increases after a certain age. This can be particularly useful for clinicians as they can use genetic testing to determine the risk that an individual may have of developing cancer at the later stages of life. Furthermore, as Canada's large baby-boomer generation retires, knowing the average age that UCC may occur can help prepare the medical industry with the necessary resources to help treat or avoid the progression of this cancer.

Analysis Plan

Step 1: Data Visualization

1. Use the aggregate function to condense the dataset of extended mutations
    a. Allows for the splitting of datasets similar to SQL
2. Determine interesting trends in the dataset through plots

Step 2: Data processing and preparation

1. Combine the genetic mutation data with patient clinical data into 1 dataset
2. Screen data to remove non-numeric data which the prcomp() function cannot handle
3. Remove rows and columns with incomplete data

Step 3: Clustering

1. Due to the categorical nature of the data, **k-modes** clustering will be used to determine possible subgroups

Step 5: Survival Analysis

1. Used to understand whether certain age groups are more likely to live longer than another after diagnosis of cancer and if diagnosis age influences survival in UCC patients
2. Creation of a data frame and creation of Kaplan-Meier plots using surv(), survfit, ggsurvplot functions

Step 6: Linear Regression

1. Used to analyze the relationship between mutations and diagnosis age
2. Determine the p-value and formulate graphs

Variables of Interest

Genetic Variables:

- Determined post step 1 (Data Visualisation) and PCA analysis
- These are variables that have had a significant impact on the data set

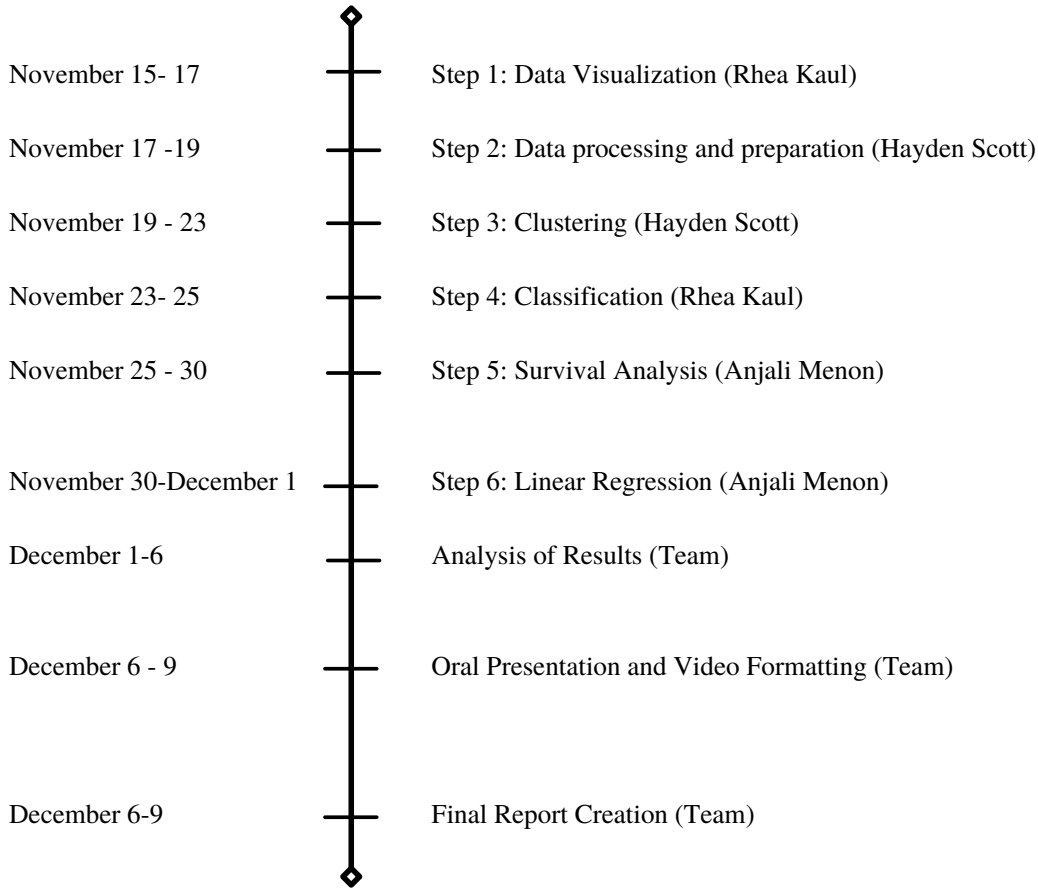
Patient Variables:

- Age, race, type of tumor, sex

Challenges

The initial challenge we faced was accessing and running the data files. The files were in an unfamiliar file format that was identified as an archive file. WinRaR was used to open the file and access the data. The next challenge we anticipate encountering is a part of our data analysis. Due to the large data set, it may be difficult to determine the variables of interest. Another limitation of the data is that the majority of the patients are male, therefore our findings may not be representative of the population as a whole. Similarly, the data does not represent all races, which will make it harder to produce conclusive results for the population as a whole. Due to use of a data set that is not very representative of the general population, the pre-processing of the data might remove the few varying factors and produce results that are only applicable for the biased majority.

Timeline and Duties



References

1. (2020, September). *Bladder Cancer: Introduction*. Cancer.net. https://www.cancer.net/cancer-types/bladder-cancer/introduction