Breast Cancer Gene Expression Profiles

Atharva Abhijit Kulkarni Hemashree Kilari Matthew Greene
Shashank Bettada Sathya Thirtha Vinu Baburaj

10/30/2022

Contents

Summary

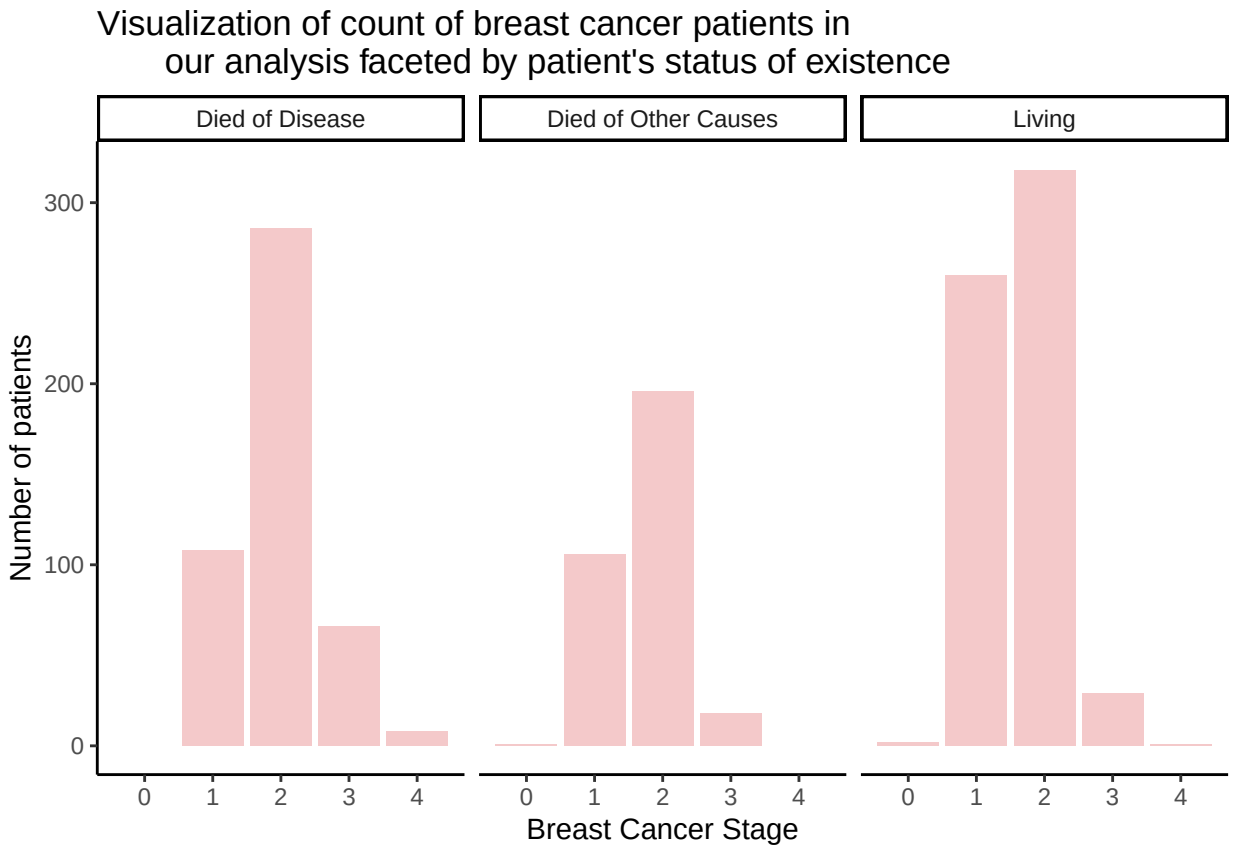
Breast cancer is the most prevalent cancer among middle aged and older women. The American Cancer Society estimates that 2.8 million women worldwide will be diagnosed with breast cancer. Our goal is to build a prediction model that predicts the chance of surviving breast cancer. The dataset for this analysis and prediction is called Breast Cancer Gene Expression Profiles (METABRIC) and is found on Kaggle.

Proposed Plan

Our goal for this project is to perform an in depth analysis of variables influencing the chance of overcoming breast cancer. We begin by doing a preliminary analysis of the available data. As there are a total of 693 variables in our data set, we only consider a few variables in the initial analysis that help us the screen the overall data set for some preliminary results. In the pre-processing stage, the nil values are eliminated and the data is streamlined for analysis that require exclusion of redundant values in categorical variables. Prior to EDA we have to understand in medical terms the numerous variables that are available to us for analysis. This knowledge can streamline the process of performing EDA and help in interpreting the results.EDA helps us in identifying key variables that have a strong correlation to higher and lower chances of survival respectively. Identifying these key variables, will help us in building various classification models which predict the chances of survival of a patient. We plan on building different classification models using Logistic Regression, K-Nearest Neighbors, Naive Bayes, Support Vector Machines (SVM) and Random Forests and compare their performances.

Preliminary Results

Firstly we need to understand how many patient's details are available to us in this data set and what their current status of existence is. In order to analyse this, a bar graph is plotted with the the count of patients in each stage of breast cancer faceted by their existence status.

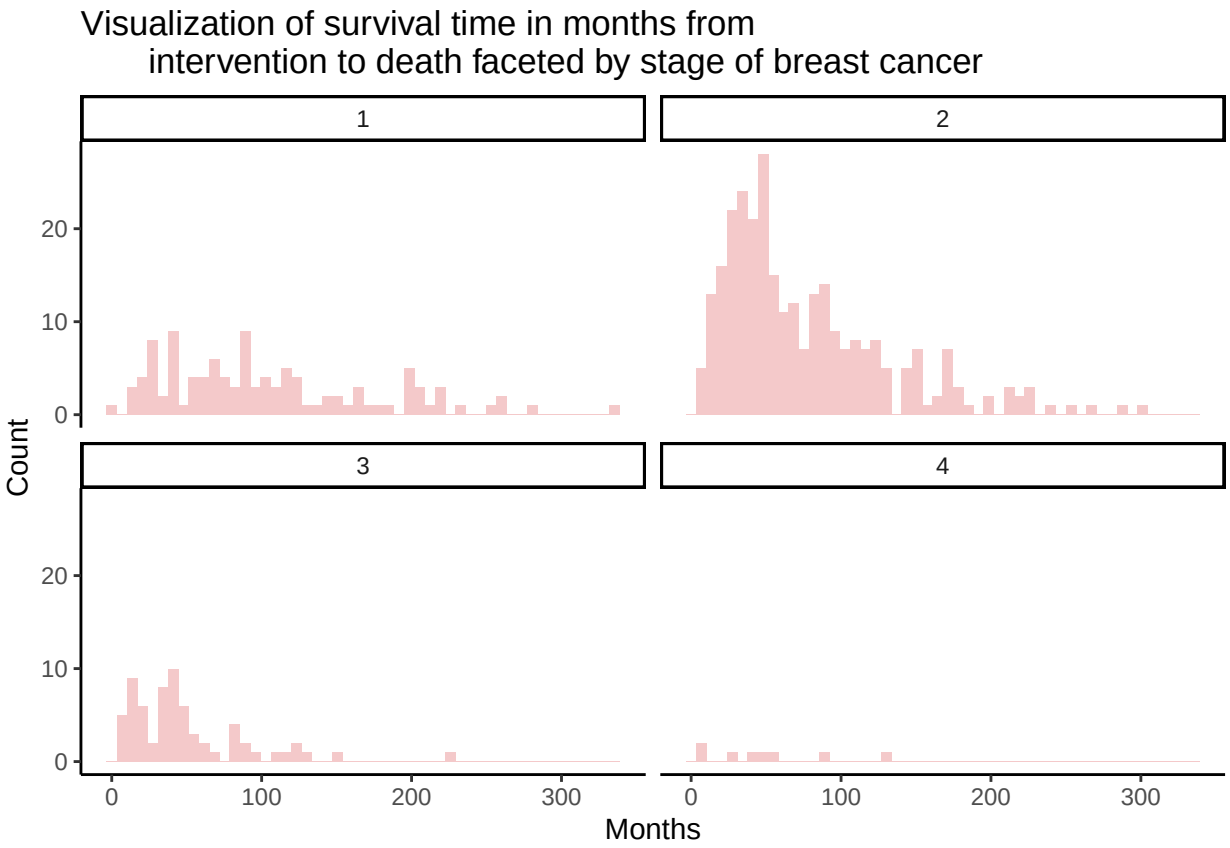


Observations:

a. Across all 3 facets, we have the most data on patients in the stage 2 of breast cancer.

b. We have the least data on the pre-cancerous stage 0 followed by stage 4.

Next in our preliminary analysis, we would like to see the duration of survival in deceased patients across different stages of the cancer. The survival period is calculated from the time of intervention to the time of death.



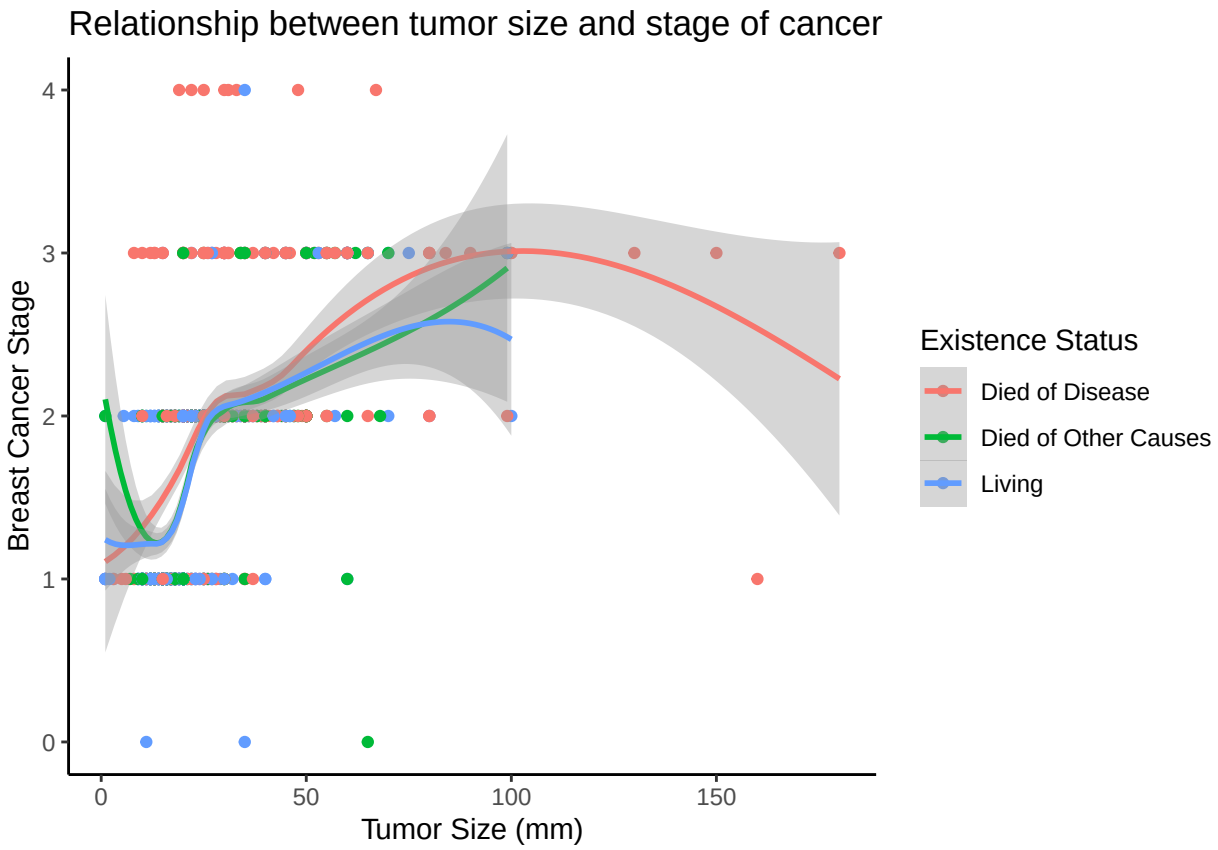
Observations:

a. We see that most of the patients in the visualization were in their second stage of cancer.

b. Across all 4 stages, the distribution seems to be right skewed.

c. Range of survival period is highest in the 1st stage of cancer and lowest in the 4th stage.

Finally we wanted to explore the relationship between the size of the tumor and stage of cancer. It's imperative to understand whether the correlation between size and stage can determine whether the patient can overcome the disease or not.



Observations:

a. We see that in patients who died of the disease, the stage of the cancer has greater influence on the survival rate than the tumor size.

b. From the above plot it is difficult to firmly establish the correlation between the size and stage and its influence on the survival rate. We wish to explore the impact of therapies (chemo, hormonal and radio) in the upcoming analysis to better understand what factors actually impact the chance of survival.

c. As we come across deaths from tumors of size less than 25 mm and also greater than 150 mm, in the preliminary analysis, we can say that the chance of survival seems to be tumor size agnostic.

References

1. Dataset link: (https://www.kaggle.com/datasets/raghadalharbi/breast-cancer-gene-expression-profiles-metabric)

2. Associations between genomic stratification of breast cancer and centrally reviewed tumour pathology in the METABRIC cohort - (https://www.nature.com/articles/s41523-018-0056-8)

3. Breast Cancer Stages - (https://www.cancer.org/cancer/breast-cancer/understanding-a-breast-cancer-diagnosis/stages-of-breast-cancer.html)