A Study To Evaluate Breast Cancer In Women

Vamshi Reddy Madem, Deepak Gugulla, Mrudula Nimmala.

1.Introduction:

Breast cancer is a common and potentially life-threatening disease characterized by abnormal cell growth in the breast tissue. While it predominantly affects women, it can also occur in men. Early detection through screenings like mammograms is crucial for successful treatment. Understanding the various risk factors, genetic predispositions, and lifestyle choices associated with breast cancer is essential for prevention and effective management

This study was conducted to examine the status of individuals who have previously experienced breast cancer. Our focus lies in assessing whether these individuals are currently alive without breast cancer recurrence or with breast cancer recurrence or death. The findings of the study might also have a big impact on future studies, public health policies and initiatives that attempt to lessen the impact of Breast Cancer in society.

2.Data Description:

The information came from Kaggle, a website for data science and machine learning enthusiasts to find and exchange datasets, kernels, and contests.The data set contains patient records from a 1984-1989 trial conducted by the German Breast Cancer Study Group (GBSG) of 720 patients with node positive breast cancer; it retains the 686 patients with complete data for the prognostic variables.These data set is generated and used by Royston and Altman.

Used Observations	686
Missing Values & Duplicates	NA
Training Data	492
Testing Data	164

Table i

Variables	***Type***	***Description***
meno	Categorical	Menopausal Status
size	Numeric	Tumor size(mm)
grade	Numeric	Tumor Grade(mm Hg)
nodes	Numeric to categorical	Number of positive lymph nodes
pgr	Categorical	Progesterone receptors(fmol/l)
er	Numeric	Estrogen receptors(fmol/l)
hormon	Categorical	Hormonal Therapy (0 = no, 1= yes)
rfstime	Numeric	Recurrence free survival time; Days to first of occurrence, 1 = recurrence or death
status	Categorical	0 = alive without recurrence , 1 = recurrence or death

Table ii

With one response variable Outcome “status” contains a value of “0” for patients alive without recurrence of breast cancer, Type 2 with “1” for the person with breast cancer recurrence or death.

It is important to note that the dataset contains no “NA” or missing or incomplete data points.

The summary statistics of the variables is provided in *Table* ii *i.*

Variables	*Min.*	*Max.*	*Median*	*Mean*
pid	1	1819	1015.1	966.1
age	21	80	53	53.05
meno	0	1	1	0.5773
size	3	120	25	29.33
grade	1	3	2	2.117
er	0	1144	36	96.25
hormon	0	1	0	0.386
rfstime	8	2659	1084	1124.5
status(0)	387			
status(1)	299			

Table ii *i*

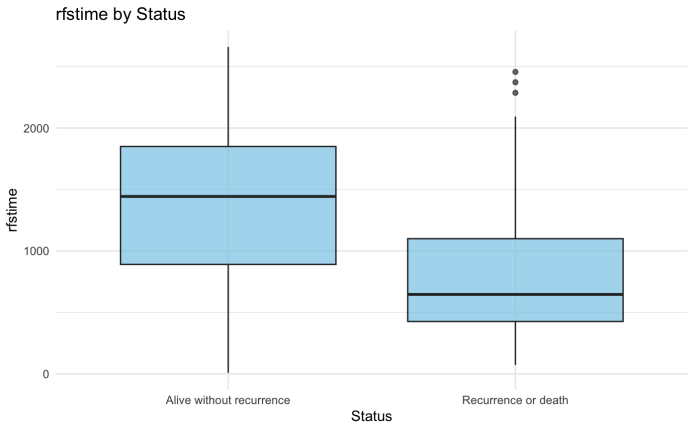
3. Methods:

3.1 Data Cleaning and Modification:

In the German breast cancer study group(gbsg) some meaningful categorization has been done. This categorization provides a more meaningful representation of harmon for analysis. Additionally, the Outcome variable, which originally had categorical values “0” for patients alive without recurrence of breast cancer, Type 2 with “1” for the person with breast cancer recurrence or death., has been transformed into more descriptive categories: "No" for individuals who are alive without recurrence and "Yes" for individuals with breast cancer recurrence or death. This change enhances the interpretability and clarity of the variable's meaning in the dataset.

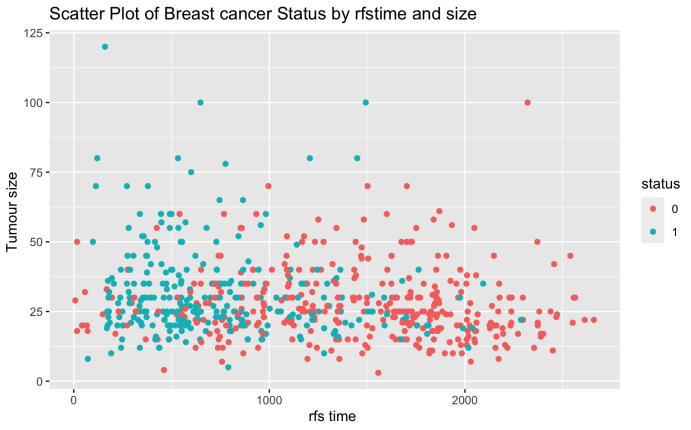
3.2 Graphical Analysis:

Based on the rfstime and outcome graph, it appears that individuals with more survival time have less chance of breast cancer recurrence or death when compared to individuals with low survival time(rfstime). See *Graph i.*



Graph i

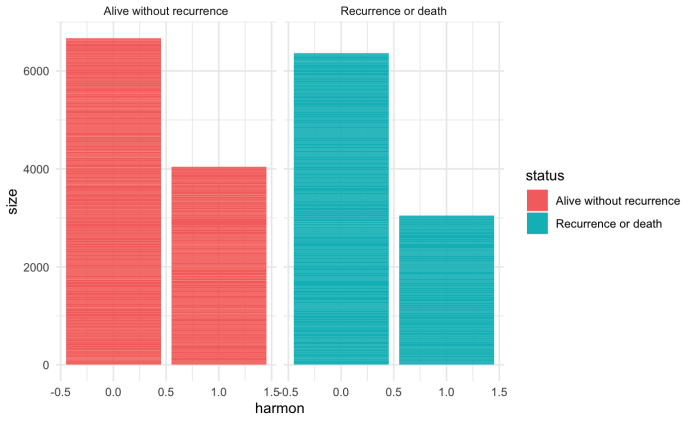
The graph depicting the relationship between rfstime and tumor size suggests that there is a higher likelihood of breast cancer recurrence or death when rfstime decreases and tumor size increases with exception of one or two outliers. See *Graph ii.*



Graph ii

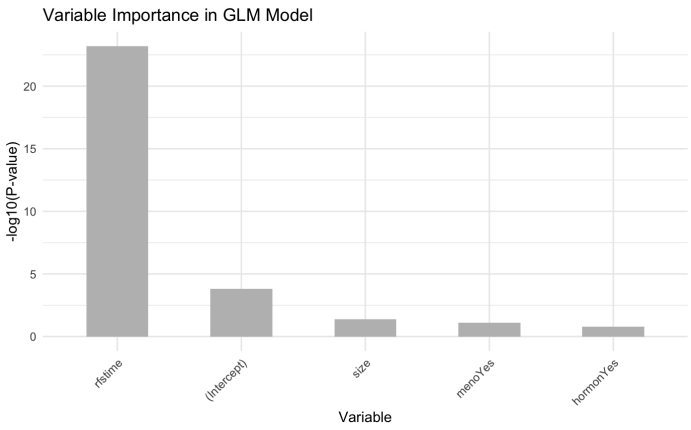
According to the bar plot of hormonal therapy, size and breast cancer status, we can depict that

people who have taken hormonal therapy have slightly less chances of getting breast cancer recurrence or death, Therefore, while the plot provides some insight into the association between hormonal therapy and size and response variable, it is not definite. See *Graph iii.*



Graph iii

The bar graph displays the significance of various variables in the final model. Rfs Time is the most significant variable, indicating that it has a strong predictive power for breast cancer status. Other variables such as meno, harmon and tumor grade size also show significance in predicting Breast cancer recurrence or death. See *Graph iv.*



Graph iv

3.3 Primary analysis:

Multiple logistic regression (MLR) is a statistical method used to analyze how a binary outcome variable is related to several predictor variables. In MLR, the outcome variable is binary, meaning it can have values of 0 or 1, or it can be categorical. Meanwhile, the predictor variables can take on different types, including continuous, categorical, or binary.

Multiple Logistic Regression (MLR) models the relationship between an outcome variable and predictor variables, calculating the probability of a binary outcome using a logit equation. It's widely used in fields like healthcare and business to predict and understand important factors influencing outcomes.

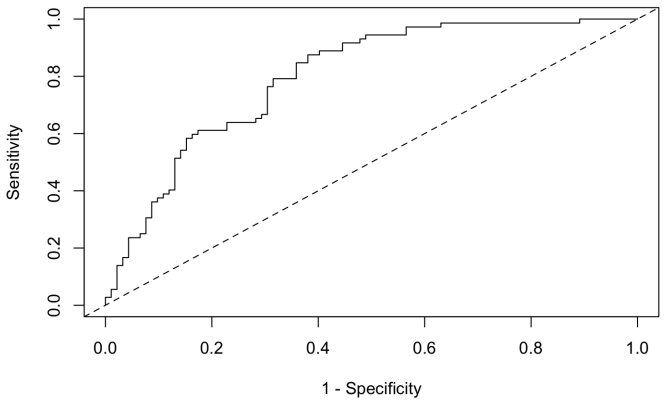
The initial model with eight predictors was insignificant for determining breast cancer status. Akaike Information Criteria (AIC) was used to select a model that balances complexity and fit, aiming to identify the most significant predictors while avoiding overfitting.

3.4 Secondary Analysis:

Initially, we used multiple logistic regression to identify the key predictors for status of breast cancer. However, we found that four of the predictor age ,grade, nodes, er were highly insignificant and cannot explain the status of breast cancer. Therefore, we performed a variable selection method called Akaike Information Criteria (AIC) to identify the best model for predicting status.

The AIC analysis helped us to determine the most significant predictors, including menopause, breast size, hormonal influences. There was no significant correlation between the variables. There was no necessary interaction in the variables.

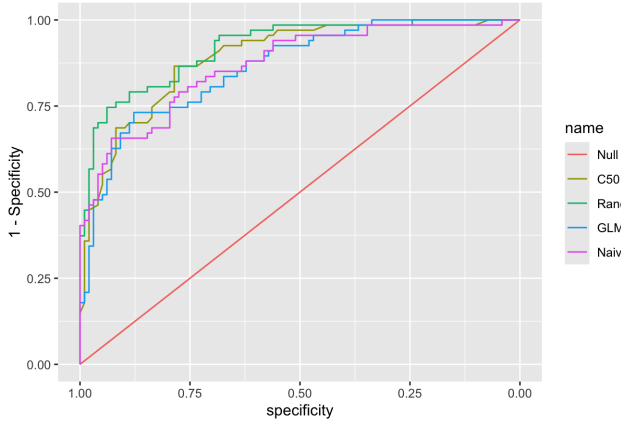
This allowed us to create a final and reliable model that can accurately predict the status of breast cancer, with an accuracy of 68%. The model has a pretty good ROC(Receiver Operating Characteristic) curve and an AUC value of 0.79 as seen in the graph. Our findings demonstrate the importance of carefully selecting and analyzing predictor variables to improve the accuracy of predictive models**.** See *Graph v*



Graph v

3.5 Other Models:

To improve accuracy, we tested models like Random Forests, C50, Naive Bayes, and a Null model using all eight predictors. Random Forests showed the highest accuracy, followed by C50, as indicated by the largest area under the curve. Exploring variable interactions may further refine model selection.



Graph vi

4. Results:

When evaluating binary classification models like logistic regression, important metrics include accuracy, sensitivity, specificity, AUC, and ROC. Accuracy tells us how often the model makes correct predictions, sensitivity measures how well it identifies true positives, and specificity gauges its ability to spot true negatives. AUC, which stands for area under the ROC curve, gives us an overall assessment of the model's predictive power for binary outcomes. Below, you'll find the model and its corresponding ROC curve.

The logit Equation of the model is $log(\frac{\hat{p}(x_1,x_2,x_3,x_4)}{1-\hat{p}(x_1,x_2,x_3,x_4)}) = 1.36 +$ 0.42*meno+0.016*size - 0.35*hormone -0.02*rfstime

The model showed an overall accuracy of 68%, sensitivity of 62%, specificity of 74%, and an area under the curve of 0.79.

5. Conclusion:

After looking at the study's data, it's clear that breast cancer is a big worry for women. The research, which used multiple logistic regression to study breast cancer, achieved a 68% accuracy rate and an AUC of 0.7986. While these results show that the study can predict reasonably well, there's still room to make it better. In the future, researchers could look into more factors and try different ways of analyzing the data to improve the results. The study found that menopause, breast size, hormones, and how much time has passed since a woman's last reproductive event are important factors in breast cancer risk. Overall, this study gives us important information about how breast cancer effects in women, showing why we need to keep studying this topic.

7. Reference:

https://www.kaggle.com/datasets/utkarshx27/breast-cancer-dataset-used-royston-and-altman