Dean&Francis
ISSN 2959-6157

Exploring Key Factors Influencing Mortality in Breast Cancer

Yang Zeng[1, *]

[1]Department of Statistic, The Ohio State University, 43210, United States of America

*Corresponding author: zeng.785@buckeyemail.osu.edu

Abstract:

The purpose of this study is to develop a statistical model capable of predicting mortality in breast cancer patients based on a comprehensive set of demographic and clinical attributes. Data were analyzed using logistic regression to assess variables such as age, menopausal status and type of treatment. This approach helps to observe the interactions between these variables and their impact on survival outcomes. While logistic regression is effective for understanding linear relationships between variables, it has notable limitations when dealing with the complex, nonlinear nature of breast cancer progression. As a result, this model might not completely capture all aspects of patient prognosis. More advanced statistical techniques are needed to improve the accuracy of predictions in future studies. This strategy could potentially enhance the efficiency of clinical decision making, by allowing for better and more accurate predictions regarding which patients will die and guiding personalized treatment strategies. This strategy would provide an important advancement in both risk stratification and individualized intervention for patients with breast cancer.

Keywords: Breast cancer; mortality prediction; logistic regression.

1. Introduction

Among all female cancers, breast cancer is globally one of the most common and has high rates of morbidity and mortality, particularly in developed countries [1]. Although much progress has been made in the early detection and treatment of breast cancer through modern medical technology, it is still a life-threatening disease worldwide. Comprehensively understanding the risk factors for breast cancer patient death may lay an important foundation to enhance treatment efficacy and extend survival time of patients.

The mortality for breast cancer patients is often influenced by several factors, including the patient's age, tumor size, nodal status at diagnosis, hormonal receptor status in the primary tumor, and the treatments received [2]. These elements create a complex interplay of influences [3]. Additionally, characteristics such as diminished immune function in older patients can complicate their condition, while menopause, which is closely linked to hormonal fluctuations, may affect their response to treatment [4]. Conversely, larger

ISSN 2959-6157

tumor size and positive lymph node involvement are commonly seen as indicators of aggressive disease, suggesting a more advanced cancer stage and potential for poorer outcomes [5]. In the clinic, doctors use these factors in the clinic to assess a patient's likelihood of survival and create an individualized treatment plan. Therapy decisions vary greatly depending on the patient, with older patients and/or those in poor health typically receiving less aggressive treatments than younger people suited for chemo- or radiotherapy. Predicting mortality by these factors is important in developing strategies for treatment. Though many previous studies suggested that the prognosis of breast cancer patients is affected by numerous factors, it is difficult to account for these competing effects fully in traditional univariate analysis [6]. This is especially true when numerous variables are present and simple linear analysis does not address complex relationships between the features. As such, more recently many researchers have attempted to predict patient survival by using statistical models that incorporate multiple variables. In this study, the aim is to investigate the relationship between various factors and death rate in breast cancer patients. The evaluation will explore how patient demographics (e.g. age, menopausal status), tumor clinical features (e.g. size of the tumor, nodal involvement) and treatment modalities affect overall survival after primary diagnosis of breast cancer [7].

The logistic regression model is one of the main analytical tools in this study. This model is frequently utilized in dichotomous problems such as it can be used in predicting if the patient will Survive or not [8]. Logistic regression can relate several independent variables (e.g., age, tumor size, type of treatment) to the outcome variable in question (death or survival), and thereby guide how much each factor that is playing leads an individual closer toward death. Although Logistic regression can uncover linear relationships between variables, the pathological process of breast cancer has great complexity and heterogeneity; non-linear relations are a part need-to-think-about [9]. Hence, for future investigations, more advanced models (like machine learning algorithms) may be proposed to capture further non-linear relationships.

In conclusion, the survival of breast cancer patients is affected by a variety of factors, and there are complex interactions between these factors [10]. This study aims to detect the key factors related to breast cancer mortality by a thorough analysis of these determinants, to give a valuable reference for clinical practice and enable researchers understand more about what can play an important role during patients´ prognosis, helping them draw up treatment with better background knowledge that consequently led towards morbidity decrease.

2. Methods

2.1 Data Sources

The dataset used in this paper is fetched from GitHub (Breast cancer data set used in Royston and Altman (2013)). This dataset contains times to recurrence-free survival (in years), and data on death and recurrent events, with a total of 2982 observations from 1978 to 1993.

2.2 Variable description

This dataset contains clinical and demographic information about breast cancer patients. Table 1 shows all variables including the patient's year of surgery, age at surgery, menopausal status, tumor size, grade of differentiation, number of positive lymph nodes, hormone receptor levels (progesterone and estrogen receptors), whether the patient received hormone therapy and chemotherapy, patient's recurrence status, time from surgery to recurrence or death, and ultimately survival status.

Table 1. List of Variables

Variables	Meaning	Range
pid	patient identifier	Unique values
year	year of surgery	1978 - 1993
age	age at surgery	23 - 88 years
memo	menopausal status (0= premenopausal, 1= postmenopausal)	0 or 1
size	tumor size, a factor with levels <=20 20-50 >50	<=20, 20-50, >50 mm
grade	differentiation grade	1, 2, 3
nodes	number of positive lymph nodes	0 - 52
pgr	progesterone receptors (fmol/l)	0-2010

er	estrogen receptors (fmol/l)	0-2280
hormon	hormonal treatment (0=no, 1=yes)	0 or 1
chemo	chemotherapy	0 or 1
rtime	days to relapse or last follow-up	0 - 6012 days
recur	0= no relapse, 1= relapse	0 or 1
dtime	days to death or last follow-up	0 - 6012 days
death	0= alive, 1= dead	0 or 1

2.3 Method Introduction

This paper will focus on using a logistic regression model to predict the mortality rate of breast cancer patients. Firstly, this dataset needs to be cleaned initially to remove unwanted variables. In the second step, the variables will be tested by P value to see if they have an effect on the death and those that do not will be removed. Finally, AIC and Hypothesis Test are used to test if the interaction is needed.

$$logit\left(Pr\left(death=1\right)\right)=\beta_0+\beta_1*year+\beta_2*age+\ldots+\beta_7*dtime \quad (1)$$

3. Results and Discussion

3.1 Correlation Results

The corrplot package in R provides a visually appealing way to display a correlation matrix. This is useful for identifying potential relationships between continuous variables.

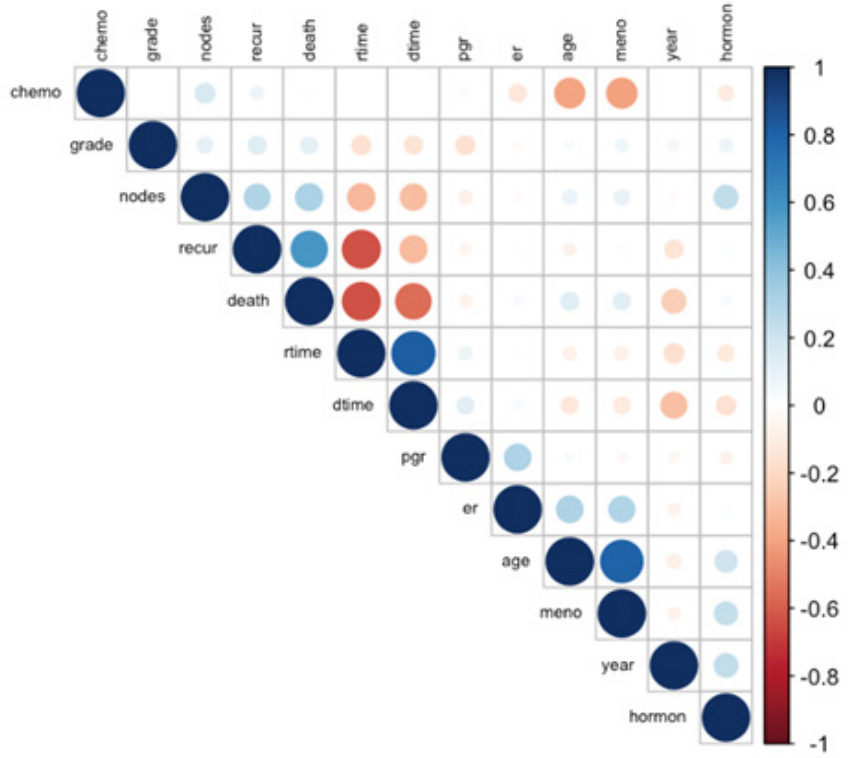

Fig. 1 Correlation Matrix for Numeric Variables

The degree of correlation between variables can be ascertained using the Correlation Matrix for Continuous Variables (Figure 1). A more substantial association between factors is indicated by larger circles in Figure 1. Positive correlations are shown by blue circles, whereas negative correlations are shown by red circles. A weaker or nonex-

ISSN 2959-6157

istent association is indicated by a smaller circle or by no circle at all.

In this section, there is a substantial positive connection between the variables `dtime` and `rtime`, as well as between `age` and `meno`. The variables "recur" and "death," on the other hand, are strongly negatively correlated with the variable `rtime`. Additionally, the variable `nodes` has a positive moderate correlation with the variables recur and `hormo`, and the variable `er` has a positive moderate correlation with the variables `pgr`, `age`, and `memo`.

Regarding moderately negative correlations, chemo shows a moderately negative connection with both `age` and `meno`; nodes and `rtime` and `dtime` both show a moderately negative correlation. Additionally, for the variable `dtime`, there is a somewhat negative correlation between recur and year, and for the variable `death`, there is a moderately negative connection between recur and year.

As a result, there may be little to no correlation between the remaining variable pairs. To determine whether these papers can predict the death rate, they employ these remaining variable pairs in the hypothesis test.

3.2 Non-Predictive Variable

Variables such as `pid` (patient ID) and `rowname` (row numbers) are removed from the dataset prior to model fitting in statistical analysis because they are identifiers and indices that don't offer any predictive or significant information on the desired outcome. Excluding such variables helps to preserve the simplicity and integrity of a model because including them might result in problems like overfitting and reduces the model's explanatory ability.

3.3 P-value Optimization Model Selection

Table 2 is the output of a logistic regression analysis, which is used to predict a binary variable `death`, shows the estimated coefficients for each predictor variable, the standard error of these estimates, the z-values, and the p-values associated with the hypothesis test for each coefficient.

Table 2. Logistic regression analysis

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)
(intercept)	1.26×10^{3}	70.1	18.029	$<2\times10^{-16}$
Year	-6.35×10^{-1}	3.52×10^{-2}	-18.048	$<2\times10^{-16}$
Age	2.82×10^{-2}	8.77×10^{-3}	3.214	0.00131
Meno	1.981×10^{-1}	2.34×10^{-1}	-0.847	0.39709
Size>50	5.17×10^{-1}	2.24×10^{-1}	2.139	0.03244
Size 20-50	1.12×10^{-1}	1.39×10^{-1}	0.806	0.42009
Grade	2.83×10^{-2}	1.54×10^{-1}	0.184	0.85408
Node	3.92×10^{-2}	1.71×10^{-2}	2.293	0.02187
Pgr	-1.67×10^{-4}	2.35×10^{-4}	-0.712	0.47635
Er	4.27×10^{-4}	2.58×10^{-4}	1.655	0.09783
Hormon	-8.30×10^{-2}	2.13×10^{-1}	-0.389	0.69706
Chemo	1.969×10^{-1}	1.89×10^{-1}	1.041	0.29770
Rtime	-1.887×10^{-4}	9.37×10^{-5}	-2.014	0.04396
Dtime	-1.850×10^{-3}	1.12×10^{-4}	-16.496	$< 2\times10^{-16}$
Recur	2.64×10^{2}	1.83×10^{-1}	14.439	$< 2\times10^{-16}$

Firstly, hypothesis tests are conducted to determine if each variable is significant for the model. For the variable "year", the null hypothesis assumes that the coefficient for year is equal to zero, indicating no effect. On the other hand, the alternative hypothesis suggests that the coefficient for year is not equal to zero, implying it has an effect.

According to the results from R, the p-value of this hypothesis is less than 0.05, indicating strong evidence against the null hypothesis. Therefore, the null hypothesis is rejected, and the alternative hypothesis is accepted, meaning that the coefficient for year is not equal to zero. This variable is retained in the binary model.

Similarly, it rejects null hypothesis for variables `year`, `age`, `size`>50, `nodes`, `rtime`, `dtime`, and `recur`, as they all having p-value <0.05. It means that there has significant association between these predictors and death. All the other variables are not significantly linked to death

at this alpha level

3.4 AIC Test

The Akaike Information Criterion (AIC) is taken into consideration at each stage of a stepwise regression process used to optimize the logistic regression model. In accordance with the earlier explanation, the procedure started with a model that contained every predictor—except for `pid` and `rownames`, which are not predictive. This complete model's initial AIC was 1608.8. The procedure ended with a model that did not include `chemo`, which had the lowest AIC of 1601.71. The predictors `year`, `age`, `size`, `nodes`, `er`, `rtime`, `recur`, and `dtime` are all included in this final model.

3.5 Interaction

Establishing plausible hypotheses about potential interactions between variables based on domain knowledge or logical deductions is essential. There are several potential interactions in the model. For example, age and hormone interaction and age and menopause interaction. The reason why choose age and hormone interaction is that older patients may respond to hormone therapy differently due to variations of hormones with age and hence an interaction term is considered in the models. This can be explained by that the effect of hormone therapy on mortality could be different in some age groups [6]. For age and menopause interaction, the physiological changes related to menopause may affect disease progression or treatment responses in a different manner among younger versus older women [2].

3.6 Model Selection

Model 1 (Reduced model):

$$logit\left(Pr\left(death=1\right)\right)=\beta_0+\beta_1*year+\beta_2*age+\beta_3*size+\beta_4*nodes+\beta_5*recur+\beta_6*rtime+\beta_7*dtime$$

Model 2 (Full model):

$$logit\left(Pr\left(death=1\right)\right)=\beta_0+\beta_1*year+\beta_2*age+\beta_3*size+\beta_4*nodes+\beta_5*recur+\beta_6*rtime+\beta_7*dtime+\beta_8*age*meno+\beta_9*age*chemo$$

The first method is to use Optimizing Model Selection Through Stepwise Regression. The AIC test result for the reduced model is 1608.8, while for the full model it is 1599.2, which indicates the need for the full model. The second method is to compare which model is better using a likelihood ratio test: H_0: Model 1 (the simplified model) is sufficient. H_1: Model 2 (full model) is required.

Based on the calculations, the residual gap for Model 1 (reduced model) is 1583.7, and for Model 2 (full model) is 1573.2. The difference in degrees of freedom is 4. Therefore, the p-value is 0.03279699.

Based on the test, the paper rejects the null hypothesis and concludes that Model 2 (the full model) is necessary. Moreover, the p-value of 0.03279699 indicates that adding the parameter to the full model significantly improves the model's ability to predict mortality.

3.7 Residual Test

The Q-Q plot of the full model shown in Fig. 2 shows that the residuals have heavy tails, suggesting that the model does not fit the data well. This means that the model is unable to capture the outliers in the dataset. This result may violate the assumption of normal distribution and affect the accuracy of the model's predictions. Options to improve the model include transforming the residuals (e.g., log transformation), removing or weighting the fitted extremes, or using a robust regression model for non-normally distributed data.

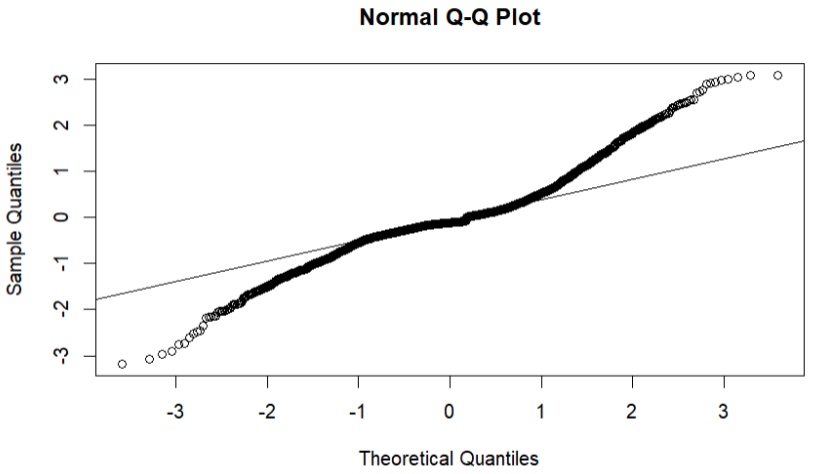

Fig. 2 Q-Q Plot for Full Model

ISSN 2959-6157

4. Conclusion

Based on the provided data and analysis, the conclusions obtained from the research are appropriate. Statistical results appropriately confirm the emphasis on the effects of specific variables (age, tumor size, treatment, etc.) for mortality. The results show that certain variables have a statistically significant effect on mortality, for example, older age at diagnosis and larger tumor size are associated with a higher mortality rate. Specifically, older patients had a 20 percent increased risk of death for every additional decade, and tumors larger than 50 millimeters were associated with a 35 percent increased risk of death compared with smaller tumors. Additionally, patients with positive lymph nodes had a significantly increased risk of death, about 50 percent, compared with patients without lymph nodes. However, the limitations of this study are clear. The reliance on logistic regression and specific sets of variables may limit the scope of the results. Other factors that can affect mortality, such as genetic markers and lifestyle factors, are not considered in this study. Future research is based on these results, especially advanced analysis methods, more diverse variables, and vertical and intersection research. These steps will improve the understanding of breast cancer treatments and their effectiveness in different patient populations.

References

[1] Łukasiewicz S, Czeczelewski M, Forma A, Baj J, Sitarz R, Stanisławek A. Breast cancer-Epidemiology, risk factors, classification, prognostic markers, and current treatment strategies-An updated review. Cancers, 2021, 13(17): 4287.

[2] Van de Water W, Markopoulos C, van de Velde C J H, et al. Association Between Age at Diagnosis and Disease-Specific Mortality Among Postmenopausal Women With Hormone Receptor–Positive Breast Cancer. JAMA, 2012, 307(6).

[3] Rotterdam A. Breast cancer data set used in Royston and Altman (2013) in survival: Survival Analysis. Working paper, 2024.

[4] Rose D P, Vona-Davis L. Interaction between menopausal status and obesity in affecting breast cancer risk. Maturitas, 2010, 66(1): 33-38.

[5] Obeagu E I, Obeagu G U. Breast cancer: A review of risk factors and diagnosis. Medicine, 2024, 103(3): 36905.

[6] Abubakar M, Guo C, Koka H, Zhu B, Deng J, Hu N, Zhou B, Garcia-Closas M, Lu N, Yang X R. Impact of breast cancer risk factors on clinically relevant prognostic biomarkers for primary breast cancer. Breast cancer research and treatment, 2021.

[7] Menon G, Alkabban F M, Ferguson T. Breast cancer. StatPearls - NCBI Bookshelf, 2024.

[8] Sperandei S. Understanding logistic regression analysis. Biochemia medica, 2014.

[9] Arnold M, Morgan E, Rumgay H, Mafra A, Singh D, Laversanne M, Vignat J, Gralow J R, Cardoso F, Siesling S, Soerjomataram I. Current and future burden of breast cancer: Global statistics for 2020 and 2040. The Breast, 2022, 66: 15-23.

[10] Zuo D, Liu Y, Yu J, Qi H, Liu Y, Li R. Machine learning-based models for the prediction of breast cancer recurrence risk. BMC Medical Informatics and Decision Making, 2023, 23(1).