P8130 Final Project Report

Yuechu Hu, Leyang Rui, Yifei Yu, Jinghan Zhao

Abstract

This study analyzes a breast cancer survival dataset of 4024 observations to develop predictive models for patient survival. Using logistic regression, we identified several significant predictors of survival times through model selection and validation. Survival analysis was applied to explore the changes in patients' survival status over time and its multiple risk factors. Results indicate that higher odds of death are associated with older age, greater regional node positivity, undifferentiated tumors, and advanced cancer stages, while marital status and race interact significantly. The findings highlight key risk factors while recognizing the study's observational limitations. Future research could focus on interventional research addressing these risk factors to optimize patient outcomes.

Introduction

The data we used for this analysis originates from a breast cancer survival dataset collected from a prospective study. The dataset contains 14 important predictor variables, including patients' age, race, marital status, tumor size, cancer stages (T Stage, N Stage, A stage, and 6th Stage), differentiated grade, estrogen and progesterone status, and regional node involvement. The dataset also records patients' survival times in months and their final survival status (dead or alive), which are the outcome variables of interest. The variable descriptions are shown in Tables 1, 2, and 3. Our objective is to develop models that predict the risk of death among breast cancer patients using these features. Specifically, we aim to identify which variables significantly impact patients' survival outcomes, determine potential interactions among the variables, and assess the performance of the models. Additionally, we will investigate fairness in the model predictions to ensure equitable accuracy between the majority race group and the minorities.

Methods

The dataset comprises 4024 observations and 16 variables, with no missing values. Initially, we cleaned and tidied the data using data-wrangling techniques such as mutating. Then, graphical tools such as histograms and box plots helped us to figure out the distributions of the variables and check for potential outliers or influential points. We also examined the pairwise relationships between variables through scatter plots. We tried logarithmic transformation for variables with right-skewed distributions, but ultimately used the original data because the impact of the transformation was not obvious.

Since the response variable, *status*, is a binary categorical variable that indicates the survival result of the patients, we chose to build a logistic regression model to predict its estimated probability. Since breast cancer grade (i.e. variable *grade*) is exactly determined by the degrees of the variable *differentiate*, we removed *grade* from the model. In addition, the level IIIC of variable *x6th_stage* also correlates with other variables in the dataset. However, the AJCC system is a complex criterion based on multiple aspects, so we cannot simply remove this variable as part of its information may still be useful.

To find the best model for predicting the patients' survival status, we started by using all 3 automated procedures–forward selection, backward elimination, and stepwise regression to choose models with statistically significant predictors respectively. Next, the criterion-based procedure–AIC and BIC are applied to compare the values among the three automatically generated models and the full models to choose the best final model. For model diagnosis, the variance inflation factor (VIF) and the generalized variance inflation factor (GVIF) were used to detect multicollinearity among numerical and categorical variables. Since the residual versus fitted values plot always shows a pattern in logistic regression because of the binary response variable, we randomized the quantile residuals to examine the assumption regarding residuals. Additionally, a residual versus leverage plot is used to explore potential outliers. Finally, we employed a 10-fold cross-validation and evaluated the goodness-of-fit of the model by log loss and AUC. Similarly, to test the performance of the model among the majority white and other races, we did another stratified cross-validation and found potential space for improvement.

We also performed the survival analysis with patients' status and survival months as response variables. The Kaplan-Meier survival time curve represents the survival rate over time, and the log-rank test tells us whether there is a difference in survival times between patients whose cells have different degrees of differentiation. Since both methods are limited in that only one variable can be tested at a time, we computed Cox proportional hazard models to adjust for multiple risk factors simultaneously.

Results

As shown in Figure 1, most patients are between 40 and 70 years old, and the most frequent survival times are larger than 45 months. The number of examined regional nodes for most patients is smaller than 30, and the majority of patients have nearly 12 examined regional nodes. It is worth noting that the distributions of both variables *regional_node_positive* and *tumor_size* are significantly skewed to the right. However, since logistic regression does not require all predictors to be normally distributed and we reexamined their distributions after transformation and confirmed that skewness was not significantly reduced, we would not use the transformation in further analysis. Over 2500 subjects only have 1 or 2 positive regional nodes, the most frequent number of positive regional nodes. Most tumor sizes are smaller than 50 mm, and we found that the most frequent size is around 19 mm, followed by around 14 mm. Figure 2 distributed the survival time by the status, the "dead" group is concentrated in the shorter survival months, while the "alive" group is predominant in longer survival months, particularly beyond 60 months. According to Figure 3, as the *t_stage* changes from stage 1 to stage 4, the size of the tumor also increases. We also noticed some potential outliers both in the T1 stage and T3 stage. Looking at Figure 4, the survival time is longer in the regional stage, and the "alive" group shows higher survival times across both stages. In Figure 5, the undifferentiated group has larger tumor sizes compared to the other categories, while the well, moderately, and poorly differentiated groups all display similar distributions with the majority of tumor sizes being small except for numerous high-value outliers. Figure 6 highlights the differences in tumor size distribution and trends with age between individuals who are "alive" and those who are "dead". While the "alive" group shows no significant relationship between age and tumor size, the "dead" group exhibits a pattern where larger tumors are associated with younger ages. Finally, according to Figure 7, as it changes from well-differentiated to undifferentiated, the negative correlation between the number of positive regional nodes and the patients' survival months strengthens.

After comparison, the stepwise regression model was chosen as the final model since it has the smallest AIC and BIC value, shown in Table 4, and the results of the final model are represented in Table 5. All but *marital_status* are highly significant variables with very small p-values. For example, people with undifferentiated tumors have 6.46 times the odds of death compared to those with well-differentiated tumors. With respect to model diagnostic, as Table 6 reveals, all the adjusted GVIFs (a measure corrected for the degree of freedom and provides a scale similar to VIF for continuous variables) are less than or not much different from 2, implying the absence of multicollinearity. The randomized quantile residuals versus fitted values plot (Figure 8) displays a pattern of randomized residuals equally distributed around the 0.5 line, satisfying the assumptions of linearity and residual equal variance. Moreover, the residual versus leverage plot (Figure 9) indicates that observations 3527, 1561, and 3074 may be potential outliers, but they are not necessarily influential and we will keep them for future attention. The results of the 10-fold cross-validation, displayed in Table 7, show the goodness of fit by log loss and area under curve (AUC). The mean of log loss is 0.372, and the mean of AUC is 0.742. The prediction performance is better in the majority race group "White" than the minority "Black" and "Other" as shown in the results of stratified validation by levels of race (Table 8), where lower log loss and larger AUC indicate better test performance. Since the distribution of survival months is different between races with different marital statuses (Figure 10), we added an interaction term *marital_status*race* in the model, which further reduced the gap between race groups and improved the performance of our model (Table 9).

For survival analysis, the Kaplan-Meier curve (Figure 11) shows the overall survival rate over months. We found a significant difference in survival between patients whose cancer cells have different degrees of differentiation in the log-rank test (Figure 12). To further discuss the multiple risk factors to survival time, we performed the Cox proportional hazard model. The assumption of the Cox model was tested based on the scaled Schoenfeld residuals, that is, the survival curves of the two different strata of the risk factor must have a proportional hazard function that varies over time. Table 10 shows that *a_stage*, *estrogen_status,* and *progesterone_status* are not constant over time, so we removed these variables in further Cox model analysis. The forest plot (Figure 13) shows the results of the Cox model, that is, the relationship between multiple variables and the probability of death. A hazard ratio greater than 1 indicates an increased probability of death, and a hazard ratio less than 1 indicates a decrease. The smaller the p-value, the greater the weight of evidence that there is a difference between the groups.

Conclusion/Discussion

Through the visualization of the variables, survival months were significantly higher for the “alive” group compared to the “dead” group and were longer in the regional stage compared to the distant stage. Additionally, as the T stage progresses from stage 1 to stage 4, tumor size consistently increases. Meanwhile, the undifferentiated group has larger tumor sizes compared to the other categories. As tumor differentiation shifts from well-differentiated to undifferentiated, the negative correlation between the number of positive regional nodes and patients’ survival months becomes stronger. While the “alive” group shows no significant relationship between age and tumor size, the “dead” group exhibits a pattern where larger tumors are associated with younger ages.

The overall survival rate in this dataset decreased over time and finally remained above 75%. According to the results of our final model (Table 5), the odds of death increase as age and number of positive regional nodes increase, and decrease as the number of examined regional nodes increases. The odds of death are higher for black people, separated couples, and widowed individuals, with higher T stages, and higher N stages. In addition, the more undifferentiated the tumor is, the higher the odds of death is. The results of survival analysis show that the hazard of death is significantly higher in black people, separated couples, patients with higher N stages, and lower differentiated cancer cells.

Although many variables are significant in predicting the odds of death, it is important to note that this dataset is from an observational study, so the conclusion is limited to correlations between these predictors and survival status and no causation can be ascertained. Further research could focus on prescribing drugs to patients with different stages and types of tumors and reassessing their survival status afterwards to effectively find the most appropriate drugs for patients with different conditions.

Group members’ Contributions

Leyang and Jinghan focused on statistical methods, building models to extract meaningful insights. Yuechu and Yifei concentrated on data description and visualization, presenting information through clear visuals. All four of us contributed to the writing of the report, integrating our individual efforts into a cohesive final report that reflects both analytical depth and visual clarity.