# Prediction of Breast Cancer Survival using Machine Learning Techniques

Ye Yao, Zhilin Zhang, Hengde Ouyang

12/17/2022

## Introduction

Breast cancer is the most common cancer among women in the United States. According to the survey, the average risk of a woman in the United States developing breast cancer is about 13%[1]. The survival prediction, especially 5-year survival analysis, was largely used to help doctors understand the prognosis, predict treatment efficiency and develop individualized treatment plans. In previous studies, machine learning was largely applied for its high accuracy and fewer assumption requirements. In this study, the survival prediction model was built based on data with 300998 cases and 9 variables extracted from the SEER research database.

## Data Description

Nine variables with potential effects on survival were selected, including Age, Race, ER_PR, HER, Tumor Size, surgery, sex, Reginal and Stage. The Race was categorized into Asian, Black, White, America Indian, Native Hawaii and other Pacific islanders and others. ER_PR, the score of estrogen receptor/progesterone receptor tests. was reflected as 0, 1, 2, indicating ER-/PR-, ER+/PR- or ER-/PR+, ER+/PR+. And receiving surgery of removing the tissue in the primary site was also represented as Surgery. The outcome variable is survival time in months.

## Data Preprocessing

Data from 2004-2014 was treated as train data while data from 2015 was treated as test data. Multivariate imputation by chained equations (MICE) was used to impute the covariates, assuming that the missing data were missing at random (MAR). Incomplete variables were imputed by separate models, including predictive mean matching for numeric variables,

logistic regression for binary variables, bayesian polytomous regression for factor variables, and proportional odds model for ordinal variables. There are 93617 cases in the complete data after mice imputation.

# Method

Survival analysis is known to explore the occurrence of an event of interest, such as alive, death or recurrence. In this study, the years of survival were calculated from the date of diagnosis to the date of death. Patients' five years of survival status is the main interest, only patients dead 5 years after diagnosis or lost follow-up would be considered. Among the total of 94326 patients, 74313 were alive after five years of diagnosis, 658 were censored and 20013 were dead. The survival package was used to plot the stratified survival curve for each important categorical variable of interest to identify its impacts on the survival rate. Based on the Kaplan-Meier plots and the descriptive analysis, two predicting models were constructed.

Cox regression analysis was used to estimate hazard ratios (HRs), 95% confidence intervals and p-values. The study used a Forest plot to specify the efficiency of predicting variables. To observe the patient's 5-year survival rate, we used the "rms" package to build a model for the filtered variables and draw a nomogram to visualize the prediction results of the independent variables and survival outcomes.

The random survival forest takes the binary survival tree as the basic unit and is an extension of the traditional binary decision tree. When the input data passes through the nodes of the binary decision tree, it will be divided into two groups of data by the judgment conditions of the nodes until the input data are classified into the same category. Unlike the usual decision tree, the survival decision tree uses a log-rank test to determine the splits. The study first constructed a random survival forest with 100 decision trees. Based on the decreasing rate of error, 60 trees were determined to build our model. The information on the RSF model is given in the table.

# Result

### Cox Proportional Hazard Model

#### fig1

Adjusting for other variables, it was determined that the key predictor variables influencing survival probability includes sex (HR=0.78, 95%CI=[0.68,0.88]), age(HR=1.03, 95% CI=[1.03,1.04]), all groups in ER_PR(HR=0.70, 95% CI=[0.66,0.73] and HR=0.51,95% CI=[0.49,0.53]), surgery (HR=0.34, 95% CI=[0.32,0.35]), and HER2 (HR=1.25, 95%CI=[1.20,1.31]).

For binary variables, females were predicted to have longer survival time than males. Having surgery and a positive HER2 test result also indicate a higher chance of survival. For the continuous variables, it was observed that as age increases, the predicted survival time decreases. And patients with more ER and PR tests being positive were predicted to have a longer survival time.

What's more, for categorical variables, the relationship between survival months and each variable becomes more complicated. Compared to the reference group, stage I (HR=0.82, 95% CI=[0.59,1.14]) and stage II (HR=1.26, 95% CI=[0.97,1.76]) were observed to be insignificant. However, patients at stage III(HR=1.95, 95% CI=[1.40,2.72]) and IV (HR=3.98, 95% CI=[2.86,5.55]) were predicted to have shorter survival time. Moreover, from stage II to stage IV, the hazard ratio increases dramatically.

The effect of tumor size on survival months is also complicated. Compared to the reference group, a tumor size of 40-50 (HR=0.28, 95%CI=[0.01,6.47]) and a tumor size of 50-60 (HR=1.84, 95% CI=[0.98,3.47]) has an opposite effect on survival months. And the interaction term between tumor size and potential confounder lymph nodes was still insignificant.

Lastly, when comparing survival months among different races, Asian is treated as the reference group. Black (HR=1.31, 95% CI=[1.23,1.40]), White (HR=1.06, 95% CI=[1.00,1.12]), America Indian(HR=1.22, 955CI=[1.04, 1.42] and Native Hawaii and other Pacific islanders (HR=1.18, 95% CI=[1.05,1.33]) have a significantly higher hazard ratio, reflecting a higher risk of death.

# fig2

The significance of variables is determined by the effect estimates and is influenced by covariables. In this figure, age contributes the most to the outcome thus it is assigned 100 points. In proportion to the most effective variable, the rest are assigned a point according to their effect size. Relative importance could be understood by comparing the most significant variable (Age) to the least significant variable (HER). # fig3

Kaplan-Meier plots showed the survival curves using the log-rank test for the comparison of HER and ER_PR variables correspondingly. The p-value in each graph indicates a significant difference in survival curves among groups. For the ER_PR variable, three survival curves have clear separation without intersection. For the HER variable, a separation is not clear enough, we can still make the inference that a positive HER indicates a smaller risk of death.

**Model Result**

# fig4

The OOB(Out-of-bag) scores for the Cox proportional hazard model and random forest model are 0.256 and 0.255, which indicates that they predicted the OOB samples with

74.4% accuracy and 74.5% accuracy, respectively. The C-index calculated from the Cox proportional hazard model is 0.745 in training data and 0.539 in test data. The C-index calculated from the random forest model is 0.801 in training data and 0.671 in test data.

# Conclusion

The C-index in both Cox proportional hazard model and the random forest model is bigger than 0.5, which reflects that they are prediction models with good fits. And the C-index score of the random survival forest is higher than the score of the Cox proportional hazard model, which also indicates that the random survival forest has a higher model fit and prediction accuracy. Machine learning methods have higher accuracy, fewer required assumptions, and a higher capacity to deal with big data. And that's why it was applied in this study. On the other hand, compared to train data, the C-index in test data is lower. With the development of treatment, early detection of breast cancer and improved diagnosis accuracy, there might be bias by using train data from 2004 to 2014 to predict test data from 2005.

Here are the limitations of the study. The tumor size was observed to be insignificant, which is inconsistent with previous studies. The effect of tumor size on survival prediction is associated with the number of positive lymph nodes(LNs). It has been indicated that small tumors with more than four positive LNs might be more aggressive diseases compared to bigger tumors[2]. In addition, stage I showed an insignificant result. In previous studies, stage 0 and stage I were combined together and treated as one reference group because both survival rate is almost 100%[4]. That might explain why it's insignificant in this study.

# Reference

[1]Breast cancer statistics: How common is breast cancer? American Cancer Society. (n.d.). Retrieved December 17, 2022, from https://www.cancer.org/cancer/breast-cancer/about/how-common-is-breast-cancer.html

[2]Liu Y, He M, Zuo WJ, Hao S, Wang ZH, Shao ZM. Tumor Size Still Impacts Prognosis in Breast Cancer With Extensive Nodal Involvement. Front Oncol. 2021 Apr 9;11:585613. doi: 10.3389/fonc.2021.585613. PMID: 33898305; PMCID: PMC8064390.

[3]Balabram D, Turra CM, Gobbi H. Survival of patients with operable breast cancer (Stages I-III) at a Brazilian public hospital–a closer look into cause-specific mortality. BMC Cancer. 2013 Sep 24;13:434. doi: 10.1186/1471-2407-13-434. PMID: 24063763; PMCID: PMC3849091.

[4]Elobaid, Y., Aamir, M., Grivna, M., Suliman, A., Attoub, S., Mousa, H., Ahmed, L. A., & Oulhaj, A. (n.d.). Breast cancer survival and its prognostic factors in the United Arab Emirates: A retrospective study. PLOS ONE. Retrieved December 17, 2022, from https://journals.plos.org/plosone/article?id=10.1371%2Fjournal.pone.0251118

[5]Understanding breast cancer survival rates. Susan G. Komen®. (2022, July 20). Retrieved December 17, 2022, from https://www.komen.org/breast-cancer/facts-statistics/breast-cancer-statistics/survival-rates/