Survival Analysis For Breast Cancer

Survival and Longitudinal Data Analysis Project

Kimmeng HONG

1rd December 2025

1 Introduction

Breast cancer is one of the most common cancers in women and can vary a lot between patients. Some patients live many years after diagnosis, while others experience a relapse or die sooner. Studying which factors affect survival really helps doctors predict outcomes and plan treatments. In this report, we use a public breast cancer METABRIC dataset, which contains clinical information (like age, tumor stage, and hormone receptor status). The goal of this study is to explore how patient characteristics relate to Overall Survival (OS), to compare survival between different patient groups, and to build survival models using both traditional (Cox) and machine learning (Random Survival Forest) methods. We also interpreted the results and did survival prediction. Lastly, we discussed ways to improve prediction by using additional genetic information.

2 Data Description

In this study, we used METABRIC [1] dataset, which contains information on about 2509 breast cancer patients. It includes clinical variables such as age at diagnosis, tumor stage, hormone receptor status (ER, PR), HER2 status, type of surgery, and treatment information. Moreover, it also contains high-dimensional genetic data.

Two survival outcomes are considered: Overall Survival (OS), defined as the time from diagnosis to death or last follow-up, and Relapse-Free Survival (RFS), defined as the time from diagnosis to cancer recurrence, death, or last follow-up. Patients who have not experienced the event at the last follow-up are treated as censored. However, we only focused on the OS in this study.

3 Exploratory Data Analysis

Before performing exploratory data analysis of the clinical data, categorical variables were converted into factors to ensure that variables are in correct format.

3.1 Descriptive Statistics

The histogram in Figure 1 shows the ages of patients diagnosed with breast cancer in the METABRIC dataset. The distribution is approximately bell-shaped and centered around the early 60s. Very few patients were diagnosed before 40 or after 90. The chart shows that breast cancer is most commonly diagnosed in older women aged around 50–75.

According to Figure 2, the distribution of tumor size is non-normal and right-skewed. Most tumors are small to medium (about 10–40 mm), but there are some extreme values, where some patients have much larger tumors exceeding 60mm, which creates a long right tail. This shows that tumor sizes vary a lot accross patients.

In Figure 1, the bar plots show the number of patients by ER, HER2, and PR status. Most patients are ER-positive and PR-positive, while HER2-positive cases are less frequent. Each features also has some missing (NA) values, which we will handle later before fitting into models.

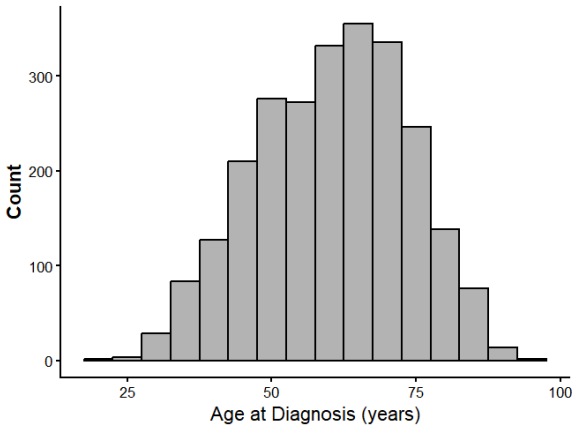

Figure 1: Age Distribution

3.2 Correlation Analysis

To explore dependencies between variables, Spearman's rank correlation was calculated for all variable pairs. We preserved the factor order of ordinal variables, while nominal categorical variables were converted into dummy variables.

Spearman correlation was chosen because it captures monotonic relationships and it is robust to non-normal distributions like Tumor Size as we see in Figure 2. Figure 4 shows the resulting correlation heatmap. Most other variables showed weak to moderate correlations but some variables show strong Spearman correlations (above 0.6). This mainly happens because several variables describe very similar biological or clinical information. ER and HER2 status are strongly linked to the molecular subtype labels that are based on these mark-

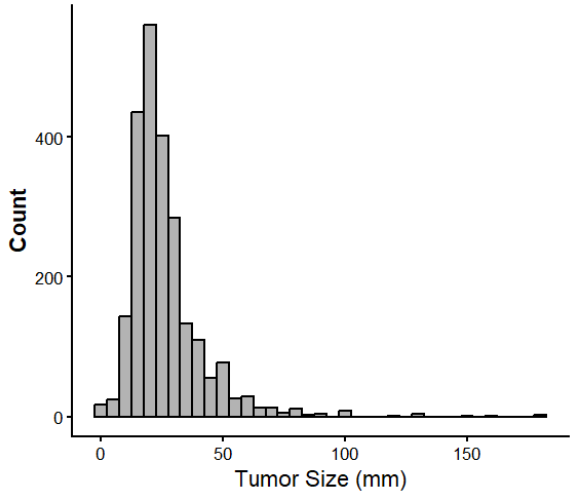

Figure 2: Tumor Size Distribution

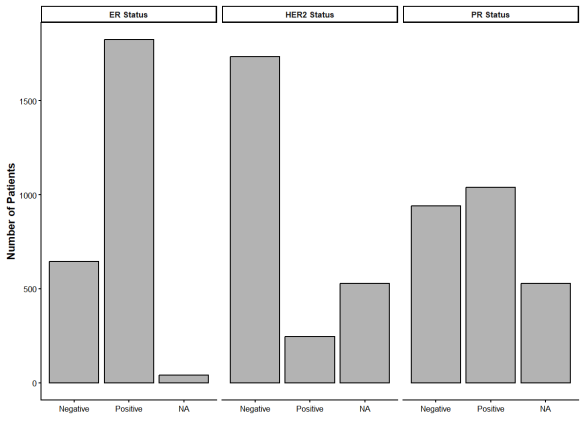

Figure 3: Distribution of ER, PR, and HER2 Status

ers. Tumor histologic subtype also matches closely with the cancer type categories and Oncotree codes, which describe the same tumor types in different ways. Clinical measures such as lymph-node positivity, tumor stage, and the Nottingham Prognostic Index are also related because they use overlapping information. These strong correlations show that some variables in the dataset provide similar information, so it is important to notice about this when choosing variables for survival analysis to avoid multicollinearity problem.

3.3 Kaplan–Meier Curves

In Figure 5, we look at how long patients survive depending on the stage of their tumor. Patients with early-stage cancer (Stage 0 or 1) have the best survival over time. As the tumor stage becomes more advanced (Stages 3 and 4), survival drops faster. This shows that the stage of the cancer at diagnosis is very important for predicting patient outcomes. In Figure 6, we compare survival for two types of breast surgery: breast-conserving surgery and mastectomy. The two lines are quite close to each other, meaning both treatments provide similar long-term survival. There is a small advantage for breast-conserving surgery, but the difference may not be large enough to be statistically meaningful.

Moreover, the Kaplan-Meier curves for both HER2

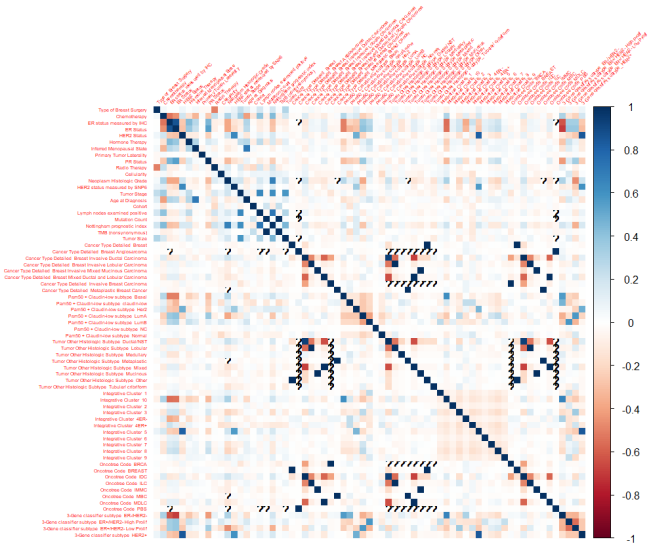

Figure 4: Spearman's Rank Correlation Plot

status and ER stage in Figure 7 and 8 demonstrate distinct differences in overall survival. Patients with HER2-negative status and those with ER-positive status tend to have better overall survival rates. Further statistical analysis in the next section, a log-rank test, was performed to confirm whether these differences are statistically significant.

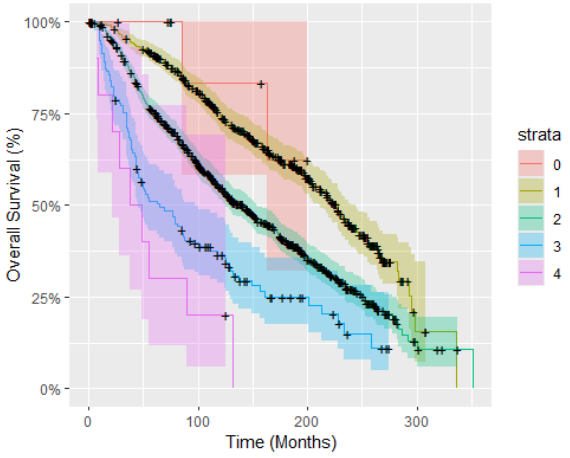

Figure 5: Kaplan-Meier Plot of Tumor Stage

3.4 Log Rank Test for Categorical Variables

Log-rank tests were performed to compare overall survival between groups for all categorical variables. The results in Table 1 show that Tumor Stage has the strongest effect on OS, followed by Type of Breast Surgery, Integrative Cluster, and Pam50 + Claudin-low subtype. Other variables, such as HER2 Status and ER Status, also show significant but with weaker associations.

Variables with p-values greater than 0.05, such as Cellularity, Primary Tumor Laterality, and Cancer Type Detailed, do not significantly influence the overall survival.

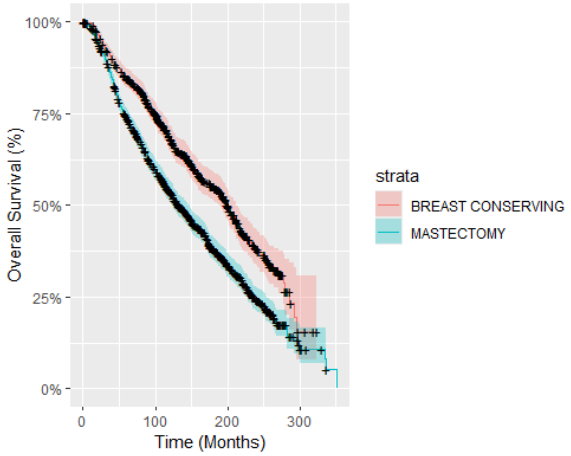

Figure 6: Kaplan-Meier Plot of Type of Breast Surgery

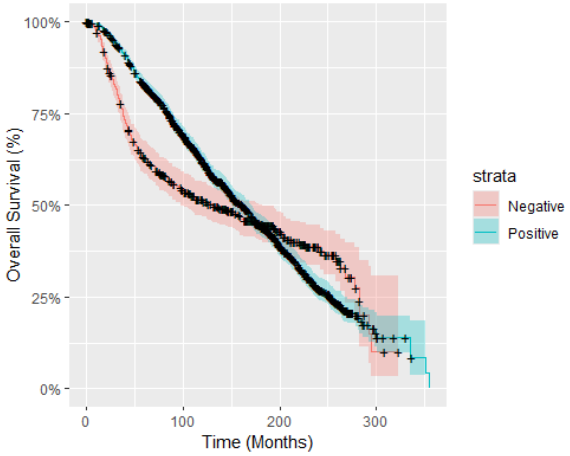

Figure 8: KM of ER Stage

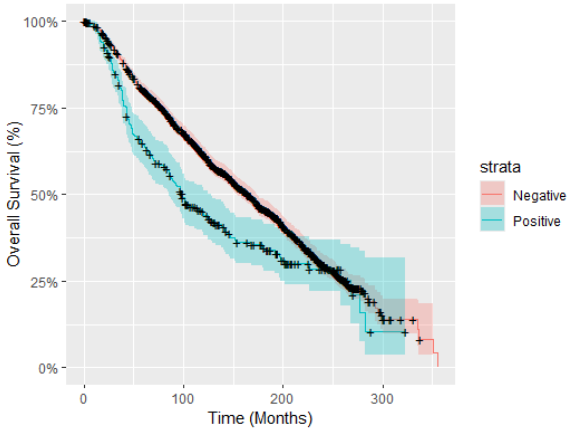

Figure 7: Kaplan-Meier Plot of HER2 Status

4 Methodology

4.1 Data Preprocessing

In order to fit the data into models, missing values have to be handled. First, 528 missing values of Overall Survival Months and Status which are the time and event outcomes, were removed. Next, if the missing values of any columns were less than 5% of total missing values in that columns, they were imputed by median for numerical variables and by mode for categorical variables. For columns having more than 5% of missing values, they were imputed using MICE (Multiple Imputation by Chained Equation) [2] to better preserve relationships between variables and reduce bias.

4.2 Cox Proportional Hazard Model

4.2.1 Linear Cox Models

First, a Cox proportional hazards model was fitted using all available clinical variables to see how each of them relates to Overall Survival (OS). To reduce the number of variables and avoid overfitting, stepwise selection with AIC was used, which keeps only the variables that improve the model. Using the selected variables, final linear Cox model was fitted, which serves as the baseline model for comparison.

Variable	Log-rank p-value
Tumor Stage	9.349e-27
Type of Breast Surgery	1.932e-12
Integrative Cluster	1.370e-11
Inferred Menopausal State	2.124e-10
Pam50 + Claudin-low subtype	2.862e-10
3-Gene classifier subtype	6.964e-09
Neoplasm Histologic Grade	1.851e-06
HER2 Status	8.568e-06
HER2 status measured by SNP6	3.220e-05
PR Status	8.277e-05
Hormone Therapy	1.434e-04
Chemotherapy	1.794e-03
Radio Therapy	6.537e-03
Tumor Other Histologic Subtype	1.098e-02
ER Status	3.505e-02
ER status measured by IHC	1.102e-01
Cancer Type Detailed	1.728e-01
Oncotree Code	1.728e-01
Primary Tumor Laterality	4.101e-01
Cellularity	8.895e-01

Table 1: Log Rank Test p-value of Categorical Variables

4.2.2 Model Diagnostics

The Cox proportional hazards model assumes that continuous variables have a linear effect on the hazard, and that the hazard ratio is constant over time. To check linearity, the Martingale residual was used and plotted as in Figure 9. The figure suggest that the residual of Age at Diagnosis is not linear since it shows a curvature. Also, Tumor Size shows clear nonlinear patterns, with the smoothed curves bending, showing that the linear assumption may be violated for these variables. Moreover, to check time invariance assumption, Schoenfeld residuals was analysed and plotted for all variables. As the results, all variables are likely time-independents. So the time invariance assumption for the cox model is hold.

4.2.3 Non-Linear Cox Model

Based on the diagnostic results, a more flexible survival model was constructed by applying non-linear functional forms to selected covariates. The model uses the

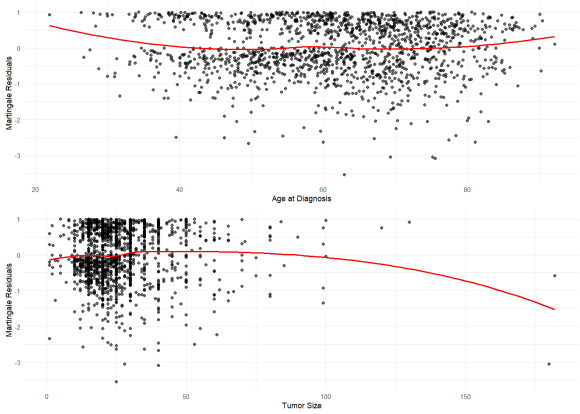

Figure 9: Martingale Residual Plots for Continous Variables

same set of variables identified through stepwise AIC in the linear Cox model. Tumor Size was log-transformed to address right skewness and improve linearity in the hazard relationship. Age at Diagnosis was modeled using natural cubic splines (ns) with 3 degrees of freedom (df), allowing flexible curvature in how it influence hazard without imposing a strict linear form.

4.3 Random Survival Forests

To complement the Cox proportional hazards approach, Random Survival Forests (RSF) was applied. RSF is a non-parametric tree-based model that can capture complex interactions and non-linear relationships between variables, without requiring proportional hazards or linearity assumptions. First, a baseline RSF model was trained using default parameters to establish a reference model for comparison and to observe variable importance without tuning. Next, the model was calibrated to improve performance. The number of trees was kept the same (ntree = 200). The number of variables tried at each split (mtry = 7) was reduced to helped control overfitting. Also, a smaller nodesize = 30 made trees slightly more flexible, and using nsplit = 20 to make finer splitting during survival estimation.

5 Results

5.1 Model Interpretation

5.1.1 Cox Proportional Hazard Model

Table 2 shows the hazard ratios and p-values from the linear Cox model after stepwise AIC selection. Several factors were strongly associated with overall survival. Age at diagnosis increased the risk slightly but steadily (HR = 1.054 per year), showing that older patients had lower survival chance. LumB and HER2-positive subtypes increased the hazard compared to a reference subtype, Basal. Number of positive lymph nodes, tumor size, and Nottingham prognostic index were also associated with higher risk. In contrast, ER-positive status and post-menopausal state showing a decrease in the risk of death. Meanwhile, Tumor Stage produced extremely large Hazard Rates with large p-value, likely due to small numbers of events in some categories, so its effect is insignificant and unstable in this model.

Variable	Hazard Rate	p-value
Age at Diagnosis	1.054	2e-16
Type of Breast Surgery (Mastectomy)	1.140	0.146987
Chemotherapy (Yes)	1.326	0.021335
Pam50 + Claudin-low subtype (claudin-low)	0.8311	0.273727
Pam50 + Claudin-low subtype (Her2)	1.194	0.291749
Pam50 + Claudin-low subtype (LumA)	1.197	0.302421
Pam50 + Claudin-low subtype (LumB)	1.474	0.028700
Pam50 + Claudin-low subtype (NC)	1.607	0.320954
Pam50 + Claudin-low subtype (Normal)	1.462	0.058379
ER Status (Positive)	0.6545	0.002913
HER2 Status (Positive)	1.336	0.016038
Inferred Menopausal State (Post)	0.5564	2.12e-05
Lymph nodes examined positive	1.036	2.67e-05
Nottingham prognostic index	1.157	0.000327
Radio Therapy (Yes)	0.8442	0.054975
Tumor Size	1.006	0.003365
Tumor Stage 1	8.733e+05	0.985645
Tumor Stage 2	1.023e+06	0.985479
Tumor Stage 3	9.776e+05	0.985526
Tumor Stage 4	1.672e+06	0.984963

Table 2: Hazard Ratios and p-values of Cox Model

5.1.2 Random Survival Forests

The variable importance results in Figure 10 show that a few key factors have the biggest impact on predicting survival. Age at diagnosis is the strongest predictor, followed by number of positive lymph nodes and the Nottingham prognostic index, meaning these clinical features give the most useful information for the model. Tumor related factors such as stage, size, and mutation count, also matter but with less effect. Other features, including molecular subtypes, treatment types, and receptor statuses, contribute only small amounts.

5.2 Model Performence

In Table 3, the Integrated Brier Score (IBS), which measures overall prediction error across time (lower is better), shows that the calibrated Random Survival Forest (RSF) performs best, followed closely by the default RSF and the non-linear Cox model. The linear Cox model performs worse than these, and the null model (Kaplan-Meier) has the highest error. The Brier score plot in Figure 11 supports these IBS results: both RSF models maintain lower error curves across most time points, while the null model consistently shows the worst performance. The results show that the models that capture non-linear relationships, especially RSF, provide the most accurate survival predictions for this dataset.

Model	IBS
Null model	0.174
Linear Cox Model	0.152
Non-linear Cox Model	0.149
RSF Default	0.148
RSF Calibrated	0.147

Table 3: Integrated Brier Score (IBS) across different models.

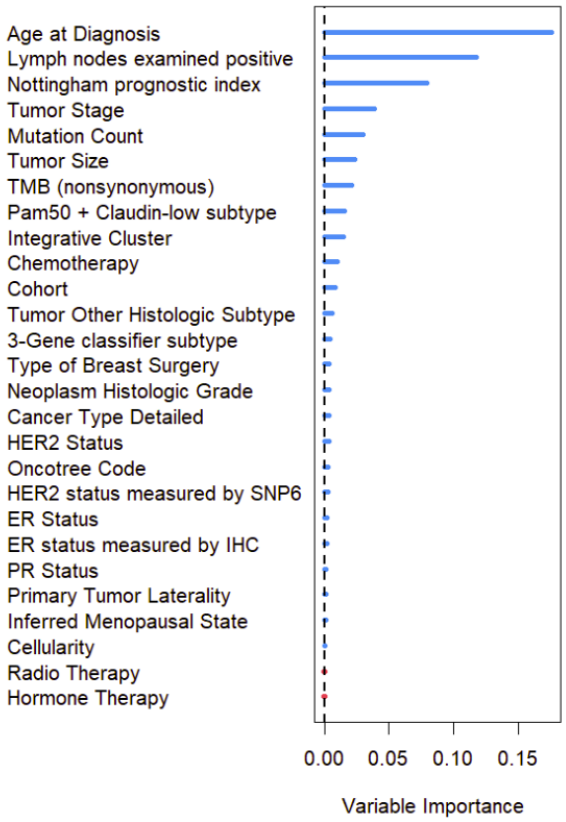

Figure 10: Random Forest Variable Importance

5.3 Survival Prediction

5.3.1 Predicted 12-Year Survival for Representative Patients

We estimated 12-year survival for three different simulate patients. In order to provide a fair comparison, some factors were fixed. All patients were assumed to have ER-positive, HER2-negative tumors, be in the pre-menopausal inferred state, have two positive lymph nodes, Nottingham Prognostic Index = 2, tumor size = 25 mm, no radiotherapy, and be treated with breast-conserving surgery. We are interested in studying the main survival difference of these patients in term of their ages, subtypes, chemotherapy, and tumor stages.

Under this scenario, by using the non-linear cox model, we obtained survival estimation of these patients as in Figure 12. The 40-year-old patient with LumA subtype and Stage 1 tumor (Patient 1) had the highest estimated 12-year survival at about 81%. Meanwhile, The 55-year-old LumB patient with Stage 3 tumor and chemotherapy (Patient 2) had a moderate survival of around 62%. Lastly, the 70-year-old Basal subtype patient with Stage 2 tumor and chemotherapy (Patient 3) had the lowest estimate of about only 50% of being alive beyond 12 years.

In summary, the model suggests that younger patients with LumA and early-stage tumours are most likely to survive long-term, while older patients and those with Basal subtype have lower survival chance, even when other factors were constant.

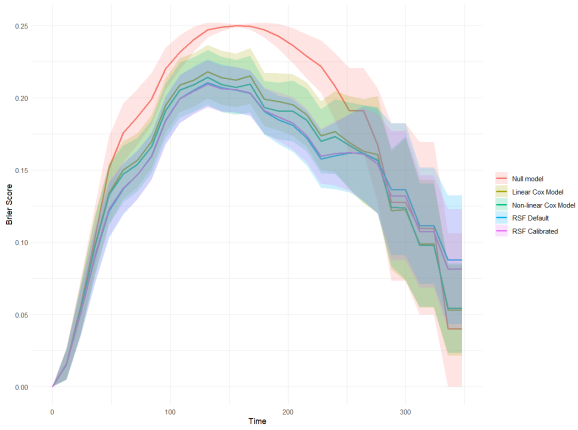

Figure 11: Comparison of Brier Scores for Different Models

5.3.2 Conditional Survival Beyond 7 Years

We next considered the conditional probability of survival for patients who have already survived 7 years since diagnosis. Using the previously defined patient profiles, we calculated the probability of surviving an additional 5 years. This was done by taking the ratio of the survival probability at 12 years to that at 7 years, using the non-linear Cox model predictions, as indicated below.

$$P(T > 12\,\text{years} \mid T > 7\,\text{years}) = \frac{P(T > 12\,\text{years})}{P(T > 7\,\text{years})}$$

According to Table 4, the conditional probabilities are all slightly higher than the corresponding absolute 12-year survival probabilities because they account for having already survived 7 years. Theoretically, this reflects the definition of conditional survival, which updates the predicted survival based on the patient's current status. Practically, this means that a patient who has already survived several years since diagnosis has a slightly better chance of being alive for the next 5 years than the original 12-year survival probabilities. As we can see in Table 4, Patient 1's chance of surviving from year 7 to 12 is about 90%, slightly higher than the overall 12-year probability of 81%.

Patient	$P(T > 12)$	$P(T > 12 \mid T > 7)$
Patient1	0.814	0.904
Patient2	0.621	0.792
Patient3	0.504	0.715

Table 4: Survival probabilities and conditional probabilities for selected patients.

6 Alternative Method Using High Dimensional Gene Expression

If we want to maximize the performance of model prediction, clinical data can be combined with high dimensional gene expression data, which is also available in

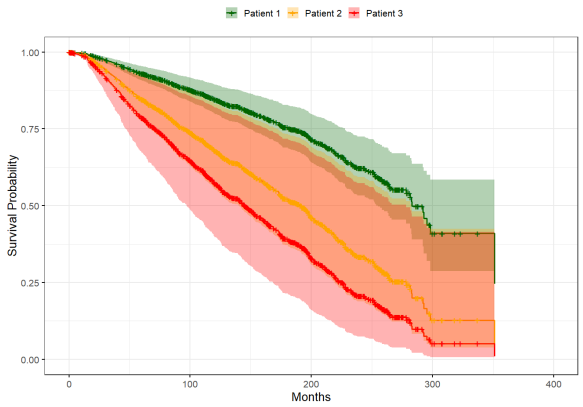

Figure 12: Survival Estimation of Different Patients

METABRIC dataset. The genetic data had to be transposed into columns in order to combine with clinical data by Patient ID. In order to deal with high dimensional gene data, we can select top 500 gene expression with high variance. After this preprocessing, we can apply an Elastic Net or Lasso Cox model to build a predictive survival model that combine both clinical and genomic information.

7 Limitations

The stepwise AIC is a good method for interpretation when we want to do variable selection. However it has several limitations. It can overfit the data because the selection is often unstable, so small changes in the data can give different results. Important variables might be left out just because they don't improve the AIC. Also, the estimated effects of the selected variables can be biased, and standard errors and p-values may not be accurate. Lasso or Elastic-Net are usually better for building reliable cox models. However, in R, the riskRegression library does not support computing the Brier Score for Elastic Net or Lasso Cox models but we can still evaluate model performance using the concordance index instead.

8 Conclusion

In conclusion, the optimized Random Survival Forest was the best model for predicting clinical data. It has the flexibility to capture nonlinear effects and interactions automatically, and typically gives better predictive accuracy (lower Brier score) than the Cox models. The RSF model is also robust to complex relationships in the data. However, Cox models (especially the nonlinear one) remain very useful for interpretation, since hazard ratios and survival curves are easier to explain. Hence, in practice, we could use RSF for prediction and the Cox model for understanding the effect of individual covariates.

References

[1] Curtis, C. and Shah, S. P. and Chin, S.-F. and others, *Breast cancer (metabric) dataset*, Nature 2012; Nat Commun 2016, 2509 samples, 2012-2016. [Online]. Available: `https://www.cbioportal.org/study/summary?id=brca_metabric`.

[2] S. van Buuren and K. Groothuis-Oudshoorn, "Mice: Multivariate imputation by chained equations in r," *Journal of Statistical Software*, vol. 45, no. 3, pp. 1–67, 2011.