Survival Final Project

Nisha Lingam, Siqi Wang, Mia Yu, Kindle Zhang, Sitian Zhou

11/28/2024

1 Background

Breast cancer is the leading cancer worldwide and was diagnosed in an estimated 2.3 million patients in 2020, which is 11.7% of all new cancer cases. These incidence rates are comparatively high in the developed part of the world like North America and Europe; however, incidence rates are shifting to developing world low- and middle-income countries due to changes in lifestyles and reproductive behaviors. There is higher mortality in least-developed regions because screening and treatment options are less available [1, 2]. Statistics show that, in the United States alone, about 12% of women will have been diagnosed with breast cancer at some point in the duration of their lives [3].

Several factors influence the prognosis of breast cancer. Tumor size plays a significant role, as tumors in localized or lower stages are associated with better outcomes. Lymph node involvement also impacts survival, with the presence of cancer in regional lymph nodes linked to lower survival rates. Additionally, receptor status is crucial: ER-positive, PR-positive, and HER2-positive tumors often respond well to treatment with certain targeted agents. In contrast, breast cancer lacking these receptors, known as triple-negative breast cancer, is more dangerous and has limited treatment options [2, 4]. Breast cancer subtypes play a critical role in determining prognosis and treatment options. Luminal A, characterized by ER/PR positivity and HER2 negativity, is slow-growing and has the best survival rates. Luminal B, which is ER-positive and HER2-positive, is more malignant than Luminal A. The HER2-enriched subtype, marked by ER-/PR-/HER2+ status, is aggressive but responds well to HER2-targeted therapies. In contrast, triple-negative breast cancer, lacking all markers, is associated with the worst overall survivability, particularly in its most advanced stages [2, 3].

Treatment modalities for breast cancer encompass both standard and emerging approaches. Standard treatments include surgery, chemotherapy, radiation therapy, hormonal therapies such as tamoxifen, and HER2-directed therapies like trastuzumab [1, 2]. Emerging treatments, such as immunotherapy with checkpoint inhibitors and precision medicine, are advancing care by utilizing genomic data to create personalized treatment plans for each patient [3, 4]. Survival rates of breast cancer vary greatly based on stage at diagnosis and access to care. Early detection leads to a five-year survival rate of nearly 99%, compared to 32% for advanced-stage diagnoses. Disparities persist globally due to inequities in healthcare access, socioeconomic factors, and resource availability [1, 2].

2 Exploratory Data Analysis

This study utilized the Rotterdam dataset from the survival package in R, which includes records of 2,982 primary breast cancer patients from the Rotterdam tumor bank. The dataset consists of 15 variables: patient identifier (pid), year of surgery (year), age at surgery (age), menopausal status (meno: 0 = premenopausal, 1 = postmenopausal), tumor size categorized into three levels (<=20 mm, 20-50 mm, >50 mm), differentiation grade (grade), number of positive lymph nodes (nodes), progesterone receptors (pgr, measured in fmol/l), estrogen receptors (er, measured in fmol/l), hormonal treatment status (hormon: 0 = no, 1 = yes), chemotherapy status (chemo), days to relapse or last follow-up (rtime), relapse status (recur: 0 = no relapse, 1 = relapse), days to death or last follow-up (dtime), and vital status (death: 0 = alive, 1 = dead). Vital status information was obtained from various sources, such as patient visits for other conditions, correspondence, financial transactions, or even social media, raising important considerations about censoring. Death records, however, are often centralized within electronic health records, ensuring accuracy.

As part of the exploratory data analysis (EDA), the dataset's distributions were visualized to gain insights. Base on Figures 1 and 2, the age and nodes variables showed concentrated distributions, with most patients aged between 40 and 60 and having fewer than 10 positive lymph nodes. Both ER and PGR variables were found to be right-skewed, prompting the application of logarithmic transformations to these variables. The recur and death variables exhibited relatively balanced distributions, indicating that the dataset is well-balanced between outcomes. No missing data were identified, as this is a well-curated dataset. A correlation analysis as shown in Figure 3 revealed a positive correlation between age and ER (0.31),

as well as between ER and PGR. These relationships, including the correlation between ER and age, are further reflected in the later analysis.

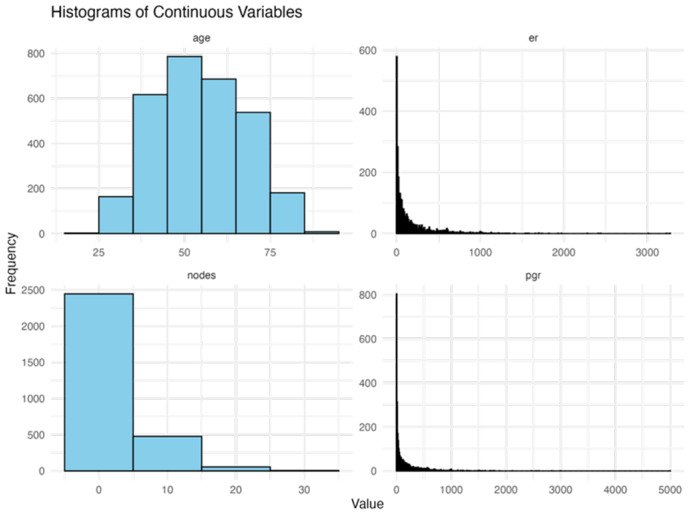

Figure 1: Visualization of Continues Variables

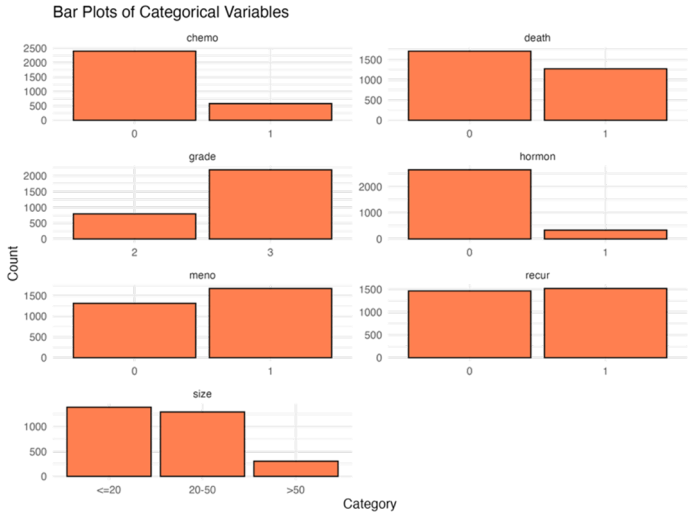

Figure 2: Visualization of Categorical Variables

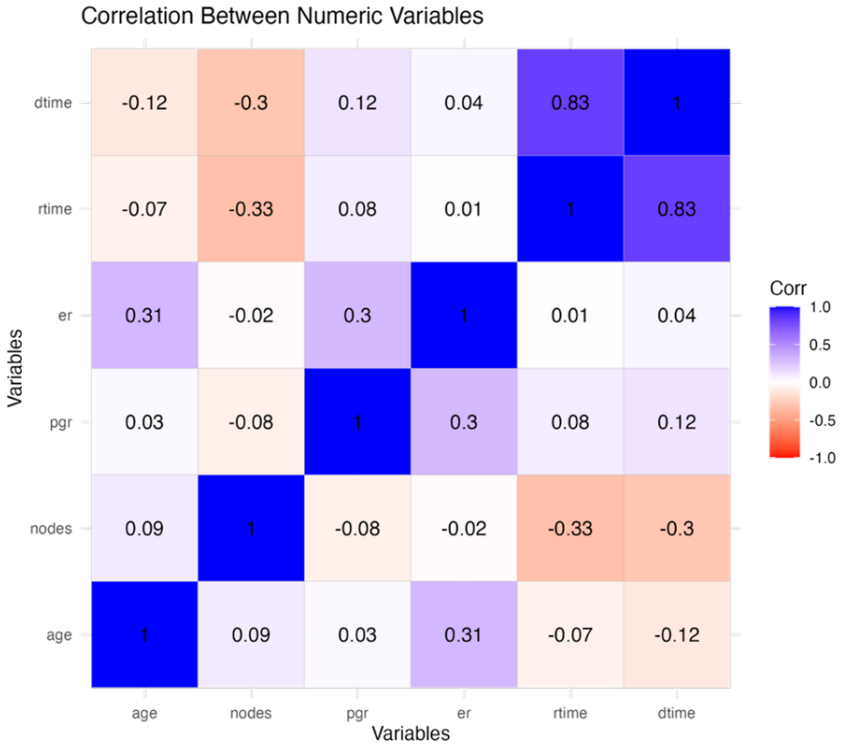

Figure 3: Correlation Plot

3 Methods

3.1 Kaplan-Meier Analysis

3.1.1 Data Source and Preparation

The *Rotterdam breast cancer dataset* was used to analyze *overall survival*, defined as the time from initial observation to death, with censoring applied for individuals alive at the end of follow-up.

- Menopause status (*meno*), tumor size (*size*), and tumor grade (*grade*) were converted to categorical variables. Also,

Chemotherapy treatment (*chemo*) and hormone therapy (*hormo*) were included as covariates for analysis.

- Progesterone receptor (*pgr*) and estrogen receptor (*er*) levels were dichotomized into "High" or "Low" categories based on their median values.

3.1.2 Kaplan-Meier Estimation

The survival time T and event status δ ($\delta = 1$ for death, $\delta = 0$ for censoring) were combined to construct the survival function:

$$S(t) = P(T > t). \tag{1}$$

The survival function $S(t)$ was estimated using the Kaplan-Meier method:

$$\hat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{n_i}\right), \tag{2}$$

where d_i is the number of events at t_i, and n_i is the number at risk just before t_i.

3.1.3 Statistical Comparison

Survival differences across groups were assessed using the log-rank test, comparing observed and expected events over time.

3.2 Cox PH Model

3.2.1 Data Cleaning

To prepare the data for analysis, unnecessary variables were removed. Specifically, **pid**, **rtime**, and **recur** were excluded, as they are not relevant to the outcome of interest. The variable **year**, being a covariate not of interest, was also deleted. Based on the exploratory data analysis (refer to **Figure** 1), **pgr** and **er**, which exhibit extreme right skewness, were log-transformed. Additionally, **dtime** was divided by 30 to facilitate interpretation as a monthly survival rate. The models were trained on a training dataset comprising 70% of the original data.

According to the time ties table (**table** 3), we found that there are some ties in this data, so we used an "efron" method instead of "exact" in the following the Cox model.

3.2.2 Variable Selection

In this part, we used three methods to select the main effect of our model, the first Collett's method, the second stepwise method, and the third Lasso.

3.2.3 Check Interaction

The next step involves using backward selection starting from a full model, focusing on selecting interaction terms. To simplify the problem, we only consider the interactions between the two categorical variables (**size**, **grade**) and the three continuous variables (**pgr**, **nodes**, **age**).

3.2.4 Check Linearity

Next, we verified whether the three continuous variables in the model satisfy the linearity assumption. If they do not, we transformed these variables into other forms to explore which transformation best captures the nonlinear relationship between these variables and the output. Here, we divided the continuous variable into three equal parts to convert it into a categorical variable.

3.2.5 Check Proportional Assumption

In this part, we examined whether the dependent variable satisfies the proportional hazard assumption, with a particular focus on the two categorical variables, **size** and **grade**. We first ploted the log cumulative hazard curves to visually assess whether size and grade violate the proportional hazards (PH) assumption. Next, we analyzed the weighted Schoenfeld residuals to quantitatively evaluate whether size and grade deviate from the PH assumption. Following that, we included a time interaction term in the model and examined whether its coefficient is equal to zero as a formal test. Finally, we revisited the cumulative hazard plots and attempt to address the question from the perspective of a piecewise Cox model.

3.2.6 Influence Diagnostics

To check for potential outliers in the data, an influence diagnosis based on DFBeta was conducted. If a large number of such outliers are identified, we will consider removing these points and refitting the model.

3.3 Parametric models

Parametric models were built to predict breast cancer survival and compare with the Cox Proportional Hazards model. Specifically, Exponential, Weibull, and Log-Normal distributions were used. The models were trained on a training dataset comprising 70% of the original data.

Initially, covariate selection was performed for each of these three models using Likelihood Ratio Tests comparing univariate models for each covariate to the null model. After selecting covariates based on the univariate models, in attempt to reduce the model complexity, chemo and ER were manually removed from models that did not have any nonsignificant covariates based on the LRT univariate model comparisons. Chemo and ER were both nonsignificant covariates which was determined from the Kaplan-Meier curves and logrank tests performed earlier. To further verify the removal of these two covariates, the full models were compared against the trimmed models for each parametric distribution (Full Weibull vs Trimmed Weibull and Full Lognormal vs Trimmed Lognormal) using Likelihood Ratio tests.

After obtaining the trimmed models for the three parametric distributions, the Exponential model was compared with the Weibull model using a Likelihood Ratio Test. The Weibull and Log-Normal models were compared in terms of their -2*log-likelihoods and AIC values and graphically, using their corresponding survival curves as compared to the empirical (Kaplan-Meier) survival curve.

3.4 Validation

Data were split into 70% training data and 30% test data, and Cox PH models and AFT models were fitted using training data only. We conducted cross-validation to assess the predictive performance of the models.

3.4.1 Time-dependent Brier Score

The Brier score measures the average distance between the observed outcomes and the predicted probabilities. In survival analyses, a time-dependent Brier score is commonly used to assess the accuracy of a predicted survival function at a specific time [5]. To account for right-censoring in survival data, Graf et al. and Gerds & Schumacher introduced the inverse probability of censoring weighting (IPCW) technique [6, 7].

Let T_i^* represent the event time and C_i^* denote the censoring time for the i^{th} individual. Let $\pi(t|x_i)$ denote the estimated survival for the i^{th} individual, $G(t|x_i)$ be the survival distributing of the censoring time, and D_i be the event indicator.

For non-censored observations, we have

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}(I(T_i^* > t) - \pi(t|x_i))^2 \tag{3}$$

For right-censored observations, we have

$$BS(t) = \frac{1}{n}\sum_{i=1}^{n}[\frac{\pi(t|x_i)I(T_i^* \leq t, D_i = 1)}{G(T_i^-|x_i)} + \frac{[1-\pi(t|x_i)]^2 I(T_i^* > t)}{G(t|x_i)}], \tag{4}$$

where $G(t^-|x_i) = P(C_i^* \geq t|x_i)$ and $D_i = I(T_i^* \leq C_i^*), i = 1, 2, ..., n$.

The time-dependent Brier score values were computed in R using the **Score()** function in **riskRegression** package.

4 Results & Interpretation

4.1 Kaplan-Meier Analysis

4.1.1 Survival by Tumor Size and Grade

The results of Figures 4 and 5 indicate that tumor **size** and **grade** have significant impacts on recurrence and survival outcomes. Patients with smaller tumors (size <= 20) and lower tumor grades (grade = 1) show longer survival times, while

those with larger tumors (size > 50) and higher grades (grade $= 3$) have shorter survival times. These findings suggest that tumor size and malignancy are critical prognostic factors for breast cancer patients. This aligns with common knowledge, as larger tumors and higher levels of malignancy are typically associated with poorer outcomes.

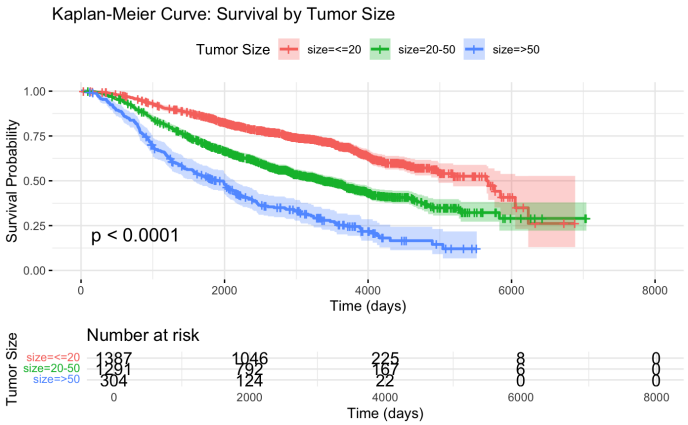

Figure 4: Kaplan-Meier Curve: Survival by Tumor Size

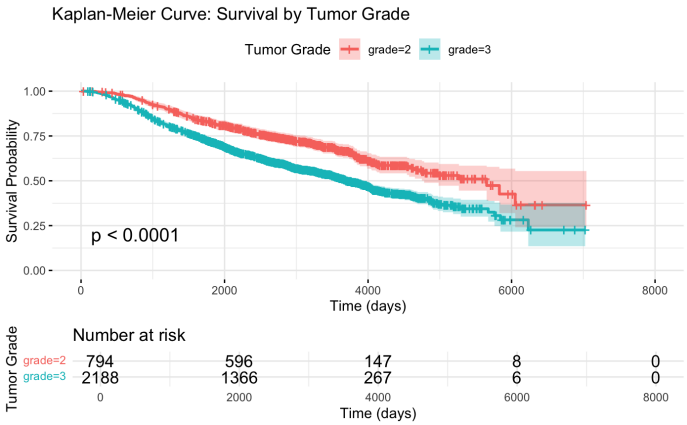

Figure 5: KM Curve: Survival by Tumor Grade

4.1.2 Survival by Menopausal Status

Menopausal status refers to whether a woman has stopped having menstrual periods (postmenopausal) or still has them (premenopausal).

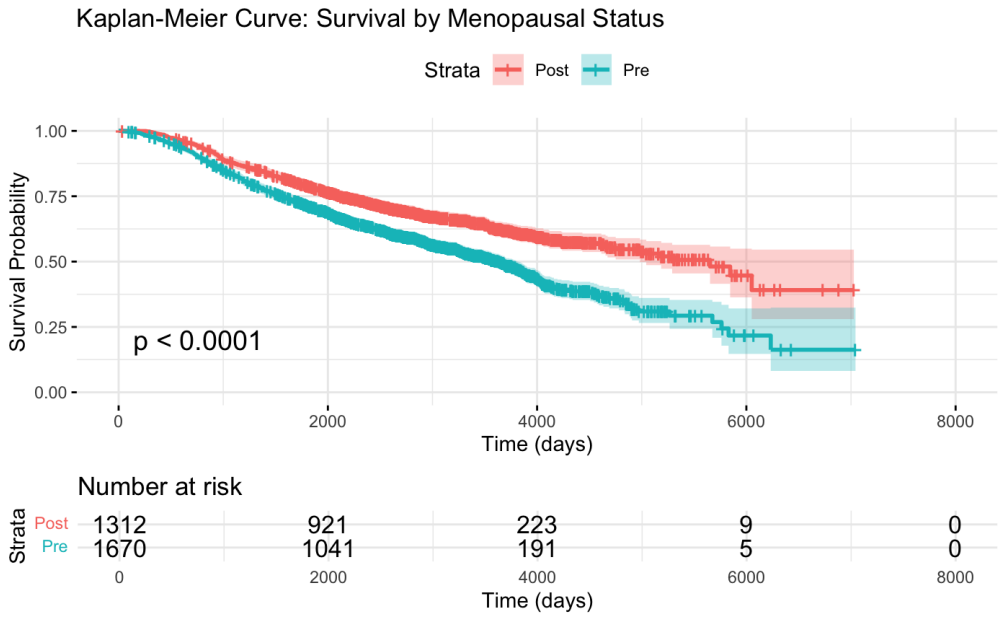

Figure 6: KM Curve: Survival by Menopausal Status

Kaplan-Meier analysis shows a highly significant difference in survival probability between these groups ($p < 0.0001$)(Figure 6). Post-menopausal patients exhibit better overall survival compared to pre-menopausal patients, suggesting that menopausal status might be an important factor influencing survival outcomes.

4.1.3 Survival by PGR and ER Level

PGR Level refers to the level of progesterone receptor expression in tumor cells, which can influence tumor growth. While **ER Level** refers to the level of estrogen receptor expression in tumor cells, another factor influencing tumor growth. To investigate the survival of these two continuous variables, we divided them into two groups, high and low, according to the median.

The results of Figures 7 show that the level of PGR (progesterone receptor) significantly affects overall survival, with p-values

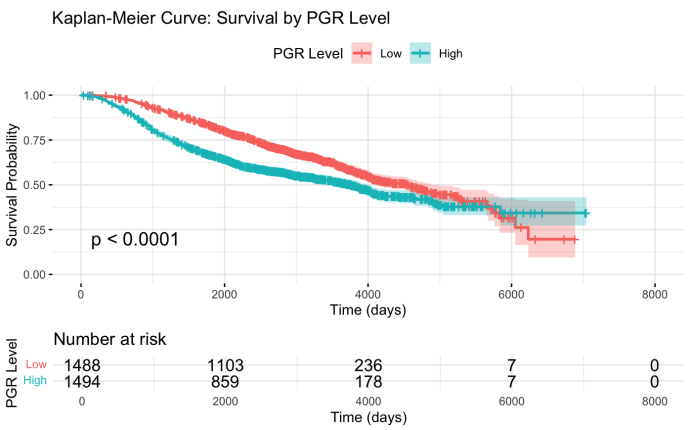

Figure 7: KM Curve: Survival by PGR Level

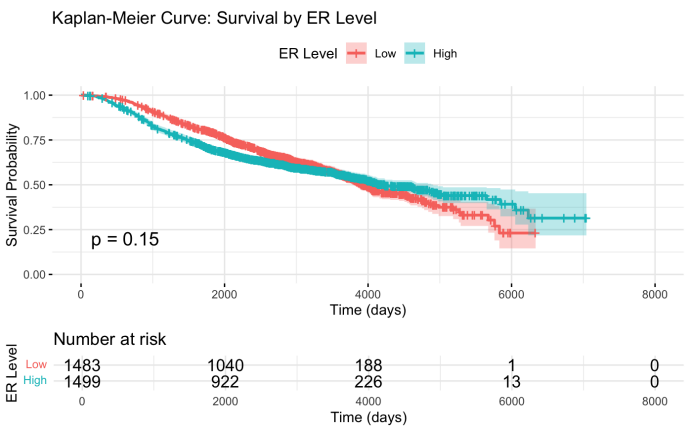

Figure 8: KM Curve: Survival by ER Level

less than 0.0001. Patients with high PGR levels consistently demonstrate better survival outcomes compared to those with low PGR levels. These findings potentially reflect better responsiveness to hormone-related therapies or less aggressive tumor behavior.

Figure 8 shows ER (estrogen receptor) level has no statistically significant impact on overall survival ($p = 0.15$). These findings imply that ER level may not be a decisive prognostic factor for survival in this dataset, though its clinical relevance could warrant further investigation in larger or more specific cohorts.

4.1.4 Survival by Chemotherapy and Hormone Treatment

Chemotherapy means a treatment that uses drugs to destroy or slow the growth of cancer cells, often used for advanced or aggressive cancers.

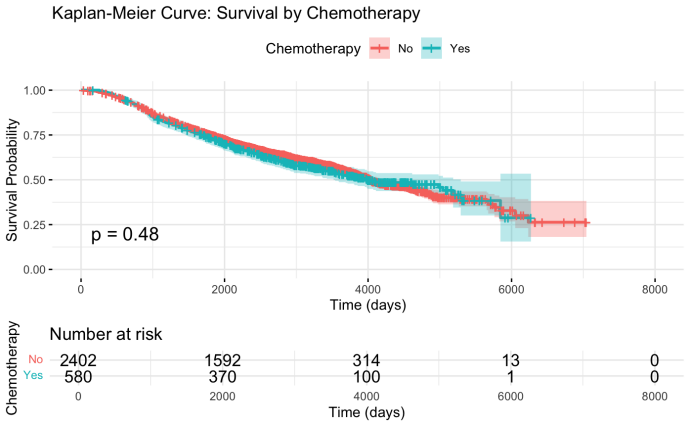

Figure 9: KM Curve: Survival by Chemotherapy

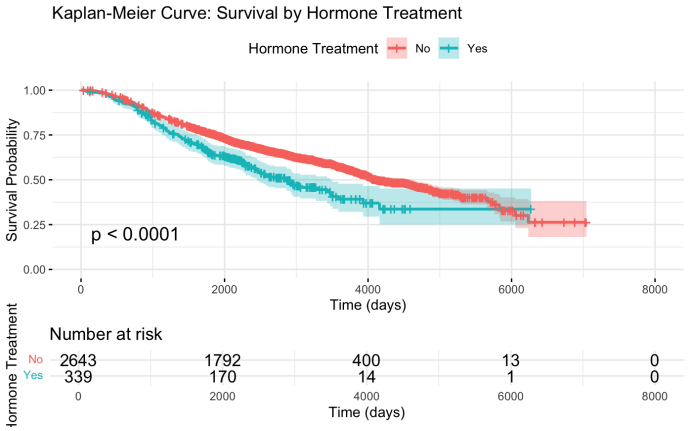

Figure 10: KM Curve: Survival by Hormone Treatment

Chemotherapy	Observed	Expected	Chisq	Pvalue
No	1014	1023.93	0.49	0.482
Yes	258	248.07	0.49	0.482

Table 1: Log-rank Test Results for Chemo Groups

The results of Table 1 and Figure 9 show that for overall survival, there is no statistically significant difference between the chemotherapy and non-chemotherapy groups ($p = 0.48$), as indicated by both the survival curve and the results of the logarithmic rank test. This suggests that the effect on overall survival on chemotherapy within this dataset is not evident.

Hormone Treatment refers to a therapy that uses hormones or hormone-blocking drugs to treat certain cancers, especially hormone-sensitive ones.

HormonTreatment	Observed	Expected	Chisq	Pvalue
No	1113	1161.63	23.69	1.1e-06
Yes	159	110.37	23.69	1.1e-06

Table 2: Log-rank Test Results for Hormon Groups

The results of Table2 and Figure 10 indicate that hormone therapy has a significant impact on overall survival. Patients receiving hormone therapy show significantly higher survival probabilities ($p < 0.0001$) compared to those who did not receive hormone therapy. The logarithmic ranking test further confirms these differences, highlighting a strong association between hormone therapy and improved survival outcomes.

4.1.5 Survival by Combined Analysis of Chemo and Hormone Treatment

To examine the survival impact of chemotherapy and hormone therapy together, rather than individually, we conducted Kaplan-Meier analysis on both two variables.

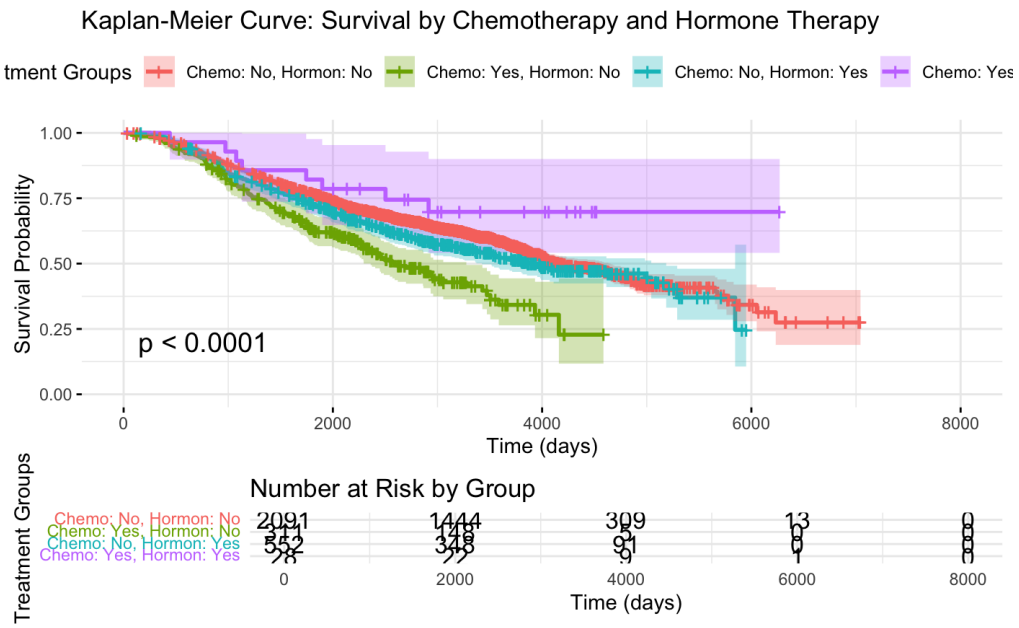

Figure 11: KM Curve: Survival by Chemotherapy and Hormone Therapy

Patients receiving both treatments (purple group) show the highest survival probabilities, while those receiving no treatment (red group) have the lowest. The log-rank test ($p < 0.0001$) confirms significant differences between the treatment groups. Combined therapy appears to offer the greatest survival benefit, highlighting the importance of using these treatments together for certain patients.

4.2 Cox PH Models

4.2.1 Data Cleaning

Value	1	2	3	4
Frequency	1367	279	46	6

Table 3: Time ties Frequency Table

4.2.2 Collett's Method Selection

According to the **Table** 4, We could identify all the relevant models required for Collett's method along with their corresponding -2 log-likelihood and AIC values. In the first step, by comparing model 1 and model 2 to model 10, we could find variable **er** and **chemo** cannot make -2 log-likelihood decreases more than 1.643 (a chi-squared test statistics with 1 df and

num	type	terms_in_model	-2LogL	AIC(alpha = 2)
1	null	1	12824.48	12824.48
2	univariate	age	12768.57	12770.57
3		meno	12782.17	12784.17
4		size	12651.65	12655.65
5		grade	12793.06	12795.06
6		nodes	12586.68	12588.68
7		pgr	12781.42	12783.42
8		er	12822.22	12824.22
9		hormon	12811.10	12813.10
10		chemo	12823.98	12825.98
11	6-variables	meno + size + grade + nodes + pgr + hormon	12445.52	12459.52
12		age + size + grade + nodes + pgr + hormon	12434.07	12448.07
13		age + meno + grade + nodes + pgr + hormon	12494.27	12506.27
14		age + meno + size + nodes + pgr + hormon	12442.36	12456.36
15		age + meno + size + grade + pgr + hormon	12572.77	12586.77
16		age + meno + size + grade + nodes + hormon	12457.46	12471.46
17		age + meno + size + grade + nodes + pgr	12435.48	12449.48
18	7-variables	age + meno + size + grade + nodes + pgr + hormon	12433.97	12449.97
19	5-variables	age + size + grade + nodes + pgr	12435.51	12447.51
20	6-variables	age + size + grade + nodes + pgr + er	12434.99	12448.99
21		age + size + grade + nodes + pgr + chemo	12434.05	12448.05

Table 4: Model Comparison Table

0.2 significance level). In the next step, we constructed a model with all useful variables in step 1(model 18), and compared it with model 11 to model 17. We found variable **meno** and **hormon** can't make -2 log-likelihood decreases more than 2.706 (a chi-squared test statistics with 1 df and 0.1 significance level). Finally, we put the useless variables in step 1 back to check whether the model gets better. Comparing model 19 with model 20 and model 21 separately, we found the model 19 is the best, so up till now, we found our interim optimal model, model 19.

4.2.3 Stepwise

As shown in **Figure** 23, the stepwise selection approach gave us the same result.

4.2.4 Lasso

Using the Lasso method, we found that when the number of variables decreased from 6 to 5, the partial likelihood deviance increased only slightly. Therefore, we chose to use the model with 5 variables (**Figure** 12). The λ is 0.015478942, and we used it to find the optimal model. The result was the same as the previous two methods.

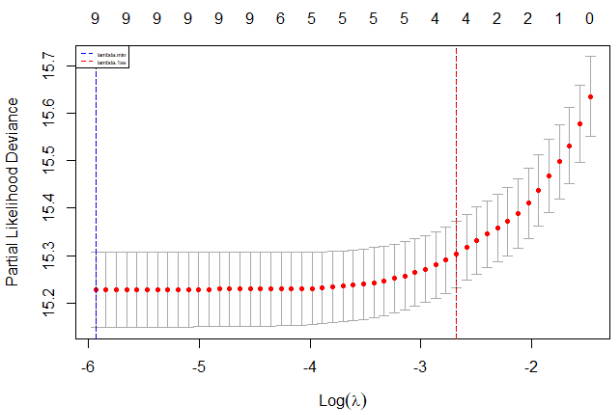

Figure 12: Partial likelihood deviance vs log(λ)

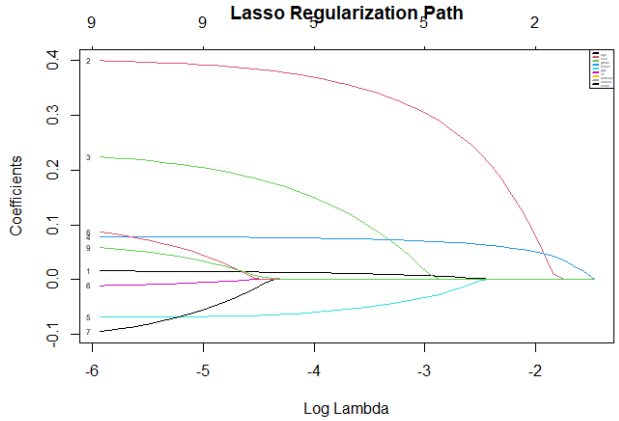

Figure 13: coefficient trajectory along log(λ)

So, we can retain 5 variables in our model so far. **age**, **size**, log(λ), log(λ) and log(λ).

4.2.5 Check Interaction

The **Figure** 28 showed the final step of the procedure. Based on the p-value from the likelihood ratio test (LRT), the model with the **size** and **nodes** interaction term was better. Wald chi-squared test had the same result.

4.2.6 Check Linearity

Here are the log-likelihood and AIC values of different models obtained after transforming the three continuous variables: **Table** 5, **Table** 6, and **Table** 7.

Type of Model	**Terms in Model**	**-2LogL**	**AIC**
Null Model	size + grade + age + pgr	12573	12583
Continuous	size * nodes + grade + age + pgr	12428	12444
Sqrt Scale	size * sqrt(nodes) + grade + age + pgr	12386	12402
Categorical	size * nodes_cat + grade + age + pgr	12451	12467
Quadratic	size * (nodes + nodes^2) + grade + age + pgr	12387	12409
Log	size * log(nodes) + grade + age + pgr	**12382**	**12398**

Table 5: Model Comparison Based on -2LogL and AIC for nodes

Type of Model	**Terms in Model**	**-2LogL**	**AIC**
Null Model	size * nodes + grade + pgr	12414	12428
Continuous	size * nodes + grade + age + pgr	12382	12398
Sqrt Scale	size * nodes + grade + sqrtage + pgr	12386	12402
Ordinal	size * nodes + grade + agecat + pgr	12389	12405
Categorical	size * nodes + grade + age48_62 + age_up62 + pgr	12385	12403
Quadratic	size * nodes + grade + age + agessquare + pgr	**12357**	**12375**

Table 6: Model Comparison Based on -2LogL and AIC for age

Type of Model	**Terms in Model**	**-2LogL**	**AIC**
Null Model	size * nodes + grade + age + agessquare	12385	12401
Continuous	size * nodes + grade + age + agessquare + pgr	**12357**	**12375**
Ordinal	size * nodes + grade + age + agessquare + pgrcat	12357	12375
Categorical	size * nodes + grade + age + agessquare + pgr_med + ...	12368	12386

Table 7: Model Comparison Based on -2LogL and AIC for pgr

According to the results, we made a log transform for variable **nodes** and added a quadratic **age** in the model. There was nothing to do with variable **pgr**.

However, it was worth noting that after applying a log transformation to the nodes variable, the interaction term in the original model was no longer significant. Based on the results of the likelihood ratio test (LRT) and the Wald chi-squared test, we should remove the interaction term from the model. (the P-value for LRT was 0.4348 and for Wald test was 0.44 which were both larger than 0.05)

4.2.7 Check Proportional Assumption

In this part, we examined whether the dependent variable satisfies the proportional hazard assumption, with a particular focus on the two categorical variables, **size** and **grade**.

Plot a log(-logS(t)) versus log(t) First, to visually assess whether the variables size and grade violate the proportional hazards (PH) assumption, we plotted the log cumulative hazard curves against log time for different subgroups of grade and size (**Figures** 14 and 15).

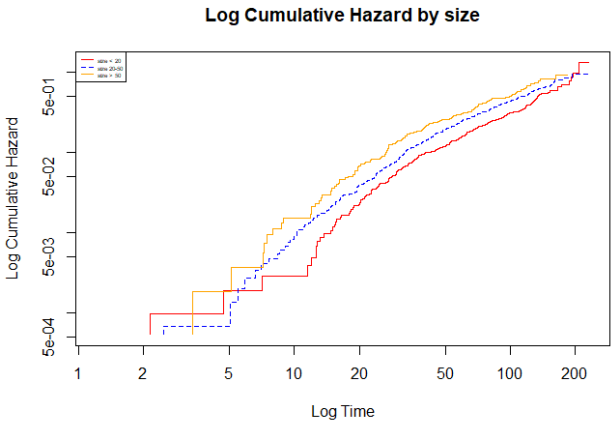

Figure 14: log cumulative hazard by size

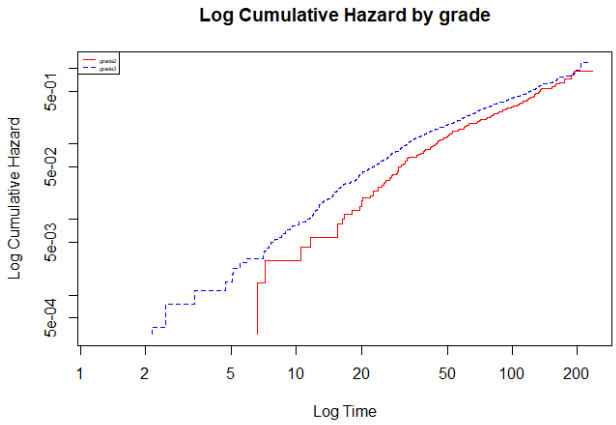

Figure 15: log cumulative hazard by grade

From the plot, it could be observed that the curves crossed at the beginning and the end of the time period, while they appeared approximately parallel in the middle. This suggested that, overall, grade and size may violate the proportional hazards (PH) assumption. To further confirm this, we need to conduct statistical tests.

Weighted Schoenfeld residuals By calculating the Weighted Schoenfeld residuals and performing residual analysis, we obtained the following results (**Table** 8):

Variable	Chisq	df	p-value
size	7.02	2	0.02986
nodes	11.06	1	0.00088
grade	4.62	1	0.03163
age	14.84	1	0.00012
pgr	65.88	1	4.8e-16
agesquare	15.87	1	6.8e-05
GLOBAL	92.57	7	<2e-16

Table 8: Schoenfeld Residuals Analysis

The p-values for the **size**and **grade** variables are both less than 0.05, indicating that these two variables do violate the proportional hazards (PH) assumption. To better visualize how the coefficients of these variables change over time, we plotted the coefficients against time (**Figures** 16 and 17).

Interestingly, although both variables violated the PH assumption, the two solid lines in the plot could approximately be regarded as horizontal, indicating that their influence over time is minimal. To further quantified this effect, we decided to introduce a time-dependent variable in the next part.

Time Dependent Variable We added an interaction term between size and time, and grade and time, respectively, into the model. Using the Wald test, we evaluated whether the coefficients of these interaction terms are zero. The results for the two tests are shown in **Figures** 24 and 25.

From the results, we could see that the p-values for *grade:time* and *size:time* are 0.0429 and 0.0082, respectively, both of which are less than 0.05. This indicated that we have sufficient evidence to reject the null hypothesis and conclude that the coefficients of these two interaction terms are not equal to zero. In other words, we should retain these interaction terms in the model.

However, considering that the hazard ratios for both variables are 0.996, which is very close to 1, the effects of these interactions on the survival rate are minimal. To simplify the interpretation of the model, we decided not to retain these interaction terms.

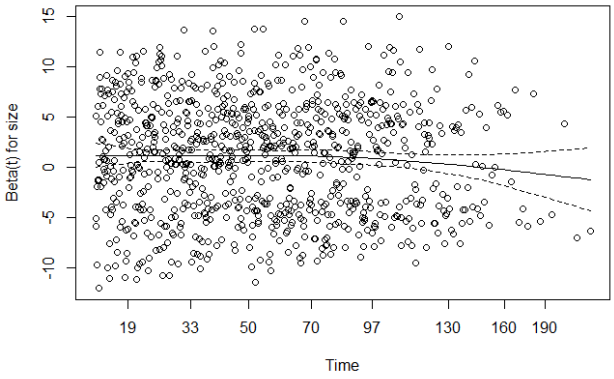

Figure 16: beta for size

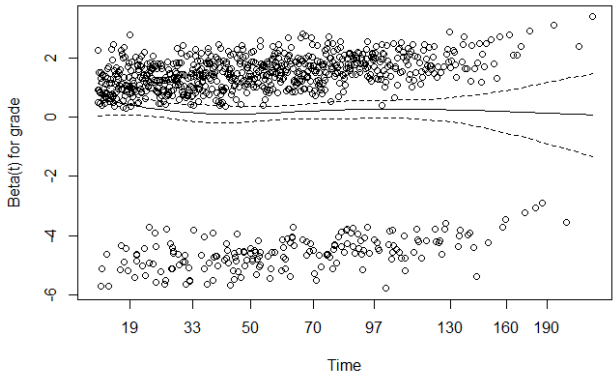

Figure 17: beta for grade

4.2.8 Influence Diagnostics

At the same time, to identify outliers in the data, we performed influence diagnostics using the DFBeta metric to identify the points that have the greatest influence on the coefficients. Upon inspection, we found that such points are very few. Taking the variable **size** as an example, the plot showed that only four points fall outside the normal range for the influence on the **size** coefficient. The patient IDs for these four points are 683, 1756, 1762, and 1962. Therefore, it is not necessary to remove these points, as their number is small and their overall impact is minimal (**Figure** 18).

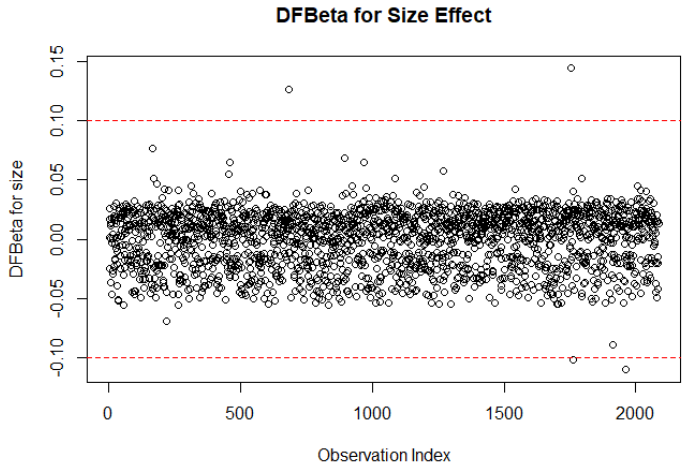

Figure 18: DFBeta for size vs observation index

4.2.9 Extension of Cox PH Model

From the log cumulative hazard along log time plot, we observed that the curves intersect at the beginning and end of the time period. This naturally led us to hypothesize whether size and grade follow the PH assumption within three separate time intervals. If we convert the time variable into a categorical variable and reconsider the interaction terms between the time variable and size or grade, we may achieve a better model for describing their impact on the survival rate.

To test this hypothesis, we plotted the cumulative hazard along time (in months) curves under different levels of size and grade. Using **Figure** 19 and **Figure** 20, we aimed to identify two time-axis cutoff points for segmenting the time variable.

We found that 50 and 150 months are two suitable cutoff points. Before 50 months, the cumulative hazard curves diverge rapidly. Between 50 and 150 months, the curves for different groups were approximately parallel and grew at a similar rate. In the later period, the curves quickly converged and intersected.

We divided the time variable into three levels based on the cutoff points at 50 and 150 months, converting it into a categorical variable. Then, we introduced interaction terms between this time variable and size or grade into the piecewise cox model

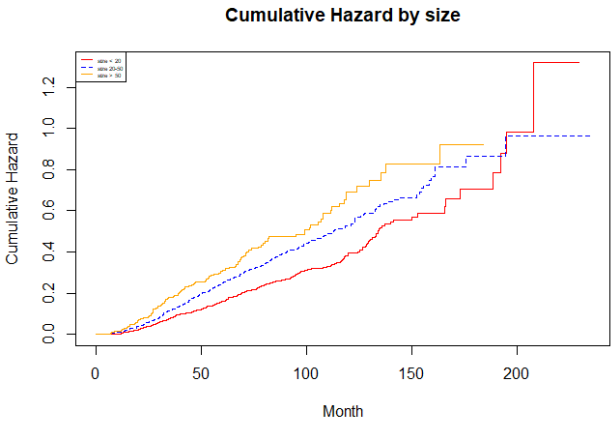

Figure 19: cumulative hazard for size along time

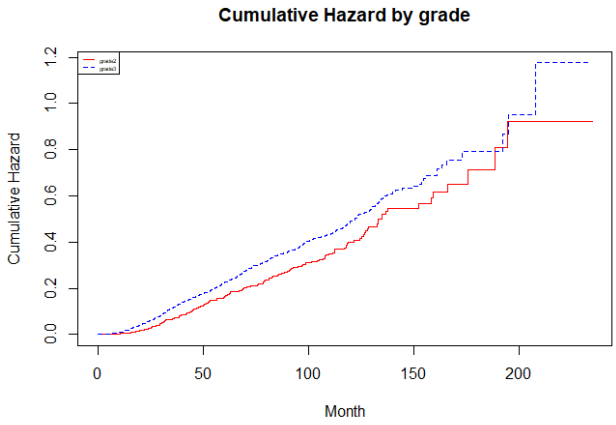

Figure 20: cumulative hazard for grade along time

and performed regression analysis again. The results are showed in **Figure** 26 and **Figure** 27

According to the result, we found both of these two models perform better than the models simply including a time*size or time*grade interaction which we got from time-dependent part. The AIC for size piecewise Cox model is 11444.977 which is smaller than 12367.882, the AIC for a Cox model with size*time. Similarly, the AIC for the piecewise model by grade is 11156.914 but 12370.993 for the Cox model with grade*time.

At the same time, after adjusted by other factors, we calculated the Hazard Ratios (HR) for different levels of size or grade across various time intervals based on the coefficients of the interaction terms. The results are presented in the following two tables.**Table** 9 and **Table** 10

HR	**Beg (<20)**	**Med (20 50)**	**End (>50)**
<20	1	1	1
20 50	28.285	0.990	1.1314
>50	30.310	1.2124	0.091

Table 9: HR Across Different Time Periods for Size

HR	**Beg**	**Mid**	**End**
grade2	1	1	1
grade3	48.587	0.8746	0.1458

Table 10: HR Across Different Time Periods for Grade

Limitation However, it is worth noting that we can observe some discrepancies between the HR results and our expectations. While it aligns with our expectations that patients with higher tumor size and grade have a greater risk of death in the early stages, the influence of size and grade on increasing mortality decreases rapidly over time. Surprisingly, it even resulted in the opposite conclusion in later stages. This indicated that our model still requires further refinement.

We believe the main issues may stem from the following aspects:

1. Piecewise Cox Model: If the piecewise Cox model holds, there might be better choices for the time cut-off points, or it might be necessary to consider more than three time intervals.

2. Interaction Between Size/Grade and Time: The interaction between size or grade and time might be better explained after applying transformations to the time variable. For instance: If we hypothesize that the impact of the interaction between size and time on the survival rate increases with time, we could try applying a logarithmic transformation to the time variable. (log(t+1)) Conversely, if we believe the impact decreases with time, we could consider transformations like exp(-t) to better capture this relationship.

Further exploration and testing of these alternatives could help improve the model's interpretability and performance.

4.2.10 Final Cox PH Model and Interpretation

We have selected the Cox PH model as our final model, which is presented below. The estimated coefficents, hazard ratios, and p-values are shown in Table 15.

$$h_i(t, Z_i) = h_0(t) \exp\big(0.3069 \cdot I(\text{size20} \sim 50) + 0.5681 \cdot I(\text{size} > 50) + 0.5244 \cdot \log(\text{nodes} + 1) + 0.2360 \cdot I(\text{grade} = 3) + (-0.0839) \cdot \text{age} + 0.0009 \cdot \text{age}^2 + (-0.0793) \cdot \log(\text{pgr} + 1)\big)$$

From the model, we observed that an increased tumor size is associated with a higher hazard ratio. Specifically, the hazard ratio for a tumor size of 20-50 mm compared to a tumor size of $\leq$ 20 mm is 1.36. This indicates that individuals with a tumor size between 20 mm and 50 mm have a 36% higher risk of death compared to those with a tumor size less than 20 mm, holding other covariates constant. Furthermore, the risk of death increases by 76.5% for individuals with a tumor size greater than 50 mm compared to those with a tumor size less than 20 mm. Moreover, a one-unit increase in log(node+1) results in a hazard ratio of 1.69. For individuals with a tumor differentiation grade of 3, the hazard ratio is 1.27, suggesting a 27% higher risk of death compared to the reference group with a tumor differentiation grade of 2, keeping other factors fixed. Other factors such as age and progesterone receptor concentration also play a role in the survival of breast cancer patients. One-year increase in age is associated with an 8% decreases in the risk of death, and one-unit (fmol/l) increase in progesterone receptor concentration decreases the risk of death by 7.6%.

4.3 Parametric Models

4.3.1 Covariate Selection

Covariate	**Age**	**Meno**	**Size**	**Grade**	**Nodes**	**Pgr**	**Er**	**Hormon**	**Chemo**
Exponential P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.468	0.001	0.469
Weibull P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Log-Normal P-value	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000

Table 11: P-values of LRT comparing 1-covariate models with null model

All covariates ended up being significant in the Weibull and Log-Normal models based on the LRTs performed with the univariate models (p-values < 0.05), which shows the necessity of the manual covariate removal step.

The Likelihood Ratio tests comparing the full models versus the trimmed models resulted in the failure to reject the null hypothesis that the full models have a better fit than the models with ER and chemo removed. Additionally, the -2*log-likelihood values of these trimmed models have not changed significantly from the full models, only experiencing an increase of 1 to 2 points.

4.3.2 Model Comparisons

Terms	**Resid. Df**	**-2*LL**	**Test**	**Df**	**Deviance**	**Pr(>Chi)**
age + meno + size + grade + nodes + pgr + hormon	2078	17111.44	NA	NA	NA	NA
age + meno + size + grade + nodes + pgr + hormon	2077	17006.08	=	1	105.3629	0.0000

Table 12: ANOVA Results: Comparison between Exponential and Weibull Models

Model	**Log-Likelihood**	**AIC**
Weibull	-8503.041	17026.08
Log-Normal	-8469.002	16958.00

Table 13: Weibull vs Log-Normal Models

We reject the null hypothesis that the less complex, Weibull model better fits the data compared to the Exponential model (p-value < 0.05).

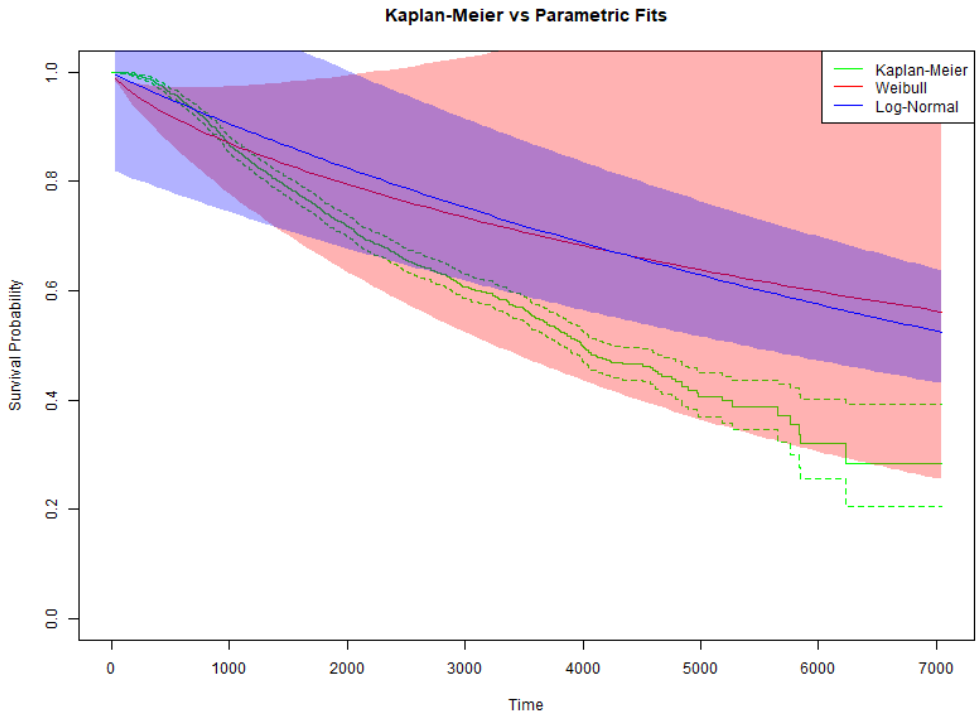

Figure 21: Kaplan-Meier vs Exponential vs Weibull Survival Curves

Comparing -2*log-likelihood and AIC values, the Log-Normal model appears to have a better fit than the Weibull model to the data, especially when considering the qualitative shape of the Kaplan-Meier (empirical) curve, the Log-Normal survival curve seems to match most closely, overall.

4.3.3 Final Log-Normal Model

$$S(t) = 1 - \Phi\left(\frac{\begin{array}{r}\log(t) - (9.31549 + (-0.0078424) \cdot \text{age} + (-0.1066110) \cdot \text{meno1} + (-0.6341278) \cdot \text{size}{>}50 \\ +(-0.3436372) \cdot \text{size20-50} + (-0.2499225) \cdot \text{grade3} + (-0.0773849) \cdot \text{nodes} \\ +(0.0003652) \cdot \text{pgr} + (0.1780863) \cdot \text{hormon1})\end{array}}{1.0409}\right)$$

Each additional year of age decreases the log survival time by 0.0078424. Being postmenopausal decreases the log survival time by 0.1066110. Having a tumor size of 20-50 decreases log survival time by 0.3436372. Having a tumor size > 50 decreases log survival time by 0.634. Having a differentiation grade of 3 decreases log survival time by 0.250. Each additional node decreases log survival time by 0.077. For every unit (fmol/l) increase of progesterone receptors, log survival time increases by 0.000365. Being in the hormonal treatment group increases log survival time by 0.0402010. The estimated coefficents are also shown in Table 16.

4.4 Validation

4.4.1 AIC Comparison

Table 14: Comparison of model fit

Model	Weibull	Log-normal	Cox PH
-2log(L)	17111.44	16938.00	12359.07
AIC	17129.44	16958.00	12373.07

Among the models we fitted, the Cox PH model has the lowest AIC value, indicating a better fit to the data compared to the others.

4.4.2 Time-dependent Brier Score

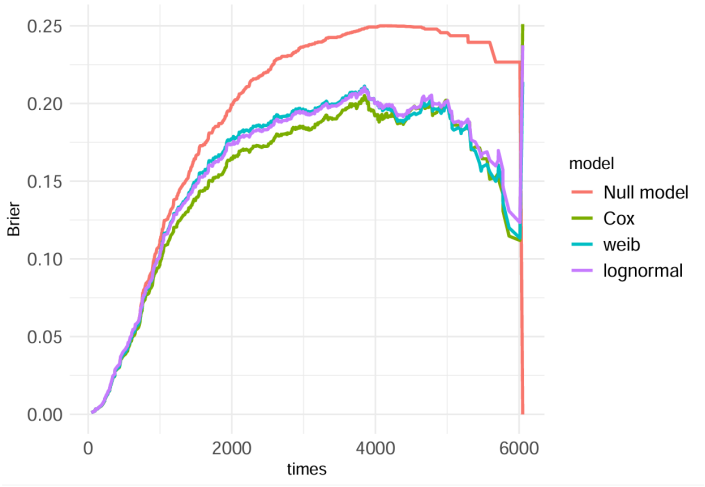

Figure 22: Time-dependent Brier score curves

The Brier score curves plotted in Figure 22 show that all models tend to make less accurate predictions as time progresses. Notably, the Cox PH model's curve is slightly lower than the others, suggesting it provides the best predictive performance among the models.

5 Conclusion

In this project, we utilized Kaplan-Meier analysis and developed various models to explore treatment effects and predict survival probabilities for breast cancer patients whose records were included in the Rotterdam tumor bank.

The KM analysis highlights several key factors influencing survival outcomes. Menopausal status significantly impacts survival, with post-menopausal patients showing better overall survival. Furthermore, higher PGR levels are strongly associated with improved survival. Hormone therapy significantly improves survival, and combining chemotherapy and hormone therapy provides the highest survival probabilities, emphasizing their synergistic benefit.

The Cox PH model suggests that tumor characteristics and the number of positive lymph nodes were significantly associated with survival probabilities among breast cancer patients. Individuals with a tumor size between 20 mm and 50 mm have a 36% higher risk of death and individuals with a tumor size above 50 mm have a 76.5% higher risk of death compared to those with a tumor size less than 20 mm, holding other covariates constant. Poor tumor differentiation, grade 3 vs. grade 2, further increased the risk by 27%. Similar findings were obtained from the log-normal model, with larger tumor sizes, poorer differentiation, and more positive lymph nodes being associated with shortened survival time.

6 Discussion

In this study, we identified several limitations in our analysis. When grouping the patients by treatments (no treatment, hormonal treatment, chemotherapy, hormonal treatment and chemotherapy combined), we observed an imbalance in the data across the treatment groups. Only 28 patients received both treatments, while several hundred patients received either hormonal treatment or chemotherapy, and about 2000 patients received neither. As a result, the survival probabilities calculated for the combined treatment group are less reliable. More data are needed to accurately assess the survival outcomes for the combined treatment group, and further research is required to better understand the effects of combined therapies.

For the Cox PH model, the model diagnostic results suggest that the proportional hazard assumption might be violated. Additionally, we observed a decline in model performance over time, as indicated by the Brier score curves. This may be due to the violation of the proportional hazards assumption and suggests the presence of time-dependent variables. To address time-varying covariates, as discussed in **section 4.2.9**, we built piecewise Cox PH models. However, the results obtained from those models are inconsistent with our previous findings, suggesting that the time cutoff we specified for the model is

not optimal or that other unaccounted factors are influencing the outcomes. This highlights the need for further exploration of time-dependent variables in breast cancer research.

7 References

[1] World Health Organization. *Global breast cancer report*. en. https://www.who.int/news-room/fact-sheets/detail/breast-cancer. Accessed: 2024-12-13.

[2] CDC. *Breast cancer basics*. en. https://www.cdc.gov/breast-cancer/about/index.html. Accessed: 2024-12-13. Nov. 2024.

[3] National Cancer Institute. *Breast cancer statistics and trends*. en. https://seer.cancer.gov/statfacts/html/breast.html. Accessed: 2024-12-13.

[4] Johns Hopkins Pathology. *Staging and prognostic factors*. en. https://pathology.jhu.edu/breast/staging-grade/. Accessed: 2024-12-13.

[5] Håvard Kvamme and Ørnulf Borgan. "The Brier score under administrative censoring: Problems and solutions". In: (2019). eprint: 1912.08581 (stat.ML).

[6] E Graf et al. "Assessment and comparison of prognostic classification schemes for survival data". In: *Stat. Med.* 18.17-18 (1999), pp. 2529–2545.

[7] Thomas A Gerds and Martin Schumacher. "Consistent estimation of the expected Brier score in general survival models with right-censored event times". In: *Biom. J.* 48.6 (Dec. 2006), pp. 1029–1040.

8 Appendix

```
Step:  AIC=12447.51
Surv(dtime, death) ~ age + size + grade + nodes + pgr

         Df   AIC
<none>      12448
+ chemo   1 12448
+ hormon  1 12448
+ er      1 12449
+ meno    1 12450
- grade   1 12454
- pgr     1 12470
- age     1 12481
- size    2 12503
- nodes   1 12584
```

Figure 23: the last step in stepwise

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
size_cat	2	1	0.30665	0.07845	15.2780	<.0001	1.359	size_cat 2
size_cat	3	1	0.56603	0.10975	26.6004	<.0001	1.761	size_cat 3
nodes		1	0.52506	0.03666	205.1408	<.0001	1.691	
grade	2	1	0.50715	0.16117	9.9016	0.0017	1.661	grade 2
age		1	-0.08498	0.01902	19.9561	<.0001	0.919	
pgr		1	-0.07956	0.01521	27.3543	<.0001	0.924	
agesqure		1	0.0008786	0.0001667	27.7747	<.0001	1.001	
gradetime		1	-0.00427	0.00211	4.0999	0.0429	0.996	

Figure 24: result with grade*time

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
size_cat	2	1	0.54747	0.12177	20.2148	<.0001	1.729	size_cat 2
size_cat	3	1	0.98697	0.19172	26.5006	<.0001	2.683	size_cat 3
nodes		1	0.52211	0.03670	202.4414	<.0001	1.686	
grade	2	1	0.23827	0.08404	8.0389	0.0046	1.269	grade 2
age		1	-0.08488	0.01906	19.8313	<.0001	0.919	
pgr		1	-0.07976	0.01522	27.4551	<.0001	0.923	
agesqure		1	0.0008763	0.0001670	27.5502	<.0001	1.001	
sizetime		1	-0.00374	0.00141	6.9953	0.0082	0.996	

Figure 25: result with size*time

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
size_cat	2	1	3.34235	0.13971	572.3012	<.0001	28.285	size_cat 2
size_cat	3	1	3.41146	0.16658	419.3831	<.0001	30.310	size_cat 3
nodes		1	0.37942	0.03837	97.7576	<.0001	1.461	
grade	2	1	0.26959	0.08435	10.2157	0.0014	1.309	grade 2
age		1	-0.06010	0.01939	9.6046	0.0019	0.942	
pgr		1	-0.04916	0.01545	10.1281	0.0015	0.952	
agesqure		1	0.0006417	0.0001697	14.3048	0.0002	1.001	
sizemid_timemid		1	-3.34326	0.13836	583.8930	<.0001	0.035	
sizemid_timehigh		1	-5.20054	0.34940	221.5370	<.0001	0.006	
sizehigh_timemid		1	-3.22463	0.19670	268.7404	<.0001	0.040	
sizehigh_timehigh		1	-5.67515	1.01459	31.2880	<.0001	0.003	

Figure 26: result of piecewise cox model by size

Analysis of Maximum Likelihood Estimates								
Parameter		DF	Parameter Estimate	Standard Error	Chi-Square	Pr > ChiSq	Hazard Ratio	Label
size_cat	2	1	0.19559	0.07984	6.0005	0.0143	1.216	size_cat 2
size_cat	3	1	0.31028	0.11160	7.7303	0.0054	1.364	size_cat 3
nodes		1	0.36895	0.03770	95.7738	<.0001	1.446	
grade	2	1	3.88335	0.16055	585.0166	<.0001	48.587	grade 2
age		1	-0.07533	0.01908	15.5899	<.0001	0.927	
pgr		1	-0.05305	0.01537	11.9169	0.0006	0.948	
agesqure		1	0.0007984	0.0001672	22.8026	<.0001	1.001	
grade3_timemid		1	-3.99924	0.14817	728.4798	<.0001	0.018	
grade3_timehigh		1	-5.86603	0.32108	333.7851	<.0001	0.003	

Figure 27: result of piecewise cox model by grade

Backward selection

```
                                                    Wald test:
                                                    ----------

                                                    Chi-squared test:
                                                    X2 = 8.3, df = 2, P(> X2) = 0.016
Call:
coxph(formula = Surv(dtime, death) ~ size * nodes + grade + age + pgr,
      data = training_data, ties = "efron")

                                        coef exp(coef)  se(coef)      z        p
size20-50                           0.527998  1.695534  0.089015  5.932 3.00e-09
size>50                             0.861344  2.366339  0.144352  5.967 2.42e-09
nodes                               0.116742  1.123829  0.014684  7.950 1.86e-15
gradeGrade 3: Poor Differentiation  0.228449  1.256650  0.084133  2.715  0.00662
age                                 0.015618  1.015741  0.002649  5.895 3.74e-09
pgr                                -0.075736  0.927061  0.015259 -4.963 6.93e-07
size20-50:nodes                    -0.047142  0.953952  0.016451 -2.866  0.00416
size>50:nodes                      -0.034341  0.966242  0.019563 -1.755  0.07920

Likelihood ratio test=396.6  on 8 df, p=< 2.2e-16
n= 2087, number of events= 899
```

Figure 28: final step for backward selection

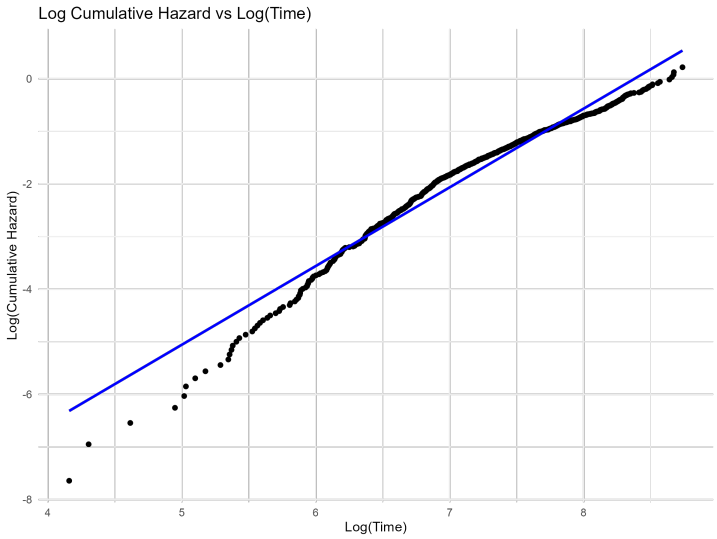

Figure 29: Log Cumulative Hazard Plot for Survival

Covariate	Coefficient (coef)	exp(coef)	SE (coef)	P-value
size20-50	0.3069	1.3592	0.0785	9.17e-05
size>50	0.5681	1.7649	0.1098	2.26e-07
log_nodes	0.5244	1.6895	0.0366	< 2e-16
grade 3: Poor Differentiation	0.2359	1.2661	0.0841	0.0049
age	-0.0839	0.9195	0.0191	1.02e-05
log_pgr	-0.0793	0.9238	0.0152	1.96e-07
age_sqr	0.0009	1.0009	0.0002	1.83e-07

Table 15: Results from the Cox Proportional Hazards Model

Covariate	Value	Std. Error	z	P-value
Intercept	9.32	0.166	56.23	< 2e-16
age	-0.00784	0.00343	-2.29	0.0221
meno1	-0.107	0.0925	-1.15	0.2489
size20-50	-0.344	0.0595	-5.78	7.5e-9
size>50	-0.634	0.0928	-6.83	8.4e-12
grade 3: Poor Differentiation	-0.250	0.0646	-3.87	0.00011
nodes	-0.0774	0.00637	-12.14	< 2e-16
pgr	0.000365	0.0000998	3.66	0.00025
hormon1	0.178	0.0880	2.02	0.0431
Log(scale)	0.0402	0.0257	1.56	0.1177

Table 16: Results from the Log-normal Model