**Background**

Breast cancer is one of the most common malignancies in women worldwide and a leading cause of cancer-related mortality1. Despite advances in early detection and treatment, breast cancer remains a highly heterogeneous disease, with subtypes such as luminal A, luminal B, HER2-enriched, and triple-negative breast cancer (TNBC) showing variable prognosis and response to therapy2. While chemotherapy is a pillar of breast cancer treatment, its efficacy varies widely across tumor subtypes, stages, and patient populations. In particular, aggressive subtypes such as TNBC and advanced-stage breast cancer often have poor outcomes despite intensive chemotherapy regimens. In addition, chemotherapy can cause significant toxicity, impacting patients' quality of life and adherence3. The variability in response underscores the need for better predictive markers and tailored therapeutic strategies to optimize outcomes.

Survival analysis provides a robust framework for evaluating the impact of chemotherapy on long-term outcomes and identifying prognostic factors that can guide personalized treatment. By examining the relationship between survival, tumor characteristics, patient background and treatment regimens, this report aims to provide insight into the real-world efficacy of chemotherapy. Understanding these dynamics is critical to improving therapeutic strategies, particularly for patients with aggressive subtypes or poor prognosis, ultimately advancing efforts to reduce recurrence and improve survival in breast cancer patients.

**Methods**

***Exploratory Data Analysis***

To investigate the baseline characteristics and potential predictors of survival in breast cancer patients receiving chemotherapy, we first performed exploratory data analysis (EDA). The dataset included 1,977 patients stratified into groups with and without chemotherapy. Summary statistics for continuous variables, such as age and tumor size, were calculated and visualized using histograms and box plots (Figure 1) to identify central tendencies and distributions. Categorical variables, such as menopausal status and tumor stage, were summarized as proportions and analyzed for group differences using Pearson's chi-squared tests and Fisher's exact tests.

Correlations between continuous variables were assessed using pairwise Pearson correlation coefficients, as shown in a heat map (Figure 2). Statistical tests, including Wilcoxon rank-sum tests for non-normally distributed continuous variables, were used to assess differences between the chemotherapy and non-chemotherapy groups.

***Kaplan-Meier Survival Analysis and (Weighted) Log-rank Test***

Kaplan-Meier survival analysis was used to compare overall survival (OS) and relapse-free survival (RFS) for chemotherapy and non-chemotherapy groups. Censoring was applied to individuals lost to follow-up or surviving beyond the study period without experiencing the event of interest. The log-rank test and Wilcoxon test was used to evaluate whether there were significant differences in survival curves between the chemotherapy and non-chemotherapy groups.

***Cox Proportional Hazards (CoxPH) model***

CoxPH model was used to assess the relationship between multiple covariates and the risk of death or relapse in breast cancer patients. Univariate Cox models were initially conducted for each predictor, including chemotherapy, age at diagnosis, tumor size, hormone and radiation therapy, HER2 status, tumor stage, and mutation count, to identify variables with p-values below 0.20 for inclusion in the multivariable model. The multivariable analysis was performed using backward selection and model selection methods, with the final model incorporating significant predictors identified from the univariate analysis. The model was assessed for proportional hazards assumptions using log-log survival plots and Schoenfeld residuals, with interaction terms introduced when violations of the assumption were detected, particularly for chemotherapy. This method allowed for a comprehensive evaluation of the factors influencing survival and relapse, adjusting for potential confounders and time-varying effects. Piecewise Cox models were also applied to account for potential changes in hazard ratios over time by dividing the follow-up period into distinct intervals. This approach provided a more flexible framework to capture time-varying effects and evaluate the impact of covariates within specific time windows.

**Results**

***Exploratory Data Analysis***

EDA revealed significant differences in patient characteristics between the groups (Table 1). Patients receiving chemotherapy were younger (mean age 50.5 vs. 63.9 years, p<0.001), more likely to be premenopausal (48% vs. 14%, p<0.001) and had larger tumors on average (mean size 32.3 mm vs. 24.7 mm, p<0.001). The distribution of tumor stages was skewed, with stage 2 and 3 being predominant among chemotherapy patients (74% and 19%, respectively) (Figure 1).

Molecular subtype distributions also differed: basal-like, HER2-enriched, and claudin-low subtypes were overrepresented in the chemotherapy group, whereas luminal A dominated the non-chemotherapy group. Significant associations were observed between chemotherapy and key tumor characteristics, including ER status (negative: 62% vs. 14%, p<0.001) and histologic grade (grade 3: 79% vs. 42%, p<0.001).

Figure 2 shows some correlations between tumor stage, tumor size, lymph node involvement and Nottingham prognostic index, possibly because stage classification and prognostic index include these factors in their definition.

A collage of multiple graphs

Description automatically generated

Figure 1. Histograms and Boxplots of Continuous Background Variables

A screenshot of a graph

Description automatically generated

Figure 2. Correlations among Continuous Background Variables

***Kaplan-Meier Survival Analysis with Log-rank and Wilcoxon Tests***

The Kaplan-Meier curve for OS demonstrated a significant difference between the two groups (Log-rank test p = 0.002, Wilcoxon test p < 0.001), with the chemotherapy group initially showing a lower overall survival probability compared to the non-chemotherapy group (Figure 3). There’s also a greatest separation between the two groups early in follow-up time. The survival curves crossed after approximately 180 months. After the cross, a lower survival probability for the non-chemotherapy group compared to the chemotherapy group.

For RFS, the chemotherapy group consistently exhibited a lower relapse free survival probability than the non-chemotherapy group across all follow-up periods (Figure 4). The gap between the two groups widens over time and with the greatest separation between the two groups around the median follow-up time. The log-rank test revealed a significant difference (Log-rank test p < 0.001, Wilcoxon test p < 0.001), indicating that patients receiving chemotherapy experienced disease progression or death at a higher rate compared to those who did not receive chemotherapy.



Figure 3. Overall survival (OS) by treatment groups



Figure 4. Progression-free survival (PFS) by treatment groups

***Cox Proportional Hazards (CoxPH) model***

The Cox Proportional Hazards model revealed significant predictors of death, with chemotherapy increasing the hazard by 54.5% (HR = 1.545, p < 0.001). This likely reflects its use in more advanced cases. Age at diagnosis increased the hazard of death by 70.6% per year (HR = 1.706, p < 0.001), while tumor size raised it by 14.7% per unit increase (HR = 1.147, p < 0.001). Radiation therapy reduced the hazard by 15.5% (HR = 0.845, p = 0.029), and HER2-positive status increased the hazard by 72.0% (HR = 1.720, p < 0.001). Tumor stage contributed a 40.4% higher hazard per unit increase (HR = 1.404, p < 0.001). Hormone therapy had a modest, non-significant protective effect (HR = 0.910, p = 0.251), and mutation count showed minimal impact (HR = 1.007, p = 0.395).

The Cox Proportional Hazards model identified chemotherapy as a significant predictor of relapse, increasing the hazard of recurrence by 80.1% (HR = 1.801, p < 0.001). This strong association likely reflects the use of chemotherapy in patients with more aggressive or advanced disease. Mutation count had a minimal impact, with a 0.7% increase in relapse hazard per unit increase (HR = 1.007, p = 0.3963), which was not statistically significant.

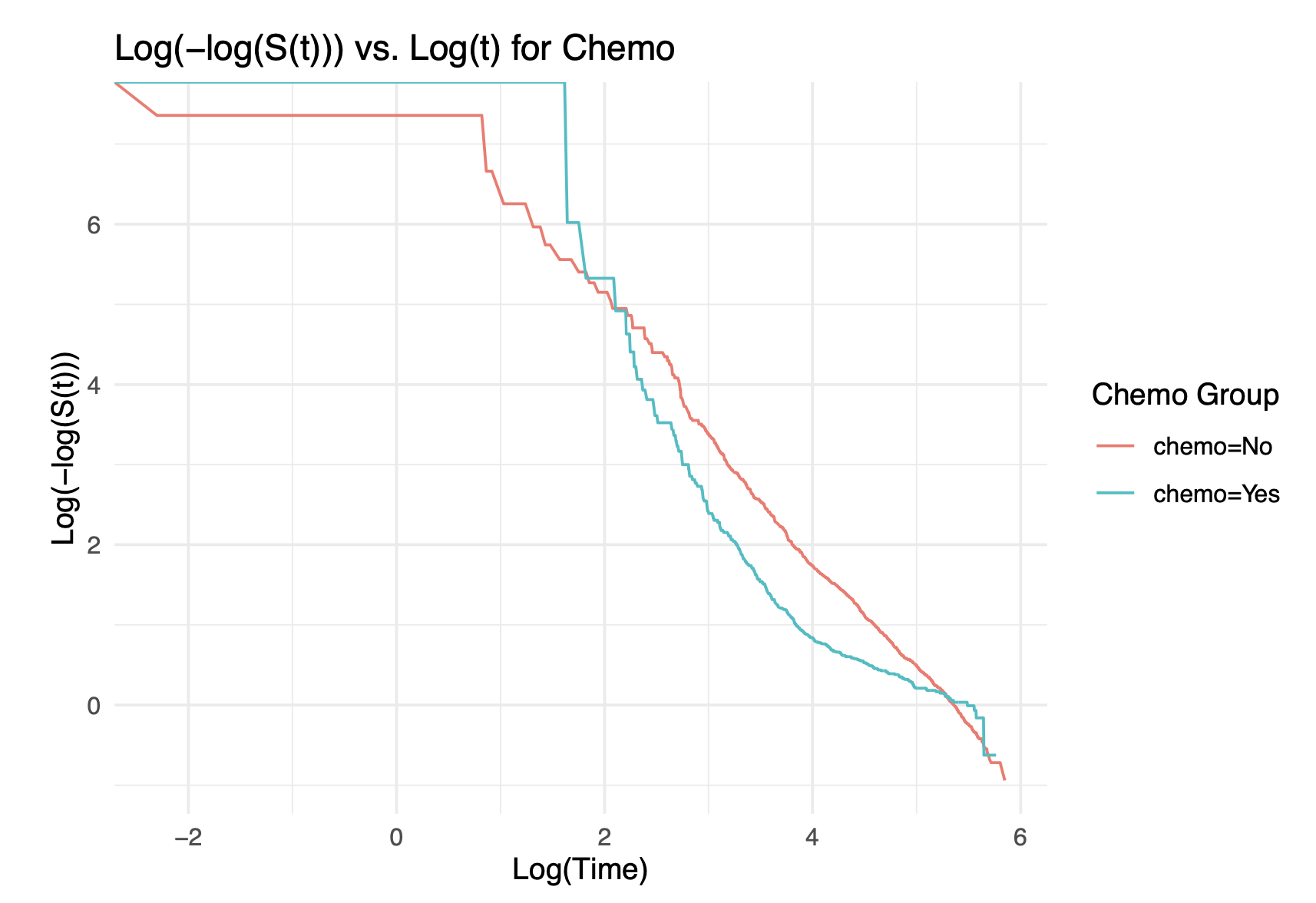


Figure 5. Cox PH Analysis of the Risk of Death

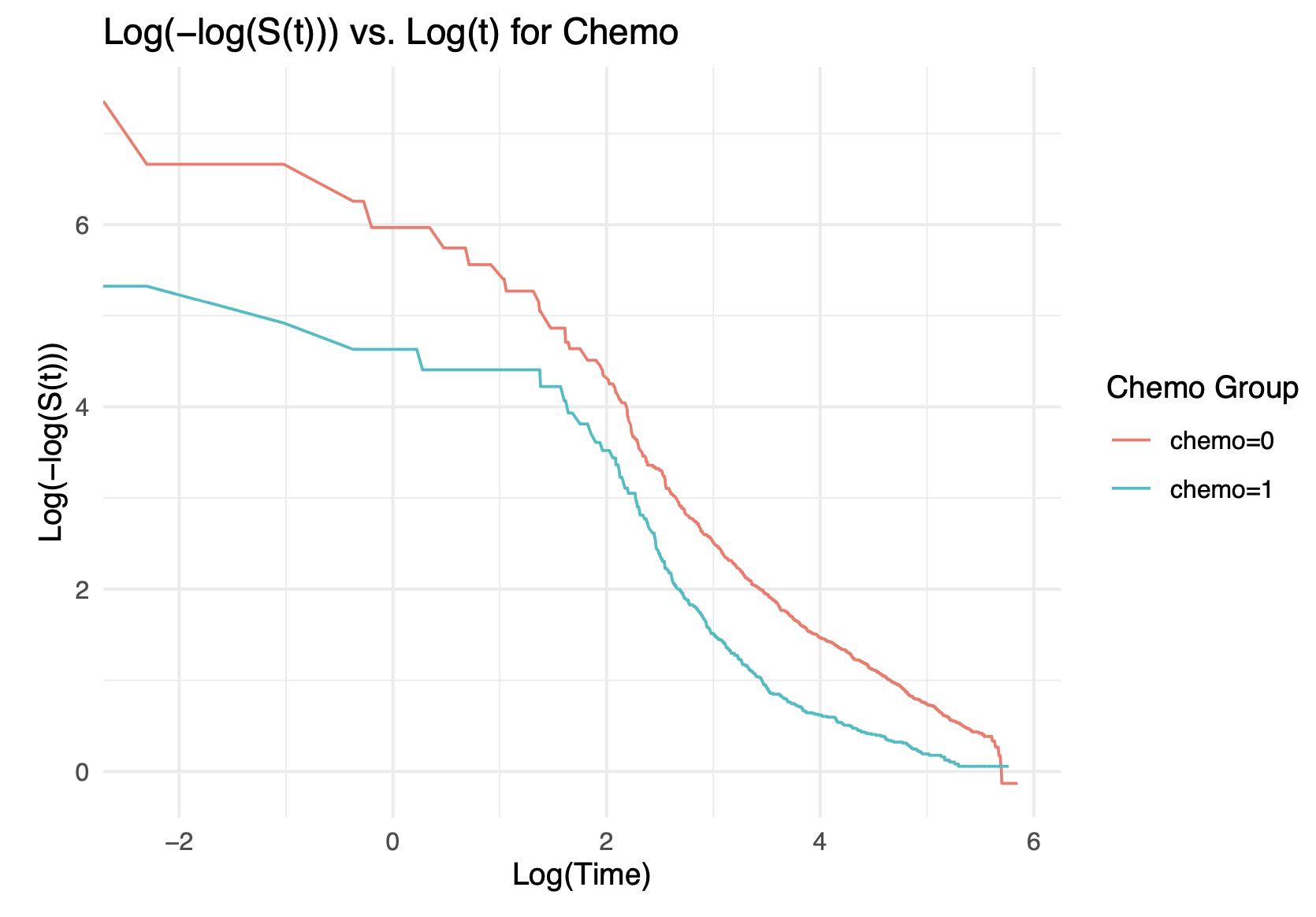


Figure 6. Cox PH Analysis of the Risk of Relapse

The piecewise Cox model for the risk of death showed chemotherapy was not significant (p=0.463). Significant predictors included age at diagnosis (p<0.001), tumor size (p=0.001), hormone therapy (p<0.001), radiotherapy (p=0.043), HER2-positive status (p<0.001), and tumor stage (p<0.001). Although the 180+ months time interval was not significant, the piecewise model outperformed the mutation count model with a lower AIC (9027.52 vs. 9861.14) and a higher concordance index (0.79 vs. 0.66), highlighting its superior performance in capturing survival dynamics.

The piecewise Cox model for the risk of relapse showed chemotherapy was not significant (p=0.606). Significant predictors included age at diagnosis (p=0.006), tumor size (p=0.009), hormone therapy (p<0.001), HER2-positive status (p<0.001), tumor stage (p<0.001), and mutation count (p=0.015). Despite the insignificance of the 180+ months time interval, the piecewise model outperformed the mutation count model with a lower AIC (6996.59 vs. 7543.27) and a higher concordance index (0.76 vs. 0.58), indicating better fit and performance.

**Discussion**

***Kaplan-Meier Survival Analysis with Log-rank and Wilcoxon Tests***

For OS, the crossing of the curves highlights a time-dependent effect of chemotherapy.The Non-Chemotherapy group has a higher survival probability early on, and at around median follow-up time (around 180 months), the two survival curves cross, after which a higher survival probability for patients receiving chemotherapy. This suggests that chemotherapy might be associated with higher early mortality risks but potentially offers long-term benefits for certain patients. In the short term, chemotherapy may initially have a negative effect on overall survival, potentially due to its toxicity, side effects, or because patients receiving chemotherapy may be more severely ill at baseline. Over time, the survival benefit of chemotherapy becomes more apparent, possibly due to its effectiveness in controlling disease progression in the long run. The observed crossing of survival curves highlights a potential violation of the proportional hazards assumption, necessitating further exploration through alternative statistical approaches.

For RFS, the significant log rank test and Wilcoxon test indicates that patients receiving chemotherapy experienced disease progression or death at a higher rate compared to those who did not receive chemotherapy. Moreover, a consistent pattern is observed that patients not receiving chemotherapy are less likely to relapse. This outcome is counterintuitive, as chemotherapy is generally expected to improve survival outcomes. Potential reasons for this unexpected outcome may include patient selection bias, that patients receiving chemotherapy may have had more severe or advanced disease at baseline, increasing their risk of relapse. The adverse effects of chemotherapy might also have outweighed its benefits for certain patients, leading to worse outcomes. Moreover, frequent monitoring in the chemotherapy group may lead to earlier relapse detection, while less intensive monitoring in the no-chemotherapy group could delay detection, inflating relapse-free survival. Lastly, treatment-related toxicity or comorbidities in the chemotherapy group may reduce recorded relapses, obscuring its true impact on relapse-free survival.

***Cox Proportional Hazards (CoxPH) model***

***Risk of Death***

The Schoenfeld residuals analysis indicated significant violations of the proportional hazards assumption for chemotherapy, with (χ² = 44.305, p < 2.8e-11). This suggests that the effect of chemotherapy on survival changes over time, highlighting the need to address its time-dependent effects for better model accuracy. The global Schoenfeld test confirmed that the proportional hazards assumption was violated for at least one variable (χ² = 116.809, p < 2e-16), emphasizing the necessity for time-dependent covariates to improve the model's validity.

Table 5 presents a comparison between the Original Final Model and the Interaction Model, which incorporates time-dependent effects. The Interaction Model demonstrates a lower AIC (9601.453 vs. 9861.136) and a reduced Negative 2 Log Likelihood (9587.453 vs. 9845.136), indicating an improved model fit. The Likelihood Ratio Test (LRT) yields a p-value of 0, reflecting a statistically significant improvement in model fit with the inclusion of interaction terms. These findings underscore the importance of accounting for interactions between time and chemotherapy to better capture time-dependent effects and provide a more accurate representation of the risk of death.

***Risk of Relapse***

The Schoenfeld residuals analysis indicated potential violations of the proportional hazards assumption for chemotherapy, HER2 status, tumor stage, and mutation count. Chemotherapy demonstrated significant time-varying effects (χ² = 13.920, p = 0.00019), indicating that its impact on relapse risk changes dynamically over time. HER2 status also showed potential time dependence (χ² = 3.924, p = 0.0476), suggesting its effect on relapse risk may vary. Tumor stage (χ² = 7.082, p = 0.00779) and mutation count (χ² = 5.933, p = 0.01486) also violated the PH assumption, indicating time-dependent effects. The global test confirmed that at least one variable in the model violated the PH assumption (χ² = 28.911, p = 8.6e-5). Addressing these violations with time-dependent covariates improves the model’s accuracy in predicting relapse risk.

To address these time-dependent effects, the Interaction Model was developed and compared to the Original Final Model (Table 13). The Interaction Model provides a significantly better fit for predicting relapse risk, as evidenced by a lower AIC (7283.058 vs. 7543.268) and reduced Negative 2 Log Likelihood (-2LogL; 7269.058 vs. 7531.268). The Likelihood Ratio Test (LRT) yielded a highly significant p-value (0), confirming the superior fit of the Interaction Model. These results underscore the importance of accounting for time-varying effects through interaction terms, enabling the model to more accurately capture the dynamic nature of relapse risk. By addressing the time-dependent effects identified in the Schoenfeld residuals analysis, the Interaction Model achieves a more accurate and nuanced representation of relapse risk. This improvement ensures a robust and valid prediction framework, aligning with the evidence that variables like chemotherapy, HER2 status, tumor stage, and mutation count exhibit dynamic influences on relapse risk over time.

***Piecewise Cox Model***

The results indicate that chemotherapy was not a significant predictor of either death or relapse risk in the piecewise Cox models. This lack of significance might stem from potential confounders that have not been accounted for in the model. Additionally, there might be significant interactions between chemotherapy and other variables, such as tumor stage or hormone therapy, making it challenging to isolate the independent effect of chemotherapy. Another plausible explanation is that the effects of chemotherapy vary over time. Initial toxicity and short-term adverse impacts, including increased mortality due to side effects, may counteract the long-term survival benefits of the treatment. If the time intervals chosen in the piecewise model do not align well with these temporal dynamics, the significance of chemotherapy could be diluted, masking its true impact.

The current cutpoint for the time intervals in the piecewise model was based on observations from the Kaplan-Meier curve. While this approach provides a practical starting point, it may not fully capture the dynamic effects of chemotherapy. Refining these cutpoints based on clinical reasoning and expert input could provide a more accurate representation of treatment dynamics. Additionally, further exploration using data-driven approaches, such as change-point detection or spline-based methods, may better identify key transitions in treatment effects over time. These refinements could improve the model's ability to detect and interpret the complex temporal effects of chemotherapy on survival and relapse risk.

**Conclusion**

This study highlights the interplay of both protective and adverse factors influencing survival and relapse risks in breast cancer patients. Chemotherapy showed a mixed effect on breast cancer prognosis. Negative predictors such as older age, larger tumor size, and HER2-positive status were associated with worse outcomes. In contrast, protective factors like hormone therapy, radiotherapy, and a lower Nottingham prognostic index significantly improved survival and reduced relapse risks. These findings underscore the importance of tailoring treatment strategies based on patient and tumor characteristics to optimize outcomes, especially for patients with aggressive subtypes or poor prognosis. The piecewise Cox model further emphasized the importance of considering time-specific dynamics in treatment effects to better capture the nuanced impact of therapeutic interventions.

**References**

1. Trapani D, Ginsburg O, Fadelu T, et al. Global challenges and policy solutions in breast cancer control. *Cancer Treat Rev*. 2022;104(102339):102339.

2. Vranic S, Cyprian FS, Gatalica Z, Palazzo J. PD-L1 status in breast cancer: Current view and perspectives. *Semin Cancer Biol*. 2021;72:146-154.

3. Lake DE, Hudis CA. High-dose chemotherapy in breast cancer. *Drugs*. 2004;64(17):1851-1860.

Table 1. Baseline Characteristics of Patients with or without Chemotherapy

| **Characteristic** | **Without Chemotherapy**  N = 1,5651 | **With Chemotherapy**  N = 4121 | **P-value**2 |
| --- | --- | --- | --- |
| Age at diagnosis | 63.9 (12.0) | 50.5 (11.0) | <0.001 |
| Sex |  |  |  |
| Female | 1,565 (100) | 412 (100) |  |
| Menopausal status |  |  | <0.001 |
| Post | 1,339 (86) | 214 (52) |  |
| Pre | 226 (14) | 198 (48) |  |
| Cancer type |  |  |  |
| Breast | 15 (1.0) | 2 (0.5) |  |
| Breast Angiosarcoma | 0 (0) | 0 (0) |  |
| Breast Invasive Ductal Carcinoma | 1,169 (75) | 367 (89) |  |
| Breast Invasive Lobular Carcinoma | 129 (8.2) | 17 (4.1) |  |
| Breast Invasive Mixed Mucinous Carcinoma | 22 (1.4) | 1 (0.2) |  |
| Breast Mixed Ductal and Lobular Carcinoma | 193 (12) | 18 (4.4) |  |
| Invasive Breast Carcinoma | 36 (2.3) | 6 (1.5) |  |
| Metaplastic Breast Cancer | 1 (<0.1) | 1 (0.2) |  |
| Primary tumor laterality |  |  | 0.13 |
| Left | 783 (53) | 188 (49) |  |
| Right | 698 (47) | 199 (51) |  |
| Missing | 84 | 25 |  |
| Tumor size | 24.7 (12.6) | 32.3 (22.1) | <0.001 |
| Missing | 17 | 6 |  |
| Tumor stage |  |  | <0.001 |
| 0 | 11 (1.0) | 0 (0) |  |
| 1 | 480 (42) | 20 (6.3) |  |
| 2 | 591 (52) | 234 (74) |  |
| 3 | 58 (5.1) | 60 (19) |  |
| 4 | 7 (0.6) | 3 (0.9) |  |
| Missing | 418 | 95 |  |
| Cellularity |  |  | 0.004 |
| High | 739 (49) | 225 (56) |  |
| Low | 165 (11) | 50 (13) |  |
| Moderate | 611 (40) | 125 (31) |  |
| Missing | 50 | 12 |  |
| PAM50 claudin-low subtype |  |  |  |
| Basal | 96 (6.1) | 113 (27) |  |
| claudin-low | 142 (9.1) | 73 (18) |  |
| HER2 | 136 (8.7) | 88 (21) |  |
| LumA | 644 (41) | 56 (14) |  |
| LumB | 428 (27) | 47 (11) |  |
| NC | 6 (0.4) | 0 (0) |  |
| Normal | 113 (7.2) | 35 (8.5) |  |
| ER status |  |  | <0.001 |
| Negative | 218 (14) | 254 (62) |  |
| Positive | 1,347 (86) | 158 (38) |  |
| Histologic grade |  |  | <0.001 |
| 1 | 161 (11) | 8 (2.0) |  |
| 2 | 695 (47) | 76 (19) |  |
| 3 | 629 (42) | 323 (79) |  |
| Missing | 80 | 5 |  |
| HER2 status |  |  | <0.001 |
| Negative | 1,421 (91) | 309 (75) |  |
| Positive | 144 (9.2) | 103 (25) |  |
| Other histologic subtype |  |  |  |
| Ductal/NST | 1,136 (74) | 354 (87) |  |
| Lobular | 129 (8.4) | 17 (4.2) |  |
| Medullary | 13 (0.9) | 12 (3.0) |  |
| Metaplastic | 1 (<0.1) | 1 (0.2) |  |
| Mixed | 193 (13) | 18 (4.4) |  |
| Mucinous | 22 (1.4) | 1 (0.2) |  |
| Other | 15 (1.0) | 2 (0.5) |  |
| Tubular/ cribriform | 20 (1.3) | 1 (0.2) |  |
| Missing | 36 | 6 |  |
| Hormone therapy | 1,025 (65) | 191 (46) | <0.001 |
| Breast surgery |  |  | 0.032 |
| Breast Conserving | 641 (41) | 143 (35) |  |
| Mastectomy | 909 (59) | 260 (65) |  |
| Missing | 15 | 9 |  |
| Radiotherapy | 838 (54) | 335 (81) | <0.001 |
| Lymph nodes involvement | 1.4 (3.4) | 4.3 (5.5) | <0.001 |
| Missing | 58 | 16 |  |
| Mutations | 5.8 (4.2) | 5.2 (3.5) | <0.001 |
| Missing | 90 | 29 |  |
| Nottingham prognostic index | 3.8 (1.1) | 5.0 (0.9) | <0.001 |
| PR status |  |  | <0.001 |
| Negative | 635 (41) | 302 (73) |  |
| Positive | 930 (59) | 110 (27) |  |

1Mean (SD); n (%)

2Wilcoxon rank sum test; Pearson's Chi-squared test; Fisher's exact test

Abbreviations: PAM50 (prediction analysis of microarray 50); ER (estrogen receptor); HER2 (human epidermal growth factor receptor 2); LumA (luminal A); LumB (luminal B); NC (not classified); NST (no special type); PR (progesterone receptor)

