**Survival Analysis Report: RADCURE Dataset**

**Course:** BINF5501  
**Assignment 4**  
**Student Name:** Suma Ambati  
**Date:** July 14, 2025

**1. Introduction**

This report analyzes survival data from the RADCURE clinical dataset to understand the effects of cancer stage and patient age on survival outcomes. We use Kaplan-Meier (KM) survival curves to visualize differences between Stage III and IV patients, perform a log-rank test for statistical comparison, fit a Cox Proportional Hazards (CoxPH) model to identify significant covariates, and finally validate findings with a Random Survival Forest (RSF) model using scikit-survival.

**2. Kaplan-Meier Survival Analysis**

The KM survival curves below compare patients in Stage III and Stage IV over a 12-month follow-up period: **A graph of a graph with blue and orange lines

AI-generated content may be incorrect.**

**Figure 1:** Kaplan-Meier Survival Curve

**Interpretation for Fig 1:**

Stage IV patients show a steeper decline in survival probability, especially in the first 6 months. The survival probability drops significantly faster compared to Stage III.

**3. Log-Rank Test**

To assess whether the survival difference between Stage III and IV is statistically significant, a log-rank test was conducted.

|  |  |  |
| --- | --- | --- |
| **Test Statistic** | **p-value** | **-log2 (p)** |
| 8.47 | <0.005 | 8.12 |

**Table 1:** Log-Rank Test Results

**Interpretation for Table 1:** The p-value < 0.005 indicates a significant difference in survival between Stage III and IV patients.

**4. Cox Proportional Hazards Model**

We fitted a CoxPH model using Age and Stage\_num (Stage III = 3, Stage IV = 4).

| **Variable** | **HR (exp(coef))** | **95% CI Lower** | **95% CI Upper** | **p-value** |
| --- | --- | --- | --- | --- |
| Age | 1.06 | 1.04 | 1.07 | <0.005 |
| Stage\_num | 1.61 | 0.74 | 3.50 | 0.23 |

**Table 2:** Cox Model Summary

Although **Age** was significantly associated with increased hazard (p < 0.005), **Stage\_num** was not statistically significant (p = 0.23). However, it still had a notable hazard ratio above 1.

A graph of a graph with numbers

AI-generated content may be incorrect.📈 **Figure 2:** Cox Model – Hazard Ratios with 95% CI

**Model Metrics:**

* Concordance index: 0.70
* Partial AIC: 2118.28
* Log-likelihood ratio test: 89.97 on 2 df (p < 0.005)

**5. Random Survival Forest (RSF) Validation**

* We trained an RSF model using the same covariates (Age, Stage\_num) on an 80/20 train-test split.
* A graph with blue squares

  AI-generated content may be incorrect.**Concordance Index (C-index):** 0.602

**Figure 3:** RSF Feature Importance

**Interpretation for Fig 3 :** The RSF model suggests that Age is the more influential predictor of survival outcome compared to Stage\_num, based on the mean decrease in C-index. The lower C-index (0.602) compared to the Cox model (0.70) indicates RSF performed less effectively in this dataset, potentially due to the limited number of covariates or smaller sample size after filtering.

**6. Conclusion**

* This survival analysis confirms that **Age is a strong and statistically significant predictor** of poorer survival outcomes in the RADCURE dataset. Although Stage\_num showed elevated hazard ratios, its p-value was not significant in the Cox model—possibly due to small sample size or wide confidence intervals. The KM curve and log-rank test still show clear survival differences by stage, which is supported by RSF validation.
* **Recommendations:** For future clinical studies, more granular staging and longer follow-up may help clarify the role of staging. Additionally, other clinical features like treatment type, comorbidities, or genetic markers could improve model performance.