**Part 4: Short-Answer Questions (Upload to Blackboard as a .PDF)**

Answer the following based on your work in **Part 3**:

1. Please provide the link to your public **GitHub repository**.
2. Based on your work in Part 3, please fill out the following table:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Strengths** | **Limitations** | **Key Use Cases** | **Your Observations** |
| **Kaplan- Meier** | A simple way to visualize survival over time. | Limited to comparing categorical groups (usually two). Cannot adjust for multiple covariates simultaneously. | Understanding survival patterns across patient groups. | KM curves showed clear survival differences between tumor stages (Stage I vs. Stage IV). It revealed poorer survival in advanced stages, consistent with clinical expectations. |
| **Cox Proportional Hazards** | Adjusts for multiple predictors simultaneously. Provides interpretable hazard ratios. | Assumes proportional hazards (constant effects over time). Sensitive to missing data. May miss complex interactions. | Identifying independent predictors of survival. | The Cox model showed moderate predictive ability (C-index = 0.66). Age, stage, and treatment were significant predictors. HPV was excluded due to substantial missing data to maintain power and validity. |
| **Random Survival Forests** | Captures nonlinear effects and interactions without proportional hazards assumption. | Harder to interpret clinically. Requires large datasets. Generally lower predictive power in smaller cohorts. | Exploring complex relationships and validating predictors. | RSF confirmed age and stage as key predictors, with a C-index of 0.60. Variable importance ranking aligned with Cox results, adding confidence to the findings despite slightly lower predictive performance. |

1. What are the primary differences between Kaplan-Meier (KM) analysis, Cox regression, and Random Survival Forests?

|  |  |
| --- | --- |
| **Method** | **Strengths** |
| **Kaplan-Meier** | Plots survival curves for categorical groups; excellent for univariate visualization but cannot adjust for covariates. It clearly showed survival differences by tumor stage. |
| **Cox Proportional Hazards** | Models the effect of multiple predictors jointly with interpretable hazard ratios; assumes proportional hazards over time. It quantified the impact of age, stage, and treatment on survival. |
| **Random Survival Forests** | A machine learning approach capturing complex nonlinear relationships and interactions without parametric assumptions. It supported Cox results, highlighting age and stage importance. |

1. What assumptions are made by Cox Proportional Hazards regression? How can these be evaluated?

* **Proportional Hazards Assumption**:  The hazard ratios for covariates are constant over time. Evaluated using Schoenfeld residuals tests or graphical diagnostics. In our analysis, stable model performance (C-index = 0.66) and consistent coefficients indicate this assumption likely holds.
* **Linearity Assumption**:  Continuous predictors (e.g., age) have a linear relationship with the log hazard. Supported by the agreement between Cox and RSF (which does not assume linearity), suggesting minimal nonlinearity in age effects.
* **Missing Data and Bias**:  Assumes missing data do not bias estimates. HPV status was missing in ~48.7% of patients. We excluded HPV from the Cox model to avoid loss of power and potential bias.
  + validity. Therefore, we excluded HPV from the multivariable analysis to avoid introducing bias and to maintain statistical power.

We carefully assessed the assumptions within the scope of our dataset and analysis approach. The steps we took helped ensure that our Cox model results were valid, interpretable, and clinically meaningful.

**5. Which method provided the best balance between interpretability and predictive performance?**

The **Cox Proportional Hazards model** provided the best balance. It yielded clinically meaningful hazard ratios with moderate predictive accuracy (C-index = 0.66). Compared to Random Survival Forests (C-index = 0.60) and Kaplan-Meier (limited to univariate analyses), Cox was the most practical and interpretable for clinical research.

**6. Identify any features that consistently demonstrate predictive power across different methods and highlight their potential clinical significance.**

* **Age:** The strongest predictor. Older patients (≥60 years) had significantly worse survival, consistent across KM, Cox, and RSF. This reflects biological aging, comorbidities, and treatment tolerance.
* **Tumor Stage:** Advanced stages (IVA/IVB) consistently predicted poorer outcomes across methods, reflecting tumor burden and disease extent.
* **Treatment Modality:** Multimodal treatments (e.g., surgery plus chemoradiotherapy) associated with better survival than single modalities, supporting comprehensive therapy benefits.
* **HPV Status:** Strongly predictive in KM analysis, with HPV-positive patients showing superior survival. Due to high missingness, HPV was excluded from multivariable models but remains biologically important, consistent with literature showing better prognosis and treatment response.

These factors are biologically and clinically relevant, influencing tumor behavior, treatment response, and survival.

**Reference**

Kuhn, J. P., Schmid, W., Körner, S., Bochen, F., Wemmert, S., Rimbach, H., Smola, S., Radosa, J. C., Wagner, M., Morris, L. G., Bozzato, V., Bozzato, A., Schick, B., & Linxweiler, M. (2021). HPV status as prognostic biomarker in Head and Neck Cancer—Which method fits the best for outcome prediction? *Cancers*, 13(18), 4730. https://doi.org/10.3390/cancers13184730