# Prognostic Modeling of Overall Survival in HNSCC

## A Comparative Analysis of Traditional Statistics vs. Machine Learning

Ruchitha Uppuluri

## Contents

# 1 Project Overview

Head and Neck Squamous Cell Carcinoma (HNSCC) presents a significant clinical challenge due to variable survival rates across tumor stages. This project evaluates the efficacy of modern Machine Learning approaches—specifically Random Survival Forests (RSF)—against traditional Cox Proportional Hazards models to improve prognostic accuracy.

## 1.1 Key Findings:

- **Random Survival Forests (RSF)** outperformed traditional models, achieving a Concordance Index (C-Index) of **0.74**.
- **Tumor Stage IVC** was identified as the single strongest predictor of mortality.
- **Time-to-Event models** proved superior to binary classification (2-year survival), highlighting the critical importance of handling censored data in clinical analytics.

# 2 Data & Methodology

Using clinical data from **The Cancer Genome Atlas (TCGA)** ($N = 528$), we conducted a comparative analysis using three distinct modeling approaches: 1. **Traditional Statistics:** Kaplan-Meier estimation and Cox Proportional Hazards. 2. **Binary Machine Learning:** Random Forest, Logistic Regression and XGBoost (predicting survival $>$ 24 months). 3. **Survival Machine Learning:** Random Survival Forests (RSF) and Penalized Cox Regression (CoxNet).

# 3 Exploratory Data Analysis

## 3.1 Demographics

The patient cohort exhibits a typical age distribution for HNSCC onset, with a median age of 61 years. The gender distribution reflects known epidemiological prevalence (Male $>$ Female), as shown below.

## 3.2 Clinical Characteristics

A significant portion of the dataset presents with advanced-stage disease (Stage IVA), which strongly influences the modeling strategy.

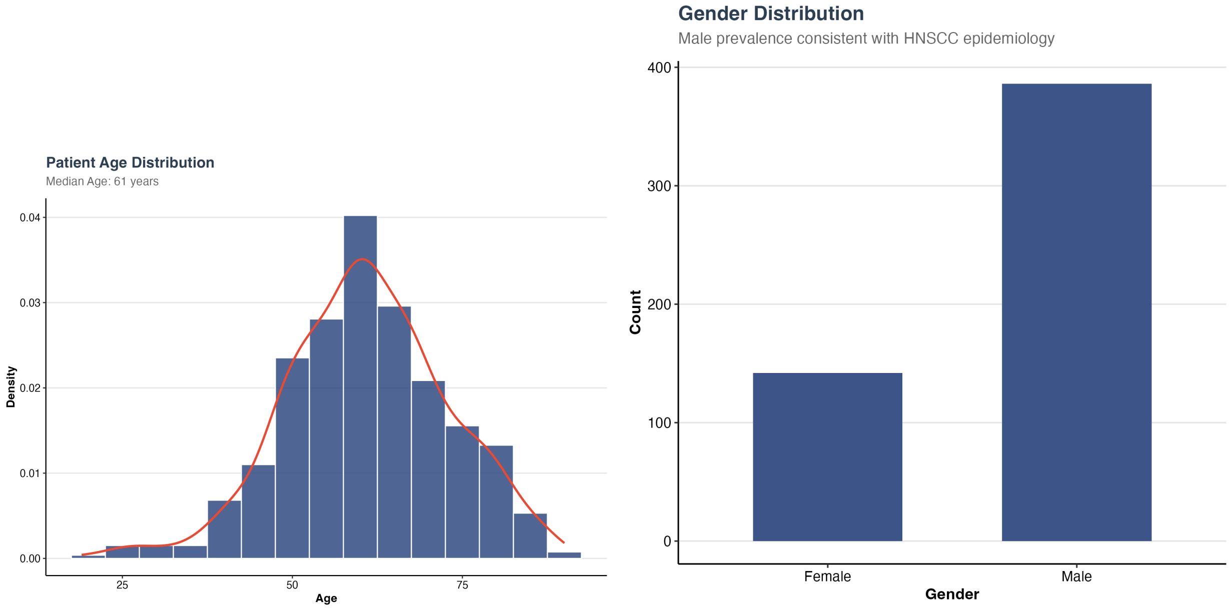
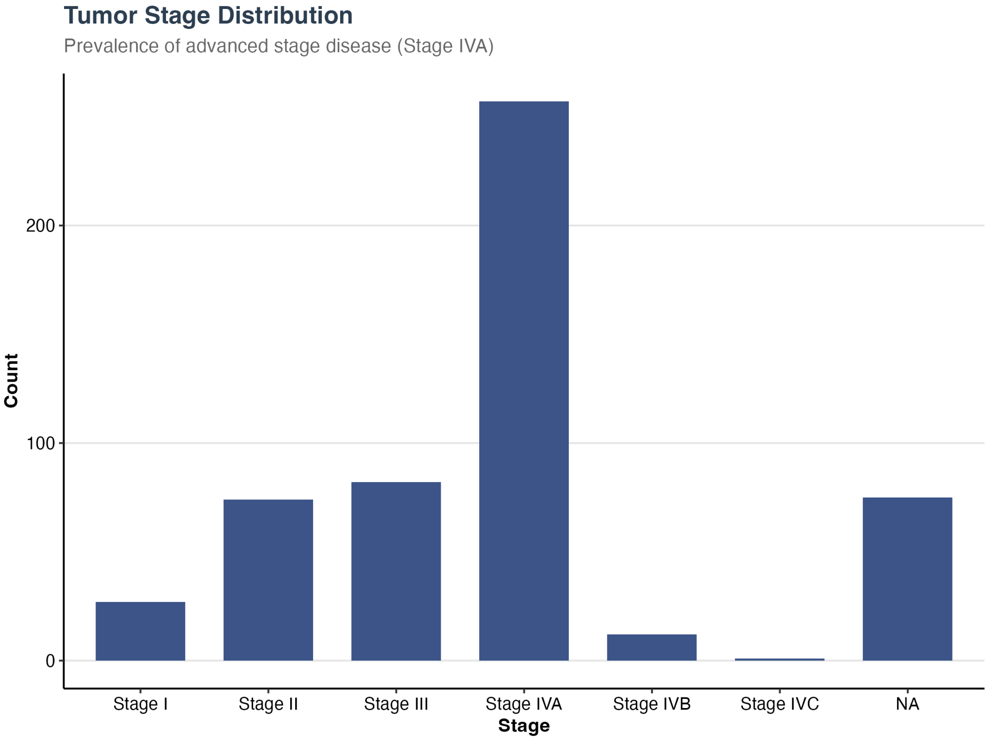
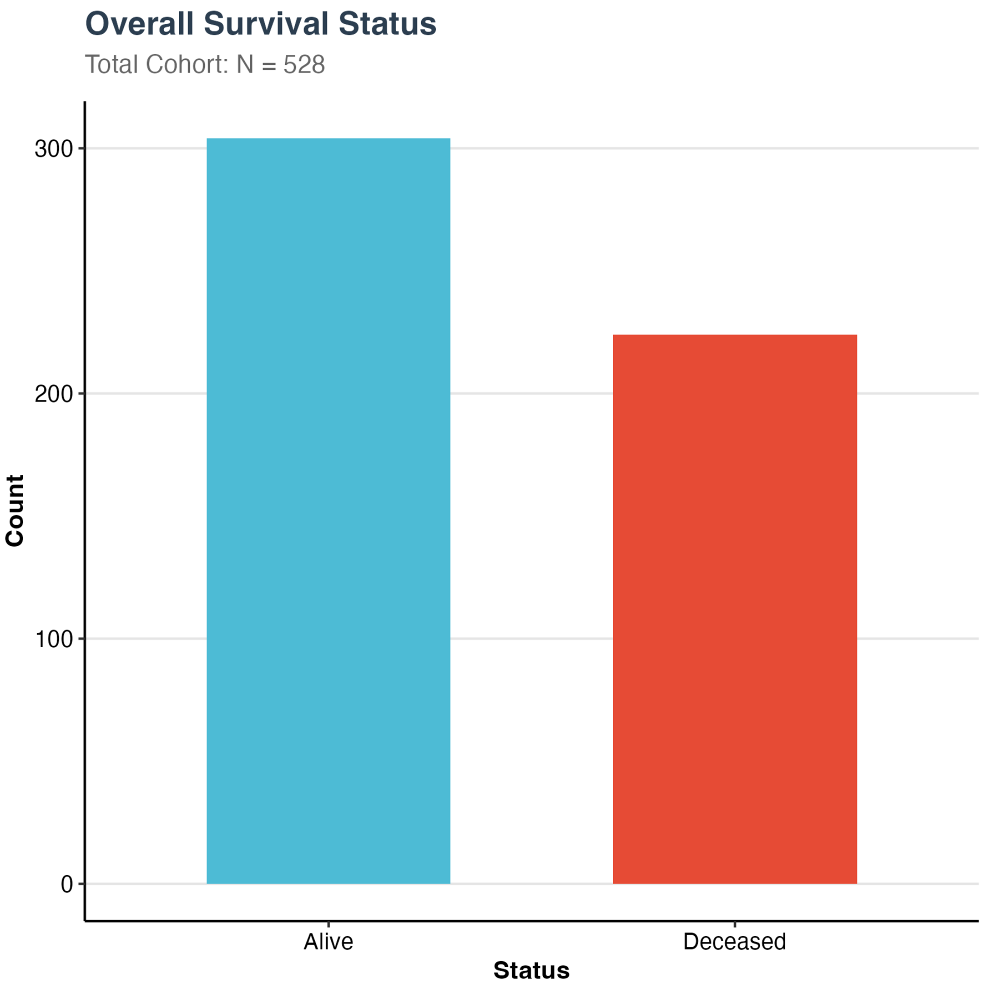


Figure 1: Patient Age Distribution



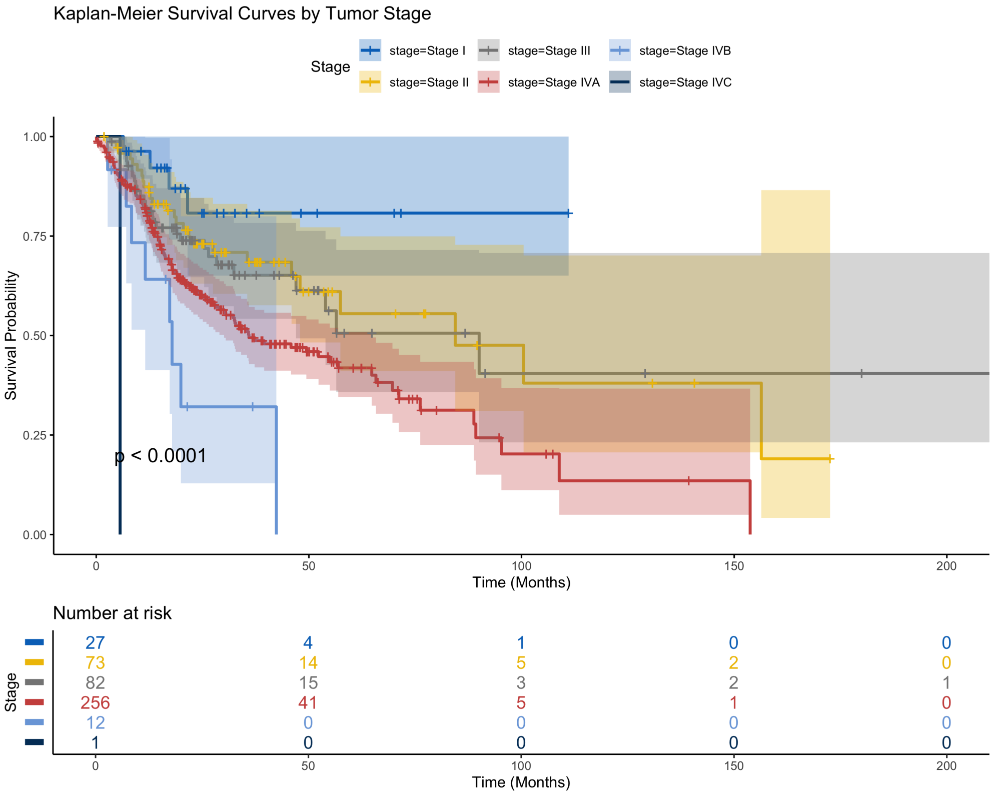


## 4 Traditional Survival Modeling

We established a baseline using Kaplan-Meier estimators and Cox Proportional Hazards models.

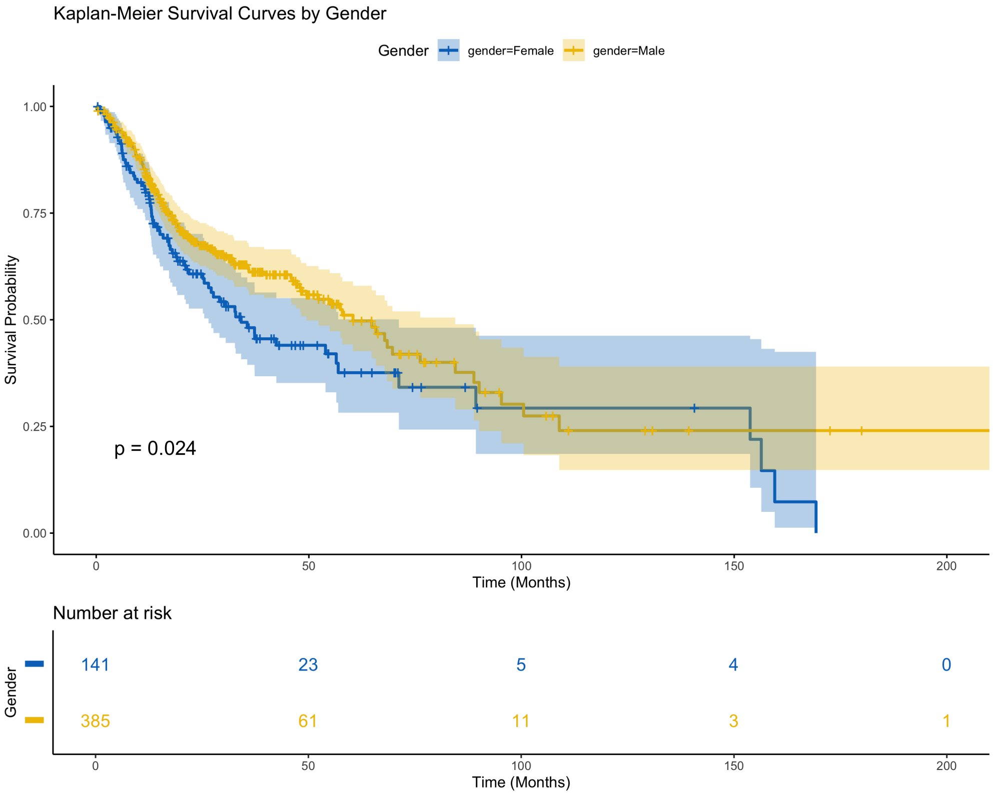
### 4.1 Survival by Tumor Stage

There is significant prognostic separation between early and late-stage disease ($p < 0.0001$), validating the clinical data quality.



## 4.2 Survival by Gender

Gender differences in survival were observed, though less pronounced than tumor stage.

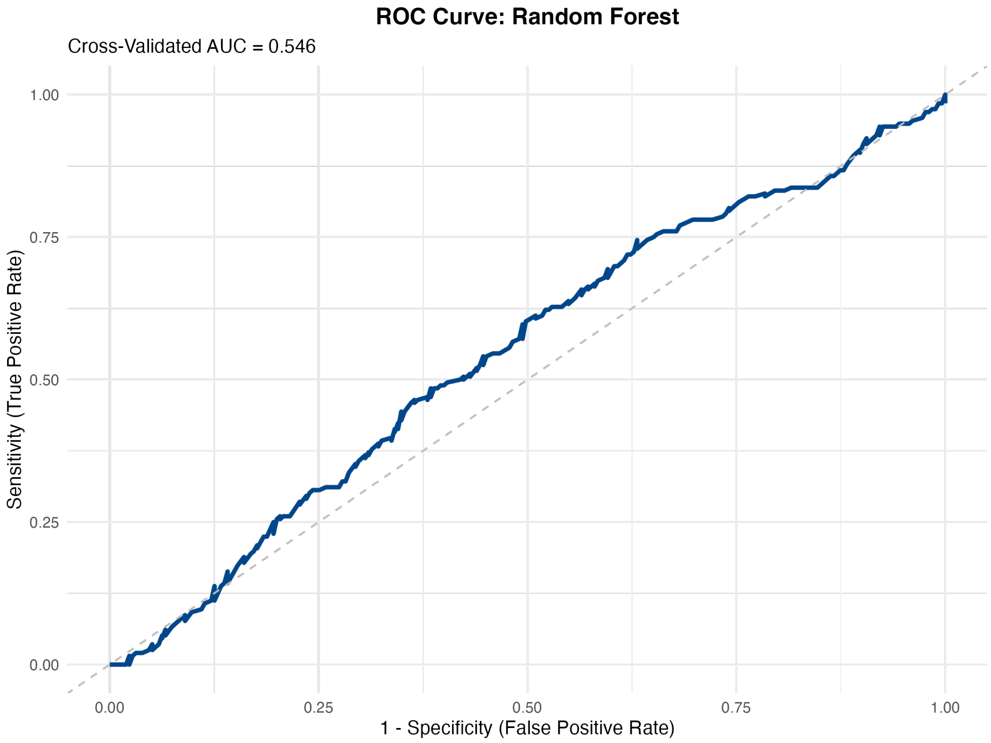


# 5 Machine Learning Models (Binary Classification)

We initially modeled the problem as a binary classification task (Survival > 24 Months). However, this approach discards valuable time-to-event information for censored patients.

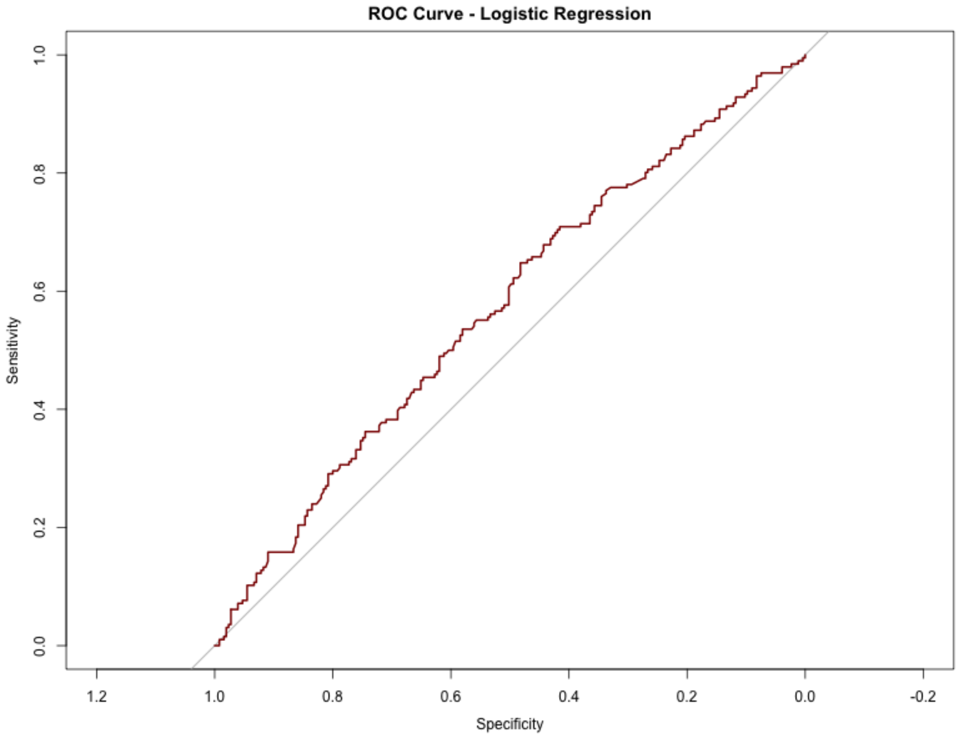
## 5.1 Random Forest Classifier

The standard Random Forest achieved moderate performance but struggled with the class imbalance inherent in the dataset.



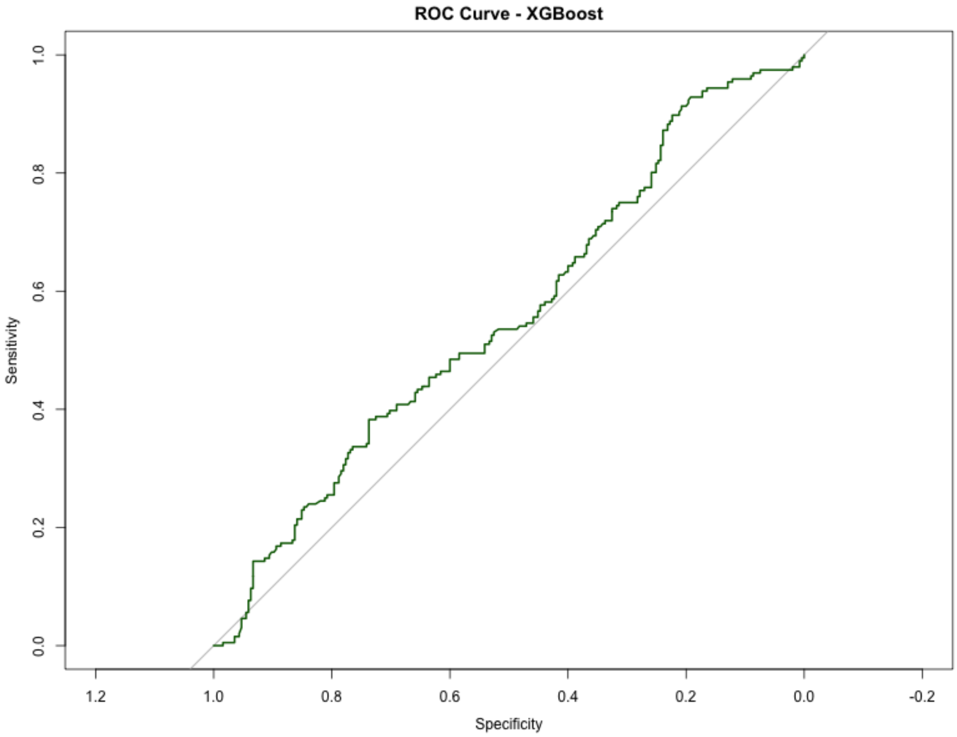
## 5.2 Logistic Regression with L1 Regularization

Logistic regression with Lasso regularization was implemented to handle potential multicollinearity, but predictive power remained comparable to the Random Forest



## 5.3 XGBoost Classifier

XGBoost results were similar to other binary classifiers, suggesting that the limitation lies in the binary formulation of the problem rather than the algorithm choice.

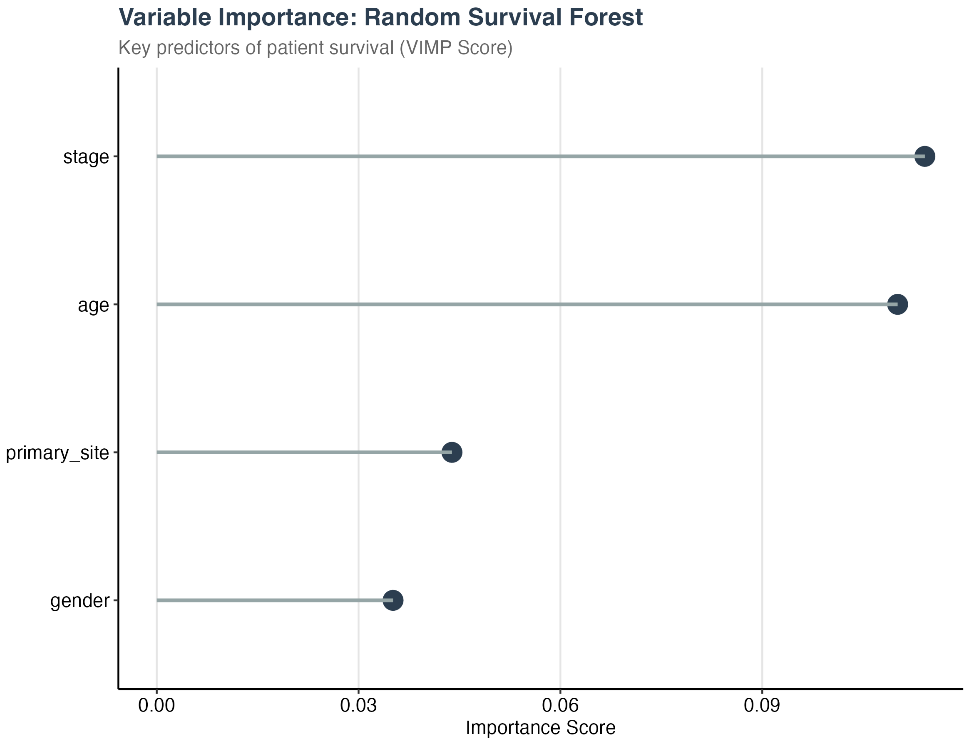


# 6 Advanced Survival Machine Learning

To address the limitations of binary classification, we implemented models capable of handling censored time-to-event data directly.

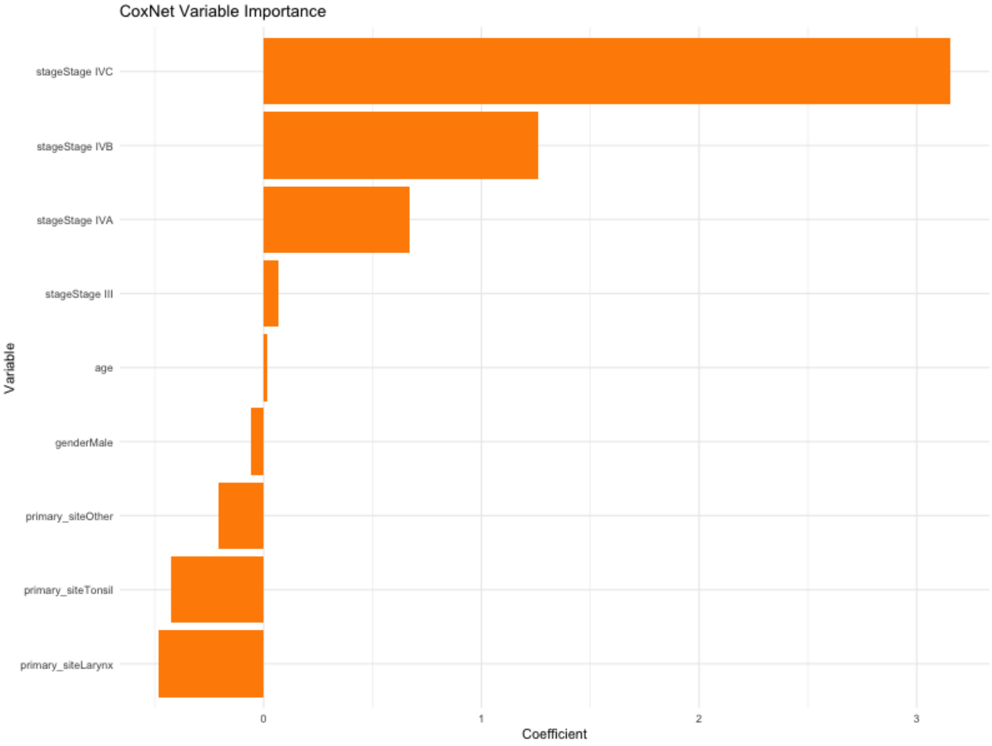
## 6.1 Random Survival Forest (RSF)

The RSF model achieved the highest predictive performance. The Variable Importance (VIMP) plot below highlights Tumor Stage and Age as the dominant predictors.



## 6.2 Penalized Cox Regression (CoxNet)

L1-regularization (Lasso) was used to perform feature selection within the Cox framework.

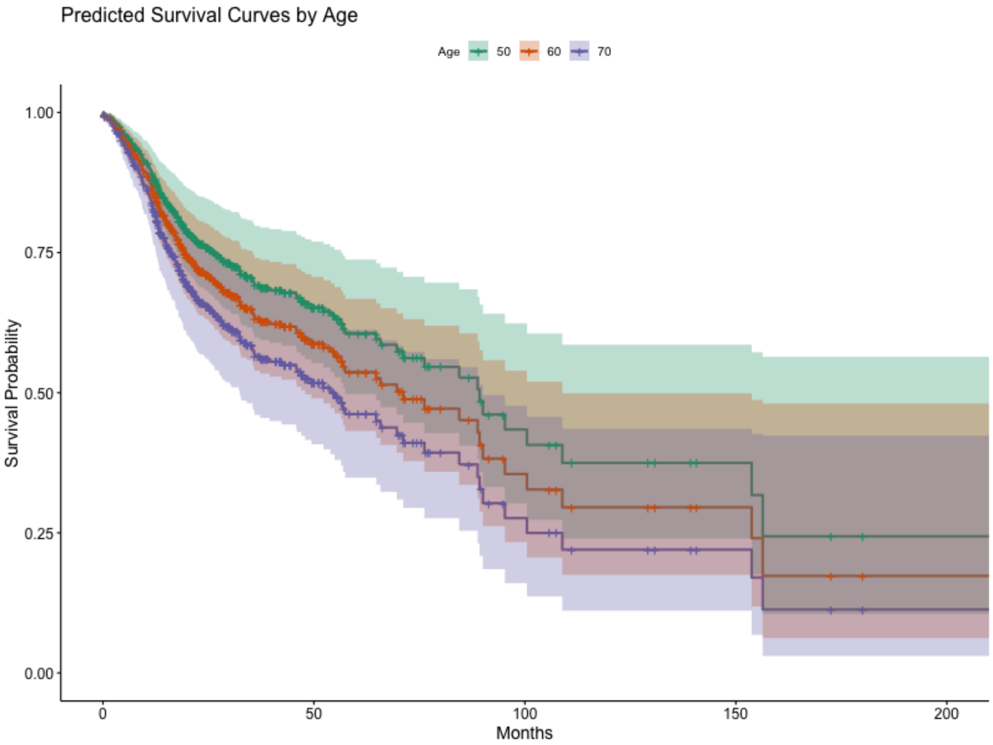


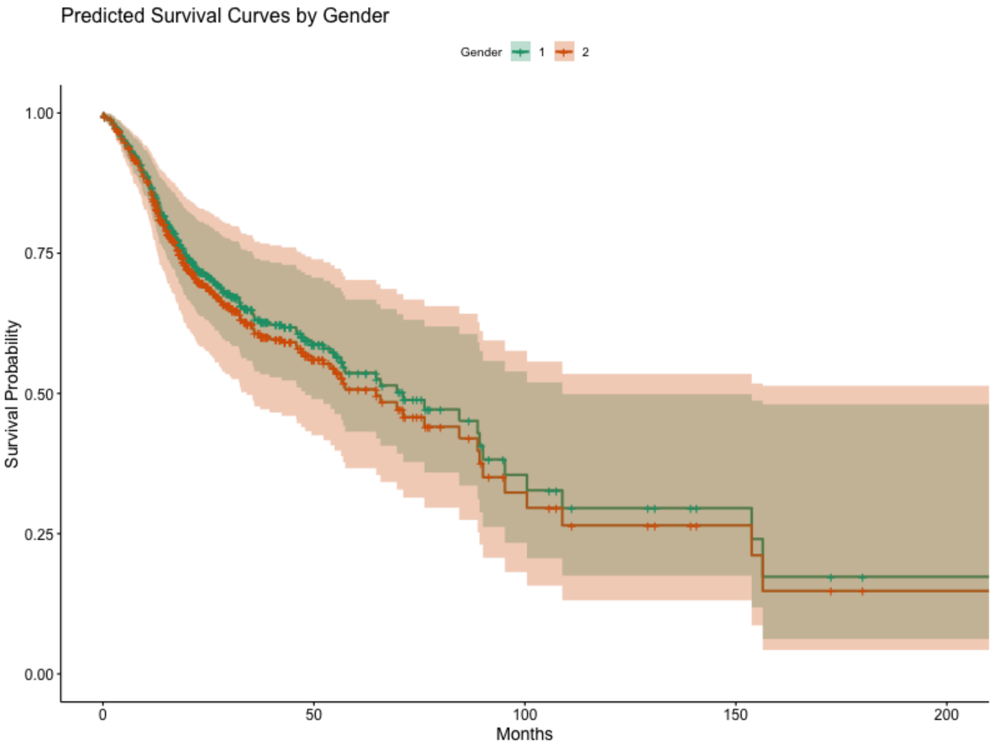
# 7 Clinical Utility: Subgroup Predictions

Using the CoxNet model, we generated predicted survival curves for hypothetical patient profiles to demonstrate clinical utility.

Table 1: Predicted Median Survival by Subgroup (CoxNet)

| Group | Median_Survival_Months |
|---|---:|
| Age 50 | 110.0 |
| Age 60 | 90.8 |
| Age 70 | 75.0 |
| Male | 90.8 |
| Female | 87.8 |



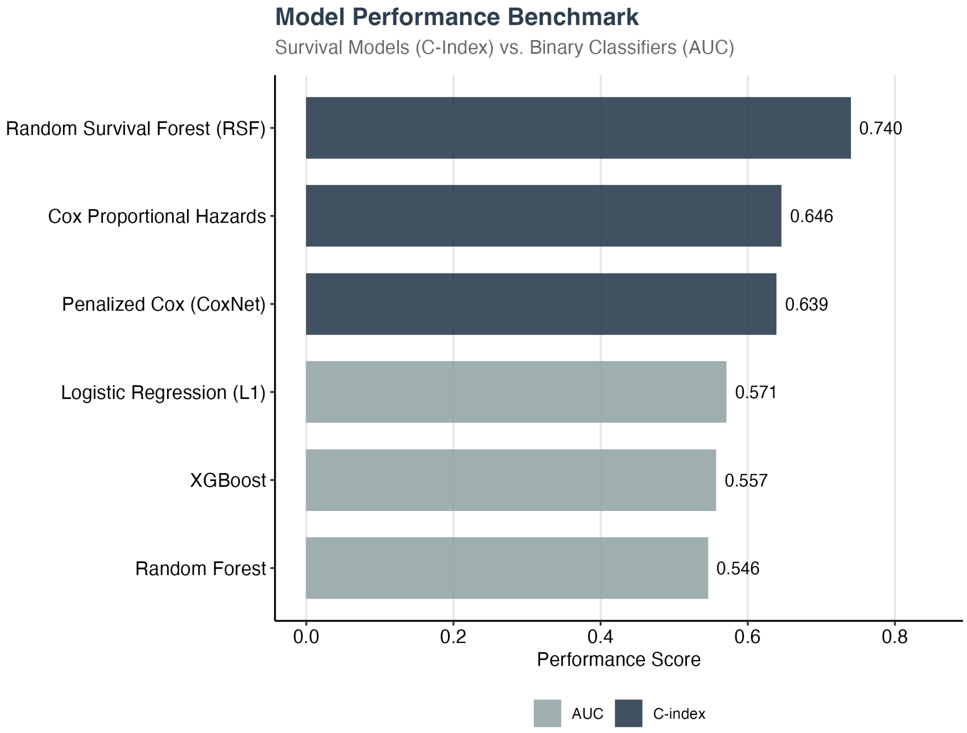


# 8 Model Performance Comparison

- The table and chart below summarize the performance across all modeling techniques.
- The Random Survival Forest demonstrated superior discrimination compared to both traditional Cox models and binary classifiers.

Table 2: Performance Benchmarking: C-Index (Survival Models) vs AUC (Binary Classifiers)

| Model | Metric | Value |
|---|---|---|
| Cox Proportional Hazards | C-index | 0.646 |
| Random Forest | AUC | 0.546 |
| Logistic Regression (L1) | AUC | 0.571 |
| XGBoost | AUC | 0.557 |
| Random Survival Forest (RSF) | C-index | 0.740 |
| Penalized Cox (CoxNet) | C-index | 0.639 |



# 9 Conclusion

This analysis demonstrates that Time-to-Event Machine Learning models (RSF) provide a quantifiable improvement over traditional statistical methods for HNSCC prognosis.

## 9.1 Strategic Implications:

1. **Methodology:** Binary classification proved insufficient for this clinical dataset due to heavy censorship and class imbalance. Survival Forests should be considered the standard for high-dimensional clinical survival data.

2. **Clinical:** While Tumor Stage remains the primary risk factor, Age and Primary Site contribute non-linear risk effects that are captured by ensemble methods but missed by linear Cox models.

3. **Future Work:** Integrating genomic data (e.g., TP53 mutation status) with this clinical RSF model could further enhance predictive precision.