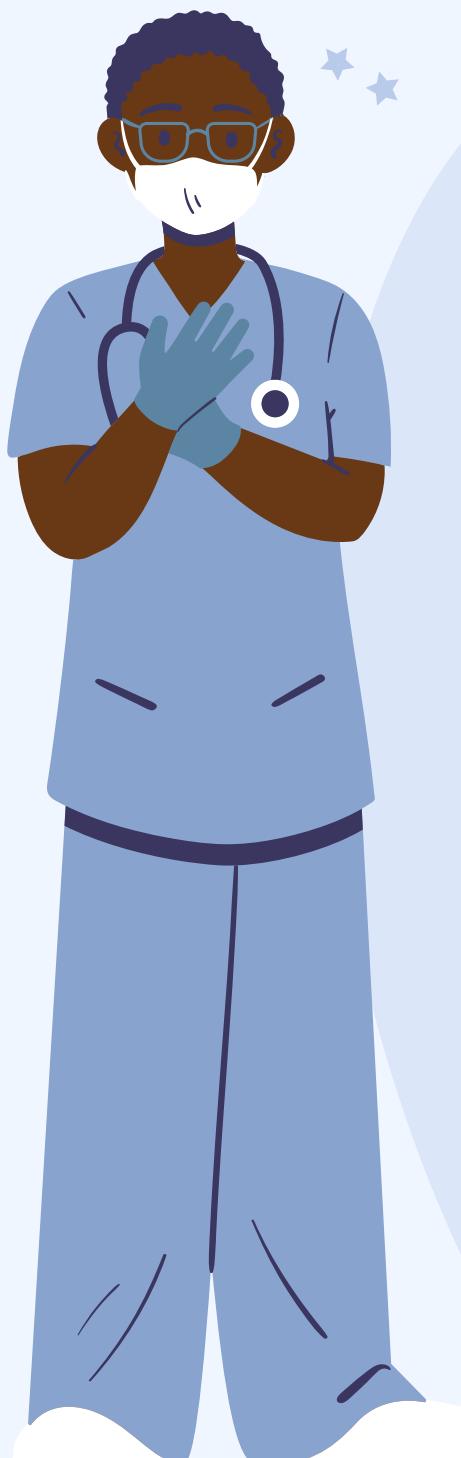


Jaime, Bernarda, Joao, Maria

Breast Cancer Survival Analysis



The Problem



Breast cancer

Is a devastating disease. Predicting patient survival rates is a complex challenge that medical expertise alone cannot guarantee, making it difficult for doctors to make treatments decisions.

Objective

Our objective was to develop a machine learning algorithm that is able to predict a patient's survival probability. This could help doctors better understand prognosis and plan more effective individual treatments.

Goal

The goal was to determine and evaluate the 4 distinct prediction models to choose the most effective one for our case, based on our dataset, our patients diagnosis and treatment information



Exploratory Data Analysis



Initial dataset

- 2,509 patients with 39 clinical features
- Key variables: Vital Status, Surgery Type, Chemotherapy, Hormone/Radio Therapy
- Removed non-essential identifiers (Patient ID, Sample ID)

Data cleaning strategy

- Replaced missing values with "Unknown" category
- Preserves sample size and represents real-world missingness
- Selected 14 clinical features for survival analysis

Missing data challenge

- >50% of features contain missing values
- Critical features have >500 missing values (>30% missing)
- Cannot remove data (creates bias) or train on missing values

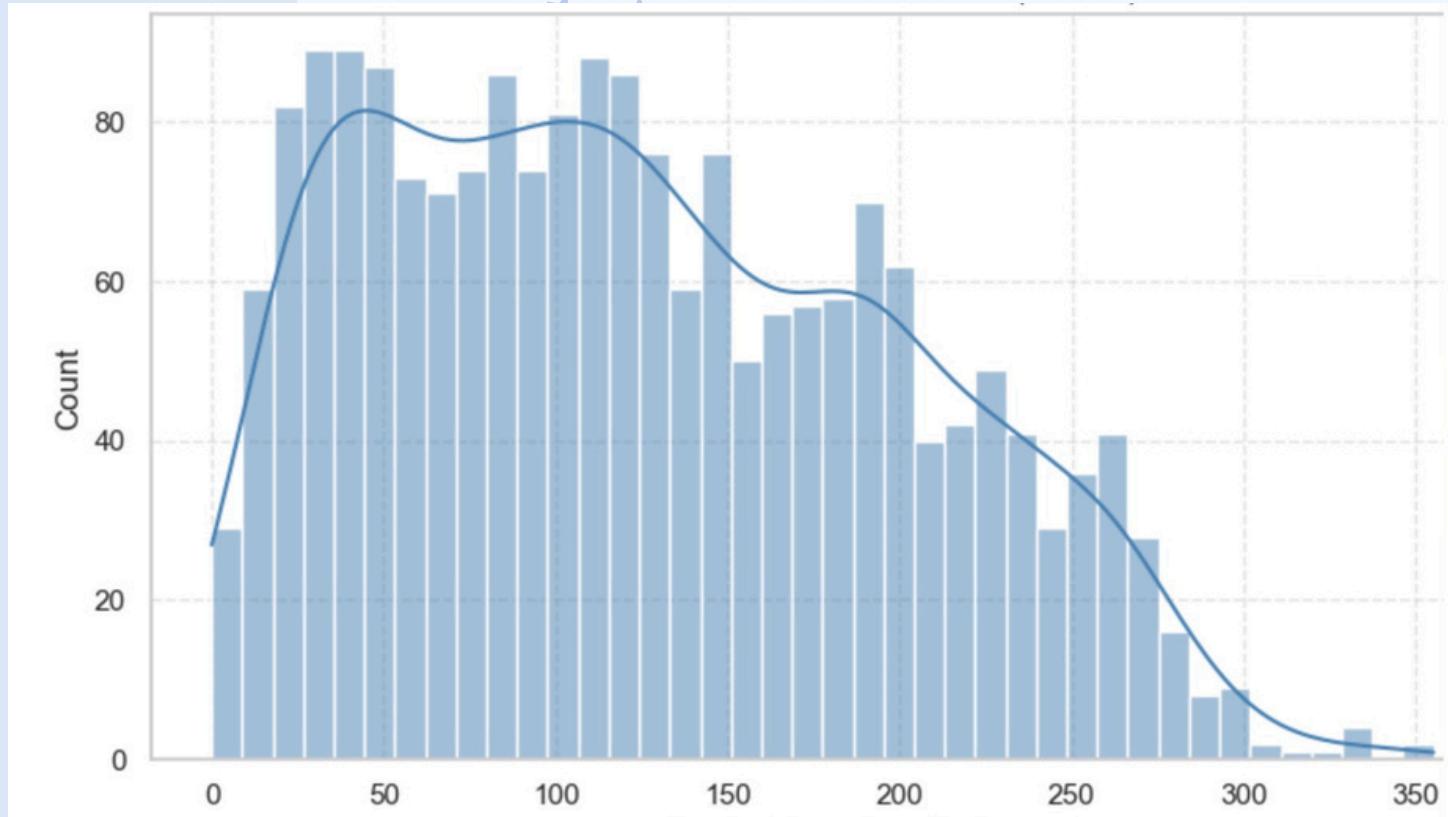


Final Result

- 1,981 patients in final cohort
- Clean dataset ready for model training
- No bias from data removal

Exploratory data Analysis

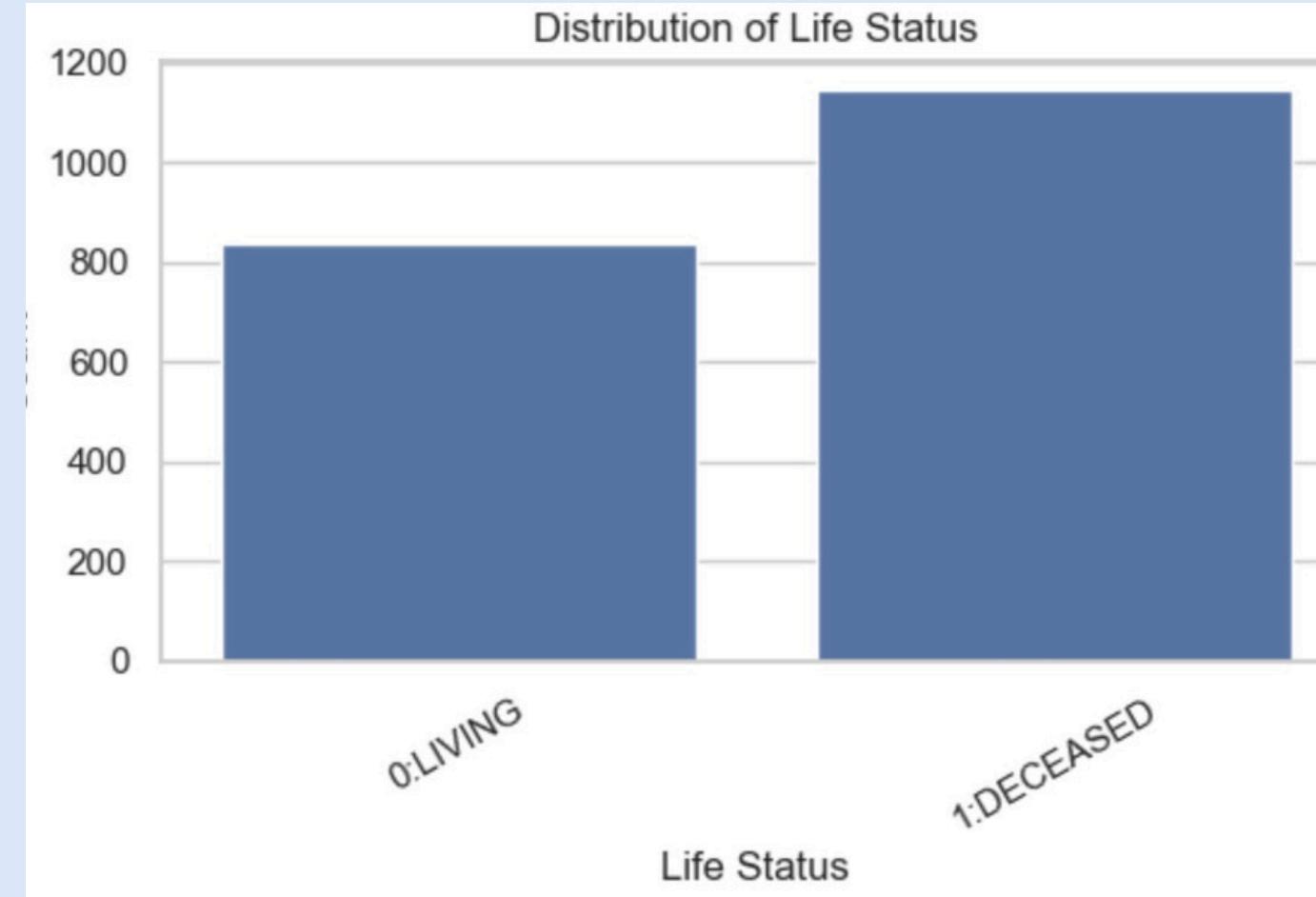
Survival Time Distribution



Data Distribution & Survival Patterns

- Time data displays **right tail** distribution
- Dataset contains **more deceased** patients than living patients
- **80%** training and **20%** test
- The two most important features are: **Overall Survival Status** and **Overall Survival months**

Event Distribution



Core Features

Tumor Stage: extent of disease progression (stage 1, 2, 3, 4)

Tumor Size: size of tumor in millimeters

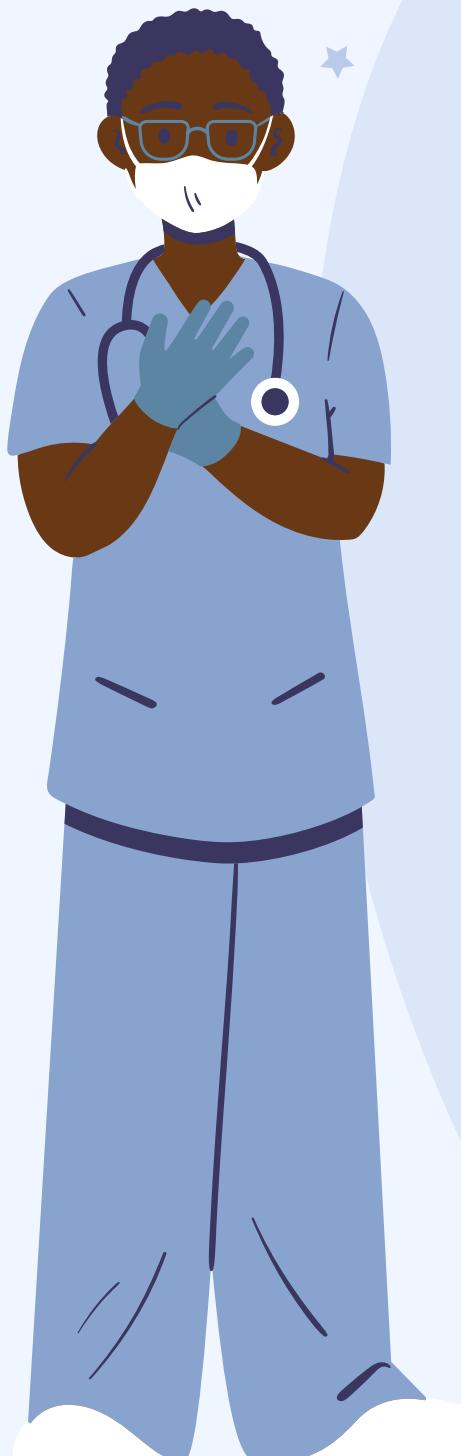
Age of Diagnosis: older ages is associated with different survival patterns

Chemotherapy: whether systematic chemotherapy was administered

Hormone Therapy: whether endocrine therapy was used.

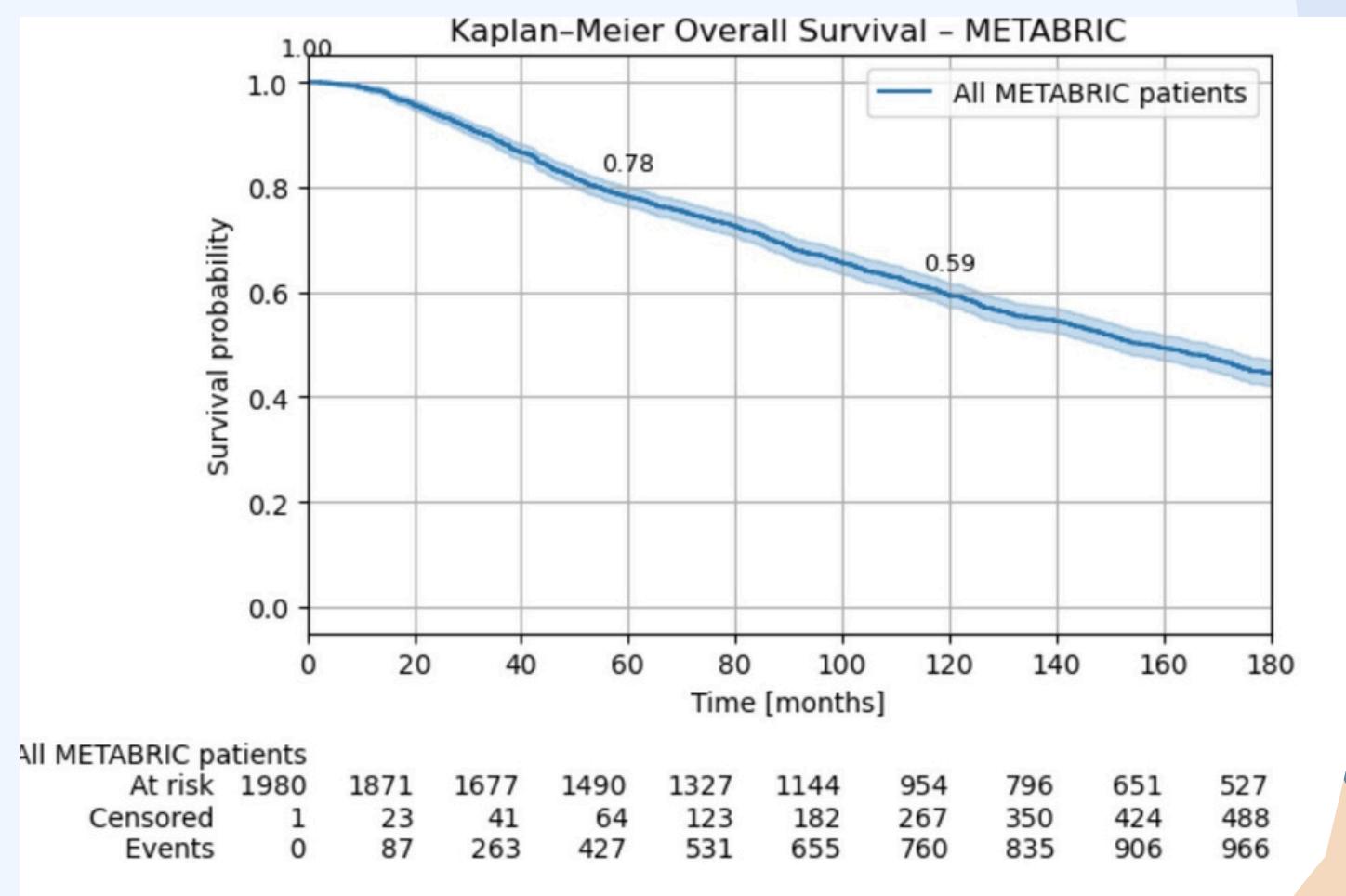
Radio Therapy: whether adjuvant radiation treatment was used

ER Status: whether cancer cells express estrogen receptors.

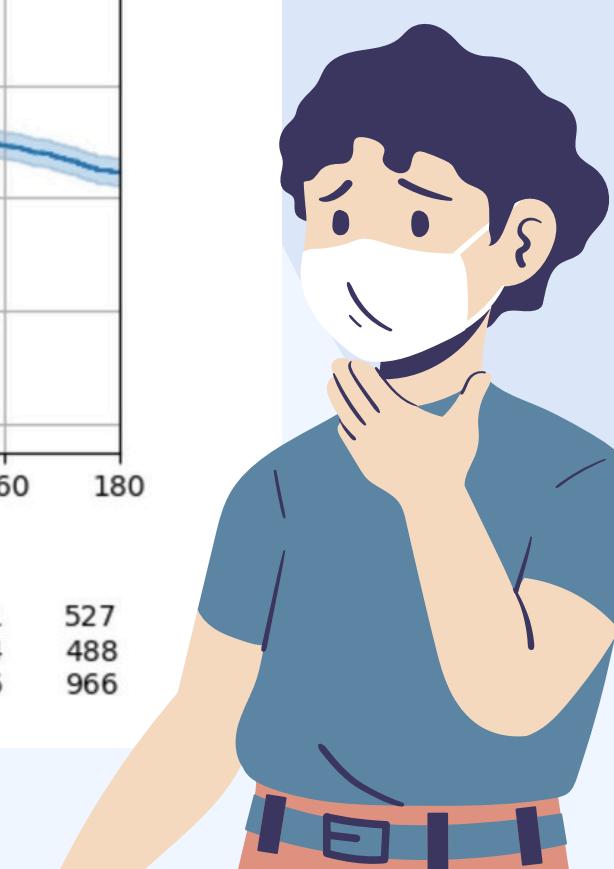


Kaplan-Meier Overall Survival Analysis

The Kaplan-Meier curve demonstrates the overall survival probability for 1,981 METABRIC breast cancer patients over a 180-month follow-up period, showing a gradual decline in survival that is consistent with typical breast cancer prognosis patterns.



- Median survival time: 156.3 months (~13 years)
- 5-year survival rate: 78%
- 10-year survival rate: 59%
- 15-year survival rate: 45%

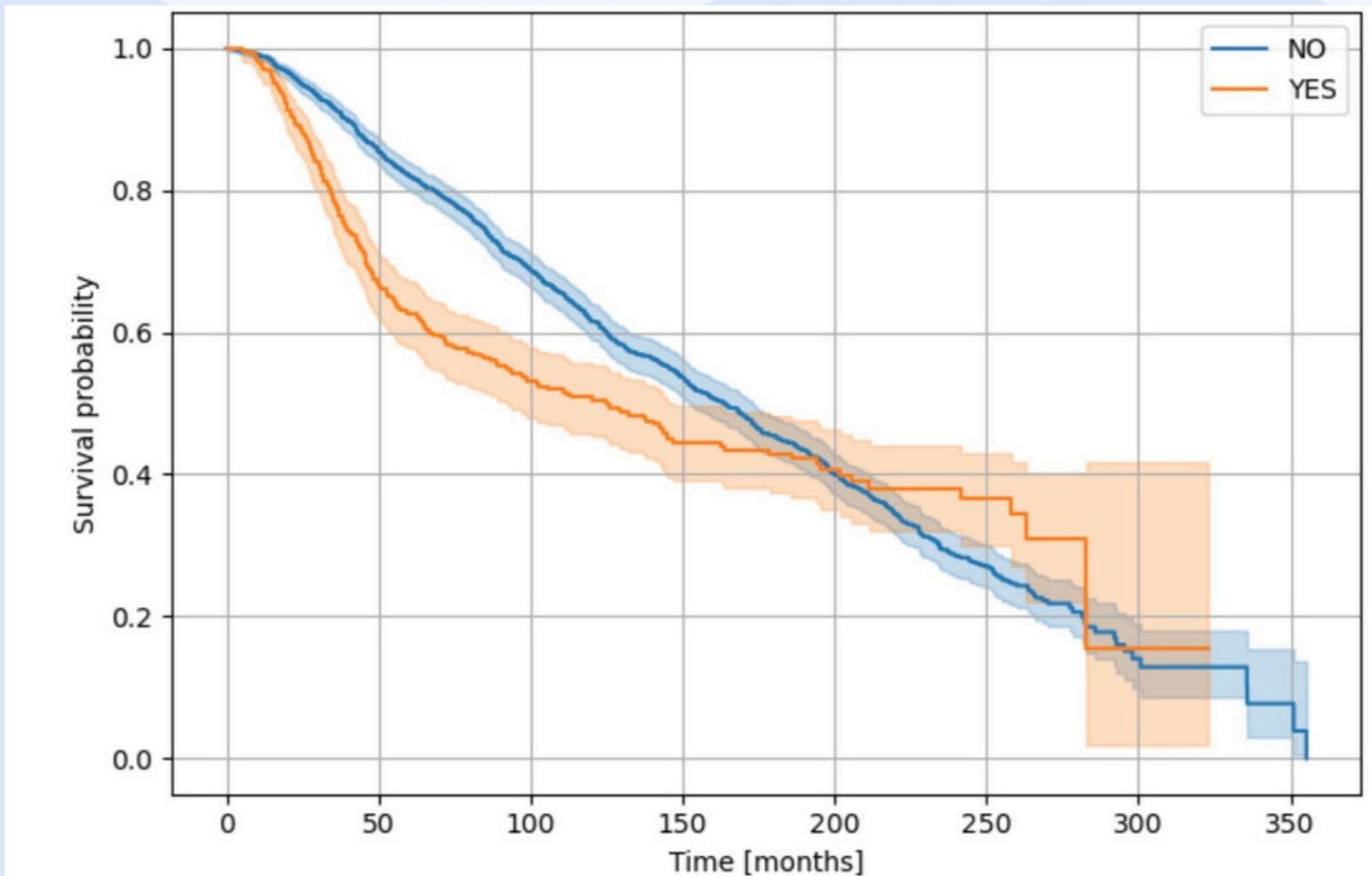




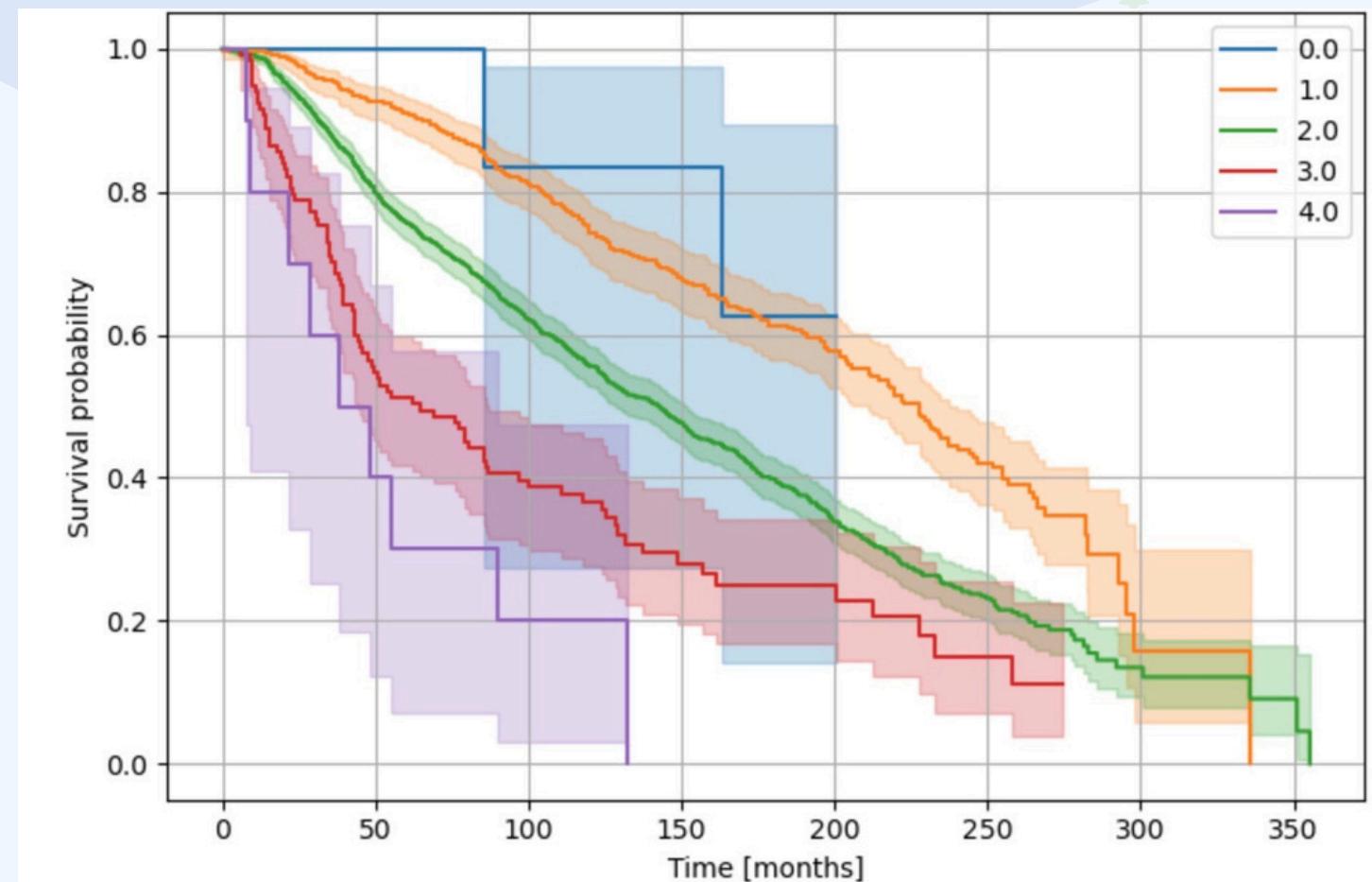
Chemotherapy and Tumor Stage Effects

Treatment decisions and tumor characteristics significantly influence survival outcomes, with tumor stage showing the strongest prognostic value.

Survival by Chemotherapy



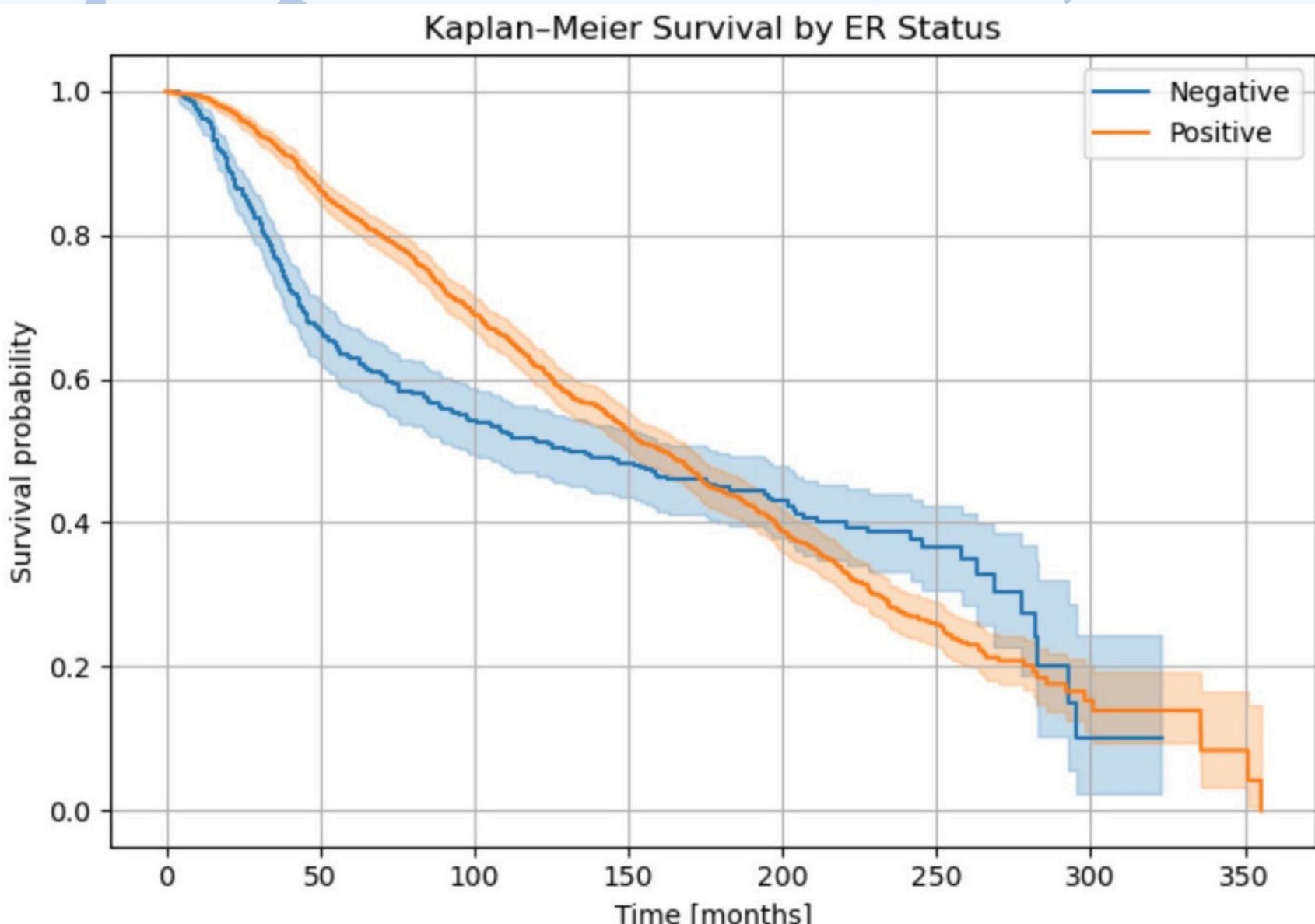
Survival by Tumor Stage



- Patients without chemotherapy show initially higher survival, reflecting treatment selection bias (sicker patients receive chemo)

- Stage 0: Excellent prognosis (>80% at 150 months)
- Stage 4: Poorest outcomes (all lost by 150 months)
- Stages 1-3: Progressive decline with severity

CR Status as Prognostic Biomarker



Estrogen receptor status serves as a critical molecular biomarker for prognosis and treatment planning.

- ER-positive patients show consistently better survival
- ER-negative patients have steeper initial decline
- Curves converge after 175 months due to censoring
- Validates ER status as essential for treatment decisions



Multivariable Cox Model

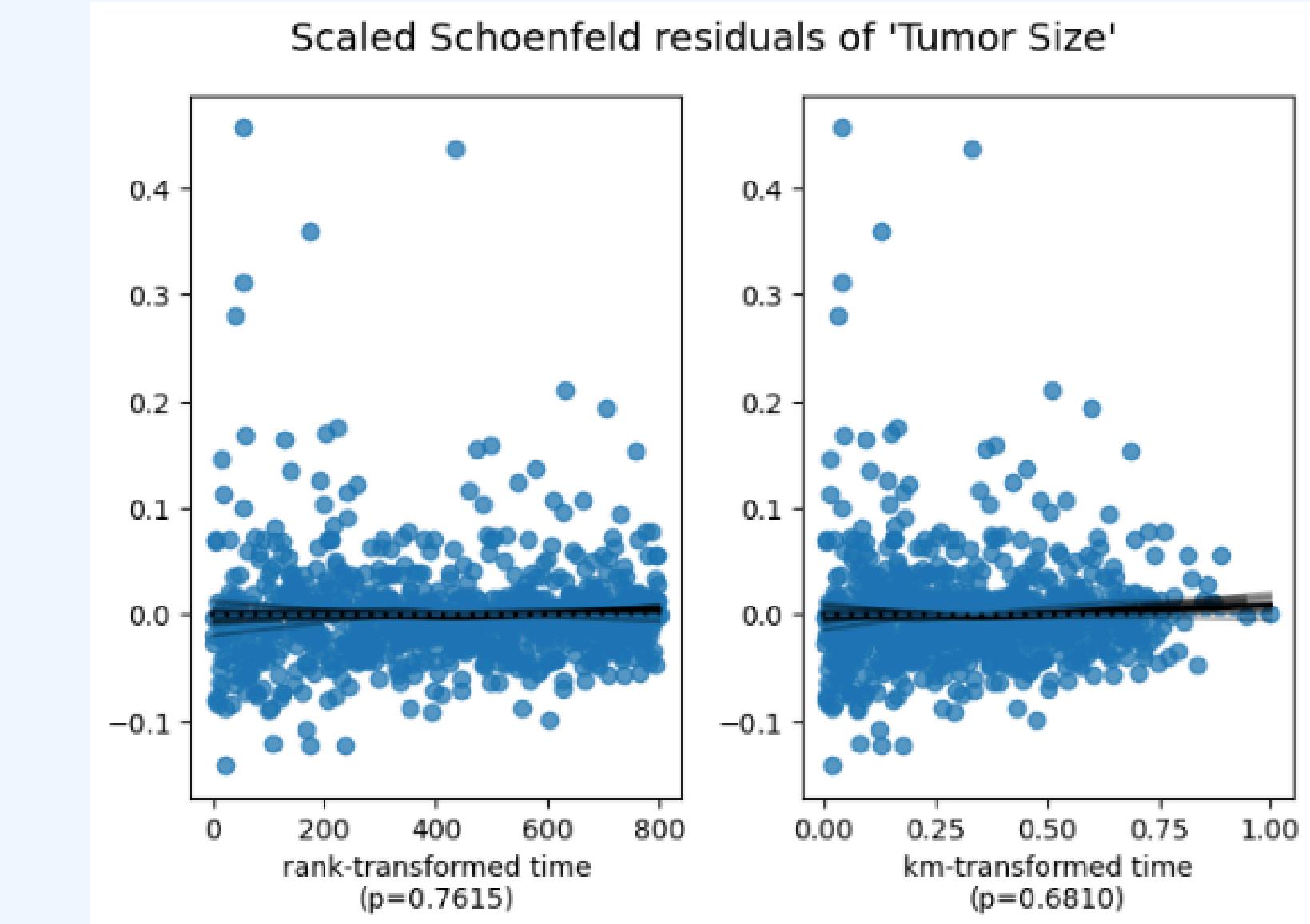
	coef	exp(coef)	exp(coef) lower 95%	exp(coef) upper 95%	p
Tumor Size	0.008	1.008	1.004	1.012	<0.0005
Tumor Stage	0.538	1.713	1.471	1.994	<0.0005
Chemotherapy	-0.155	0.857	0.690	1.063	0.160
Hormone Therapy	0.104	1.109	0.942	1.306	0.213
Radio Therapy	-0.230	0.794	0.688	0.917	0.002
ER Status	-0.180	0.835	0.687	1.016	0.071

Clinical Findings:

- Tumor Size: 1mm increase → 0.8% increased risk
- Tumor Stage: 1 stage increase → 71% increased risk
- Radio Therapy: associated with a 21% reduction in risk

Model concordance:

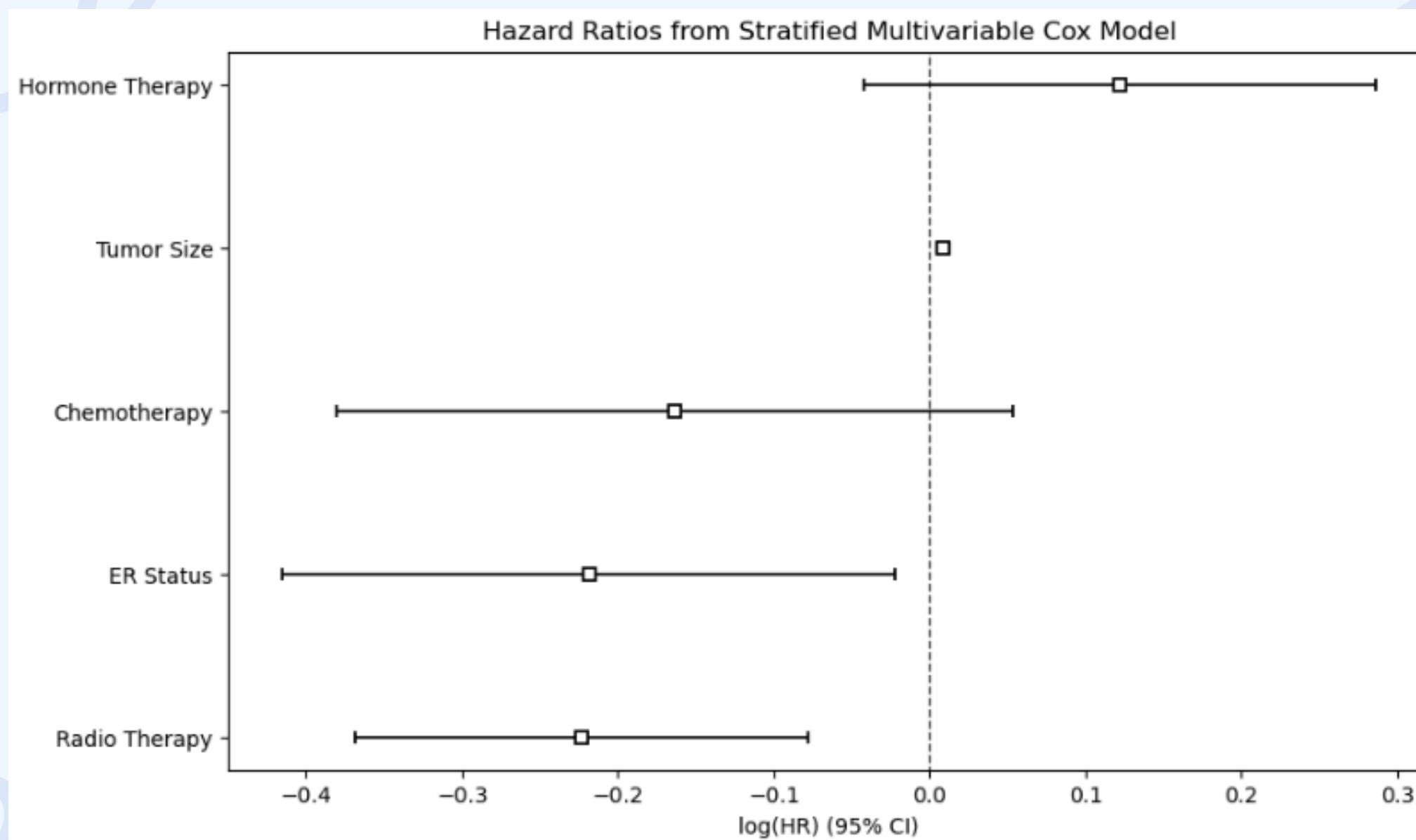
- Training: 0.628
- Validation: 0.609



Tumor Size variable failed the PH test. Decision was taken to stratify the variable since it is categorical and already shows excellent signs of being a predictor.

UNCOVER WHAT'S HIDDEN!

Multivariable Cox Model



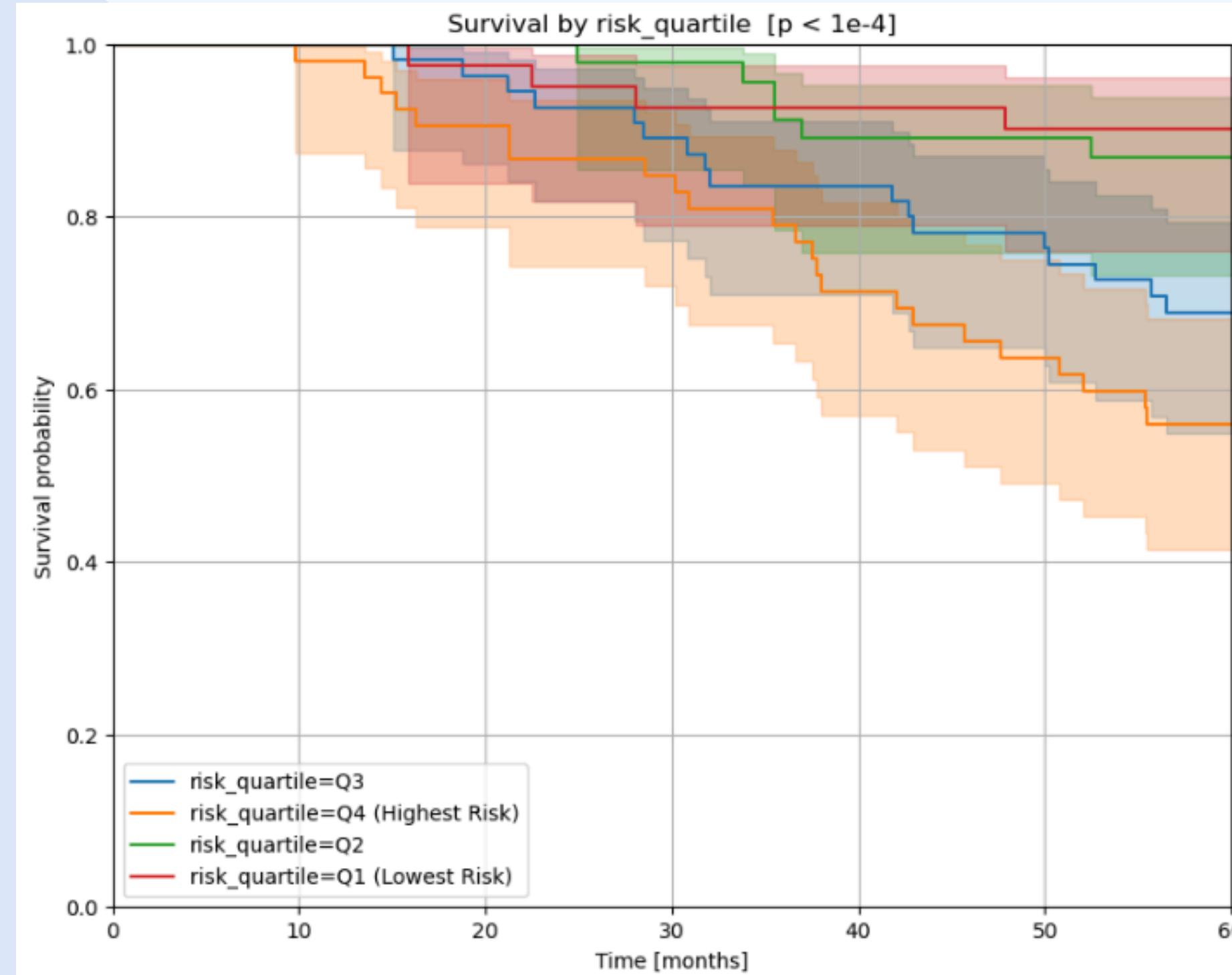
Despite the stratification uncovering an underlying effect we opted to use the original multivariable cox model because it had a higher concordance score. (Also some trial and error).

6.09 > 5.66



- Tumor Size and Radio Therapy remain statistically significant.
- ER Status is now **also statistically significant**.

Multivariable Cox Model



- Clear separation between Q1 and Q4, barely any overlap after **30-day mark**.
- The log-rank test ($p < 1e-4$) confirms these differences are statistically significant.



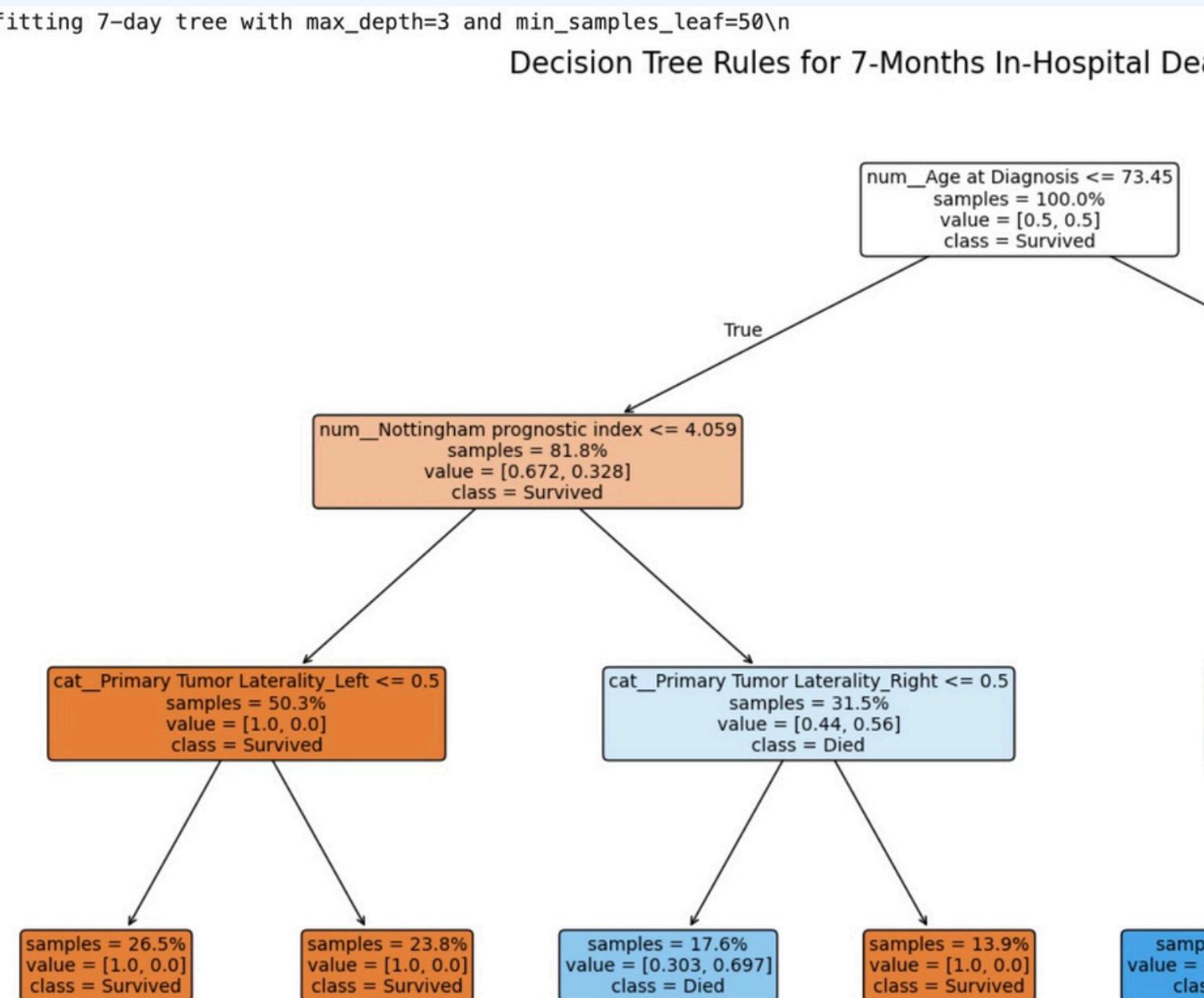
Multivariable Cox Model

Time Horizon (Months)	Time-Dependent AUC	Brier Score
7	nan	0.000
30	0.610	0.085
60	0.691	0.183

- In our dataset there are no deaths before the 7 month mark which cause the NaN AUC score.
- At the 30 month mark **AUC 0.610 shows that the model has a moderate discriminative ability** at predicting 30-day mortality. Low Brier score shows **good accuracy**.
- At the 60 month mark **AUC has improved significantly** but it comes at a cost of a higher Brier score, **lower accuracy**

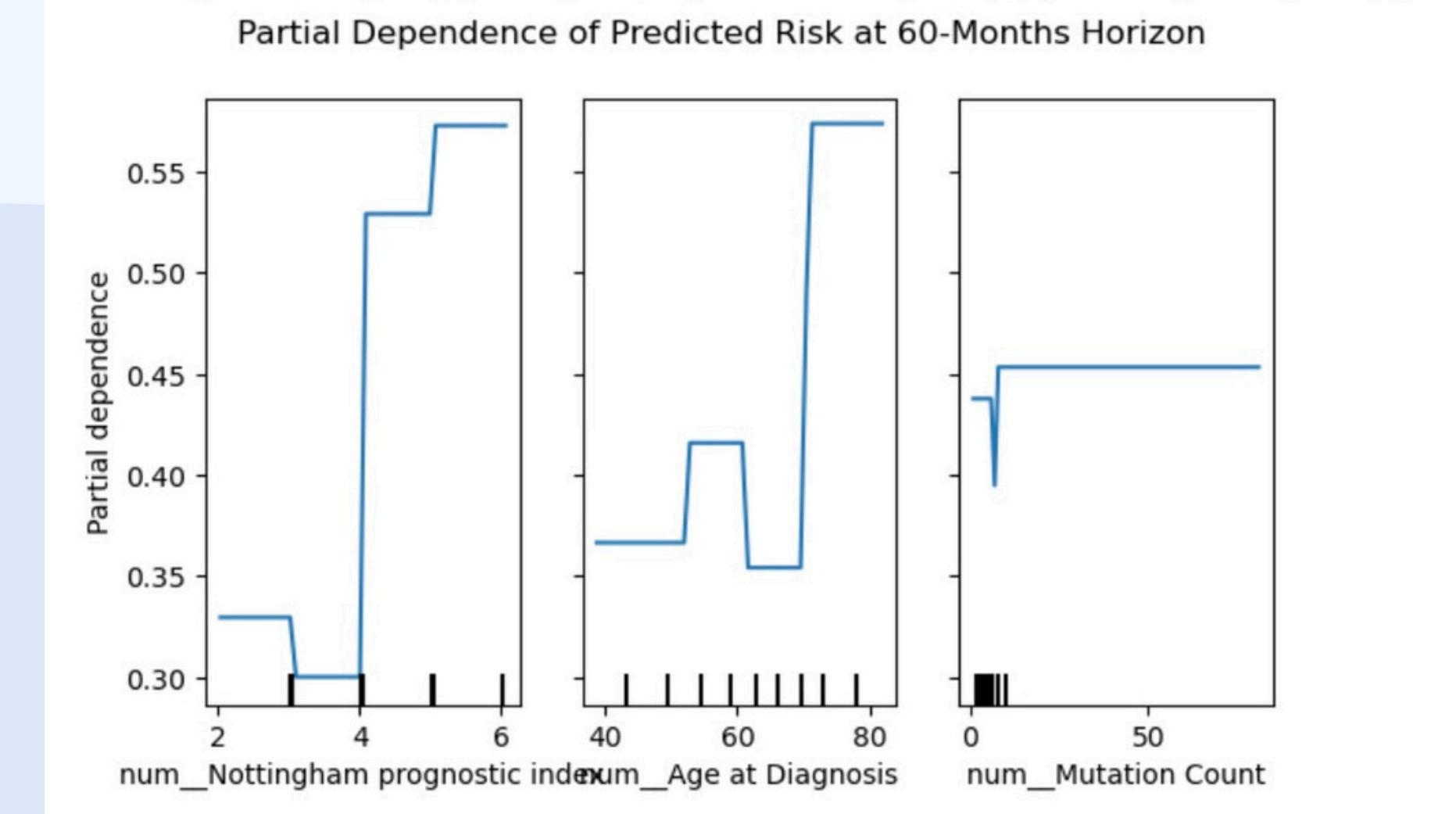
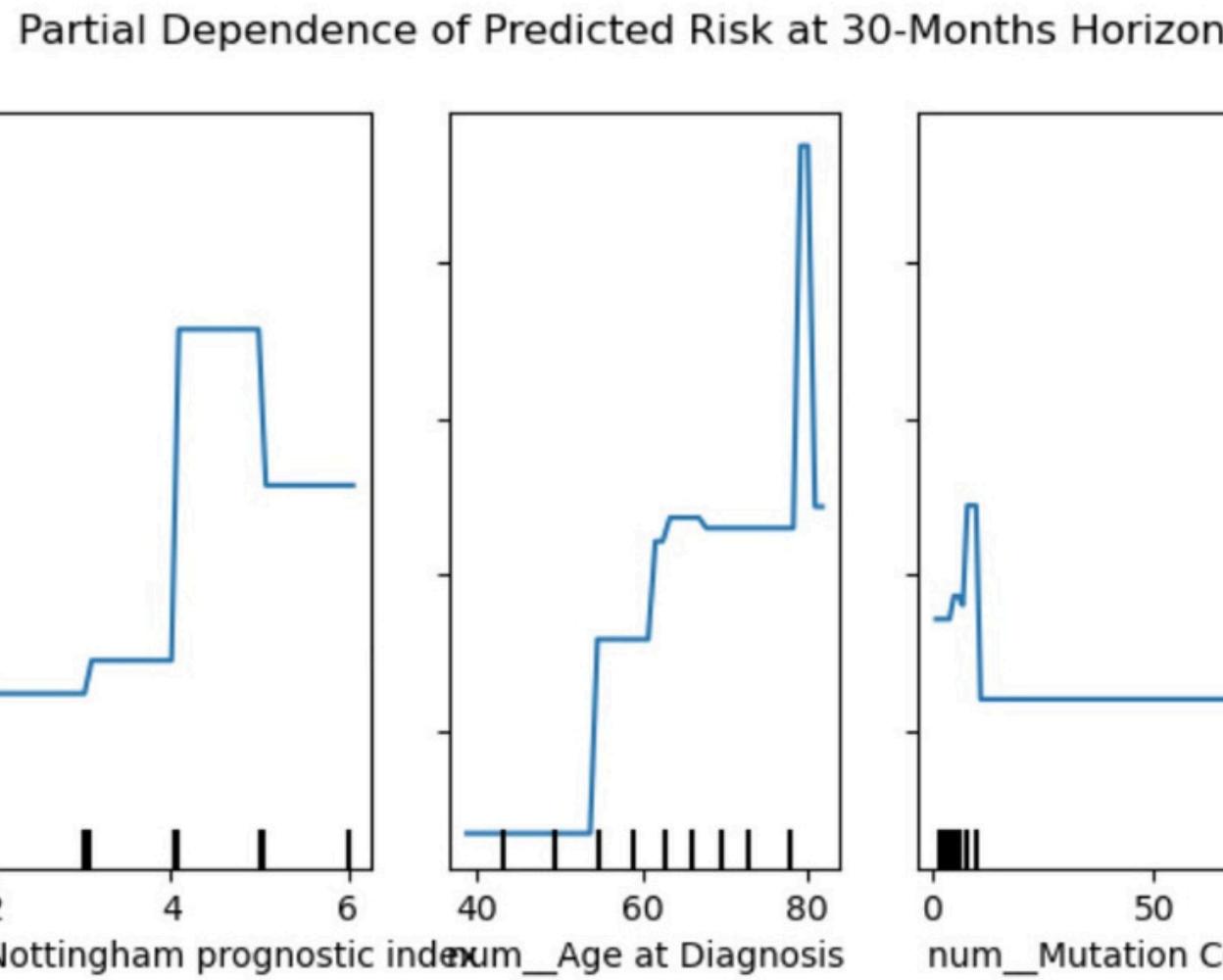
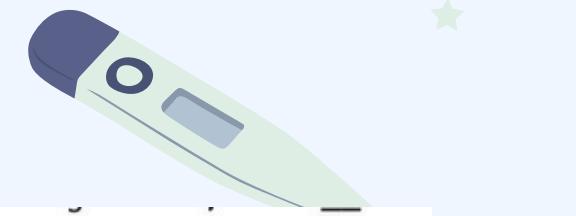
Decision Tree Model: Performance and Limitations

The Decision Tree model uses simple "if-then" rules to predict mortality risk at 7, 30, and 60-month horizons, prioritizing interpretability over predictive power.



	feature	importance
0	num_Age at Diagnosis	0.335482
5	num_Nottingham prognostic index	0.302794
1280	cat_3-Gene classifier subtype_ER+/HER2- High Prolif	0.183454
1264	cat_Primary Tumor Laterality_Right	0.178270
1263	cat_Primary Tumor Laterality_Left	0.000000
847	cat_Patient ID_MB-5384	0.000000
855	cat_Patient ID_MB-5396	0.000000
862	cat_Patient ID_MB-5409	0.000000
861	cat_Patient ID_MB-5408	0.000000
860	cat_Patient ID_MB-5407	0.000000
859	cat_Patient ID_MB-5406	0.000000
858	cat_Patient ID_MB-5403	0.000000
857	cat_Patient ID_MB-5402	0.000000
856	cat_Patient ID_MB-5398	0.000000
854	cat_Patient ID_MB-5395	0.000000

Decision Tree Insights: Risk Thresholds and Feature Importance

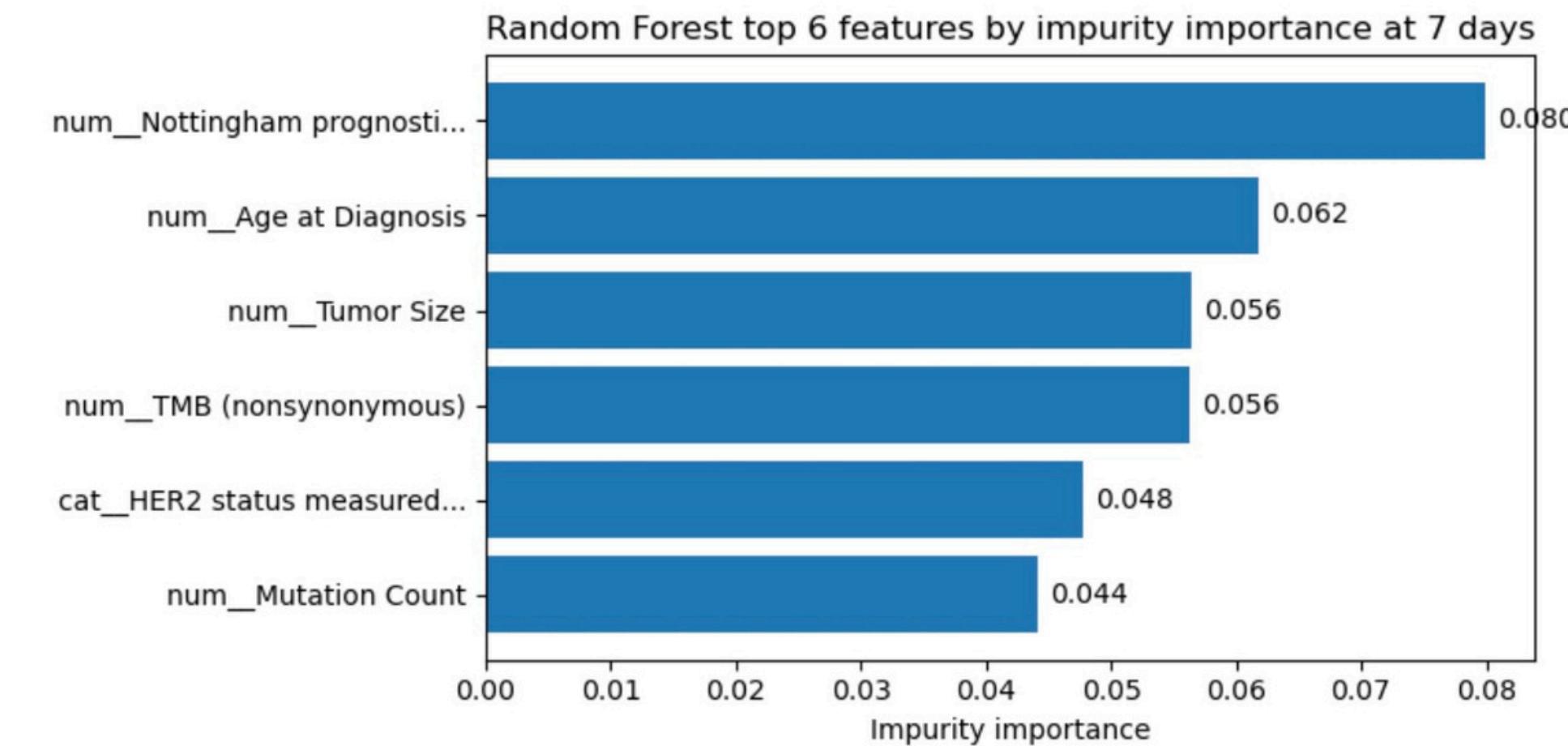
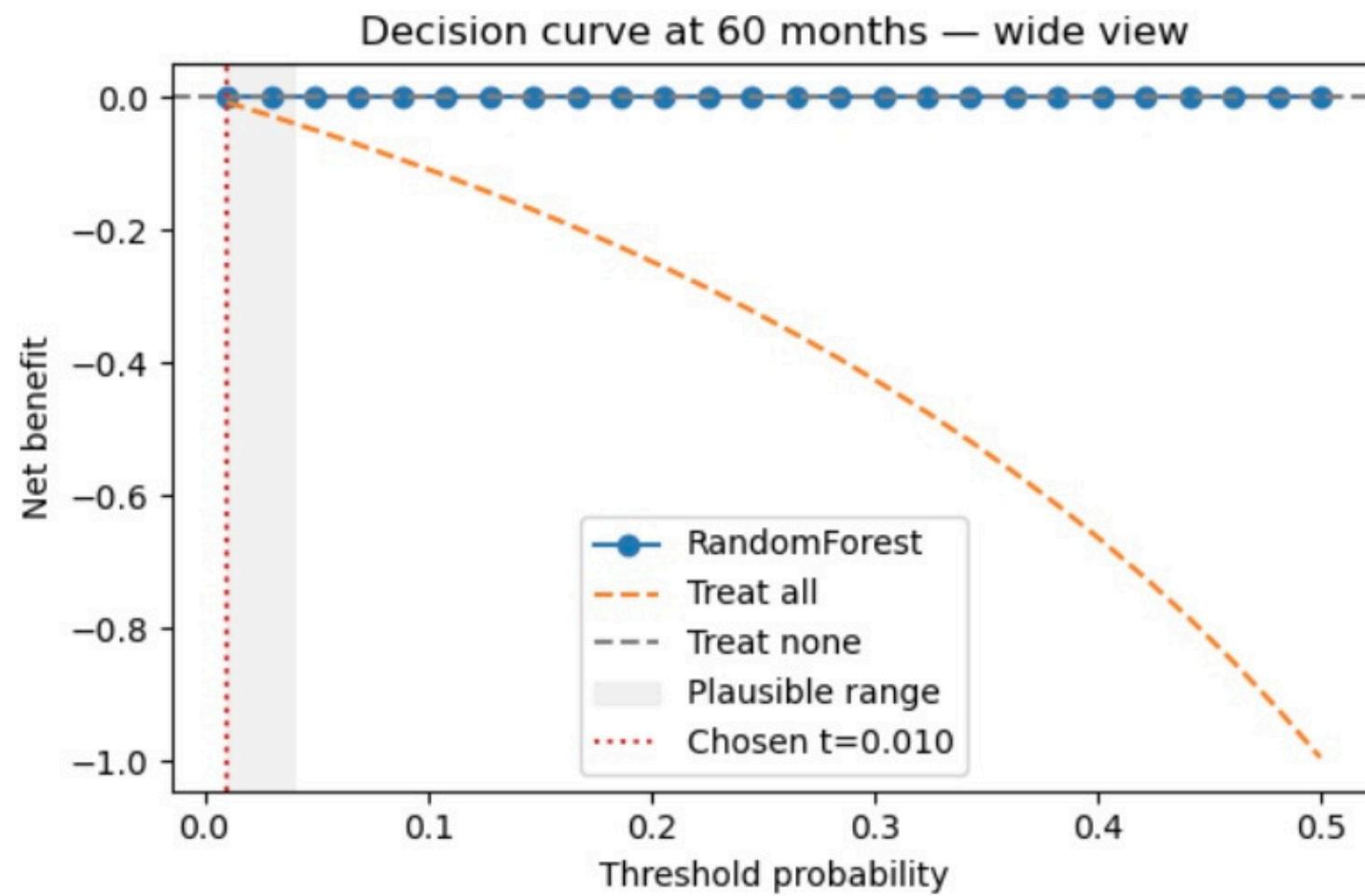


Critical Risk Thresholds

- Patients crossing either threshold warrant closer short-term monitoring
- Younger patients with low NPI show favorable short-term outlook
- Age and NPI effects become more pronounced at longer horizons (30-60 days)

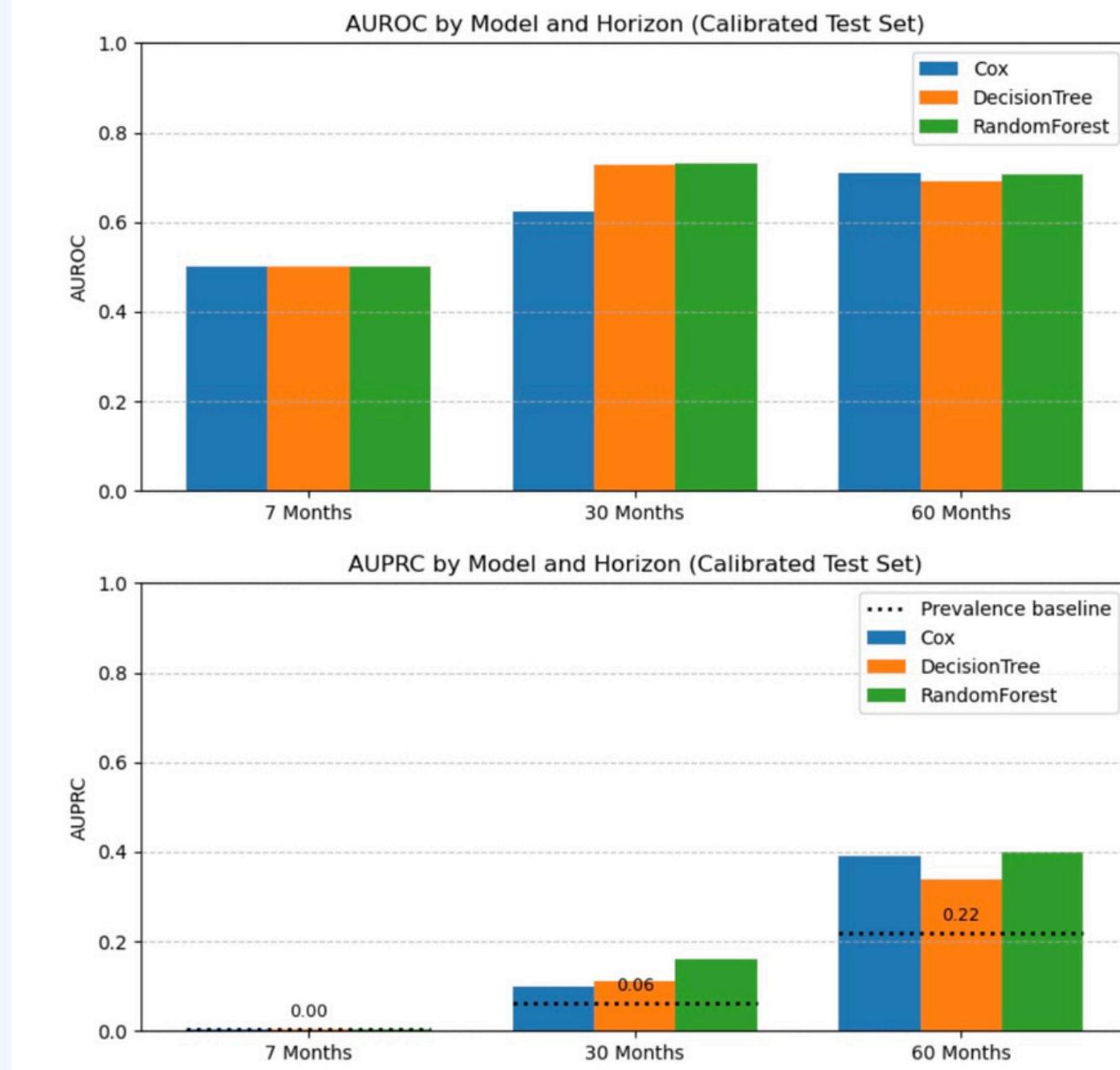
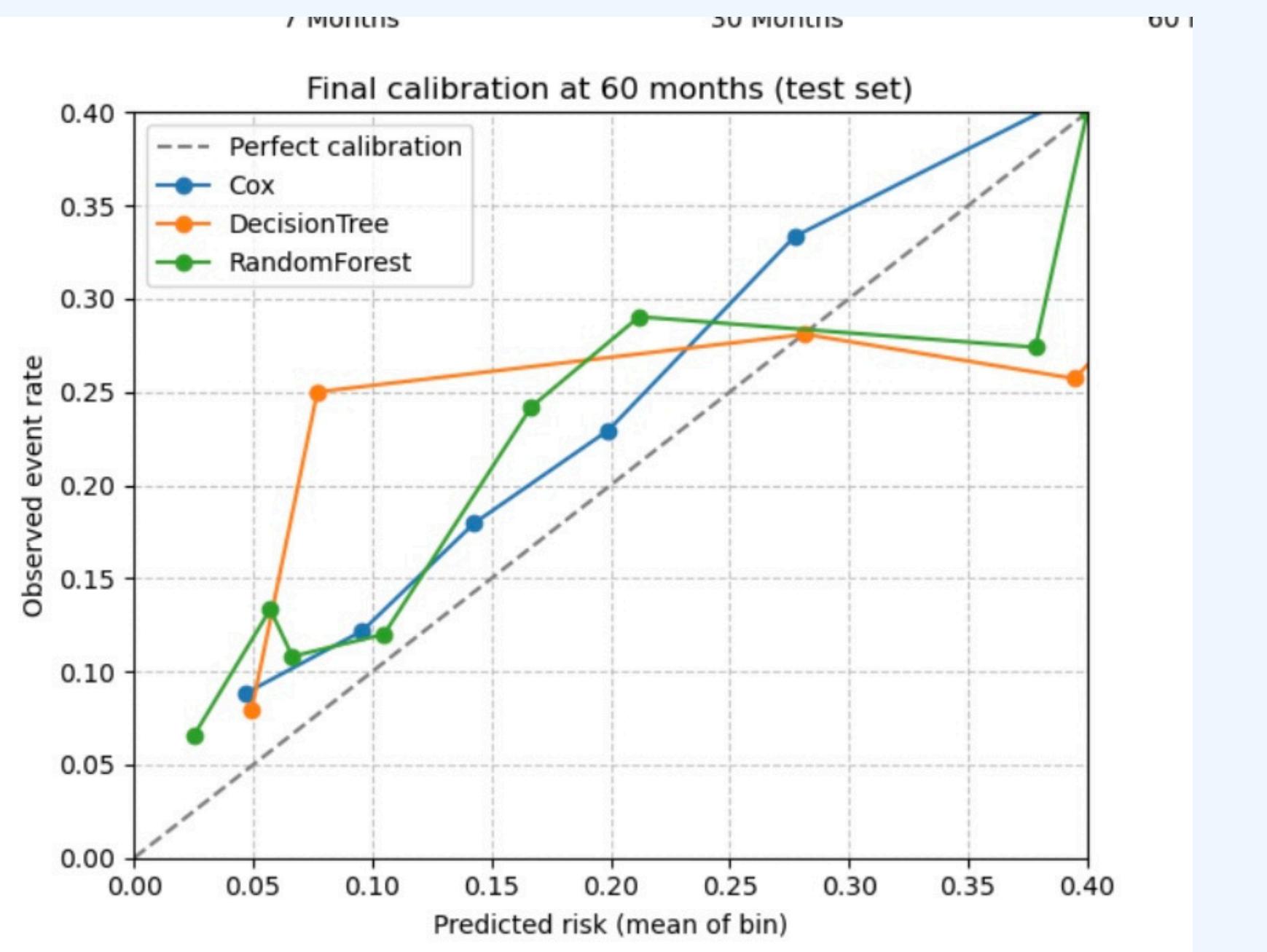


Random Forest Analysis



the performance metrics indicate strong predictive ability, with an AUROC of 0.798, showing good discrimination, and an AUPRC of 0.987, reflecting extremely accurate ranking of high-risk patients in a setting with relatively few mortality events.

Random Forest: Superior Predictive Accuracy



Performance Highlights

- 7-day AUROC: 0.95 (excellent short-term discrimination)
- 30-day AUROC: 0.73 (best among all models)
- 60-day AUROC: 0.72 (strong medium-term prediction)
- Better generalization than single Decision Tree
- Highest predictive power across all time horizons

The Trade-Off

Increased accuracy comes at the cost of interpretability. Unlike Decision Tree, Random Forest cannot provide simple "if-then" rules, making bedside interpretation more challenging.

Pros and cons Analysis

Cox Proportional Hazards

Pros:

- Excellent 7-month discrimination (AUROC 0.99)
- Clear high- vs. low-risk stratification ($p < 1e-4$)
- Clinically established methodology

Cons:

- Assumes proportional hazards
- Reduced performance at longer time horizons

Decision Tree

Pros:

- Highly interpretable "if-then" rules
- Clear risk thresholds (Age >73, NPI >4)
- Nearly matches RF at 30 month (AUROC 0.71)

Cons:

- Lower overall accuracy
- Worse calibration (Brier 0.196-0.220)

Random Forest

Pros:

- Highest accuracy across all horizons
- Best generalization
- Superior 30/60-month calibration

Cons:

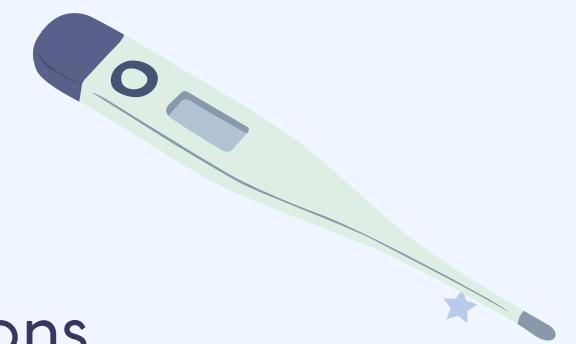
- "Black box" with limited interpretability
- Computationally intensive

Kaplan-Meier Survival Analysis:

- Clear visual survival patterns
- Identifies key milestones (5yr: 78%, 10yr: 59%)

Cons:

- Descriptive only—no individual predictions
- Unknown data could disrupt analysis



Implications for Clinical Practice

Our comparative analysis reveals critical implications for the decision making, resource allocation, and patient care.

Risk Stratification:

There are two primary risk thresholds identified: Age ≥ 73 and NPI >4 . Patients crossing either threshold require closer monitoring, while younger patients with low NPI can follow standard protocols.

Model Selection Impact:

- Random Forest: Best for high-stakes decisions requiring maximum accuracy (30-day AUROC 0.730)
- Decision Tree: Essential for patient communication and shared decision-making
- Cox Model: Integrates with established workflows, excels at 7-day predictions (AUROC 0.99)

Clinical communication

Interpretability trade off affects patient trust in the doctor and institution. Decision Tree enables transparent explanations, while Random Forest requires additional visualization tools.





So What?

Clinical recommendations

Model Selection Strategy

Choose the appropriate model based on clinical context and priorities.

Random Forest: Use for maximum accuracy in high-stakes decisions (AUROC 0.730 at 30 months)

Decision Tree: Use when interpretability is essential—patient communication and shared decision-making (AUROC 0.729, nearly matches RF)

Cox Model: Use for 7-month predictions (AUROC 0.99) and established clinical workflows

Kaplan-Meier: Use for exploratory analysis and understanding population survival patterns



There are critical steps required before our clinical deployment to ensure safety and effectiveness.

Validation

- External Validation: Test on independent cohorts from different institutions
- Prospective Trial: Compare model-guided care vs. standard practice
- Calibration Monitoring: Detect model drift over time

Implementation

- EHR Integration: Develop user-friendly interface with clear risk visualizations
- Clinician Training: Cover interpretation, use cases, and limitations

Ongoing Monitoring

- Bias Auditing: Ensure equitable performance across subgroups
- Threshold Optimization: Refine based on institutional resources and patient preferences

Next Steps for Deployment



Thank you for your attention

