Breast Cancer Survival Analysis

Jaime, Maria, Bernarda, João

Machine Learning for Healthcare

Project: Survival analysis

Use Case: Breast Cancer

Source: METABRIC - https://www.cbioportal.org/study/summary?id=brca_metabric

Problem

Breast Cancer has been a horrific disease affecting individuals and families for years. Hospitals struggle to effectively understand the necessary requirements to ensure a patient's survival. Determining the patient's survival rate is a challenge that doctors alone can not guarantee

Objective

Build a machine learning algorithm that can accurately predict the patient's survival rate based on the several factors that contribute to the patient's health and/or status. The algorithm will allow doctors to know the likelihood a patient will survive given the characteristics of the disease, the treatment used and other factors regarding health.

Goal

Create three different models that can predict the survival rate of the patients given the variables related to the patients health, diagnosis and treatment and compare the performance of all three to determine the most effective model for survival analysis.

Exploratory Data Analysis:

The dataset contains 39 columns (features) and 2509 rows (values). A great amount of data of data with some variables that will not be necessary for our analysis such as Patient ID and Sample ID while others such as Patient Vital Status, Type of Breast Surgery, Relapse Free Status, Chemotherapy and so on are valuable information that our models will require in order to predict survivability.

Unfortunately the dataset suffers from a great deal of missing values. More than half the features contain missing values with many of the most important features having over 500 missing values, more than 30% of the values of the variables are missing. We can not train

our model on missing values but we also cannot simply remove all the N/A values of these variables as there would be a great deal of bias or imbalance towards certain variables. To fix this we replace the missing values with unknown, an often meaningful category to represent real-world missingness. This allows us to train our models without destroying the sample size and creating bias results. After we perform this while eliminating the values that we will not require, our dataset contains 1981 values and 39 features. We begin splitting our model with 80% going into the training set and 20% into the testing set and capture the two most relevant features in our dataset: Overall Survival Status and Overall Survival months.

Core Variables:

Variable: Overall Survival (Months)
Meaning: How long each patient was followed until death or last known followup, measured in months.

Variable: Overall Survival Status/Patient's Vital Status
Meaning: Indicates whether a patient died (event = 1) or is still alive/censored (event = 0) at the end of follow-up.

Variable: Tumor Stage
Meaning: Extent of disease progression (stage 1, 2, 3, 4). Higher stage = worse prognosis.

Variable: ER Status
Meaning: Whether cancer cells express estrogen receptors. ER plus tumors tend to respond to endocrine therapy and have better outcomes.

Variable: PR Status
Meaning: Hormone receptor expression. Progesterone Receptor which coincides with ER often predicts improved prognosis

Variable: HER2 Status
Meaning: Overexpression of Human Epidermic Growth Receptor 2+. Normally HER2+ has poor prognosis, but treatment can dramatically improve outcomes

Variable: Cellularity
Meaning: How packed the tumor cells are ("High", "Moderate", "Low"). High cellularity leads to more aggressive tumors

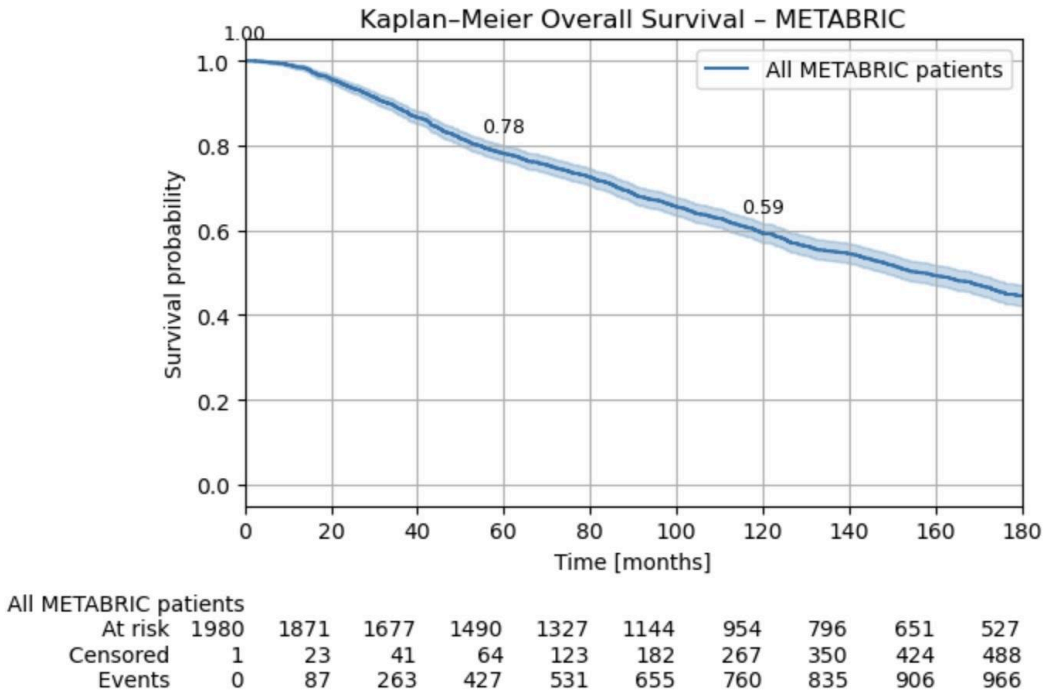
Variable: Tumor Size
Meaning: Size of tumor in millimeters. Larger tumors give worse prognosis.

Variable: Chemotherapy
Meaning: Whether systemic chemotherapy was administered. Higher-risk patients often receive chemo.

Variable: Type of Breast Surgery

Meaning: Whether the patient received breast-conserving surgery or Mastecomy. Survival curves often differ by treatment type.

Kaplan–Meier curves Model

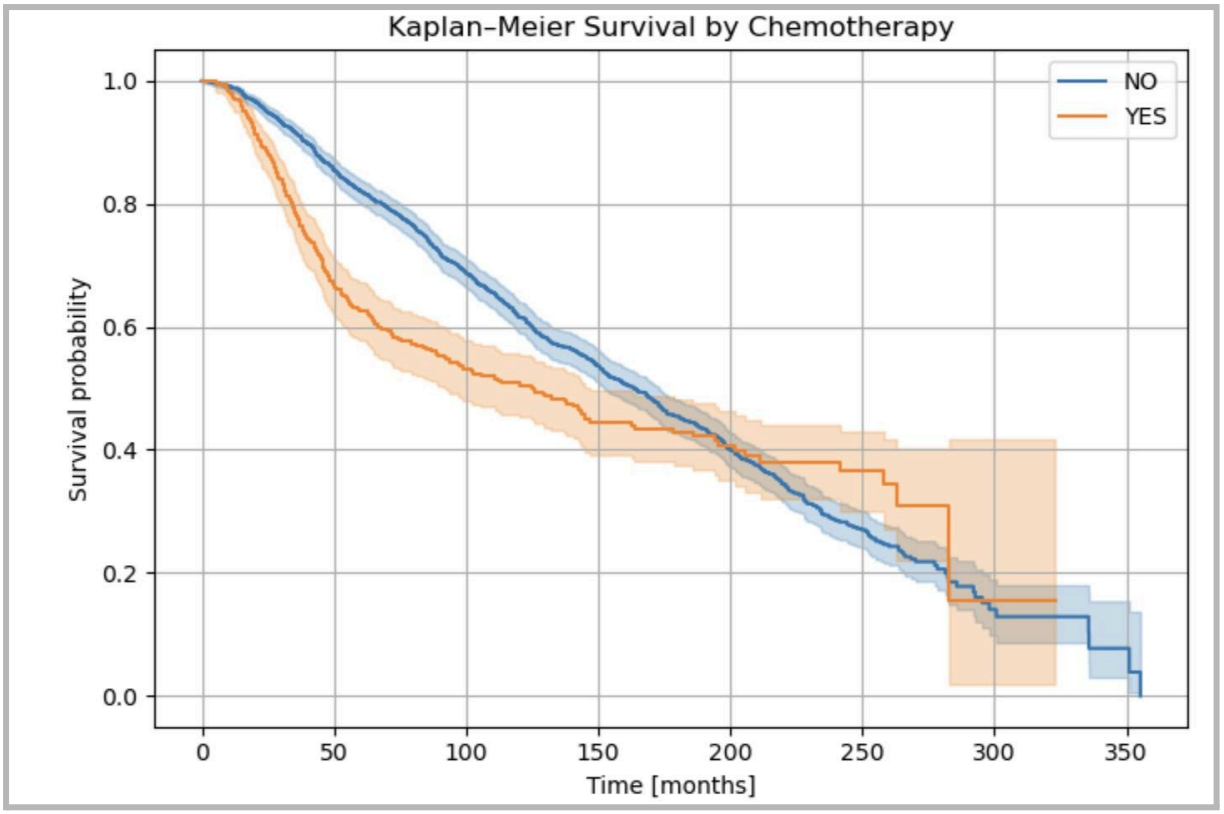


All METABRIC patients										
At risk	1980	1871	1677	1490	1327	1144	954	796	651	527
Censored	1	23	41	64	123	182	267	350	424	488
Events	0	87	263	427	531	655	760	835	906	966

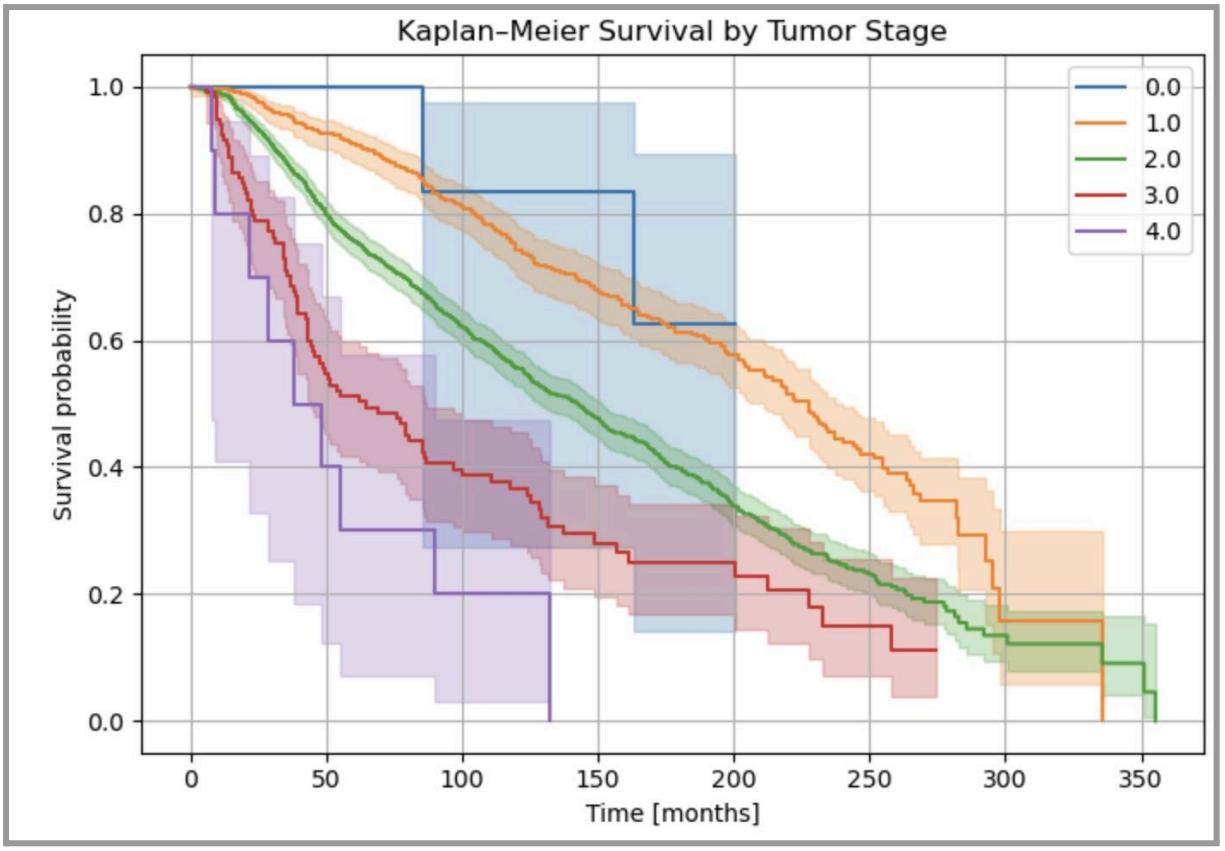
Using the Kaplan Meier model we analyze the overall survival probability among the patients up to 180 months. The survival rate gradually decreases in the course of time as in 60 months (5 years) the survival rate is 78%, in 120 months (10 years) the survival rate is 59% and at 180 months (15 years) about 45% remain alive.

The model also provides us with the amount of patients who are at risk, the patients who are alive at their last follow up or were lost to follow up (censored), and the patients who were confirmed dead from the cancer (events) over the months. The patients at risk gradually decrease as they are either successfully treated, lost or have passed away.

We also capture the Median Survival Time which for patients in METABRIC is ~156.3 months and the Restricted Mean Survival Time which we found out is on average ~ 54 patients survive up for 60 months (5 years), ~ 95 patients survive up for 120 months (10 years), and ~ 126 patients survive for 180 months (15 years).



The first subgroup shows the survival rate of patients who have been treated with chemotherapy. The patients without chemotherapy treatment have a higher survival rate than those with treatment up until 200 months where the survival rate of the patients with treatment begins to plateau indicating that there could be either there are cases of a patient's status becoming unknown or the patient was discharged.

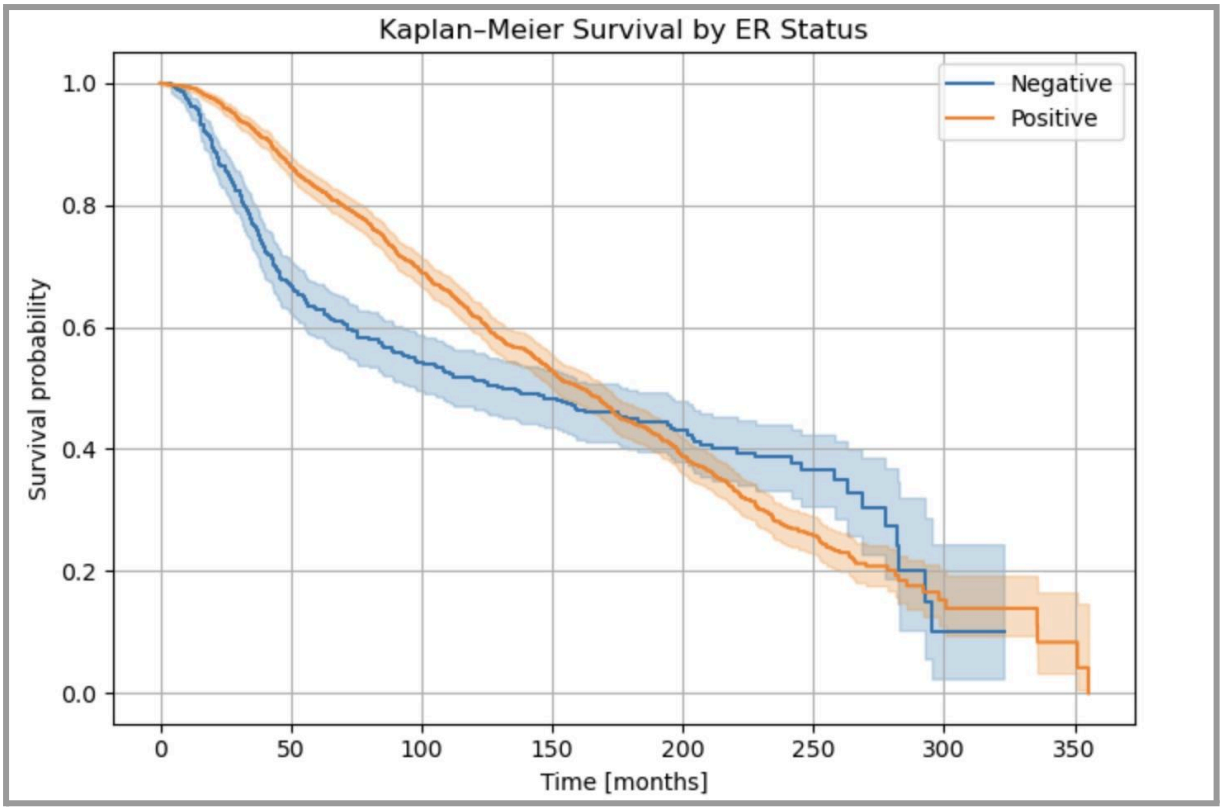


The second subgroup provides the survival rate of patients with different tumor stages. here all types of patients start decreasing gradually with the exception of stage 0, stage 0 tumors are uninvasive meaning it has a better prognosis.

We can find that patients with stage 0 breast cancer hardly die as more than 80% last 150 months and by 200 months more than 60% survive with the graph no longer tracking them, this does not mean that by 200 months patients that survive will no longer die to breast cancer rather it means by this time they were either lost, discharged or simply it was their most recent survival confirmation which is the mostly given the plateaus._

Another stage to understand is the stage 4 breast cancer, as patients with this tumor stage have the lowest survival rate. Before 150 months all the patients with this stage are either dead or lost due to the various plateaus.

For the rest of the patients stages 1, 2, and 3, the survival rate decreases with few plateaus (unknown data) although stage 3 patients are no longer tracked after 250 months by then they have less than 20% chance of survival and stages 1 and 2 continue until they are all deceased or lost with stage 1 patients reaching a survival rate slightly higher than stage 2 and 3 patients but still less than 20% at 300 months



The last subgroup allows us to see the survival rate of the patients based on their Estrogen Receptor status. As time goes on the survival rates of both types of patients decreases gradually however the patients with a negative status or without endocrine therapy have lower survival rate than patients whose cancer cells express estrogen receptors or positive status. However, until 175 months pass, then patients without endocrine therapy have a slightly higher survival rate until they plateau at 300 months revealing either discharge or loss of patient data.

Calibration:

Using the three cohorts 60 months (5 years), 120 months (10 years), and 180 months (15 years), we are able to understand how stable the calibration of the model is. We do this by using the metric RMST (Restricted Mean Survival Time) and comparing them across the training and testing set. We found that on average patients survive ~54-55 months in the first 60 months, patients survive ~ 96 months out of the first 120 months, and patients survive ! 127 months out of the first 180 months. Seeing how the RMST values are nearly identical in both training and testing sets among the three cohorts indicates that the model has calibrated effectively with low variability between splits and strong stability for the survival estimates.

Cox Proportional Hazard Model

For the Cox Proportional Hazard Model the decision was to use a 70-20-10 train, validate, test split, this would allow us to tune our model ensuring there was no data leakage by isolating the last 10% for testing only.

```
--- Data Splitting Report ---
Training set   : 1386 rows (70%)
Validation set : 398 rows (20%)
Test set       : 197 rows (10%)
```

model	lifelines.CoxPHFitter
duration col	'time'
event col	'event'
penalizer	0.01
l1 ratio	0.0
baseline estimation	breslow
number of observations	1386
number of events observed	800
partial log-likelihood	-5147.935
time fit was run	2025-12-04 19:45:22 UTC

	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

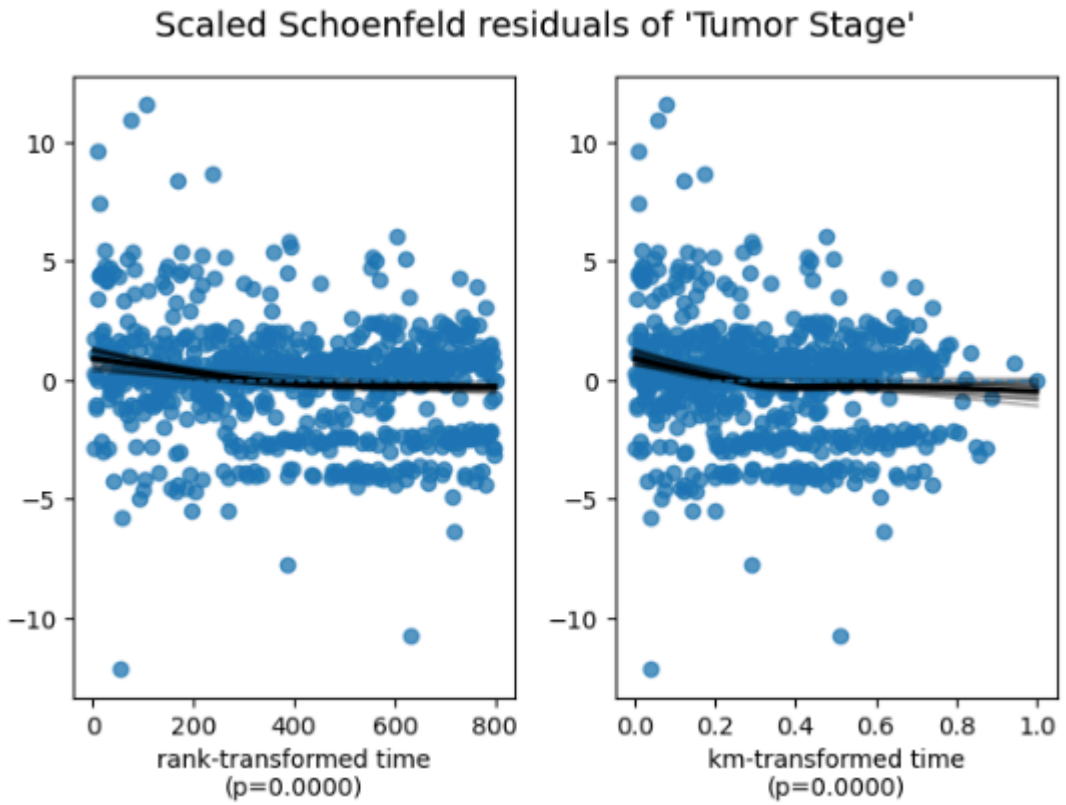
Concordance	0.628
Partial AIC	10307.871
log-likelihood ratio test	115.886 on 6 df
-log2(p) of ll-ratio test	72.831

```
Concordance | train 0.628 | validation 0.609
```

We fit a multivariable cox proportional hazard model to our data and obtain the above results. The model shows a concordance score of 0.628 in the training set and 0.609 in the validation set, a moderate score. The significant variables, those with a p value inferior to the threshold of 5%, are Tumor Size, Tumor Stage and Radio Therapy.

The "*exp(coef)*" value tell us the effect of each variable on the "death chance" increase per unit increase of the variable. For tumor size (1.008) the value shows that for each 1mm increase, there is a 0.8% increase in the mortality. This may seem small but this means a 10mm increase leads to an 8% increase in mortality and tumor sizes can vary from 10-70mm.

For tumor stage a one level or one stage increase means a 71% increase in mortality, to be expected. Radio Therapy was the only therapy type that passed our significance threshold and shows a 21% decrease in mortality.



Next was to take a look at the PH test for all of the variables while mainly keeping an eye on our significant variables. Tumor size and radio therapy passed the PH test but tumor stage did not. The above graph is the tumor stage PH test with a p value of 0.6810.

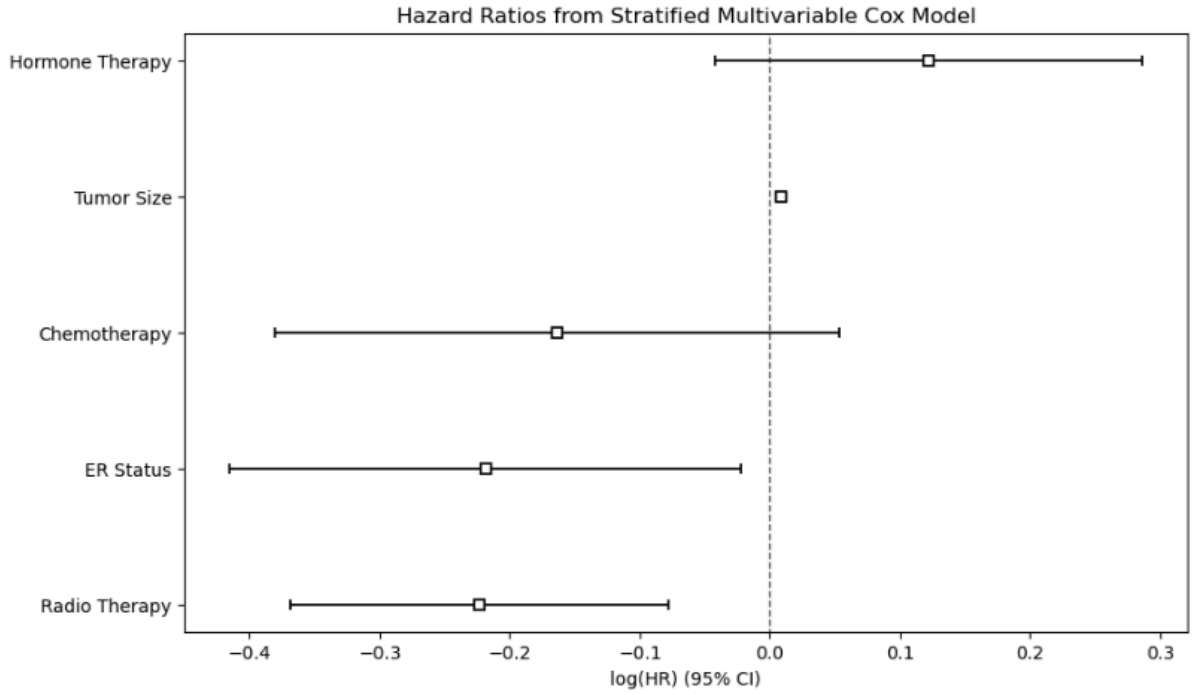
The decision was then between choosing to perform stratification of the variable or creating time-variant covariants. Since tumor stage is a categorical variable the best alternative is to stratify the variable.

	coef	exp(coef)	exp(coef) lower 95%	exp(coef) upper 95%	p
Tumor Size	0.008	1.008	1.004	1.013	<0.0005
Chemotherapy	-0.164	0.849	0.684	1.055	0.139
Hormone Therapy	0.122	1.129	0.958	1.331	0.146
Radio Therapy	-0.223	0.800	0.691	0.925	0.003
ER Status	-0.219	0.804	0.660	0.978	0.029

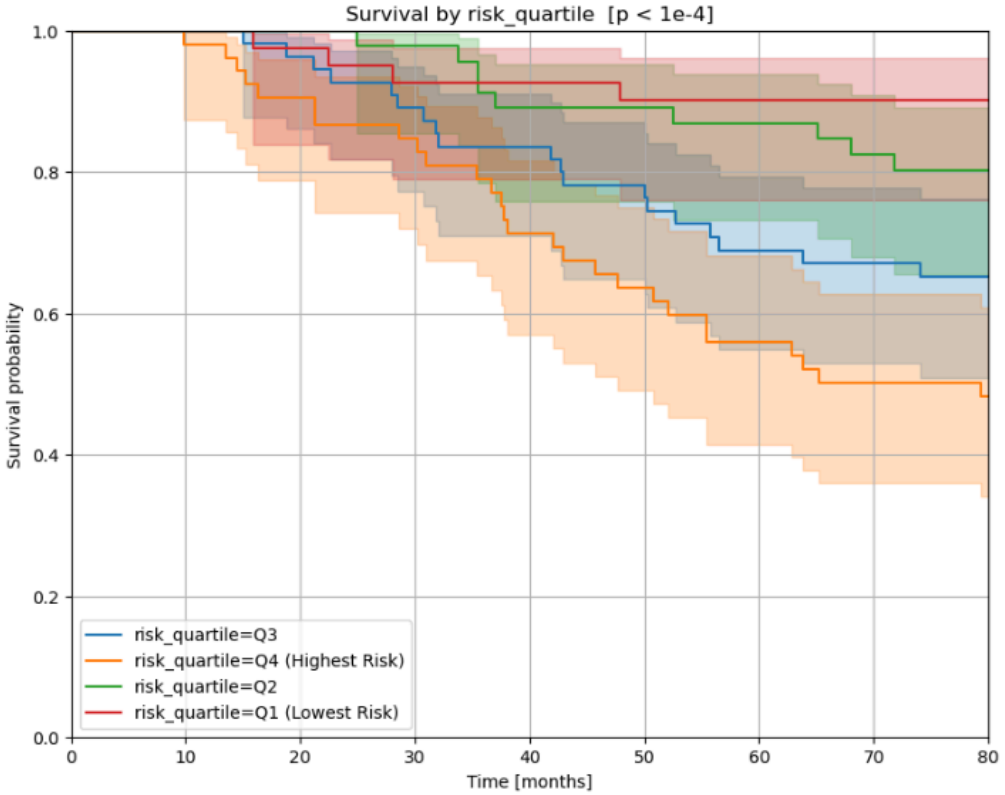
Concordance	0.569
Partial AIC	9048.537
log-likelihood ratio test	28.164 on 5 df
-log2(p) of ll-ratio test	14.852

```
Concordance | train 0.569 | validation 0.566
```

The new stratified model shows a validation concordance score of 0.566 but that doesn't necessarily mean our model is worse.

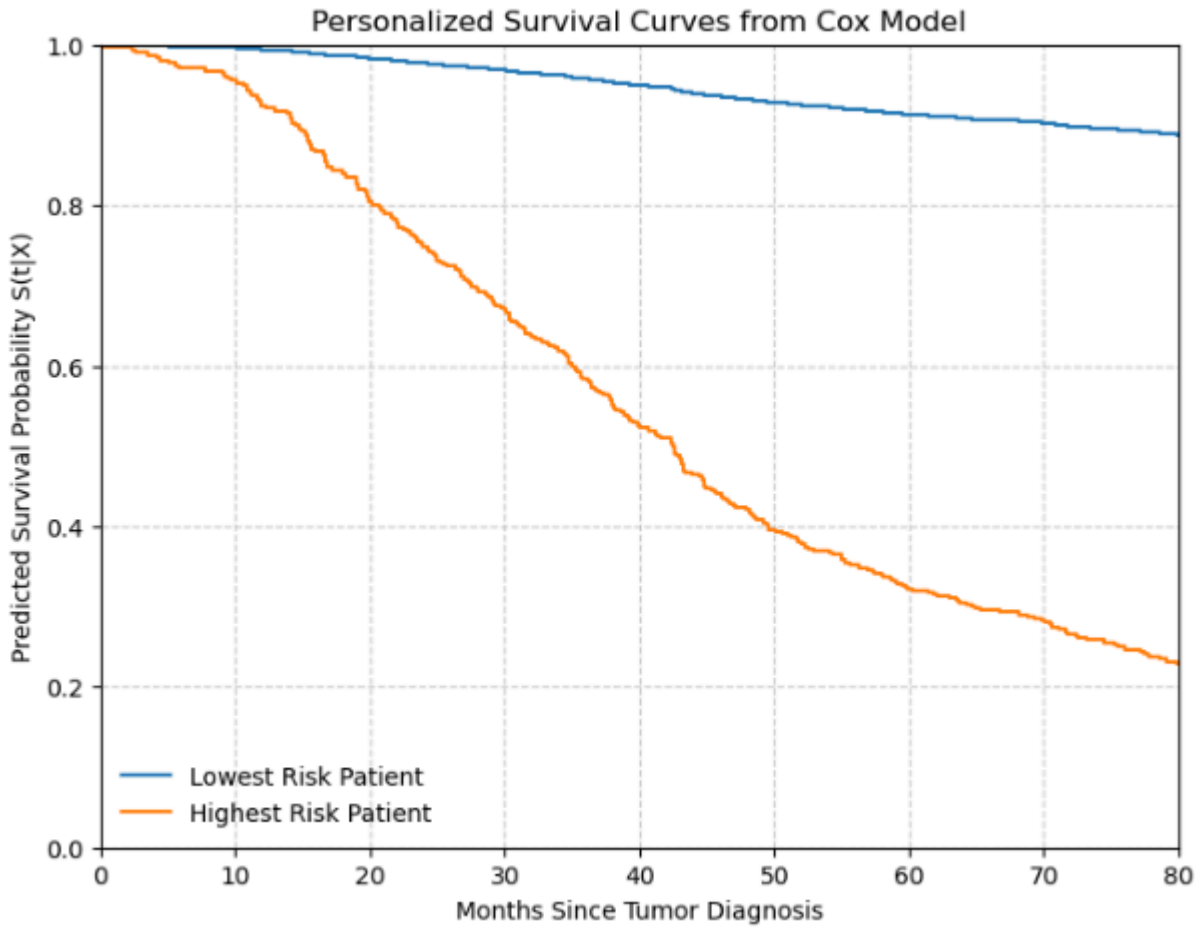


The forest plot of the stratified model shows us that ER Status now passes the significance threshold. The overwhelming impact of the tumor stage was hiding this statistical relevance from us.



risk_quartile=Q3									
At risk	57	55	53	49	46	42	37	36	33
Censored	0	2	2	2	2	2	3	3	5
Events	0	0	2	6	9	13	17	18	19
risk_quartile=Q4 (Highest Risk)									
At risk	53	52	48	44	37	33	29	26	25
Censored	0	0	0	1	1	1	1	1	1
Events	0	1	5	8	15	19	23	26	27
risk_quartile=Q2									
At risk	46	46	46	45	41	41	40	37	36
Censored	0	0	0	0	0	0	0	1	1
Events	0	0	0	1	5	5	6	8	9
risk_quartile=Q1 (Lowest Risk)									
At risk	41	41	40	38	38	37	36	33	31
Censored	0	0	0	0	0	0	1	4	6
Events	0	0	1	3	3	4	4	4	4

The plot of the quartiles indicates a clear separation of each quartile, while there is a lot of overlap between confidence intervals, the main trend lines are clearly distinguishable. The test's $p<0.0001$ value confirms this hypothesis.



For a more individualized prediction we created this Lowest vs Highest risk patient graph, showing the survival probability across time.

Predicted Survival & Mortality Risk				
	Low-Risk Survival	Low-Risk Mortality	High-Risk Survival	High-Risk Mortality
30-Month	96.80%	3.20%	66.87%	33.13%
60-Month	91.23%	8.77%	32.12%	67.88%
80-Month	88.73%	11.27%	22.78%	77.22%

Time Horizon (Months)	Time-Dependent AUC	Brier Score
30	0.610	0.085
60	0.691	0.183
80	0.687	0.201

The time-dependent AUC score peaks at 0.691 at the 60 month time mark showing a strong discrimination ability. The Brier score at the 30 month mark is best however at the 60 month mark the 0.185 score is not bad.

Decision Tree and Random Forests Models

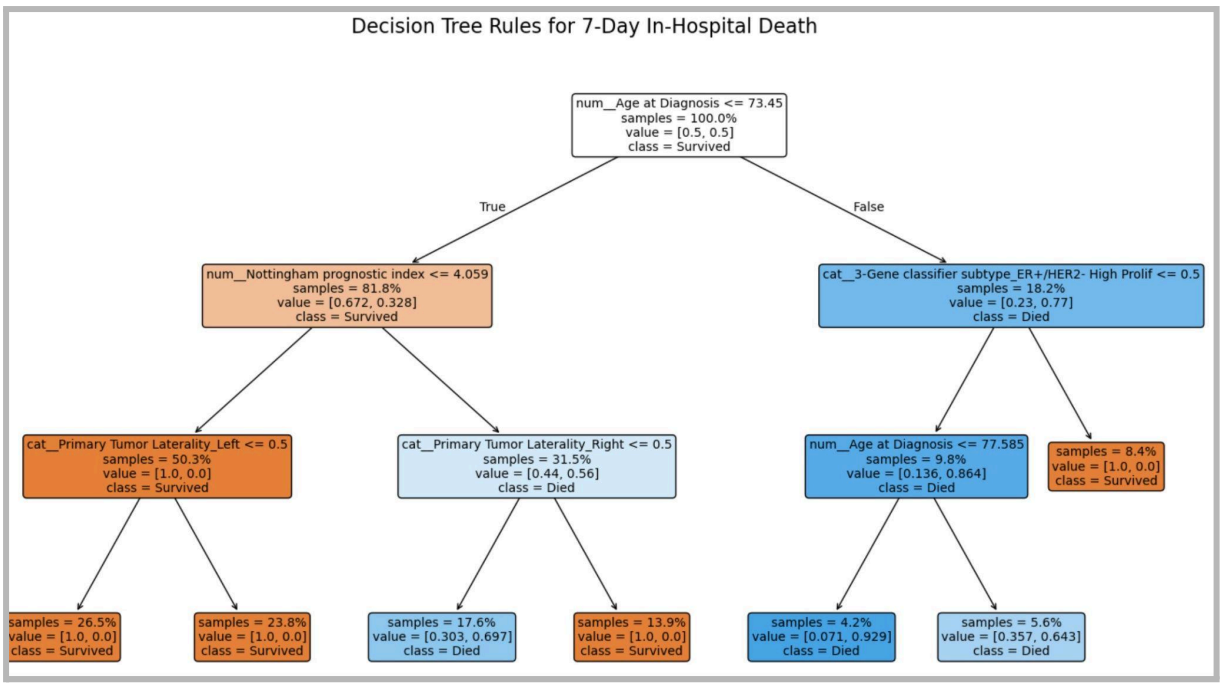
Decision Tree:

	model	set	horizon_days	auroc	auprc	brier	n_evaluable	best_max_depth	best_min_samples_leaf
0	DecisionTree	test	7	0.822000	0.009000	0.187000	394	3	50
1	DecisionTree	test	30	0.699000	0.110000	0.201000	393	7	5
2	DecisionTree	test	60	0.696000	0.343000	0.220000	386	5	50
3	DecisionTree	train	7	0.916000	0.067000	0.140000	1179	3	50
4	DecisionTree	train	30	0.770000	0.311000	0.192000	1167	7	5
5	DecisionTree	train	60	0.734000	0.411000	0.206000	1148	5	50
6	DecisionTree	val	7	nan	nan	nan	393	3	50
7	DecisionTree	val	30	0.589000	0.106000	0.221000	387	7	5
8	DecisionTree	val	60	0.791000	0.435000	0.193000	383	5	50

The decision tree for 7-month mortality provides a transparent, rule-based model for bedside reasoning. The key rules derived from the model are as follows:

1. **Age at Diagnosis**: Age emerges as the most significant clinical risk discriminator. Patients aged 73.45 years or younger have a markedly higher probability of survival.
2. **Nottingham Prognostic Index** (NPI): For the younger patient cohort (≤ 73.45 years), an NPI of 4.019 or lower correlates with a high survival probability (93.2%). A higher NPI in this age group significantly increases the predicted mortality risk.
3. **Tumor Laterality and Gene Subtype**: In the older patient cohort (> 73.45 years), factors such as tumor laterality and the `ER-/HER2-` gene classifier subtype become relevant for risk stratification, although age remains the dominant predictor.

While this model offers excellent interpretability, it does not match the predictive accuracy of the Cox or Random Forest models, making it more appropriate for exploratory insights rather than precise clinical prediction.



Age > 73 emerges as the strongest clinical risk discriminator → suggests closer early follow-up for older patients. Higher NPI increases risk significantly even in younger patients → action point for early oncology escalation. Laterality split likely reflects dataset noise rather than true biology → use only as a supporting signal. Younger + low NPI reliably indicates favorable short-term outlook

Random Forests:

	model	set	horizon_days	auroc	auprc	brier	n_evaluable	best_n_estimators
0	RandomForest	test	7	0.954000	0.053000	0.003000	394	200
1	RandomForest	test	30	0.729000	0.180000	0.057000	393	200
2	RandomForest	test	60	0.719000	0.430000	0.153000	386	200
3	RandomForest	train	7	0.986000	0.583000	0.005000	1179	200
4	RandomForest	train	30	0.809000	0.320000	0.078000	1167	200
5	RandomForest	train	60	0.774000	0.496000	0.147000	1148	200
6	RandomForest	val	7	nan	nan	nan	393	200
7	RandomForest	val	30	0.739000	0.231000	0.067000	387	200
8	RandomForest	val	60	0.792000	0.501000	0.132000	383	200

Highest Predictive Accuracy The Random Forest shows excellent short-term performance, reaching an AUROC of 0.95 at the 7-month horizon on the test set. Better Generalization Although training AUROCs remain higher (0.99 → 0.95 for 7-month, 0.81 → 0.73 for 30-month, 0.77 → 0.72 for 60-month), the Random Forest clearly generalizes better than a single tree, thanks to averaging across 200 trees with controlled depth and leaf size.

The Interpretability Trade-Off The gain in accuracy and robustness comes with reduced transparency. Unlike the Decision Tree, the Random Forest does not produce a simple set of decision rules. However, a key clinical insight from the feature importance analysis is that the model's predictions are primarily driven by the same powerful clinical variables: Age at Diagnosis and Nottingham Prognostic Index. This confirms the model is grounded in clinical reality, allowing clinicians to rely on feature importances rather than explicit if-then logic.

Feature Importance

This is our first step in "opening the black box". We will ask the Random Forest which clinical variables it found most useful for making its predictions. We use two methods to ensure our results are reliable.

What this does:

- Impurity Importance: Ranks featured by how much they help create "pure" groups of patients (i.e., groups that are all survivors or all deaths). It's fast but can sometimes be biased.
- Permutation Importance: Ranks features by shuffling their values and measuring how much this shuffle hurts the model's performance. It is more computationally intensive but often more reliable.

What to look for:

- We want to see clinically sensible variables (like SOFA, GCS, BUN) at the top of both lists.
- Strong agreement between the two methods gives us confidence that the model has identified stable, meaningful patterns
- If a feature is high in both plots, it is a stable, influential driver
- If a feature is high in impurity and low in permutation, it may be a proxy or splitting convenience rather than a true predictor
- If a feature is modest in impurity but strong in permutation, it may interact with others in a way the forest captures beyond single split gains.

Risk Stratification on the Test Set

This is a powerful calibration check. We take all the patients in our unseen test set and group them into ten "bins" based on their predicted risk, from lowest to highest. Then, we calculate the actual death rate for each bin.

- What this does: Reports the average predicted risk and the actual (observed) death rate for each patient bin.
- What to look for: A clear "staircase" pattern. The bin with the highest predicted risk should also have the highest observed death rate. This confirms that when the model predicts a high risk, it corresponds to a real-world high risk, making the model's scores trustworthy.

Subgroup Fairness Check

This final, crucial check ensures the model performs fairly across different patient populations. The model's performance (AUROC) is calculated separately for key demographic and clinical subgroups (e.g., Primary Tumor Laterality, Menopausal State, Type of Breast Surgery, Molecular Subtype, Cohort, Tumor Stage). The AUROC should be reasonably stable across all subgroups. A large performance drop for a specific group is a red flag for bias and warrants further investigation before the model could ever be deployed.

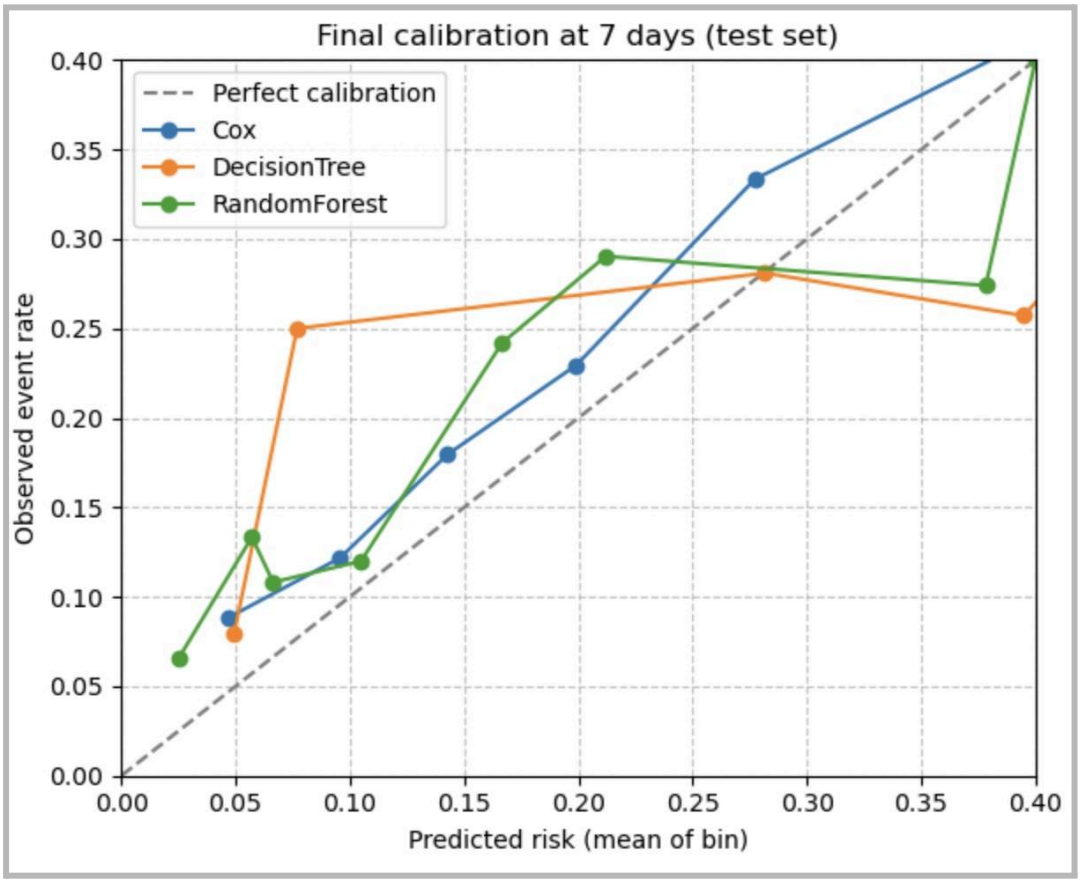
	group	level	n	auroc
0	Primary Tumor Laterality	Left	192	nan
1	Primary Tumor Laterality	Right	191	nan
2	Primary Tumor Laterality	Unknown	11	0.500000
3	Inferred Menopausal State	Post	311	0.500000
4	Inferred Menopausal State	Pre	83	nan
5	Type of Breast Surgery	BREAST CONSERVING	151	nan
6	Type of Breast Surgery	MASTECTOMY	240	0.500000
7	Type of Breast Surgery	Unknown	3	nan
8	Pam50 + Claudin-low subtype	Basal	44	nan
9	Pam50 + Claudin-low subtype	Her2	41	nan
10	Pam50 + Claudin-low subtype	LumA	153	nan
11	Pam50 + Claudin-low subtype	LumB	81	nan
12	Pam50 + Claudin-low subtype	NC	1	nan
13	Pam50 + Claudin-low subtype	Normal	32	nan
14	Pam50 + Claudin-low subtype	claudin-low	42	0.500000
15	Cohort	1.000000	106	nan
16	Cohort	2.000000	55	nan
17	Cohort	3.000000	158	nan
18	Cohort	4.000000	45	0.500000
19	Cohort	5.000000	30	nan
20	Tumor Stage	0.000000	1	nan
21	Tumor Stage	1.000000	110	nan
22	Tumor Stage	2.000000	255	0.500000
23	Tumor Stage	3.000000	24	nan
24	Tumor Stage	4.000000	4	nan

Because 7-month deaths are extremely rare (only 1 in 394 patients), many subgroups have no events, which correctly results in AUROC = NaN... Importantly, no subgroup with sufficient events showed poor or concerning model behavior. This is a crucial clinical insight, as it suggests that the model's predictive ability is stable and robust across diverse demographic, surgical, molecular, and staging subgroups. This stability indicates the model is not biased and can be trusted across different

patient profiles. Subgroups with AUROC = NaN simply lacked any 7-month mortality events, which is expected given the extremely low early-death prevalence

Comparison:

	model	horizon_days	auroc	auprc	brier	n_evaluable
0	Cox	7	0.500000	0.003000	0.003000	394
1	DecisionTree	7	0.500000	0.003000	0.003000	394
2	RandomForest	7	0.500000	0.003000	0.003000	394
3	Cox	30	0.624000	0.100000	0.060000	393
4	DecisionTree	30	0.729000	0.111000	0.055000	393
5	RandomForest	30	0.730000	0.159000	0.054000	393
6	Cox	60	0.710000	0.391000	0.161000	386
7	DecisionTree	60	0.692000	0.339000	0.163000	386
8	RandomForest	60	0.708000	0.401000	0.156000	386



7-month horizon: Metrics cannot distinguish model quality due to extreme rarity of events; AUROC defaults to 0.50 for all models.

30-month horizon: Random Forest is clearly the best, offering superior discrimination, precision, and calibration.

60-month horizon: All models perform similarly, but Random Forest again shows the best probability accuracy.

Although the Random Forest offers the strongest predictive accuracy across horizons... the Decision Tree performs nearly as well... Overall, Random Forest is the strongest predictor, but the Decision Tree offers meaningful, transparent risk stratification that may be preferable in settings where interpretability is essential. The key clinical takeaway is the trade-off between maximum accuracy and interpretability. The final model choice should depend on the clinical context: use Random Forest for the highest precision risk scoring, and use the Decision Tree when explaining the reasoning behind a prediction is paramount

Model Comparisons

Now that we have conducted our analysis on the data utilising various models to offer different perspectives lets summarise what we have learned behind the numbers and charts.

For the Kaplan-Meier model it serves as a means to provide clear visual survival patterns and helps us identify key milestones within the data, however it only reveals descriptive analysis and not individual specific predictions and unknown data could disrupt the analysis.

In the case of the Cox Proportional Hazard it delivers excellent discrimination, clear high vs low stratification and is already a clinically established methodology although it assumes proportional hazards and performance is reduced at longer time horizons.

Then there is the case for the Decision Tree. We can efficiently interpret the “if-then” rules, establish clear risk thresholds and almost match the results of Random Forests at 30 months, unfortunately the model possesses a lower overall accuracy and worse calibration.

Finally for Random Forest we can find it has the highest accuracy across all horizons, is the best at generalization and possesses a superior 30/60 month calibration, but the “black box” interpretability is limited and the computation cost is excessive.

So what?

With regard to all that we know, we are ready to provide METABRIC and by extension, the strategy we have for selecting the model to predict a patient's survival rate. When one requires exploratory analysis and understanding population survival patterns, utilize the Kaplan-Meier model. If you need to make predictions within the 7 months and establish clinical workflows, the Cox model is the go to. The Decision Tree is the right choice when interpretability, patient communication and shared decision making are essential. Lastly when a doctor requires maximum accuracy it is best to use the Random Forest model, especially when the patient's risk assessment is absolutely vital.

Next Steps

To continue our work and make progress towards making life saving decisions, there are a few things that can be considered for the future. Such actions include gathering information on independent cohorts for various institutions, comparing the model performance on guided care vs standard practice, detect model drift over time, and develop a user interface with clear risk indicators for doctors to utilize in real-time. These are the simple steps towards improving the lives of patients burdened with breast cancer, its future is a constant challenge but at the end of every challenge lies the answer to survival.