Methodology Report

Kevin Zheng

May 2024

1 Introduction

The dataset used in this project is derived from the I-SPY 2 breast cancer trial, which assesses the efficacy of experimental drug treatments in neoadjuvant chemotherapy. Each participant in the trial undergoes a series of four breast tumor MRIs, from which various radiomic features are extracted. Additionally, the dataset includes survival data and breast cancer status features. For this project, we focus on the radiomic feature "FTV3", representing the functional tumor volume extracted from the fourth MRI taken at the end of the chemotherapy treatment. It is used in conjunction with the variable "combined_status", which denotes which of the four sub-types of breast cancer the patient falls under ("HR-/HER2-", "HR+/HER2-", "HR-/HER2+", or "HR+/HER2+"). Our goal is to combine statistical modeling with machine learning methodologies to determine whether a log or cube root transformed version of FTV3 performs better than untransformed FTV3 (all in conjunction with cancer status) in predicting breast cancer recurrence. [Li]

2 Methods

2.1 Evaluation Metric: Concordance Index

Concordance index (c-index) was chosen as the metric for evaluating model performance. The c-index, introduced by Dr. Frank Harrell and his colleagues in 1982, is a commonly used statistic for survival model validation. It is analogous to an area under the ROC curve (AUC) that takes into account censorship. We define the c-index as the proportion of pairs for which the model correctly predicts the relative survival time, among all pairs for which this can be determined.

[Harrell, 2015]

The C-index is defined as:

$$C = \frac{\sum_{i,i':y_i>y_{i'}} I(\hat{\eta}_{i'} > \hat{\eta}_i)\delta_{i'}}{\sum_{i,i':y_i>y_{i'}} \delta_{i'}}$$

$I(\hat{\eta}_{i'} > \hat{\eta}_i)$ is the indicator variable (equaling 1 if $\hat{\eta}_{i'} > \hat{\eta}_i$ and 0 otherwise) that determines whether the predicted risk for i' is greater than the predicted risk for i. $\hat{\eta}$ are the predicted values, y are the observed times, and δ is the censoring indicator. [James]

2.2 Data Preprocessing 1: Transformations For Skewed Data

Fig. 1 is a distribution plot of untransformed FTV3, with the red line denoting the mean. As illustrated, the distribution of FTV3 is very right skewed, with most of the tumor volumes falling between 0 and 20. To address this, we are employing two different normalizing transformations: cube root and log. Both of these methods are commonly used in the normalization of continuous data [Osborne 2002]. The cube root FTV3 model is particularly interpretable for volume as it is a three dimensional measurement. We are, in a sense, reducing the dimensionality of FTV3 to one by taking the cube root of the volume.

Fig. 2 and 3 illustrate the distribution of cube root FTV3 and log transformed FTV3 respectively. The result is a much more normalized distribution, with cube root FTV3 values ranging from 0 to 7 and log-transformed FTV3 values ranging from 0 to 6, compared to the original right-skewed distribution where values extend beyond 250.

In previous studies, we have found that a cube root transformation of FTV3 is beneficial in the prediction of breast cancer pathologically complete response. We are hypothesizing that similar benefits will be found in the context of survival analysis.

2.3 Data Preprocessing 2: Multiple Imputation With Predictive Mean Matching

FTV3 has 76 missing observations out of 899 total (8.45%). To address this, we utilized multiple imputation with the MICE package in R with Predictive Mean Matching (PMM), which is the default method for handling continuous vari-

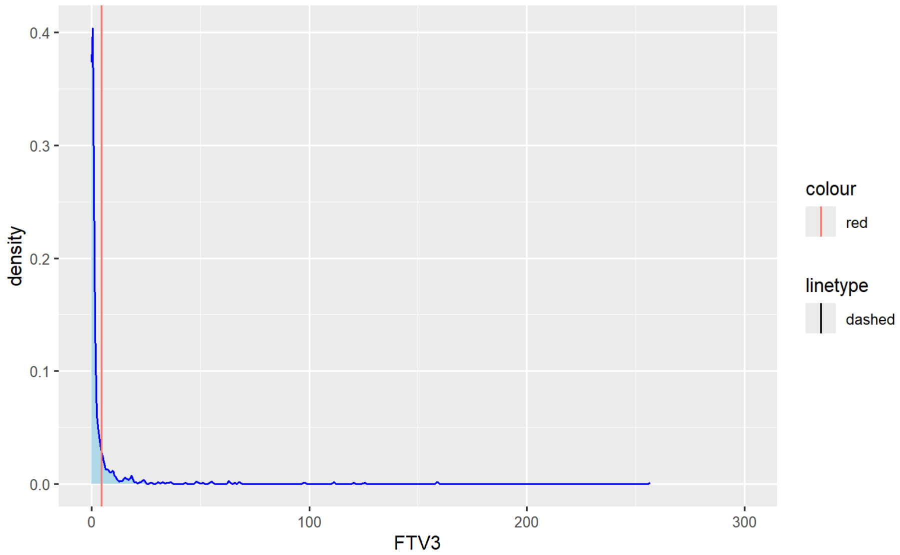


Figure 1: Raw FTV3 Distribution

ables. PMM works by creating a set of imputed values, then for each imputed value, we identify a set of real observed values from the dataset that are similar to each imputed value. That imputed value is substituted with a real value drawn from that observed set. PMM is effective for imputing numeric missing data with realistic values because we are ensuring that the imputed values are consistent with the distribution of the observed data points.

We have chosen PMM for this study because it has been demonstrated to yield the least biased estimates and superior model performance metrics, as reported by Marshall, Altman, and Holder (2010). This makes it a robust and reliable method for addressing missing data. It is also robust to transformations of the target feature, meaning that similar imputed data will be obtained from imputing cube root FTV3 or log FTV3. However, it is important to note that PMM is not recommended when greater than 50 percent of the data is missing (Marshall, et al. 2010), or if the data is particularly skewed (Kleinke 2017). The cross validated model (Section 2.4) was also evaluated on data with dropped missing values as a sensitivity analysis. [Buuren]

Convergence of imputations were assessed using convergence plots. According to van Buuren and Groothuis-Oudshoorn (2011), when assessing convergence, “the different streams should be freely intermingled with each other,

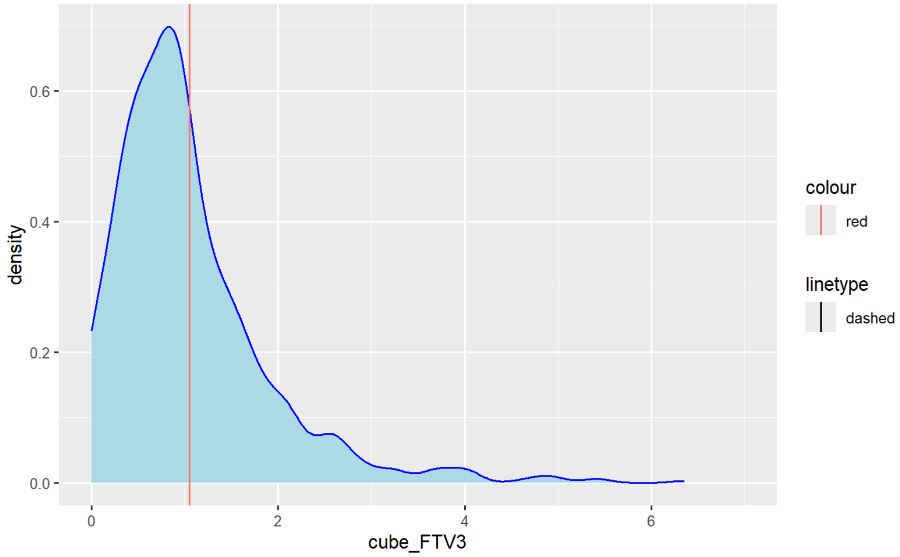


Figure 2: Cube Root FTV3 Distribution

without showing any definite trends. Convergence is diagnosed when the variance between different sequences is no larger than the variance within each individual sequence" (van Buuren and Groothuis-Oudshoorn, 2011).

The convergence plot for multiple imputation of FTV3 is shown in Fig. 4. Fig. 5 is an example of bad convergence, with the streams not adequately mixing in a steady manner. Fig. 6 is an example of good convergence, with streams mixing steadily.

2.4 Survival Model: Cox Proportional Hazards

The primary survival model chosen for this project is the Cox Proportional Hazards model, which was first introduced by Sir David Cox in 1972. This model is the widely used standard procedure for regression analysis in survival studies, and is known for its robustness and ability to handle right-censored data effectively. Another strength of the Cox model is that it does not specify the baseline hazard function, which allows for flexibility in its application and reduces complexity. Additionally, the model allows for the use of multiple covariates. This is particularly important for our application, as we are interested in FTV3 in conjunction with breast cancer status. In our analysis, the assumptions of the cox model (such as proportional hazards assumption) were tested

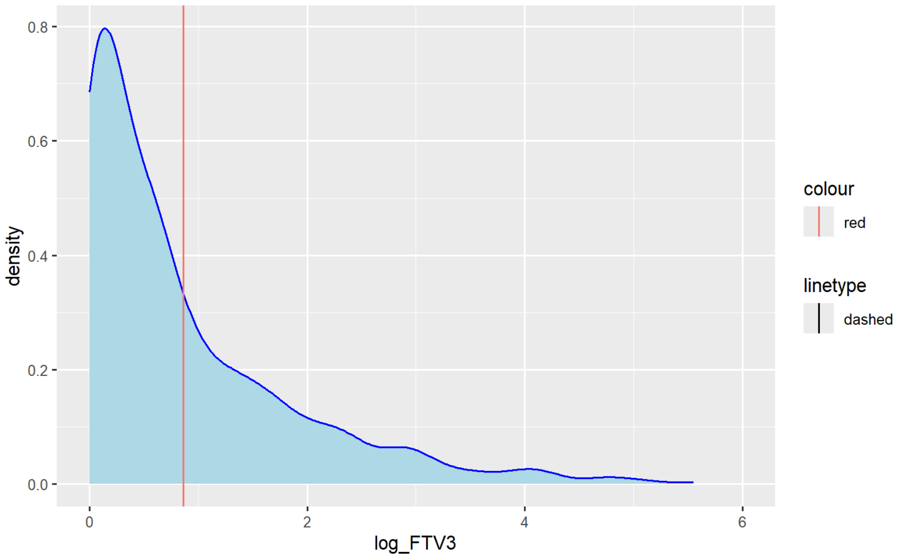


Figure 3: Log FTV3 Distribution

using Schoenfeld residuals.

The Cox proportional hazards model is expressed as:

$$h(t|\mathbf{X}) = h_0(t)\exp(\mathbf{X}^\top \boldsymbol{\beta})$$

where:

- $h(t|\mathbf{X})$ is the hazard function at time t given covariates $\mathbf{X}$.
- $h_0(t)$ is the baseline hazard function.
- $\mathbf{X}$ is the vector of covariates.
- $\boldsymbol{\beta}$ is the vector of coefficients.

[Abd ElHafeez et al. 2021]

2.5 Model Assessment: K-Fold Cross Validation

We utilized k-fold cross-validation, a well established validation technique in machine learning, to compare the transformed FTV3 Cox models (log or cube root) with the non-transformed FTV3 Cox model. The breast cancer status variable ”combined_status” was incorporated in each model.

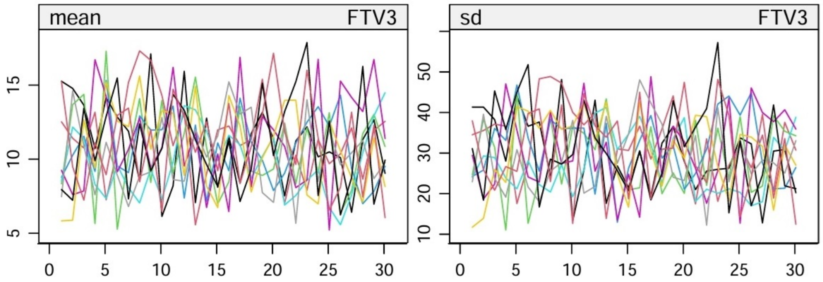


Figure 4: FTV3 Convergence Plot

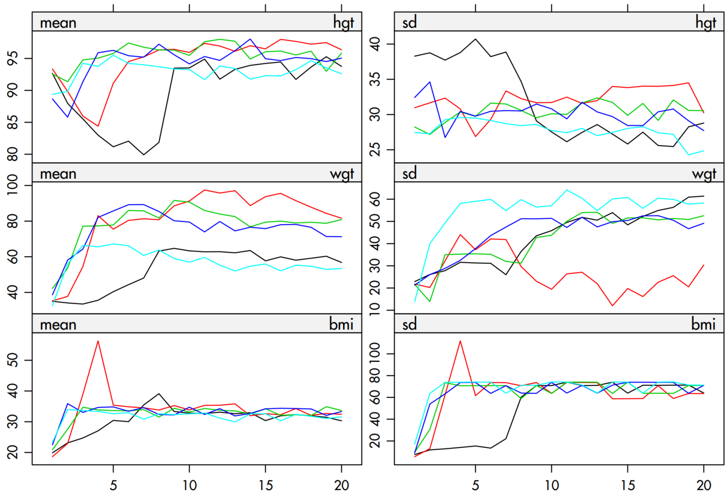


Figure 5: Example of Non Convergence

In this technique, the original dataset is split into “k” folds (groups) of roughly equal size, with k usually being 5 or 10. Each of the folds takes a turn as the validation set, with the remainder of the data being used as the training set. The results of all the validation sets are then averaged, yielding the final k-fold cross validation estimate (James et al. 2013). A diagram illustrating this process is depicted in Fig. 7 (Al-Issa, Alqudah, Alquran, & Al Issa, 2022).

We have chosen k-fold cv due to its efficient utilization of our limited dataset (n = 899). Each part of the data is utilized as training and validation, which reduces variance and provides a more generalizable estimate of model performance. When performed with k = 5 or k = 10, it “has been shown empirically

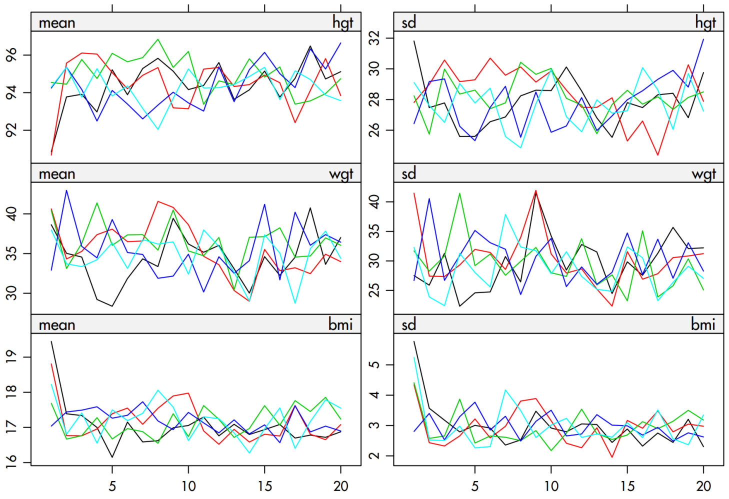


Figure 6: Example of Healthy Convergence

to yield estimates that suffer neither from excessively high bias nor from very high variance" (James et al. 2013). However, a notable drawback is that it is computationally expensive when applied to particularly large datasets or very complex models. In our case, this concern is mitigated by our relatively modest dataset size.

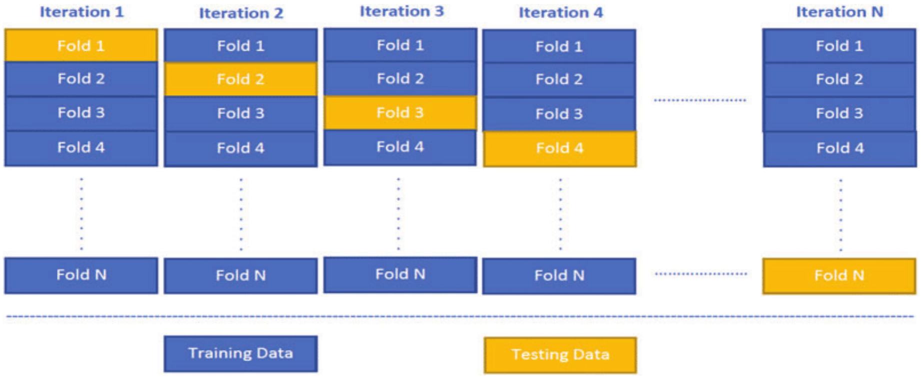
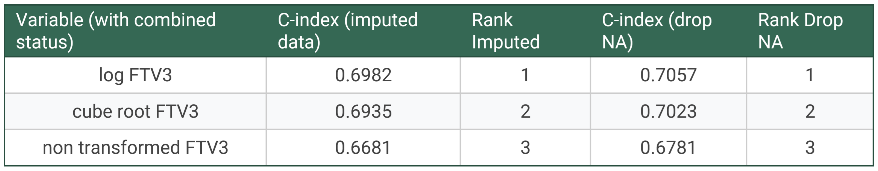


Figure 7: K-Fold Cross Validation Diagram

For our analysis, we ran 5-fold cross validated cox proportional hazards on 3 models with differing forms of FTV3. The best model was chosen based on both the highest cross validated concordance index as well as on model interpretability. As show n in Fig. 8, the models with transformed data had very similar results, and both yielded higher cross validated c-indices compared to the non transformed model (log: 0.698, cube: 0.694, raw: 0.668). In both the imputed data analysis as well as the dropped missing sensitivity analysis, the log FTV3 model performed slightly better than the cube root FTV3 model, though the differences were too small to be meaningful. The cube root model was ultimately selected as the winner due to better interpretability (as discussed in section 2.2).



Variable (with combined status)	C-index (imputed data)	Rank Imputed	C-index (drop NA)	Rank Drop NA
log FTV3	0.6982	1	0.7057	1
cube root FTV3	0.6935	2	0.7023	2
non transformed FTV3	0.6681	3	0.6781	3

Figure 8: 5 Fold CV Results

2.6 Time-Dependent ROC Analysis: Additive Model

In addition to the k-fold Cox analysis, we performed time-dependent ROC analysis to assess survival at different time points. We evaluated survival at 1 year, 2 years, and 5 years, focusing on 5-year survival due to its prominence in cancer research.

The time-dependent ROC curves and AUCs were generated using the PROC PHREG procedure in SAS and the timeROC package in R, both employing the IPCW estimation method. Fig. 9 through 11 display overlapped ROC curves for 1 year, 2 years, and 5 years, respectively, while Fig. 12 through 14 show ROC curves with three overlapped years by model.

Both of these implementations perform the ”estimation of time-dependent ROC curve and area under time dependent ROC curve (AUC) in the presence of censored data” (Blanche 2013).

2.7 6 Pairwise Comparison for Cox Models

Using the winning variables identified in section 2.5 (cube root FTV3 with breast cancer status), we implemented non cross validated cox proportional hazards to create a pairwise comparison table. We evaluated permutation-style comparisons by changing the reference group to each of the breast cancer status

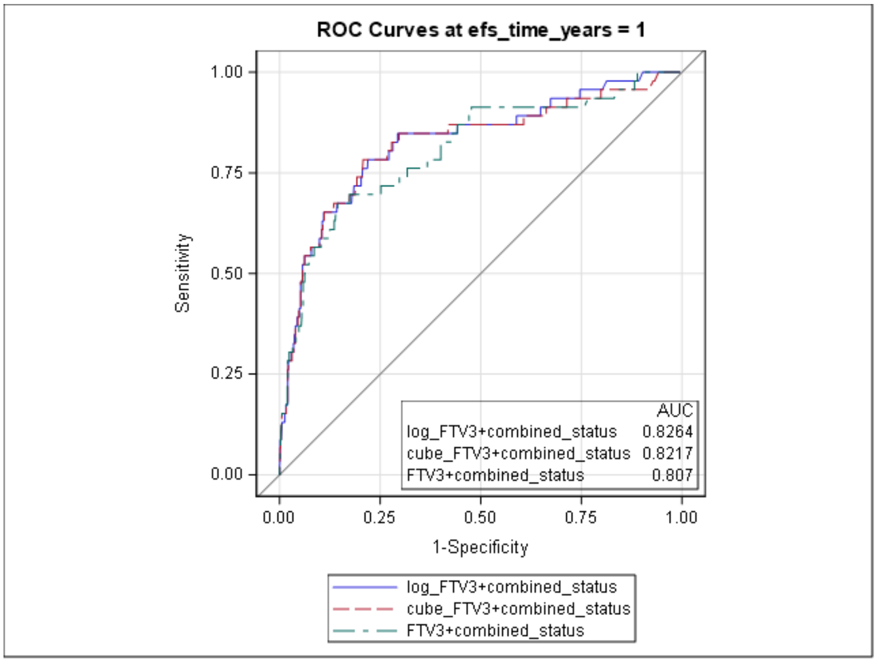


Figure 9: 1 year Overlapped ROC Curve

subtypes (“HR-/HER2-”, “HR+/HER2-”, “HR-/HER2+”, “HR+/HER2+”). This approach yielded six pairwise comparisons, as shown in Fig. 15.

2.8 Test for Interactions

To evaluate interaction, we first fit a Cox model without specifying the interaction term (cube_FTV3 + combined_status) and another model including the interaction term (cube_FTV3 + combined_status + cube_FTV3:combined_status). An ANOVA test was performed to determine if there was a significant difference between the two models. As shown in the ANOVA output in Fig. 16, the p-value was 0.83, indicating that the interaction term does not significantly improve the model.

2.9 Cox Models by Subgroup

We are interested in assessing differences in survival modeling between the additive model with all the breast cancer subgroups together (full dataset) versus models fit on datasets separated by subgroup. Separate univariate cox models

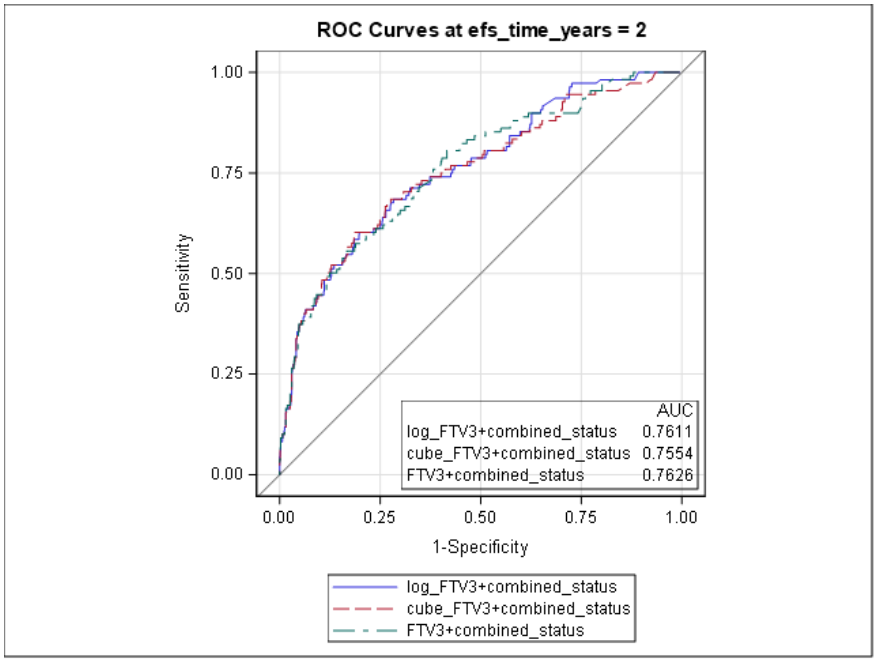


Figure 10: 2 year Overlapped ROC Curve

were fit within each breast cancer subgroup. The dataset was split into 4, one for each subgroup. A regular cox model was then fit for each subgroup, with cube root FTV3 as the only predictor. Fig. 17 shows the hazard ratios, p values, 95 percent confidence intervals, and c-indices of each of the four cox models. As detailed in the table, the HR+ groups had higher hazard ratios (HR+/HER2-: hazard ratio = 2.11 (95% ci 1.65-2.69), p value 0.00; HR+/HER2+: hazard ratio = 2.17 (95% ci 1.49-3.16), p value 0.00) than the HR- groups (HR-/HER2-: hazard ratio = 1.82 (95%ci 1.53-2.16), p value 0.00; HR-/HER2+ hazard ratio = 1.82 (95% ci 1.32-2.52, p value 0.00).

2.10 Time-Dependent ROC Analysis: Separated By Subtype

Finally, time-dependent ROC analysis was performed to evaluate survival models on datasets separated by subgroup. Fig. 18 through 21 are overlapped time-dependent ROC curves (1,2, 5 year), utilizing datasets of their respective breast cancer subtypes. In the HR-/HER2- ROC analysis, the 1 year AUC was 0.73, 2 year AUC was 0.73, and 5 year AUC was 0.7. With HR-/HER2+, the

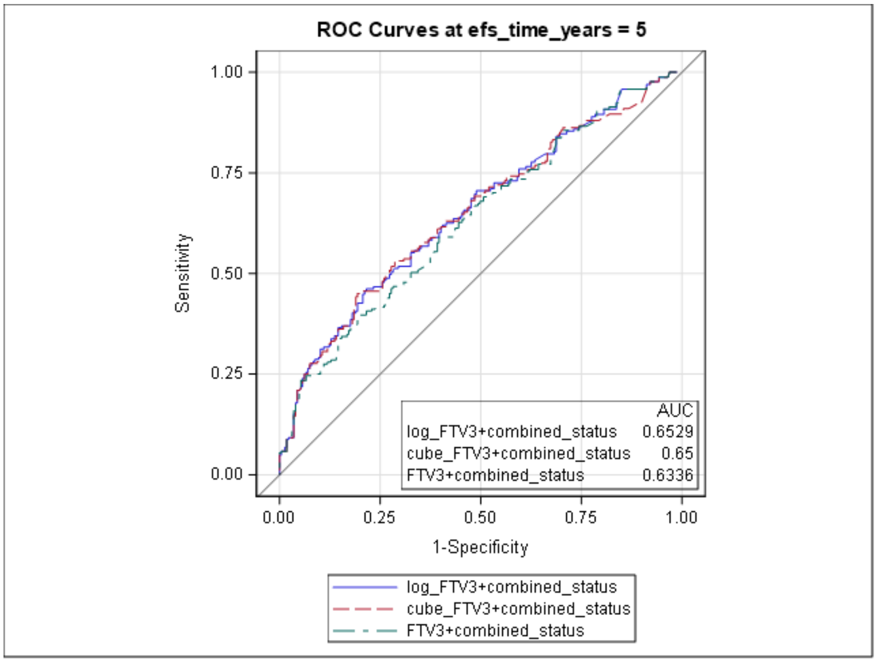


Figure 11: 5 year Overlapped ROC Curve

1 year AUC was 0.85, 2 year AUC was 0.78, and 5 year AUC was 0.61. With HR+/HER2-, the 1 year AUC was 0.8, 2 year AUC was 0.69, and 5 year AUC was 0.65. Lastly, in HR+/HER2+, the 1 year AUC was 0.81, 2 year AUC was 0.8, and 5 year AUC was 0.68. These ROC curves appear to generally follow the same trend as the full additive dataset ROC analysis depicted in Fig. 13, with a decreasing trend from 1, 2, to 5 years.

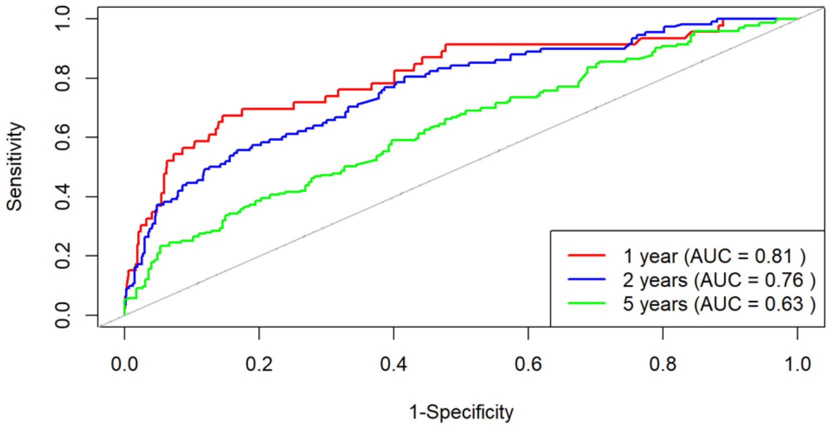


Figure 12: Raw FTV3 (with status) 1,2,5 year Overlapped ROC Curve

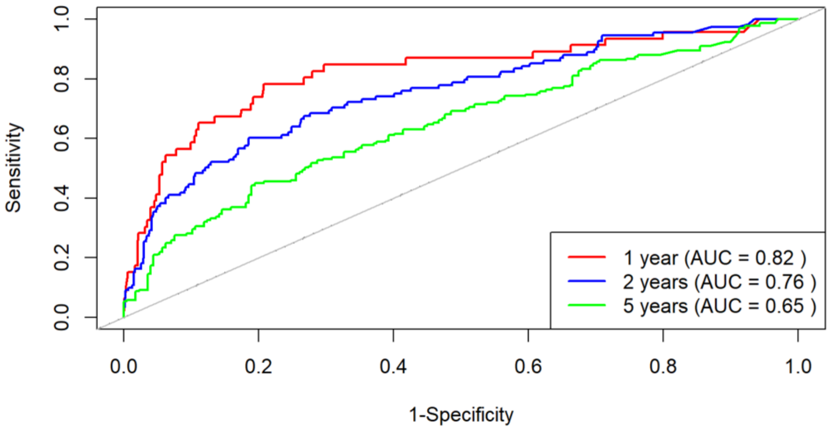


Figure 13: Cube Root FTV3 (with status) 1,2,5 year Overlapped ROC Curve

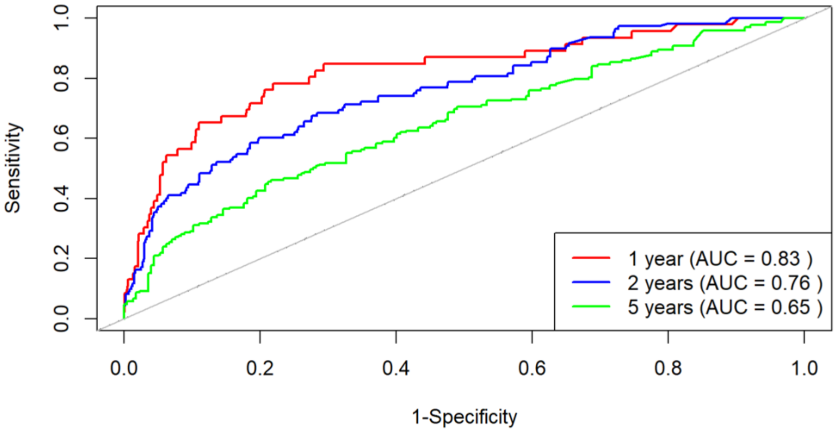
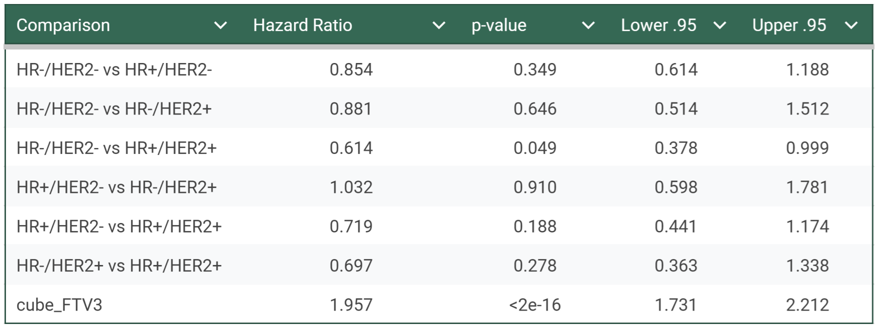
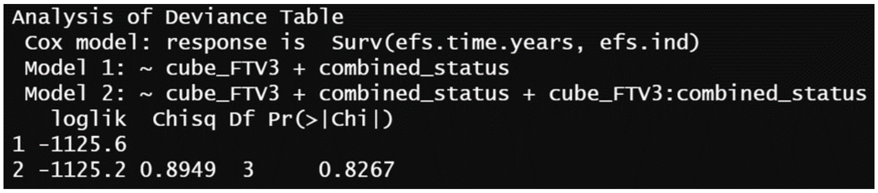


Figure 14: Log FTV3 (with status) 1,2,5 year Overlapped ROC Curve



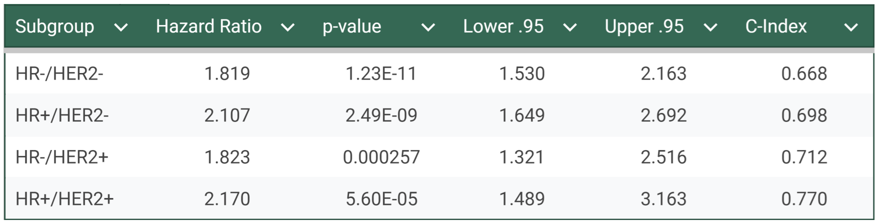
Comparison	Hazard Ratio	p-value	Lower .95	Upper .95
HR-/HER2- vs HR+/HER2-	0.854	0.349	0.614	1.188
HR-/HER2- vs HR-/HER2+	0.881	0.646	0.514	1.512
HR-/HER2- vs HR+/HER2+	0.614	0.049	0.378	0.999
HR+/HER2- vs HR-/HER2+	1.032	0.910	0.598	1.781
HR+/HER2- vs HR+/HER2+	0.719	0.188	0.441	1.174
HR-/HER2+ vs HR+/HER2+	0.697	0.278	0.363	1.338
cube_FTV3	1.957	<2e-16	1.731	2.212

Figure 15: 6 Pairwise Cox Model Comparison Table



```
Analysis of Deviance Table
 Cox model: response is  Surv(efs.time.years, efs.ind)
 Model 1: ~ cube_FTV3 + combined_status
 Model 2: ~ cube_FTV3 + combined_status + cube_FTV3:combined_status
   loglik  Chisq Df Pr(>|Chi|)
1 -1125.6
2 -1125.2 0.8949  3     0.8267
```

Figure 16: ANOVA Test Output



Subgroup	Hazard Ratio	p-value	Lower .95	Upper .95	C-Index
HR-/HER2-	1.819	1.23E-11	1.530	2.163	0.668
HR+/HER2-	2.107	2.49E-09	1.649	2.692	0.698
HR-/HER2+	1.823	0.000257	1.321	2.516	0.712
HR+/HER2+	2.170	5.60E-05	1.489	3.163	0.770

Figure 17: Cox Models by Subgroup

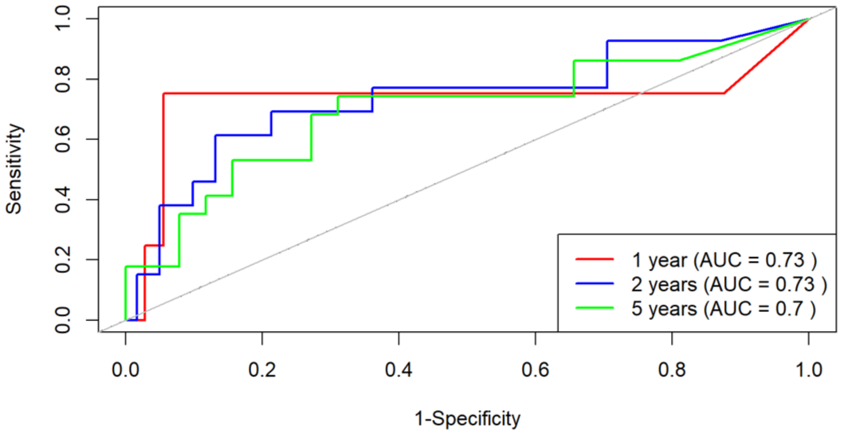


Figure 18: HR-/HER2-

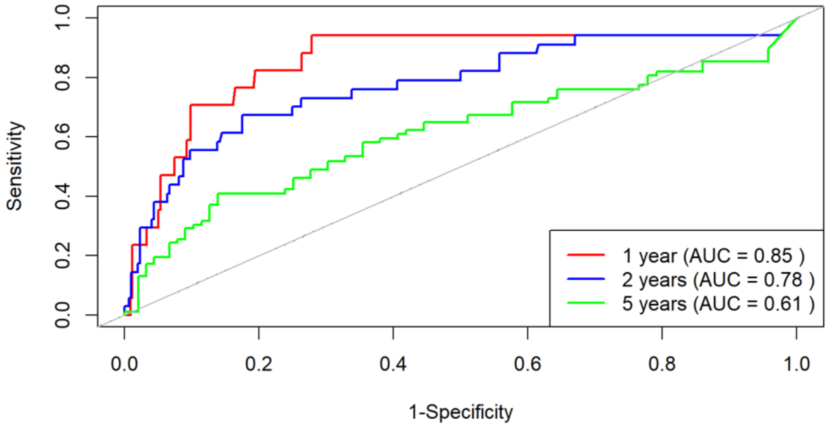


Figure 19: HR-/HER2+

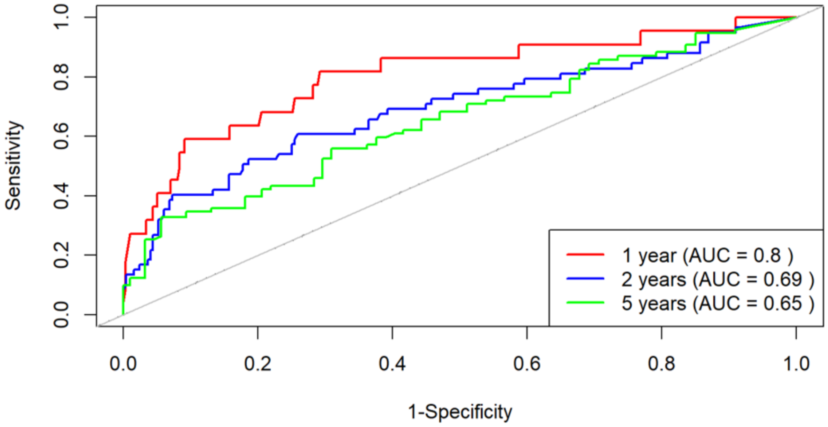


Figure 20: HR+/HER2-

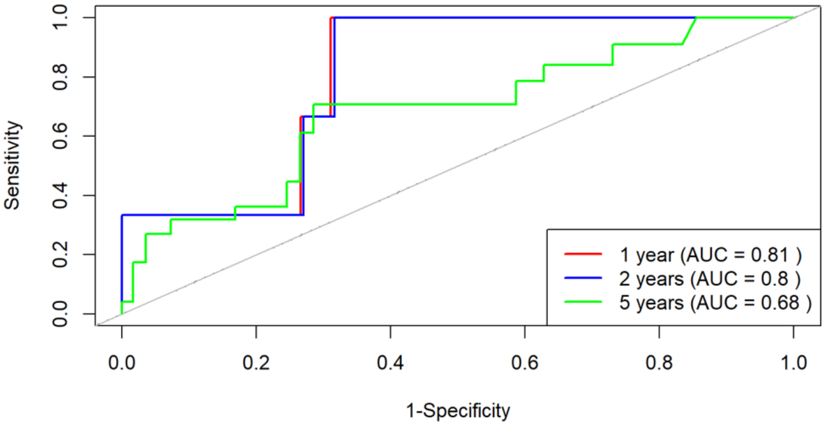


Figure 21: HR+/HER2+