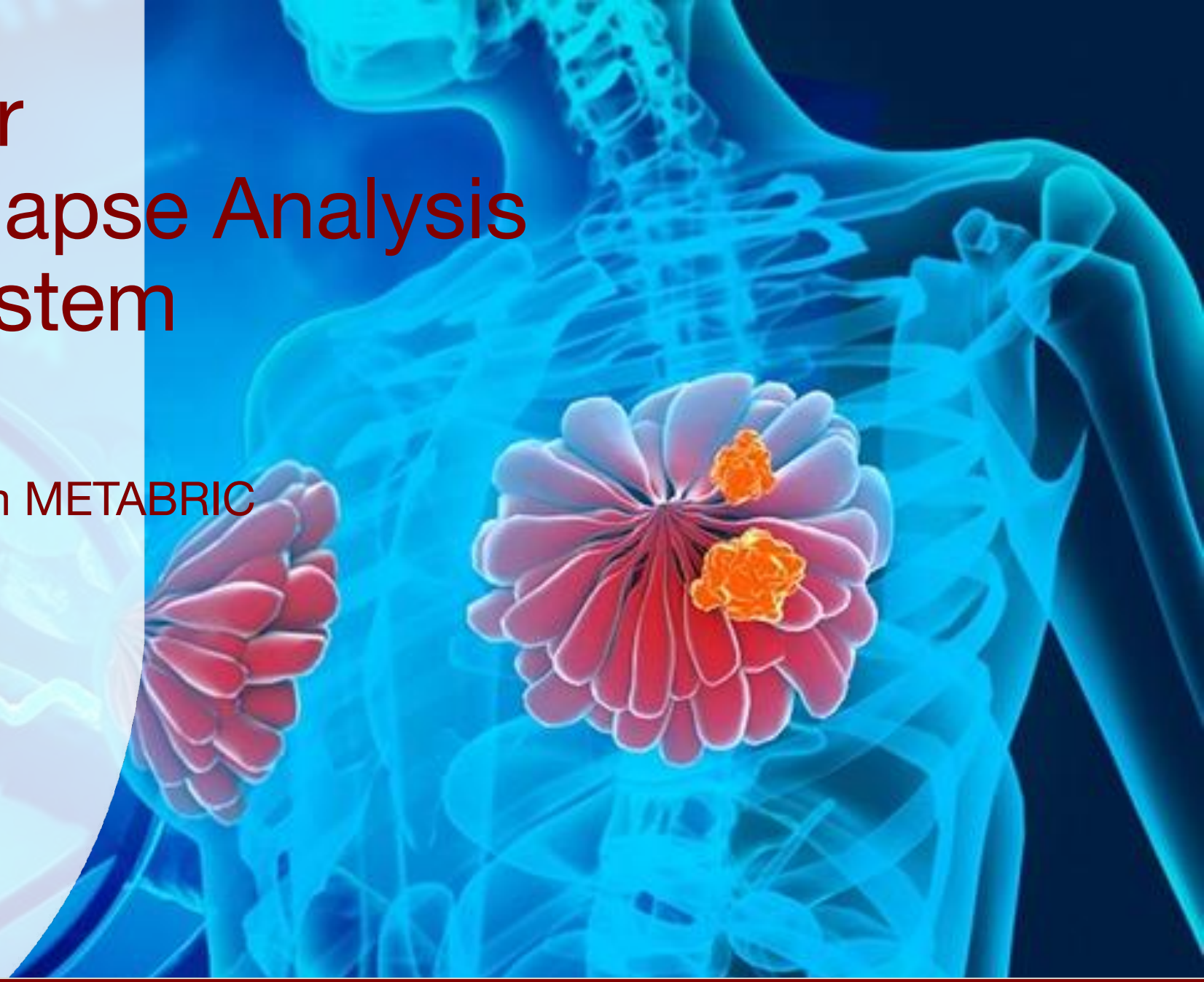
Breast Cancer Survival & Relapse Analysis and Score System Construction

– 1980 patients data from METABRIC



December 7

Qing Chen and Freya Wang

https://github.com/ffyyww/Breast-Cancer-Survival-and-Relapse-Analysis

Table of Contents

Overview

- **Overview:** In this project, we analyzed the clinical and genomics data of **1,980 breast cancer** patients, drawing **overall survival and relapse curves** using **Kaplan-Meier**. **Predictions** were made using **random survival forests**, and the top-ranking variables were selected and integrated into a **Cox regression model**, thereby constructing two **risk scoring systems** for better prediction in clinical practice.

- **Data Source:** The **METABRIC** project, funded by Cancer Research UK, the British Columbia Cancer Foundation and Canadian Breast Cancer Foundation BC/Yukon. Over **2,000** clinically annotated primary fresh-frozen breast cancer specimens were collected between **1977–2005** from five centers in the UK and Canada.

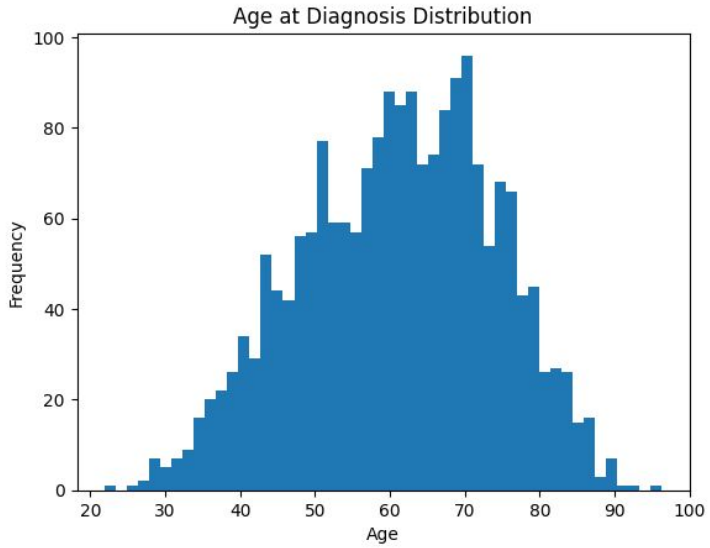
- **Main Variables:** Overall Survival (Months), Overall Survival Status, Relapse Free (Months), Relapse Free Status, Age, Mutation Count, Tumor Size, Tumor Stage, Surgery and Therapy

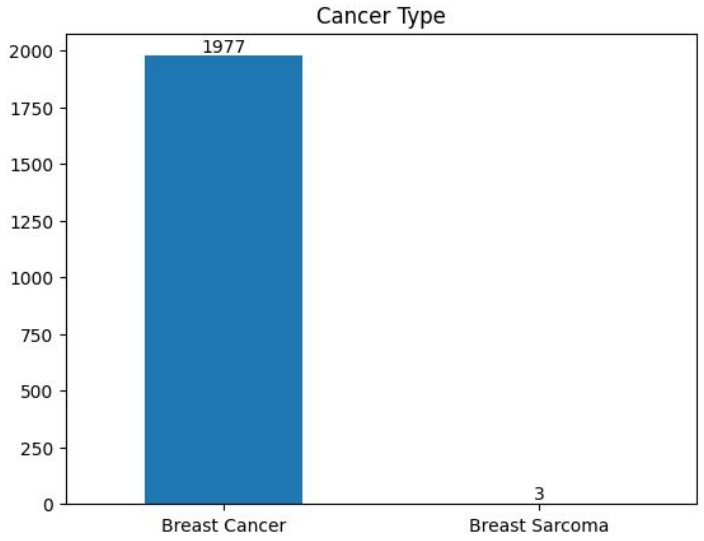
Overview

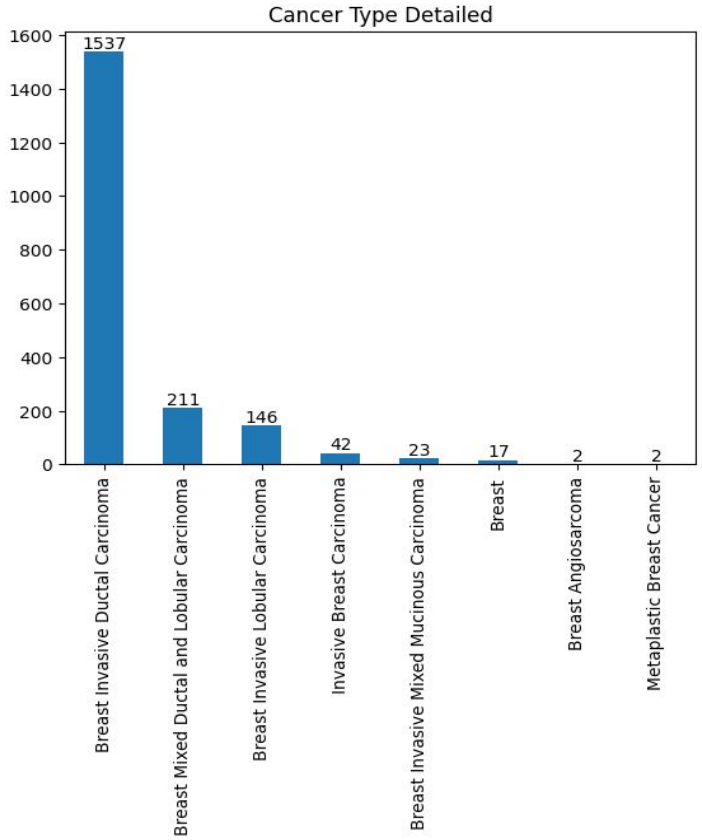
- **Existing solutions for breast cancer survival and relapse analysis:**
  - Commonly used methods: cox proportional hazards regression, decision trees, neural networks, and support vector machines
  - Predictive factors often include: nodal status, tumor size, tumor grade, age at diagnosis, and estrogen receptor status
  - Models like Nottingham Prognostic Index (NPI), and PREDICT v1.3 are frequently validated and used [1,2]
- **How our model is different than existing model / Why our model is state-of-the-art:**
  - Combining Random Survival Forests and Cox regression modeling
  - 2 scoring systems covering survival and relapse
  - Not only using clinical data but also genomics data
- **Assumptions about data:**
  - The data is representative of the underlying population and that the relationships captured in the data are stable over time
- **Health care impact - how our solution is to be used in health care decision system workflow:**
  - In clinical decision-making, the risk scoring system can aid in tailoring treatment plans, identifying high-risk patients, and potentially improving patient outcomes. By providing a risk score system, clinicians can make more informed decisions
- **Weaknesses in our solution and how it can be improved in next release:**
  - **External Validation:** Our model, built on a specific dataset, needs validation in different populations to ensure its generalizability
  - **Data Diversity:** Including data from more varied populations. Including other crucial factors, such as **genetic mutational profiles**, although the purpose of this study was to construct a prognostic model using clinically easy-to-use variables

[1] Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. PloS one, 16(4), e0250370. https://doi.org/10.1371/journal.pone.0250370
[2] Kalafi, E. Y., Nor, N. A. M., Taib, N. A., Ganggayah, M. D., Town, C., & Dhillon, S. K. (2019). Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. Folia biologica, 65(5-6), 212–220.

EDA



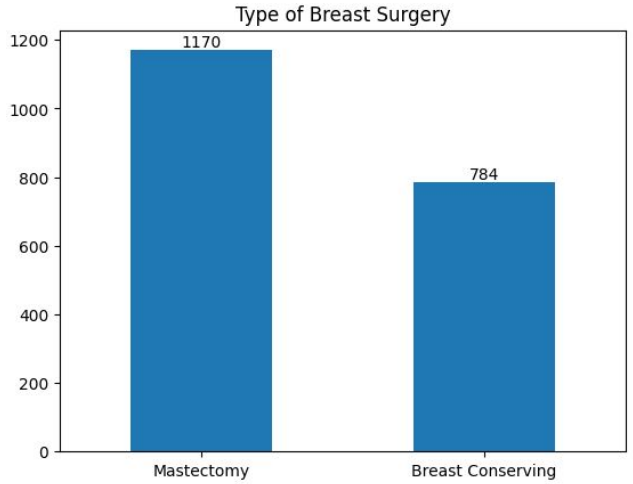


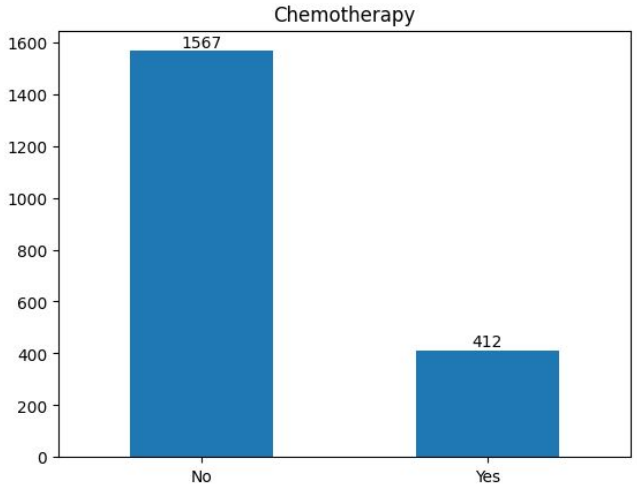


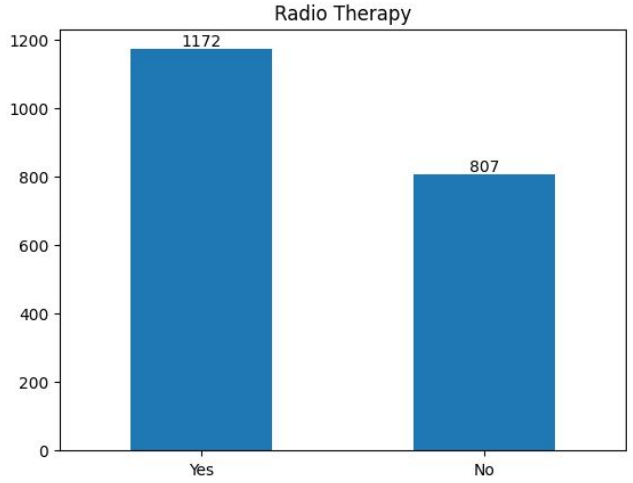
Age between 21.9-96.3
Mean: 60.4

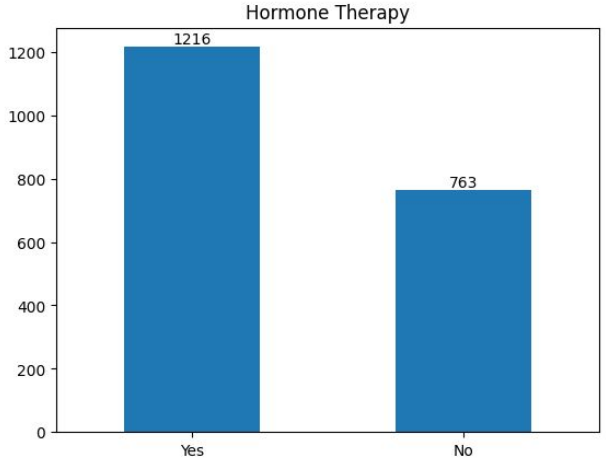
Cancer type distribution reflects real world scenarios accurately.

EDA - Surgery and Therapy

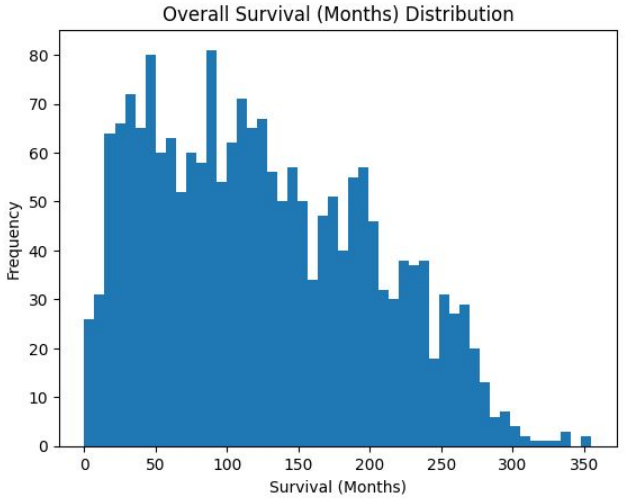


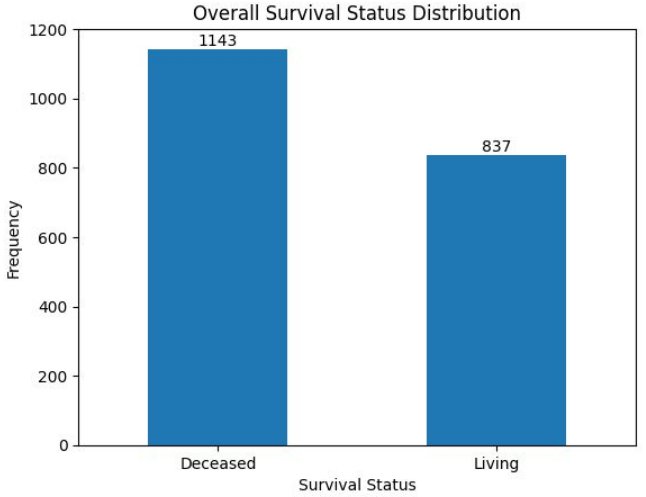


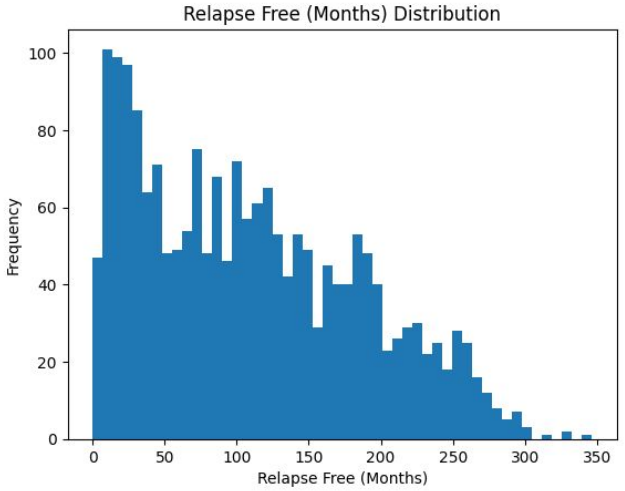


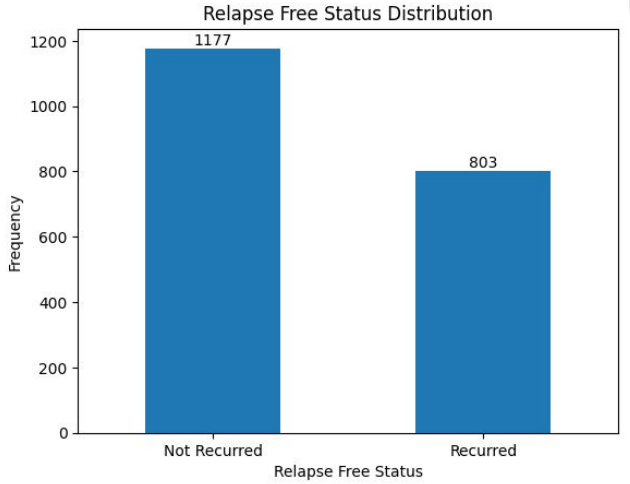


EDA - Events (Survival and Relapse)

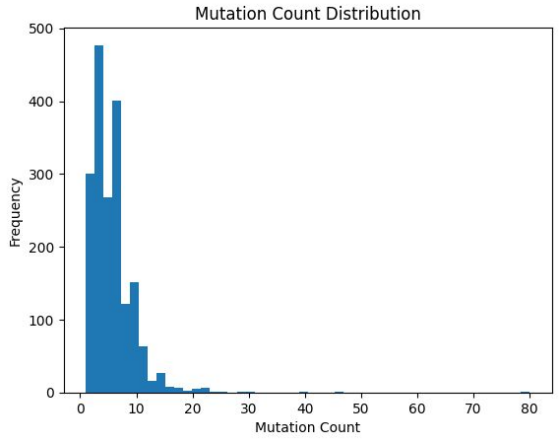


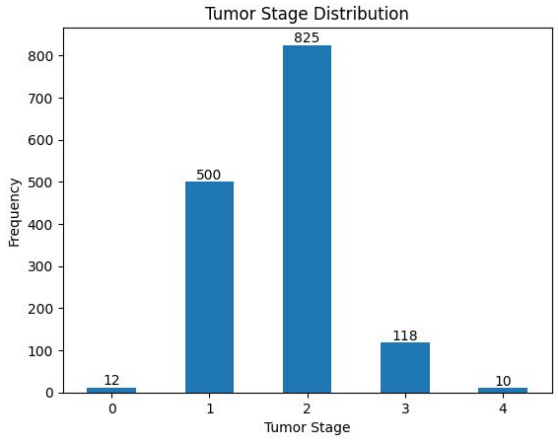


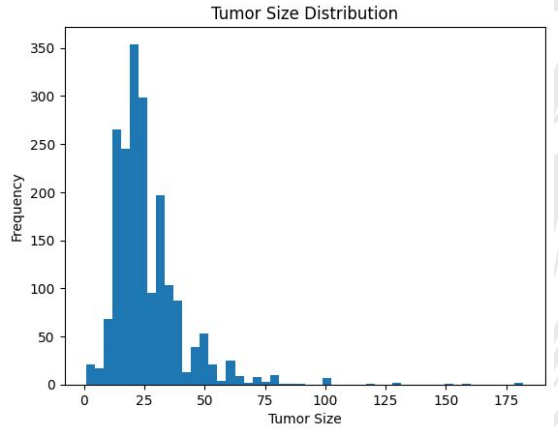


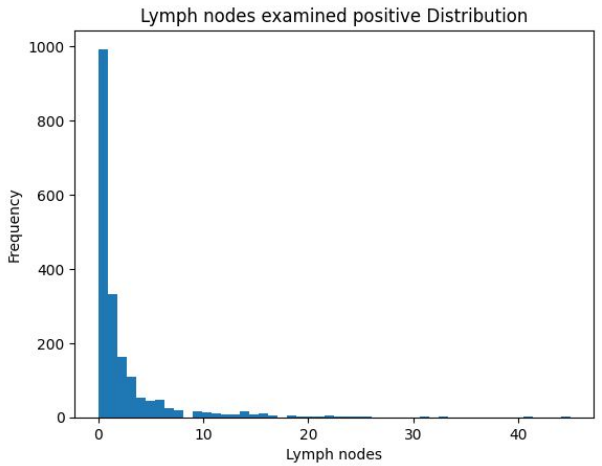


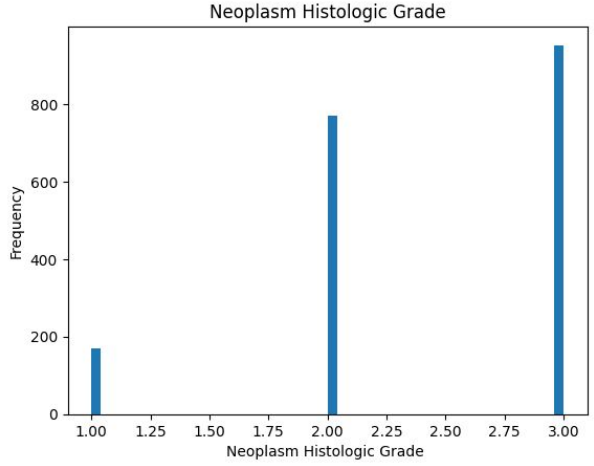
EDA - Important Predictors

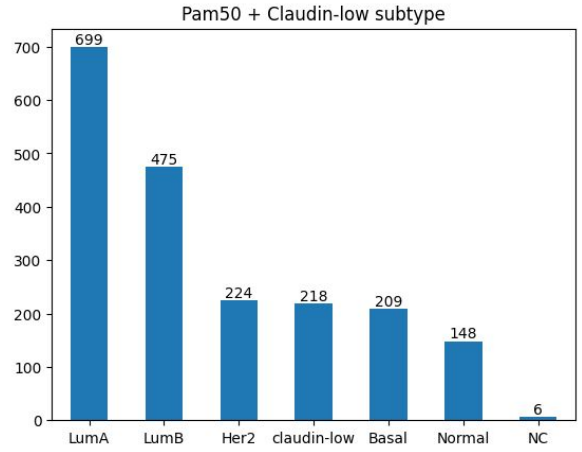




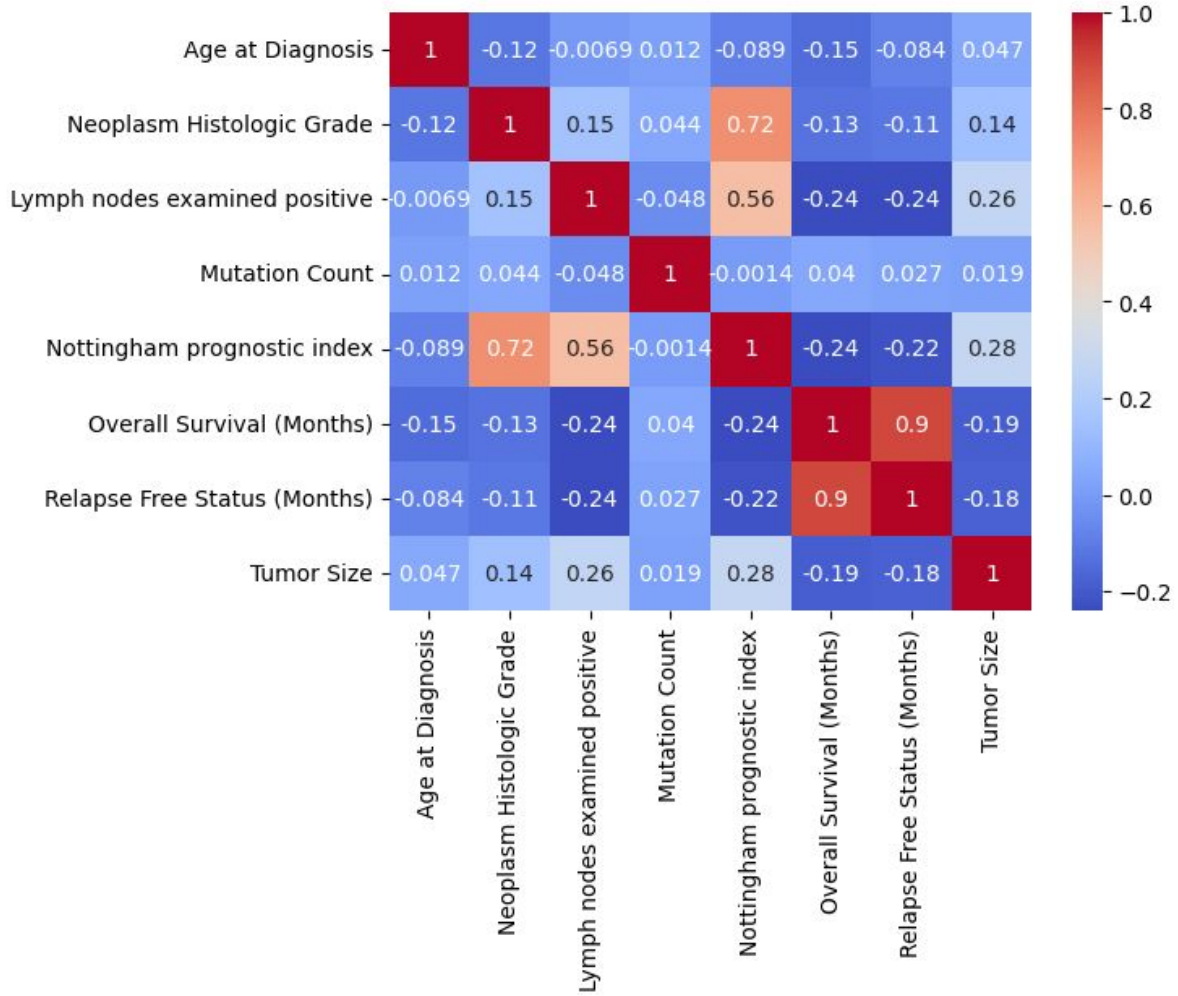




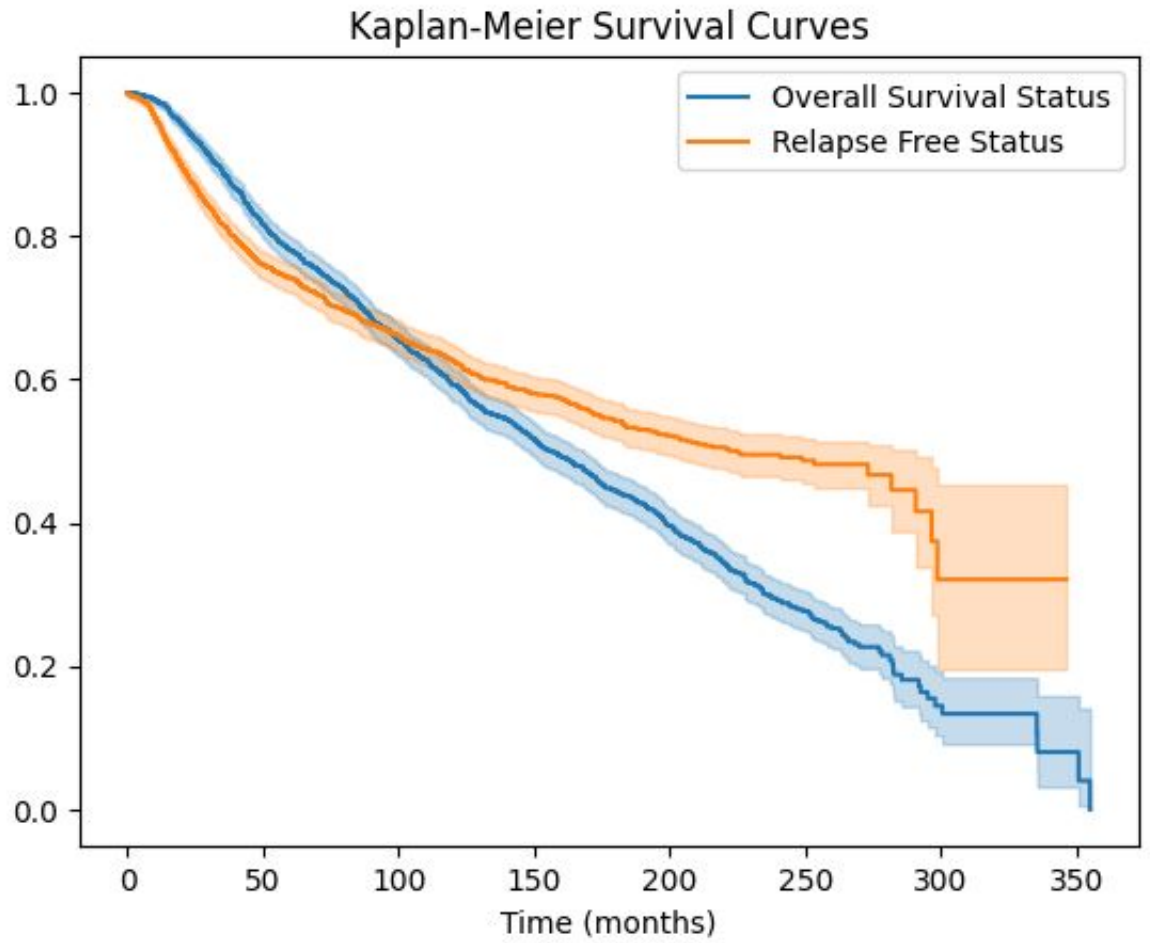




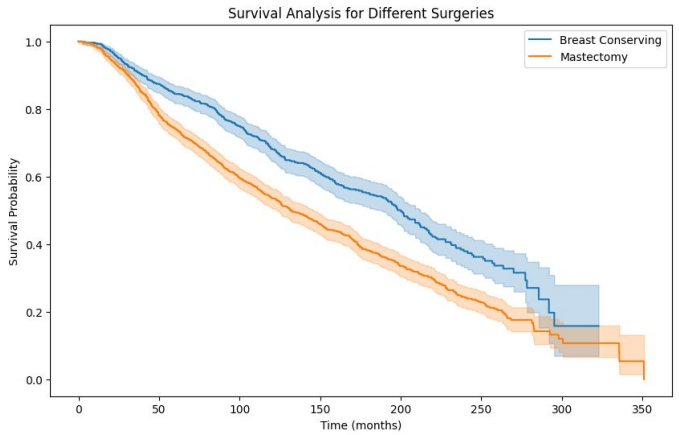
EDA - Correlation Matrix

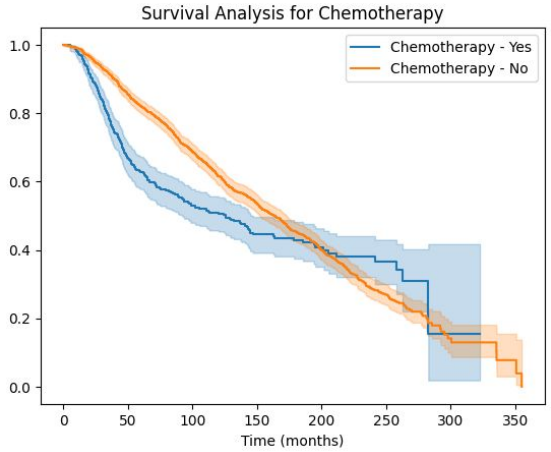


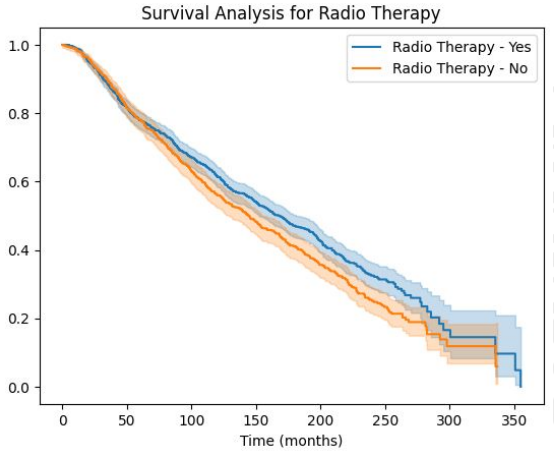
Kaplan Meier Survival & Relapse Curves

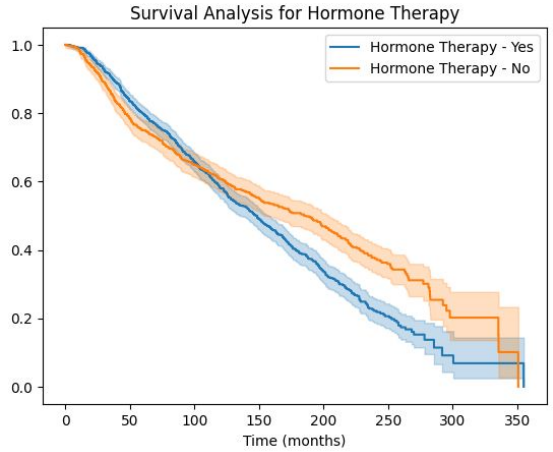


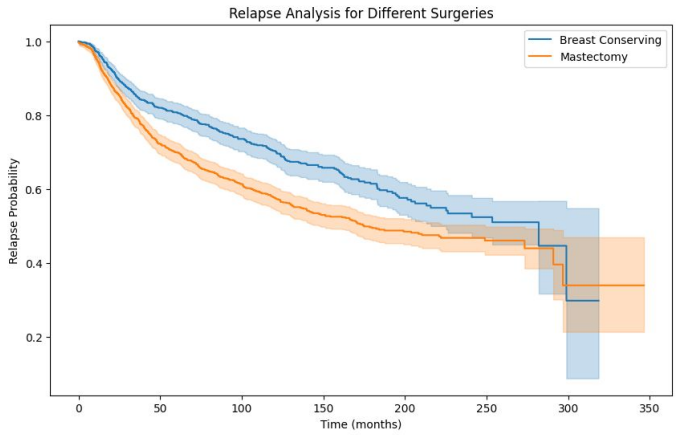
Kaplan Meier Curves - In terms of Treatments

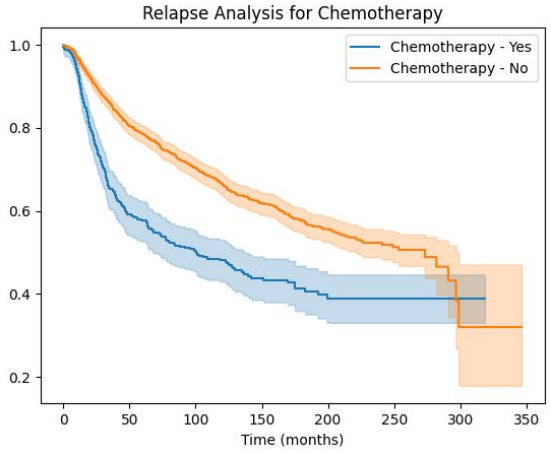


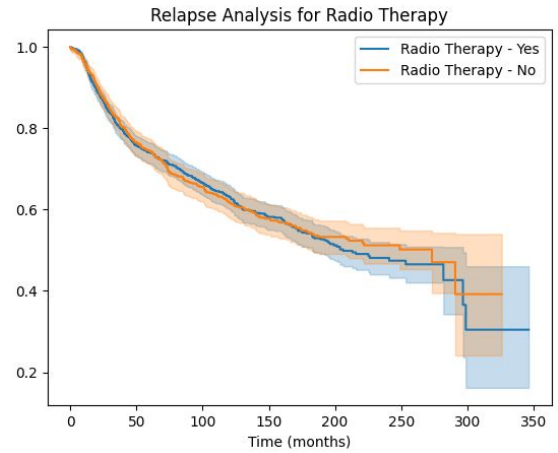


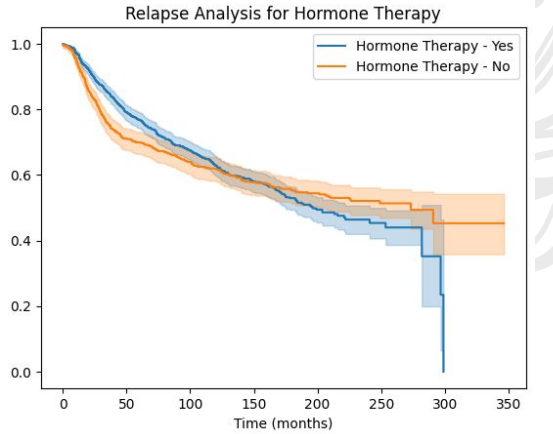












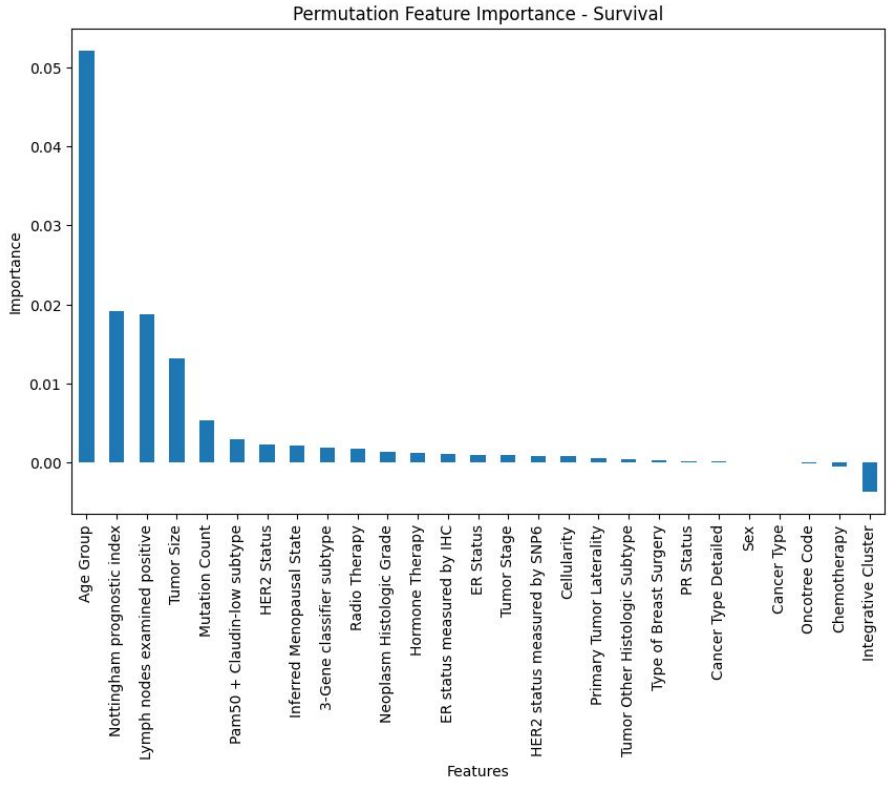
*The curves can not reflect any casual relationship or the effectiveness of the treatments.

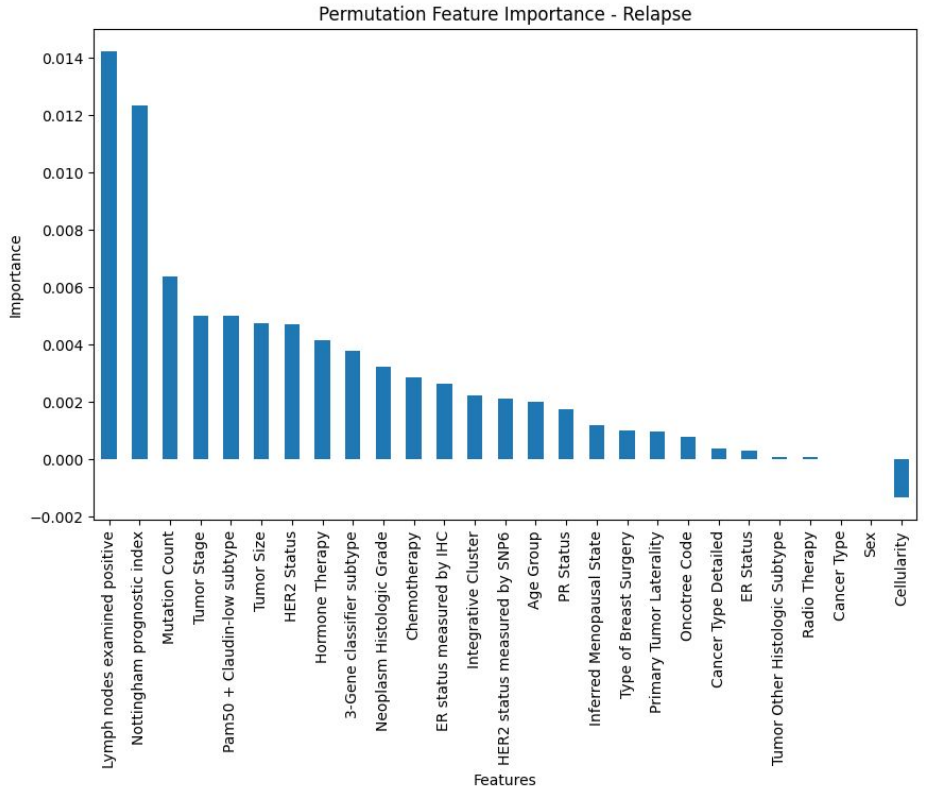
Random Survival Forest - Prediction

- Delete Missing Values
- Generate Age Groups for Age
- Train set (n=873) and Test set (n=219)
- Cross Validation
- Metric: Concordance Index

- For Survival, Concordance Index on Test Set: 69.29%
- For Relapse, Concordance Index on Test Set: 64.74%

Random Survival Forest - Feature Importances

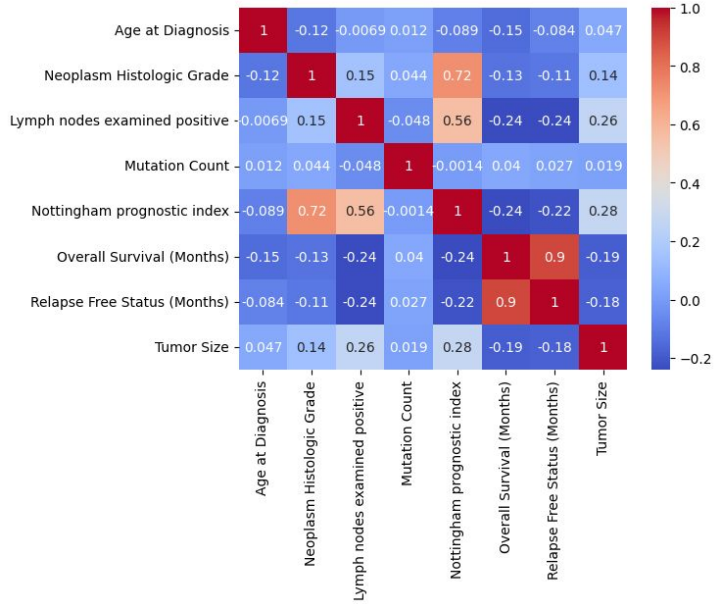




Re-train using top-10 features: Survival 70.2%, Relapse 65.13%

Variable Selection for Cox Model

- Identified variables with positive importance scores
- Remove variables with high correlation to prevent multicollinearity
- Remove variables calculated based on multiple other clinical or genomic markers



Relapse:

	Relapse Free Status	Relapse Free Status (Months)	Age Group	Lymph nodes examined positive	Mutation Count	Pam50 + Claudin-low subtype	Neoplasm Histologic Grade	3-Gene classifier subtype	PR Status	Tumor Stage	HER2 Status	Chemotherapy	Hormone Therapy	Radio Therapy	Inferred Menopausal State
1	0	83.52	40.0-50.0	0.0	2.0	LumA	3.0	ER+/HER2- High Prolif	Positive	1	Negative	No	Yes	Yes	Pre
4	1	18.55	70.0-80.0	8.0	2.0	LumB	3.0	ER+/HER2- High Prolif	Positive	2	Negative	Yes	Yes	Yes	Post
5	1	2.89	70.0-80.0	0.0	4.0	LumB	3.0	ER+/HER2- High Prolif	Positive	4	Negative	No	Yes	Yes	Post
10	0	36.09	80.0-90.0	1.0	4.0	LumB	3.0	ER+/HER2- High Prolif	Negative	2	Negative	No	Yes	Yes	Post
11	1	35.79	80.0-90.0	0.0	5.0	Her2	2.0	ER+/HER2- High Prolif	Negative	2	Negative	No	No	No	Post
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1697	0	83.88	70.0-80.0	4.0	11.0	LumA	2.0	ER+/HER2- Low Prolif	Positive	2	Negative	No	Yes	No	Post
1698	0	198.52	70.0-80.0	0.0	9.0	LumB	1.0	ER+/HER2- Low Prolif	Positive	1	Negative	No	No	Yes	Post
1700	0	103.82	70.0-80.0	0.0	4.0	Basal	3.0	ER-/HER2-	Negative	1	Negative	No	No	No	Post
1702	0	197.70	50.0-60.0	6.0	5.0	Normal	2.0	ER+/HER2- High Prolif	Positive	2	Negative	Yes	No	Yes	Post
1743	0	277.80	60.0-70.0	0.0	3.0	LumB	2.0	ER+/HER2- High Prolif	Positive	2	Negative	No	No	Yes	Post

Survival

	Overall Survival Status	Overall Survival (Months)	Age Group	Lymph nodes examined positive	Mutation Count	Pam50 + Claudin-low subtype	Neoplasm Histologic Grade	3-Gene classifier subtype	PR Status	Tumor Stage	HER2 status measured by SNP6	Radio Therapy	Inferred Menopausal State
1	0	84.633333	40.0-50.0	0.0	2.0	LumA	3.0	ER+/HER2- High Prolif	Positive	1	Neutral	Yes	Pre
4	1	41.366667	70.0-80.0	8.0	2.0	LumB	3.0	ER+/HER2- High Prolif	Positive	2	Neutral	Yes	Post
5	1	7.800000	70.0-80.0	0.0	4.0	LumB	3.0	ER+/HER2- High Prolif	Positive	4	Neutral	Yes	Post
10	1	36.566667	80.0-90.0	1.0	4.0	LumB	3.0	ER+/HER2- High Prolif	Negative	2	Gain	Yes	Post
11	1	36.266667	80.0-90.0	0.0	5.0	Her2	2.0	ER+/HER2- High Prolif	Negative	2	Loss	No	Post
...	...	...	...	...	...	...	...	...	...	...	...	...	...
1697	1	85.000000	70.0-80.0	4.0	11.0	LumA	2.0	ER+/HER2- Low Prolif	Positive	2	Neutral	No	Post
1698	0	201.166667	70.0-80.0	0.0	9.0	LumB	1.0	ER+/HER2- Low Prolif	Positive	1	Neutral	Yes	Post
1700	1	105.200000	70.0-80.0	0.0	4.0	Basal	3.0	ER-/HER2-	Negative	1	Gain	No	Post
1702	0	200.333333	50.0-60.0	6.0	5.0	Normal	2.0	ER+/HER2- High Prolif	Positive	2	Neutral	Yes	Post
1743	0	281.500000	60.0-70.0	0.0	3.0	LumB	2.0	ER+/HER2- High Prolif	Positive	2	Neutral	Yes	Post

1092 rows × 13 columns

Cox Model

Relapse

```
from sklearn.preprocessing import StandardScaler
df_relapse = pd.get_dummies(df_relapse, drop_first=True)

scaler = StandardScaler()
df_relapse_scaled = scaler.fit_transform(df_relapse)
df_relapse_scaled = pd.DataFrame(df_relapse_scaled, columns=df_relapse.columns)
cph = CoxPHFitter(penalizer=0.1)
cph.fit(df_relapse, duration_col='Relapse Free Status (Months)', event_col='Relapse Free Status'
print(cph.summary)
cph.check_assumptions(df_relapse, p_value_threshold=0.05)
```

```
                                      coef  exp(coef)  se(coef)  \
covariate
Neoplasm Histologic Grade             0.078281   1.081427  0.080537
Age Group_30.0-40.0                   0.004037   1.004045  0.221906
Age Group_40.0-50.0                  -0.104562   0.900719  0.175522
Age Group_50.0-60.0                  -0.060983   0.940839  0.139044
Age Group_60.0-70.0                  -0.027626   0.972752  0.134432
Age Group_70.0-80.0                   0.048072   1.049247  0.148278
Age Group_80.0-90.0                  -0.146121   0.864053  0.230743
Age Group_90.0-100.0                  1.019093   2.770679  0.604426
Pam50 + Claudin-low subtype_Her2     -0.022473   0.977778  0.169560
Pam50 + Claudin-low subtype_LumA     -0.119334   0.887512  0.132520
Pam50 + Claudin-low subtype_LumB      0.102534   1.107975  0.137660
Pam50 + Claudin-low subtype_NC       -0.127654   0.880158  0.913223
Pam50 + Claudin-low subtype_Normal    0.295173   1.343359  0.189499
Pam50 + Claudin-low subtype_claudin-low -0.108107   0.897532  0.167337
3-Gene classifier subtype_ER-/HER2-  -0.180833   0.834575  0.155313
3-Gene classifier subtype_HER2+       0.025666   1.025998  0.185294
Tumor Stage_1                        -0.296515   0.743405  0.145165
Tumor Stage_2                         0.067480   1.069809  0.139614
Tumor Stage_3                         0.498517   1.646277  0.184379
Tumor Stage_4                         1.782273   5.943350  0.465355
HER2 Status_Positive                  0.249984   1.284005  0.176301
Chemotherapy_Yes                      0.112868   1.119484  0.124035
Radio Therapy_Yes                    -0.127526   0.880271  0.094544
Inferred Menopausal State_Pre         0.073377   1.076136  0.171889
Mutation_Count_Group_(10.0, 19.0]     0.221739   1.248245  0.182663
Mutation_Count_Group_(19.0, 28.0]    -0.441652   0.642973  0.504304
Mutation_Count_Group_(28.0, 37.0]     1.081278   2.948444  1.027069
Mutation_Count_Group_(37.0, 46.0]    -1.215305   0.296619  1.507201
lymph_nodes_group_Group 2             0.815500   2.260306  0.186735
lymph_nodes_group_Group 3             0.807245   2.241725  0.339174
lymph_nodes_group_Group 4             1.231974   3.427990  0.639903
lymph_nodes_group_Group 5            -0.735393   0.479317  2.403384
```

Survival

```
from sklearn.preprocessing import StandardScaler
df_survival = pd.get_dummies(df_survival, drop_first=True)

scaler = StandardScaler()
df_survival_scaled = scaler.fit_transform(df_survival)
df_survival_scaled = pd.DataFrame(df_survival_scaled, columns=df_survival.columns)
cph = CoxPHFitter(penalizer=0.1)
cph.fit(df_survival, duration_col='Overall Survival (Months)', event_col='Overall Survival Stat
print(cph.summary)
cph.check_assumptions(df_survival, p_value_threshold=0.05)
```

```
                                                 coef  exp(coef)  se(coef)  \
covariate
Neoplasm Histologic Grade                    0.099953   1.105118  0.071867
Age Group_30.0-40.0                         -0.139593   0.869712  0.230363
Age Group_40.0-50.0                         -0.239916   0.786694  0.178822
Age Group_50.0-60.0                         -0.279494   0.756166  0.147217
Age Group_60.0-70.0                          0.145438   1.156546  0.141801
Age Group_80.0-90.0                          0.813623   2.256067  0.193883
Age Group_90.0-100.0                         1.420569   4.139476  0.546264
Pam50 + Claudin-low subtype_Her2             0.105675   1.111461  0.159937
Pam50 + Claudin-low subtype_LumB            -0.093186   0.911024  0.143363
Pam50 + Claudin-low subtype_NC              -0.757021   0.469062  0.874016
Pam50 + Claudin-low subtype_Normal           0.152191   1.164383  0.196218
Pam50 + Claudin-low subtype_claudin-low     -0.152819   0.858285  0.161531
3-Gene classifier subtype_ER+/HER2- Low Prolif -0.079090   0.923957  0.106064
3-Gene classifier subtype_ER-/HER2-          0.043246   1.044195  0.144026
3-Gene classifier subtype_HER2+              0.088148   1.092150  0.153872
Tumor Stage_1                               -0.286419   0.750948  0.140067
Tumor Stage_2                                0.105667   1.111452  0.136823
Tumor Stage_3                                0.420071   1.522070  0.176369
Tumor Stage_4                                0.909557   2.483223  0.473438
HER2 status measured by SNP6_Loss           -0.116971   0.889611  0.197340
HER2 status measured by SNP6_Neutral        -0.184654   0.831392  0.112029
HER2 status measured by SNP6_Undef          -0.005446   0.994569  0.704563
Radio Therapy_Yes                           -0.225006   0.798511  0.082309
Inferred Menopausal State_Pre               -0.144443   0.865504  0.173385
Mutation_Count_Group_(19.0, 28.0]            0.368498   1.445561  0.399617
Mutation_Count_Group_(28.0, 37.0]            1.244159   3.470014  1.034598
Mutation_Count_Group_(37.0, 46.0]           -0.435179   0.647149  0.712986
lymph_nodes_group_Group 2                    0.778957   2.179198  0.170153
lymph_nodes_group_Group 3                    0.615128   1.849893  0.346309
lymph_nodes_group_Group 4                    0.296379   1.344979  0.733762
lymph_nodes_group_Group 5                    3.069610  21.533499  1.167769
```

Scoring System Based on HR:

Intuitive and Effective

Hazed Ratio:

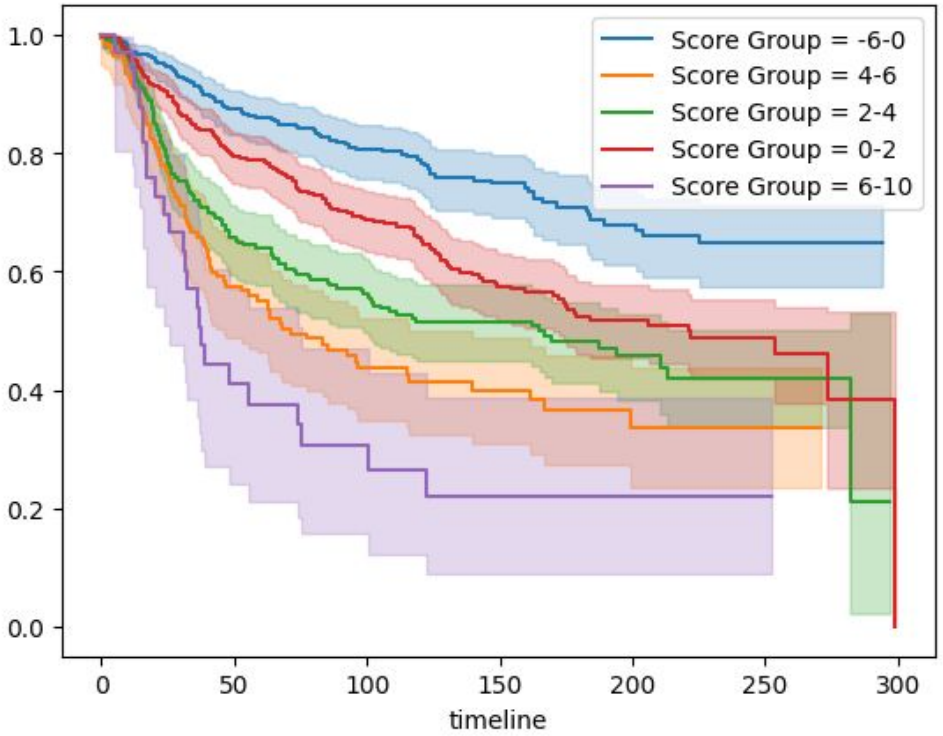
- HR > 1: Indicates increased risk associated with the covariate.
- HR < 1: Suggests reduced risk.
- HR = 1: No significant impact on risk.

Score Assignment:

- Score +1 if HR > 1
  Indicates that the covariate increases the hazard, or risk of the event.
- Score -1 if HR < 1
  Indicates that the covariate decreases the hazard, or risk of the event.
- A HR of 1 implies no effect and hence, no score is assigned.

Grouping and Kaplan-Meier Analysis: Relapse

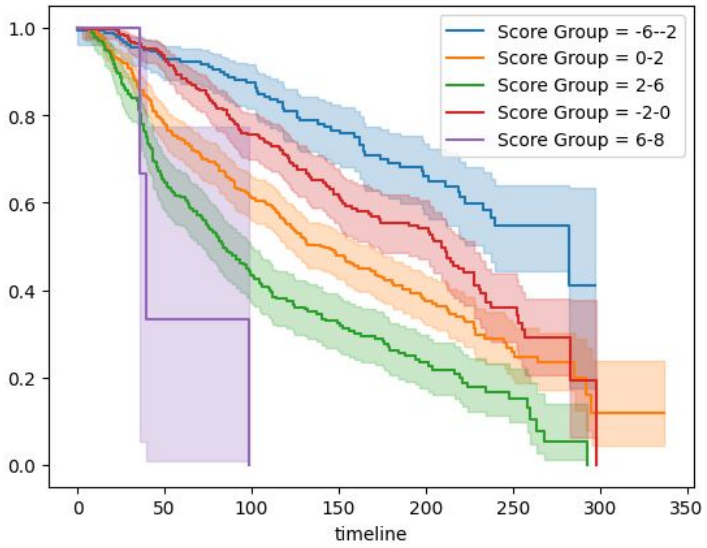
- dividing patients into groups based on overall scores.
- Kaplan-Meier curves for each group.
- Comparison of median survival times.



Median Survival Time for Score Group -6-0: inf months
Median Survival Time for Score Group 0-2: 220.89 months
Median Survival Time for Score Group 2-4: 165.07 months
Median Survival Time for Score Group 4-6: 71.12 months
Median Survival Time for Score Group 6-10: 37.53 months

Grouping and Kaplan-Meier Analysis: Survival

- dividing patients into groups based on overall scores.
- Kaplan-Meier curves for each group.
- Comparison of median survival times.

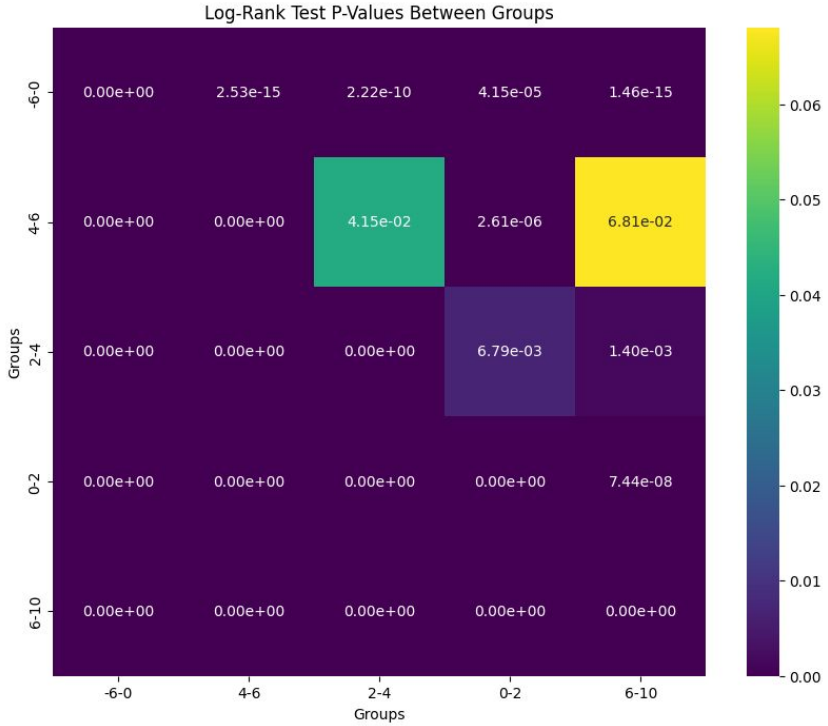


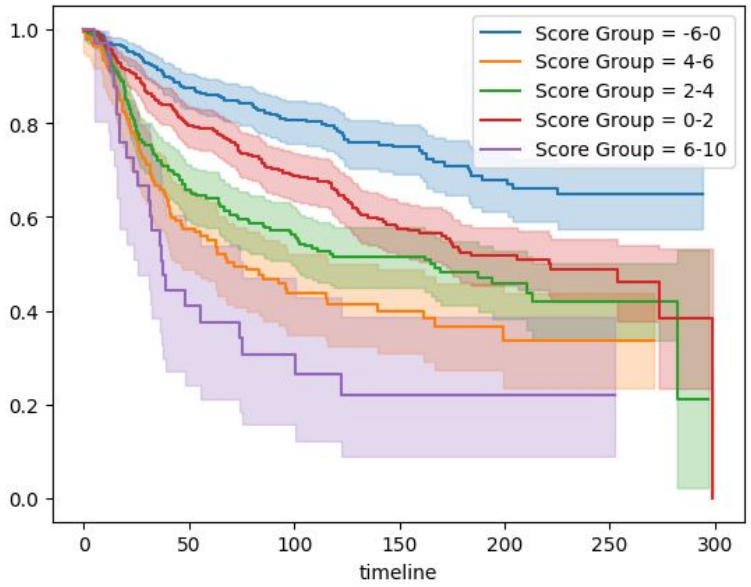
Median Survival Time for Score Group -6--2: 282.5666667 months
Median Survival Time for Score Group -2-0: 209.0333333 months
Median Survival Time for Score Group 0-2: 141.56666669999998 months
Median Survival Time for Score Group 2-6: 83.66666667 months
Median Survival Time for Score Group 6-8: 39.53333333 months

Statistical Significance and Log Rank Test: Relapse

log rank test: test the null hypothesis that there is no difference in the survival experiences of the different groups being compared.

Presentation of p-values and their significance



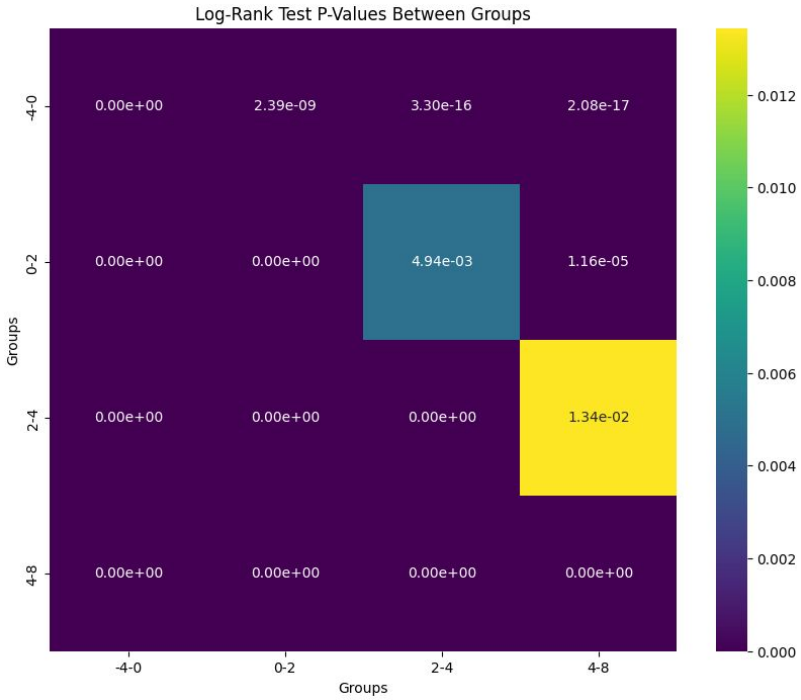


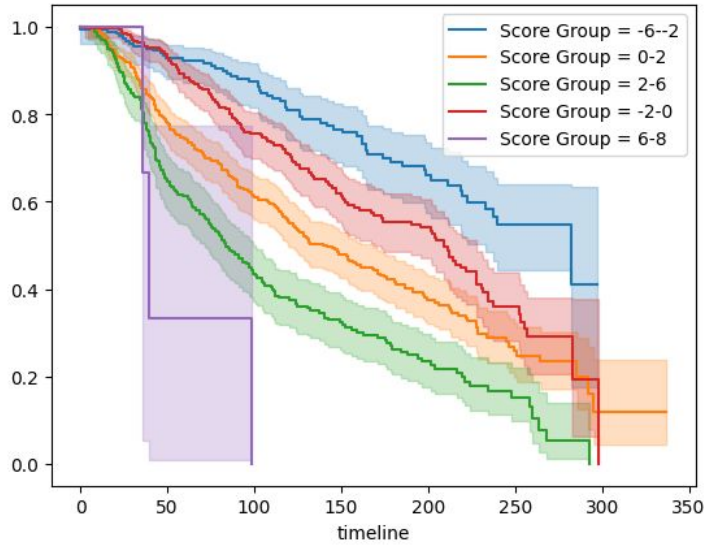
Median Survival Time for Score Group -6-0: inf months
Median Survival Time for Score Group 0-2: 220.89 months
Median Survival Time for Score Group 2-4: 165.07 months
Median Survival Time for Score Group 4-6: 71.12 months
Median Survival Time for Score Group 6-10: 37.53 months

Statistical Significance and Log Rank Test: Survival

log rank test: test the null hypothesis that there is no difference in the survival experiences of the different groups being compared.

Presentation of p-values and their significance





Median Survival Time for Score Group -6--2: 282.5666667 months
Median Survival Time for Score Group -2-0: 209.0333333 months
Median Survival Time for Score Group 0-2: 141.56666669999998 months
Median Survival Time for Score Group 2-6: 83.66666667 months
Median Survival Time for Score Group 6-8: 39.53333333 months

Accuracy examples:

Relapse:

Concordance Index for Score Group -6-0: 0.8208516886930984

Concordance Index for Score Group 4-6: 0.7414448669201521

Concordance Index for Score Group 2-4: 0.7615715823466093

Concordance Index for Score Group 0-2: 0.7431597528684908

Concordance Index for Score Group 6-10: 0.7894736842105263

Survival:

Concordance Index for Score Group -6--2: 0.7772435897435898

Concordance Index for Score Group 0-2: 0.7888157894736842

Concordance Index for Score Group 2-6: 0.886039886039886

Concordance Index for Score Group -2-0: 0.8089330024813896

Conclusion

Our approach has successfully integrated variables identified through a Random Forest Survival algorithm with practical clinical knowledge into a Cox model. This has enabled us to develop straightforward yet clinically impactful scoring models.

The scoring models enhances clinical decision-making processes. For example the relapse scoring system is useful in selecting therapies that potentially extend the duration of relapse-free periods, thereby improving patient outcomes.

Moreover, it provides valuable prognostic information, allowing for a more personalized approach to patient care.

Future Work

Data Source	The scoring systems will benefit from more clinically used information such as images
Healthcare Professionals	We see immense value in collaborating with clinical practitioners. Their insights and validation would be instrumental in refining our models and also bridge the gap between data-driven research and practical clinical application
Real-Time Integration	Explore ways to integrate the scoring system into real-time clinical workflows, such as electronic health record systems, to enhance its practical utility

Reference

[1] Li, J., Zhou, Z., Dong, J., Fu, Y., Li, Y., Luan, Z., & Peng, X. (2021). Predicting breast cancer 5-year survival using machine learning: A systematic review. PloS one, 16(4), e0250370. https://doi.org/10.1371/journal.pone.0250370
[2] Kalafi, E. Y., Nor, N. A. M., Taib, N. A., Ganggayah, M. D., Town, C., & Dhillon, S. K. (2019). Machine Learning and Deep Learning Approaches in Breast Cancer Survival Prediction Using Clinical Data. Folia biologica, 65(5-6), 212–220
[3]Kim J, Hong JY, Kim ST, Park SH, Jekal SY, Choi JS, Chang DK, Kang WK, Seo SW, Lee J. Clinical scoring system for the prediction of survival of patients with advanced gastric cancer. ESMO Open. 2020 Mar;5(2):e000670. doi: 10.1136/esmoopen-2020-000670. PMID: 32188716; PMCID: PMC7078777.
[4]Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, Michael S. Lauer "Random survival forests," The Annals of Applied Statistics, Ann. Appl. Stat. 2(3), 841-860

Extra:

- HER2 Status: This indicates the overexpression of the Human Epidermal growth factor Receptor 2 in cancer cells. HER2-positive breast cancers tend to be more aggressive but may respond to targeted therapies like trastuzumab (Herceptin).
- Hormone Therapy: This indicates whether a patient has received treatment that lowers hormone levels or blocks their effects on breast cancer cells. Hormone therapies are typically used for cancers that are hormone receptor-positive.
- Pam50 + Claudin-low subtype: PAM50 is a test that categorizes breast cancer into one of several molecular subtypes for prognostic and predictive purposes. The Claudin-low subtype is characterized by low expression of claudin genes, associated with more aggressive disease and a stem cell-like phenotype.
- 3-Gene classifier subtype: This refers to a prognostic tool based on the expression of three genes. It helps predict if certain early-stage breast cancer patients will benefit from chemotherapy in addition to hormonal therapy.
- PR Status: The Progesterone Receptor Status of the cancer, indicating whether the cancer cells have receptors for the hormone progesterone. PR-positive cancers are typically more responsive to hormone therapy.
- Tumor Stage: This provides a description of the extent of cancer, based on the size of the tumor, and whether it has spread to nearby lymph nodes or other parts of the body. Staging helps determine the severity of the cancer and guide treatment.
- Oncotree Code: A comprehensive oncology data standard that provides a detailed classification system for tumor morphology, including anatomic site, histology, and subtype.
- Tumor Other Histologic Subtype: Beyond the main histologic types of breast cancer (like ductal or lobular), there are several subtypes based on the cancer cells' microscopic appearance, which can have different prognostic and treatment implications.
- Nottingham Prognostic Index: A clinical tool used to predict survival outcomes of breast cancer patients. It is based on tumor size, lymph node status, and histological grade.
- PREDICT v1.3: Predict is an online tool. Predict asks for some details about the patient and the cancer. It then uses data about the survival of similar women in the past to show the likely proportion of such women expected to survive up to fifteen years after their surgery with different treatment combinations.

Random Survival Forest:

- "Random Survival Forest is a machine learning technique specifically designed for survival data. It extends the Random Forest approach, combining the strengths of decision trees with the complexities of survival analysis."

Building Decision Trees:

- "The process starts by creating multiple decision trees. Each tree is constructed using a different bootstrap sample from the original dataset. This sampling is done with replacement, meaning some observations may be repeated in each sample."

Splitting Nodes Using Survival Analysis:

- "In each tree, nodes are split based on criteria that best separate the survival outcomes. Unlike standard decision trees, RSF uses measures like the difference in survival times or hazard ratios to determine the best split at each node."

Aggregating Tree Predictions:

- "After building numerous trees, RSF aggregates their predictions. This aggregation is typically done using a method called ensemble learning, where the survival outcomes from all trees are combined to produce a more accurate and robust prediction."

Estimating Survival Function:

- "The final step involves estimating the survival function for each individual in the dataset. The RSF model does this by averaging the survival functions across all the trees, providing a comprehensive survival curve for each individual."

Advantages of RSF:

- "Random Survival Forest is beneficial due to its accuracy, ability to handle large and complex datasets, and its non-reliance on underlying assumptions about the distribution of survival times, unlike some traditional survival analysis methods."

Numerical variables group:

```
bins = [0, 8, 16, 24, 32, 41]

df_survival['lymph_nodes_group'] = pd.cut(df_survival['Lymph nodes examined positive'], bins, include_lowest=True, labels=['Group 1', 'Group 2', 'Group 3', 'Group 4', 'Group 5'])

print(df_survival[['Lymph nodes examined positive', 'lymph_nodes_group']])
```

```
mutation_count_min = df_survival['Mutation Count'].min()
mutation_count_max = df_survival['Mutation Count'].max()

bins = pd.interval_range(start=mutation_count_min, end=mutation_count

df_survival['Mutation_Count_Group'] = pd.cut(df_survival['Mutation Co

print(df_survival[['Mutation Count', 'Mutation_Count_Group']])
```

```
      Mutation Count Mutation_Count_Group
.                2.0          (1.0, 10.0]
4                2.0          (1.0, 10.0]
5                4.0          (1.0, 10.0]
.0               4.0          (1.0, 10.0]
.1               5.0          (1.0, 10.0]
..               ...                  ...
.697            11.0         (10.0, 19.0]
.698             9.0          (1.0, 10.0]
.700             4.0          (1.0, 10.0]
.702             5.0          (1.0, 10.0]
.743             3.0          (1.0, 10.0]
```

Log-Rank:

The logrank test statistic compares estimates of the hazard functions of the two groups at each observed event time. It is constructed by computing the observed and expected number of events in one of the groups at each observed event time and then adding these to obtain an overall summary across all-time points where there is an event.

Consider two groups of patients, e.g., treatment vs. control. Let $1, \ldots, J$ be the distinct times of observed events in either group. Let $N_{1,j}$ and $N_{2,j}$ be the number of subjects "at risk" (who have not yet had an event or been censored) at the start of period j in the groups, respectively. Let $O_{1,j}$ and $O_{2,j}$ be the observed number of events in the groups at time j. Finally, define $N_j = N_{1,j} + N_{2,j}$ and $O_j = O_{1,j} + O_{2,j}$.

The null hypothesis is that the two groups have identical hazard functions, $H_0 : h_1(t) = h_2(t)$. Hence, under H_0, for each group $i = 1, 2$, $O_{i,j}$ follows a hypergeometric distribution with parameters N_j, $N_{i,j}$, O_j. This distribution has expected value $E_{i,j} = O_j \frac{N_{i,j}}{N_j}$ and variance

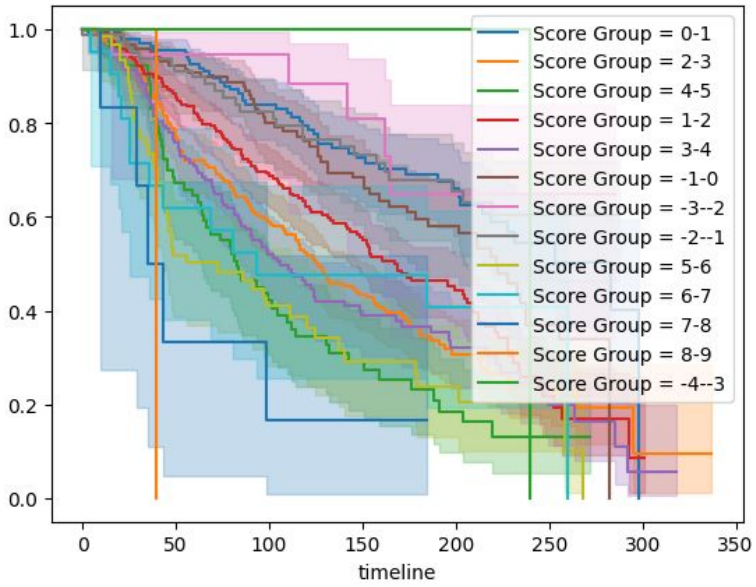
$$V_{i,j} = E_{i,j} \left(\frac{N_j - O_j}{N_j} \right) \left(\frac{N_j - N_{i,j}}{N_j - 1} \right).$$

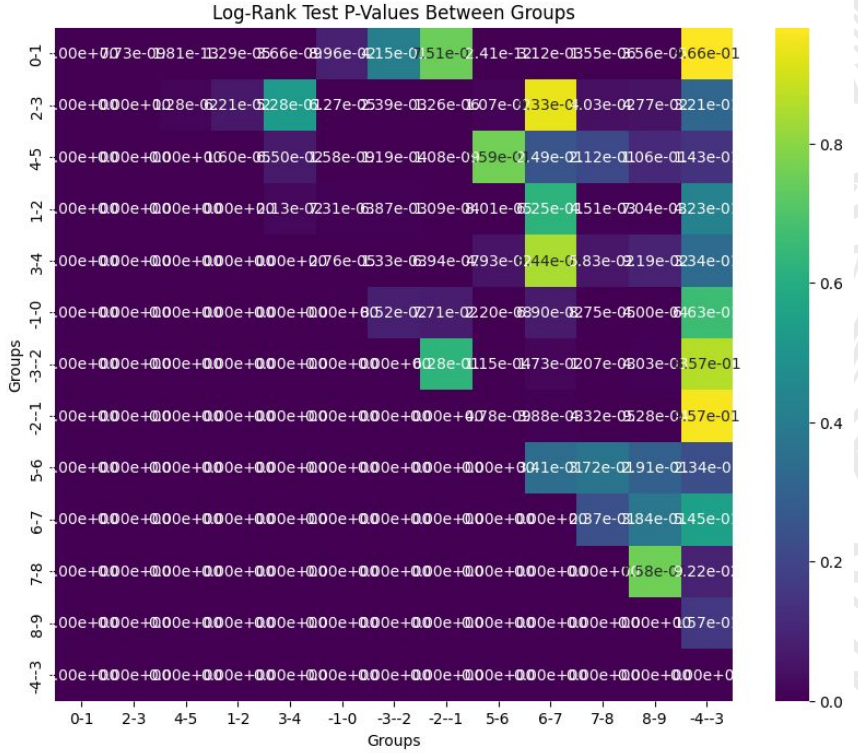
For all $j = 1, \ldots, J$, the logrank statistic compares $O_{i,j}$ to its expectation $E_{i,j}$ under H_0. It is defined as

$$Z_i = \frac{\sum_{j=1}^{J} (O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0,1) \qquad \text{(for } i = 1 \text{ or } 2\text{)}$$

By the central limit theorem, the distribution of each Z_i converges to that of a standard normal distribution as J approaches infinity and therefore can be approximated by the standard normal distribution for a sufficiently large J. An improved approximation can be obtained by equating this quantity to Pearson type I or II (beta) distributions with matching first four moments, as described in Appendix B of the Peto and Peto paper.[2]

Trial and Errors





Standard Cox and Stratified Cox

Standard Cox proportional hazards model:

$$h(t|Z) = h_0 \exp(\beta_{\text{age}} Z_{\text{age}} + \beta_{\text{rx}} Z_{\text{rx}})$$

Stratified Cox proportional hazards model (you're actually fitting two equations that share a single β_{age}):

$$h_{\text{rx}=2}(t|Z) = h_{0,\text{rx}=2} \exp(\beta_{\text{age}} Z_{\text{age}})$$
$$h_{\text{rx}=1}(t|Z) = h_{0,\text{rx}=1} \exp(\beta_{\text{age}} Z_{\text{age}})$$

Thank you!