Survival Analysis – Breast Cancer
Team 7: Nathan Brunet, Aurélie Maugein, Hugo Boileau, Brandt Olson
2024-08-11

Introduction

For this project we have modelled breast cancer survival. We have employed a data cleansing techniques, complete case analysis, and utilised several know survival analyses techniques, such as, Cox Regressions and Random Survival Forest.

Part 1: Data

1.1 Source

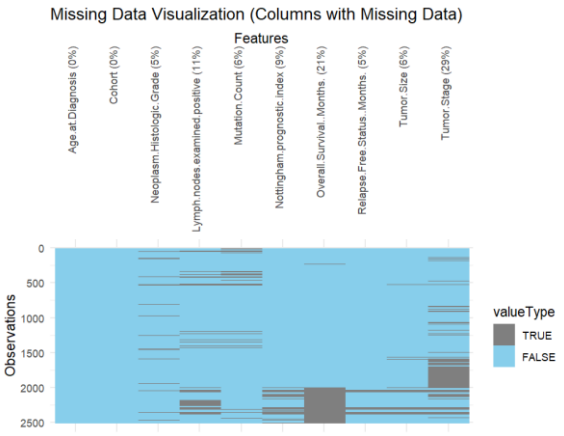
- Breast Cancer METABRIC (Molecular Taxonomy of Breast Cancer International Consortium)

1.2 Overview

- Clinical profiles of 2,509 breast cancer patients with 34 variables, such as age at diagnosis, estrogen receptor (ER) status and positive lymph node count

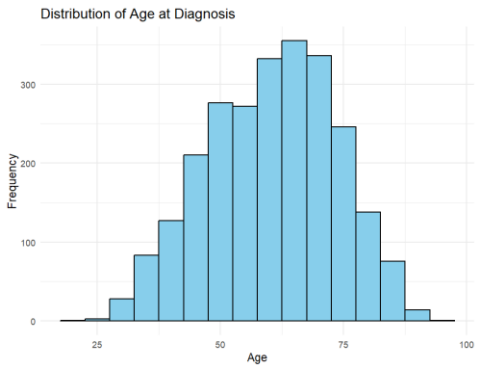
1.3 Data Cleansing

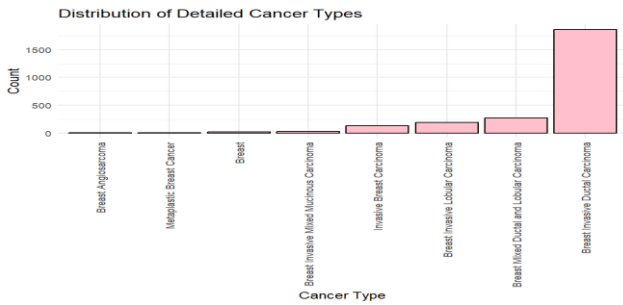
- Several variables had missing data, notably Overall Survival Months (528 missing) and Tumor Stage (721 missing)
- Complete case analysis was used, meaning that any row with missing values in the selected important variables was removed
- After cleaning and selecting important variables, the analysis dataset contains 1833 observations, reducing the dataset by 676 observations (about 27% of the original data)



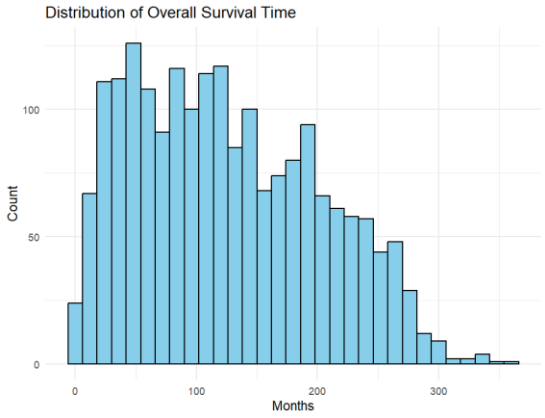
1.4 Key Data

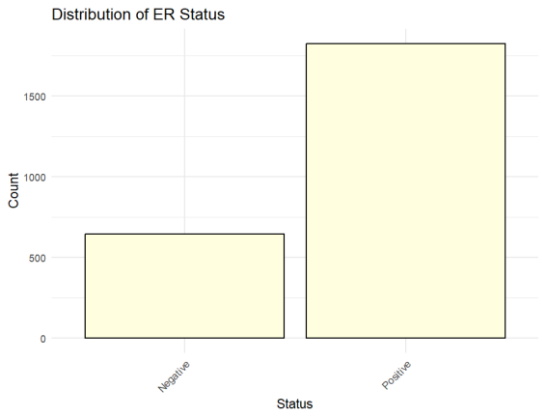
- Age at Diagnosis: Approximately normally distributed, centred around 60 years.
- Overall Survival Time: Right-skewed distribution, with a median around 116 months.
- Cancer Type: All patients have breast cancer





- Survival Status: More patients were living than deceased at the end of the study.
- Estrogen Receptor (ER) Status: Majority of patients are ER-positive (1350) compared to ER-negative (420)

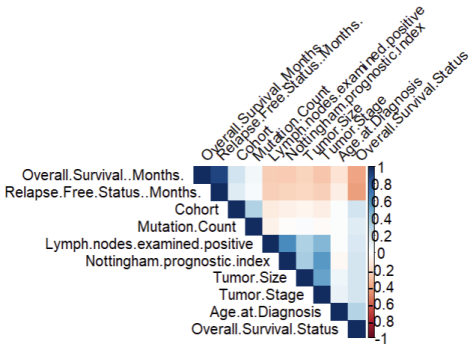


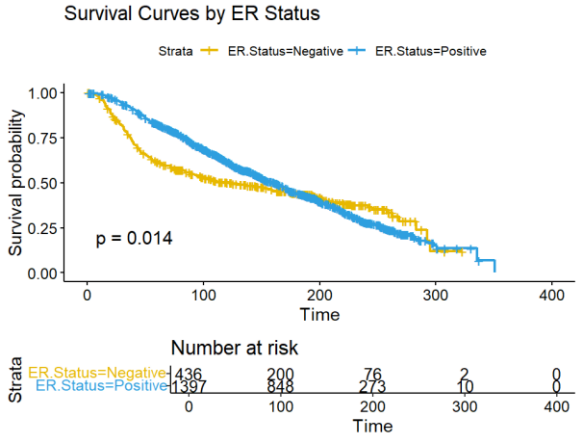


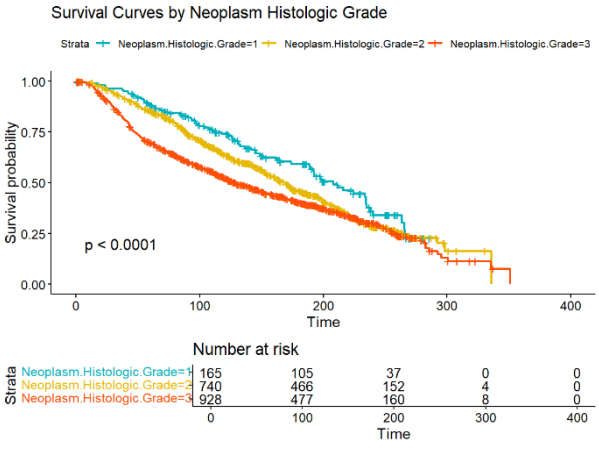
Part 2: Survival Analysis

2.1 Cox Proportional Hazards (PH) Model:

- The variables: Age, ER Status, Neoplasm Grade 3, and Lymph nodes, are significant predictors
- Each year increase in age increases hazard by 4% (HR = 1.0397)
- ER-positive status decreases hazard by 30% (HR = 0.7008)
- Grade 3 tumours increase hazard by 43% compared to Grade 1 (HR = 1.4269)
- Each additional positive lymph node increases hazard by 6.3% (HR = 1.0632)
- Concordance = 0.65, indicating moderate predictive ability

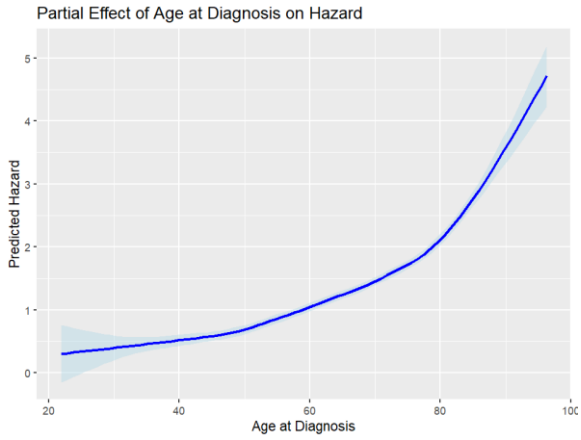


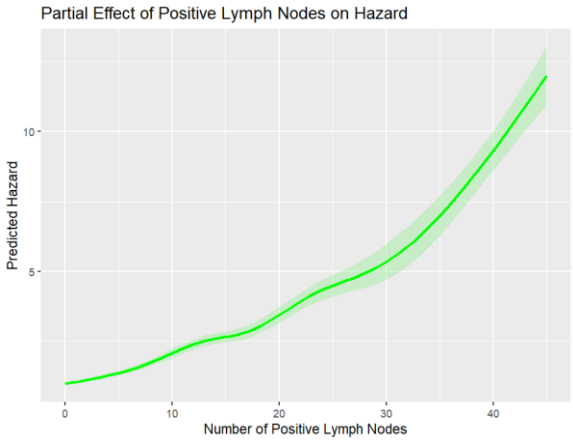




2.2 Stratified Cox Model (stratified by ER Status)

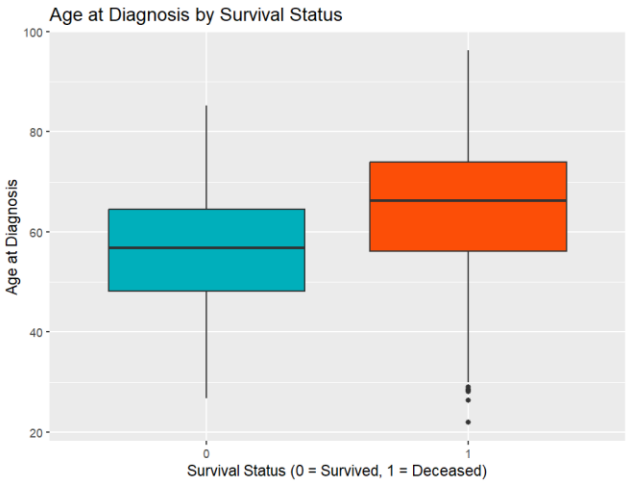
- Similar effects to the original model, but slightly better concordance (0.66)
- Age at Diagnosis: HR = 1.040
- Lymph nodes examined positive: HR = 1.061
- Neoplasm Histologic Grade 3: HR = 1.463

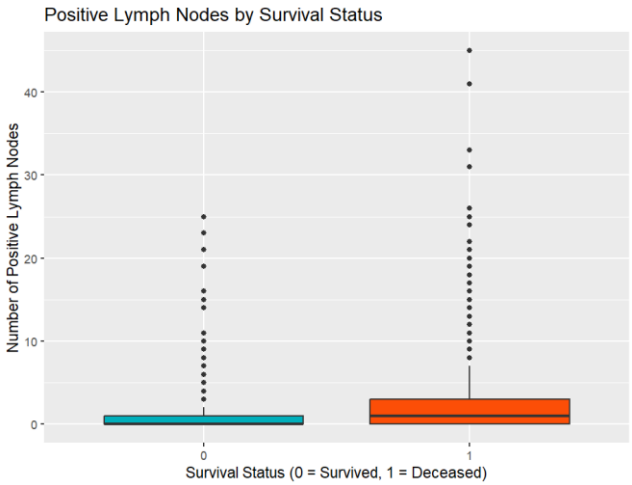




2.3 Time-dependent Cox Model

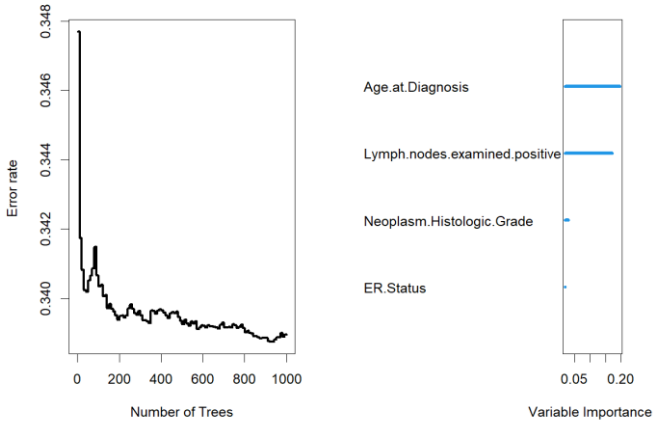
- Shows similar main effects but also indicates that the effect of age slightly increases over time, while the effect of positive lymph nodes slightly decreases.
- Concordance = 0.658





2.4 Random Survival Forest

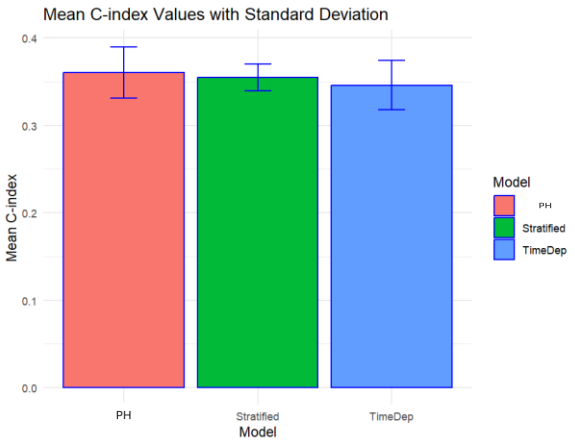
- Age at diagnosis and number of positive lymph nodes found, again shown to be key variables
- Variable importance: Age at Diagnosis > Lymph nodes examined positive > Neoplasm Histologic Grade > ER Status

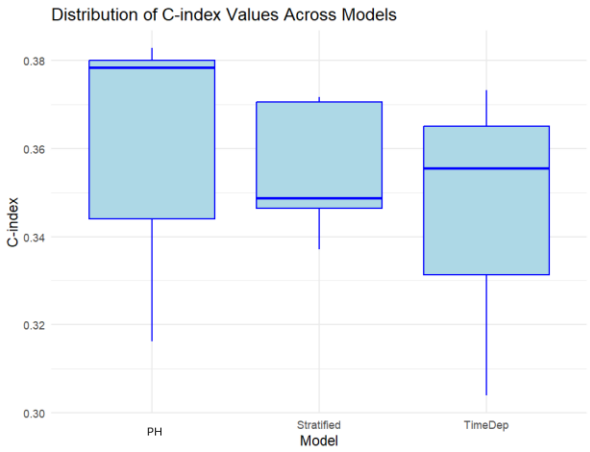


Part 3: Model Comparison and Validation

3.1 AIC and Cross-validation comparison

- Cox Stratified model has the lowest AIC (12898.40), suggesting it's the best fit among the Cox models (PH, 14035.75, Time Dependant, 14000.00)
- All models show poor performance in cross-validation, with C-index values around 0.35-0.36
- The Cox PH model has slightly higher mean C-index (0.3603) compared to the stratified (0.3549) and time-dependent (0.3458) models



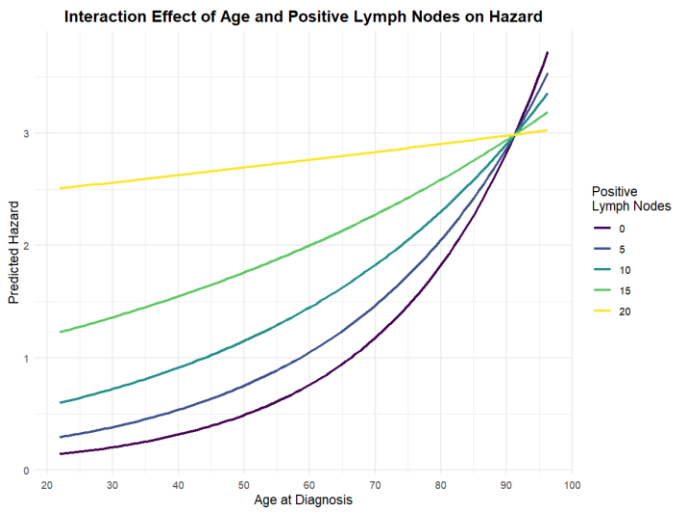


3.2 Final Model Interpretation (Stratified Cox Model)

- The model stratifies by ER Status, allowing for different baseline hazards for ER-positive and ER-negative patients, (3.2:1, for all patients with cancer in study)
- Age at Diagnosis: For each year increase, the hazard increases by 4% (HR = 1.040)
- Neoplasm Histologic Grade 3: 46.3% higher hazard compared to Grade 1 (HR = 1.463)
- Lymph nodes examined positive: Each additional positive node increases hazard by 6.1% (HR = 1.061)

Part 4: Key Findings

- Age, tumour grade, and number of positive lymph nodes are consistently important predictors of survival across all models
- ER-positive status is associated with better survival outcomes
- The effect of positive lymph nodes on survival risk decreases slightly with age
- The stratified Cox model, which accounts for different baseline hazards by ER status, provides the best fit according to AIC



- The proportional hazards assumption is violated for all variables in the original Cox model, justifying the use of stratified and time-dependent models
- The low C-index values in cross-validation (around 0.35) are concerning and suggest that the models may not generalize well

Conclusion

The models provide insights into important prognostic factors for breast cancer survival consist with the limited literature review made by the team. The poor cross-validation performance suggests that caution should be exercised in applying these models to new data. Further refinement and validation of the models would be necessary before they could be considered reliable.

References

1. Yazdani A, Haghighat S. Determining Prognostic Factors of Disease-Free Survival in Breast Cancer Using Censored Quantile Regression. Breast Cancer: Basic and Clinical Research. 2022;16. doi:10.1177/11782234221108058
2. Tan KF, Adam F, Hussin H, Mohd Mujar NM. A comparison of breast cancer survival across different age groups: a multicentric database study in Penang, Malaysia. Epidemiol Health. 2021;43:e2021038. doi: 10.4178/epih.e2021038. Epub 2021 May 25. PMID: 34044478; PMCID: PMC8510833.

Data Source

https://www.kaggle.com/datasets/gunesevitan/breast-cancer-metabric

Code Location

https://github.com/Brandt-DSTI/Breast_Cancer_Survival-Analysis