# Effect of Prognostic Factors on Recurrence-Free Survival in Node-Positive Breast Cancer Patients: A Survival Analysis using Multiple Imputation

StduentNo : 2332635

April 26, 2023

**Abstract**

The research aims to investigate the prognostic factors influencing recurrence-free survival in breast cancer patients. To develop a prognostic time-to-event model to assess the impact of various clinical and pathological factors on recurrence-free survival in patients with primary node-positive breast cancer. To identify significant prognostic factors and to assess the effects of hormonal therapy. The research will employ multiple imputations to handle missing data and backward elimination to identify significant prognostic factors. A Cox proportional hazard model will be fitted on the imputed dataset, adjusting for significant prognostic factors. Hormonal therapy showed a significant positive effect on recurrence-free survival time, with a 36.5% lower risk of recurrence than those who did not undergo the treatment. Other significant prognostic factors include progesterone receptor, tumour size, tumour grade and transformation of number of positive nodes. This research will contribute to the existing knowledge on breast cancer prognosis and recurrence-free survival. The findings could be used to make clinical decisions and in developing personalised treatment plans for breast cancer patients. Additionally, the importance of using appropriate statistical techniques and addressing missing data in survival analysis is highlighted in the research. This will encourage further investigations of alternate time-to-event models for similar studies.

## 1 Introduction

Breast cancers is classified as node-positive when the cancer cells from the tumour in the breast are found in the lymph nodes in the armpit area, it is called Node-positive breast cancer. Although surgery is effective in removing the tumour containing the cancer cells, its presence in the lymph nodes imposes a higher chance of recurrence and spreading [5].

The recurrence of breast cancer is a huge issue, despite advancements in treatment options and early detection methods. Understanding the factors that influence recurrence-free survival is crucial for improving treatment outcomes [1]. This study aims to investigate the impact of various prognostic factors on recurrence-free survival in patients with primary node-positive breast cancer, with a focus on the role of hormonal therapy. The study also aims to investigate the significant factors in breast cancer recurrence rates. Data collected from the German Breast Cancer Study Group between July 1984 and December 1989 for a Comprehensive Cohort Study is used in this research. In the original study, 720 participants were randomised into groups of different numbers of cycles of chemotherapy [2]. The primary eligibility criterion was primary node-positive breast cancer. This study uses a subset of data from the original study group. Prognostic factors included in the study are age, tumour size, tumour grade, number of positive lymph nodes, progesterone and oestrogen receptor status and menopausal status. This research employed a Cox proportional hazard model [citation]to investigate and assess the effects of various prognostic factors on recurrence-free survival in primary node-positive breast cancer patients. On initial observations of the dataset, missingness in the dataset is identified. Multiple imputations by chained equations (MICE) are employed to account for this missingness [3][4]. Due to the presence of missingness in the event time, a cumulative baseline hazard function of time, Nelson and Alan's estimate will be included in the imputation model to account for this uncertainty [6]. Each imputed dataset is fitted with Cox proportional hazard model and the results are pooled according to the Rubins rule [3] to obtain the final estimates. The fitted model can be further used for the prognosis of recurrence-free survival times in breast cancer patients. Additionally, this research will provide valuable insights into the influence of hormonal therapy on recurrence-free survival in breast cancer patients. The results of this study will give clinicians and researchers valuable insights that would help them in developing personalised treatment plans and conducting further research, improving patient outcomes in breast cancer patients.

# 2 Methods

## 2.1 Overview and Data

The dataset consists of 686 participants with data on recurrence-free survival and prognostic factors including age, indicator of hormonal therapy, tumour size, number of positive lymph nodes, progesterone and estrogen receptor status, menopausal status, and tumour grade. Indicator of hormonal therapy, Menopausal status, Tumour grade, Indicators of tumour grade are categorical variables and the remaining variables are continuous. The event indicator and event time are represented by censrec and rectime in weeks and recyear in years. All the other variables represent the prognostic factors.

The analysis is done in R studio using R programming language. The required R packages include readstata13, survival, survminer, naniar, tidyverse, vim, haven, tinytex lmtest, mice, MASS, dplyr, and ggplot2, are installed and loaded for analysis, imputation, variable selection, model building and visualisation of the data at different points of the research. The dataset available in the file assessment.rds was read into R using readRDS() function.

## 2.2 Data Pre-processing

The first step was to explore the data stored in a variable, "dat" using head() and summary() functions, to view the first five data rows and to summarise each variable in the dataset in detail. The variables were then formatted to represent the appropriate type (numeric and factor) using transmute() function. The transmute() function transformed the data as well as selected the required variables for further processing. At this stage, all the variables in the initial dataset are selected.

## 2.3 Handling missing data

The data is further explored to assess missingness by counting the number of completed cases and a number of incomplete cases. The patterns of missingness in the data is explored using vis$_{m}iss()andmiss_{v}ar_{s}ummary()func$

An approximate compatible approach was used to address missingness in the data as the Substantive model compatible fully conditional specification (SMCFCS) approach is computationally intensive and the risk of bias is high if the model assumptions are not met. It uses multiple imputations by chained equations (MICE). This method created multiply imputed datasets, each with plausible values estimated from observed data. The imputation methods used were predictive mean matching (PMM) for continuous variables and logistic regression for binary variables.

### 2.3.1 Missing data mechanism

The association between missing data and observed data was explored using a dummy array with 0's changed to 1's for rows without any missingness. This dummy array is then added to the dataset and ran a logistic regression, keeping it as the outcome variable and all the other variables except the event and time indicators, as the explanatory variables.
**Missing Completely At Random (MCAR)**: None of the variables shows statistical significance with missingness.
**Missing At Random (MAR)**: There is some association between missingness in the data with the observed outcome.
**Missing Not At Random (MNAR)**: The association between missingness in the data with the observed outcome can be explained.

### 2.3.2 MICE Algorithm

Multiple imputations by Chained Equations (MICE) perform m, number of imputations on the dataset. The mice algorithm predicts one variable missing data by using a model where the missing variable is treated as the outcome and all other variables as predictors. Based on this prediction, the algorithm moves on to the next variable and predicts its missing values based on all the others. These regression equations are strung together like a chain.

By default, the value of m is set to 5. To get a less biased, plausible value for missing data, the number of imputations needs to be consistent with the proportion of missingness. The value for m should be at least equal to the percentage of incomplete cases. The proportion of cases with any missingness is calculated using ici() from the mice package. The mean of this value is multiplied by 100 and rounded off to get the value of m.

The regression method used for each variable depends on the type of that variable. The predictor matrix value is set to 0 for all the variables with complete values. With time-to-event data, the outcome is a pair of two variables, event indicator, censrec and event time, recyear. As there is missingness in both these variables, both are added to the imputation model. The baseline hazard is the hazard function when all predictors are at zero or their reference level is calculated using the Nelson-Aalen estimate of the cumulative hazard function and added to the dataset instead of event time. After the imputation, the event time is put back in the dataset and the missing values in the variable recyear is filled with the corresponding imputed values of the cumhzd variable.

## 2.4 Covariates selection

Variable selection is done after multiple imputation because it takes into account the uncertainty associated with missing data and is better to assess the stability of the selected variables across the imputed datasets compared to when covariate selection is done before that.

For each imputed dataset, variable selection is performed separately with backward elimination approach using the Akaike Information Criterion (AIC), to identify the most significant predictors for survival. The selected variables are compared across imputed datasets and the consistently selected covariates (selected more than m/2 times) saved in a variable, selected$_covariatesareusedforfurtheranalysis.$

Each of the imputed datasets is iteratively completed using complete() from mice package and fitted on a Cox model, with censrec and recyear as the event indicator and event time and all the other variables as the covariates. The backward elimination procedure starts with fitting the full Cox proportional hazards model, including all available covariates, and iteratively removed the least significant variable until the AIC could not be reduced further. The stepAIC() function from the MASS package is used for this with a backward elimination option to identify the most significant covariates in each imputed dataset.

## 2.5 Survival Analysis Model

The Survival Analysis on the time-to-event data is done using Cox proportional hazards model, adjusted for the selected variables for each of the 4 models. For each imputed dataset, cox model was fitted and predictions were made in the form of survival probabilities. These predictions were then combined across the imputed datasets by averaging the survival probabilities. This is done using pool() from MICE package and summary(). The estimates obtained on log-odds ratio scale was converted to odds ratio scale using exp() function. The cox model is fitted using coxph() with() functions.

The final model, model 4 was fitted after further refining, based on results from model 2 (with all covariates) and model 3 (with selected covariates). In models 2 and 3, the same set of covariates showed statistical significance. Thus, model 4 is fitted on those significant covariates.

## 2.6 Model diagnostics and validation

The validity of the fitted cox model was ensured by checking the proportional hazards assumption using the Schoenfeld residuals test. Each imputed dataset fitted on the coz model is checked for proportional hazards assumption using cox.zph() function.

### 2.7 Prediction and Interpretation

The survival probabilities for each participant in the dataset is predicted after fitting the survival analysis model. The median survival time was calculated for each participant based on the average survival probability. This provides a measure of the central tendency of survival time.

The results of the survival analysis were interpreted in terms of hazard ratios for the significant predictors. This interpretation gives an estimate of the relative risk associated with each significant predictor. The survival probabilities and median survival time were used to draw conclusions to be used in clinical decision-making and improving patient outcomes.

## 3 Results

### 3.1 Variables in the Dataset

All the variables in the datset is transformed to apropriate type as shown in the table.

Table 1: Description of Variables in the dataset

| Variable | Description | type |
|---|---|---|
| id | ID of study participants | double |
| hormon | Indicator of hormonal therapy (0 no, 1 yes) | factor |
| age | Age in years | double |
| menostatus | Menopausal status (1 premenopausal, 2 postmenopausal) | factor |
| tsize | Tumour size in mm | double |
| tgrade | Tumour grade ($1 > 2 > 3$) | factor |
| posnodes | The number of positive lymph nodes | double |
| progrec | Progesterone receptor, fmol | double |
| estrec | Estrogen receptor, fmol | double |
| rectime | Recurrence free survival in days | double |
| recyear | Recurrence free survival in years | double |
| censrec | Censoring indicator (0 censored, 1 event) | double |
| x4a | Indicator of tumour grade $\geq 2$ | factor |
| x4b | Indicator of tumour grade $= 3$ | factor |
| x5e | Transformation of number of positive nodes, $\exp(0.12 \times \text{posnodes})$ | double |

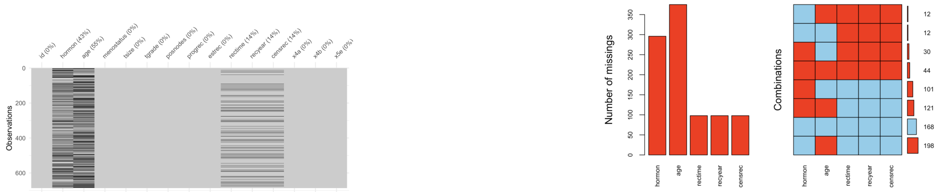
### 3.2 Data Patterns and Missingness

The dataset is found to contain missingness in covariates representing age, hormonal therapy status, event time variables such as recurrence-free survival in weeks and years and the event indicator, the status of being censored. Out of the total 686 observations, more than 75% of observations have missingness in at least one of the variables. This is shown in Table 2 :

Table 2: Number and percentage of missing values per Variable

| Variable | Number of Missing Values | Percentage of Missing Values |
|---|---|---|
| age | 375 | 54.66 |
| hormon | 296 | 43.15 |
| rectime | 98 | 14.29 |
| recyear | 98 | 14.29 |
| censrec | 98 | 14.29 |
| id | 0 | 0.00 |
| menostatus | 0 | 0.00 |
| tsize | 0 | 0.00 |
| tgrade | 0 | 0.00 |
| posnodes | 0 | 0.00 |
| progrec | 0 | 0.00 |
| estrec | 0 | 0.00 |
| x4a | 0 | 0.00 |
| x4b | 0 | 0.00 |
| x5e | 0 | 0.00 |

It is observed from 1, that recurrence-free time in days and weeks and indicator for censoring have missingness for the same participants. The following plots gives the patterns of missingness in the dataset :



(a) Pattern of missingness in the dataset

(b) Missingness per covariate and accross covariates

Figure 1: Missing data patterns

### 3.3 Missing Data Mechanism

Association between missing data and observed data is analysed and tumour size is found to have strong associations with missingness in the dataset. This invalidates Missing completely at random (MCAR) mechanism. Thus, the missing data mechanism in the dataset will either be Missing at random(MAR) or Missing not at random (MNAR). In the following table, the p-value of tumour size is 0.0024 is statistically significant and explains this association.

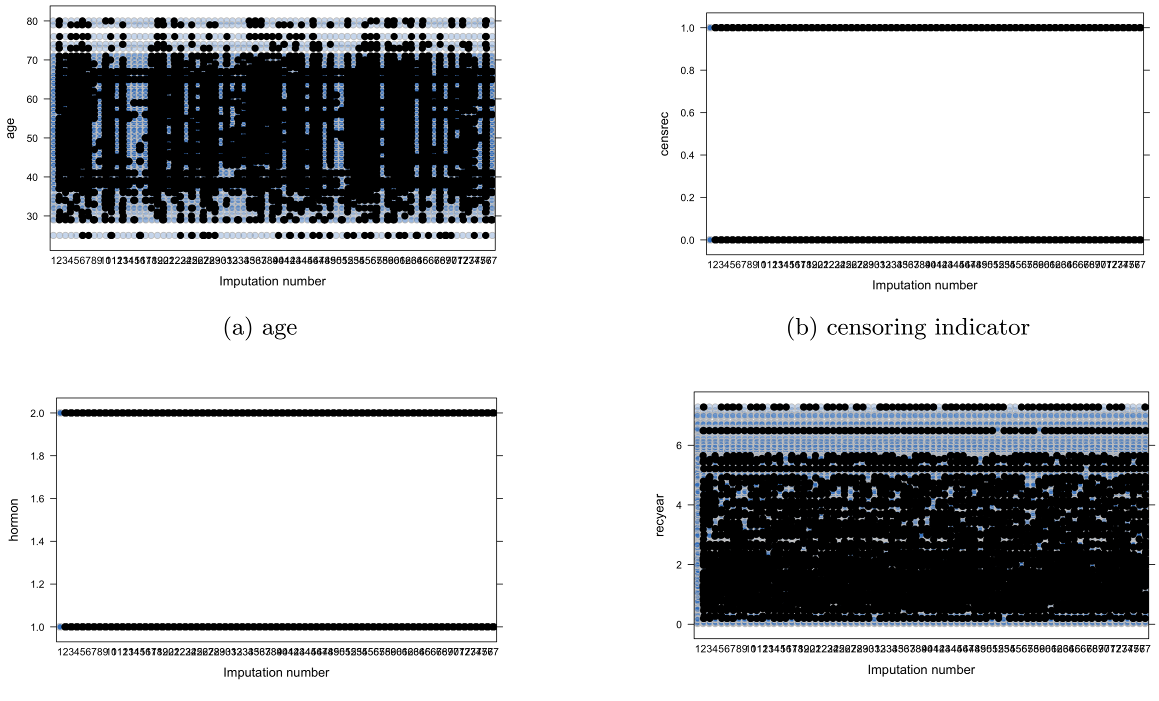
Table 3: Association between observed and missing data

| Term | Estimate | Std. Error | p-value |
|---|---|---|---|
| (Intercept) | 2.469e+01 | 2.614e+03 | 0.9925 |
| hormon1 | -2.650e-01 | 8.623e-01 | 0.7586 |
| age | 9.692e-02 | 7.482e-02 | 0.1952 |
| menostatus2 | -2.159e+01 | 2.614e+03 | 0.9934 |
| tsize | -2.346e-01 | 7.728e-02 | 0.0024** |
| tgrade2 | 3.768e-01 | 9.844e-01 | 0.7019 |
| tgrade3 | 9.743e-01 | 1.635e+00 | 0.5513 |
| posnodes | -3.878e-02 | 2.611e-01 | 0.8820 |
| progrec | 1.772e-04 | 1.391e-03 | 0.8986 |
| estrec | -2.064e-03 | 1.961e-03 | 0.2925 |
| x4a1 | NA | NA | NA |
| x4b1 | NA | NA | NA |
| x5e | -1.071e+00 | 5.049e+00 | 0.8321 |

[maybe add the table created with the huge code]

### 3.4 Multiple Imputation Results

The number of imputations to be performed should be atleast equal to the percentage of incomplete cases. Based on this, the number of imputations is calculated as 76. The plots given below explains the imputations. Figure 2 shows strip plots for all imputed variables in the dataset. When the number of missing values is large, stirpplot may not be very informative as the imputed values are superimposed over the observed values and we use box plots to explain the imputations as shown in Figure 3.



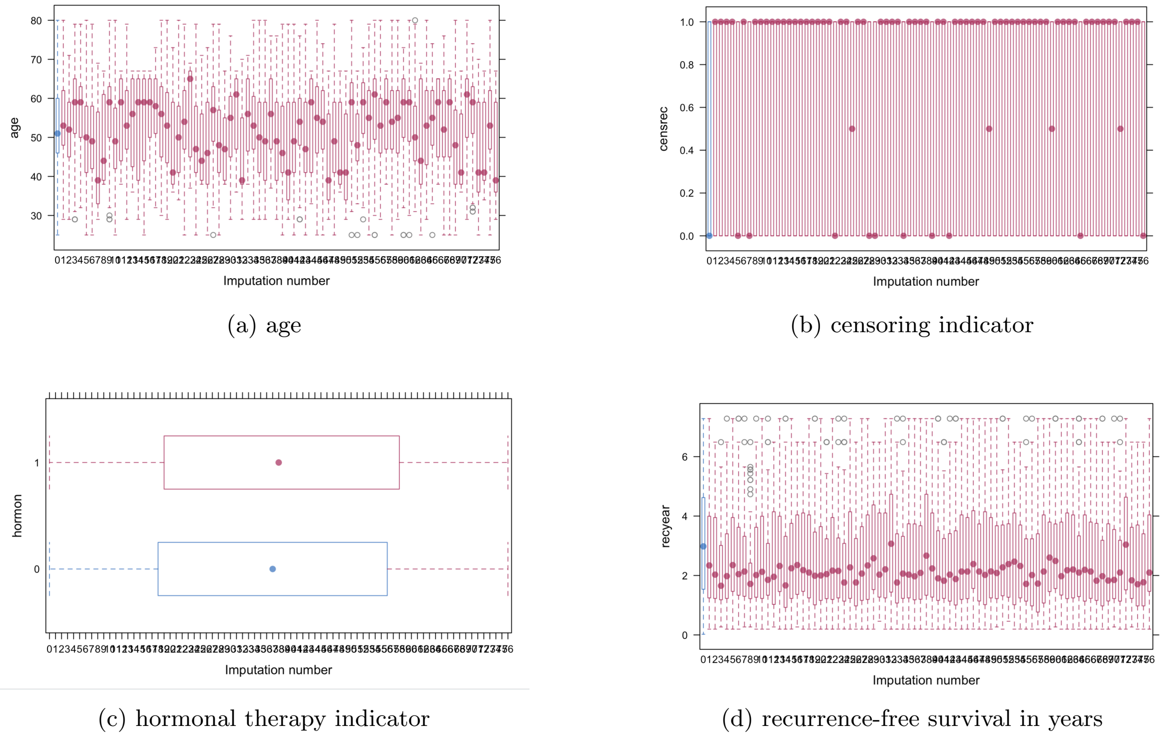
(a) age

(b) censoring indicator

(c) hormonal therapy indicator

(d) recurrence-free survival in years

Figure 2: Strip plot for imputed vs observed values



(a) age

(b) censoring indicator

(c) hormonal therapy indicator

(d) recurrence-free survival in years

Figure 3: Box plot for imputed vs observed values

## 3.5 Model 1

The first model, model 1 was fitted on all the covariates in the dataset, with missing values. The association between survival outcome(recyear), accounting for censoring (censrec) in the presence of missing data is presented in the results. Patients who underwent hormonal therapy is found to have 51.3% lower risk of breast cancer recurrence compared to those who did not underwent hormonal therapy. The hazard ratio is given by 0.49, with a 95% confidence interval of 0.26 and 0.896. The p-value of 0.0206 indicates a significant association between hormonal therapy and recurrence-free survival time.

Age, Menopausal status, tumour size, and transformation of number of positive nodes are seen to have significant associations with recurrence-free survival, while tumour grade, number of positive lymph nodes, estrogen and progesterone receptors shows no significant association. The goodness of fit for the model was found to be 40.82 for Likelihood ratio test, 36.55 for Wald test and 41.9 for logrank test on 10 degrees of freedom.

The results for model 1 is explained in table 1 :

Table 4: Regression Results for Model 1

| Term | Estimate | Std. Error | p-value | 2.5% | 97.5% |
|---|---|---|---|---|---|
| hormon1 | 0.4869 | 0.3109 | 0.0206 | 0.2648 | 0.8955 |
| age | 0.9532 | 0.0221 | 0.0302 | 0.9128 | 0.9954 |
| menostatus2 | 2.7596 | 0.4578 | 0.0266 | 1.1251 | 6.7686 |
| tsize | 1.0427 | 0.0204 | 0.0403 | 1.0019 | 1.0852 |
| tgrade2 | 1.6105 | 0.4860 | 0.3268 | 0.6213 | 4.1747 |
| tgrade3 | 2.8061 | 0.6114 | 0.0915 | 0.8467 | 9.3004 |
| posnodes | 0.9534 | 0.0406 | 0.2403 | 0.8804 | 1.0324 |
| progrec | 0.9980 | 0.0014 | 0.1479 | 0.9953 | 1.0007 |
| estrec | 1.0011 | 0.0011 | 0.3292 | 0.9989 | 1.0033 |
| x5e | 0.0376 | 1.1771 | 0.0053 | 0.0037 | 0.3774 |

### 3.6 Model 2

Model 2 was fitted on all the covariates in the imputed dataset. The uncertanities caused by missingness in the data is accounted for by fitting each imputed dataset on the cox model and the pooling the result. Patients who underwent hormonal therapy is found to have 40.8% lower risk of breast cancer recurrence compared to those who did not underwent hormonal therapy. The hazard ratio is given by 0.592, with a 95% confidence interval of 0.415 and 0.844. The p-value of 0.004 indicates a significant association between hormonal therapy and recurrence-free survival time.

Tumour grade, tumour size, progesterone receptors and transformation of number of positive nodes are seen to have significant associations with recurrence-free survival, while age, Menopausal status, number of positive lymph nodes, and estrogen receptors shows no significant association.

The results for model 2 is explained in table 2 :

Table 5: Regression Results Model 2

| Term | Estimate | Std. Error | p-value | 2.5% | 97.5% |
|---|---|---|---|---|---|
| hormon1 | 0.5923 | 0.1786 | 0.0042 | 0.4156 | 0.8441 |
| age | 0.9955 | 0.0119 | 0.7065 | 0.9722 | 1.0193 |
| menostatus2 | 1.3249 | 0.2034 | 0.1691 | 0.8859 | 1.9814 |
| tsize | 1.0135 | 0.0055 | 0.0162 | 1.0025 | 1.0247 |
| tgrade2 | 1.8686 | 0.2814 | 0.0274 | 1.0730 | 3.2538 |
| tgrade3 | 2.3934 | 0.3169 | 0.0064 | 1.2812 | 4.4710 |
| posnodes | 0.9689 | 0.0291 | 0.2795 | 0.9147 | 1.0264 |
| progrec | 0.9981 | 0.0006 | 0.0023 | 0.9970 | 0.9993 |
| estrec | 1.0001 | 0.0006 | 0.8839 | 0.9989 | 1.0012 |
| x5e | 0.0793 | 0.6489 | 0.0001 | 0.0220 | 0.2860 |

### 3.7 Model 3

Model 3 was fitted on the selected covariates menopausal status, number of positive lymph nodes, indicator of hormonal therapy, tumour size, tumour grade, progesterone receptors and Transformation of number of positive nodes. Patients who underwent hormonal therapy is found to have 40.5% lower risk of breast cancer recurrence compared to those who did not underwent hormonal therapy. The hazard ratio is given by 0.595, with a 95% confidence interval of 0.418 and 0.846. The p-value of 0.004 indicates a significant association between hormonal therapy and recurrence-free survival time.

Tumour grade, tumour size, progesterone receptors and transformation of number of positive nodes are seen to have significant associations with recurrence-free survival, while Menopausal status, and number of positive lymph nodes shows no significant association.

The results for model 3 is explained in table 3 :

Table 6: Regression Results for Model 3

| Term | Estimate | Std. Error | p-value | 2.5% | 97.5% |
|---|---|---|---|---|---|
| hormon1 | 0.5948 | 0.1777 | 0.0043 | 0.4182 | 0.8461 |
| menostatus2 | 1.2480 | 0.1362 | 0.1053 | 0.9541 | 1.6324 |
| tsize | 1.0134 | 0.0055 | 0.0177 | 1.0024 | 1.0245 |
| tgrade2 | 1.8894 | 0.2786 | 0.0233 | 1.0912 | 3.2716 |
| tgrade3 | 2.3656 | 0.2986 | 0.0043 | 1.3133 | 4.2610 |
| posnodes | 0.9687 | 0.0288 | 0.2716 | 0.9149 | 1.0256 |
| progrec | 0.9982 | 0.0006 | 0.0016 | 0.9970 | 0.9993 |
| x5e | 0.0796 | 0.6381 | 0.0001 | 0.0225 | 0.2810 |

### 3.8 Model 4

The final model, model 4 was fitted after further refining, based on results from model 2 (with all covariates) and model 3 (with selected covariates). Thus, model 4 is fitted on the significant covariates tumour size, tumour grade, progesterone receptors and a transformation of number of positive nodes. Patients who underwent hormonal therapy is found to have 36.6% lower risk of breast cancer recurrence compared to those who did not underwent hormonal therapy. The hazard ratio is given by 0.634, with a 95% confidence interval of 0.45 and 0.893. The p-value of 0.0097 indicates the significant association between hormonal therapy and recurrence-free survival time.

All the covariates, tumour grade, tumour size, progesterone receptors and transformation of number of positive nodes are seen to have significant associations with recurrence-free survival.

The results for model 4 is explained in table 4 :

Table 7: Regression Results for Model 4

| Term | Estimate | Std. Error | p-value | 2.5% | 97.5% |
|---|---|---|---|---|---|
| hormon1 | 0.6345 | 0.1727 | 0.0097 | 0.4504 | 0.8937 |
| tsize | 1.0118 | 0.0053 | 0.0298 | 1.0012 | 1.0225 |
| tgrade2 | 1.9129 | 0.2778 | 0.0204 | 1.1065 | 3.3068 |
| tgrade3 | 2.3518 | 0.2974 | 0.0044 | 1.3090 | 4.2254 |
| progrec | 0.9981 | 0.0006 | 0.0017 | 0.9970 | 0.9993 |
| x5e | 0.1571 | 0.2739 | 1.7463e-10 | 0.0915 | 0.2696 |

The estimates for x4a1 and x4b1 were not available in any of the models. This is likely due to lack of variation in the data for these variables or multi-collinearity. The goodness of fit for the models 2, 3 and 4, done on imputed data is not significant as there is no single model fit to assess these tests.

The cox models fitted on imputed datasets accounting for refined and selected covariates are checked against the proportionality hazards assumption. It is found that nearly 60% of the models lend support to cox proportional hazards assumption. The remaining 31 models contain at least one variable that violates cox proportional hazards assumption. In all the 76 imputed datasets, hormonal therapy is seen to follow the PH assumption.

The fitted coz model on each imputed dataset is used for prediction of recurrence-free survival time in years. Figure 4 shows the average predicted survival probability.

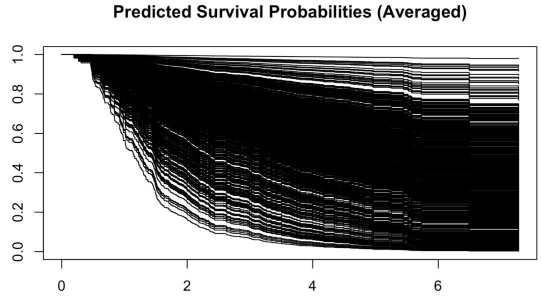


Figure 4: Average predicted survival probabilities

# 4 Discussion

The choice of the appropriate survival analysis method depends on the research question, the nature of the data, and the underlying assumptions. It was not possible to assume a distribution for this data as the dataset contained missingness in survival time. Even though multiple imputations was employed, it was not possible to predict the survival time accurately as there was further missingness in the event indicator and other covariates in the model. Considering all of these issues, it was decided to choose Cox proportional hazards model for this research. Few points of comparison with Weibull, Exponential, and Kaplan-Meier models for survival analysis:

- The Cox model does not require any assumptions for the survival time, whereas Weibull and exponential models assume parametric distribution.
- The Cox model can handle multiple covariates, whereas the Kaplan-Meier method is a univariate technique and does not account for covariates.
- The Cox model assumes that the hazard ratios are proportional over time. In practice, this assumption is reasonable and easier to interpret than more complex methods with time-varying effects.
- The Cox model is a semi-parametric model, as it uses partial likelihood estimation to estimate the regression coefficient.
- The Cox model is popular and can handle complex data structures, there is a wealth of resources and software to support its implementation

The current study aimed to investigate the association between various clinical and pathological factors on recurrence-free survival in patients with node-positive breast cancer. This provided valuable insights for clinicians and researchers, aiding them in developing personalised treatment plans and conducting further research.

Multiple Imputation was employed to address the issue of missingness in data, resulting in a more accurate and unbiased estimate of the relationship between the prognostic factors and patient outcomes. The impact of undergoing hormonal therapy on recurrence-free survival is found to be significant. This finding is consistent with previous research [7][8].

It is essential to acknowledge the limitations of the current study. The non-random nature of missingness in observational studies will result in the findings being subjected to some degrees of bias, despite employing multiple imputations to handle missing data. Additionally, the sample size and the retrospective nature of the study may limit the generalisability of the results. Further research can be conducted on larger, prospective cohorts to validate and expand upon the findings from this study. Further investigations of alternate time-to-event models for similar studies can be conducted.

## 5 References

[1] Mengjuan Wu, Ting Zhao, Qian Zhang, Tao Zhang, Lei Wang, and Gang Sun. (2022). Prognostic analysis of breast cancer in Xinjiang based on Cox proportional hazards model and twostep cluster method, 12, https://doi.org/10.3389/fonc.2022.1044945

[2] W. Sauerbrei, P. Royston. (1999). Building Multivariable Prognostic and Diagnostic Models: Transformation of the Predictors by Using Fractional Polynomials, 162(1): 71-94, https://www.jstor.org/stable/2680468

[3] Jonathan A C Sterne, Ian R White, John B Carlin. (2009). Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls, https://doi.org/10.1136/bmj.b2393

[4] M R Baneshi and A R Talei. (2011). Multiple Imputation in Survival Models: Applied on Breast Cancer Data, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3371994/

[5] Victoria Sopik and Steven A. Narod. (2018). The relationship between tumour size, nodal status and distant metastases: on the origins of breast cancer, 170(3): 647–656, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6022519/

[6] Saranya1 and Karthikeyan. (2015). A Comparison study of Kaplan Meier and NelsonAalen Methods in Survival Analysis, 2(11).

[7] C. Rauh, F. Schuetz, B. Rack, E. Stickeler. (2015). Hormone Therapy and its Effect on the Prognosis in Breast Cancer Patients, 75(6): 588–596, https://doi.org/10.1055/s-0035-1546149

[8] Akram Yazdani and Shahpar Haghighat. (2022). Determining Prognostic Factors of Disease-Free Survival in Breast Cancer Using Censored Quantile Regression, 16, https://doi.org/10.1177/11782234221108058

# 6 Appendix

~~~
---
title: "R code"
author: 'StduentNo : 2332635'
date: "2023-04-18"
---

```{r , include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

```{r}
rm(list=ls())

Sys.Date()
dir <- getwd()
setwd(dir)
```

```{r setup, include=FALSE}
# Required packages and libraries
install.packages('tinytex')
install.packages("knitr")
install.packages("naniar")


#tinytex::tlmgr_install('multirow')
#tinytex::reinstall_tinytex(repository = "illinois")

```

```{r}
library(tinytex)
library(knitr)
library(kableExtra)
library(haven)
library(VIM)
library(gtsummary)
library(tidyverse)
library(naniar)
library(dplyr)
library(mice)
library(survival)
library(lmtest)
library(survminer)
```



```{r}
# read data
dat <- readRDS("assessment.rds")

# inspect the data
head(dat)
summary(dat)
```

```{r}
# pre-processing data
~~~

````

dat0 <- dat %>% transmute(
  id = as.numeric(id),
  hormon = as.factor(hormon),
  age = as.numeric(age),
  menostatus = as.factor(menostatus),
  tsize = as.numeric(tsize),
  tgrade = as.factor(tgrade),
  posnodes = as.numeric(posnodes),
  progrec = as.numeric(progrec),
  estrec = as.numeric(estrec),
  rectime = as.numeric(rectime),
  recyear = as.numeric(recyear),
  censrec = as.numeric(censrec),
  x4a = as.factor(x4a),
  x4b = as.factor(x4b),
  x5e = as.numeric(x5e)
)

head(dat0)

class(dat0$hormon)
‘‘‘

‘‘‘{r}
# investigating the presence of missing data
missing_dat0 <- sapply(dat0, function(x) sum(is.na(x)))
missing_dat0




# missingness
dat0 %>%
tbl_summary(missing = "ifany", missing_text = "Missing")


‘‘‘

‘‘‘{r}
# explore patterns of missingness in the data

# examine missingness at induvidal participant level
vis_miss(dat0)
miss_var_summary(dat0)

# examine missingness patterns per covariates and accross covariates with missing values
# excluded unique identifier, id
missplot_all <- aggr(
dat0[, c(
  "hormon",
  "age",
  "rectime",
  "recyear",
  "censrec"
)],
prop = FALSE, numbers = TRUE, sortCombs = TRUE,
cex.axis = 0.75, cex.numbers = 0.75
)

‘‘‘
````

~~~~

```{r}
# further exploration of missingness

# no of complete cases
misscount <- numeric(nrow(dat0))
for (i in 1:nrow(dat0)) {
misscount[i] <- countNA(dat0[i, c(
  "id",
  "hormon",
  "age",
  "menostatus",
  "tsize",
  "tgrade",
  "posnodes",
  "progrec",
  "estrec",
  "rectime",
  "recyear",
  "censrec",
  "x4a",
  "x4b",
  "x5e"
)])
}

table(misscount)
```

# no of complete cases : 168

```{r}
round(table(misscount) / sum(table(misscount)) * 100, 2)
```

```{r}
# checking associations between missing data and observed data

indic_comp <- rep(0, nrow(dat0))
indic_comp[which(misscount == 0)] <- 1
dat0$indic_comp <- indic_comp
assoc_comp <- glm(
indic_comp ~ hormon + age + menostatus + tsize + tgrade + posnodes + progrec + estrec  + x4a + x4b + x5e,
data = dat0, family = binomial)

summary(assoc_comp)

# do not include outcome variables, rectime, recyear, censrec for convergence

# result : tsize is find to be associated with the mnissingness. And give the statistical interpretation.

exp(coef(assoc_comp))
```


```{r}
# updated datset after dropping id and rectime

dat1 <- dat0 %>% dplyr::select(hormon, age, menostatus, tsize, tgrade, posnodes, progrec, estrec, x4a, x4b, x5e, recyear, censrec)
~~~~

~~~
head(dat1)

‘‘‘



‘‘‘{r}
# model 1
# cox model on the initial data(data with misingness) with all covariates
# censrec - event of interest
# recyear - time of event
model_cox <- coxph(Surv(recyear, censrec) ~ hormon + age + menostatus + tsize + tgrade + posnodes + progrec + estrec + x4a + x4b + x5e, data = dat1)

summary(model_cox, exponentiate = TRUE, conf.int = 0.95)
‘‘‘


‘‘‘{r}
# Multiple imputation

# cumulative baseline hazard - converting linear time to a function of time to go with substantive model assumptions

dat1$cumhzd <-
  nelsonaalen(dat1, recyear, censrec)

head(dat1)
# nelsonaalen is not dependent on cox hazard model


# plot for cumulative hazard function
plot(x = dat1$recyear, y = dat1$cumhzd, ylab = "Cumulative hazard", xlab = "Time")

‘‘‘

‘‘‘{r}
# imputing


# data for imputation
dat_imp_incomplete <- dat1 %>%
  dplyr::select(cumhzd, hormon, age, menostatus, tsize, tgrade, posnodes, progrec, estrec, censrec, x4a, x4b, x5e)

# conduct a dryun of mice with default settings
dryrun <- mice(dat_imp_incomplete[,c("cumhzd", "hormon", "age", "menostatus", "tsize", "tgrade", "posnodes", "progrec", "estrec", "censrec", "x4a", "x4b", "x5e")],
               maxit = 0, seed = 987)

dryrun
# explain why dryrun is conducted

‘‘‘

‘‘‘{r}
# change the predictor matrix
pred <- dryrun$pred

~~~

~~~

# set the rows of the fully observed variables to 0 - this to avoid predicting the already complete variales, thus to avoid unwanted processing time
pred["tsize",] <- 0
pred["menostatus",] <- 0
pred["tgrade",] <- 0
pred["posnodes",] <- 0
pred["progrec",] <- 0
pred["estrec",] <- 0
pred["x4a",] <- 0
pred["x4b",] <- 0
pred["x5e",] <- 0

pred
```

```{r}
# save method
method <- dryrun$method
method
```


```{r}
# calculate the number of imputations, m
# calculation for m

# Proportion of complete cases
mean(cci(dat1))
# Proportion of cases with any missing value
p <- mean(ici(dat1))

m <- round(100*p)
m

# Proportion of cases with missing values for each variable

# average over cases with missing data for each variable and then taking the highest average value available. Multiplying this with 100 and rounding off
P <- sapply(dat1, function(x) mean(is.na(x)))
P

m <- max(5, round(100*p))
m

```

```{r}
# impute the data
dat_imp <- mice(
  dat_imp_incomplete[, c("cumhzd", "hormon", "age", "menostatus", "tsize", "tgrade", "posnodes", "progrec", "estrec", "censrec", "x4a", "x4b", "x5e")],
  method = method,
  pred = pred,
  m = m,
  maxit = 10,
  seed = 987
)

```
~~~

~~~

```{r}
# Imputed datasets in long form
completedData <- complete(dat_imp, "long", include = TRUE)

# Replacing missing values in recyear with imputed values from cumhzd

# Repeat time variable m + 1 times
# includes the original data as well as m imputations
completedData$recyear <- rep(dat1$recyear, dat_imp$m + 1)


# Replace missing recyear values with corresponding imputed cumulative hazard value, cumhzd

#  .imp > 0 prevents replacing missing values in the original data

sub_data <- completedData$.imp > 0 & is.na(completedData$recyear)
if(sum(sub_data) > 0) {

  # Create a look-up table with the event times and corresponding cumulative hazards
  look_up <- data.frame(time   = dat1$recyear,
                        cumhzd = dat1$cumhzd)

  # Sort and remove duplicates
  look_up <- look_up[order(look_up$time),]
  look_up <- look_up[!duplicated(look_up) & !is.na(look_up$time),]

  for(i in 1:sum(sub_data)) {
    # Use max since last 2 times have the same cumhaz
    completedData$recyear[sub_data][i] <-
      max(look_up$time[look_up$cumhzd == completedData$cumhzd[sub_data][i]], na.rm = T)
  }
}

# Convert back to a mids object
completedData <- as.mids(completedData)


#completedData$imp[[1]]

```

# completedData contains all the completed data for m number of imputations, ie, m datasets

```{r}
# stripplot
# examine the imputed dataset using plots
# hormon - factor
stripplot(completedData, hormon ~ .imp,
          col = c("gray", "black"),
          pch = c(21, 20),
          cex = c(1, 1.5))
# age - continuous
stripplot(completedData, age ~ .imp,
          col = c("gray", "black"),
          pch = c(21, 20),
          cex = c(1, 1.5))

~~~

```
# recyear - continuous
stripplot(completedData, recyear ~ .imp,
          col = c("gray", "black"),
          pch = c(21, 20),
          cex = c(1, 1.5))
# censrec - continuous
stripplot(completedData, censrec ~ .imp,
          col = c("gray", "black"),
          pch = c(21, 20),
          cex = c(1, 1.5))
‘‘‘

‘‘‘{r}
# bwplot
# if the no of missing values is large, stirpplot may not be very informative as the imputed values are plotted on top of observed values.
# use bwplot()

# hormon - factor
bwplot(completedData, hormon ~ .imp)
# age - continuous
bwplot(completedData, age ~ .imp)
# recyear - continuous
bwplot(completedData, recyear ~ .imp)
# censrec - continuous
bwplot(completedData, censrec ~ .imp)


‘‘‘


‘‘‘{r}
# model 2
# make a model with all the covariates in the imputed data.
model_cox2 <- with(completedData,coxph(Surv(recyear, censrec) ~ hormon + age + menostatus + tsize + tgrade + posnodes + progrec + estrec + x4a + x4b + x5e))

summary(pool(model_cox2), exponentiate = TRUE, conf.int = 0.95)

‘‘‘

‘‘‘{r}
# variable selection using backward elimination

library(MASS)

# variable selection
selected_vars <- lapply(1:m, function(i) {
  dataset <- complete(completedData, i)

  # Full model with all covariates
  full_model <- coxph(Surv(recyear, censrec) ~ hormon + age + menostatus + tsize + tgrade + posnodes + progrec + estrec + x4a + x4b + x5e, data = dataset)

  # Backward elimination using AIC
  step_result <- stepAIC(full_model, direction = "backward", trace = FALSE)
  vars <- names(coef(step_result))

  return(vars)
})
```

~~~

# Count the frequency of each variable being selected
var_freq <- table(unlist(selected_vars))
var_freq

# the variables that were consistently selected
# adjusted this threshold to >= m/2

selected_covariates <- names(var_freq[var_freq >= (m / 2)])
selected_covariates



```

```{r}
# model 3
# cox model with selected covariates

model_cox_fit <- with(completedData,
                coxph(Surv(recyear, censrec) ~ hormon + menostatus + tsize + tgrade + posnodes + progrec + x5e))

# odds ratio scale - exponentiate
cox_model_pool <- summary(pool(model_cox_fit), exponentiate = TRUE, conf.int = 0.95)
cox_model_pool

# log-odds ratio scale
summary(pool(model_cox_fit))

```

```{r}
# model 4
# Further refining the model - After imputation, the same set of variables showed statistical significance for model 2 (with all covariates) and models 3 (with selected covariates), we decided to refine the model further with only the significant covariates.

model_cox4 <- with(completedData,
                coxph(Surv(recyear, censrec) ~ hormon + tsize + tgrade + progrec + x5e))

# odds ratio scale - exponentiate
summary(pool(model_cox4), exponentiate = TRUE, conf.int = 0.95)

```


# Model diagnostics and validation
```{r}
# include in methods and results

# proportional hazards assumption



# Function to perform cox.zph() on each imputed dataset
ph_test_each_imputed <- function(model_cox4) {
  ph_test <- cox.zph(model_cox4)

~~~

~~~~
  return(ph_test)
}

# Apply the function to the list of fitted cox models
ph_tests <- lapply(model_cox4$analyses, ph_test_each_imputed)

# function to check how many imputations have proportional hazards assumption true.
flag <- FALSE
a<-function(ph_tests){
  each <- ph_tests$table
  if(each[6,3]<0.05)
    flag <- TRUE
  return(flag)
}

signif_test <- lapply(ph_tests, a)

count_true <- sum(sapply(signif_test, function(x) sum(x == FALSE)))
count_true

# p-value for GLOBAL variable not < 0.05 implies, no statistical significance. Thus validating proportional hazard assumption.

# Here, only 45 of the 76 imputations shows support towards proportional hazards assumption.
# This is based on the GLOBAL value. A statistically significant global value indiactes that atlest one of the variables in the model violates proportional hazard assumption.
# In all the 76 imputations , hormon representing hormonal therapy is seen to follow PH assumption.
# Check the induvidal p-values for the varaiables to assess which all variables violate proportional hazard assumption.

# plot ph_tests
# ph_test pooled
ggcoxzph(ph_tests[[1]])

# Plot the Schoenfeld residuals for each imputed dataset
for (i in seq_along(ph_tests)) {
  plot_residuals(ph_tests[[i]], i)
}

# Model diagnostics and validation for cox model
model_cox4 %>% gtsummary::tbl_regression(exp = TRUE)

# percentage of models with valid PH assumption
(45/76)*100
76-45
(count_true/length(signif_test))*100

#almost 60% of the imputed models lend support to PH assumption
```



```{r}
# predictions using cox model

# Function to fit Cox model and make predictions
fit_and_predict <- function(data) {
  # Fit the Cox model using the selected variables
~~~~

~~~
  cox_model <- coxph(Surv(recyear, censrec) ~ hormon + tsize + tgrade + progrec + x5e, data = data)

  # Make predictions
  pred_surv <- survfit(cox_model, newdata = data)
  return(pred_surv)
}

# Fit the Cox model and make predictions for each imputed dataset
predictions_list <- lapply(1:m, function(i) {
  dataset <- complete(completedData, i)
  predictions <- fit_and_predict(dataset)
  return(predictions)
})


# Compute the average of the predicted survival probabilities
avg_predictions <- predictions_list[[1]]$surv
for (i in 2:m) {
  avg_predictions <- avg_predictions + predictions_list[[i]]$surv
}
avg_predictions <- avg_predictions / m

# Create a new survfit object to store the average predictions
avg_pred_survfit <- predictions_list[[1]]
avg_pred_survfit$surv <- avg_predictions

# Plot the average predictions
plot(avg_pred_survfit, main = "Predicted Survival Probabilities (Averaged)")
```
~~~