Integration of Genetic Biomarkers in Prognostic Models for Breast Cancer Survival

A Thesis submitted by

Deirdre Wall

Supervisor: Dr John Newell

School of Mathematics, Statistics and Applied Mathematics

National University of Ireland, Galway

September 2013

Contents

Contents

List of Figures

List of Figures

List of Figures

List of Figures

List of Tables

List of Tables

Abstract

The main aim of my PhD is to create a prognostic model for invasive breast cancer patients for disease recurrence and death. The data were collected retrospectively and are comprised of 647 invasive breast cancer patients with patient characteristics and genetic markers measured. An additional complexity exists due to the presence of missing data. A complete case analysis with both clinical and pathological biomarkers reduces the number of cases to 103 patients. A major challenge is how best to build a prognostic model for breast cancer in the presence of missing data.

The Kaplan Meier estimate of the survival function is the most commonly used method for the representation of the distribution of survival times. Extensions to graphical comparisons of these survival estimates were developed.

Classical approaches to modelling survival data using complete case analysis are examined and then an empirical simulation study is used to examine the effect of missing data on variable selection and to compare the performance of variable selection techniques in imputed data.

The final model identified Bilateral, Lymph Node status, Mitotic Count, Metastasis and UICC staging as being good predictors of Disease Free Survival and a subset of these for Overall Survival (Mitotic Count, Metastasis and UICC staging). These models have good concordance and were calibrated both internally and externally.

Classification and Regression Trees (CART) are a non-parametric approach to regression modelling. The main feature of CART is the data are recursively partitioned into groups and a simple prediction model fitted to each partition. A novel approach using surrogate splits to create alternative competing trees with comparable prediction power are introduced. This helps identify underlying structure in the data.

Acknowledgements

First, I would like to thank Dr John Newell, without his help, support and guidance this work would not have been possible. He has been a brilliant supervisor. His vast amount of experience and knowledge has made this an easier experience.

I would also like to thank Dr Carl Scarrott. He first introduced the idea of using surrogates splits to identify underlying structure. I would also like to thank Dr Helen Ingoldsby and Prof Grace Callagy for providing the data for this research.

My fiance, David, thank you so much for keeping me supplied with chocolate and wine for the last few months! You have been so supportive of me and have always been there to listen to me during times when I was finding things tough.

My parents, Teresa and Mike, you have made me the person I am today. You have always been so supportive of my choices and have always been there when I needed you.

My sister, Niamh, my brothers, Richard and Michael, my sister-in-law, Martha and my gorgeous nephew Danny, you have always been there for me. Those cuddles from Danny in the last few months have cheered me up when I have really needed it. My extended family and friends, you all have been so supportive in the last few years and I really appreciate it.

Finally, I would like to thank the National Breast Cancer Research Institute for their financial assistance throughout this process.

Chapter 1

Introduction

The primary aim of my research is to create a reliable and precise prognostic model for breast cancer survival and time to recurrence using data collected for invasive breast cancer patients in the National Breast Cancer Research Institute (NBCRI), National University of Ireland, Galway. The data are comprised of 647 patients with patient characteristics and genetic markers for breast cancer (details given in `Section 1.2.1`). An additional complexity exists due to the presence of missing data as highlighted in `Section 1.2.1`. A complete case analysis would reduce the number of cases to 103 patients. An additional challenge therefore is how best to build a prognostic model for breast cancer in the presence of such missing data.

The second aim is to identify potentially useful predictors of the Oncotype DX classification, a genetic based test for prediction of breast cancer recurrence in 10 years used to guide therapy choices. Oncotype DX is an expensive patented test (it costs $\$3,800$ per patient) that analyses 16 genes in patients with Oestrogen Receptor positive and Lymph Node negative breast cancer. It assigns each patient with a *Oncotype DX Recurrence Score (RS)*, which is an estimate of their likelihood of developing a breast cancer recurrence in 10 years. The higher the RS means a higher likelihood of disease recurrence. The RS were divided into three categories low risk RS < 18; intermediate risk $18 - 30$; and high risk > 30.

Treatment after surgery is related to the Oncotype DX classification. If a pa-

tient is categorized as low risk of disease recurrence only Tamoxifen (a hormone therapy which blocks the effects of Oestrogen on the cancer cells) is required, however if a patient is classified as intermediate or high risk, Tamoxifen and chemotherapy are required. Low risk patients can therefore avoid the harmful side effects of chemotherapy.

Published research suggests that the results of Oncotype DX can be predicted just as well by routinely (and more cheaply) assessed pathological variables and biomarkers. For example, previous published studies in this area have used tree based models, however each of the papers have identified different sets of risk factors, namely

- Grade, progesterone receptor status and Ki67 level [Allison et al., 2011],
- Mitotic score (M) greater than one combined with a negative progesterone receptor result [Flanagan et al., 2008],
- Tubule formation (T), nuclear grade (N), mitotic count (M), oestrogen receptor score, progesterone receptor score and Her2/neu score [Auerbach et al., 2010],
- Oestrogen receptor score, progesterone receptor score, Her2 score and the three components of grade (T, N and M) [Geradts et al., 2010].

This has lead to considerable debate. Comparable data are available from a Galway cohort and one aim is to use these data to try to consolidate conflicting evidence.

1.1 Breast Cancer

Breast cancer is a type of cancer originating in breast tissue, most commonly from the inner lining of milk ducts or the lobules. It is caused by the uncontrolled growth of cells.

Breast cancer has the second highest mortality rate of all cancers and is the leading cause of cancer-related death in women in Ireland. `Figure 1.1` contains a scatterplot of the number of breast cancer diagnoses per 100,000 women verses breast cancer deaths per 100,000 women by country in 2002. The

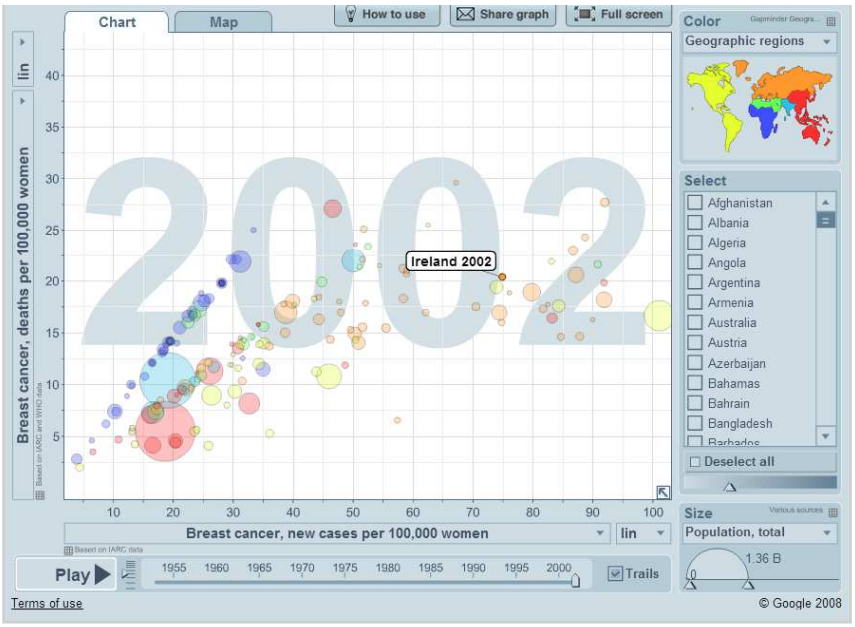


Figure 1.1: Breast cancer diagnosis per 100,000 women verses breast cancer deaths per 100,000 from *www.gapminder.com*. Over 1,800 new cases of Breast Cancer were diagnosed in Ireland in 2002.

countries are colour coded by continents and for those countries which did not have the data available, these missing values were interpolated. Ireland had one of the highest number of breast cancer cases diagnosed per 100,000 women in 2002. There were over 1,800 new cases of breast cancer diagnosed that year and this has now increased to nearly 3,000 new cases diagnosed annually in Ireland. This steady increase can be explained in part by the more stringent screening of Irish women through the *Breast Check* clinics. *Breast Check* is a Government-funded programme providing breast screening and invites women aged between 50 and 64 for a free mammogram on an area-by-area basis every two years. Despite recent reductions in mortality rates due to earlier diagnosis and improved therapies, on average over 600 Irish women die from the disease annually.

BreastCheck
The National Breast Screening Programme

In addition to taking a patient's history with regard to risk factors of breast cancer and an examination, the doctor may use a number of tests to help diagnose breast cancer including:

- Mammogram: an X-ray of the breast.
- Ultrasound: may be performed in addition to or instead of a mammogram, especially in younger women.
- Biopsy: if a lump is found on the breast on examination or in a mammogram, a biopsy is performed using a needle to remove a part of the tissue which is then examined under the microscope.

The size, stage, rate of growth, whether the cancer is sensitive to hormones and other characteristics of a breast cancer determine treatment options. Treatments include surgery, which is the most common, and a combination of radiotherapy, hormone therapies and chemotherapies.

1.1.1 National Breast Cancer Research Institute

The National Breast Cancer Research Institute (NBCRI) is a voluntary based charity founded in 1991. Their research aims are to determine the cause of breast cancer, to improve diagnosis and treatment for patients. Many of their recent research projects involve identifying micro-RNA in the blood and tissue that can distinguish between breast cancer patients and controls [Waters, Wall, et al., 2012, McDermott, Wall, et al., 2013, Khan, Wall, et al., 2013].

1.2 Datasets

The two datasets used throughout the thesis will now be introduced.

1.2.1 Galway Breast Cancer Patient Cohort

The data were collected retrospectively from the records from University Hospital, Galway (single centre retrospective study). The data set is comprised of 647 invasive breast cancer patients diagnosed between 1999 and 2006. Breast cancer can be classified using several different things such as Stage, Histopathology, Grade and Receptor Status. These data will be used to demonstrate the different survival analysis techniques. Summary statistics for the data are given in `Table 1.1 & 1.2` and a graphical representation of the data is given in `Figure 1.2`. Red indicates missing values for patients for predictors. This plot is useful to show how much missing data is scattered across the dataset.

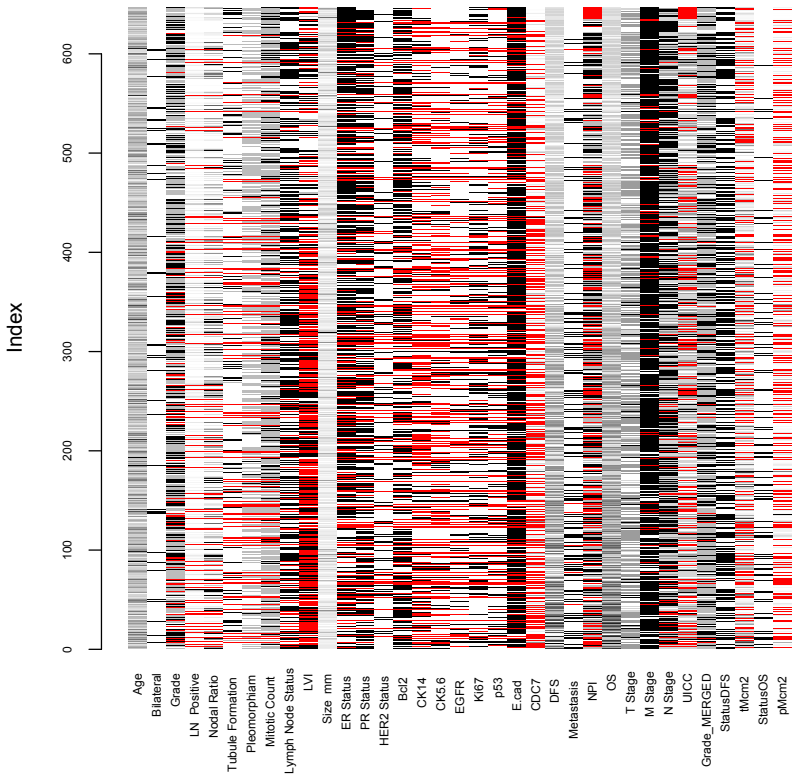


Figure 1.2: Matrix plot for the breast cancer data. Red indicates missing values and grey, black and white indicate the different levels in a predictor.The index on the y-axis are individual patients.

There are two main outcomes of interest, Disease Free Survival (DFS) and Overall Survival (OS). DFS was defined as the time from diagnosis until any recurrence of breast cancer or diagnosis of metastatic disease. OS was the time from diagnosis until death. At the time of gathering clinical information, 83% of patients were alive and 49% were alive and had no evidence of disease recurrence. The mean length follow-up time was 50 months, range < 1 to 113 months.

Tumours were graded using the TNM Classification of Malignant Tumours [Sobin et al., 2009] and their size in millimeters was measured. The average size of a tumour was $23.8mm$, ranging from $0mm - 140mm$. Grading is based on assessments of tubule/gland formation, nuclear pleomorphism and mitotic counts [Elston and Ellis, 2002]; 51% of BC cases are grade 2, with only 14% being grade 1 and 35% grade 3, with grade 1 having the best prognosis and grade 3 the worst prognosis. The cancer had spread to other organs in 20% of patients (distant metastasis).

The total number of positive lymph nodes was recorded for each patient and lymph node status was grouped as positive or negative, with any patient with at least one lymph node positive classed as positive. 49% of patients had at least one lymph node positive. Lymphovascular Invasion (LVI) was also recorded. Fifty three percent of patients were identified as having LVI, with a further 2% with probable LVI.

Biomarker development is one of the main focuses in cancer research at the moment. Biomarkers can be used for risk assessment, screening, diagnosis and prognosis. Biomarkers can be measured using the Whole Tissue Sample (WTS) or Tissue Micro Arrays (TMAs). TMAs are a powerful tool for molecular classification of breast tumours as they facilitate large-scale analysis of biomarker expression on hundreds to thousands of cases, permitting large-scale testing and validation of potential prognostic and predictive markers. They consist of paraffin blocks with up to 1000 tissue samples in an array. This allows many histological analysis of multiple biomarkers. Generally biomarkers can be assessed through both WTS and TMAs, however for the less routinely assessed biomarkers in this data only the TMA measurement is available. To be consistent across all biomarkers the results for only TMAs are used for the analysis.

Information on 13 different biomarkers was collected. `Table 1.2` classi-

Clinical Variables	
	mean (sd)
Age	57.8 (13.0)
Tumour Size (mm)	27.6 (19.8)
LN Positive	2.5 (4.5)
	n(%)
Grade	
1	80(14)
2	297(51)
3	210(35)
Lymph Node Positive	
Yes	287(49)
No	297(51)
Lymphovascular Invasion	
Yes	200(53)
No	170(45)
Probable	9(2)
Tubule Formation	
$> 75\%$	31(5)
10-75%	73(13)
$< 10\%$	467(82)
Nuclear Pleomorphism	
Mild	7(1)
Moderate	282(49)
Marked	282(49)
Mitotic count	
Low	360(63)
Moderate	100(18)
High	111(19)
Metastasis	
Yes	127(20)
No	513(80)
Events	
Alive, no disease	292(45)
Alive, locoregional disease	198(30)
Alive, distant metastasis	56(9)
Dead with evidence of disease progression	84(12)
Dead with no evidenence of disease progression	26(4)

Table 1.1: Summaries for clinical and pathological predictors for the invasive breast cancer patient sample.

fies each patient into positive and negative using varying cut-off values in the literature. Oestrogen receptor status, progesterone receptor status and human epidermal growth factor receptor 2 status (Her2) are routinely assessed in breast cancer patients. Roughly two out of every three breast cancers test positive for at least one of these hormone receptors.

Oestrogen (ER) is a female sex hormone. It stimulates some breast cancers to grow by triggering particular proteins (receptors) in the cancer cells. If breast cancer cells have oestrogen receptors, the cancer is said to be oestrogen positive. Hormone therapies can stop oestrogen from stimulating the cells to divide and grow. These therapies work best for oestrogen positive breast cancers. The majority of patients in the sample are oestrogen positive (66%).

The cancer is progesterone receptor (PR) positive if it has progesterone receptors. Fifty six percent of patients in the sample are PR positive. Again, this means that the cancer cells may receive signals from progesterone that could promote their growth.

The Her2 gene makes Her2 proteins. Her2 proteins are receptors on breast cells. Normally, Her2 receptors help control how a healthy breast cell grows, divides and repairs itself. However in about 25% of all breast cancers, the Her2 gene does not work correctly and makes too many copies of itself (known as Her2 gene amplification). All these extra Her2 genes instruct breast cells to make too many Her2 receptors (Her2 protein over expression). This makes breast cells grow and divide in an uncontrolled way. Her2 positive breast cancers tend to grow faster and are more likely to spread and return compared to Her2 negative breast cancers. Eighty-six percent of patients in the sample are Her2 negative. Patients who are ER or PR positive can receive hormonal treatments such as Tamoxofin and patients who are Her2 positive can receive Herceptin treatment.

The other biomarkers are not as routinely assessed - Bcl-2, Ki67, CK5/6, CK14, EGFR, p53, E-cad - but have been identified in the literature of having links to breast cancer. Bcl-2 has important roles in apoptosis, cell proliferation and cell differentiation in breast cancer. Cytoplasmic staining was assessed for Bcl-2. The value of Bcl-2 protein as predictive/prognostic factor for adjuvant chemotherapy treatment in breast cancer has been investigated. Fifty-five percent of patients are Bcl-2 positive.

Biomarkers	
	n(%)
Oestrogen Receptor Status	
Positive	340(66)
Negative	172(34)
Progestrogen Receptor Status	
Positive	294(56)
Negative	230(44)
Her2 Status	
Positive	83(14)
Negative	505(86)
BcI2	
Positive	288(55)
Negative	237(45)
CK14	
Positive	94(22)
Negative	341(78)
CK5/6	
Positive	66(15)
Negative	386(85)
EGFR	
Positive	74(14)
Negative	444(86)
Ki67	
Positive	161(30)
Negative	369(70)
p53	
Positive	103(20)
Negative	411(80)
E-cad	
Positive	455(88)
Negative	62(12)
tMcm2	
<1	8(2)
1-10	89(24)
11-33	133(35)
34-66	74(20)
>66	73(19)
	mean (range)
CDC7 Expression	2.59 (0.0-34.2)
pMcm2 Expression	6.19 (0.0-90.0)

Table 1.2: Summaries for pathological biomarker predictors for the invasive breast cancer patient sample.

The expression of the human Ki-67 protein is strictly associated with cell proliferation. The Ki-67, a cell proliferation associated nuclear antigen, is found in cells in nearly all stages of the cell cycle and is therefore a direct indicator of the growth fraction. Thirty percent of patients are Ki67 positive. The Ki-67 growth fraction is significantly related to the grade of most tumors, being highest in grade III invasive carcinomas. Oestrogen and progesterone receptor negative tumors tend to be Ki-67 positive and this index could be used to add adjuvant chemotherapy in both receptor negative and positive patients.

Cytokeratins are proteins of keratin-containing intermediate filaments found in the cytoskeleton of epithelial tissue. The cytokeratins can be divided into low versus high molecular weight solely based on their molecular weight. Expression of these cytokeratins is frequently organ or tissue specific. Cytokeratin (CK) 14 is an acidic cytokeratin. Twenty-two percent of patients in the sample are CK14 positive. Fifteen percent of patients are CK5-6 positive. CK5-6 are antibodies to basal cytokeratins that stain myoepithelial cells (found in the cell membrane).

Fourteen percent of patients are EGRF (Epidermal Growth Factor Receptor) positive. Mutations affecting EGFR expression or activity could result in cancer. p53 (protein 53) is crucial in multicellular organisms, where it regulates the cell cycle and thus, functions as a tumor suppressor that is involved in preventing cancer. Twenty percent of patients are p53 positive. Eighty-eight percent of patients are E-cadherin positive. E-cadherin is a calcium-dependent cell-cell adhesion molecule with pivotal roles in epithelial cell behavior, tissue formation, and suppression of cancer. CDC7 (Cell Division Cycle 7) and pMcm2 were also measured as continuous variables. CDC7 is a gene which codes the protein cdc7 kinase. This protein regulates the cell cycle. tMcm2 and pMcm2 are mini-chromosome maintenance proteins which are involved in genome replication. The tMcm2 protein was measured as a categorical variable using cut offs for percentage staining.

Various techniques will be investigated to identify the 'best' model for the data. The dataset has a large number of variables and the aim is to identify a subset of these to create a prognostic model for DFS and OS while retaining prediction performance. During the last two decades, several clinical and pathological indicators such as histological grade, tumour size and lymph node

involvement have been used for prediction of survival of breast cancer patients independently of treatment, as known as prognostication [Haibe-Kains et al., 2008]. For some markers there have been several published studies with conflicting results [Altman and Lyman, 1998].

As can be seen in `Figure 1.2` there is a high proportion of missingness present. The most common way to deal with missing data is casewise deletion. This can reduce the sample size significantly even if there is as little as 10% missing per predictor. The sample size of 647 patients is reduced to 103 patients using casewise deletion. This may result in bias from restricting the analysis to complete data.

Prognostic Models

The development of a successful model depends on the following features [Altman and Royston, 2000]:

- the potential for accurate prognosis, which is presumably unknown;
- the intrinsic prognostic information in the available factors, which depends on many things, including the physiology of the disease in question;
- the measurement process, which converts the intrinsic information into numbers, some measurements being more reliable than others;
- and the accuracy with which the model converts the measurements into predictions.

The idea of a transparent, simple model is not necessarily a virtue; performance of the model is more important, and simplicity over complexity should not be the primary consideration in the model building process [Taylor et al., 2012]. The model needs to be as simple as possible but as complex as necessary.

The advantages of retrospective studies are its simplicity and feasibility. The disadvantages include identifying patients, missing data and incorrect information on the patients' records.

1.2.2 Oncotype DX Classification Data

The second dataset, the Oncotype DX data, is now introduced and a description of the variables collected given.

The Galway (West of Ireland) dataset contains 52 patients with their Oncotype DX score and categorization and 32 useful clinical and pathological variables. `Table 1.3` contains a summary of the predictors from the Oncotype DX data.

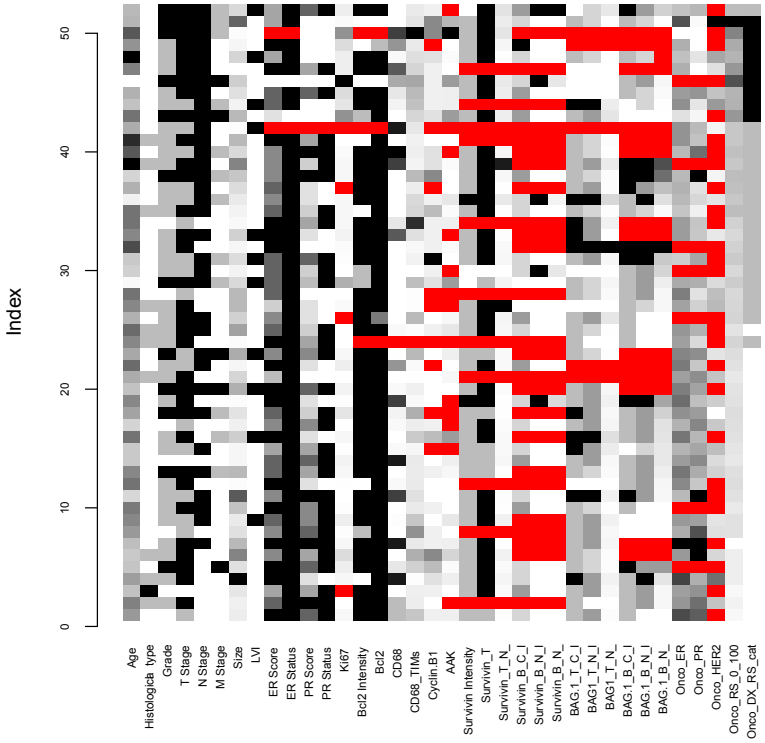


Figure 1.3: Matrix plot for the Oncotype DX data. Red indicates missing values and grey, black and white indicate the different levels in a predictor.

The aim of this study is to identify a set of useful clinical and pathological variables which can predict the risk of recurrence and consolidate previous conflicting results in the literature. However due to the small sample and the large number of useful predictors (some with missing data), the results from classical approaches such as logistic regression are unstable. A matrix plot is given in `Figure 1.3`, which highlights the high proportion of missing data present. From this plot, it is obvious there is missing values for majority of

Clinical/Pathological Variables	
	mean (sd)
Age	57.8 (9.7)
Tumour Size (mm)	22.4 (9.8)
Ki67 Staining	9.4 (14.5)
CD68 Staining	18.5 (15.2)
Cyclin B1 Staining	0.9 (1.0)
AAK Staining	2.5 (3.7)
Survivin TN Staining	10.5 (19.2)
	n(%)
Grade	
1	7(13)
2	33(63)
3	12(24)
Tumour Stage	
1	2(4)
2	17(33)
3	33(63)
Lymph Node Stage	
1	0(0)
2	18(35)
3	34(65)
Metastasis Stage	
1	35(67)
2	12(23)
3	5(10)
Lymphovascular Invasion	
Yes	11(21)
No	41(79)
Oestrogen Receptor Status	
Positive	48(96)
Negative	2(4)
Progestrogen Receptor Status	
Positive	31(60)
Negative	20(40)
Bag1 Staining	
0	8(17)
1	15(32)
2	19(40)
3	5(11)

Table 1.3: Summaries for clinical and pathological predictors for the patient sample for the Oncotype DX data.

the patients. An alternative to the classical approaches, the non-parametric approach of Classification and Regression Trees can be used to identify useful predictors. A novel application of using surrogate splits in the tree building process to identify underlying structure in the data will be explored.

The results of this study have been published in *The Breast* [Ingoldsby, Webber, Wall, Scarrott, Newell, and Callagy, 2013].

1.3 Structure of Thesis

The thesis contains six chapters; summaries of Chapters 2 to 6 now follow.

Chapter 2: Survival Analysis

Chapter 2 starts with an introduction to survival data and examines graphical and numerical summaries of survival estimates. The Kaplan Meier estimate of the survival function is the most common method used for the representation of the distribution of survival times. It is generally used for graphical comparisons of survival for two or more groups of patients.

The Log-rank test performs a hypothesis test to compare the survival estimates of two groups. Graphical representation of survival estimates across two or more groups can be a useful tool to help interpret the Log-rank tests. Extensions to classical approaches will be presented in this chapter.

Chapter 3: Non-parametric Tree based Methods

Tree based methods were first introduced by Breiman et al. [1984] and will be introduced using examples based on both the Oncotype DX classification and score. Also examples of survival trees will be shown using the Breast Cancer data from UCH Galway.

Trees were used to identify useful predictors for the Oncotype DX classification and score. However, these trees were created using a small sample. A novel use of surrogate splits is introduced where the primary surrogates are used to create competing or comparable trees which may have the same predictive power as the original tree. To create these extra trees an interactive surrogate plot developed in R will be demonstrated.

Chapter 4: Classical Approaches to Modelling

This chapter focuses on classical approaches to modelling survival data. As the response of interest is survival, one choice is the Cox proportional hazards model. An introduction to the theory underpinning the model is given and then various modelling approaches examined. Firstly, a model containing all predictors was created. Next various variable selection techniques are applied, including backward selection, ridge regression and the LASSO (Least Absolute Shrinkage and Selection Operator). Splines were applied to relax the linearity assumptions and the need for interaction terms.

Chapter 5: Variable Selection Techniques in Imputed Data

The Cox model typically deals with missing data by casewise deletion. Casewise deletion can result in over half the cases being deleted even if there is as little as 10% missing per variable. This results in a smaller dataset to perform the analysis with loss of power as a consequence. Chapter 5 discusses an alternative approach by performing variable selection on imputed data. Obviously complete data would be the best scenario, however since this is a retrospective study, it is not possible to retrieve the data that are missing. The lost information, caused by casewise deletion, can be "reclaimed" somewhat by performing variable selection techniques on imputed data.

The results of an empirical simulation study used to assess the performance of these techniques are given.

Chapter 6: Validation

In the end of chapter 5 a summary of all the models assessed in the previous chapters and their performances is given. Final models for disease free survival and overall survival are reported. The next step in the process is to validate the models to see how well they perform.

Interval validation will be performed using bootstrap resamples of the data examining the discrimination and calibration of the models. External validation will be performed for the OS model using a dataset from 10 European based breast cancer clinics. Finally some visualization tools for the models will be presented in the chapter.

Conclusions, an overall summary and suggested future work is given to conclude the thesis in Chapter 7.

Chapter 2

Survival Analysis

2.1 Introduction

This chapter will introduce survival analysis and examine the graphical and numerical summaries of survival estimates. Survival data measures time from some origin point to some event. There is one major difference between survival data and other types of continuous responses: the time to the event occurring is not necessarily observed in all subjects [Machin et al., 2006].

Survival analysis is the name given in statistics to the analysis of such lifetime data. Survival analysis can be applied in many different areas of research. In medicine, for example a study may follow disease free patients until they develop heart disease. Another study may follow a patient from diagnosis of cancer to death or the recurrence of the disease. An example in criminology would be parolees, following people who are released from prison for weeks to see if they are rearrested. An example in engineering, is the lifetime of components for machines; testing the components to see how long they will last until they fail.

There are three special principles central to survival analysis [Van Houwelingen and Putter, 2011]

1. It takes time to observe time; studies tend to have fixed termination time.
2. The event might never happen; subjects may experience the event after the study terminates or not at all.

3. You only die once; once a subject experiences the event the survival time is not measured any longer.

The basic goals of survival analysis are to [Kleinbaum and Klein, 2005]:

- estimate and interpret survival and/or hazard functions from survival data;
- compare survival and/or hazard functions;
- assess the relationship of explanatory variables to survival time.

The typical graphical summaries in survival analysis are plots of the parametric or non-parametric estimated survival function. Generally comparisons of survival estimates are preformed using the Log-rank test or variations thereof. Modelling the effects of covariates and factors can be typically be performed using a Cox Proportional hazards model [Cox, 1972].

The Kaplan Meier survival estimator [Kaplan and Meier, 1958] and the Cox regression model for survival [Cox, 1972] are standard elements in the training of medical doctors, and the papers describing these statistical techniques are among the most frequently cited scientific papers. In Web of Science, for example, a search on May 12, 2011 for the papers by Kaplan & Meier and by Cox resulted in 34,946 and 25,149 hits [Van Houwelingen and Putter, 2011].

2.1.1 Censoring

As survival analysis concerns time to a particular event, the response contains both continuous (survival time) and discrete values (events). Survival times are positive and continuous. The event is a binary outcome, either the event occurs or not, typically coded as 1 and 0 respectively. Those patients who do not experience the event are referred to as censored.

There are 4 reasons why censoring may occur:

- a patient does not experience the event before the study ends;
- a patient is lost to follow up during the study period;
- a patient withdraws from the study (dropout);

- a patient may have experienced another event (competing risk) and can no longer be observed (e.g. death by accident).

The most common type of censoring is *right censoring*. This is when a patient is enrolled at the beginning of the study however they are lost to follow-up or withdrawn before the study ends. *Left censoring* occurs when a patient's true survival time is less than the patient's observed survival time. For example, if a study is following patients until they are diagnosed with HIV, an event is recorded when the patient first tests positive for the disease. However, the exact time of exposure to the disease is unknown. Another form of censoring is *interval censoring*. Interval censoring occurs when the patient experiences the event in an interval of time. For example, a patient experiences the event between two hospital appointments. Censoring may also occur due to the termination of the study, such censoring is termed *administrative censoring*.

Survival data uses information from the whole follow-up period and all patients can contribute information during their time under surveillance [Bull and Spiegelhalter, 1997]. The patients who have been censored have been observed but have not experienced the event before the end of the observation time. Basically, a censored observation is an incomplete observation; it contains only partial information about event time [Wienke, 2010].

The challenge in analyzing survival data is that the survival time for a patient that is censored must be incorporated in the analysis until censorship.

If subjects with censored survival times are removed from the analysis, it could lead to unbiased estimates of the survival time. An important assumption is that time to censoring and survival times are independent (non-informative censoring).

2.2 The Survival and Hazard Function

Throughout this thesis T will represent the random variable 'survival time' of an individual under investigation, which is the time from diagnosis to the event occurence (i.e. death for OS and recurrence for DFS). The cumulative

distribution function F is given by:

$$F(t) = P(T \leq t) = \int_0^t f(t)dt \tag{2.1}$$

where f is the probability density function. Once the probability density function is specified for survival time, the corresponding survival and hazard functions can be determined, as the probability density function can be expressed in terms of the product of the survival and hazard function.

The survival function captures the probability that a patient will survive beyond time t. The survival function is given by

$$S(t) = Prob(T > t) = 1 - F(t) \text{ for any } t > 0 \tag{2.2}$$

where $F(t)$ is the cumulative distribution function for time T.

Here are some properties of the survival function:

- $S(0) = 1$, i.e. no one experiences the event at time 0
- $lim_{t\to\infty}S(t) = 0$, i.e. everyone eventually experiences the event (death)
- $S(t_a) \geq S(t_b)$ where $t_a \leq t_b$ i.e. $S(t)$ declines monotonically
- $S(t) = 1 - F(t) = \int_t^\infty f(t)dt$
- typically, the population survivor function is smooth, however estimates of it are not.

The survival function focuses on the probability the event will not happen where as the hazard function is the opposite, it focuses on the event occurring. The hazard function $h(t)$ gives the instantaneous potential per unit time for the event to occur, given the individual has survived up to time t. The hazard function is sometimes called the conditional failure rate or instantaneous rate as it is the probability an observation will have the event in the next unit of time given they have not experienced the event up until time t. The hazard function is defined by the following equation

$$h(t) = lim_{dt\to 0}\frac{P(t \leq T < t + dt|T > t)}{dt} = \frac{f(t)}{S(t)} \tag{2.3}$$

The cumulative hazard function describes the accumulated risk up until time t. The hazard function is non-decreasing so this means it is either increasing or stable as t increases. The cumulative hazard function is given by

$$H(t) = \int_0^t h(u)du \tag{2.4}$$

The cumulative hazard function can also be written as

$$H(t) = \int_0^t \frac{f(u)}{S(u)}du = \int_0^t \frac{-S'(u)}{S(u)}du = [-logS(u)]_0^t = -logS(t) \tag{2.5}$$

Alternatively, the survival function can be written in terms of the cumulative hazard function

$$S(t) = exp\,(-H(t)) \tag{2.6}$$

2.2.1 The Likelihood Function for Survival Data

The likelihood function for survival data is

$$L(x, \delta; \theta) = \prod_{i=1}^{n} f(x_i; \theta)^{\delta_i} S(x_i; \theta)^{1-\delta_i} \tag{2.7}$$

This section explains how this is derived.

It is important to assume independence between censoring and survival times for two reasons. Firstly, the probability of being censored for any subject at time t does not depend on that subject's prognosis for event time at time t. Secondly, it simplifies the likelihood function for survival. Consider a sample of n identically independent distributed (iid) subjects. Each subject has an event time T_i, a censored time C_i, a survival time X_i, where $X_i = min(T_i, C_i)$, and a censoring indicator Δ_i, where $\Delta_i = I(T_i \leq C_i)$. Δ_i is defined to be:

$$\Delta_i = \begin{cases} 1 & \text{if } T_i \leq C_i, \text{i.e. } X_i \text{ is event time} \\ 0 & \text{if } T_i > C_i, \text{i.e. } X_i \text{ is censored time} \end{cases}$$

For T, we have the density function $f(t)$, the distribution function $F(t)$, the survival function $S(t)$ and the hazard function $h(t)$. For C which is also an event time, we have the density function $g(t)$, the distribution function $G(t)$,

the survival function $K(t)$ and the hazard function $\mu(t)$.

By assuming independent censoring the density function of (X, Δ) is as follows:

$$f(x,\delta) = lim_{h\to 0}\frac{P(x \le X < x+h, \Delta = \delta)}{h}, x \ge 0, \delta = 0, 1 \qquad (2.8)$$

Consider $Case$ 1 where $\delta = 1$, i.e. $T \le C$ and $X = min(T, C) = T$,

$$\begin{aligned} & P[x \le X < x+h, \Delta = 1] \\ = \; & P[x \le X < x+h, C \ge T] \\ \approx \; & P[x \le X < x+h, C \ge x] \\ = \; & P[x \le X < x+h]P[C \ge x] \\ = \; & f(\xi)h \times K(x), \text{where } \xi \in [x, x+h) \end{aligned}$$

Therefore

$$\begin{aligned} f(x, \delta = 1) &= \lim_{h\to 0}\frac{P[x \le X < x+h, \Delta = 1]}{h} \\ &= \lim_{h\to 0}\frac{f(\xi)h \times K(x)}{h} \\ &= f(x) \times K(x) \end{aligned}$$

Similarly, consider the case where the observation is censored:

$$\text{Case 2: } \delta = 0 \text{ then } f(x, \delta = 0) = S(x) \times g(x)$$

Combining $Case$ 1 and $Case$ 2, the probability density function becomes:

$$f(x,\delta) = f(x)^{\delta}S(x)^{1-\delta} \times K(x)^{\delta}g(x)^{1-\delta} \qquad (2.9)$$

$f(x)$ depends on an unknown parameter θ, i.e. $f(x, \theta)$ and $g(x)$ depends on an unknown parameter ϕ, i.e. $g(x, \phi)$

The likelihood function is defined as

$$L(x, \delta; \theta, \phi) = \prod_{i=1}^{n} f(x_i, \delta_i; \theta, \phi)$$

$$= \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i} S(x_i;\theta)^{1-\delta_i} \times K(x_i;\phi)^{\delta_i} g(x_i;\phi)^{1-\delta_i} \quad (2.10)$$

As has been stated previously, the assumption that T the survival times and C the censoring are independent is extremely important. Here the main aim is to make inference on the parameter θ characterizing the distribution of T. Since T and C are independent, θ and ϕ have no common parameters. When maximising the log likelihood to derive the maximum likelihood estimate, ϕ will be considered a constant. The likelihood function simplifies to

$$L(x,\delta;\theta) = \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i} S(x_i;\theta)^{1-\delta_i} \quad (2.11)$$

Or equivalently to

$$L(x,\delta;\theta) = \prod_{i=1}^{n} h(x_i;\theta)^{\delta_i} S(x_i;\theta) \quad (2.12)$$

since the density function is a product of the survival and hazard functions. The maximum likelihood is used to find estimates for the model parameters.

2.3 Non-parametric Survival Estimates

A graphical representation of a survival function provides the most understandable summary of the time related data. The most commonly used estimator for $S(t)$ was derived by Kaplan and Meier [1958]. The Kaplan Meier (KM) survival estimator is a non-parametric procedure for estimating a survival function that does not make any assumptions about the shape of the underlying survival function.

Edward Kaplan died in 2006. Kaplan worked for Bell Labs where his main focus was on finding a better way to measure the survival of vacuum tubes. Paul Meier died 2 years ago aged 87. He was a well renowned statistician, who was described as the "statistician who revolutionized medical trials". Firstly, Meier was an advocate for randomization and in his obituary in the New York Times it was said "that strategic decision half a century ago has already saved millions of lives and those millions should be attributed to Paul". The other major contribution Meier made was in estimating the survival function. Indeed

the KM estimator is now synonymous with plots of the survival function.

The general formula for the KM survival estimator, given in `Equation 2.13`, is the probability of surviving past the previous failure time t_{j-1}, multiplied by the conditional probability of surviving past time t_j, given survival to least time t_j [Kleinbaum and Klein, 2005].

$$S(t_j) = S(t_{j-1}) \times Pr(T > t_j | T \geq t_j) \tag{2.13}$$

where t_j is a specified failure time j and t_{j-1} is the previous failure time.

In other words, the KM estimator is the product of the estimated survival probabilities at each distinct event time, i.e.

$$S_{KM}(t) = \prod_{j=1}^{t} (1 - \frac{e_j}{r_j}) \tag{2.14}$$

where e_j is the number of patients who experience the event at time t_j and r_j is the number of individuals still 'at risk' at time t_j. The Kaplan Meier estimate is also called the "product limit estimate"; as to calculate the probability that a patient will survive beyond time t, the patient must not have experienced the event previous to time t. It involves multiplying all the probabilities that the patient had not experienced the event for all the previous time points to obtain the final estimate at time t.

`Figure 2.1` contains the KM estimates for Overall Survival (OS, event is death) by Lymph Node Status for the UCH Galway patients with months after presentation given on the horizontal axis. At each event time the KM survival probability is calculated using `Equation 2.14`.

Confidence intervals are constructed using the Greenwood variance estimator given by the following formula [Greenwood, 1926]

$$\widehat{Var}(\hat{S}(t_j)) = (\hat{S}(t_j))^2 \sum_{t_j < T} \frac{e_j}{r_j(r_j - e_j)} \tag{2.15}$$

`Figure 2.2` contains the KM estimates by Her2 Status for OS and DFS without and with 95% confidence intervals, where the Greenwood variance estimator is again used to calculate the confidence intervals. For DFS, there Her2 negative patients seem to have a better prognosis than Her2 positive patients.

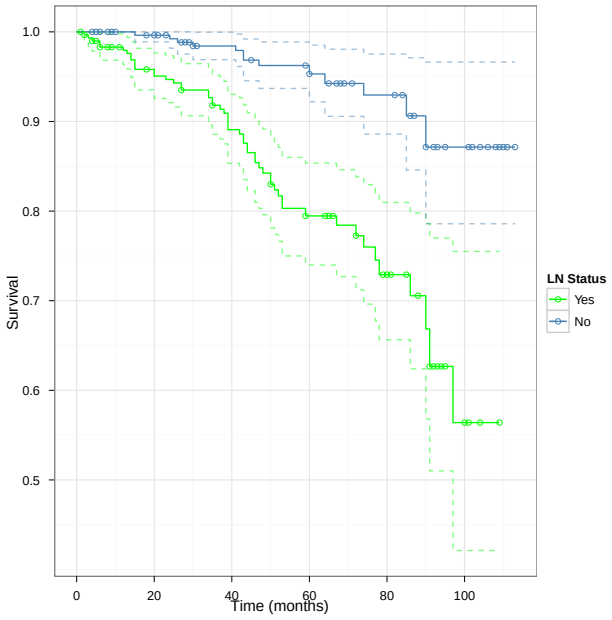
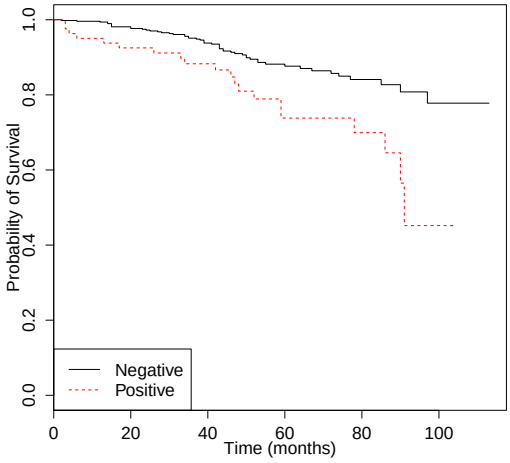


Figure 2.1: Kaplan Meier Survival Estimates for Overall Survival for Lymph Node Status.

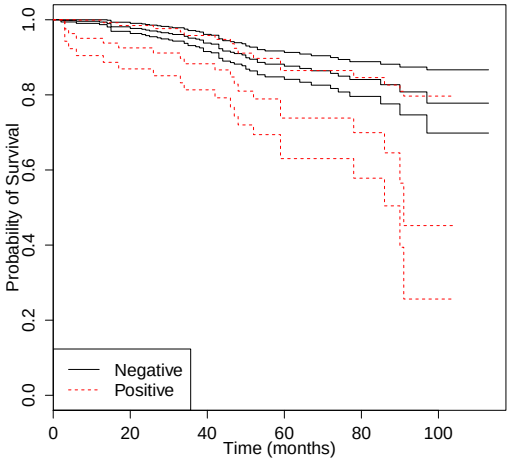
For OS, there does not seem to be much of a difference in survival for the first 5 years, however, after 5 years Her2 negative patients seem to have a better prognosis than Her2 positive patients.

The median survival time is easily read off a plot of the KM estimates by looking at the time where the probability of survival of 0.5 meets the KM estimate. For example, for patients with Her2 positive BC the median disease free survival is approximately 43 months where as Her2 negative patients have a median survival time of 67 months `Figure 2.2`. Examining overall survival, Her2 positive patients have a median survival time of 91 months, however since the KM estimates for Her2 negative patients does not fall below 0.5, it is not possible to give a median survival time for this cohort.

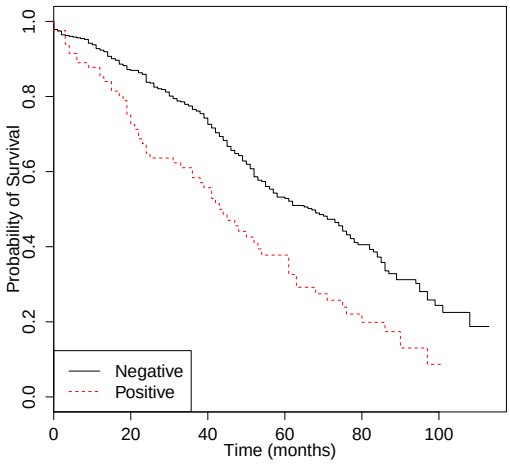
Generally, the main reason KM estimates are plotted for 2 or more groups is to graphically examine if there appears to be a difference between the groups. The Mantel-Haenszel Log-rank test performs a hypothesis test to compare the survival estimates of two groups. It is a large-sample chi-square test with the



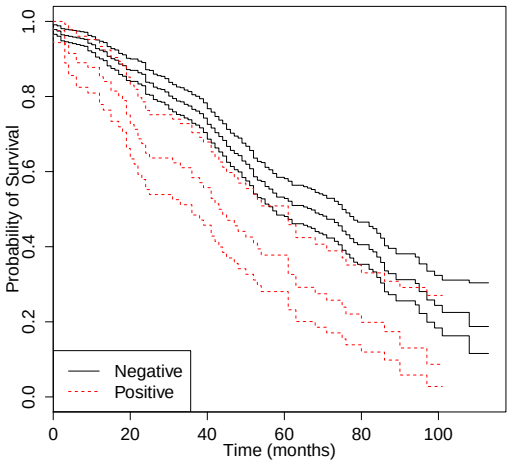
(a) Overall Survival



(b) Overall Survival



(c) Disease Free Survival



(d) Disease Free Survival

Figure 2.2: Survival Curves for Her2 status. Top row contains the survival curves for Her2 status for OS with and without confidence intervals (Log-rank p-value < 0.001). Bottom row contains the survival curves for Her2 status for DFS with and without 95% confidence intervals (Log-rank p-value < 0.001).

test statistic calculated as

$$\sum_{i=1}^{N} \frac{(Obs_i - Exp_i)^2}{Exp_i} \sim \chi^2_{N-1} \tag{2.16}$$

where N is the number of groups of the variable, Obs_i is the number of observed failures for group i, Exp_i is the expected number of failures for group i and the null hypothesis is that the survival curves of the different groups are identical. The expected number of failures for group 1 is calculated by

$$Exp_1 = \sum_{j=1}^{t} \left(\frac{r_{1j}}{r_{1j} + r_{2j}} \right) \times (e_{1j} + e_{2j}) \tag{2.17}$$

where, at time t_j, r_{1j} is the number of patients at risk in group 1, r_{2j} is the number of patients at risk in group 2, e_{1j} is the number of failures in group 1, e_{2j} is the number of failures in group 2. The Log-rank test statistic is calculated to be 27.5 using `Equation 2.16` on 1 degree of freedom (N equals 2 here as there are 2 groups) for the Kaplan Meier survival estimates in `Figure 2.1`. This corresponds to a p-value of < 0.001 which means there is a significant difference in overall survival between lymph node negative and positive patients, with lymph node negative patients having a better probability of survival.

Alternatives to the Log-rank test include the Wilcoxon, Tarone-Ware, Peto and Flemington-Harrington tests. These apply different weights at the j^{th} failure time to the Log-rank test statistic. The general form of the test statistic for a two group comparison is given by:

$$\frac{\left(\sum_j w(t_j)(Obs_{ij} - Exp_{ij})\right)^2}{var\left(\sum_j w(t_j)(Obs_{ij} - Exp_{ij})\right)} \tag{2.18}$$

where i equals 1 or 2, j is the j^{th} failure time, $w(t_j)$ is the weight at the j^{th} failure time. The weights for each of the tests are given in `Table 2.1`.

While typically, the KM survival estimates are plotted as in `Figures 2.1 and 2.2`. However another way of examining the estimates is using the cumulative hazard function which is given by $\hat{H}(t) = -log(\hat{S}(t))$. An example for OS for Her2 status is given in `Figure 2.3`.

Test Statistics	$w(t_j)$
Log-rank	1
Wilcoxon	n_j
Tarone-Ware	$\sqrt{n_j}$
Peto	$\tilde{s}(t_j)$
Flemington-Harrington	$\tilde{s}(t_{j-1})^p[1-\tilde{s}(t_{j-1})]^q$

Table 2.1: Weights used for various test statistics.

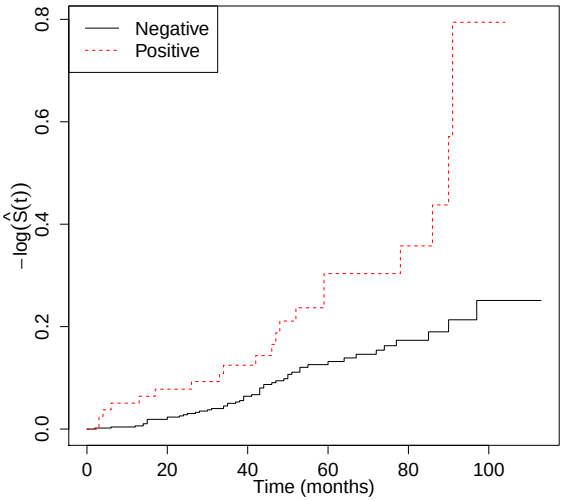


Figure 2.3: The cumulative hazard function for Overall Survival for Her2 status.

Kaplan Meier analysis provides a nonparametric method, but requires categorization of all continuous predictors. It is equivalent of cross-tabulated data for categorical outcomes for a survival context [Steyerberg, 2009].

`Section A.3` contains all the KM survival estimates and Log-rank p-values for each of the predictors in the breast cancer dataset from UCH Galway for overall survival and disease free survival. The majority of the comparisons are significant for both DFS and OS except the biomarker E-cad and Nuclear Pleomorphism. However, these tests were all performed marginally and multiple testing was not taken into account.

Variables	DFS	OS
Grade	< 0.001	< 0.001
Bilateral	< 0.001	< 0.001
Tubule Formation	0.029	0.028
Mitotic Count	< 0.001	< 0.001
Nuclear Pleomorphism	0.087	0.066
LN Status	< 0.001	< 0.001
ER Status	0.014	< 0.001
PR Status	< 0.001	< 0.001
Her2 Status	< 0.001	0.001
Metastasis	< 0.001	< 0.001
UICC	< 0.001	< 0.001
Tumour Staging	< 0.001	< 0.001
LN Staging	< 0.001	< 0.001
Metastasis Staging	< 0.001	< 0.001
NPI	< 0.001	< 0.001
Bcl2 Status	0.002	< 0.001
CK14 Status	0.033	0.571
CK5/6 Status	0.004	0.099
Ki67 Status	0.003	< 0.001
EGFR Status	0.077	0.024
E-cad Status	0.191	0.528
p53 Status	0.009	0.005

Table 2.2: Log-rank tests were performed on each of the categorical predictors. The p-values for each of these tests are given above. The Kaplan Meier estimates for each of the predictors is given in the `Appendix A.3`.

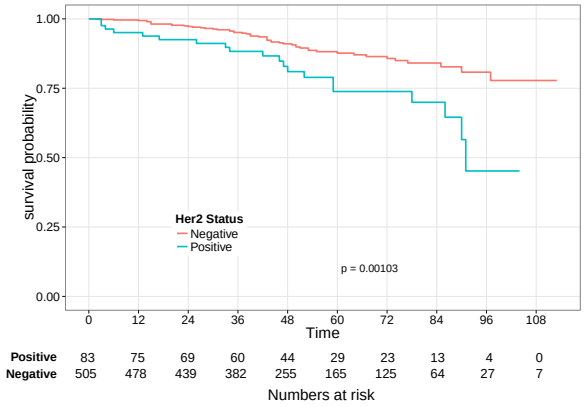


Figure 2.4: Kaplan Meier estimates for OS for Her2 Status. The number of patients at risk at each time point is given under the graph.

2.3.1 Graphical comparisons of two survival functions: Alpha Blending, Ratio and Difference of Survival Estimates for Two Groups

Visualization aims to turn data into understanding through graphical representations [Karvanen and Harrell Jr, 2009].

Obviously, as time progresses the number of patients at risk to calculate the survival estimates is decreasing as patients either experience the event or are censored. Apparent "differences" at later follow up times may be due to variability caused by smaller samples being observed rather than 'real' differences. The number of patients at risk at each time point can be added to the graphs of the KM estimates. An example is given in `Figure 2.4`. An alternative way to incorporate the number of patients at risk, is to use `alpha blending` from the `ggplot2` library in *R* for plotting. This shades the lines relative to the number of patients at risk at each time point, the lighter the colour represents a lower number of patients at risk. Hence as time increases the colour gets lighter since patients either experience the event or are censored. An example is given in `Figure 2.5` for Her2 status.

One statistical test for comparing two or more survival functions has been discussed. A graphical comparison of two survival curves has been identified as a useful tool for interpreting the results of the test. When the number of groups

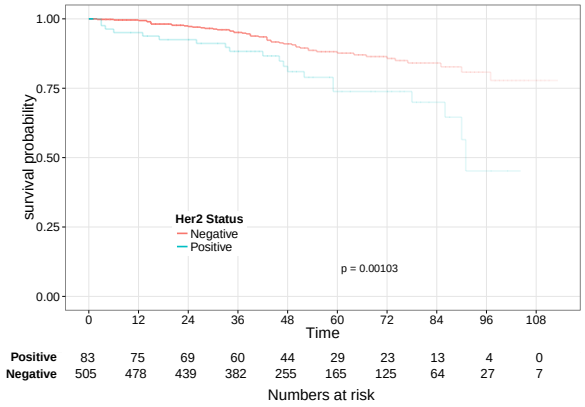


Figure 2.5: Kaplan Meier estimates for OS for Her2 Status. The number of patients at risk at each time point is given under the graph and `alpha blending` on the lines to show how many patients are at risk at the time.

increases, the plot can get cluttered once confidence intervals are included. One way of looking at the difference in the survival estimates of two curves is to examine the ratio of survival estimates for one group to the survival estimates for the other group at each failure time. This was first introduced by Newell et al. [2006]. The pointwise ratio of survival estimates is given by

$$R(t) = \frac{S_1(t)}{S_2(t)} \tag{2.19}$$

where $S_k(t)$ are estimated using the Kaplan Meier estimator for the k^{th} treatment at time t.

Confidence intervals for the pointwise difference and ratio can be created using bootstrap resamples of the data. The interval between the 2.5% and 97.5% percentiles of the bootstrap distribution of a statistic is a 95% bootstrap percentile confidence interval for the corresponding parameter [Hesterberg et al., 2007]. Under the null hypothesis, $H_0 : S_1(t) = S_2(t)$ for all t, the ratio R would be equal to one if H_0 were true. If the confidence interval for R does not contain one, it can be said there is a significant difference in the pointwise estimates for some values (or at least one) of t. Alternatively, an acceptance region can be created under the null hypothesis, and if the ratio is not within this acceptance region, it can be said there is a difference between the two survival curves.

An alternative to the ratio is to consider the pointwise difference in survival

i.e. $S_1(t) - S_2(t)$. Bootstrap resamples of the data can again be used to create confidence intervals. If the null hypothesis is true, $H_0 : S_1(t) = S_2(t)$, the confidence interval for the difference will contain zero. Harrell has already created a `survdiffplot` function in the `rms` library in r [Harrell Jr, 2001]. An example of Harrell's `survdiffplot` comparing Her2 status is given in `Figure 2.6`. Harrell uses exact asymptotic formulas to calculate the standard errors for approximate confidence bands.

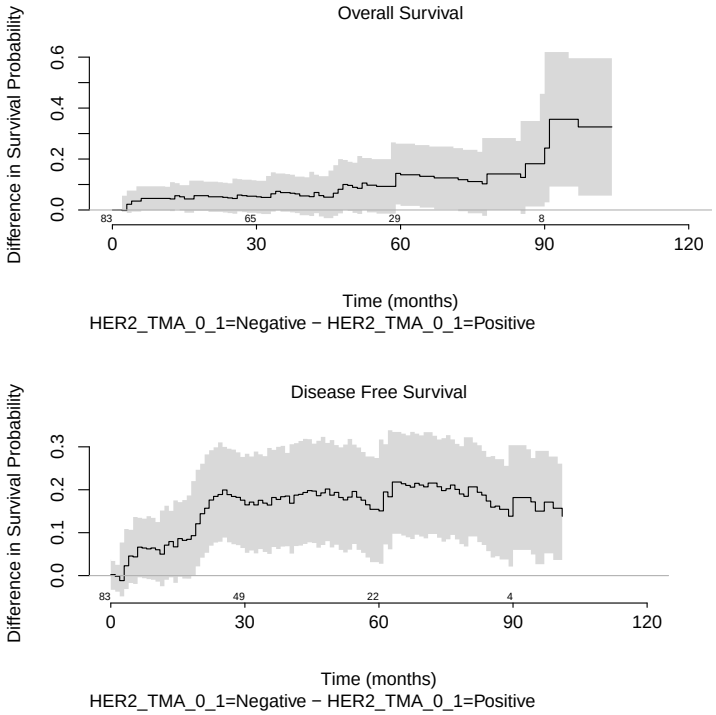


Figure 2.6: Graphical comparison of pointwise differences created by Harrell's `survdiffplot` for Her2 Status.

An extension to these plots proposed in this thesis is to incorporate `alpha blending` in the ratio and difference plot and to highlight areas where a difference may be present. An example for comparing Her2 negative and positive patients is given for OS and DFS in `Figure 2.7`. The top row contains OS and the bottom row contains DFS. The left plots contain the observed difference (black line - which is shaded using `alpha blending`, the more patients at risk the darker the line) and the confidence interval for the difference (red

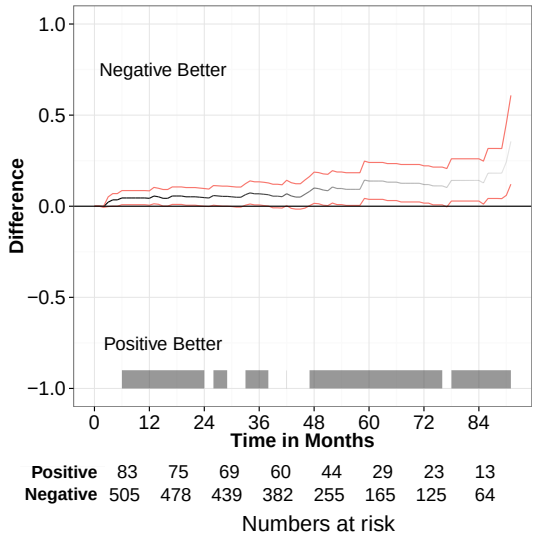
lines). If the confidence interval for the difference does not contain zero (if the null hypothesis was true, the difference between the groups would be zero), it looks as if there is significant difference between groups for some values of t. The grey bar at the bottom of the plot shows the times where the confidence interval does not contain zero. The plots on the right contain the acceptance region for the hypothesis ($S_1(t) = S_2(t)$) and the observed difference (black line). If the observed difference is not contained within the acceptance region, there is a significant difference between the groups. An example for comparing Her2 Negative and Positive patients using the ratio is given for OS and DFS in `Figure 2.8`. The top row contains OS and the bottom row contains DFS. The left plots contain the observed ratio (black line) and the confidence interval around the ratio. If the null hypothesis was true $S_1(t) = S_2(t)$, the ratio of survival estimates would be equal to one. If the confidence interval for the ratio does not contain one for at least one t, there is a significant difference between groups. The plots on the right contain the acceptance region for the hypothesis (about one) and the observed ratio (black line). If the ratio is not contained within the acceptance region, there is a difference between the groups.

The confidence intervals created here are only pointwise intervals and are not a replacement for the Log-rank test. However the graphs give more information and are a complementary graphical tool. The `alpha blending` highlight the sample size diminishing over time and the bars at the bottom of the plot give an indication of regions of time where a difference may occur.

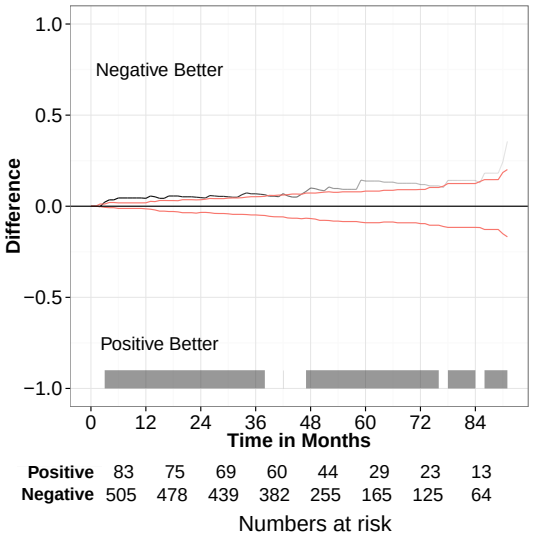
2.4 Conclusions

This chapter has introduced survival analysis. Examples of graphical and numerical summaries using the breast cancer dataset from UCH Galway have been examined.

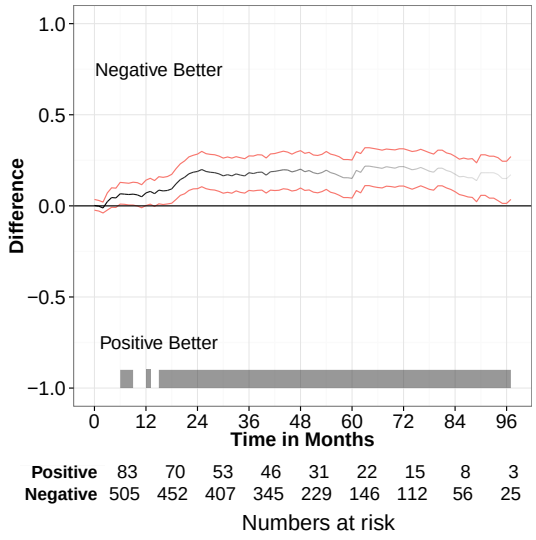
The Kaplan Meier estimate of the survival function is the most common method used for the representation of the distribution of survival times. It is generally used for graphical comparisons of survival experience of two or more groups of patients. The Log-rank test is a hypothesis test for the equality of survival functions.



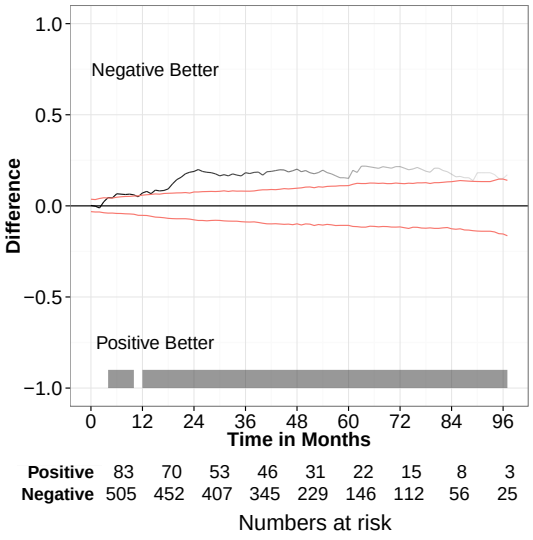
(a) Overall Survival



(b) Overall Survival

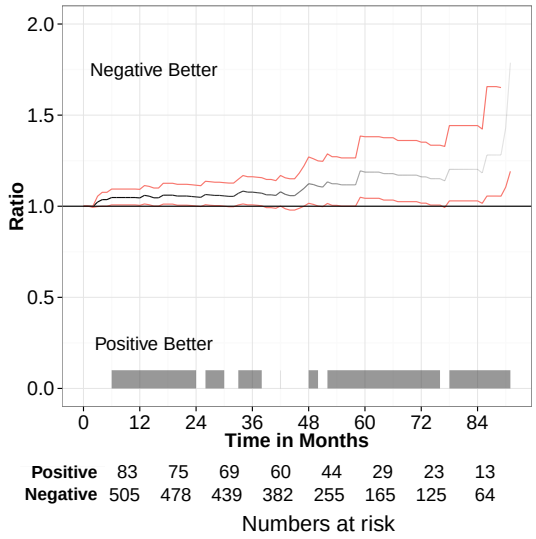


(c) Disease Free Survival

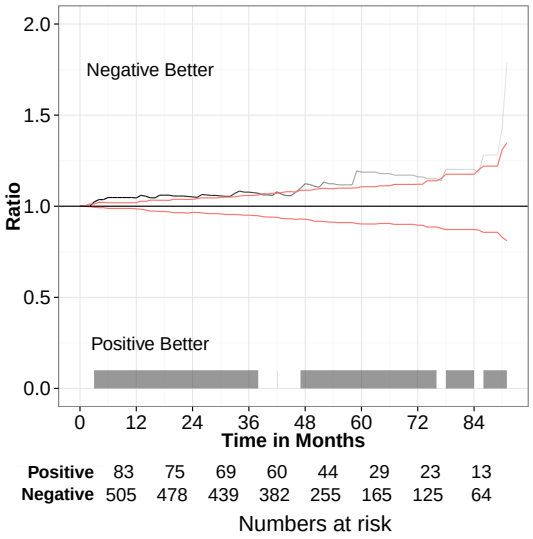


(d) Disease Free Survival

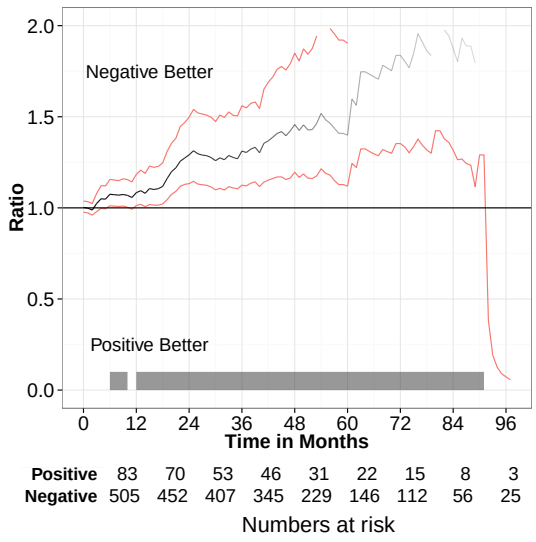
Figure 2.7: Graphical comparison of pointwise differences for Her2 positive and negative patients. Black line observed difference. The red lines in the plots on the left hand side are estimated confidence intervals for the difference. The area between the red lines in the plots on the right hand side are acceptance regions for the null hypothesis.



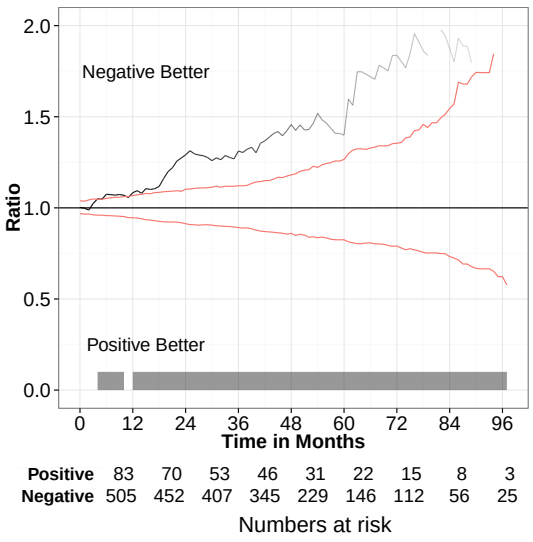
(a) Overall Survival



(b) Overall Survival



(c) Disease Free Survival



(d) Disease Free Survival

Figure 2.8: Graphical comparison of pointwise ratios for Her2 positive and negative patients. Black line observed ratio. The red lines in the plots on the left hand side are estimated confidence intervals for the ratio. The area between the red lines in the plots on the right hand side are acceptance regions for the null hypothesis.

Graphical comparisons of two groups can be a useful tool to help interpret and complement the Log-rank test results. An obvious way is to look at the pointwise difference between the groups but an alternative is to examine the pointwise ratio. In this chapter, examples of these graphical tools and new extensions are shown using the bootstrap to create pointwise confidence intervals and acceptance regions. These also include highlighting the number of patients at risk at each time point using `alpha blending` shading.

In the univariate case, testing each of the predictors using the Log-rank test suggests that the majority of the predictors look like they may be useful for predicting DFS and OS.

In the next chapter, a non-parametric approach using tree based models are examined to help identify useful predictors of the Oncotype DX test, OS and DFS. Classification and Regression Trees (CART) can handle many different types of data and responses including categorical, continuous and survival outcomes. Chapter 3 gives an introduction to tree based approaches and uses the Oncotype DX data and the breast cancer data from UCH Galway to demonstrate their usefulness.

Chapter 3

Tree based Models and Surrogate Splits

3.1 Introduction

Classification and Regression Trees (CART) were first introduced by Breiman et al. [1984] and are a simple non-parametric regression approach used for identifying useful predictors and "structure" in a given dataset. CART is an alternative to classical modelling approaches such as multiple regression, logistic regression or the Cox proportional hazards model.

In this chapter, classical tree based methods (CART - splitting based on dissimilarity), a more recent development based on conditional inference (splitting based on p-values), and extensions to tree based methods, namely forests will be examined. A novel approach using surrogate splits will be presented to help identify alternative trees with comparable prediction power as the 'best' tree. These methods will be applied to the Oncotype DX data to consolidate results of conflicting papers and to the BC data to identify potentially useful predictors.

CART models consist of a hierarchy of univariate binary decisions. These decision trees depict rules for dividing data into groups [Neville, 1999]. Recursive partitioning methods are amongst the most popular and widely used statistical learning tools for non-parametric regression and classification [Strobl et al.,

2009]. The main feature of CART is that data are recursively partitioned into groups. The partition is created such that observations with similar response values are grouped. At each node a constant value of the response variable is then predicted within each group, generally the mean for a continuous response and the level with majority vote for a categorical response (mode), i.e. CART uses a decision tree to represent how the data may be classified or predicted.

CART constructs models which are obtained by recursively partitioning the data by fitting a simple prediction model to each partition [Loh, 2011]. The algorithm creates binary splits on nominal or interval predictors for a nominal, ordinal or interval response. An exhaustive search is made for the split that maximizes the splitting measure.

The key steps for a CART analysis are as follows:

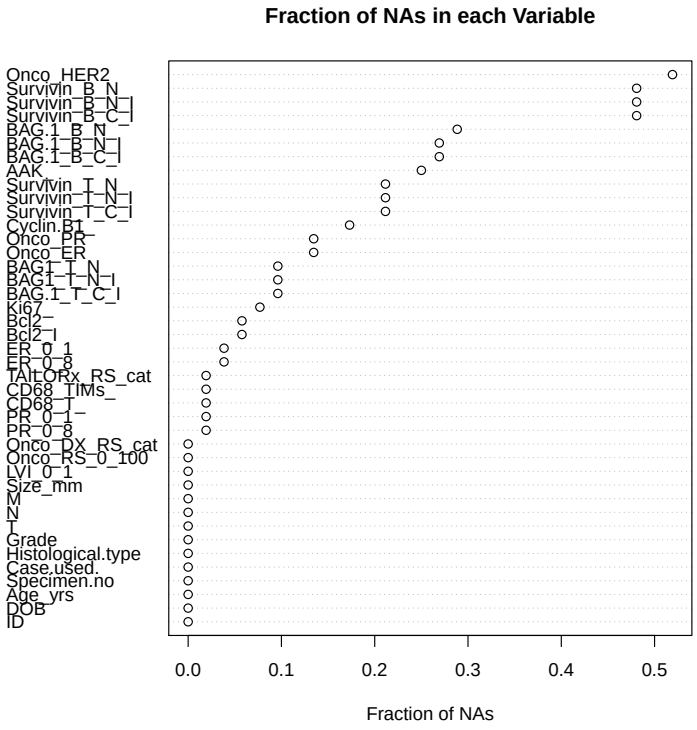
1. splitting each node in the tree;
2. deciding when a tree is 'complete' and
3. assigning each node to a class outcome for a classification tree or a predicted value for regression trees.

The appeal of CART is evident: the results are concise and easy to understand, and are geared towards decision making [Systems, 2001].

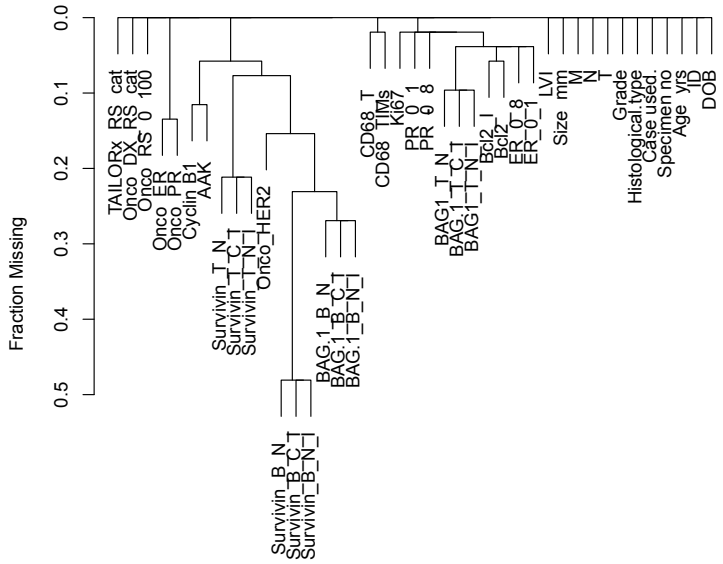
Such methods have been used when analysing data relating to Oncotype DX classification [Allison et al., 2011] and the same approach will be used for the Galway Oncotype DX data.

3.1.1 Oncotype DX data

Recall from Chapter 1, the Galway Oncotype DX data has 52 patients. From `Figure 3.1(a)`, it can be seen that there is a high proportion of missing values present in the Oncotype DX. Some of the variables have up to 50% missing. Examining the patterns in the missing data in `Figure 3.1(b)`, if a patient is missing in one of the *Survivin* variables it is very likely it will be missing in the other *Survivin* variables. Classical approaches such as logistic regression do not work well due to the high proportions of missing data (see results in `Section A.1.1`). Using casewise deletion reduces the sample size from 52 down



(a) Proportions of missingness in each predictor of the Oncotype DX dataset. (NAs missing values)



(b) Cluster analysis showing which predictors tend to be missing on the same patient for the Oncotype DX dataset. Complete-linkage clustering was used to cluster the predictors.

Figure 3.1: Missing values in the Oncotype DX data

to 7. Variable selection cannot be performed on the data if cases with missing values are deleted and this is why tree based models have been used.

3.1.2 Advantages of CART

There are numerous advantages to the predictive models created by CART such as:

1. CART makes no distributional assumptions.
2. The explanatory variables used in CART can be of any form, continuous, interval or categorical.
3. CART uses surrogates splits (these are the next best splits to the best split) to deal with missing data. This means all cases can be included as it uses the surrogate splits to pass the cases down the tree if missing on a particular predictor.
4. CART is not effected by outliers, collinearities, heteroscedasticity, or distributional error structures that affect parametric procedures.
5. CART can detect interactions in a data set and uncover hidden structure in a highly complex data set.
6. Transformations of the data, such as a natural logarithm, does not change the structure of the tree.
7. CART can use the same variable at different branches in the tree where as most analyses only use a variable once unless specified as an interaction.

One limitation of CART is that one cannot force variables into the model. This is a problem when there is a need to control for certain risk factors. Another disadvantage is that some trees may be unstable.

3.1.3 Types of Trees

There are many different trees to deal with different types of responses, for example regression trees, classification trees and survival trees. Regression trees involve response variables that take continuous or ordered discrete values, with prediction error typically measured by the squared difference between

the observed and predicted values. Classification trees are designed for response variables that take a finite number of unordered values, with prediction error measured in terms of misclassification costs. Survival trees have a response with both the survival time and censoring information. The prediction error is measured by a concordance index.

3.2 Classification and Regression trees

Recursive partitioning builds trees by finding the variable that bests splits the data into two groups, this is achieved by searching iteratively over all possible splits and the data in the subgroups are partitioned into two subgroups using this criteria. Next the best split for these two subgroups are found and these subgroups are partitioned. This continues until no improvement can be made in the impurity or a minimum number of individuals is achieved in each terminal node. The impurity of a node is a measure of the misclassification in a node.

3.2.1 Splitting Criterion

A tree is formed by iteratively splitting nodes to minimize an impurity measure, $I(T)$ [Breiman et al., 1984]. The best split s^* of node t is the split in S, where S is the set of all possible splits, which most decreases $I(T)$, where $I(T)$ is the node impurity. More precisely, for any split s of t into t_L and t_R, left and right nodes respectively, let

$$\Delta I(s,t) = I(t) - I(t_L) - I(t_R) \tag{3.1}$$

Take the best split s^* to be a split such that

$$\Delta I(s^*,t) = max_{s\in S}\Delta I(s,t) \tag{3.2}$$

Node impurity is largest when all classes of the response variable are equally mixed together in a node and is smallest when the node contains only one class. The best split is always chosen such that the variable produces the highest reduction in impurity.

The split, i.e. the variable and threshold, that leads to the split that is the

most "pure", that is the most homogeneous groups with respect to the response variable is selected.

After a split is selected, the observations are then divided into each node using the splitting variable and threshold and the splitting of each node continues until some stopping criteria is achieved. Stopping criteria include a minimum number of observations in a terminal node or a given threshold for the minimum change in impurity is not achieved by any variable.

An example of an overfitted survival tree for DFS is given in `Figure 3.2`. The splitting criteria for survival trees is based on the distance between Kaplan Meier survival estimates. Kaplan Meier estimates were introduced in the previous chapter and these give an estimated survival probability for the patients in each terminal node.

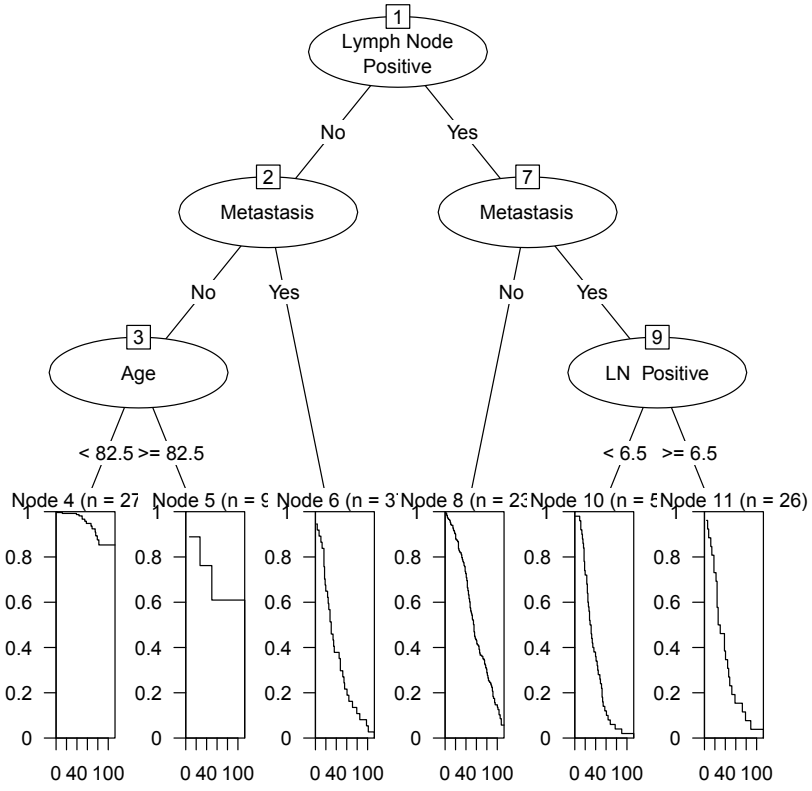


Figure 3.2: Unpruned (over-fitted) recursive partitioning survival tree for DFS for the Galway cohort.

Typically trees may not be grown to the correct size, the results are over

optimistic and the final tree needs to be pruned back as some of the extra splits may only marginally reduce the prediction error. From `Figure 3.2`, terminal nodes 10 and 11 seem very similar and pruning might be required here as the extra split may not reduce the prediction error by much.

3.3 Pruning Procedure

Starting with a large tree T_{max} and selectively pruning the tree to the right size can reduce over-fitting. The branches that do not contribute to the prediction accuracy in the cross-validation are removed. Cross-validation involves partitioning the data into a test and training set. The tree is created using the training set and validated using the test set. This is performed multiple times and the results are averaged to find the 'best' sized tree. The best size of a recursive partitioning tree can be found using information on the optimal pruning based on a complexity parameter calculated using cross validation. The complexity parameter is the complexity cost per terminal node and each value corresponds to a different sized tree. The size of the tree is the number of splits in the tree.

If trees are too small they will not make accurate predictions however if trees are too big, the terminal nodes will not have enough data in them to make any reliable predictions about the contents of the terminal node when the tree is applied to a new data set. If the splitting was carried out until only one case was in each terminal node, then each node is classified by the case it contains, the resubstitution estimate gives a zero misclassification rate. The resubstitution estimate is the probability of misclassification using the training data as the test data. On the other hand, too small a tree will not use some of the classification information available in the data set, resulting in a higher misclassification rate than the right sized tree [Breiman et al., 1984].

The complexity parameter associated with minimum error should be selected. For example, in `Figure 3.3` for DFS, the error is minimised for a tree with 4 splits. However a tree with 3 splits seems adequate since the trade off between the error an the extra split is very little. So the extra splits in this unpruned survival tree are removed to give the pruned survival tree in `Figure`

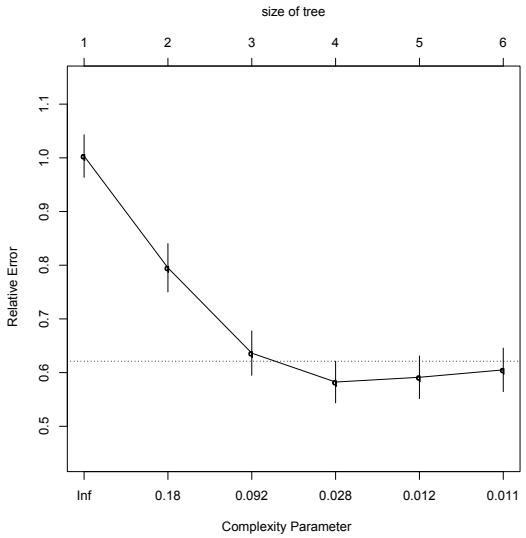
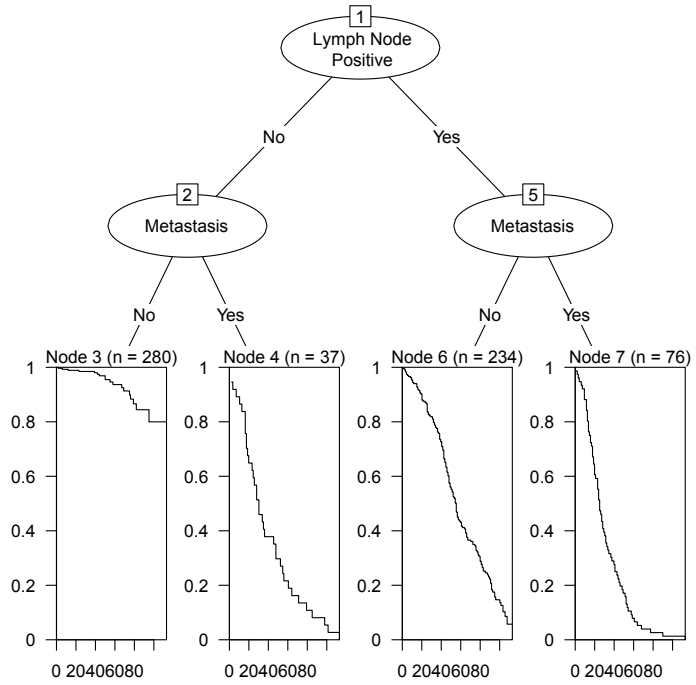


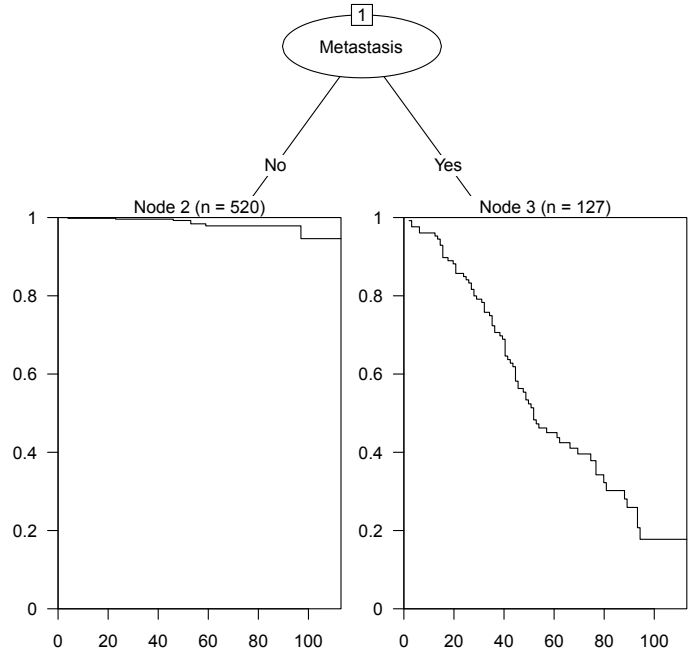
Figure 3.3: Cross validation error for the complexity parameter for the unpruned survival tree for DFS.

`3.4(a)`. This identifies Lymph Node status and Metastasis as useful predictors and may have identified an interaction effect. If a patient has no lymph nodes positive and no metastatic cancer they have the best survival. If a patient has lymph nodes positives and metastatic cancer they have the worse prognosis. The same procedure was applied for OS and the pruned tree `Figure 3.4(b)` has just one split on Metastasis. Patients diagnosed with metastatic cancer have worse prognosis than those without metastatic disease which is as expected.

The same tree growing and pruning procedure was applied to the Oncotype DX classification (`Figure 3.5`). Examining the terminal node on the right hand side of the tree, if a patient has a high progesterone score ($PR\_0\_8 > 1$) and high lymph node grade ($N = 3$) they are classed as high risk. This is what would be expected, as the more positive lymph nodes (i.e. higher lymph node stage), the higher at risk a patient would be. Some of the variables identified previously in the literature as good predictors of Oncotype DX also appear in this tree, such as Progesterone Status and N Stage (lymph node stage). However this tree is a fixed tree and was modeled using a small sample size and there are small numbers in the terminal nodes so it would not be recommended to use this as a predictive model.



(a) Disease Free Survival



(b) Overall Survival

Figure 3.4: Pruned recursive partitioning survival trees for the BC data.

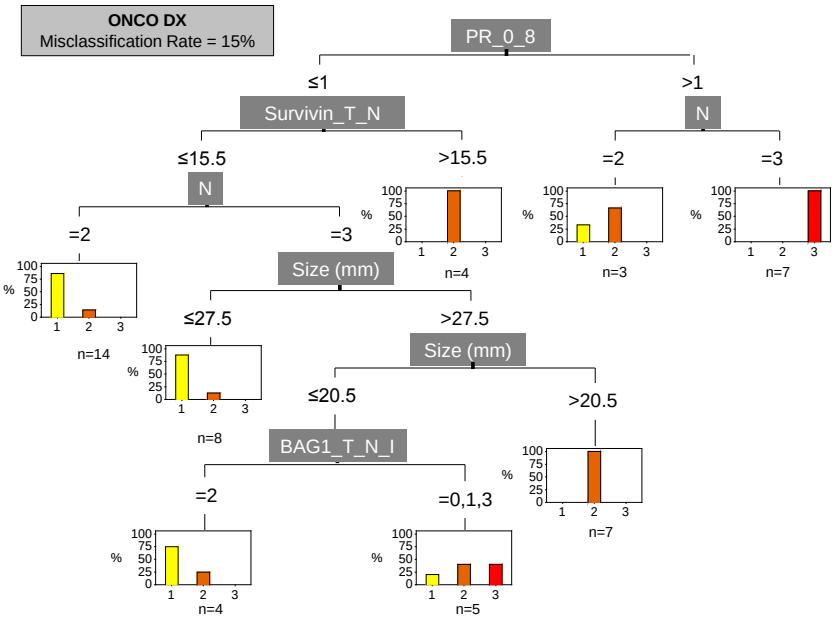


Figure 3.5: Recursive partitioning classification tree for the Galway Oncotype DX patients into low, medium and high risk.

One concern when using tree based methods is when the number of variables is large compared to the number of cases in the variable selection problem. The term variable selection bias refers to the fact that in standard tree algorithms variable selection is biased in favour of predictors offering many potential cut-points so that predictors with many categories and continuous predictors are artificially preferred [Strobl et al., 2009].

3.4 Conditional Inference Trees

An alternative approach for tree building which corrects for the variable selection problem is called conditional inference trees. Such trees are split based on a p-value from a suitable hypothesis tests depending on the context. Since conditional inference trees use statistical stopping rules, pruning is automatic. The significance level can be adjusted for the multiple testing of the splits, which controls the probability of falsely identifying one of the predictors as significant. One such stopping criterion is a simple Bonferroni correction with $\alpha = .05$ is used in Hothorn et al. [2006]. The `party` library in *R* provides

code for regression trees for nominal, ordinal, numeric, censored (survival), and multivariate responses, applying a stopping criterion based on hypothesis tests. These statistically motivated stopping criteria implemented by hypothesis tests lead to trees whose predictive performance is equivalent to the performance of optimally pruned trees [Hothorn et al., 2006].

Here is a summary of the algorithm:

1. Test the overall null hypothesis that the response variable is independent of the predictors; stop growing the tree if the null hypothesis cannot be rejected. Alternatively, choose the predictor with the strongest association to the response variable (i.e. the smallest p-value).

2. Split the data into the two subgroups using this split criteria for the selected predictor.

3. Repeat steps 1 and 2 until the overall null hypothesis of the predictors being independent of the response variable cannot be rejected.

The significant tests are used for selecting the best split at a node but also as stopping criteria. Once a p-value for a split exceeds the significance level the tree is not grown any further.

One important improvement with the introduction of conditional inference trees was that they overcame the variable selection problem of classical approaches. An advantage of conditional inference is the tree is automatically pruned. However, the tree may miss interactions. For example, the tree may not split the data because there is no significant split overall however if an interaction was present, a split on a certain predictor further down the tree may be significant.

The conditional inference tree procedures were applied to the Onoctype DX classification (`Figure 3.6`). It has a higher misclassification rate than that of the pruned classification tree (`Figure 3.5`) and has identified other potentially useful predictors. It contains predictors similar to that of the paper published by Allison et al. [2011].

Examples of survival trees via conditional inference for the BC data are given in `Figure 3.7`. For DFS, Metastasis and Lymph Node Status are selected which are also in the recursive partitioning tree (`Figure 3.4`). However, the

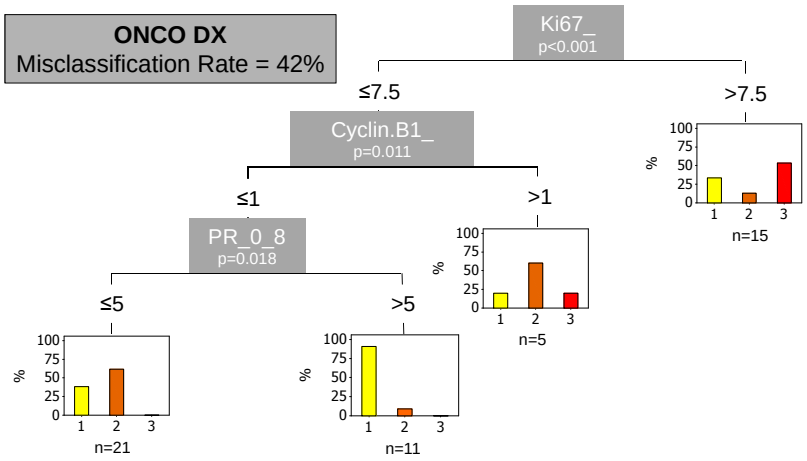


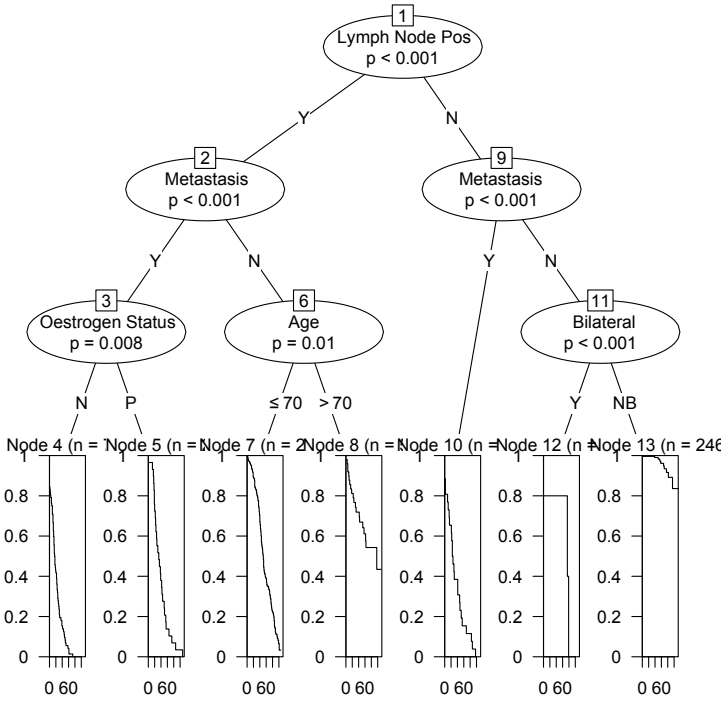
Figure 3.6: Conditional inference classification tree for the Oncotype DX risk.

conditional inference tree also identified extra splits. For OS, the first split in the conditional inference tree is Metastasis which is the same as the recursive partitioning tree, however the conditional inference tree identified extra splits, Oestrogen status and UICC staging.

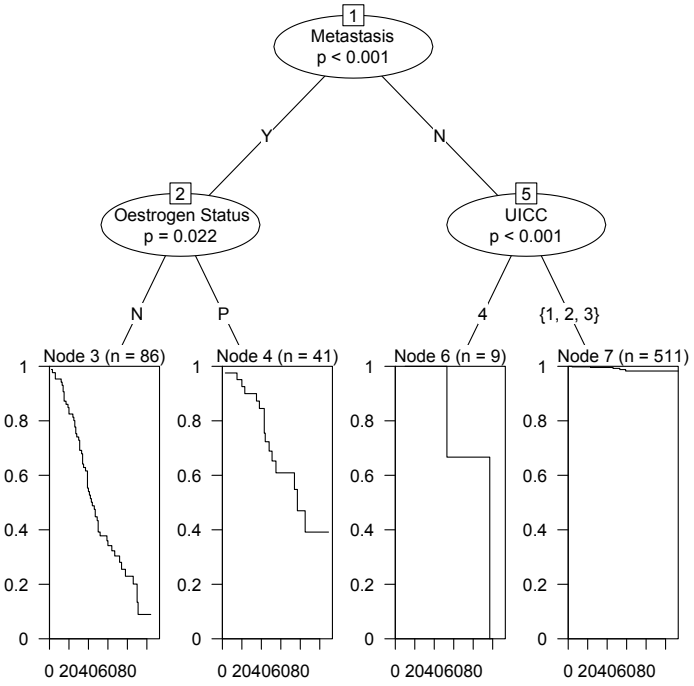
3.5 Random Forests

Random Forests, an extension to CART, are an ensemble method involving bootstrap resampling of the data [Breiman, 2001]. Random Forests work by growing a large number of trees from resampled data and prediction is obtained by averaging over all the predicted responses for each tree. In addition to constructing each tree using a different bootstrap resample of the data, each node is split using the best split from a subset of predictors randomly chosen at that node, instead of using the best split from all the predictors. If there is a particularly good predictor which is always chosen as the split, choosing a subset of predictors at each node, may help in identify other potentially useful splits hidden when using the 'best' split. The trees grown are not pruned.

Random forests can handle problems with a small sample size and large number of variables very well since it examines a subset of predictors at each node. The higher the number of trees in the forest the more reliable the prediction



(a) Disease Free Survival



(b) Overall Survival

Figure 3.7: Conditional Inference survival tree for DFS for the Galway cohort.

and interpretability of the variable importance.

The major advantage of random forests is the reduction in prediction error. However, the one disadvantage is there is no tree structure. The variable importance measure from random forests can be used to identify important predictors for the model. Large importance values indicate variables with predictive ability, whereas zero or negative values identify non-predictive variables to be filtered [Ishwaran et al., 2008].

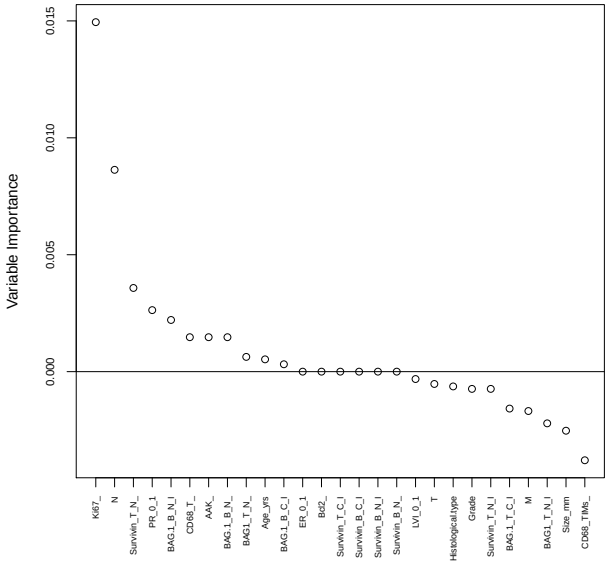
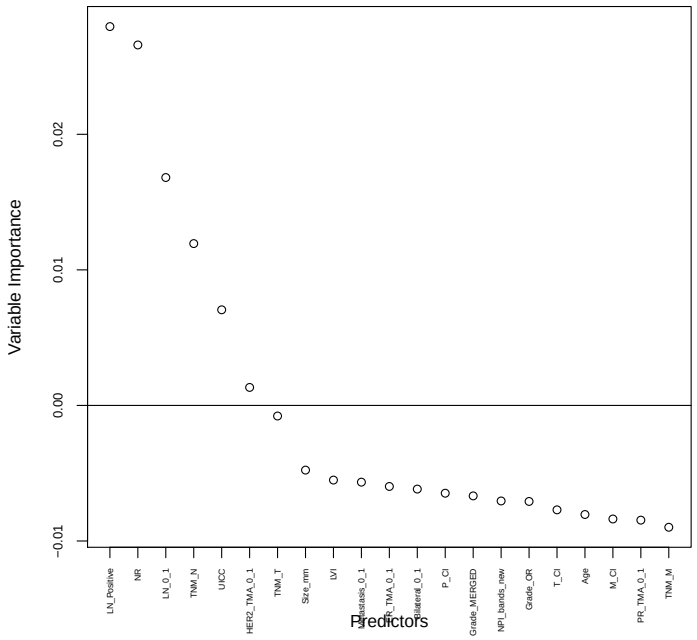


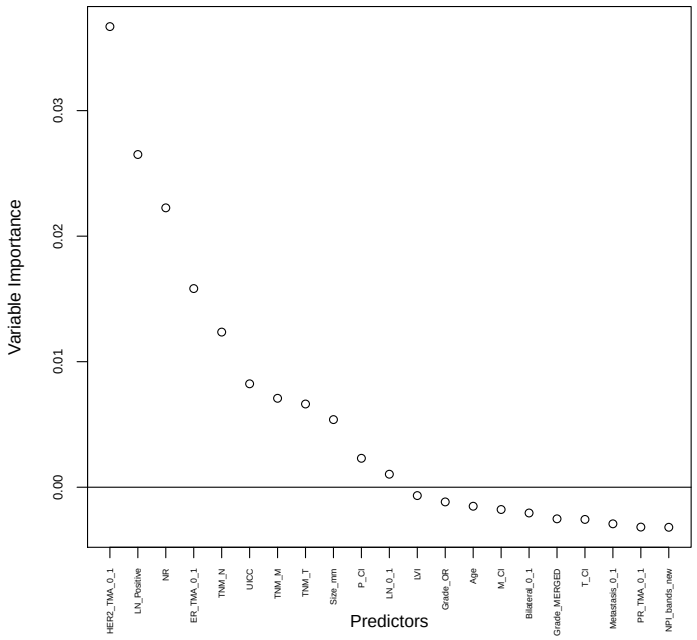
Figure 3.8: Variable importance measure for the Oncotype DX data using `cforest` in the `party` package.

The variable importance measure for the Oncotype DX data (`Figure 3.8`) identified Ki67, Lymph Node stage, Survivin and Progesterone status as being potentially useful predictors.

The `randomForest` library and the `cforest` in the `party` library in *R* do not have the capability to create a random forest or the variable importance measure for censored survival data. However, a special library, called `randomForestSRC`, is available to handle right censored survival data [Ishwaran et al., 2008]. The variable importance measure for DFS (`Figure 3.10(a)`) identified many of the lymph node predictors as useful, such as number of positive nodes, nodal ratio,



(a) Disease Free Survival



(b) Overall Survival

Figure 3.9: Variable importance measure for OS using the Random Forest package for survival tree.

status and stage. It also identified UICC staging as useful. For OS (`Figure 3.10(b)`) the variable importance measure identified again a few of the lymph node predictors such as number of positive nodes, nodal ratio and stage. Also some of the hormones such a Her2 and oestrogen status and the tumour, metastasis and UICC stages.

The next section, will introduce a novel way of identifying structure in the data using competing trees built using surrogate splits.

3.6 Surrogate Splits

At any given node in a tree, the best split s^* is chosen (i.e. the split which decreases the impurity most). A surrogate split is a split which most accurately predicts the action of s^*. There are two types of surrogates, primary and secondary. Primary surrogates are the splits with a similar performance in impurity (predictive performance) to the best split. Secondary surrogate splits resemble the best split in terms of the number of cases they send the "same way" (similar partition of units/cases) and are typically used to handle missing data.

A surrogate split is a splitting rule that closely mimics the action of the primary split. Not only must a good surrogate split the parent node into descendant nodes similar in size and composition to the primary descendant nodes, but, to the extent possible, the surrogate must also match the primary split on the specific cases that go to the left child and right child nodes. A surrogate is thus evaluated on its ability to match the primary split on a case-by-case basis. If no variable can mimic the primary splitting criterion, there will not be any surrogates listed [Systems, 2001]. The CART decision tree algorithm can use surrogate features to rank individual features by their importance [Springer and Kegelmeyer, 2008].

In Breiman's Classification and Regression Trees [Breiman et al., 1984], a surrogate is defined as follows. Let S_m be the set of all splits on x_m, any variable, and $\bar{S}_m$ the set of all splits in the complement to S_m. For any split $s_m \in S_m \cap \bar{S}_m$ of the node t into t'_L and t'_R, let $N_j(LL)$ be the number of cases in t that both s^* and s_m send left, that is, that go into $t_L \cap t'_L$. By the usual

procedure, the probability that a case falls into $t_L \cap t'_L$ is estimated as

$$p(t_L \cap t'_L) = \sum_j \pi(j) N_j(LL)/N_j \tag{3.3}$$

where j is the predicted class and $\pi(j)$ are defined as the prior class probabilities.

The estimated probability is defined as $p_{LL}(s^*, s_m)$ that both s^* and s_m send a case in t left as

$$p_{LL}(s^*, s_m) = p(t_L \cap t'_L)/p(t). \tag{3.4}$$

Similarly, define $p_{RR}(s^*, s_m)$. The probability that s_m predicts s^* correctly is estimated by

$$p(s^*, s_m) = p_{LL}(s^*, s_m) + p_{RR}(s^*, s_m). \tag{3.5}$$

The best surrogate split is defined to be

$$p(s^*, \tilde{s}_m) = \max_{s_m} p(s^*, s_m). \tag{3.6}$$

3.6.1 Traditional uses of surrogates

Missing Data

Missing values are usually handled by using casewise deletion in prediction modelling. If values are missing for the response variable, the only viable strategy is case-wise deletion [Berk, 2008]. If values are missing from the predictors there are several options to deal with the missing values. First is casewise deletion, where if a patient is missing on any predictor they are removed, however this results in a reduction in sample size and power. Secondly the data could be imputed using some regression equation such as MICE (Multivariate Imputation by Chained Equations). Finally, using tree based models, surrogate splits can be used to pass cases with missing values down a tree.

Secondary surrogate splits are used to assign missing values to the correct node. It works as follows: suppose that the best split s^* has been found for a node. The cases are split using this best split if they have a value for the variable. However, if there are missing values present in the variable, these cases are split using the next best secondary surrogate split to s^*.

Variable Ranking

The critical issue is how to rank those predictors that, while not giving the best split of a node, may give the second or third best [Breiman et al., 1984]. For example, a variable may not appear in the tree but if the first split variable is removed from the model, the variable may appear and have a similar tree structure or performance.

Another use for surrogate splits is in identifying variable importance. For example, if x_1 and x_2 are predictors, x_1 may not appear anywhere in a tree, but if variable x_2 is removed and another tree is grown, x_1 may occur often in this second tree. This could be as accurate as the first tree and in this situation variable ranking would be required to identify x_1 as important.

3.6.2 Novel use of Surrogates

Surrogates could be used to create other trees that have comparable predictive performance as the 'best' tree. Obviously when a tree is created, it uses the best split as the splitting criteria. However, the surrogates could be used to create alternative trees, by selecting a surrogate and growing a tree using this as the first split.

This results in several trees, which is similar to Best Subsets in regression modelling. Best Subsets involves examining all of the models created from all possible combination of the predictors, so there are many possible models and the best one is chosen based on it's predictive ability. The trees created using the surrogate splits may have competing predictive ability of that of the original tree.

Here we propose a novel approach of using surrogate splits to identify underlying structure in the data (i.e. a set of potentially useful predictors).

3.7 Surrogate Plot

As the Oncotype DX data has a small sample size and large number of predictors, it would not be recommended to use the tree (`Figure 3.10`) as a prediction model. There may be more underlying structure if we look at the surrogate splits. Surrogates are useful for identifying variables, which may not appear

in the original tree structure, but if the first split is removed from the model, the variable that had not previously been included may now appear and have a similar tree structure. The idea of an interactive surrogate plot is to help identify underlying structure using surrogate splits. `Figure 3.10` contains the classification tree for the Oncotype DX Galway data, however we have also identified some primary surrogate splits at the first few nodes. To represent the information about the primary and secondary surrogate splits, I developed an interactive surrogate plot package in *R*. Instructions how to implement this code is given in the `Appendix A.1.3`.

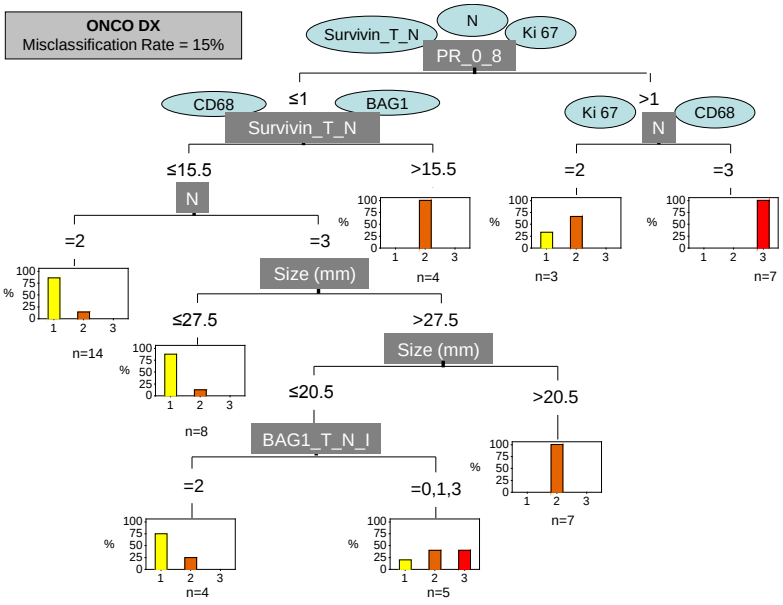


Figure 3.10: Recursive partitioning classification tree for the Galway Oncotype DX patients into low, medium and high risk. Some primary surrogates are identified as each of the nodes.

An example of a surrogate plot applied to the Oncotype DX data is given in `Figure 3.11`. This interactive plot was created using the `tcl-tk` library in R [Grosjean, 2012]. The x-axis is divided into three parts, nodes (the number of times the variable appears as a split in the original tree), primary surrogates and secondary surrogates for the different nodes in the original tree. All the variables in the dataset are given along the y-axis and have been ordered in terms of importance, with the most important risk factors close to the top and

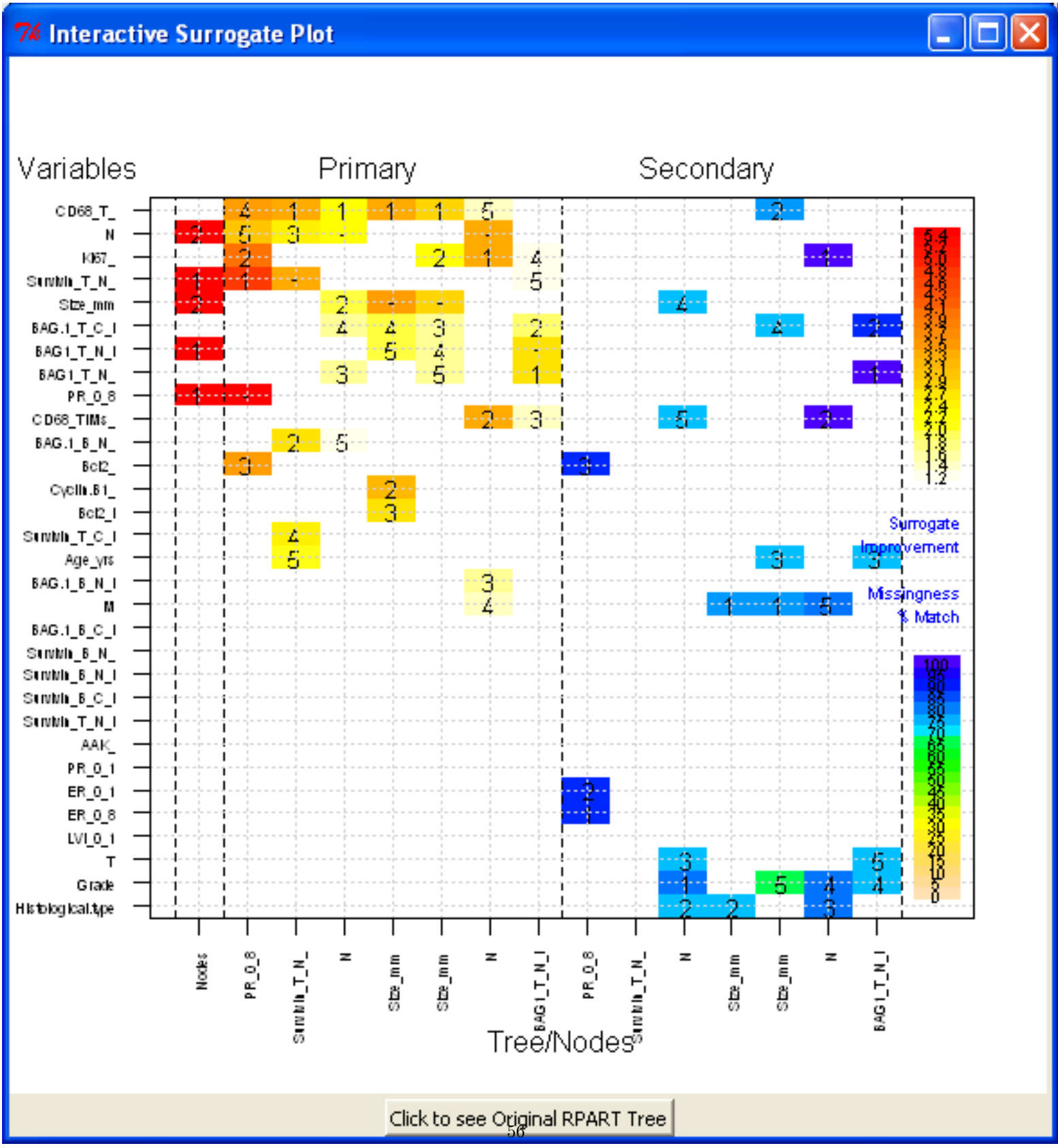


Figure 3.11: Surrogate Plot for the Classification tree of the Galway Oncotype DX classification.

the risk factors which do not seem to be as important closer to the bottom. Five primary surrogates were identified for each of the nodes in the tree. A heat-map is used to compare the predictive power of each surrogate to the best split.

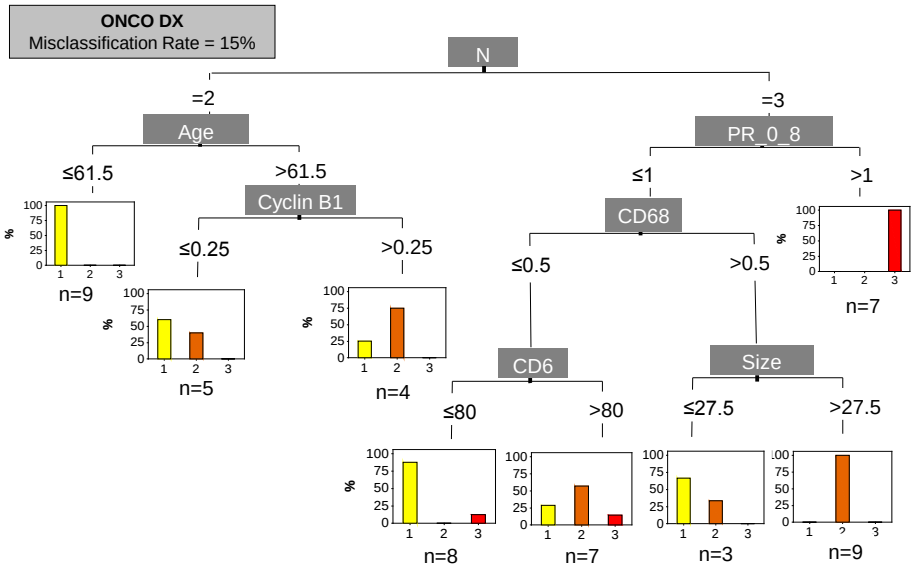


Figure 3.12: Tree for Surrogate N for Oncotype DX classification.

For example, for the first split in the Galway Oncotype DX tree in `Figure 3.10` is *Progesterone Score* $(0-8)$, however the surrogates for this split include *Survivin*, *Ki*67, *BcI*2, *CD*68 and *N* (Lymph node staging). If one of these surrogates is selected, lets say N (lymph node staging), a tree can be created using this surrogate as the first split. The resultant tree for this surrogate is given in `Figure 3.12`. The first split is on the surrogate selected, N, then a tree is grown using all the predictors for each partition of the parent node. This tree has a misclassification rate of 15%, the same as the original tree, and contains some of the predictors identified as useful predictors of Oncotype DX in the paper by Geradts et al. [2010]. If a tree is created for surrogate CD68, similar risk factors identified by Auerbach et al. [2010] appear in the tree. These trees have identified more potentially useful variables and biomarkers but they have comparable predictive power to that of the original tree.

This analysis identified risk factors using surrogate splits which consolidated

results from previous studies of Oncotype DX classification. Using the tree in `Figure 3.5` as a prediction model on an individual level is not very sensible (due to small sample size). However, these risk factors can now be used on a much larger sample to build a better prediction rule.

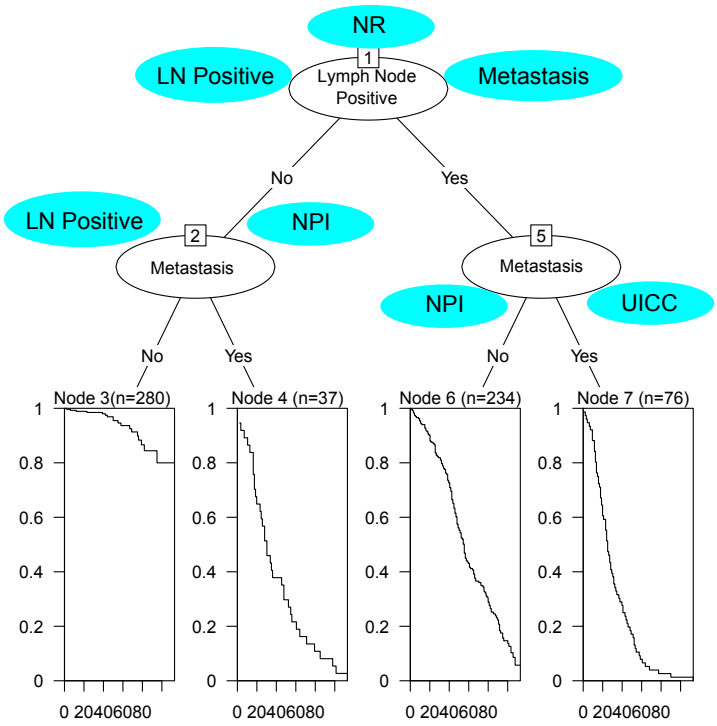
Similarly, the surrogates can be used to identify other structure in the Breast Cancer data. In the pruned DFS survival tree, `Figure 3.13(a)`, other lymph node predictors were identified as alternatives to the best first split of the tree, lymph node status. The surrogate tree using Nodal Ratio as the first split is given in `Figure 3.14`. It identified Metastasis as the split again on the left hand side and Bilateral on the right hand side. For OS, `Figure 3.13(b)`, metastasis was identified as the only useful split, however, if this predictor was removed, nodal ratio, number of positive lymph nodes or UICC staging would be useful to create other comparable trees.

3.8 Conclusions

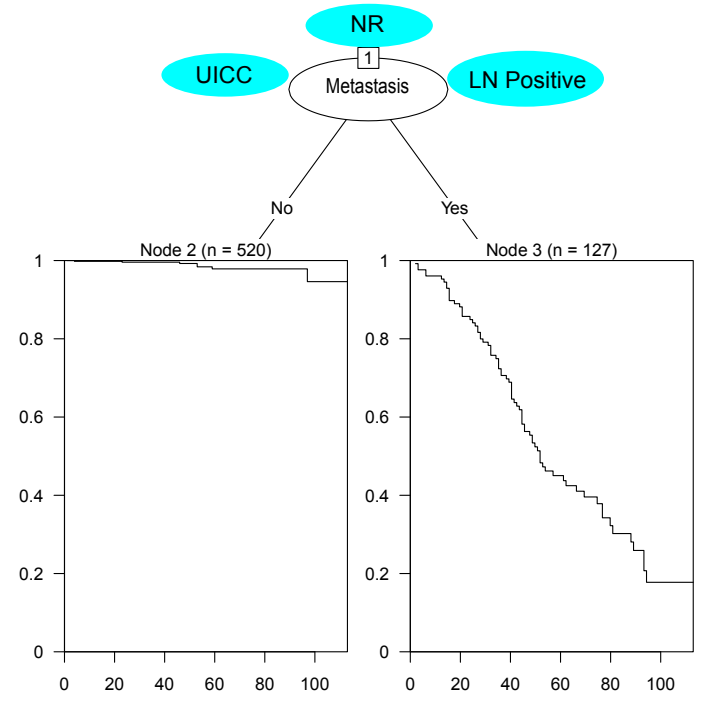
Tree based methods are useful alternatives to classical approaches of modelling. Both CART and conditional inference trees have identified useful predictors in both the Oncotype DX and BC data. CART grows a large tree and then prunes back but this overfitting identifies more structure in the data. Conditional inference has automatic pruning however it may miss interactions. Extensions to trees include bootstrapping in random forests, which is good for prediction however no structure is identified. Random Forests use a variable importance measure for identifying potentially useful predictors.

The novel way of examining the surrogate splits may be the equivalent to Best Subsets in regression modeling, as it results in many different trees with comparable predictive power. An analysis of the surrogates can be useful to identify underlying structure in the data. This novel approach for identifying structure has been presented at the International Workshop for Statistical Modelling [Wall et al., 2012].

Examining the different techniques for the Oncotype DX data, the recursive partitioning tree identified *Progesterone*, *Survivin*, N (Lymph Node Stage), *Size* and $Bag1$ as good predictors for classification into low, medium and high



(a) Disease Free Survival



(b) Overall Survival

Figure 3.13: Pruned survival tree for DFS with some of the surrogates for each node.

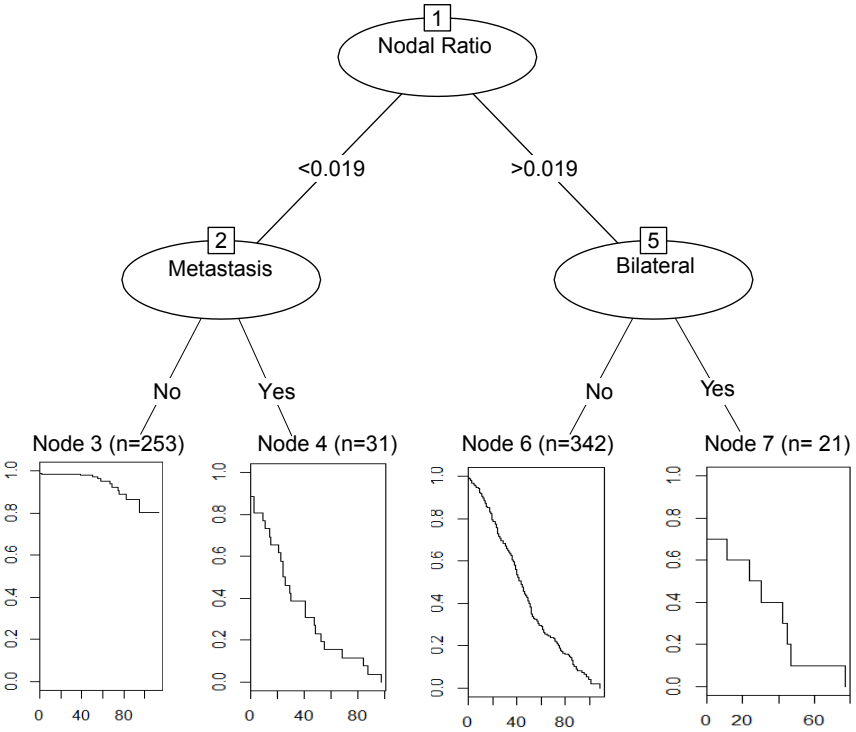


Figure 3.14: Tree for Surrogate Nodal Ratio for DFS in the BC data.

risk. However, surrogate splits can be used to create other trees with comparable prediction power. The surrogate plot identifies other comparable and competing trees which in turn identifies other potentially useful variables and underlying structure but with comparable predictive power. These surrogates were used to create competing trees which consolidated the results from the previous research of Oncotype DX classification.

Many different predictors have been identified as good predictors of both OS and DFS (summarized in `Tables 3.1-3.4`). The predictive ability of the models are measured using the concordance index (also known as the c-index). This is a measure of the probability of agreement between the predicted and observed survival. The concordance is 0.5 for random predictions and 1 for perfectly discriminating model. This will be discussed in more detail in Chapter 6.

Pruned recursive partitioning trees identified Lymph Node status and Metastasis as good predictors of DFS and just Metastasis as a good predictor of OS. The conditional inference trees identified LN status, Metastasis, Oestrogen status and Age as good predictors of DFS and it identified Metastasis, Oestrogen status and UICC staging as good predictors of OS. The random survival forests identified several of the lymph node variables such as Nodal Ratio, Lymph Node status, number of positive lymph nodes and N (Lymph Node staging) as good predictors of DFS as well as Her2 status and UICC staging. The random survival forests again identified several of the Lymph Node variables as good predictors of OS as well as the size of the tumour, Lymphovascular invasion, P (Nuclear Pleomorphism), Oestrogen status, Her2 status, T, N & M staging (Tumour, Lymph Node and Metastasis Staging) and UICC staging. Even though there are different combinations of predictors identified by the CART techniques, there are a few predictors that appear across all CART techniques, such as Metastasis and LN status for DFS and Metastasis and ER status for OS.

In the following chapter, a classical approach, the Cox proportional hazards model, for modelling survival data will be introduced. Variable selection techniques will be investigated to find a set of potentially useful predictors.

Predictors	CART	Conditional Inference	Random Forests
Age		X	
Bilateral		X	
Grade			
Tubular Formation			
Nuclear Pleomorphism			
Mitotic Count			
Tumour Size			
Lymphovascular Invasion			
No of Lymph Nodes positive			X
Lymph Node status	X	X	X
Nodal Ratio			
Metastasis	X	X	
Tumour Staging			
Metastasis Staging			
Lymph Node Staging			
UICC			X
NPI			
Oestrogen Status		X	
Progesterone Status			
Her2 Status			X
Concordance Index	0.826	0.762	0.777

Table 3.1: Summary of clinical predictors selected through various tree techniques for DFS. X means the predictor is included in the model.

Predictors	CART	Conditional Inference	Random Forests
Age		X	
Bilateral		X	
Grade			
Tubular Formation			
Nuclear Pleomorphism			
Mitotic Count			
Tumour Size			X
Lymphovascular Invasion			
No of Lymph Nodes positive			X
Lymph Node status	X	X	X
Nodal Ratio			X
Metastasis	X	X	
Tumour Staging			
Metastasis Staging			
Lymph Node Staging			X
UICC			
NPI			
Oestrogen Status		X	
Progesterone Status			
Her2 Status			
Bcl2 Status			
CK14 Status			
CK5/6 Status			
EGFR Status			
Ki67 Status			
p53 Status			
E-cad Status			
tMcm2 Status			
CDC7			
pMcm2			
Concordance Index	0.762	0.777	0.765

Table 3.2: Summary of clinical and pathological predictors selected through various tree techniques for DFS. X means the predictor is included in the model.

Predictors	CART	Conditional Inference	Random Forests
Age			
Bilateral			
Grade			
Tubular Formation			
Nuclear Pleomorphism			X
Mitotic Count			
Tumour Size			X
Lymphovascular Invasion			
No of Lymph Nodes positive			X
Lymph Node status	X		X
Nodal Ratio			X
Metastasis	X	X	
Tumour Staging			X
Metastasis Staging			X
Lymph Node Staging			X
UICC		X	X
NPI			
Oestrogen Status		X	X
Progesterone Status			
Her2 Status			X
Concordance Index	0.909	0.922	0.911

Table 3.3: Summary of clinical predictors selected through various tree techniques for OS. X means the predictor is included in the model.

Predictors	CART	Conditional Inference	Random Forests
Age			
Bilateral			X
Grade			
Tubular Formation			X
Nuclear Pleomorphism			X
Mitotic Count			
Tumour Size			X
Lymphovascular Invasion			X
No of Lymph Nodes positive			X
Lymph Node status	X		X
Nodal Ratio			X
Metastasis	X	X	X
Tumour Staging			
Metastasis Staging			X
Lymph Node Staging			X
UICC		X	X
NPI			X
Oestrogen Status		X	X
Progesterone Status			X
Her2 Status			X
Bcl2 Status			X
CK14 Status			X
CK5/6 Status			X
EGFR Status			
Ki67 Status			
p53 Status			X
E-cad Status			X
tMcm2 Status			X
CDC7			
pMcm2			
Concordance Index	0.909	0.922	0.790

Table 3.4: Summary of clinical predictors selected through various tree techniques for OS. X means the predictor is included in the model.

Chapter 4

Classical Approaches to Modelling Survival Data

4.1 Introduction

In previous chapters, some non parametric approaches have been examined for variable selection. To develop models for survival data in a population, a simple way of describing the variation in survival among individuals is needed. A popular model is to consider the individual specific hazard function $h_i(t)$ and to make a proportional hazards assumption such as

$$h_i(t) = c_i h_0(t) \tag{4.1}$$

where c_i is a constant and $h_0(t)$ is the baseline hazard left unspecified. The effects of covariates on the hazard can be modelled by letting $c_i = exp(\beta' X)$, this is the Cox proportional hazards model.

Classical approaches for modelling survival data typically assume some underlying parametric distribution or alternatively a flexible semi-parametric approach such as the Cox proportional hazards could be used.

It is important for any prediction model that the model is as simple as possible but as complex as necessary.

An alternative to creating a model with all predictors is to use a smaller

subset of predictors that perform just as well as the full model. This is more cost and time effective as less predictors need to be measured and recorded. Variable selection techniques will be applied to the BC data to find the most parsimonious set of predictors. Classical approaches such as backward variable selection and new shrinkage methods such as Ridge Regression and the LASSO will also be applied.

Non-linear effects in the model may add more complexity so interactions and the use of splines to relax the linearity assumption will be investigated.

4.2 Parametric Models for Survival Data

A parametric model is one in which survival time follows a known distribution and its functional form is completely specified, except for the values of the unknown parameters [Kleinbaum and Klein, 2005]. Parametric survival models may be useful for predictive purposes because of their parsimony and robustness, for example at the end of follow up, or even beyond the observed follow up [Steyerberg, 2009].

The more complex parametric models have more free parameters which can be used to incorporate different survival patterns for early and late mortality. These tend to produce more precise estimates than non-parametric or semi-parametric analysis if the 'correct' distribution is specified.

Two of the most common distributions used to model survival data are described in the next few sections.

4.2.1 Exponential

The exponential model is the simplest parametric survival model with only one parameter. As shown previously, there is a relationship between the survival and the hazard function. For the exponential distribution they are given as $\exp(-\lambda t)$ and λ, respectively (`Table 4.1`).

The exponential distribution is a special case of the Weibull distribution which is discussed in the next section.

Distribution	$S(t)$	$h(t)$
Exponential	$\exp(-\lambda t)$	λ
Weibull	$\exp(-\lambda t^p)$	$\lambda p t^{p-1}$

Table 4.1: Parameters for Parametric Models.

4.2.2 Weibull

The Weibull distribution is the most commonly used parametric model. The hazard is given by $\lambda p t^{p-1}$. λ is the scale parameter and p is the shape parameter (where p and λ are greater than zero). When p is equal to one the Weibull distribution simplifies to the exponential distribution. If p is greater than one the hazard increases as the time increases and if p is less than one the hazard decreases over time.

The confidence limits for a parametric survival function are narrower than for the corresponding non-parametric function; greater precision has been obtained at a cost of having to make assumptions which may be unattainable and may lead to additional bias [Bull and Spiegelhalter, 1997].

4.3 Cox Proportional Hazards Model

The most popular regression model for time to event data is the semi-parametric Cox proportional hazards model.

The Cox Proportional Hazards (CPH) regression model is a semi-parametric procedure; a parametric model for the hazard is combined with a non-parametric estimate of the underlying hazard. The Cox model is a model that provides simultaneous estimates of hazard ratios while adjusting for multiple explanatory variables and is used extensively in the analysis of survival data. The Cox regression model provides a default framework for prediction of long-term prognostic outcomes [Steyerberg, 2009].

The use of the model includes assessing treatment effects in studies and in particular adjusting these comparisons for baseline characteristics [Machin et al., 2006]. The Cox model is given by the following formula:

$$h(t, \mathbf{X}) = h_0(t) \times \exp\left(\sum_{i=1}^{p} \beta_i X_i\right) \tag{4.2}$$

where $h_0(t)$ is the baseline hazard function that determines the shape of the survival function, $\mathbf{X}$ are the predictor variables and the $\beta's$ are the regression coefficients. Note: there is no constant term in the model, it is 'absorbed' into the baseline hazard. The model can also be written as

$$S(t|X) = S_0(t)^{exp(\beta' X)} \tag{4.3}$$

where $exp(\beta' X)$ is called the prognostic index. It is common practice not to define a parametric model for the baseline hazard in a similar manner with the practice of displaying the time estimate of the survival function.

An important feature of this formula, which concerns the proportional hazards (PH) assumption, is that the baseline hazard is a function of t, but does not involve the X's. An appealing property of the Cox model is that, even though the baseline hazard part of the model is unspecified, it is still possible to estimate the $\beta's$ in the exponential part of the model. The Cox model is widely used since it can approximate the results for the correct parametric model. The Cox model does not assume anything about the shape of the survival function $S(t)$ across t for an individual, but it does consider how survival estimates for different subjects are related. Specifically, it assumes that $log[-log(S(t))]$ for different subjects are equidistant over time, or equivalently that hazard functions for any two subjects are proportional over time. This proportional hazards assumption can be checked using smoothed plots of a special type of residual from the model called the Schoenfeld residual [Harrell Jr, 1996].

The hazard at time t is related to the probability that the event will occur in a small interval around t, given the event hasn't occurred before time t. The $h(t)$ part of $h(t|\mathbf{X})$ is sometimes called an underlying hazard function or hazard function for a standard subject, which is a subject with $\mathbf{X}\beta = 0$.

Some assumptions of the Cox PH model include:

- The true form of the underlying functions ($h(t)$, $H(t)$ *and* $S(t)$) should be specified correctly;
- The relationship between the predictors and log hazard or cumulative hazard should be linear in its simplest form;
- The way in which the predictors influence the distribution of the response

will be by multiplying the hazard by $exp(\mathbf{X}\beta)$ or equivalently by adding $\mathbf{X}\beta$ to the log hazard or log cumulative hazard at each t. The effect of the predictors is assumed to be the same at all values of t since $log\ h(t)$ can be separated from $\mathbf{X}\beta$. In other words, the PH assumption implies no t by predictor interaction [Harrell Jr, 2001].

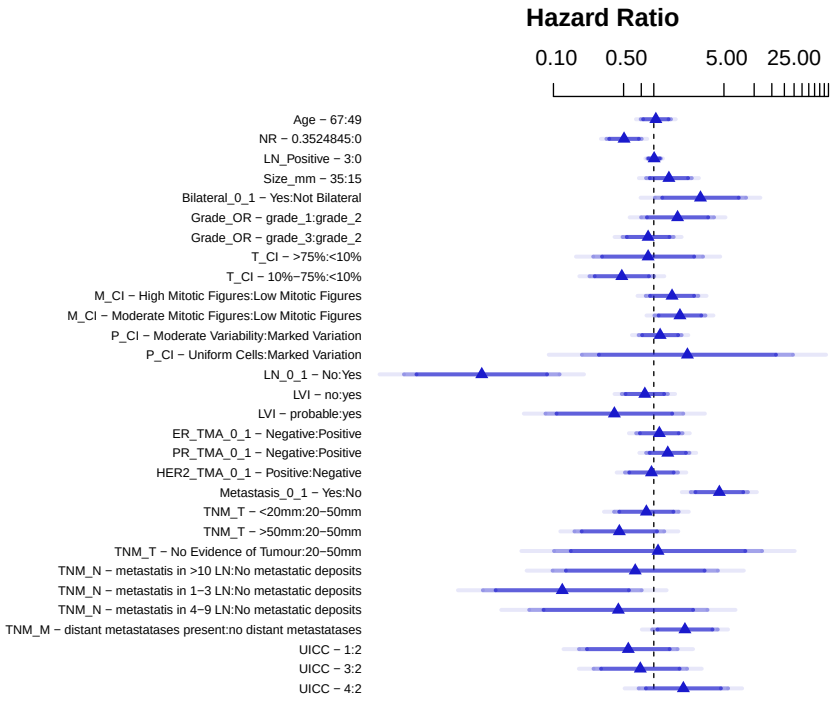
A multivariable survival model for time until the event would be more beneficial than just looking at one predictor (as in Chapter 2) with the comparison of survival estimates marginally for each explanatory variable. Examining a model with all the routinely assessed clinical predictors the sample size of 647 patients is reduced to 221 due to the missing data present, see `Table 4.2`. The estimates are given in terms of the coefficients. Typically these are expressed as $exp(\hat{\beta})$ which are the hazard ratios. For categorical predictors, this is a comparison to the baseline level. For continuous predictors this is a multiple effect per unit increase. For example, for the model for DFS in `Table 4.2`, the coefficient for Bilateral category Yes is 1.071. The hazard ratio is the exponential of this value which is 2.918, this means those patients with Bilateral breast cancer have a worse prognosis than those without Bilateral breast cancer. Similarly with the size of the tumour, as the size of the tumour increases the patient has a worse prognosis. These models are presented with point and interval estimates in a hazard ratio chart in `Figures 4.1(a) and 4.1(b)`.

For DFS, adjusting for all other clinical predictors, Nodal Ratio, Bilateral, LN Status, Metastasis and N stage (labelled as TNM_N in `Figure 4.1`) seem to be good predictors adjusting for all other clinical predictors. For OS, adjusting for all other clinical predictors, Metastasis, T stage and N stage are good predictors adjusting for all other clinical predictors. However, the OS model has large standard errors for some of the coefficients in the model. This results in wide confidence intervals in the hazard ratio plot.

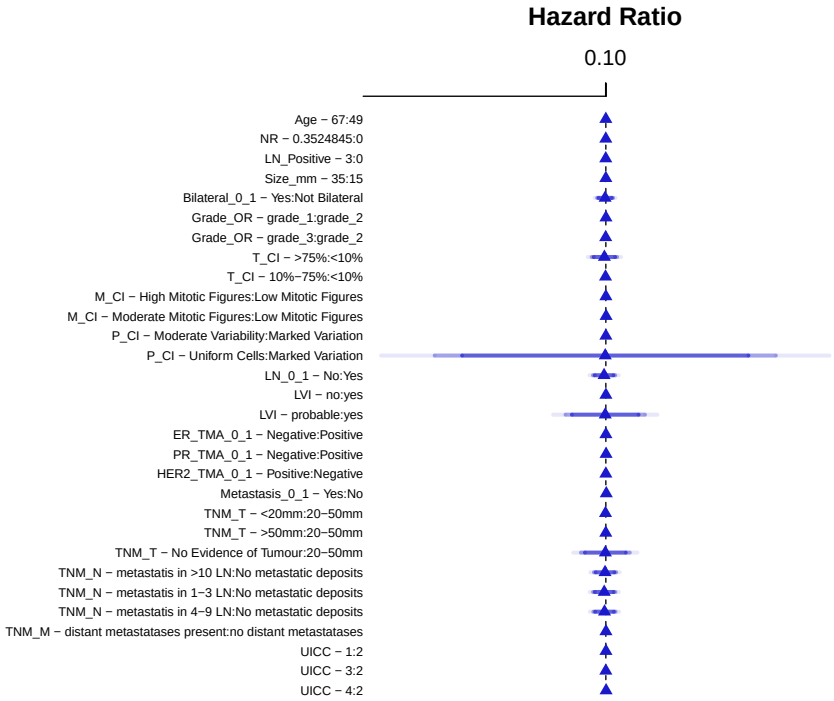
Models for DFS and OS including both the clinical and pathological biomarkers were also fitted `Table 4.3`. However, when these extra predictors were included, the sample size was reduced to 103 patients. Resulting in regression coefficients with large estimated standard errors.

Table 4.2: Cox Proportional Hazards model for clinical predictors for DFS and OS for the BC data.

	Dependent variable:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Age	0.003 (0.010)	−0.009 (0.036)
Nodal Ratio	−1.937*** (0.585)	−1.238 (1.627)
Bilateral (Yes)	1.071** (0.534)	−7.781 (43.556)
Grade (grade 2)	−0.545 (0.427)	−1.062 (2.463)
Grade (grade 3)	−0.676 (0.496)	−2.711 (2.373)
Tubule Formation ($> 75\%$)	−0.131 (0.643)	−15.593 (69.758)
Tubule Formation ($10\% - 75\%$)	−0.734* (0.378)	−3.469 (2.331)
Mitotic Count (Low)	−0.419 (0.308)	−0.507 (0.978)
Mitotic Count (Moderate)	0.181 (0.317)	1.932 (1.274)
Nuclear Pleomorphism (Moderate)	0.145 (0.251)	0.345 (0.946)
Nuclear Pleomorphism (Uniform Cells)	0.771 (1.235)	−6.305 (**935.901**)
No. LN Positive	0.003 (0.026)	−0.306 (0.227)
LN Status (Positive)	3.947*** (0.911)	18.913 (61.179)
Size (mm)	0.017 (0.013)	−0.004 (0.029)
Lymphovascular Invasion (probable)	−0.699 (0.841)	−7.316 (**217.782**)
Lymphovascular Invasion (yes)	0.206 (0.270)	−0.702 (0.912)
ER Status (Positive)	−0.127 (0.271)	−0.762 (1.438)
PR Status (Positive)	−0.320 (0.253)	−2.528** (1.038)
HER2 Status (Positive)	−0.054 (0.310)	0.169 (0.798)
Metastasis (Yes)	1.504*** (0.334)	4.790*** (1.245)
Tumour Staging (≥ 50)	−0.620 (0.718)	2.924 (2.004)
Tumour Staging (20-50mm)	0.171 (0.378)	3.587** (1.800)
Tumour Staging (No Tumour)	0.268 (1.198)	−1.636 (**133.947**)
LN Staging (metastasis in ≥ 10 LN)	1.674** (0.672)	8.637** (3.934)
LN Staging (metastasis in $4 - 9$ LN)	1.287** (0.539)	3.184 (2.416)
LN Staging (No metastatic deposits)	2.100** (0.929)	17.416 (61.162)
Metastasis Staging (no metastases)	−0.714* (0.384)	−1.225 (1.068)
UICC=2	0.585 (0.577)	−1.206 (2.513)
UICC=3	0.276 (0.792)	−1.363 (3.238)
UICC=4	1.264* (0.753)	1.508 (2.795)
Observations	221	221
R^2	0.527	0.652
χ^2 (df = 32)	163.937***	127.970***
Concordance Index	0.826	0.969
Note:		*p<0.1; **p<0.05; ***p<0.01



(a) Disease Free Survival



(b) Overall Survival

Figure 4.1: Hazard ratios and multilevel confidence bars for the effects of predictors in the full model for DFS and OS. The large estimated standard errors leads to wide confidence intervals.

Table 4.3: Cox Proportional Hazards model for clinical and pathological biomarkers predictors for DFS and OS for the BC data.

	Dependent variable:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Age	0.022 (0.020)	0.182 (0.667)
Nodal Ratio	-5.925^{***} (1.956)	3.730 (52.396)
Bilateral (Yes)	−1.616 (2.385)	16.183 (**591.096**)
Grade (grade 2)	−1.216 (0.750)	−2.623 (29.538)
Grade (grade 3)	−0.657 (0.886)	5.784 (28.852)
Tubule Formation (> 75%)	2.162 (1.601)	25.913 (93.243)
Tubule Formation (10%-75%)	0.715 (0.674)	12.498 (31.324)
Mitotic Count (Low)	1.305 (0.861)	5.275 (30.830)
Mitotic Count (Moderate)	0.097 (0.584)	15.099 (18.485)
Nuclear Pleomorphism (Moderate)	−0.270 (0.441)	1.191 (19.136)
Nuclear Pleomorphism (Uniform Cells)	0.000 (0.000)	0.000 (0.000)
No LN Positive	−0.071 (0.051)	−0.856 (1.389)
LN Status (Positive)	−6.233 (**140.931**)	−4.010 (**585.793**)
Size (mm)	-0.048^{*} (0.027)	0.047 (0.847)
Lymphovascular Invasion (probable)	−1.000 (1.467)	−1.879 (**268.702**)
Lymphovascular Invasion (yes)	0.360 (0.597)	−3.244 (18.421)
ER Status (Positive)	0.405 (0.488)	−8.568 (13.002)
PR Status (Positive)	−0.771 (0.597)	−10.137 (12.901)
HER2 Status (Positive)	−0.742 (0.789)	−8.044 (22.701)
Metastasis (Yes)	2.271^{***} (0.706)	27.175 (21.486)
Tumour Staging (≥ 50)	4.094^{***} (1.266)	7.403 (21.901)
Tumour Staging (20-50mm)	2.744^{***} (0.904)	7.649 (27.809)
Tumour Staging (No Tumour)	0.256 (2.631)	34.501 (92.571)
LN Staging (metastatis in 1-3 LN)	-9.812^{***} (2.497)	−32.785 (54.541)
LN Staging (metastatis in 4-9 LN)	-2.896^{**} (1.362)	−21.026 (30.035)
LN Staging (No metastatic deposits)	−20.042 (140.961)	−40.618 (**587.978**)
Metastasis Staging (no metastatases)	−1.388 (0.854)	10.601 (21.326)
UICC=2	−0.286 (1.383)	−8.885 (28.653)
UICC=3	-3.599^{*} (1.916)	−14.841 (33.221)
UICC=4	1.090 (1.917)	−14.751 (48.334)
Bcl2 Status (Positive)	0.549 (0.532)	18.023 (20.156)
CK14 Status (Positive)	−1.149 (0.700)	−2.876 (28.242)
CK5/6 Status (Positive)	-3.179^{**} (1.290)	−10.669 (37.729)
Ki67 Status (Positive)	−0.530 (0.801)	14.214 (24.442)
EGFR Status (Positive)	−0.391 (0.738)	5.284 (18.539)
E-cad Status (Positive)	−0.321 (0.763)	9.069 (17.638)
p53 Status (Positive)	0.260 (0.651)	−0.105 (20.154)
CDC7 Expression	0.054 (0.048)	0.344 (1.328)
tMcm2 Expression	0.013 (0.013)	−0.133 (0.439)
pMcm2 Expression	−0.003 (0.029)	−0.024 (1.098)
Observations	103	103
R^2	0.719	0.458
Concordance Index	0.884	0.969
Note:		*p<0.1; **p<0.05; ***p<0.01

4.4 Variable Selection

One benefit of simpler models is a logistical one, there may be reduced expense in omitting difficult to measure predictors. In the breast cancer dataset, where there are so many biological predictors, the cost and time benefit from only measuring a subset of the predictors is huge. The fact that a marker is significantly associated with outcome does not necessarily mean that it is important. Importance depends on the degree to which the marker influences patient outcome. Statistical significance is merely an indicator of whether the hypothesis of no prognostic effect can be ruled out [Simon and Altman, 1994]. We have already examined variable importance measures using random forests and also examined variable selection from the predictors that appear as splits in CART. These identified predictors such as Lymph Node status, Metastasis, Oestrogen status, UICC staging just to name a few as good predictors of DFS. For OS, predictors such as Lymph Node status, Metastasis and Oestrogen status were identified.

Here backward elimination was used for variable selection, this involves starting with all predictors included in the model. Each of the predictors is then tested using some model comparison criteria and any predictor which does not improve the model will be deleted. This process is continued until no further improvement in the model can be made. `fastbw` from the `rms` library performs a slightly inefficient but numerically stable version of backward elimination. This method uses the fitted complete model and computes conditional (restricted) Maximum Likelihood Estimates assuming multivariate normality of estimates [Harrell Jr, 2001].

The cost and time benefits have been mentioned previously. However, there are a few disadvantages of variable selection:

- The R^2 values are biased.
- The standard errors of the regression coefficients are biased downwards.
- The p-values often are too low as they are not corrected for the multiple comparisons of models.

Variable selection using backward elimination was applied to the full models (excluding the biomarkers) in `Table 4.2`. Lymph Node status, Nodal Ratio,

Metastasis and Lymph Node Stage were selected as the most parsimonious set of predictors for DFS (`Table 4.4`). When variable selection was applied to the OS model no predictors were selected due to the small sample size available. Also no predictors were identified using variable selection techniques on the CPH models with both clinical and pathological biomarkers.

Table 4.4: The predictors chosen by variable selection techniques on the DFS model with clinical predictors.

	Dependent variable:
	DFS
	$\hat{\beta}(ESE)$
Nodal Ratio	-1.251^{***} (0.299)
LN Status (Positive)	3.035^{***} (0.624)
Metastasis (Yes)	1.589^{***} (0.141)
N Stage (metastatis in 1-3 LN)	-1.077^{***} (0.257)
N Stage (metastatis in 4-9 LN)	-0.570^{***} (0.215)
N Stage (No metastatic deposits)	-0.163 (0.589)
Observations	576
R^2	0.434
Concordance Index	0.781
Note:	*p<0.1; **p<0.05; ***p<0.01

An alternative technique, LASSO using shrinkage methods, will be examined next.

4.5 Least Absolute Shrinkage and Selection Operator (LASSO)

The *LASSO* is a shrinkage and variable selection method for linear regression models. It minimizes the usual sum of squared errors with a bound on the sum of the absolute values of coefficients.

Given a set of predictors $x_1, x_2 \dots x_p$ and a response variable y, the lasso fits a linear model

$$\hat{y_i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}. \tag{4.4}$$

It uses the criteria that minimizes the $\sum (y_i - \hat{y}_i)^2$ subject to $sum[|\beta_j|] \leq s$. The

first sum is taken over cases in the dataset. The bound s is a tuning parameter. If s is chosen to be larger than $s_0 = \sum_1^p |\hat{\beta}_j|$ (where $\hat{\beta}_j = \hat{\beta}_j^{ls}$ the least squares estimates), then the LASSO estimates are the $\hat{\beta}_j s$. On the other hand, for $s = s_0/2$ say, then the least squares coefficients are shrunk by about 50% on average [Hastie et al., 2008]. It shrinks some coefficients and sets others to zero and hence tries to retain the good features of subset selection [Tibshirani, 1996]. This makes the final model easier to interpret. Choosing s is like choosing the number of predictors to include in the model and cross validation is a good tool for estimating the best value for s.

LASSO, on the other hand, is somewhat indifferent to highly correlated predictors, and will tend to pick one and ignore the rest. The LASSO penalty corresponds to a Laplace prior, which expects many coefficients to be close to zero, and a small subset to be larger and nonzero [Friedman et al., 2010].

The method for the linear case is adapted to accommodate the Cox proportional hazards model. The LASSO has identified useful clinical predictors for both DFS and OS (`Table 4.5`). For OS, Bilateral, Oestrogen status, Progesterone status, Metastasis and Metastasis staging have been identified as useful and all other coefficients have been shrunk to zero. From the plots of the coefficients verses the L1 norm in `Figure 4.2`, the non-zero coefficients are moving away from zero. As the L1 norm increases the coefficients increase towards the least squares coefficients. Take for example the plot for the clinical OS model (`Figure 4.2(c)`), the green line is the coefficient for Metastasis, the pink line is for Oestrogen status, the blue line is the coefficient for Metastasis stage, the black line is for Progesterone status and the blue line is for Bilateral status. These coefficients move away from zero quicker than those who are penalised.

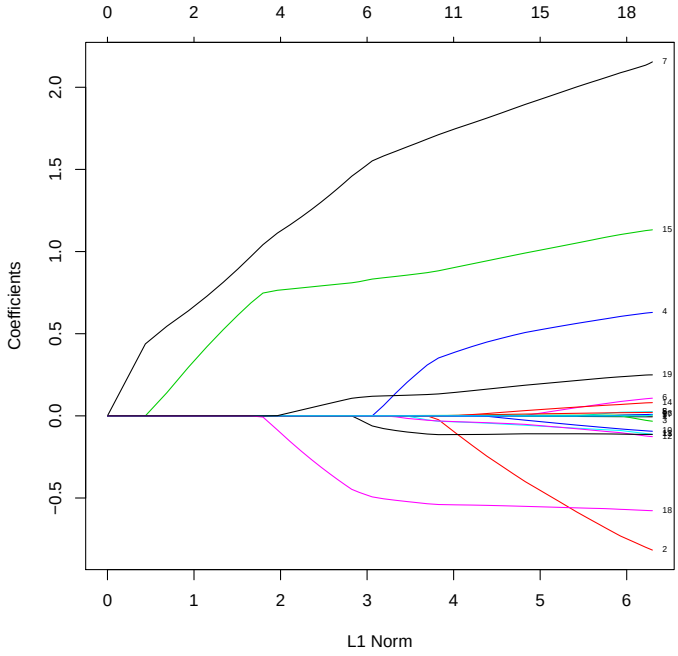
`Table 4.6` contains the LASSO results for both the clinical and pathological predictors.

4.6 Ridge Regression

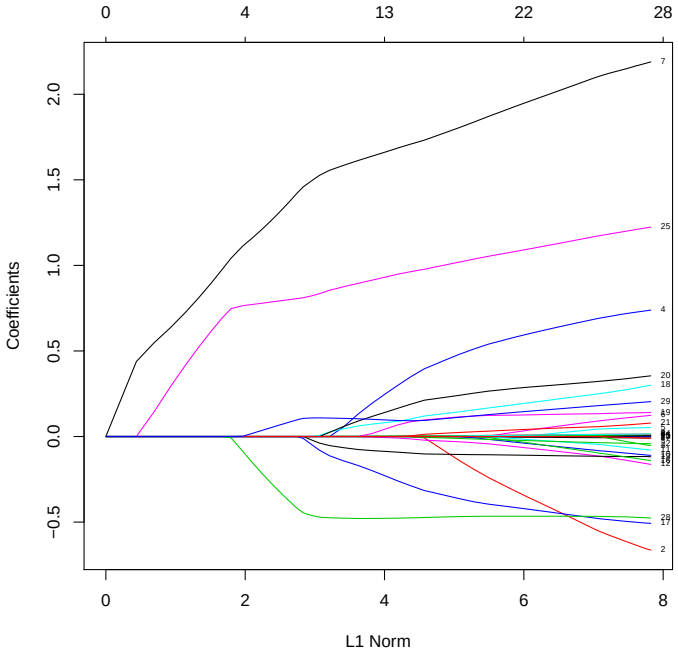
Ridge Regression is an alternative approach to alleviate multicollinearity amongst predictor variables in a model. The undesirable symptoms of correlated predictors can be reduced by retaining the size of the parameter estimates or shrinking

Variables	DFS $\hat{\beta}$	OS $\hat{\beta}$
Age		
Nodal Ratio		
Grade		
Bilateral	0.309	-0.163
Tubule Formation		
Mitotic Count		
Nuclear Pleomorphism	-0.024	
Ln Positive		
LN Status	1.685	
Size	0.002	
LVI		
ER Status	-0.027	-0.612
PR Status	-0.111	-0.311
Her2 Status		
Metastasis	0.872	2.817
UICC	0.130	
Tumour Staging		
LN Staging		
Metastasis Staging	- 0.535	0.515
Concordance Index	0.802	0.919

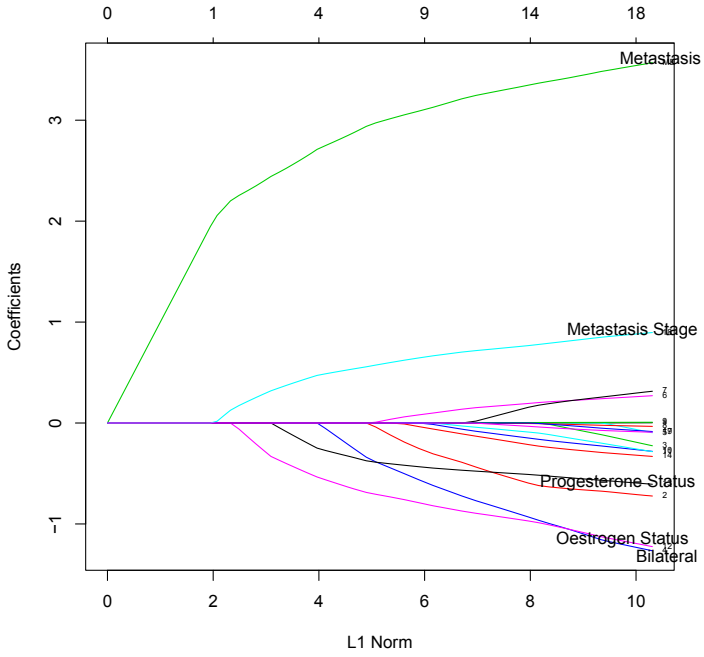
Table 4.5: Coefficients for Cox proportional hazards model with clinical predictors using LASSO.



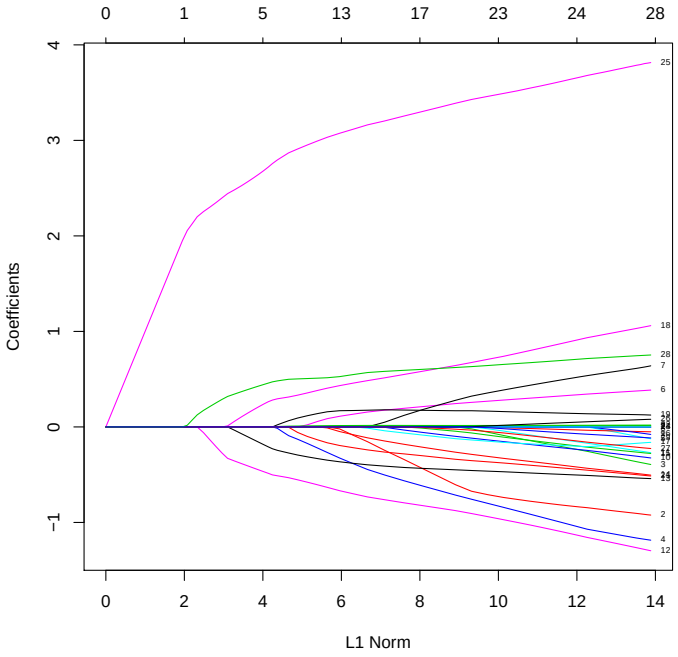
(a) DFS Clinical



(b) DFS Clinical & pathological



(c) OS Clinical



(d) OS Clinical & pathological

Figure 4.2: Plots of coefficients for the LASSO.

Variables	DFS $\hat{\beta}$	OS $\hat{\beta}$
Age	-0.001	
Nodal Ratio		
Grade		
Bilateral	0.349	-0.129
Tubule Formation		
Mitotic Count		
Nuclear Pleomorphism	-0.024	
Ln Positive	0.001	
LN Status	1.709	
Size	0.004	0.003
LVI		
ER Status	-0.014	-0.554
PR Status	-0.097	-0.294
Her2 Status		
Metastasis	0.872	
UICC	0.087	0.093
Tumour Staging		
LN Staging	-0.288	
Metastasis Staging	0.103	0.332
Bcl2 Status	0.189	
CK14 Status	0.004	-0.057
CK5/6 Status	-0.002	0.009
Ki67 Status		
EGFR Status		
E-cad Status	0.964	2.910
p53 Status		
CDC7 Expression		
tMcm2 Expression	-0.475	0.503
pMcm2 Expression	0.092	
Concordance Index	0.819	0.895

Table 4.6: Coefficients for Cox proportional hazards model with clinical and pathological predictors predictors using LASSO.

parameter estimates. Ridge regression is a linear regression technique which modifies the residual sum of squares to include a penalty for large parameter estimates.

The LASSO minimizes the sum of the square errors with a bound on the sum of the the absolute value of values of the coefficients where as ridge regression uses a bound on the sum of the square of the values of the coefficients. Ridge regression scales all the coefficients towards 0, but sets none to exactly zero like the LASSO. This helps to regularize in problems with $p > n$, but does not give a sparse solution. However, ridge regression better handles correlated predictors. If two predictors are highly correlated, ridge regression will tend to give them equal weight [Simon et al., 2011].

Variables	DFS $\hat{\beta}(ESE)$	OS $\hat{\beta}(ESE)$
Age	0.005 (0.01)	0.013 (0.01)
Nodal Ratio	0.012 (0.25)	0.164 (0.54)
Grade	0.017 (0.12)	-0.120 (0.27)
Bilateral	0.342 (0.34)	-0.771 (0.78)
Tubule Formation	-0.061 (0.11)	-0.091 (0.26)
Mitotic Count	0.057 (0.11)	0.102 (0.25)
Nuclear Pleomorphism	-0.032 (0.14)	0.226 (0.33)
Ln Positive	0.003 (0.01)	0.015 (0.03)
LN Status	0.574 (0.17)	0.201 (0.39)
Size	0.004 (0.01)	0.003 (0.01)
LVI	0.042 (0.08)	-0.001 (0.19)
ER Status	-0.120 (0.16)	-0.465 (0.36)
PR Status	-0.238 (0.15)	-0.825 (0.33)
Her2 Status	-0.057 (0.19)	-0.039 (0.41)
Metastasis	0.667 (0.19)	1.841 (0.41)
UICC	0.209 (0.10)	0.309 (0.21)
Tumour Staging	0.056 (0.08)	0.127 (0.19)
LN Staging	-0.097 (0.08)	-0.112 (0.17)
Metastasis Staging	-0.617 (0.23)	-0.477 (0.42)
Concordance Index	0.788	0.952

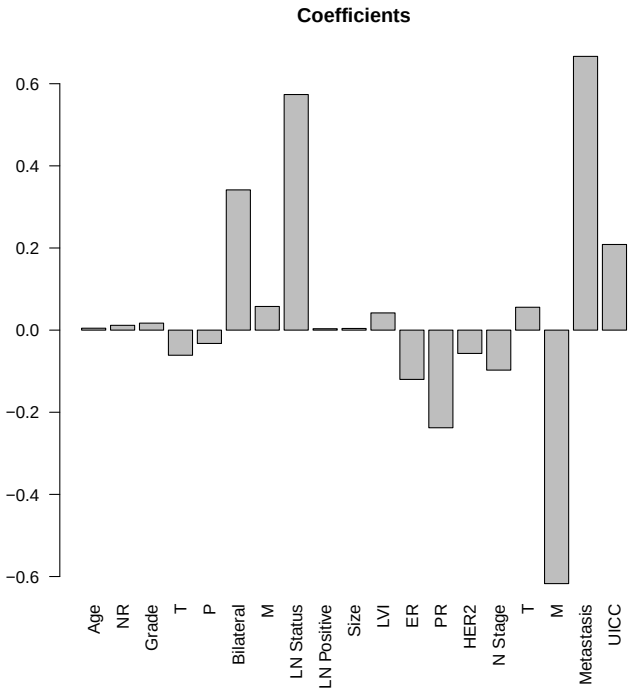
Table 4.7: Coefficients for Cox proportional hazards model using Ridge Regression with clinical predictors.

Ridge regression was applied to the UCH Galway breast cancer data. Firstly examining the routinely assessed predictors (`Table 4.7`) and then with both the routinely assessed and not so routinely assessed predictors (`Table 4.8`) for both DFS and OS. Plots of the coefficients for the four models are given in `Figure`

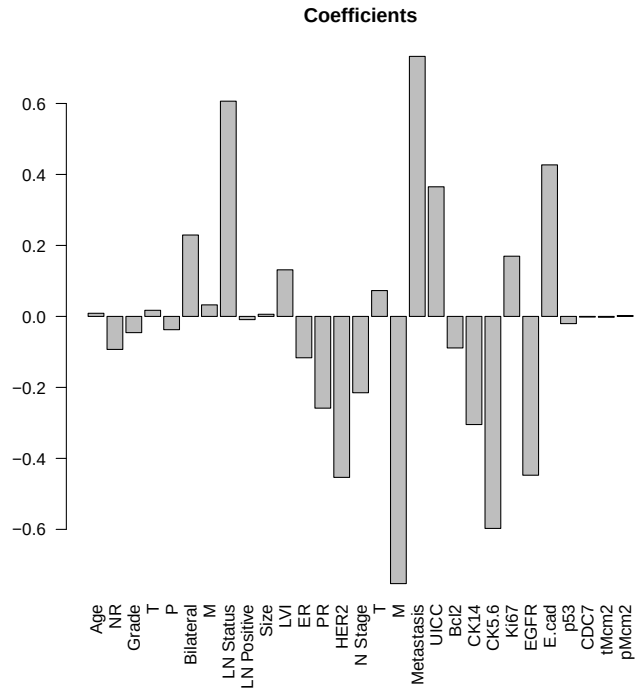
4.3. This shows the predictors which contribute most to the model. For the clinical DFS model, Bilateral, Lymph Node status, Metastasis stage, Metastasis and UICC staging seem to be the predictors strongest predictors. For the OS model, Bilateral, Oestrogen status, Progesterone status, Metastasis stage, Metastasis and UICC staging seem to be the strongest predictors. Similarly the stronger predictors can be identified for the models with all the clinical and pathological biomarkers.

Variables	DFS $\hat{\beta}(ESE)$	OS $\hat{\beta}(ESE)$
Age	0.008 (0.01)	-0.012 (0.03)
Nodal Ratio	-0.093 (0.40)	0.243 (1.35)
Grade	-0.046 (0.20)	0.095 (0.70)
Bilateral	0.229 (0.67)	-0.020 (2.53)
Tubule Formation	0.017 (0.18)	-0.381 (0.76)
Mitotic Count	0.033 (0.16)	0.080 (0.57)
Nuclear Pleomorphism	-0.037 (0.23)	0.240 (0.78)
Ln Positive	-0.009 (0.02)	0.027 (0.07)
LN Status	0.606 (0.29)	0.291 (0.97)
Size	0.006 (0.01)	0.004 (0.02)
LVI	0.131 (0.13)	0.330 (0.47)
ER Status	-0.117 (0.25)	-1.118 (0.89)
PR Status	-0.258 (0.25)	-0.863 (0.88)
Her2 Status	-0.453 (0.33)	-1.484 (1.04)
Metastasis	0.732 (0.29)	3.331 (0.92)
UICC	0.365 (0.17)	0.415 (0.57)
Tumour Staging	0.073 (0.14)	0.274 (0.46)
LN Staging	-0.215 (0.13)	-0.207 (0.45)
Metastasis Staging	-0.752 (0.37)	0.315 (1.00)
Bcl2 Status	-0.888 (0.25)	-0.304 (0.89)
CK14 Status	-0.305 (0.31)	-0.716 (1.21)
CK5/6 Status	-0.597 (0.37)	-1.021 (1.39)
Ki67 Status	0.170 (0.26)	0.559 (0.89)
EGFR Status	-0.447 (0.34)	0.249 (0.97)
E-cad Status	0.427 (0.37)	1.480 (1.39)
p53 Status	-0.020 (0.30)	-0.828 (0.92)
CDC7 Expression	-0.001 (0.02)	0.089 (0.06)
tMcm2 Expression	-0.002 (0.01)	-0.013 (0.02)
pMcm2 Expression	-0.003 (0.01)	0.027 (0.04)
Concordance Index	0.841	0.998

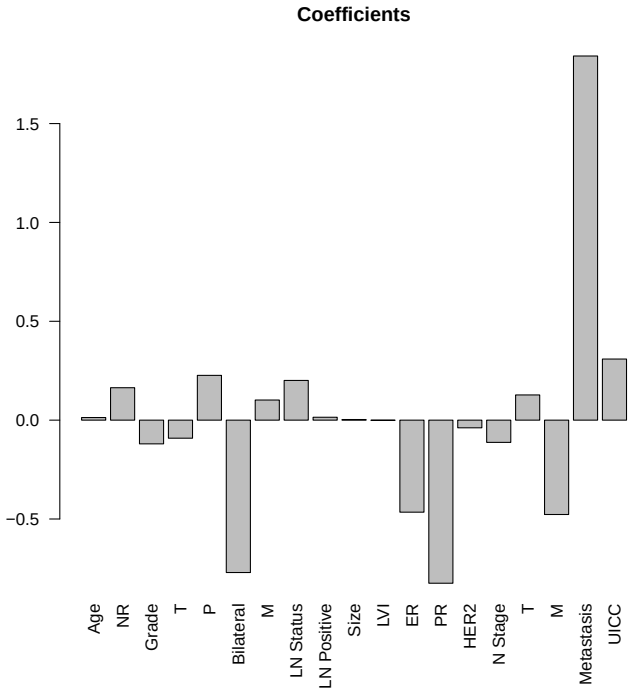
Table 4.8: Coefficients for Cox proportional hazards model using Ridge Regression with clinical and pathological predictors.



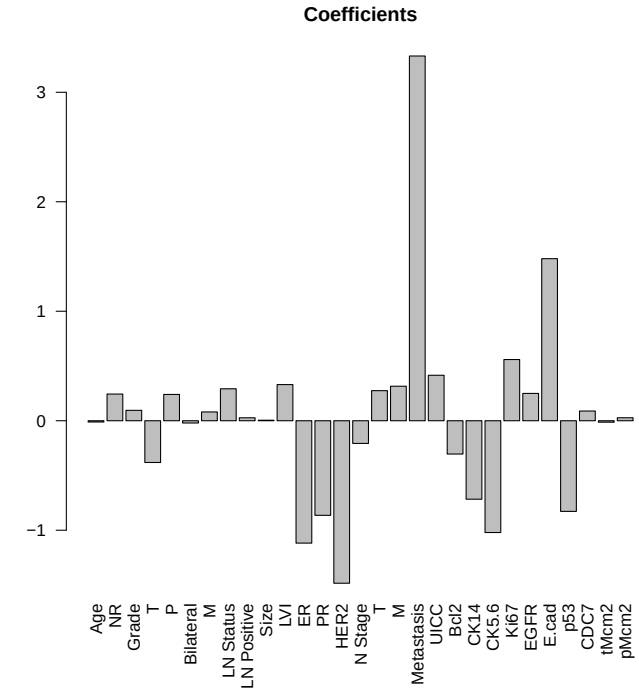
(a) DFS Clinical



(b) DFS Clinical & pathological



(c) OS Clinical



(d) OS Clinical & pathological

Figure 4.3: Bar charts of coefficients from Ridge Regression for DFS and OS.

4.7 Non-Linear Effects

4.7.1 CPH with splines

In modeling the functional relationship between a response Y and a predictor X, often the relationship is non-linear. Regression splines offer a convenient way to examine the functional relationship. Spline functions are piecewise polynomials used in curve fitting. That is, they are polynomials within intervals of X that are connected across different intervals of X [Harrell Jr, 2001].

Table 4.9: Wald Statistics for examining non-linear effects in both DFS and OS.

	DFS			OS		
	χ^2	$d.f.$	$p-value$	χ^2	$d.f.$	$p-value$
Age	0.15	2	0.9260	1.77	2	0.4136
Nonlinear	0.03	1	0.8693	1.70	1	0.1916
Nodal Ratio	14.37	2	0.0008	6.40	2	0.0408
Nonlinear	2.56	1	0.1095	6.33	1	0.0119
Bilateral	3.56	1	0.0591	0.10	1	0.7528
Grade	2.01	2	0.3666	0.35	2	0.8406
Tubule Formation	5.32	2	0.0701	1.49	2	0.4736
Mitotic Count	2.43	2	0.2967	1.50	2	0.4727
Nuclear Pleomorphism	0.07	2	0.9642	0.52	2	0.7716
LN Positive	3.06	2	0.2167	3.04	2	0.2192
Nonlinear	3.05	1	0.0807	0.81	1	0.3674
LN Status	10.81	1	0.0010	0.23	1	0.6338
Size (mm)	4.44	2	0.1084	1.43	2	0.4882
Nonlinear	1.47	1	0.2251	1.36	1	0.2439
LVI	1.44	2	0.4855	0.64	2	0.7248
ER Status	0.03	1	0.8605	0.54	1	0.4608
PR Status	1.91	1	0.1670	9.65	1	0.0019
HER2 Status	0.00	1	0.9440	0.38	1	0.5379
Metastasis	20.48	1	< 0.0001	7.82	1	0.0052
Tumour Staging	5.08	3	0.1660	8.00	3	0.0459
LN Staging	5.68	3	0.1282	6.38	3	0.0947
Metastasis Staging	2.27	1	0.1323	2.09	1	0.1478
UICC	5.39	3	0.1452	4.41	3	0.2201
TOTAL NONLINEAR	4.88	4	**0.2997**	7.42	4	**0.1153**
TOTAL	131.99	34	< 0.0001	34.44	34	0.4468

The simplest spline function is the linear spline function which divides the x-axis into intervals at different points called *knots* and fits piecewise linear functions. A draw back of linear splines is that they generally do not fit curved

functions well. This obstacle can be overcome using a cubic spline function. A cubic spline is a spline constructed of piecewise third-order polynomials. Cubic spline functions however have a drawback that they can behave poorly in the tails. Restricted cubic splines fit a linear function in the tails and a cubic function in between the tails.

Non-linear effects for the continuous predictors are examined using cubic spline functions with 3 knots for both the OS and DFS survival models. These were tested using Wald Statistics for all the effects in the model. However none of the continuous predictors were associated with non-linear effects (p-values > 0.05 for non-linear effects) and the p-values for the TOTAL non-linear effects are not significant (0.2997 and 0.1153 in `Table 4.9` respectively). These TOTAL non-linear and TOTAL are pooled Wald statistics for the combined effects for non-linear effects and all effects. As these non-linear effects do not have a significant impact on the model they will not be included in further analysis.

4.7.2 Interaction Terms

Similarly to the testing for non-linear terms using cubic splines in the continuous predictors in the previous section, the effects of interactions are tested using Wald Statistics. This includes joint tests of all interaction terms in the model and all non-linear terms in the model performed. The inclusion of interaction terms in a model can be guided by the clinician (what is biologically plausible) and the data. The clinicians have specified that there should be no need for interaction terms. In Chapter 3, examining the survival trees for DFS and OS, there looks as if there may be an interaction present for DFS between Lymph Node status and Metastasis (`Figure 3.13(a)`). The Kaplan Meier estimate for this interaction is given in `Figure 4.4`. The interaction seems obvious from this as if we examine the Lymph Node negative patients (black and green lines), there seems to be a difference between the patients with and without Metastasis. The same can be said for lymph node positive patients. The over-all Log-rank test yielded a p-value < 0.001. Many different interactions were examined however only the interactions between Lymph Node status and Metastasis, Metastasis and UICC staging and Bilateral and Lymph Node Status staging were significant

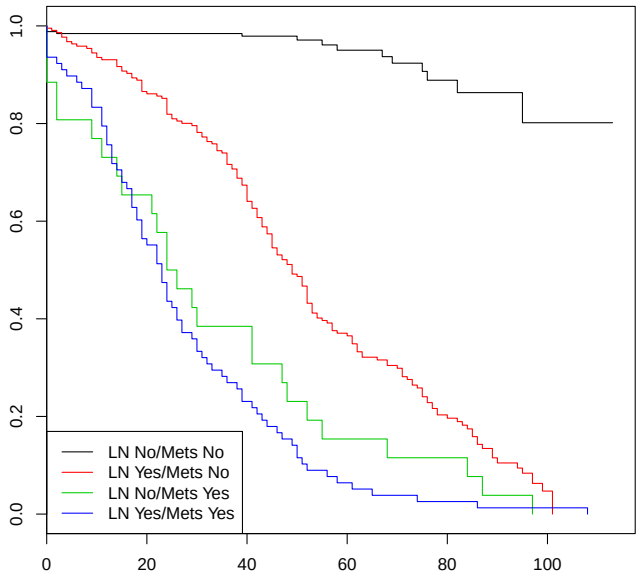


Figure 4.4: Kaplan Meier estimates for groups for the interaction between Lymph Node status and Metatasis. (Legend: LN- Lymph Node status and Mets- Metatasis.)

for DFS (`Table 4.10`).

For the OS model, no interactions were significant since none of the interaction terms are significant and the TOTAL interaction is not significant. No interaction terms were included in any of the other models fitted.

4.8 Conclusions

An overview of classical approaches to modelling survival data, specifically Cox proportional hazards, have been examined in this chapter. More candidate models have been identified (`Tables 4.11-4.14`) using all predictors and a subset of predictors. For DFS, Lymph Node Status, Nodal Ratio, Metastasis and Lymph Node staging were identified as useful predictor using variable selection. For OS and the models with both the clinical and pathological biomarkers variable selection could not be applied. The LASSO was also applied which identified a

Table 4.10: Wald Statistics for the effects of interactions. (*Factor+Higher Order Factors - tests the combined main effect and interaction effects)

	DFS			OS		
	χ^2	$d.f.$	$p-value$	χ^2	$d.f.$	$p-value$
Age	0.65	1	0.4198	0.00	1	0.9557
Nodal Ratio	4.17	1	0.0412	4.92	1	0.0265
Bilateral (Factor+Higher Order Factors)	29.85	2	$<$ **0.0001**	2.46	2	0.2928
All Interactions	9.09	1	0.0026	0.02	1	0.8820
Grade	1.23	2	0.5404	1.70	2	0.4267
Size (mm)	6.71	1	0.0096	4.76	1	0.0291
ER Status	0.01	1	0.9226	6.41	1	0.0113
PR Status	1.17	1	0.2789	0.64	1	0.4252
HER2 Status	0.01	1	0.9094	2.30	1	0.1295
LN Status (Factor+Higher Order Factors)	40.47	3	$<$ **0.0001**	0.16	3	0.9835
All Interactions	21.59	2	$<$ **0.0001**	0.03	2	0.9836
Metastasis (Factor+Higher Order Factors)	55.84	5	$<$ **0.0001**	20.01	5	0.0012
All Interactions	21.59	4	0.0002	0.84	4	0.9329
Tumour Staging	2.56	3	0.4649	6.64	3	0.0844
LN Staging	6.54	3	0.0881	4.86	3	0.1821
Metastasis Staging	0.55	1	0.4575	0.01	1	0.9399
UICC (Factor+Higher Order Factors)	16.72	6	0.0104	11.28	6	0.0801
All Interactions	7.93	3	0.0476	0.82	3	0.8436
LN $\times$ Metastasis (Factor+Higher Order Factors)	9.19	1	0.0024	0.01	1	0.9160
Bilateral $\times$ LN (Factor+Higher Order Factors)	9.09	1	0.0026	0.02	1	0.8820
Metastasis $\times$ UICC (Factor+Higher Order Factors)	7.93	3	0.0476	0.82	3	0.8436
TOTAL INTERACTION	32.19	5	$<$ **0.0001**	0.86	5	0.9729
TOTAL	171.60	26	$<$ 0.0001	43.80	26	0.0159

subset of predictors for each of the models. Investigations into non-linear effects using splines and interactions did not improve the models.

Some of these classical approaches did not work well since there are a lot of missing values present in the data and this reduces the sample size used in the analysis. This caused particular issues using variable selection techniques.

As there is such a high proportion of missing data present in the data, the next chapter will examine the effect of such missing data in identifying good predictors. Variable selection techniques applied to imputed datasets will be examined using a simulation study based on synthetic data from the BC data and all the results reported.

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO
Age		X		X		X	
Bilateral		X		X		X	X
Grade				X		X	
Tubular Formation				X		X	
Nuclear Pleomorphism				X		X	X
Mitotic Count				X		X	
Tumour Size				X		X	X
Lymphovascular Invasion				X		X	
No of Lymph Nodes positive			X	X		X	
Lymph Node status	X	X	X	X	X	X	X
Nodal Ratio				X	X	X	
Metastasis	X	X		X	X	X	X
Tumour Staging				X		X	
Metastasis Staging				X		X	X
Lymph Node Staging			X	X	X	X	
UICC			X	X		X	X
NPI				X		X	
Oestrogen Status		X		X		X	X
Progesterone Status				X		X	X
Her2 Status			X	X		X	
Concordance Index	0.826	0.762	0.777	0.766	0.781	0.788	0.802

Table 4.11: Summary of clinical predictors selected in techniques explored so far for DFS. X means the predictor is included in the model.

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO
Age		X		X		X	X
Bilateral		X		X		X	X
Grade				X		X	
Tubular Formation				X		X	
Nuclear Pleomorphism				X		X	X
Mitotic Count				X		X	
Tumour Size			X	X		X	X
Lymphovascular Invasion				X		X	
No of Lymph Nodes positive			X	X		X	X
Lymph Node status	X	X	X	X		X	X
Nodal Ratio			X	X		X	
Metastasis	X	X		X		X	X
Tumour Staging				X		X	
Metastasis Staging				X		X	X
Lymph Node Staging			X	X		X	X
UICC				X		X	X
NPI				X		X	
Oestrogen Status		X		X		X	X
Progesterone Status				X		X	X
Her2 Status				X		X	
Bcl2 Status				X		X	X
CK14 Status				X		X	X
CK5/6 Status				X		X	X
EGFR Status				X		X	
Ki67 Status				X		X	
p53 Status				X		X	
E-cad Status				X		X	X
tMcm2 Status				X		X	X
CDC7				X		X	
pMcm2				X		X	X
Concordance Index	0.762	0.777	0.765	0.884	NA	0.841	0.819

Table 4.12: Summary of clinical and pathological predictors selected in techniques explored so far for DFS. X means the predictor is included in the model.

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO
Age				X		X	
Bilateral				X		X	X
Grade				X		X	
Tubular Formation				X		X	
Nuclear Pleomorphism			X	X		X	
Mitotic Count				X		X	
Tumour Size			X	X		X	
Lymphovascular Invasion				X		X	
No of Lymph Nodes positive			X	X		X	
Lymph Node status	X		X	X		X	
Nodal Ratio			X	X		X	
Metastasis	X	X		X		X	X
Tumour Staging			X	X		X	
Metastasis Staging			X	X		X	X
Lymph Node Staging			X	X		X	
UICC		X	X	X		X	
NPI				X		X	
Oestrogen Status		X	X	X		X	X
Progesterone Status				X		X	X
Her2 Status			X	X		X	
Concordance Index	0.909	0.922	0.911	0.969	NA	0.952	0.919

Table 4.13: Summary of clinical predictors selected in techniques explored so far for OS. X means the predictor is included in the model.

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO
Age				X		X	
Bilateral			X	X		X	X
Grade				X		X	
Tubular Formation			X	X		X	
Nuclear Pleomorphism			X	X		X	
Mitotic Count				X		X	
Tumour Size			X	X		X	X
Lymphovascular Invasion			X	X		X	
No of Lymph Nodes positive			X	X		X	
Lymph Node status	X		X	X		X	
Nodal Ratio			X	X		X	
Metastasis	X	X	X	X		X	
Tumour Staging				X		X	
Metastasis Staging			X	X		X	X
Lymph Node Staging			X	X		X	
UICC		X	X	X		X	X
NPI			X	X		X	
Oestrogen Status		X	X	X		X	X
Progesterone Status			X	X		X	X
Her2 Status			X	X		X	
Bcl2 Status			X	X		X	
CK14 Status			X	X		X	X
CK5/6 Status			X	X		X	X
EGFR Status				X		X	
Ki67 Status				X		X	
p53 Status			X	X		X	
E-cad Status			X	X		X	X
tMcm2 Status			X	X		X	X
CDC7				X		X	
pMcm2				X		X	
Concordance Index	0.909	0.922	0.790	0.995	NA	0.998	0.895

Table 4.14: Summary of clinical and pathological predictors selected in techniques explored so far for OS. X means the predictor is included in the model.

Chapter 5

Variable Selection techniques with imputed data

5.1 Introduction

As we have seen in the previous chapter, missing data can be a serious problem, particularly in retrospective observational studies where the percentage of subjects with complete data can be of concern. Missing values in the BC dataset presents such challenges when considering variable selection techniques, as there are a large number of clinical and pathological variables and there is some proportion of missing data present in the majority of these variables. `Figure 5.1` displays the proportions of missingness (recorded as NAs in stored data) in each of the clinical and pathological predictors. Some of the clinical predictors have no missing data, such as bilateral and age, while others, such as CK14 and CK5/6, have more than 30% missing. The genetic predictors generally have the highest proportion of missing data as some of these are not routinely assessed.

The most common way to deal with missing data is casewise deletion. Casewise deletion is applied when a subject is missing in one predictor, the whole case (subject) is omitted as a consequence. The consequent drop in sample size

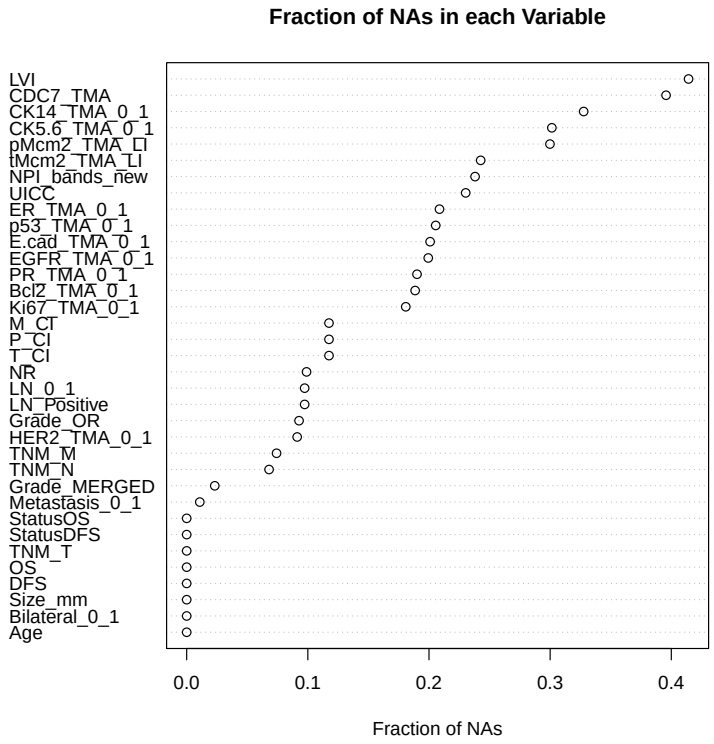


Figure 5.1: Proportion of missingness for each clinical and pathological predictors for the BC data.

will reduce the power of such studies to demonstrate a given effect [Altman and Lyman, 1998]. Using complete cases could mean the target population has been changed. Casewise deletion results in regression coefficient estimates that can be terribly biased, imprecise, or both [Harrell Jr, 2001]. The inefficiency comes from the reduction in sample size, which causes standard errors to increase, confidence intervals to widen, and power of tests of association and tests of lack of fit to decrease [Harrell Jr, 2001]. Deletion of cases with missing predictors causes bias and increased variance. Even though caution should be taken when imputing missing values, it is usually better to estimate selected data values than to delete an entire subject's record [Janssen et al., 2010].

To "provide" data using multiple imputation is a better alternative than discarding valuable observed data. The purpose of multiple imputation is not to make up or gain data but to preserve real, observed data [Janssen et al., 2010].

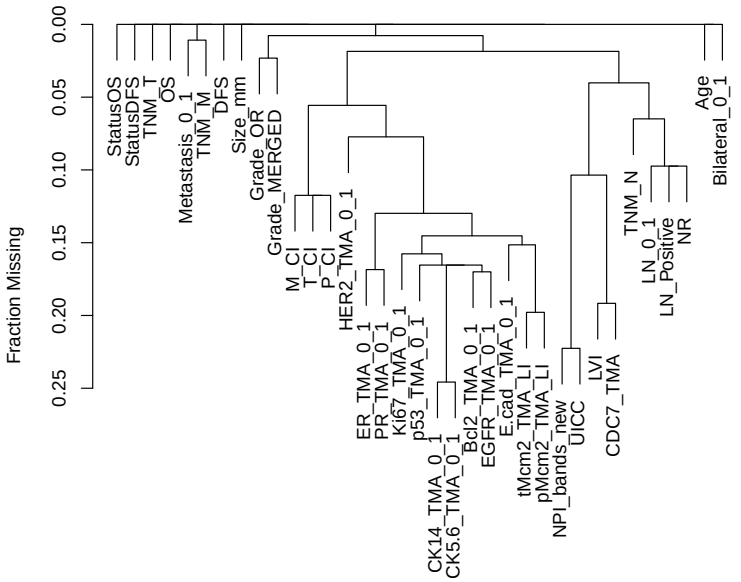


Figure 5.2: Cluster analysis showing which predictors tend to be missing on the same patients for the BC data.

`Figure 5.2` contains a cluster analysis that shows which predictors tend to be missing on the same patients. If a patient is missing in one of the biomarker predictors it is quite likely they are missing the rest of the biomarker information. If a patient gets one of the biomarker predictors measured, they are more likely to have all the biomarker variables measured. This is due to reasons such as patient samples not being available or not sent for analysis of genetic markers.

As seen in the previous chapter, this high proportion of missing data in the predictors results in a small complete case sample size, which causes convergence issues when identifying useful predictors for a prognostic index for breast cancer. In this chapter, the effects of missing data on the predictors selected into the final model will be examined and an alternative approach for model selection by identifying the final model based on the imputed datasets will be examined.Variable selection techniques using imputed data will be presented

and their performance evaluated using simulated (synthetic) data based on the BC data. The synthetic data are simulated using a proportional hazards model based on the BC data. Missing data are then induced under a variety of mechanisms and assumptions. The missing data are then imputed using multivariate imputation by chained equations (MICE) and random forests. Variable selection techniques are then applied to the imputed data in a variety of ways in a similar manner to Wood et al. [2008].

Variable selection techniques can be applied to each imputed dataset. An appealing attribute of this approach is that power can be retained by avoiding casewise deletion. However, an extra level of complexity is added in terms of identifying a consistent set of predictors. There are currently no guidelines for variable selection in multiply imputed data sets. The usual practice is to perform variable selection among the complete cases, a simple but inefficient and potentially biased procedure [Wood et al., 2008]. Methods for variable selection in multiply imputed data in the literature suggests selecting predictors using a voting system or by stacking the imputed datasets and performing weighted regression. Another approach is to impute using random forests, where only one dataset is imputed.

5.2 Simulation Study Set up

A comparison of variable selection techniques in multiply imputed data was performed using a simulation study. Here is a summary of the simulation procedure:

- Started by simulating a complete set of predictors X, with both "good" and "poor" (noise) predictors of DFS;
- Simulated the response Y which is time to event using a Cox proportional hazards model based on a predefined model consisting of a subset of X deemed useful for predicting DFS;
- Induced missing data in X under a variety of missing data patterns and assumptions;

- Imputed the missing data in X using Multivariate Imputation by Chained Equations and Random Forests;
- Performed the model selection procedures;
- Repeated these steps 1000 times for each of the simulated datasets and compare the results across the simulations.

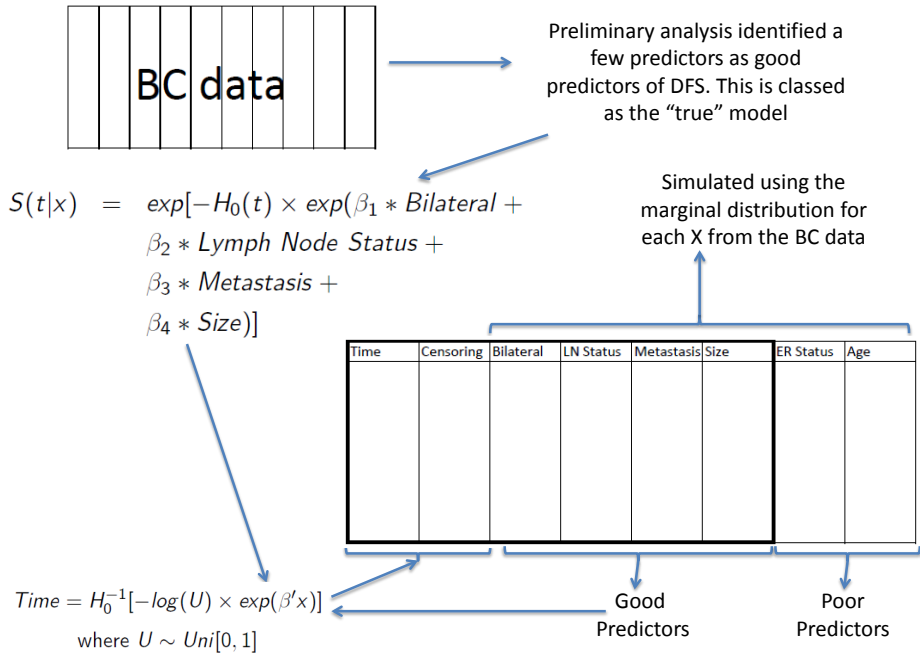


Figure 5.3: Flow chart showing how the data was simulated.

5.2.1 Simulation of Predictors

Data was simulated using parameter estimates based on the Galway breast cancer data. From preliminary analysis, Bilateral (Yes/No), Lymph Nodes Positive (Yes/No), Metastasis (Yes/No) and size of the tumour were identified as good predictors of DFS. Data for Bilateral, Lymph Node Positive and Metastasis were simulated using a binomial distribution with $p = 0.07,\ 0.55$ and 0.17 respectively, based on the marginal distribution of the data in the Galway breast cancer dataset. The size of the tumour was simulated using a normal distribution with $\mu = 2.6$ and $\sigma = 1.36$. These four predictors are classed as good

Table 5.1: 'True' model for DFS using Cox proportional hazards model.

	Dependent variable:
	DFS
	$\hat{\beta}(ESE)$
Bilateral (Yes)	1.132***
	(0.233)
LN Status (Yes)	1.616***
	(0.186)
Metastasis (Yes)	1.123***
	(0.149)
Size	0.226**
	(0.094)
Observations	464
R^2	0.419
Note:	*p<0.1; **p<0.05; ***p<0.01

predictors of DFS and are classed as our 'true' model (see `Table 5.1`). *Noise* variables, such as Age and Oestrogen Status, were also simulated, which are not good predictors of DFS.

The sample size of the simulated data was varied to examined the effect with $n = 1000,\ 700$ and 100.

5.2.2 Simulation of time to event data using the Cox PH model

Recall from Chapter 4 that the survival function of the Cox proportional hazards model is given by

$$S(t|x) = exp[-H_0(t) \times exp(\beta' x)] \tag{5.1}$$

where

$$H_0(t) = \int_0^t h_0(u)du \tag{5.2}$$

is the cumulative baseline hazard function. The distribution function of the Cox model is

$$F(t|x) = 1 - exp[-H_0(t) \times exp(\beta' x)] \tag{5.3}$$

Let Y be a random variable with distribution function F, then U=F(Y) follows a uniform distribution on the interval from 0 to 1. Let T be the survival time of the Cox model (5.1), then it follows from (5.3) that

$$U = exp[-H_0(t) \times exp(\beta' x)] \sim Uni[0, 1] \tag{5.4}$$

If $h_0(t) > 0$ for all t, then H_0 can be inverted and the survival time T of the Cox Model (5.1) can be expressed as

$$T = H_0^{-1}[-log(U) \times exp(\beta' x)] \tag{5.5}$$

where U is a random variable with $U \sim Uni[0, 1]$.

An uncensored survival time was simulated using a Cox proportional hazards model and the parameters given in the *true* model (`Table 5.1`). A censored time for each subject was simulated using the exponential distribution as an independent censoring distribution. The survival time was the minimum of each of these two times and the censoring indicator was coded as 1 if the uncensored survival time was smaller or 0 if the censored survival time was smaller. This creates a similar censoring mechanism to that in the BC cohort in Galway. The simulation study datasets had 17% censored observations on average.

5.2.3 Missing Data

Missing Data were induced in the explanatory variables using a variety of missing data patterns and assumptions. As the main aim is to examine variable selection, no missingness was induced in the response. As shown in `Figure 5.1`, there was no missingness data present in Age, Size or Bilateral, so no missing data will be induced into these variables. Missing data was induced under 3 different missing data mechanisms as outlined below.

1. Missing At Random (MAR)

MAR is where the missing values are random conditional on the other available information in the data [Janssen et al., 2010]. Missingness in a predictor is MAR if it does not depend on the actual value of the the predictor itself once the other predictors in the data are available.

The BC data was used to identify which predictors were related to missingness in a particular predictor. These models were used to induce missingness in the simulated data. Missingness was induced in LN status, Metastatic (Y/N) and ER status using the logistic models in `Table 5.2`. Each model can be used to create event probabilities using the simulated data. Missing values can be imposed on the predictor for individuals with fitted probabilities falling in the top 10%, 20% and 30%, the proportion of missingness is varied using this.

2. Missing Completely at Random (MCAR)

Missing data were simulated using the mechanism MCAR, which is a special case of MAR. With MAR, missingness has a purely random component and a systematic component that depends on some variables in the dataset, but not on the actual values of the variable with missingness. With MCAR, the missingness has a purely random component [Paul et al., 2008]. This means that the missing values are completed independent of the predictors and the response, i.e. the patients with missing values in the predictor do not differ systematically to those with the predictor observed.

This can be easily implemented by randomly selecting values which can be replaced by missingness.

3. Missing Not At Random (MNAR)

The third missingness mechanism is known as Missing Not At Random (MNAR), also referred to as "non-ignorable" in much published research. If missingness on the predictor is MNAR, it depends on the actual level of the predictor and potentially other variables not available in the data. Note that MNAR does not mean that missingness lacks a random component, only that its systematic component is a function of the actual values of the variable with missingness [Paul et al., 2008]. This was induced by using logistic models with predictors

Table 5.2: Logistic models used to induce missingness for MAR.

	Dependent variable:		
	Missing LN Status	Missing Metastasis	Missing ER Status
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
DFS Time	0.024^{**}	0.150^{*}	0.002
	(0.009)	(0.087)	(0.005)
DFS Status (1)	-0.334	0.850	-1.042^{**}
	(0.676)	(1.926)	(0.427)
Bilateral (Yes)	0.510	6.431^{*}	0.777
	(0.863)	(3.808)	(0.497)
Metastasis (Yes)	0.956		0.467
	(0.739)		(0.420)
LN Positive (Yes)		0.937	0.571
		(2.201)	(0.387)
Size	-0.484^{**}	-0.290	-0.827^{***}
	(0.230)	(1.067)	(0.113)
Age	0.087^{***}	0.149	-0.023^{**}
	(0.022)	(0.116)	(0.011)
ER Status (Positive)	-1.179^{**}	-1.531	
	(0.500)	(1.964)	
Constant	-7.460^{***}	-25.731^{**}	2.274^{***}
	(1.759)	(12.854)	(0.760)
Observations	376	358	464
Log Likelihood	-63.025	-8.437	-207.248
Akaike Inf. Crit.	142.050	32.874	430.496

Note: *p<0.1; **p<0.05; ***p<0.01

such as Progesterone status, Her2 status and BcI2 status which are not present in the simulated dataset of predictors X given in `Table 5.3`. Again the amount of missingness is imposed depended on a percentage cases in the top fitted probabilities.

Table 5.3: Logistic models used to induce missingness for MNAR.

	Dependent variable:		
	Missing LN Status	Missing Metastasis	Missing ER Status
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
PR Status (Positive)	−0.279	−0.067	0.008
	(0.474)	(1.052)	(0.592)
HER2 Status (Positive)	−1.696	−17.056	−0.248
	(1.062)	(2,330.575)	(0.828)
Bcl2 Status (Positive)	−0.092	−1.518	−0.214
	(0.473)	(1.208)	(0.584)
Ki67 Status (Positive)	0.421	1.199	0.054
	(0.486)	(1.020)	(0.625)
Constant	−2.420***	−4.023***	−2.992***
	(0.425)	(0.919)	(0.533)
Observations	357	357	357
Log Likelihood	-83.148	-19.407	-62.125
Akaike Inf. Crit.	176.297	48.813	134.249
Note:			*p<0.1; **p<0.05; ***p<0.01

5.2.4 Multiple Imputation

Multiple Imputation (MI) is a popular technique in missing data problems [Rubin, 1987]. MI uses models based on observed data to replace missing values with credible values. This process is repeated a number of times to create several imputed datasets. To "provide" data using multiple imputation is a better alternative than discarding valuable observed data. The purpose of multiple imputation is not to make up or gain data but to preserve real, observed data [Janssen et al., 2010].

Multivariate Imputation by Chained Equations (MICE)

The philosophy behind MICE is that multiple imputation is best done as a sequence of small steps, each of which may require diagnostic checking [van Buuren and Groothuis-Oudshoorn, 2011]. For each missing variable, a conditional distribution for the missing data given the other data can be specified [Van Buuren and Oudshoorn, 2000]. Initially, all the missing values are filled at random. The first variable with at least some missing values are regressed on all other variables in the dataset. Missing values in the predictor are replaced by simulated draws from the posterior predictive distribution of the predictor, an important step known as proper imputation [Royston and White, 2011]. This is then repeated for all other predictors with missing values.

Random Forest Imputation

Random Forest Imputation is a non-parametric method of imputation where random forests are used to impute a single dataset, where values are imputed by averaging over many unpruned classification or regression trees. For each variable it fits a random forest on the observed part and then predicts the missing part. Random Forest Imputation can cope with a mixed type data of both continuous and categorical predictors, interactions and large numbers of predictors with a small sample size (large p small n). The main advantages of this technique are it is quicker and there is only one dataset imputed which means applying variable selection techniques is made easier. The disadvantage is that there is no measure of variability available across the imputations.

5.2.5 Classical Variable Selection Techniques

This topic has already been discussed in Chapter 4. The classical variable selection techniques include stepwise, backward, forward and shrinkage methods such as ridge regression and the LASSO. CART can also be used for variable selection, where the splits at each node in the tree are deemed useful predictors. Here, in this study, stepwise, backward and tree based methods are examined.

5.2.6 Variable Selection in Imputed Data

Typically multiple imputation is used at the end of the analysis where the final model generated from the original data are fitted to each of the imputed datasets and the results combined using Rubin's Rules [Rubin, 1987]. Typically, several datasets are imputed and Rubin [1987] developed a method to average the outcomes across each of the imputed datasets. Each imputed data set is analyzed separately and the parameter estimates are averaged except for the standard error term (SE). The combined SE is calculated by the within variance of each dataset as well as the variance between imputed items on each data set. An alternative use for MI is to identify the final model based on the imputed datasets. For example variable selection techniques can be applied to each imputed dataset. An appealing attribute of this approach, is that power can be retained by avoiding casewise deletion. However, an extra level of complexity is added in terms of identifying a consistent set of predictors.

Methods for variable selection in multiply imputed data in the literature [Wood et al., 2008] include selecting predictors using a voting system: variable selection would be performed on each of the imputed datasets and predictors are selected based on whether they appear in any, half or all the models.

Another method suggested is to stack the imputed datasets and perform weighted regression using weights related to the amount of missingness present. Three different weights are used,

$$\begin{aligned} W1 &= \frac{1}{M} \\ W2 &= \frac{1-f}{M} \\ W3 &= \frac{1-f_i}{M} \end{aligned}$$

where M is the number of imputed datasets, f is the average proportion of missing data across all variables and f_i is the proportion of missingness across each subject. A regression analysis of a single stacked data set, consisting of M imputations, produces unbiased estimates of regression coefficients. This is obviously less computationally extensive than performing variable selection on each individual imputed dataset.

With random forest (RF) imputation since a single imputed dataset is obtained, classical variable selection techniques can be applied. These classical variable selection techniques were also applied to both the full dataset (before missing data was induced) and to the complete cases.

5.3 Summary of Simulation Study

The different scenarios are summarized in `Table 5.4`. Results for simulations with MAR, MCAR and MNAR, with 10%, 20% and 30% missing in the explanatory variables and for varying sample size (n=1000, 700 and 100 respectively).

Simulation Scenario	Missing Data	Sample Size
1	MAR equal fractions of missing data (10%)	1000
2	MAR equal fractions of missing data (20%)	1000
3	MAR equal fractions of missing data (30%)	1000
4	MAR equal fractions of missing data (10%)	700
5	MAR equal fractions of missing data (10%)	100
6	MCAR equal fractions of missing data (10%)	1000
7	MCAR equal fractions of missing data (20%)	1000
8	MCAR equal fractions of missing data (30%)	1000
9	MCAR equal fractions of missing data (10%)	700
10	MCAR equal fractions of missing data (10%)	100
11	MNAR equal fractions of missing data (10%)	1000
12	MNAR equal fractions of missing data (20%)	1000
13	MNAR equal fractions of missing data (30%)	1000
14	MNAR equal fractions of missing data (10%)	700
15	MNAR equal fractions of missing data (10%)	100

Table 5.4: Different scenarios examined in the simulation study.

For each of the scenarios, the model selection approaches were run 1000 times and the predictors that were selected were recorded for each of the variable selection techniques. The model selection techniques were assessed by comparing the number of times the true and noise predictors were selected. The term "power" is used to indicate the probability that a method will correctly select a given variable from the true model and "type 1 error" to indicate the probability that a method will wrongly select a given variable not from the true model.

5.4 Results of Simulation Study

Since there are many different scenarios, detail here will be provided for just one scenario with the data MAR, 10% missing per predictor, n=1000. Results for the other scenarios are given in the `Appendix A.2`. The full results are given in `Table 5.5`, however it is easier to interpret the results from the plot in `Figure 5.4`. Ideally, it would be expected that the predictors in the true model should be always selected and the noise variables should not be selected into the final model, so we would like the techniques to have a high 'power' and low 'type 1 error'. Firstly, examining the performance of trees, it can be seen these generally have a low type 1 error except for when complete cases are used. However, the tree based methods do not seem to have a high power, especially in the case of one of the voting methods.

Examining the stepwise techniques, they all have a high power except when complete cases are used. However, the type 1 error is quite high for this technique, especially for one of the voting methods. Finally examining the backward variable selection technique, this method has high power except when it is applied to complete cases. Also it seems to have a low type 1 error except for the voting. Voting does not perform well overall.

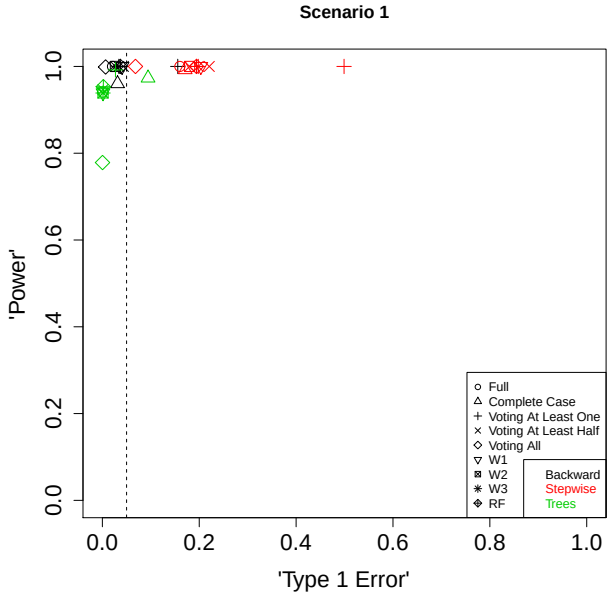
Summaries for each of the techniques are given in `Table 5.5`. Variable selection using `fastbw` (backward elimination) seems to perform the best with the average power for the full dataset is 1.00 and has an average type 1 error of 0.021 (range=0.019-0.023). In comparison, the complete cases model selection has lower power (average=0.958, range=0.834-1) and a slightly higher type 1 error (average=0.031, range=0.031-0.032) due to the casewise deletion of subjects with missing data. Using the voting method of variable selection for multiply imputed data with `fastbw`, if we choose the variables that appear at least once in the models, the power is better than that of the complete cases, however the trade off is a much larger type 1 error. Examining the voting where by choosing the variables that appear in at least half the models, again the power is better than that of the complete cases, however the type 1 error is still higher in comparison to that of the complete cases. The trade off with choosing the variables that appear in all models is a small reduction in power however it also decreases the type 1 error. The variable selection for stacking the imputed datasets and

using weighted regression has better power than that of the complete cases and similar type 1 error. Comparing the RF imputation variable selection to that of the complete cases, the power is higher for RF however the type 1 error is slightly increased.

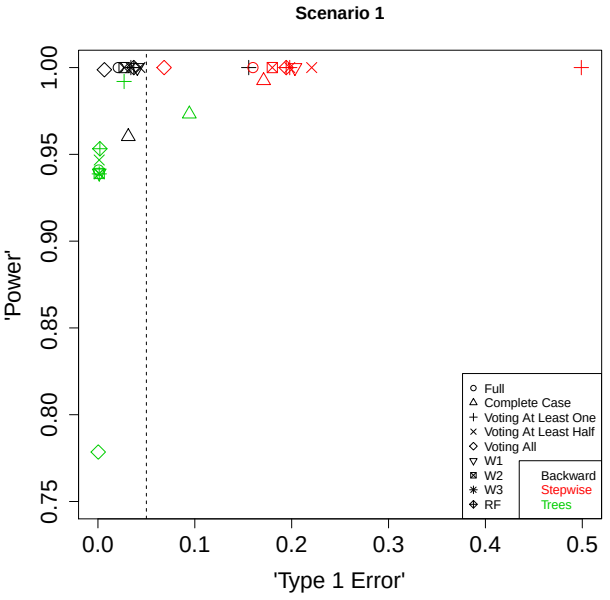
Also, the results from the other scenarios are similar to that of Scenario 1. By increasing the proportion of missing data in each variable, variable selection in imputed data performs much better than that of complete cases. The power and type 1 error are nearly as good as data with a lower proportion of missingness. Similar results are found for the other missing data mechanisms MCAR and MNAR.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.3	2.1	1.9
Backward	100	100	100	100	16.1	15.9	16.0
rpart	90.61	100	100	85.8	0.2	0	0
Complete Cases							
Fastbw	83.4	100	100	99.8	3.1	3.1	3.2
Backward	97.0	100	100	100	18.7	17.4	15.2
rpart	93.4	100	97.8	98.1	26.9	0.6	0.8
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	16.8	21.1	8.8
At Least Half	100	100	100	100	5.3	5.2	2.5
All	99.5	100	100	100	1.2	0.5	0.3
Backward							
At Least Once	100	100	100	100	53.8	62.0	33.9
At Least Half	100	100	100	100	23.3	24.6	18.3
All	100	100	100	100	7.1	6.1	7.3
rpart							
At Least Once	99.3	100	100	97.5	6.2	0.6	1.3
At Least Half	90.7	100	100	88.0	0.2	0.0	0.2
All	50.0	100	98.8	62.6	0.0	0.0	0.1
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	5.1	4.7	2.4
W2	100	100	100	100	3.7	3.3	1.1
W3	100	100	100	100	4.5	4.1	1.6
Backward							
W1	100	100	100	100	21.8	22.3	16.9
W2	100	100	100	100	19.2	19.9	14.9
W3	100	100	100	100	21.5	21.8	16.1
rpart							
W1	86.4	100	99.8	89.5	0.3	0.0	0.1
W2	86.3	100	99.8	89.5	0.3	0.0	0.1
W3	86.0	100	99.9	89.6	0.3	0.0	0.1
Imputation (RF)							
Fastbw	100	100	100	100	4.0	4.5	2.6
Backward	100	100	100	100	19.5	20.3	18.5
rpart	97.7	100	99.8	83.8	0.3	0.0	0.3

Table 5.5: Scenario 1: Number of times a variable was chosen by a simulation into the Survival Model (MAR and equal fractions of missing data (10% missing per variable) and sample size 1000). Average complete case sample size is 764.



(a) Scenario 1



(b) Scenario 1 (zoomed in)

Figure 5.4: Power and type 1 error for scenario one, MAR, sample size 1000 and 10% missing in each variable.

5.5 Random Forest Imputation Results

A small simulation study presented below compares the results of imputing data using random forests multiple times and imputing the data once using random forests (`Table 5.6 & Figure 5.5`). The data are imputed multiple times using random forest imputation. The same variable selection techniques, such as voting and stacking and weighting, are performed and then compared.

Examining the results from the graph in `Figure 5.5`, the trees and stepwise variable selection do not perform as well as backward variable selection. The blue points represent the variable selection on a single random forest imputation. It seems that, on average, the single random forest imputation performs just as well as using the random forest multiple times.

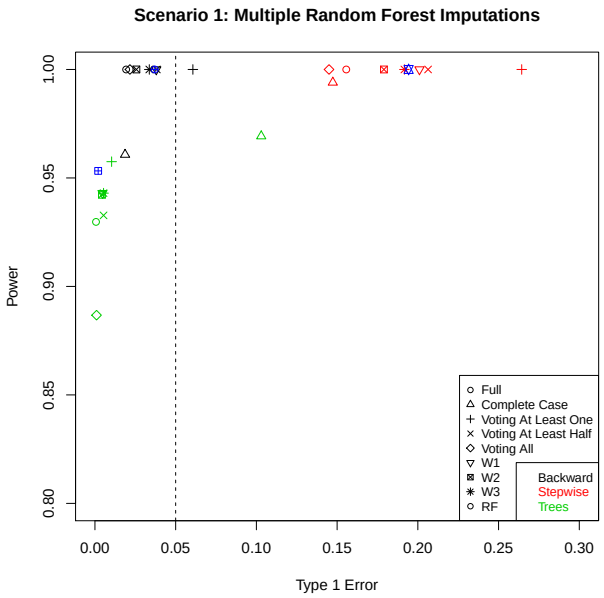


Figure 5.5: Number of times a variable was chosen by a simulation into the Survival Model using multiple random forest imputation (MAR, sample size 1000 and 10% missing in each variable). Blue points variable selection techniques using only a single imputation.

Variable selection in imputed data seems to perform better than that of complete cases. The techniques for variable selection in imputed data will now be applied to the BC data and the models examined so far will be compared in

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.0	2.2	1.6
Backward	100	100	100	100	16.6	16.3	13.8
rpart	89.4	100	100	82.5	0	0	0.2
Complete Cases							
Fastbw	84.7	100	100	99.6	2.4	2.0	1.2
Backward	97.6	100	100	100	15.1	16.0	13.1
rpart	93.6	100	97.2	96.9	29.7	0.5	0.7
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	6.4	7.7	4.1
At Least Half	100	100	100	100	4.1	5.1	2.3
All	100	100	100	100	2.3	3.2	1.0
Backward							
At Least Once	100	100	100	100	26.8	31.6	20.9
At Least Half	100	100	100	100	21.0	23.3	17.6
All	100	100	100	100	15.8	15.4	12.3
rpart							
At Least Once	91.6	100	99.9	91.5	0.9	1.2	1.0
At Least Half	86.5	100	99.9	86.7	0.4	0.5	0.7
All	79.1	100	99.6	76.0	0.1	0.1	0.1
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	4.3	5.1	2.0
W2	100	100	100	100	2.5	4.0	1.2
W3	100	100	100	100	3.5	4.6	2.0
Backward							
W1	100	100	100	100	20.4	22.6	17.3
W2	100	100	100	100	18.6	21.1	14.0
W3	100	100	100	100	20.2	21.8	15.5
rpart							
W1	87.3	100	99.7	90.2	0.3	0.5	0.6
W2	87.0	100	99.7	90.2	0.3	0.5	0.5
W3	87.3	100	99.7	90.2	0.4	0.5	0.7

Table 5.6: Scenario 1: Number of times a variable was chosen by a simulation into the Survival Model using multiple random forest imputation (MAR and equal fractions of missing data (10% missing per variable) and sample size 1000). Average complete case sample size is 764.

the next section.

5.6 Model Comparisons

A summary of the predictors selected using the various techniques for variable selection are summarized in `Tables 5.7-5.10` at the end of this chapter. The models can be compared using the concordance index which is a measure of performance (see Chapter 6). Although the models with all predictors perform slightly better than models were fewer predictors, the difference in the performance is just marginal. The cost of measuring the extra predictors, may also outweigh the slightly improved performance. The addition of the less routinely assessed biomarkers does not improve the performance much. Models with routinely assessed predictors seem more feasible since there is a cost associated with measuring the extra biomarkers. The final model for DFS will include Bilateral, Lymph Node status, Mitotic count, Metastasis and UICC staging and the final model for OS will include Mitotic count, Metastasis and UICC staging which is a subset of those predictors used for the final DFS model.

5.7 Conclusions

Results of this simulation study have been presented at the conference of the International Society for Clinical Biostatisticians [Wall et al., 2013].

The results of this simulation study suggest variable selection based on imputed data rather than on the complete cases is an attractive option for model selection. Complete cases fails to detect important predictors due to a lack of power. By using other methods the trade off is between power and type one error. By imputing the missing data and using the voting method, the voting selection is subjective (whether you take predictors that appear in one, half or all the models). If variables that appear in all models across all the imputations are chosen, the power is retained however there is a high type 1 error. If only variables that appear in all models are chosen, the power is reduced however the type 1 error is improved. Also the voting method is more computationally expensive, as the variable selection techniques need to be performed on each

imputed dataset. The voting method appears to be crude and does not perform well when the vote is being split between correlated predictors. The stacking method is 'easier', as the imputed datasets are stacked and only one analysis needs to be performed using the weights. The results of the simulation study has shown stacking and weighting has high power and low type 1 error. However, the first weight does not incorporate the amount of missing data present like weights 2 and 3 which provide more information on the proportion of missingness present. Also stacking means no matter how many imputations are performed only one analysis is needed for the variable selection.

Obviously nothing compares to complete data, however power can be "reclaimed" by using the variable selection techniques on multiple imputed data rather than complete cases and is an attractive alternative to complete cases.

In the previous chapter, classical approaches for variable selection could not be performed with complete cases since the sample size was reduced. Performing the variable selection on imputed data identified models with a subset of predictors which have good prediction power. Other predictors have been identified as potentially useful which were not identified in the previous chapters.

The final model for DFS should include Bilateral (Y/N), M (Mitotic count), LN status, Metastatic (Y/N) and UICC staging and for OS M, Metastatic, and UICC staging. The clinical pathological predictors do not seem to add anything extra to the models. The next stage in the process is to validate these models, which is examined in the next chapter.

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO	V1	V2	V3l	W1	W2	W3	RF Imputation
Age		X		X		X								
Bilateral		X		X		X	X	X	X	X	X	X	X	X
Grade				X		X								
Tubular Formation				X		X								
Nuclear Pleomorphism				X		X	X							
Mitotic Count				X		X		X	X	X	X	X	X	
Tumour Size				X		X	X	X			X			
Lymphovascular Invasion				X		X		X						
No of Lymph Nodes positive			X	X		X			X	X				
Lymph Node status	X	X	X	X	X	X	X	X	X	X	X	X	X	X
Nodal Ratio				X	X	X		X	X				X	X
Metastasis	X	X		X	X	X	X	X	X	X	X	X	X	X
Tumour Staging				X		X		X	X					
Metastasis Staging				X		X	X	X						
Lymph Node Staging			X	X	X	X								X
UICC			X	X		X	X	X	X	X	X	X	X	
NPI				X		X								
Oestrogen Status		X		X		X	X							
Progesterone Status				X		X	X							
Her2 Status			X	X		X								
Concordance Index	0.762	0.777	0.766	0.826	0.781	0.788	0.802	0.812	0.810	0.801	0.806	0.810	0.807	0.807

Table 5.7: Summary of clinical predictors selected through various techniques for DFS. X means the predictor is included in the model. V1, V2 and V3 are the voting system in multiple imputation, choosing the predictors that appear in at least 1, at least half or all the models. W1, W2, and W3 are the weights (`Section 5.2.6`).

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO	V1	V2	V3l	W1	W2	W3	RF Imputation
Age		X		X		X	X							
Bilateral		X		X		X	X	X	X	X	X	X	X	X
Grade				X		X								
Tubular Formation				X		X								
Nuclear Pleomorphism				X		X	X							
Mitotic Count				X		X		X	X	X	X	X	X	
Tumour Size			X	X		X	X	X	X		X	X	X	
Lymphovascular Invasion				X		X		X						
No of Lymph Nodes positive			X	X		X	X	X	X		X	X	X	
Lymph Node status	X	X	X	X		X	X	X	X	X	X	X	X	X
Nodal Ratio			X	X		X		X	X		X	X	X	X
Metastasis	X	X		X		X	X	X	X	X	X	X	X	X
Tumour Staging				X		X		X	X		X			
Metastasis Staging				X		X	X	X						
Lymph Node Staging			X	X		X	X	X						X
UICC				X		X	X	X	X	X	X	X	X	X
NPI				X		X								
Oestrogen Status		X		X		X	X							
Progesterone Status				X		X	X	X						
Her2 Status				X		X								
Bcl2 Status				X		X	X							
CK14 Status				X		X	X							
CK5/6 Status				X		X	X	X						X
EGFR Status				X		X								
Ki67 Status				X		X								
p53 Status				X		X								
E-cad Status				X		X	X	X	X					
tMcm2 Status				X		X	X	X						
CDC7				X		X		X						
pMcm2				X		X	X	X						
Concordance Index	0.762	0.777	0.765	0.884	NA	0.841	0.819	0.872	0.802	0.801	0.816	0.811	0.811	0.814

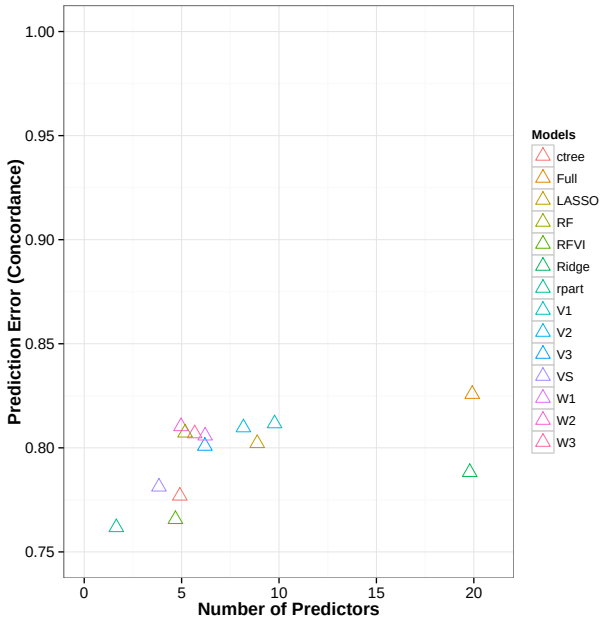
Table 5.8: Summary of clinical and pathological predictors selected through various techniques for DFS. X means the predictor is included in the model. V1, V2 and V3 are the voting system in multiple imputation, choosing the predictors that appear in at least 1, at least half or all the models. W1, W2, and W3 are the weights (`Section 5.2.6`).

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO	V1	V2	V3l	W1	W2	W3	RF Imputation
Age				X		X								
Bilateral				X		X	X							
Grade				X		X								
Tubular Formation				X		X								
Nuclear Pleomorphism			X	X		X								
Mitotic Count				X		X		X	X	X	X	X	X	X
Tumour Size			X	X		X		X						
Lymphovascular Invasion				X		X		X						
No of Lymph Nodes positive			X	X		X		X						X
Lymph Node status	X		X	X		X		X						
Nodal Ratio			X	X		X		X						
Metastasis	X	X		X		X	X	X	X	X	X	X	X	X
Tumour Staging			X	X		X		X						
Metastasis Staging			X	X		X	X							
Lymph Node Staging			X	X		X		X						
UICC		X	X	X		X		X	X	X	X	X	X	
NPI				X		X		X						
Oestrogen Status		X	X	X		X	X	X						X
Progesterone Status				X		X	X	X						
Her2 Status			X	X		X		X						
Concordance Index	0.909	0.922	0.911	0.969	NA	0.952	0.919	0.958	0.929	0.929	0.929	0.929	0.929	0.923

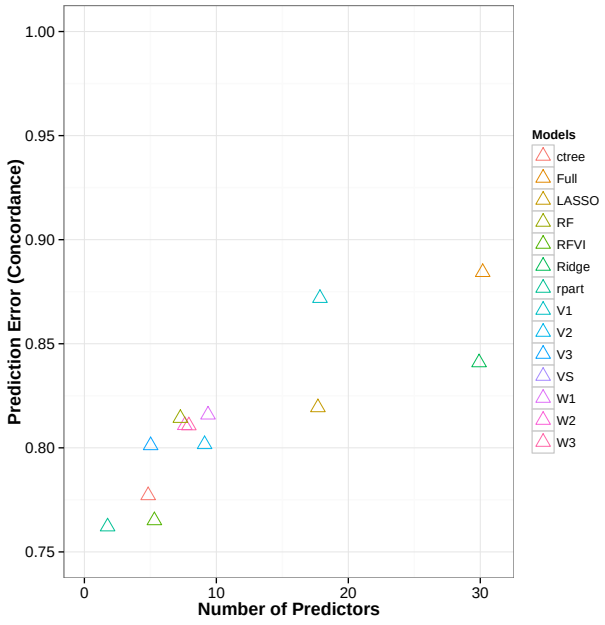
Table 5.9: Summary of clinical predictors selected through various techniques for OS. X means the predictor is included in the model. V1, V2 and V3 are the voting system in multiple imputation, choosing the predictors that appear in at least 1, at least half or all the models. W1, W2, and W3 are the weights (`Section 5.2.6`).

Predictors	CART Tree	Conditional Inference	Random Forest	Full	Variable Selection	Ridge	LASSO	V1	V2	V3l	W1	W2	W3	RF Imputation
Age				X		X								
Bilateral			X	X		X	X							
Grade				X		X		X						
Tubular Formation			X	X		X								
Nuclear Pleomorphism			X	X		X								
Mitotic Count				X		X		X	X	X	X	X	X	X
Tumour Size			X	X		X	X	X	X					
Lymphovascular Invasion			X	X		X		X						
No of Lymph Nodes positive			X	X		X								
Lymph Node status	X		X	X		X		X						
Nodal Ratio			X	X		X								
Metastasis	X	X	X	X		X		X	X	X	X	X	X	X
Tumour Staging				X		X		X						
Metastasis Staging			X	X		X	X							
Lymph Node Staging			X	X		X		X						
UICC		X	X	X		X	X	X	X	X	X	X	X	X
NPI			X	X		X		X						
Oestrogen Status		X	X	X		X	X	X						X
Progesterone Status			X	X		X	X	X						
Her2 Status			X	X		X								
Bcl2 Status			X	X		X								
CK14 Status			X	X		X	X							
CK5/6 Status			X	X		X	X							
EGFR Status				X		X		X						
Ki67 Status				X		X		X	X					
p53 Status			X	X		X								
E-cad Status			X	X		X	X							
tMcm2 Status			X	X		X	X							
CDC7				X		X		X	X					
pMcm2				X		X								
Concordance Index	0.909	0.922	0.790	0.995	NA	0.998	0.895	0.981	0.927	0.929	0.929	0.929	0.929	0.931

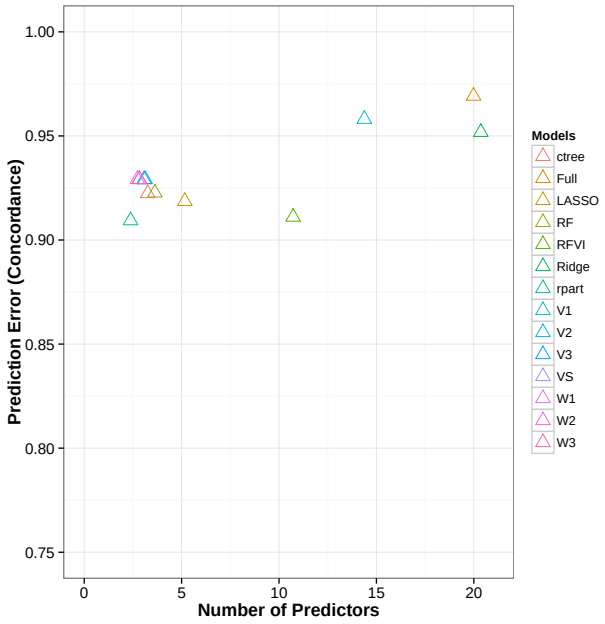
Table 5.10: Summary of clinical and pathological predictors selected through various techniques for OS. X means the predictor is included in the model. V1, V2 and V3 are the voting system in multiple imputation, choosing the predictors that appear in at least 1, at least half or all the models. W1, W2, and W3 are the weights (`Section 5.2.6`).



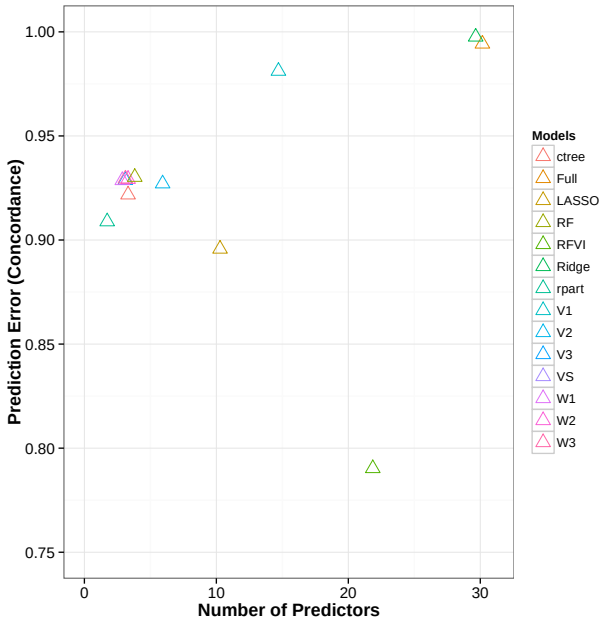
(a) DFS Clinical Models



(b) DFS Clinical & Pathological Models



(c) OS Clinical Models



(d) OS Clinical & Pathological Models

Figure 5.6: Comparison of concordance index for different models.

Chapter 6

Model Validation and Calibration

6.1 Introduction

In the previous chapter useful predictors of DFS and OS have been identified using variable selection techniques on imputed data. Also the final section in the chapter summarized the predictors selected using all the different techniques including trees, classical variable selection processes such as backward, Ridge Regression and the LASSO and variable selection with imputed data. Although the models with all predictors, penalized as needed, perform slightly better, the difference in the performance is marginal. The cost of measuring the extra predictors may outweigh the slightly improved performance. Based on all the analysis and results to date, a final model with *Bilateral, Lymph Node status, Mitotic count, Metastasis and UICC staging* seems to be the best model for DFS and a final model with *Mitotic count, Metastasis and UICC staging* seems to be the best model for OS. The models are summarized in `Table 6.1`. The next step in the process is to validate and test these models.

There can be 2 types of validated models according to Altman and Royston [2000]:

- A *statistically validated model* is one which passes all appropriate statistical checks, including goodness-of-fit on the original data set and unbiased

Table 6.1: Final Cox proportional hazards models for DFS and OS.

	Response variables:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Bilateral=Yes	1.356^{***}	
	(0.279)	
Mitotic Count=Low	-0.385^{**}	-0.709^{**}
	(0.161)	(0.302)
Mitotic Count=Moderate	0.033	0.374
	(0.211)	(0.411)
LN Status=Yes	1.389^{***}	
	(0.218)	
Metastasis=Yes	1.448^{***}	3.765^{***}
	(0.146)	(0.527)
UICC=2	0.503	0.230
	(0.320)	(0.571)
UICC=3	0.633^{*}	1.118^{**}
	(0.337)	(0.555)
UICC=4	1.044^{***}	1.942^{***}
	(0.381)	(0.584)
Observations	432	444
R^2	0.449	0.435
χ^2	255.975^{***} (df = 8)	180.939^{***} (df = 6)
Concordance Index	0.810	0.929
Note:		*p<0.1; **p<0.05; ***p<0.01

prediction on a new data set.

- A *clinically validated model* is one which performs satisfactorily on a new data set according to context-dependent statistical criteria laid down.

The utility of predictive models depends on their generalizability, which can be separated into two components: **internal validity** (reproducibility) and **external validity** (transportability - the model is valid in other breast cancer data). Internal validation is when the model retains accuracy when it is applied to another set of patients from the same underlying population as that of the development sample. Internal validity can be completed by splitting the sample into a development and test set, cross-validation or bootstrapping. External validation can be checked using a new sample of patients in different settings. External validation is when the model retains accuracy when applied to patients from a different population or location. The idea of validating a prognostic or diagnostic model is generally taken to mean establishing that it works satisfactorily for patients other than those from whose data the model was derived [Altman and Royston, 2000].

In general, there are two aspects of predictive accuracy that need to be assessed. Firstly, calibration is the ability of the model to make unbiased estimates of the outcome and secondly discrimination is the model's ability to separate subject's outcomes [Harrell Jr, 2001].

The simplest form of internal validation is data splitting. The data are randomly split into a training and test set. Bootstrapping, jackknifing and their resampling plans can be used to obtain nearly unbiased estimates of model performance without sacrificing sample size.

6.1.1 Missing data effect

The final models presented in `Table 6.1` were fitted using complete cases. By imputing the data and fitting the same model on the imputed data, the effect and the sensitivity of the missing data generating mechanism can be examined. The data have been imputed using Random Forests and MICE.

The model fitted on the random forest imputation is given in `Table 6.2`. The estimates have the same direction and similar magnitude to those obtained

from complete cases. There is a marginal difference in the performance of the models measured by examining the concordance.

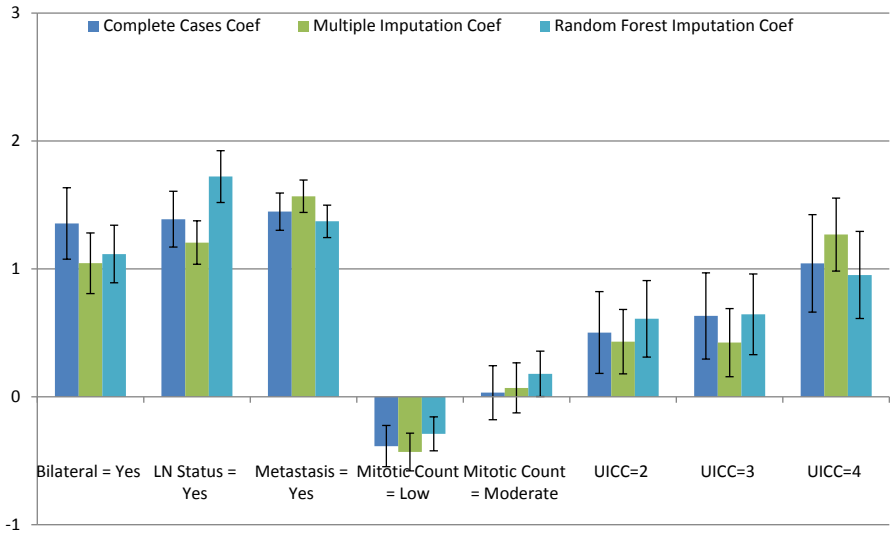
The final model was also fitted on the multiply imputed data by using MICE. The corresponding estimates were combined using Rubins Rules (`Table 6.3`). These estimates again have the same direction and similar magnitude to those obtained from complete cases. The performance is measured as the average concordance across all the models. Again there is a marginal difference in the performance of the models. `Figure 6.1` contains parameter estimates and estimated standard errors for the final models using complete cases, multiply imputed data and random forests.

Table 6.2: Final Cox proportional hazards models for DFS and OS fitted on Random Forest imputed data.

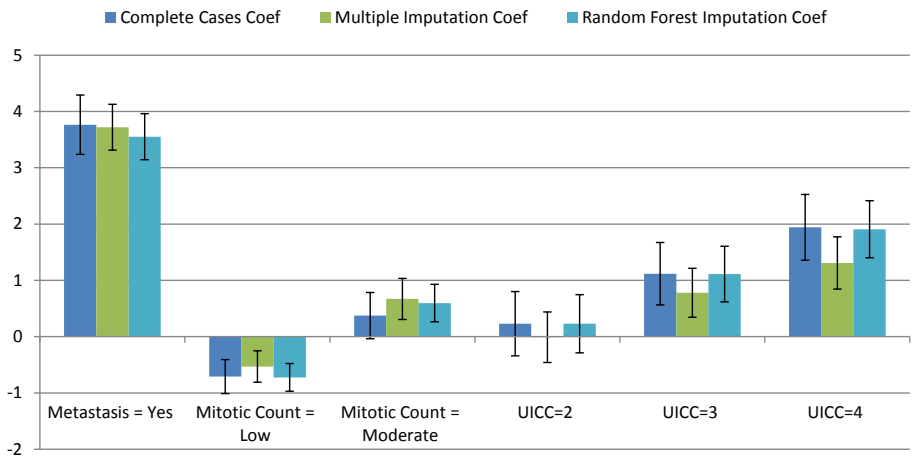
	Response variables:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Bilateral = Yes	1.116^{***} (0.225)	
Mitotic Count = Low	-0.289^{**} (0.133)	-0.723^{***} (0.246)
Mitotic Count = Moderate	0.180 (0.178)	0.597^{*} (0.333)
LN Status = Yes	1.722^{***} (0.202)	
Metastasis = Yes	1.372^{***} (0.127)	3.552^{***} (0.409)
UICC=2	0.610^{**} (0.299)	0.229 (0.517)
UICC=3	0.645^{**} (0.316)	1.112^{**} (0.494)
UICC=4	0.953^{***} (0.341)	1.908^{***} (0.507)
Observations	647	647
R^2	0.483	0.339
Concordance Index	0.812	0.933
Note:		*p<0.1; **p<0.05; ***p<0.01

6.2 Internal Validation Techniques

The main internal validation techniques include apparent validation, split-sample validation, cross-validation and bootstrap validation. These will be discussed in more detail in the next few sections.



(a) Disease Free Survival



(b) Overall Survival

Figure 6.1: Parameter estimates and corresponding estimated standard errors for the final models.

Table 6.3: Combined Estimates using Rubins Rules for the final Cox proportional hazards models for DFS and OS fitted on multiply imputed data.

	Response variables:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Bilateral = Yes	1.045*** (0.237)	
Mitotic Count = Low	−0.431*** (0.147)	−0.530* (0.279)
Mitotic Count = Moderate	0.070 (0.196)	0.670* (0.365)
LN Status = Yes	1.206*** (0.170)	
Metastasis = Yes	1.568*** (0.127)	3.720*** (0.407)
UICC=2	0.432* (0.252)	−0.010 (0.449)
UICC=3	0.424 (0.266)	0.779* (0.435)
UICC=4	1.269*** (0.285)	1.309*** (0.464)
Observations	647	647
Average Concordance	0.790	0.926
Note:		*p<0.1; **p<0.05; ***p<0.01

6.2.1 Apparent Validation

Apparent validation uses the data used to create the model to validate the model. It is well known that validation to test the performance of the model using the data that were used to create the model leads to biased assessment of performance since the model is created and tested on the same data. Naturally this leads to an optimistic estimate of performance since the model parameters were optimized for the sample [Steyerberg et al., 2001]. Apparent validation is attractive because it is easy to perform however the estimates are biased. For the BC data, this is not a good validation technique, as one aim is to make the models transferable to other centers.

6.2.2 Split-Sample Validation

In split-sample validation, the data are split into two, a development set and a test set. The model is created using the development set and the model performance is evaluated using the test set. Data splitting has the advantage of allowing hypothesis tests to be confirmed in the test sample [Harrell Jr, 2001]. However, data splitting has the following disadvantages:

- Decreases the sample size for both model development and model testing;
- Requires a larger sample to be held out than cross-validation to be able to obtain the same precision of the estimate of predictive accuracy;
- May yield different predictive accuracy when repeated;
- Does not validate the final model, but rather a model developed on only a subset of the data. The training and test sets are recombined for fitting the final model, which is not validated;
- Requires the split before the first analysis of the data. With other methods, analyses can proceed in the usual way on the complete dataset. Then, after a "final" model is specified, the modelling process is re-run on multiple resamples from the original data to mimic the process that produced the "final" model.

6.2.3 Cross-Validation

Cross-validation is an extension of split-sample validation aiming for more stability. A prediction model is again tested on a random subset of individuals from the sample that was left out from the sample. The model is developed on the remaining part of the sample. However this process is repeated for consecutive fractions of patients [Steyerberg et al., 2001].

6.2.4 Bootstrap Validation

Bootstrap resamples can be used to estimate the bias due to overfitting or optimism in the final model. A sample of size n (same size as original dataset) is drawn at random with replacement. The model is fitted in the bootstrap resample and applied to the original sample. The accuracy is calculated as the accuracy index for the bootstrap minus the accuracy index for the original sample. This is repeated for multiple bootstrap replications to estimate the optimism. This is then subtracted from estimate of performance of the model to correct for overfitting. Overfitting causes optimism about a model performance in new subjects [Steyerberg, 2009]. Bootstrapping seems to work well in large datasets, even if the number of predictors exceeds the number of samples.

For our final model we will use bootstrap validation for internal validation.

6.3 External Validation

There are a few different ways of validating the models externally. Temporal validation uses patients who are more recently diagnosed to validate the model. This is not possible here, as more recent patients are not diagnosed long enough to examine five year survival. Alternatively the model is fully validated by independent investigators. Geographical Validation uses data from another site/hospital to validate the model. ONCOPOOL is a dataset with patients from 10 breast cancer units in Europe.

ONCOPOOL

ONCOPOOL is a retrospectively compiled database of primary operable invasive breast cancers treated in the 1990s in 10 European breast cancer units [Blamey et al., 2010]. There are sixteen thousand, nine hundred and forty four patients included in the data set. The patients are all women, have tumours less than 5 cm and are aged 70 or less. The patients were diagnosed between 1990 and 1999. This dataset does not contain the time and censoring information or all the variables needed for validation of DFS however it does contain enough information for validation of OS.

6.4 Evaluation of Performance

The evaluation of model performance focused on discrimination and calibration. Discrimination refers to the ability to distinguish high-risk patients from low risk patients and is commonly quantified by a measure of concordance, the c index. Discrimination measures a predictors ability to separate patients with different responses [Harrell Jr, 1996]. The definition of the concordance probability C is based on the property that a survival model should predict a lower survival time for subjects that fail earlier and a higher survival time for subjects that fail later [Van Oirbeek and Lesaffre, 2010]. The concordance probability C is

defined as

$$C = P(\hat{T}_i < \hat{T}_j | T_i < T_j) = P(S(t|\mathbf{X}_i) < S(t|\mathbf{X}_j) | T_i < T_j) \text{for any } t > 0 \quad (6.1)$$

where $\hat{T}$ is the predicted survival time. Calibration refers to whether the predicted probabilities agree with the observed probabilities. We used one simple measure to quantify calibration, that is, the slope of the prognostic index, which was originally proposed by Cox. The slope of the prognostic index (or linear predictor) is the regression coefficient β in a logistic model with the prognostic index as the only covariate: observed mortality $= \alpha + \beta$ prognostic index. Calibration refers to the extent of bias [Harrell Jr, 1996].

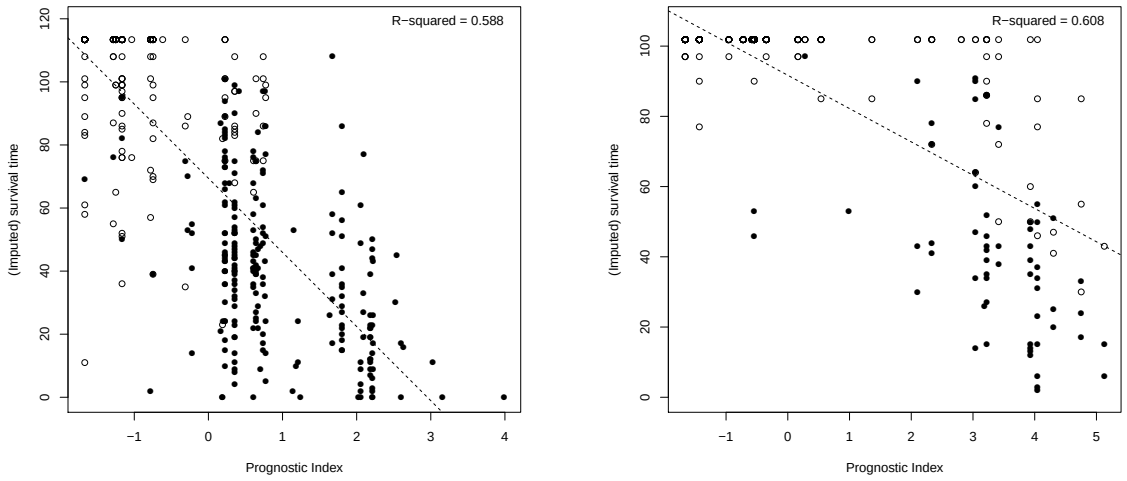
The observed mortality is a variable coded as binary (0/1) and the prognostic index is calculated as the linear combination of the regression coefficients as estimated in the subsample with the values of the predictors for each patient in the test data. The slope of the prognostic index (referred to as the calibration slope) should ideally be one, when the predicted risks agree fully with the observed frequencies. Models providing overoptimistic predictions will show a slope that is less than one, indicating that low predictions are too low and high predictions are too high.

6.4.1 Visualising the relation between predictor and survival

Royston [2001] proposed visualisation by adapting the scatterplot for survival data where the survival times are plotted against the predicted risk from a Cox PH model. The adaption is needed because such a plot is not possible due to the presence of censored data, where the exact survival time is not available. Royston [2001] suggests fitting a parametric lognormal model to the data and then imputing censored observations by randomly drawing a value from the conditional distribution. This can lead to some unrealistic imputations (i.e. survival times that are not biologically plausible). However this can easily be rectified by putting a limit on the imputation value for the censored observations. The limit is chosen by examining the longest survival time observed.

The scatterplot of the survival times and the prognostic index score is given in `Figure 6.2`. The explained variation (R^2) is 0.588 and 0.608 for DFS and

OS respectively. Although this is not the most appropriate quantification, these values indicate the models are providing relatively good prediction for DFS and OS. From the plots it seems that those patients with longer survival times have lower prognostic scores. This means patients with a higher prognostic score have a worse prognosis.



(a) Disease Free Survival

(b) Overall Survival

Figure 6.2: Partly imputed survival time verses prognostic index. Imputed values are represented by open circles, observed times by solid circles.

6.4.2 Internal Validation of Final Models

Bootstrapping was used to identify if there was significant over-fitting during the development of the model (see section 6.2.4). The accuracy of the model was assessed with measures of discrimination and calibration. Predictive discrimination (i.e. the ability of a predictive model to separate those who die early from those who die late) was assessed using the Somers' D_{xy} rank correlation coefficient. This measure quantifies the association between predicted and observed survival time [Teno, 2000]. A model which can predict low and high risk patients perfectly would have a Somers' D_{xy} of one and a weak model a value close to zero.

The models are validated using Somers' D_{xy} rank correlation between the

Table 6.4: Validation Results for D_{xy} and slope shrinkage for disease free survival using 200 bootstrap resamples of the data.

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.6029	0.6127	0.5987	0.0140	0.5889	200
R^2	0.4488	0.4659	0.4328	0.0331	0.4157	200
Slope	1.0000	1.0000	0.8938	0.1062	0.8938	200
D	0.1057	0.1119	0.1007	0.0112	0.0946	200
U	−0.0008	−0.0008	0.0046	−0.0054	0.0046	200
Q	0.1066	0.1127	0.0962	0.0165	0.0900	200
g	1.4054	1.5005	1.3352	0.1654	1.2401	200

predicted log hazard and observed survival time and for slope shrinkage. The bootstrap is used (with 200 resamples) to penalize for possible overfitting.

Examining D_{xy} and R^2 for the DFS model in `Table 6.4`, it can be seen that there is a small amount of overfitting. A shrinkage coefficient can be used to quantify over-fitting or one can go a step further and use the coefficient to re-calibrate the model [Harrell Jr, 1996]. The slope has a shrinkage factor of 0.89 which does not cause concern. The apparent D_{xy} is 0.60, however the corrected D_{xy} of 0.59 is an unbiased estimate of future predictive discrimination on similar patients. The apparent D_{xy} is only marginally optimistic.

Statistics validated include the Nagelkerke R^2, D_{xy} slope shrinkage, the discrimination index $D = (modelL.R.\chi^2 - 1)/L$ (where L is -2 log likelihood with $\beta = 0$), the unreliability index U=(difference in -2 log likelihood between uncalculated $X\beta$ and $X\beta$ with overall slope calibrated to test sample)/L, the overall quality index $Q = D - U$ and g is the g-index on the log relative hazard scale.

For the OS model, see `Table 6.5`, again there is a small amount of overfitting but not enough to be concerned about. The slope has a shrinkage factor of 0.90 which does not cause concern. The apparent D_{xy} is 0.86 while the corrected D_{xy} of 0.85 is an unbiased estimate of future predictive discrimination on similar patients. The D_{xy} is the difference between the probability of concordance and the probability of discordance of pairs of predicted survival times and pairs of observed survival times, accounting for censoring [Harrell Jr, 2001]. If this falls below 0.85, for example, it may cause some concern. However for both models this is above 0.85.

Table 6.5: Validation Results for D_{xy} and slope shrinkage for overall survival using 200 bootstrap resamples of the data.

Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.8582	0.8626	0.8524	0.0102	0.8480	196
R^2	0.4350	0.4452	0.4221	0.0231	0.4119	196
Slope	1.0000	1.0000	0.8991	0.1009	0.8991	196
D	0.2762	0.2858	0.2662	0.0196	0.2566	196
U	−0.0031	−0.0031	0.0059	−0.0091	0.0060	196
Q	0.2793	0.2889	0.2603	0.0287	0.2506	196
g	1.9698	2.1916	1.9322	0.2594	1.7105	196

Next the final models need to be validated (without the shrinkage coefficient) for calibration accuracy in predicting the probability of surviving five years. The bootstrap is used to estimate the optimism in how well predicted five year survival from the final Cox models tracks Kaplan Meier five year estimates, stratifying by grouping the patients with about 40 patients per interval of predicted five year survival. The results for calibration are given in `Tables 6.6-6.7 and Figure 6.3`. Perfect calibration would be indicated with a 45 degree line [Teno, 2000]. For all ranges of estimates, the model prediction of survival estimates are adequate (the CI contains the line at 45 degrees). Bootstrap validation performs well except for some of the groups with poorer prognosis - their survival is slightly better than predicted.

6.4.3 Measuring the Discriminative Ability

A useful summary statistic is the *area under the curve (AUC)*. If the AUC is equal to 0.5, the model has no discriminative ability. An interesting comparison is to compare those patients who experience the event and those who do not at each time point (this is similar to the concordance index). A plot of the AUC(t) against t (at the event times) for the prognostic indices for DFS and OS are shown in `Figure 6.7`. This includes a plot of the lowess smoothing curve. The AUC for DFS and OS are 0.60 and 0.62 respectively.

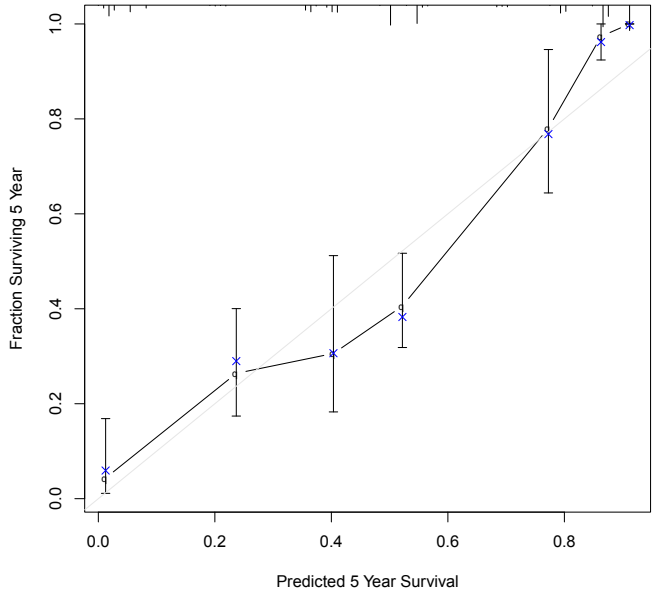
The concordance for the DFS and OS final models are 0.801 and 0.929 respectively. Dynamic versions of the concordance index can be obtained by averaging over all event times within a fixed window of time. This is shown in `Figure 6.5` for the window of w=60 (five years). Both figures for DFS (`Figure 6.4(a) and`

Table 6.6: Calibration results for Disease Free Survival.

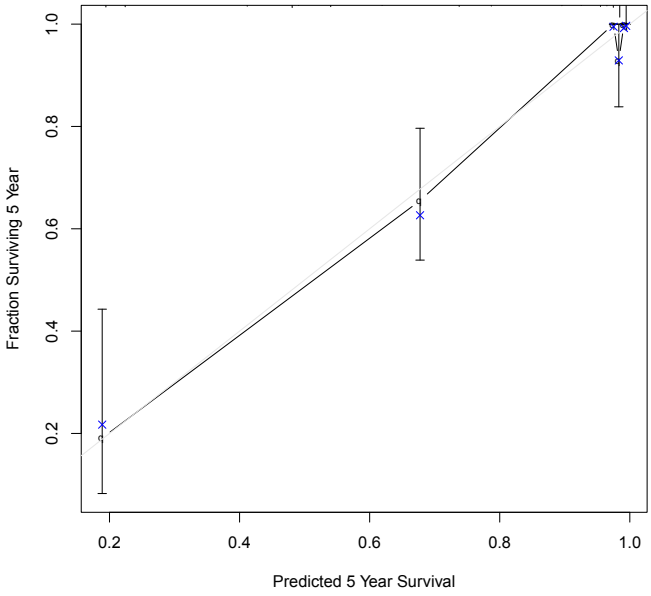
index original	training	test	mean optimism	mean corrected	n	mean predicted	KM	KM corrected	Standard Error
0.0310	0.0234	0.0425	−0.0158	0.0468	200	0.0125	0.0435	0.0593	0.6916
0.0271	0.0289	0.0327	−0.0258	0.0530	200	0.2368	0.2639	0.2898	0.2125
−0.0977	−0.1574	−0.1381	−0.0006	−0.0971	195	0.4036	0.3059	0.3064	0.2628
−0.1162	−0.0724	−0.1221	0.0230	−0.1392	200	0.5219	0.4057	0.3827	0.1236
0.0080	0.0421	0.0373	0.0127	−0.0046	197	0.7726	0.7806	0.7679	0.0981
0.1108	0.1155	0.1153	0.0118	0.0989	133	0.8629	0.9737	0.9619	0.0267
0.0881	0.0915	0.0836	0.0024	0.0857	196	0.9119	1.0000	0.9976	0.0000

Table 6.7: Calibration results for Overall Survival.

index original	training	test	mean optimism	mean corrected	n	mean predicted	KM	KM corrected	Standard Error
0.0025	0.0424	0.0505	−0.0260	0.0285	189	0.1888	0.1913	0.2172	0.4284
−0.0224	−0.0064	−0.0385	0.0286	−0.0511	200	0.6775	0.6551	0.6264	0.0997
0.0250	−0.1080	−0.0935	0.0052	0.0198	177	0.9750	1.0000	0.9948	0.0000
−0.0538	0.0154	0.0154	0.0002	−0.0540	87	0.9831	0.9293	0.9291	0.0525
0.0090	−0.0055	−0.0035	0.0077	0.0013	187	0.9910	1.0000	0.9923	0.0000
0.0056	−0.0138	−0.0364	0.0033	0.0023	87	0.9944	1.0000	0.9967	0.0000
	0.0035	0.0016	0.0027		145				



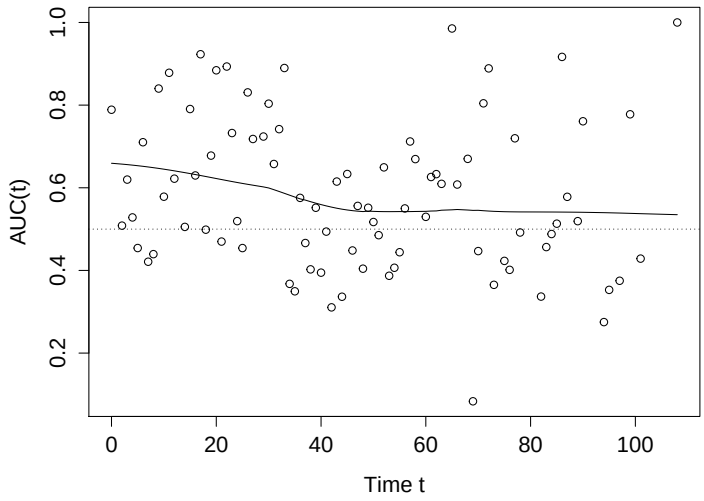
(a) Disease Free Survival



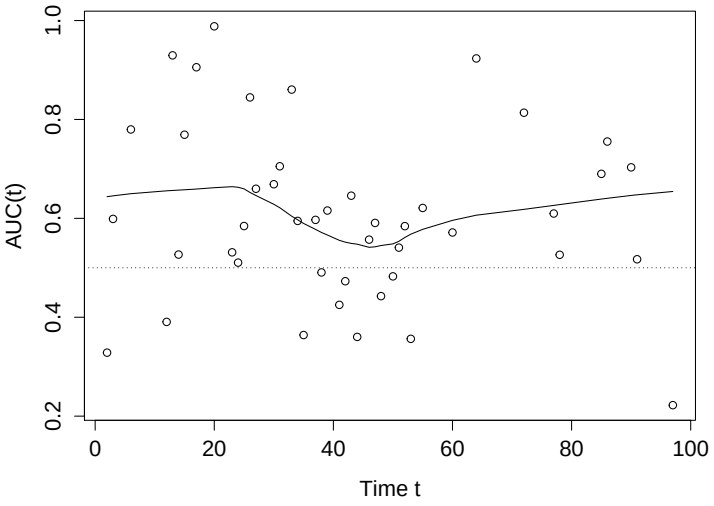
(b) Overall Survival

Figure 6.3: Bootstrap estimate of calibration accuracy for five-year estimates from the final Cox model for DFS and OS. Dots correspond to apparent predictive accuracy. X marks the bootstrap corrected estimates. (n=444, 60 patients per group, 200 bootstrap replicates)

`6.5(a)`) show that the discriminative ability is slowly decreasing over time. For OS (`Figure 6.4(b) and 6.5(b)`), the discriminative ability decreases for the first three years, however, a strange feature is that it starts to increase slightly again.

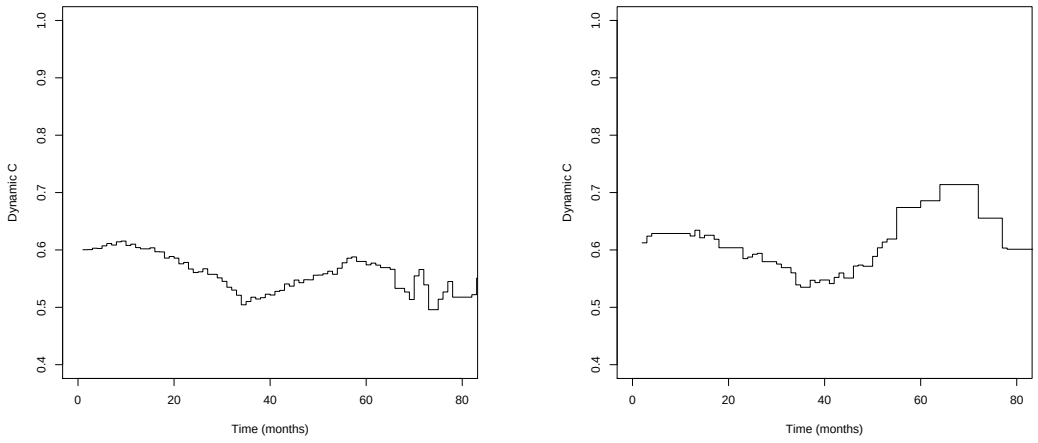


(a) Disease Free Survival



(b) Overall Survival

Figure 6.4: AUC(t) for the final Cox models for DFS and OS.



(a) Disease Free Survival

(b) Overall Survival

Figure 6.5: Dynamic C index with a window of five years for the final Cox models for DFS and OS.

6.5 DFS Model with Interactions

The models that have been discussed so far in this chapter have only included main effects. However, in Chapter 4, interactions seem to be important for the DFS model. Interactions have been added to the main effects DFS model discussed previously in this chapter (`Table 6.8`). The concordance index for this model is 0.816 which is only marginally better performance than the main effects model.

6.5.1 Validation of DFS Model with Interactions

Bootstrap validation techniques will be applied to assess the performance of the DFS model with interactions. Examining D_{xy} and R^2 for the DFS model in `Table 6.9`, it can be seen that there is a small amount of overfitting. The slope has a shrinkage factor of 0.86 which does not cause concern since it is above 0.85. The apparent D_{xy} is 0.62, while the corrected D_{xy} of 0.61 is an unbiased estimate of future predictive discrimination on similar patients. The apparent D_{xy} is only marginally optimistic. This is marginally more optimistic than the model with only main effects.

Table 6.8: Final Cox proportional hazards model for DFS including interactions.

	Response variable:
	DFS
	$\hat{\beta}(ESE)$
Bilateral=Yes	3.200^{***} (0.429)
Mitotic Count=Low	-0.436^{***} (0.166)
Mitotic Count=Moderate	0.098 (0.212)
LN Status=Yes	2.607^{***} (0.385)
Metastasis=Yes	3.751^{***} (0.577)
UICC=2	0.697 (0.541)
UICC=3	0.713 (0.562)
UICC=4	0.768 (0.671)
LN Status=Yes ∗Metastasis=Yes	-1.933^{***} (0.484)
Bilateral=Yes ∗LN Status=Yes	-2.661^{***} (0.624)
Metastasis=Yes ∗UICC=2	-1.165^{*} (0.700)
Metastasis=Yes ∗UICC=3	−0.605 (0.727)
Metastasis=Yes ∗UICC=4	0.145 (0.841)
Observations	432
R^2	0.514
Concordance Index	0.816
Note:	*p<0.1; **p<0.05; ***p<0.01

The results for calibration are given in `Figure 6.6`. Recall from earlier perfect calibration would be indicated with a 45 degree line. For all ranges of estimates, the model prediction of survival estimates are adequate (the CI contains the line at 45 degrees). Bootstrap validation performs well except for some of the groups with poorer prognosis - their survival is slightly better than predicted.

The performance is very similar to that of the DFS model with just main effects.

The AUC for the DFS model with interactions is 0.614 (`Figure 6.7(a)`). This is marginally better than that of the main effects model ($AUC_{main\ effects} = 0.60$).

The concordance for the main effects DFS model is 0.801 and this increases to 0.816 when interactions are included. The dynamic version of this concordance index for the model with interactions is shown in `Figure 6.7(b)` for the window of w=60 (five years). The discriminative ability behaves just as in the previous

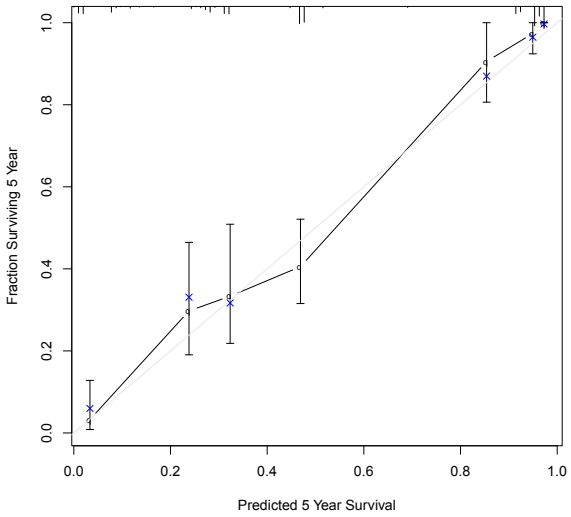


Figure 6.6: Bootstrap estimate of calibration accuracy for five-year estimates from the final Cox model for DFS with interactions. Dots correspond to apparent predictive accuracy. X marks the bootstrap corrected estimates.

model.

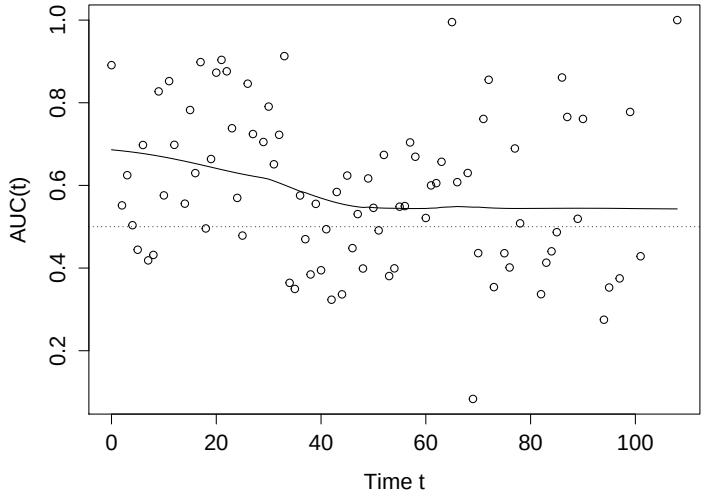
The addition of interactions into the DFS model marginally improves the performance of the DFS model.

Table 6.9: Validation Results for D_{xy} and slope shrinkage for DFS model with interactions using 200 bootstrap resamples of the data.

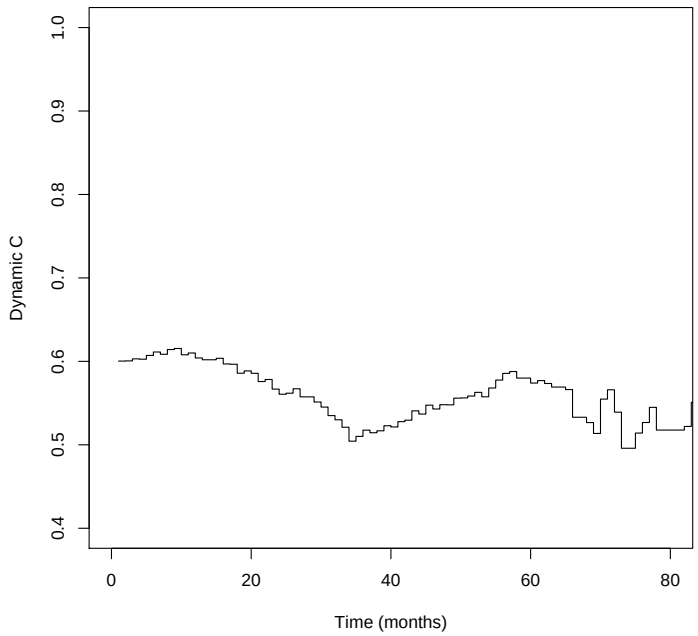
Index	Original Sample	Training Sample	Test Sample	Optimism	Corrected Index	n
D_{xy}	0.6327	0.6360	0.6158	0.0202	0.6125	192
R^2	0.5144	0.5317	0.4928	0.0388	0.4756	192
Slope	1.0000	1.0000	0.8627	0.1373	0.8627	192

6.6 External Validation of Final Models

The ONCOPOOL dataset is used for the external validation of the OS model. These data cannot be used to validate the DFS model as information on DFS time, status, Bilateral and Lymph Node status are not available in the dataset.

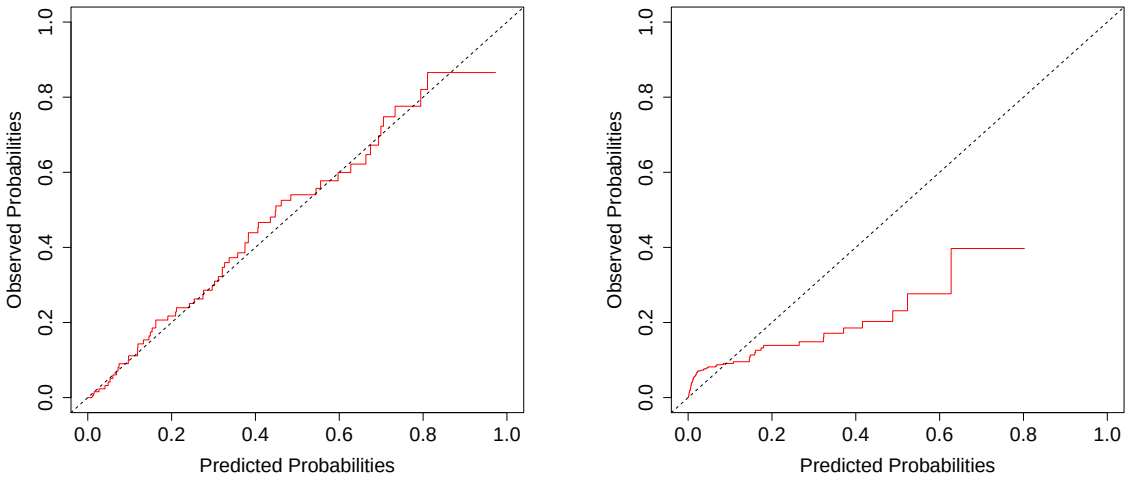


(a) AUC(t)



(b) Dynamic C

Figure 6.7: Final Cox models for DFS with interactions.



(a) BC data

(b) ONCOPOOL data

Figure 6.8: Plot of observed and fitted probabilities for the OS model.

`Figure 6.8(a)` contains a plot of the observed versus the predicted probabilities calculated for five year survival from the OS model on the original BC data. The observed and predicted probability is very similar as the estimates lie along the line of equality. `Figure 6.8(b)` contains a plot of the observed versus predicted probabilities calculated for five year survival from the OS model for the ONCOPOOL data. The predicted probabilities for the ONCOPOOL patients calculated using the OS model are higher than the observed probabilities.

One cause of the difference in probabilities is that Mitotic count may not have been measured in the same way in the ONCOPOOL data and the Galway data. Also the ONCOPOOL data did not include any patients with UICC staging 4. This means their population may be slightly different to our population or the same staging may not have been applied in both datasets. The model for OS is over predicting the five year probability in the ONCOPOOL data.

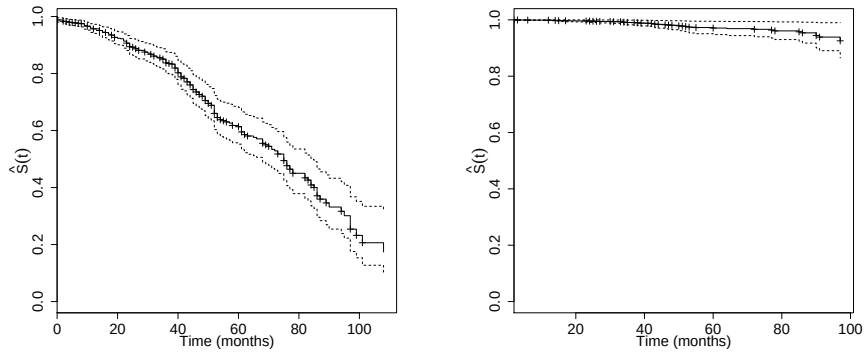
The concordance for the OS model on the ONCOPOOL data is 0.691. This is lower than the original concordance based on the Galway data (0.929).

6.7 Visualising the model

For non-statisticians, it may be easier to interpret the models using some visualisation tool. There are a few options available. The visualisations will just include the main effects models as the prediction performance is only marginally better including the interactions.

Plots of survival estimates

One option is to plot the Kaplan Meier survival estimates for the models (`Figure 6.9`). These survival estimates represents patients with baseline predictor measurements. However these are difficult for non-statisticians to interpret.



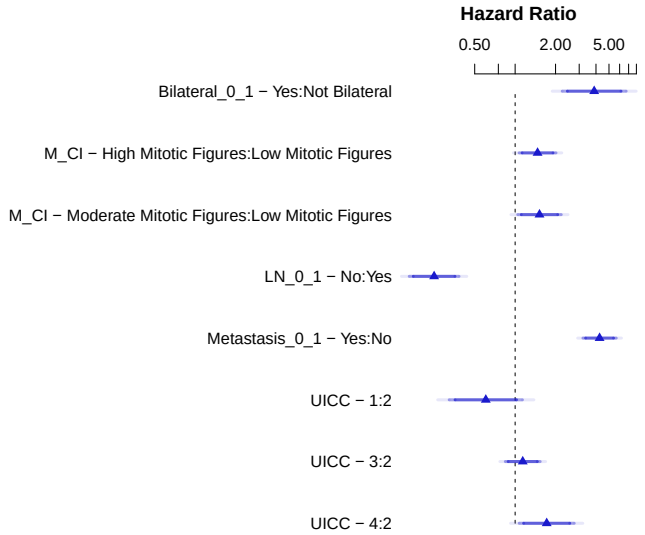
(a) Disease Free Survival

(b) Overall Survival

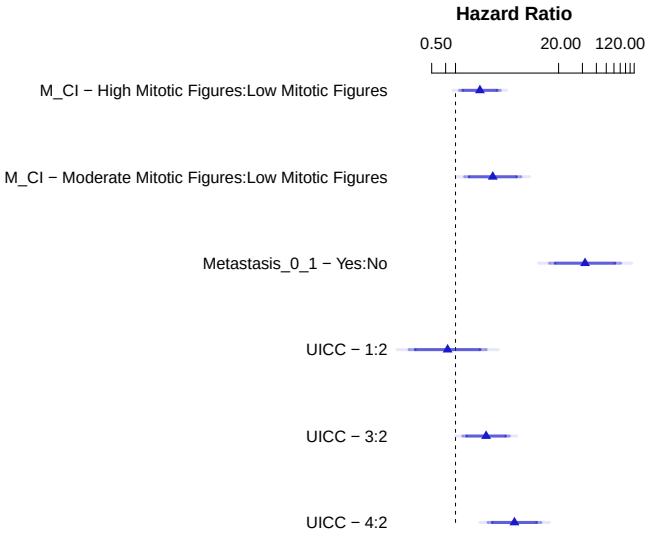
Figure 6.9: Survival Estimates the final models for DFS and OS for baseline patient characteristics.

Plots of hazard ratios

Hazard ratios with multilevel confidence bars for the effects of each of the predictors is given in `Figure 6.10`. Each of the plots contain the hazard ratio with a confidence interval for each of the levels of the predictor compared to the baseline level. Take for example Bilateral in DFS (`Figure 6.9(a)`), patients with Bilateral BC are more likely to get a disease recurrence than those patients without Bilateral BC (i.e. a worse prognosis).



(a) Disease Free Survival



(b) Overall Survival

Figure 6.10: Hazard ratios and multilevel confidence bars for the effects of predictors in the final models for DFS and OS.

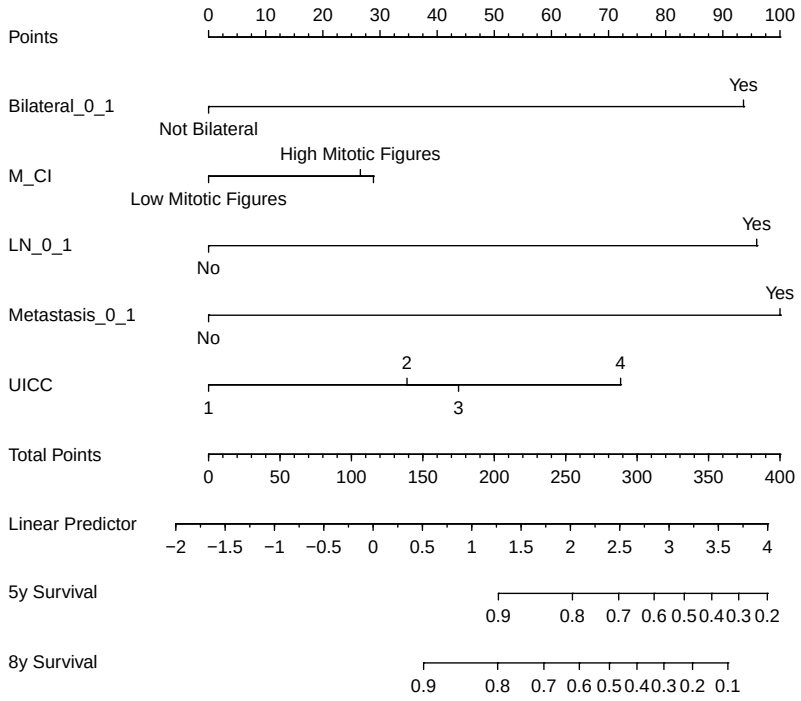


Figure 6.11: Nomogram for predicting survival probabilities for 5 year and 8 year survival for DFS models.

Nomograms

A nomogram is a graphical calculator which allows the user to calculate their probability of survival using a points scoring system based on their clinical results. Nomograms to calculate the probability of 5 year and 8 year survival are given for DFS in `Figure 6.11` and for OS in `Figure 6.12`. Take for example a patient who does not have bilateral breast cancer, with moderate mitotic count, has lymph nodes positive, does not have metastatic cancer and has stage 3 UICC staging. This patient has a probability of being disease free of 0.35 and a survival probability greater than 0.9 for five years. The higher number of Total Points the worse prognosis for a patient.

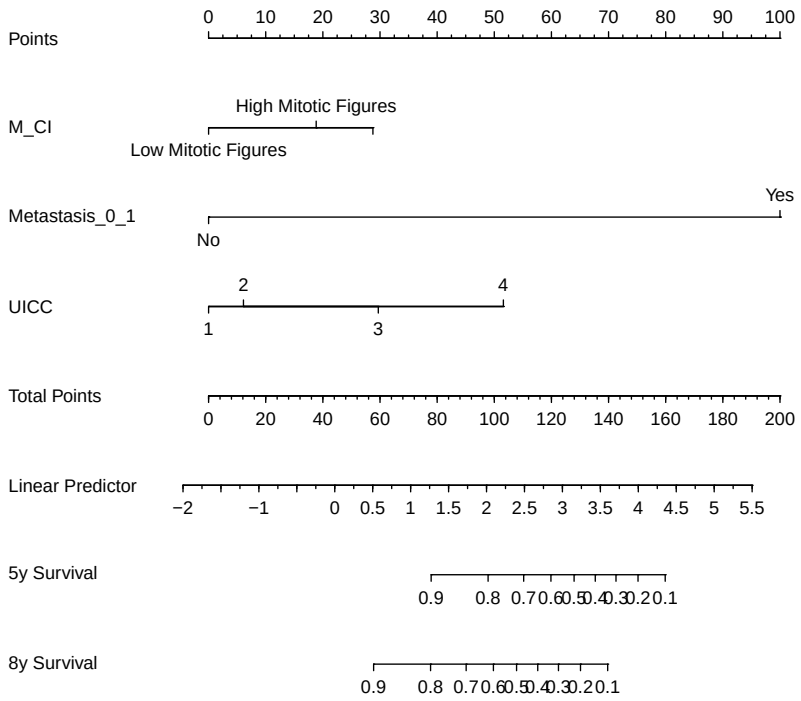


Figure 6.12: Nomogram for predicting survival probabilities for 5 year and 8 year survival for OS models.

On-line Calculator

Finally a visual front end tool available on-line for clinicians or patients to use would also be beneficial. This would save a clinical or patient the effort of calculating the 5 year estimated survival probability using the previous methods.

The `shiny` package in *R* is an easy way to turn these prognostic models into interactive web applications that anyone can use. The patients/clinicians just have to input the patient clinical and pathological details. The output includes a 5 year point survival estimate with a confidence interval for both DFS and OS. Also the output includes the median survival time. `Figure 6.13` contains the on-line calculator for OS I created using the `shiny` library. This includes the ridge regression model for all clinical predictors and also the model with a subset of predictors, namely *Mitotic count, Metastasis and UICC staging*. This figure gives the estimates for the baseline predictor level. `Figure 6.14` has a few of the clinical predictors changed to show how the estimated 5 year survival probability changes. This figure shows a patient with metastasis, stage 3 cancer, oestrogen positive and progesterone positive breast cancer has a 5 year estimated survival probability of 0.72.

6.8 Conclusions

After choosing the final models for DFS and OS using variable selection in imputed data, results presented in this chapter suggests that these models are validated and calibrated.

The model was validated internally using bootstrap resamples of the data and examining the discrimination and calibration. Both models seem to perform reasonably well.

External validation was performed using a dataset based on breast cancer patients diagnosed in 10 European breast cancer units. However, the data did not contain information on Bilateral or Lymph Node status so only the final model for OS could be validated using this data. The final OS model over-predicts the 5 year probability of survival for the patients in the ONCOPOOL data. However we cannot guarantee than Mitotic count and UICC staging may not have been measured in the same way in both datasets.

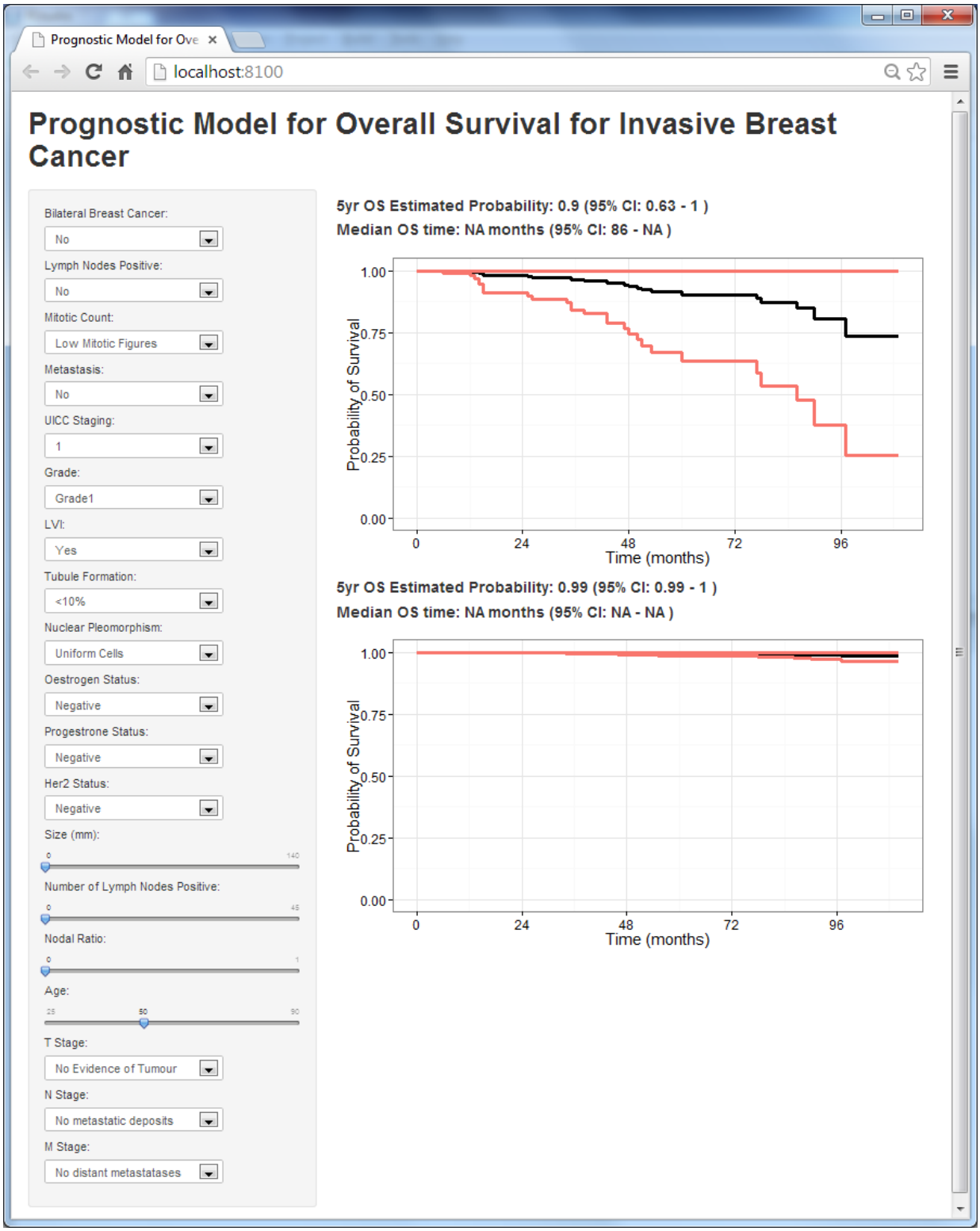


Figure 6.13: On-line calculator for calculating 5 year estimated survival probabilities using ridge regression and variable selection model.

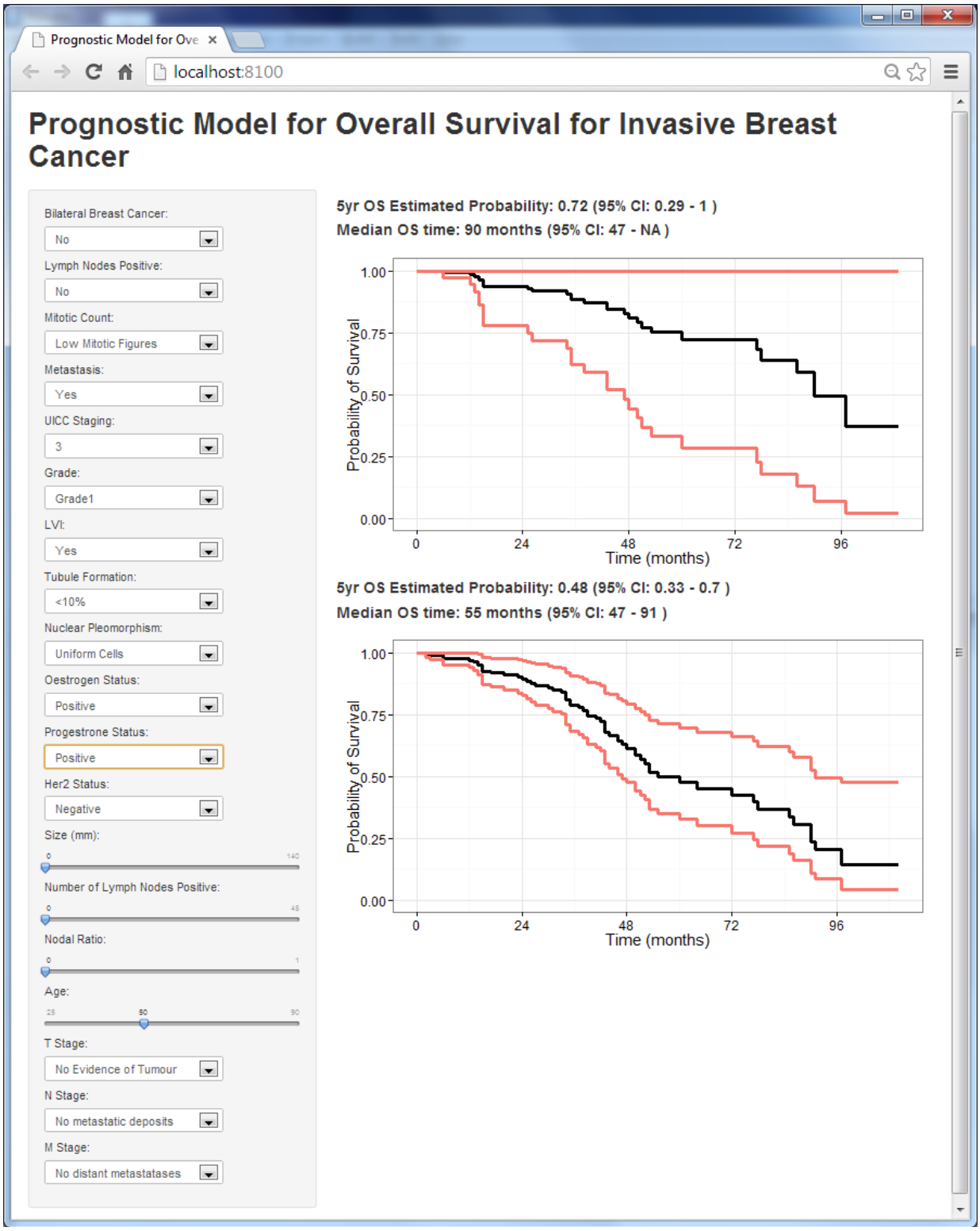


Figure 6.14: On-line calculator for calculating 5 year estimated survival probabilities using ridge regression and variable selection model.

The inclusion of interaction terms in the DFS model only marginally improved the prediction performance.

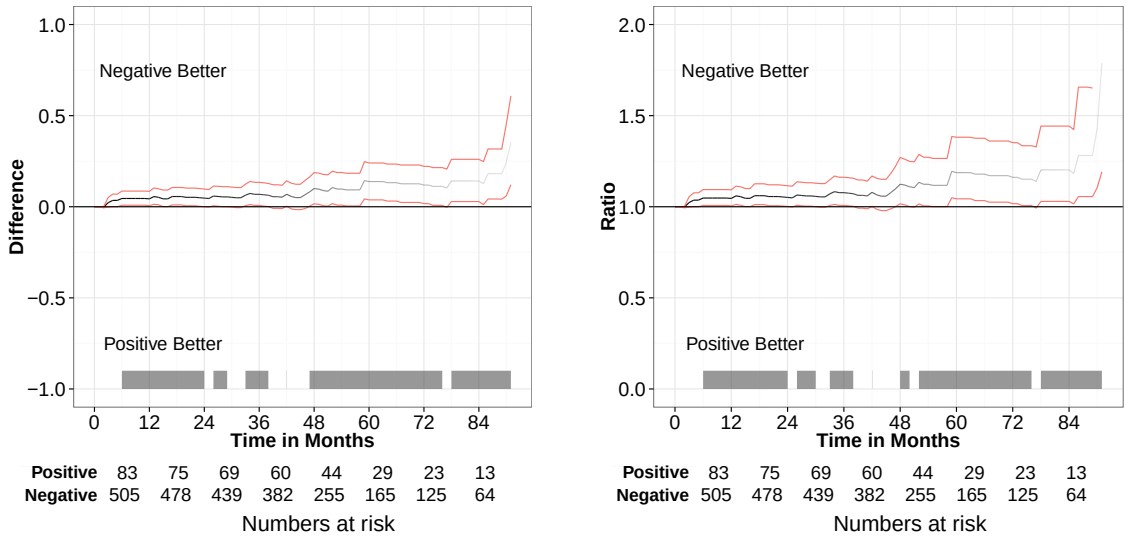
Visualisation tools are useful for making these models easier to interpret for non-statisticians. Nomograms are easy for either a clinician or patient to use, however, an on-line calculator makes it easier again as it performs the calculations to obtain the 5 year estimated survival probability by just filling in the clinical and pathological details.

Chapter 7

Conclusions and Future Work

The main aim of my PhD was to create a prognostic model for invasive breast cancer patients for DFS and OS. The data are comprised of 647 patients with patient characteristics and genetic markers for breast cancer which were collected retrospectively. An additional level of complexity existed due to the presence of missing data. A complete case analysis with both clinical and pathological biomarkers reduces the number of cases to 103 patients. A major challenge was how best to build a prognostic model for breast cancer in the presence of missing data.

Since the first aim of my PhD was to create a prognostic model for breast cancer, these data are time to event data. Chapter 2 introduced survival analysis and examined graphical and numerical summaries of survival estimates. The Kaplan Meier estimate is the most commonly used estimate for the representation of the distribution of survival times. It is generally used for the graphical comparison of survival estimates for two or more groups; for example, comparing the survival estimates for patients with and without Lymph Node disease. The Log-rank test is a hypothesis test to compare the survival estimates of groups. Graphical representations are a useful tool to complement the results of the Log-rank test. However, these graphical representations can get cluttered when confidence intervals are added to each group. An alternative ways of graphi-



(a) Difference

(b) Ratio

Figure 7.1: Graphical comparison of pointwise ratio/difference for Her2 positive and negative patients for OS.

cally representing the difference is to plot either the ratio or the difference of the survival functions. Extensions to these were introduced creating plots of both the pointwise ratio and difference with confidence intervals (created using bootstrap resamples of the data). These plots included the number of patients at risk at the bottom of the plot and also `alpha blending` which shaded the ratio/difference line relative to the number of patients at risk at each time point. This highlights the decrease in the number of patients at risk as time progresses due to patients experiencing the event or being censored.

A major challenge was how best to build a prognostic model for breast cancer in the presence of missing data. Some classical approaches were implemented in Chapter 4, however they did not perform well as the most common way to deal with missing values, casewise deletion, results in a considerably reduced sample size and possible bias. Instead of creating a model with all predictors, if a small subset of these predictors perform just as well as all the predictors, it is more cost and time effective to just use the subset of predictors. Variable selection techniques could not be performed on the original data as the missing data reduced the sample size from 647 patients to just 103. An alternative approach would be to perform variable selection techniques on imputed data. An empirical

simulation study was created using synthetic data based on the BC data to examine the effect of missing data on variable selection and the performance of variable selection in imputed data. The literature suggest using a voting system or stacking and weighting for variable selection on multiply imputed data. Obviously nothing compares with having complete data, however, if there are missing values, lost information can be "reclaimed" by performing variable selection techniques on imputed data. The results from the simulation study suggests that variable selection on imputed data ($n = 647$) performs better than that on complete cases ($n = 103$). However the stacking and weighting performs much better than the voting system. The weights include information on the proportion of missingness and the number of imputations. These methods for variable selection on imputed data identified other potentially useful predictors not identified previously for the BC data.

A summary of all the models fitted was given in the end of Chapter 5. Fitting models with both the clinical and pathological biomarkers marginally improved the performance of the model than using just the clinical predictors, so a model using just the clinical predictors seems like a logical step to reduce the cost and time of measuring all the biomarkers. The models identified using variable selection in imputed data performed better than those models with subsets of predictors identified earlier. Fitting models with all the predictors marginally improved the performance. The final model for DFS included Bilateral, Lymph Node status, Mitotic count, Metastasis and UICC staging and the final model for OS included a subset of these (Mitotic count, Metastasis and UICC staging). The models are given in `Table 7.1`. These were validated internally using bootstrapping and the OS model was validated externally using a dataset consisting of patients from 10 European breast cancer centers. Both models perform reasonably well. The addition of interaction terms in the DFS model only marginally improved the prediction performance.

The second aim of my PhD is to identify potentially useful predictors of the Oncotype DX risk and also to consolidate conflicting results from the literature. Many people believe that the results of Oncotype DX can be predicted just as well by routinely (and more cheaply) assessed pathological variables and biomarkers. However, there was just a small sample of 52 patients available

Table 7.1: Final models for DFS and OS.

	Response variables:	
	DFS	OS
	$\hat{\beta}(ESE)$	$\hat{\beta}(ESE)$
Bilateral = Yes	1.356*** (0.279)	
Mitotic Count = Low	−0.385** (0.161)	−0.709** (0.302)
Mitotic Count = Moderate	0.033 (0.211)	0.374 (0.411)
LN Status = Yes	1.389*** (0.218)	
Metastasis = Yes	1.448*** (0.146)	3.765*** (0.527)
UICC=2	0.503 (0.320)	0.230 (0.571)
UICC=3	0.633* (0.337)	1.118** (0.555)
UICC=4	1.044*** (0.381)	1.942*** (0.584)
Observations	432	444
Concordance Index	0.810	0.929
Note:		*p<0.1; **p<0.05; ***p<0.01

with the Oncotype DX risk (due to the high price of the test) and there were a large number of predictors measured. Classical approaches such as logistic regression are not suitable to use in this situation as the estimated standard errors were quite high. Alternatively, we examined a non-parametric approach using classification and regression trees. Trees were created using recursive partitioning and conditional inference. Both trees identified Progesterone status as a useful predictor however the other splits were on other predictors. Random Forests were also examined which improves the prediction error however the disadvantage is there is no tree structure. A novel approach to using surrogate splits was implemented where comparable and competing trees are created using the surrogate splits. An *R* package was created to produced an interactive surrogate plot, which allows trees to be grown using the surrogate splits as the first split in the tree. This technique was applied to the Oncotype DX data and it produced alternative trees with comparable predictive power to that of the original tree. Each of these surrogate trees identified predictors which were previously published in the literature and hence consolidating the conflicting results.

7.1 Future Work

The development of the new ideas proposed in this thesis opens up new opportunities for further analysis and research. Incorporating the number of patients and alpha blending into Kaplan Meier survival estimates. I will include confidence intervals into my code for `ggplot2` and create a package for *R*. This would make it accessible for others to use. Although there are already packages available to plot KM estimates, this new package would include adding the number of patients at risk to the bottom of the graph and also include alpha blending which incorporates those numbers of patients at risk into the actual plot. This package would allow the user to plot the KM estimator and would include the various options for plotting KM estimates like including confidence intervals, the numbers at risk and alpha blending. This package would also include the plots of the difference and ratio which were discussed in Chapter 2.

Also I plan to make the package for the interactive surrogate plot freely available. Improvements could also be made to the output, including the format of the plot for the surrogate tree and extending the code to include conditional inference trees.

In relation to variable selection techniques in imputed data, it would be interesting to examine how the LASSO and Ridge regression performs with imputed data. This could easily be added to the empirical simulation study that I had already created. The weighting system employed for variable selection in multiply imputed data examined three weights. Another weight could also be examined. The synthetic data used in the study were simulated using marginal estimates based on the BC data. Alternatively the data could be simulated by bootstrapping the original data. This would eliminate implausible biological patient characteristics simulated using the marginal estimates.

Obviously it is hard for non-statisticians to interpret probabilities with confidence intervals, so it may be useful to examine the mean residual life function so the output is in terms of time rather than risk in the on-line calculator.

Appendix A

Appendix

A.1 CART

A.1.1 Classical Approaches to Modelling for the Oncotype DX data

From `Figure 3.1(a)` in Chapter 3, it can be seen that there is a lot of missingness present in the Oncotype DX. Some of the variables have up to 50% missing. If a patient is missing in one of the Survivin variables it is very likely it will be missing in the other Survivin variables. Using complete cases this reduces the sample size down to 7. Variable selection using all the predictors could not be performed and this is why CART were used in the previous chapter.

Using a subset of the predictors and removing the predictors with large amounts of missing data, it reduces the sample size to 33. A logistic model was fitted (see `Table A.1`), however the majority of the predictors have extremely large standard errors so no accurate interpretation can be made from the model. Performing variable selection does not identify any of the predictors as important.

Due to the high proportions of missing data, a non-parametric approach, CART were implemented.

Table A.1: Logistic Model for Oncotype DX classification into Low, Medium and High risk. It is clear from the estimated coefficients and standard errors that this model cannot be interpreted accurately.

	Dependent variable:
	Oncotype DX category
	$\hat{\beta}(ESE)$
y$\geq$ 2	-84.606 $(7,562.539)$
y$\geq$ 3	-104.302 $(7,564.078)$
Age_yrs	1.658 (44.482)
Histological.type=2	69.340 (400.898)
Grade=2	-26.027 (296.386)
Grade=3	-33.620 (472.845)
Tumour Staging=2	-32.739 (802.855)
Tumour Staging=3	-33.263 $(1,543.226)$
LN Staging=3	41.697 (724.999)
Metastasis Staging=2	32.286 (320.442)
Metastasis Staging=3	14.180 (968.266)
Size (mm)	-1.019 (7.893)
LVI=1	-4.688 (823.635)
ER Score	15.094 (171.489)
ER Status (Positive)	18.721 $(3,734.411)$
PR Score	8.882 (127.518)
PR Status (Positive)	-93.389 $(1,185.893)$
Ki67	3.166 (40.792)
Bcl2=2	62.672 (710.272)
Bcl2=3	45.169 (329.754)
Bcl2 Score	-1.442 (9.250)
CD68 Score	-0.033 (1.981)
CD68 TIMs	0.538 (5.336)
Cyclin-B1	5.128 (268.145)
AAK	-7.770 (56.546)
Observations	33
R^2	1.000
χ^2	67.643
Note:	*p<0.1; **p<0.05; ***p<0.01

A.1.2 Trees and RF Variable Importance

This section shows the trees and Random Forests for models with all clinical and pathological predictors included. The pruned recursive partitioning trees are the same as the trees for the just the clinical predictors. The conditional inference trees have the same splits, however the p-values are slightly changed due to the fact more tests are being performed.

A.1.3 Surrogate Plot

RPART Code

Figure A.3 contains the tree which was over-fitted and then pruned back using `rpart.snip` and it is obtained using the following R code:

```
onco.c.RS.rpart <- rpart(Onco_DX_RS_cat ~ ., data=onco.c.RS,
control=rpart.control(minsplit=9,maxcompete=5,maxsurrogate=5))
onco.c.RS.rpart.snip=snip.rpart(onco.c.RS.rpart,toss=c(8,18,76,77))
draw.tree(onco.c.RS.rpart.snip)
```

It is possible to choose the number of primary surrogates given in the output by using `maxcompete` and also the number of secondary surrogates using `maxsurrogate`. The default for both of these is five.

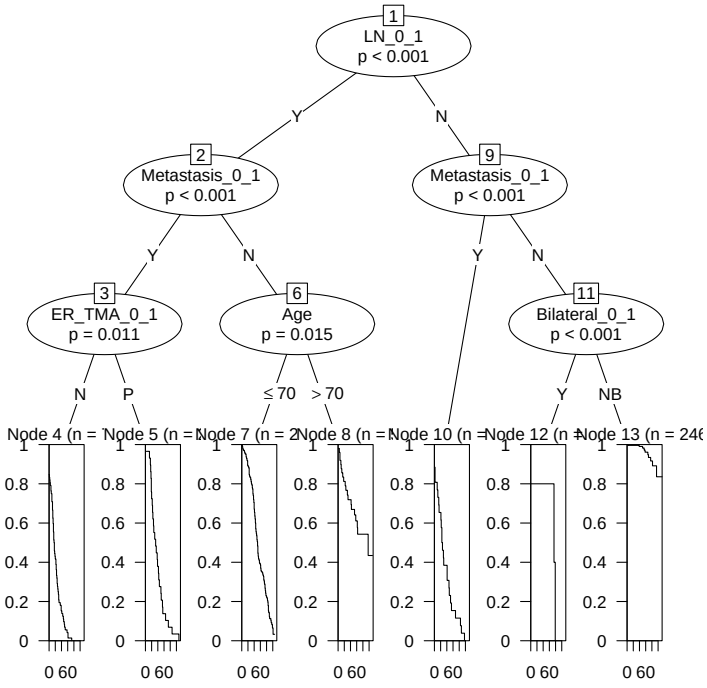
Graphs from Surrogate Plots

The original tree given in Figure 3.5 was used in the paper for the Oncotype DX research, however the output from R since the `rpart` library is being used, is in the default format. Figures A.3 and A.4 contain the actual output from the interactive surrogate plot.

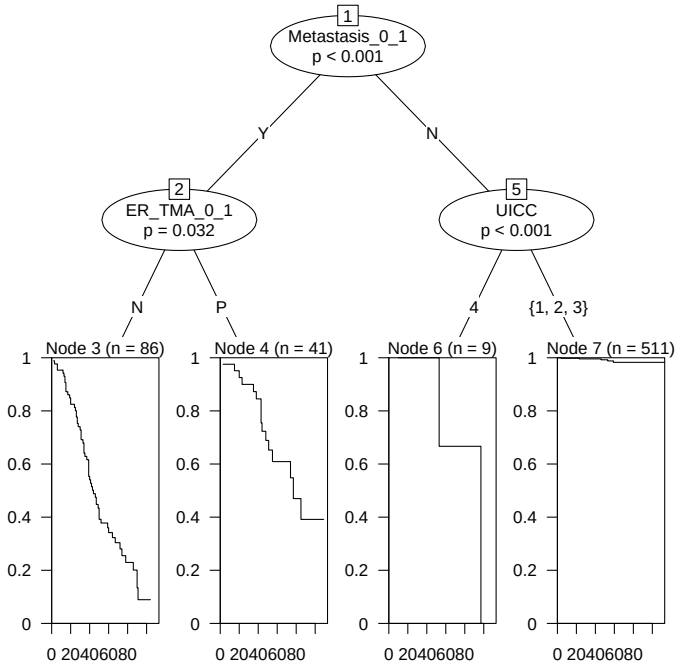
Features of the Surrogate Plot

There are certain program options for the default plot in Figure 3.11, that can be changed. Table A.2 contains some of these.

If any of the surrogates are clicked, it will ask "draw tree for surrogate?". If yes is clicked, the tree for that surrogate is drawn and outputted in a new window. If a point on the plot where there is no surrogate is clicked, an error message will appear: "No Surrogate at this point". Also there is a button to

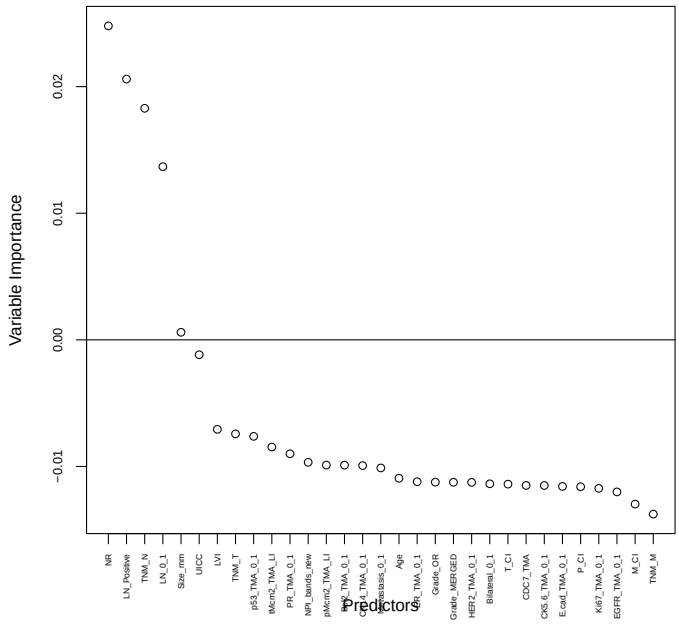


(a) DFS Clinical and Pathological Predictors

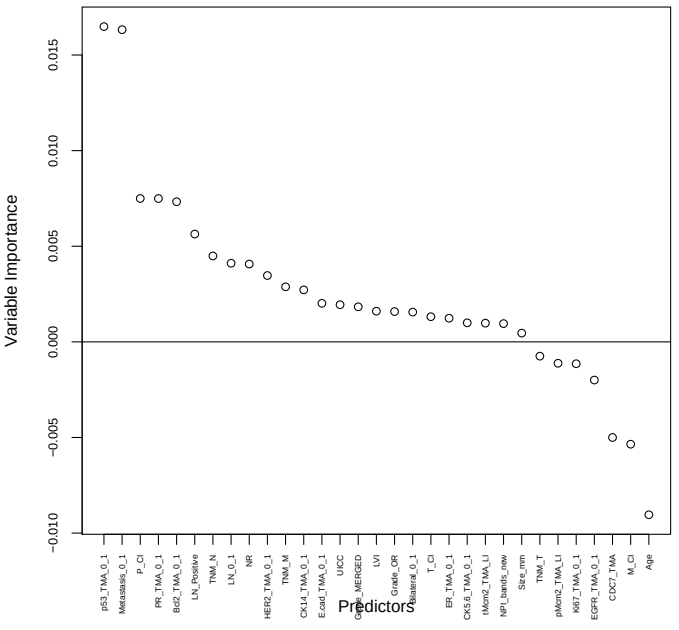


(b) DFS Clinical and Pathological Predictors

Figure A.1: Conditional Inference Tree for all clinical and pathological predictors for both DFS and OS using the Random Forest package for survival tree.



(a) DFS Clinical and Pathological Predictors



(b) DFS Clinical and Pathological Predictors

Figure A.2: Variable importance measure for all clinical and pathological predictors for both DFS and OS using the Random Forest package for survival tree.

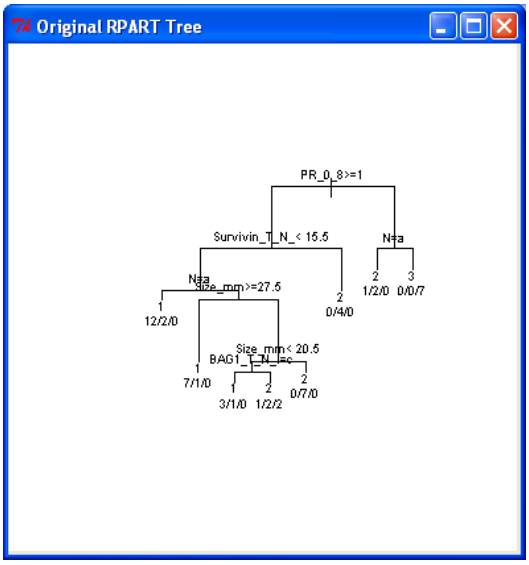


Figure A.3: Original RPART tree from Interactive Surrogate Plot Output.

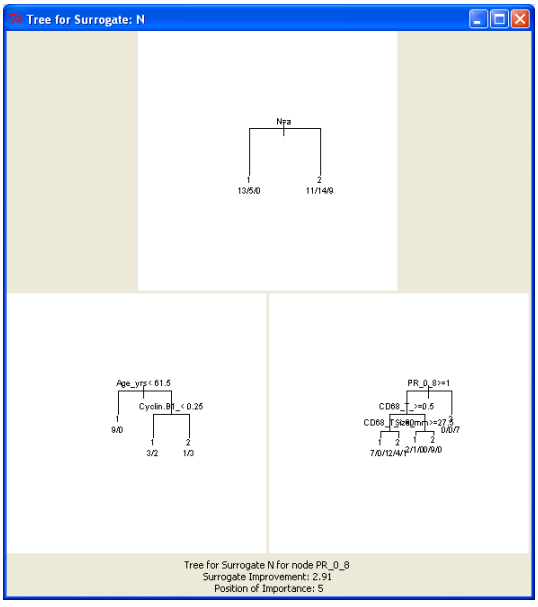


Figure A.4: Tree for surrogate N.

Table A.2: Program Options.

Argument	Explanation	Default
main	Title of plot	Interactive Surrogate Plot
ylab	Title of y axis	Variables
xlab	Title of x axis	Tree/Nodes
cex	Text size on plot	0.8
cex.axis	Text size for Axises	0.5
ordering	Orders variables according to importance	TRUE
colormap.pri	Colors for primary surrogates	heat.colors
colormap.sec	Colors for secondary surrogates	topo.colors

click to output the original rpart tree in a new window. The original tree is also drawn in a new window if the first split of the original tree is selected `Figure A.3`. An error message will also appear if there is no split for that surrogate.

A.2 Simulation Results

More results for other scenarios for the variable selection in multiply imputed data

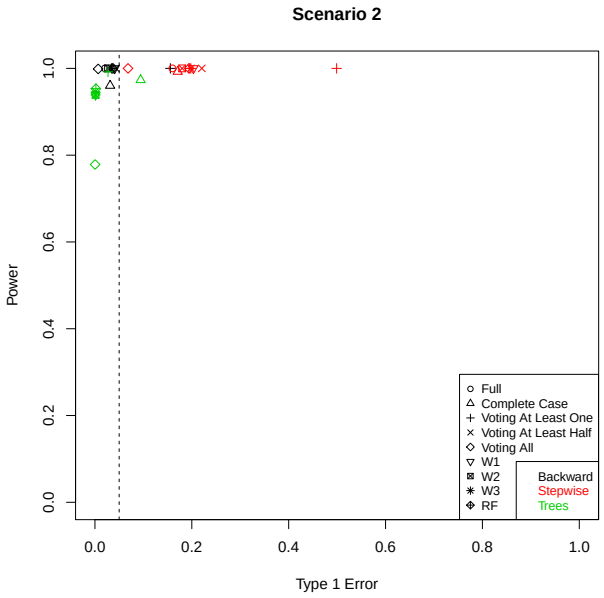


Figure A.5: Power and type 1 error for scenario two, MAR, sample size 1000 and 20% missing in each variable.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.2	1.9	1.6
Backward	100	100	100	100	16.0	17.4	16.1
rpart	90.2	100	100	82.7	0.2	0	0.2
Complete Cases							
Fastbw	2.4	100	100	89.1	2.4	2.6	1.4
Backward	3.9	100	100	99.4	17.6	16.4	17.8
rpart	74.4	100	72.2	94.2	46.0	0.9	0.9
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	98.9	100	100	100	36.5	49.9	14.0
At Least Half	97.0	100	100	100	10.3	14.8	3.9
All	75.8	100	100	100	1.6	1.1	0.6
Backward							
At Least Once	99.8	100	100	100	72.8	88.7	45.1
At Least Half	98.4	100	100	100	35.1	39.6	17.9
All	88.6	100	100	100	9.2	8.6	5.5
rpart							
At Least Once	95.8	100	100	97.0	19.4	8.5	2.3
At Least Half	72.7	100	99.7	84.1	0.9	0.4	0.1
All	15.5	100	92.7	49.6	0.0	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	97.4	100	100	100	9.3	11.6	3.9
W2	96.8	100	100	100	5.4	6.6	2.0
W3	97.1	100	100	100	7.9	10.2	3.5
Backward							
W1	98.8	100	100	100	34.7	34.7	17.2
W2	98.6	100	100	100	28.7	30.4	12.7
W3	98.7	100	100	100	32.4	32.4	16.1
rpart							
W1	64.5	100	99.0	84.7	1.2	0.2	0.1
W2	64.2	100	98.9	84.6	1.1	0.2	0.1
W3	64.6	100	99.1	84.5	1.0	0.2	0.1
Imputation (RF)							
Fastbw	100	100	100	100	5.8	9.4	3.6
Backward	100	100	100	100	23.9	32.1	21.4
rpart	95.2	100	96.6	81.6	1.5	2.9	0.5

Table A.3: Scenario 2: Number of times a variable was chosen by a simulation into the Survival Model (MAR and equal fractions of missing data (20% missing per variable) and sample size 1000). Average complete case sample size is 602.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.6	1.6	1.7
Backward	100	100	100	100	15.8	13.8	15.7
rpart	89.2	100	100	83.8	0.3	0.0	0.1
Complete Cases							
Fastbw	0.0	100	100	59.0	2.8	2.6	1.8
Backward	0.0	100	100	92.0	15.0	16.0	16.2
rpart	59.3	100	44.3	94.0	49.2	0.9	1.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	99.5	100	100	100	54.6	79.1	18.2
At Least Half	96.6	100	100	100	20.0	30.3	3.9
All	81.1	100	99.9	100	2.5	3.8	0.6
Backward							
At Least Once	100	100	100	100	85.8	97.0	51.8
At Least Half	98.7	100	100	100	48.4	59.3	18.6
All	90.0	100	100	100	13.9	12.6	3.4
rpart							
At Least Once	95.8	100	99.9	99.3	30.5	32.6	3.6
At Least Half	72.6	100	99.2	90.2	1.1	2.8	0.0
All	18.3	100	79.2	49.9	0.0	0.2	0.0
Stacked and Weighted							
Fastbw							
W1	96.9	100	100	100	18.8	27.5	3.5
W2	95.7	100	100	100	8.9	16.4	1.3
W3	96.81	100	100	100	15.7	23.0	2.8
Backward							
W1	99.1	100	100	100	45.7	54.3	15.9
W2	98.7	100	100	100	36.2	45.0	9.8
W3	98.9	100	100	100	42.7	51.4	13.8
rpart							
W1	63.7	100	98.0	86.9	1.3	1.8	0.0
W2	63.1	100	97.8	86.8	1.2	1.8	0.0
W3	63.0	100	97.9	87.6	1.5	1.8	0.0
Imputation (RF)							
Fastbw	100	100	100	100	11.0	15.0	4.9
Backward	100	100	100	100	30.6	36.9	23.7
rpart	93.3	100	87.7	87.6	6.2	6.3	0.2

Table A.4: Scenario 3: Number of times a variable was chosen by a simulation into the Survival Model (MAR and equal fractions of missing data (30% missing per variable) and sample size 1000). Average complete case sample size is 459.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.5	2.4	3.4
Backward	100	100	100	100	16.1	16.7	19.8
rpart	85.7	100	99.8	85.6	1.4	0.0	2.1
Complete Cases							
Fastbw	70.1	100	100	97.2	2.4	2.6	3.6
Backward	87.9	100	100	99.7	16.9	16.4	19.1
rpart	88.9	100	95.5	96.2	37.1	1.6	4.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	99.9	100	100	100	16.6	22.0	11.3
At Least Half	99.7	100	100	100	4.0	4.0	4.3
All	93.2	100	100	100	0.8	0.4	1.2
Backward							
At Least Once	100	100	100	100	52.6	61.8	41.0
At Least Half	99.9	100	100	100	25.0	22.7	21.0
All	98.3	100	100	100	6.1	4.1	8.6
rpart							
At Least Once	99.0	100	99.9	98.4	18.7	4.4	10.2
At Least Half	86.0	100	99.5	89.4	1.7	0.3	1.6
All	37.7	100	95.4	60.6	0.1	0.0	0.1
Stacked and Weighted							
Fastbw							
W1	99.7	100	100	100	4.1	3.2	3.8
W2	99.6	100	100	100	2.9	2.3	3.0
W3	99.7	100	100	100	3.8	2.8	3.4
Backward							
W1	99.9	100	100	100	23.2	20.4	19.8
W2	99.9	100	100	100	20.4	18.9	16.9
W3	99.9	100	100	100	22.1	20.4	18.8
rpart							
W1	84.1	100	99.6	91.5	2.3	0.3	2.6
W2	84.1	100	99.6	91.5	2.2	0.3	2.6
W3	84.3	100	99.6	91.5	2.5	0.3	2.5
Imputation (RF)							
Fastbw	100	100	100	100	3.3	4.5	4.5
Backward	100	100	100	100	2.9	2.3	3.0
rpart	94.6	100	99.0	85.2	3.7	1.0	2.9

Table A.5: Scenario 4: Number of times a variable was chosen by a simulation into the Survival Model (MAR and equal fractions of missing data (10% missing per variable) and sample size 700). Average complete case sample size is 536.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	68.9	100	88.9	51.9	4.0	3.5	4.6
Backward	90.4	100	98.4	87.5	17.4	18.0	19.6
rpart	6.3	99.6	29.9	34.1	13.0	2.5	13.1
Complete Cases							
Fastbw	2.6	99.9	72.1	20.2	3.2	4.4	2.7
Backward	4.0	100	94.3	59.5	20.3	20.9	20.0
rpart	7.5	96.2	19.3	39.9	20.1	2.3	15.6
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	75.1	100	95.6	71.6	14.3	19.1	11.9
At Least Half	51.0	100	85.4	53.9	4.7	5.4	5.0
All	19.0	99.8	57.6	33.9	0.7	0.6	1.8
Backward							
At Least Once	91.1	100	99.3	92.8	46.9	53.7	38.0
At Least Half	77.3	100	96.9	86.4	21.1	25.6	21.7
All	42.2	100	83.4	73.7	8.3	8.0	10.2
rpart							
At Least Once	17.6	100	72.2	75.9	48.9	17.5	50.6
At Least Half	5.3	99.2	33.2	34.0	10.1	1.8	10.0
All	0.9	92.9	7.5	7.0	091	0.1	0.6
Stacked and Weighted							
Fastbw							
W1	46.6	100	81.9	54.2	5.0	5.0	4.9
W2	43.4	99.9	48.9	3.7	3.6	3.2	3.7
W3	44.9	100	80.6	52.4	4.7	4.6	4.1
Backward							
W1	74.7	100	96.3	86.8	20.9	23.4	21.0
W2	71.7	100	95.5	85.7	17.6	21.3	18.3
W3	74.0	100	96.1	86.5	19.7	22.9	19.7
rpart							
W1	55.9	100	80.5	98.9	97.0	29.1	96.4
W2	55.1	100	79.9	98.7	96.0	28.0	96.0
W3	55.5	100	80.6	98.9	96.7	29.8	96.7
Imputation (RF)							
Fastbw	69.9	100	81.3	52.4	4.0	6.9	4.6
Backward	90.2	100	95.8	85.0	20.7	24.1	21.0
rpart	6.4	99.0	21.5	35.6	14.1	3.2	11.6

Table A.6: Scenario 5: Number of times a variable was chosen by a simulation into the Survival Model (MAR and equal fractions of missing data (10% missing per variable) and sample size 100). Average complete case sample size is 77.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.2	1.8	1.7
Backward	100	100	100	100	16.1	13.8	14.7
rpart	88.7	100	100	82.2	0.3	0	0.3
Complete Cases							
Fastbw	100	100	100	100	1.5	2.2	1.5
Backward	100	100	100	100	15.6	14.8	14.2
rpart	78.8	100	87.7	68.4	0.4	0.0	0.1
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	7.0	10.7	7.4
At Least Half	100	100	100	100	1.8	2.5	2.1
All	100	100	100	100	0.2	0.1	0.7
Backward							
At Least Once	100	100	100	100	34.6	45.5	35.4
At Least Half	100	100	100	100	18.3	17.3	16.0
All	100	100	100	100	6.7	4.2	5.2
rpart							
At Least Once	96.7	100	100	95.3	1.6	0.2	1.2
At Least Half	89.4	100	100	85.3	0.2	0.0	0.0
All	70.9	100	98.9	58.1	0.1	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	2.0	1.9	2.0
W2	100	100	100	100	1.2	1.4	1.6
W3	100	100	100	100	1.9	1.7	1.8
Backward							
W1	100	100	100	100	17.5	15.5	14.6
W2	100	100	100	100	14.8	13.5	12.4
W3	100	100	100	100	16.3	14.8	13.9
rpart							
W1	90.1	100	100	85.3	0.3	0.0	0.1
W2	90.1	100	100	85.2	0.3	0.0	0.1
W3	90.6	100	100	85.6	0.3	0.0	0.1
Imputation (RF)							
Fastbw	100	100	100	100	2.8	2.3	2.4
Backward	100	100	100	100	17.8	17.2	17.5
rpart	87.3	100	99.9	81.4	0.1	0.0	0.2

Table A.7: Scenario 6: Number of times a variable was chosen by a simulation into the Survival Model (MCAR and equal fractions of missing data (10% missing per variable) and sample size 1000). Average complete case sample size is 729.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.3	1.7	1.2
Backward	100	100	100	100	12.7	17.9	14.0
rpart	88.3	100	100	83.7	0.2	0.0	0.1
Complete Cases							
Fastbw	100	100	100	100	1.6	2.3	1.5
Backward	100	100	100	100	13.4	17.2	14.8
rpart	35.8	100	26.8	23.5	0.2	0.0	0.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	10.0	20.7	10.0
At Least Half	100	100	100	100	1.7	4.0	2.2
All	100	100	100	100	0.5	0.4	0.5
Backward							
At Least Once	100	100	100	100	44.5	69.3	48.3
At Least Half	100	100	100	100	16.2	21.9	16.6
All	100	100	100	100	4.3	3.4	4.1
rpart							
At Least Once	98.2	100	100	97.2	1.8	0.2	2.4
At Least Half	91.0	100	99.7	86.3	0.1	0.0	0.0
All	66.2	100	92.4	51.3	0.0	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	1.4	3.1	2.1
W2	100	100	100	100	0.8	1.3	0.9
W3	100	100	100	100	1.1	2.4	1.7
Backward							
W1	100	100	100	100	15.2	18.7	15.0
W2	100	100	100	100	10.9	14.9	10.7
W3	100	100	100	100	13.9	18.2	13.3
rpart							
W1	90.3	100	99.3	84.4	0.1	0.0	0.1
W2	90.2	100	99.3	84.4	0.1	0.0	0.1
W3	90.3	100	99.3	84.4	0.1	0.0	0.1
Imputation (RF)							
Fastbw	100	100	100	100	2.9	4.3	3.1
Backward	100	100	100	100	16.8	21.9	18.8
rpart	88.8	100	98.4	81.8	0.2	0.0	0.3

Table A.8: Scenario 7: Number of times a variable was chosen by a simulation into the Survival Model (MCAR and equal fractions of missing data (20% missing per variable) and sample size 1000). Average complete case sample size is 512.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.0	2.2	1.9
Backward	100	100	100	100	15.3	17.0	15.3
rpart	89.1	100	100	85.4	0.0	0.0	0.0
Complete Cases							
Fastbw	99.8	100	100	99.1	1.4	1.6	1.9
Backward	100	100	100	99.9	15.1	15.7	15.2
rpart	30.5	99.9	11.0	15.7	0.1	0.0	0.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	16.7	34.7	17.4
At Least Half	100	100	100	100	3.0	5.7	3.1
All	100	100	100	100	0.2	0.2	0.4
Backward							
At Least Once	100	100	100	100	55.0	81.2	55.0
At Least Half	100	100	100	100	19.2	26.7	19.2
All	100	100	100	100	4.1	3.7	3.8
rpart							
At Least Once	98.8	100	100	98.7	3.1	1.2	3.4
At Least Half	91.7	100	99.37	88.6	0.0	0.0	0.0
All	57.9	100	78.9	46.7	0.0	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	2.3	4.4	2.8
W2	100	100	100	100	0.4	1.7	0.9
W3	100	100	100	100	1.7	3.1	2.2
Backward							
W1	100	100	100	100	17.6	21.1	17.4
W2	100	100	100	100	10.5	14.6	10.6
W3	100	100	100	100	15.4	19.1	15.3
rpart							
W1	87.5	100	97.3	83.0	0.0	0.0	0.0
W2	87.3	100	97.3	82.9	0.0	0.0	0.0
W3	87.5	100	97.4	83.2	0.0	0.0	0.0
Imputation (RF)							
Fastbw	100	100	100	100	5.5	7.0	5.0
Backward	100	100	100	100	21.6	27.0	21.3
rpart	84.3	100	94.5	83.3	0.3	0.2	0.1

Table A.9: Scenario 8: Number of times a variable was chosen by a simulation into the Survival Model (MCAR and equal fractions of missing data (30% missing per variable) and sample size 1000). Average complete case sample size is 343.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.9	1.5	2.3
Backward	100	100	100	100	15.8	16.0	18.4
rpart	85.3	100	99.9	87.2	1.4	0.1	1.1
Complete Cases							
Fastbw	100	100	100	99.9	1.9	1.3	2.2
Backward	100	100	100	100	16.0	15.6	17.7
rpart	72.4	100	82.2	68.8	2.7	0.2	2.1
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	6.9	11.2	7.6
At Least Half	100	100	100	100	2.1	2.2	2.0
All	100	100	100	100	0.6	0.5	0.7
Backward							
At Least Once	100	100	100	100	35.7	48.4	36.1
At Least Half	100	100	100	100	17.4	19.2	18.9
All	100	100	100	100	6.4	5.0	7.8
rpart							
At Least Once	96.2	100	100	98.6	9.5	1.0	8.1
At Least Half	87.2	100	99.5	89.9	1.0	0.1	0.4
All	64.3	100	93.6	60.2	0.2	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	2.0	2.0	1.9
W2	100	100	100	100	1.3	1.4	1.4
W3	100	100	100	100	1.6	1.7	1.9
Backward							
W1	100	100	100	100	16.1	17.5	17.6
W2	100	100	100	100	14.2	15.1	15.6
W3	100	100	100	100	15.6	16.5	16.6
rpart							
W1	88.0	100	99.4	94.2	1.6	0.1	1.0
W2	87.9	100	99.4	94.3	1.5	0.1	1.0
W3	88.2	100	99.4	94.2	1.5	0.1	0.9
Imputation (RF)							
Fastbw	100	100	100	100	2.5	2.3	2.3
Backward	100	100	100	100	16.4	18.1	18.6
rpart	84.3	100	94.5	83.3	2.2	0.1	1.2

Table A.10: Scenario 9: Number of times a variable was chosen by a simulation into the Survival Model (MCAR and equal fractions of missing data (10% missing per variable) and sample size 700). Average complete case sample size is 510.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	72.2	100	89.6	53.2	3.0	3.0	3.6
Backward	92.8	100	99.0	86.7	16.6	17.9	19.2
rpart	7.1	99.2	30.7	34.0	12.4	2.5	12.2
Complete Cases							
Fastbw	50.1	99.7	73.7	37.9	3.8	4.2	3.7
Backward	76.9	100	95.1	74.9	18.8	18.3	18.2
rpart	8.0	94.5	18.6	30.5	12.9	2.2	11.2
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	82.1	100	94.9	67.3	9.6	15.4	11.4
At Least Half	71.6	100	86.2	52.2	3.3	4.3	3.6
All	50.6	100	56.8	32.6	0.5	1.0	1.0
Backward							
At Least Once	94.7	100	99.9	92.4	35.7	48.5	35.4
At Least Half	92.2	100	97.6	84.8	19.1	19.6	20.5
All	81.1	100	85.1	72.9	7.7	5.6	7.8
rpart							
At Least Once	18.3	99.9	62.4	70.8	48.5	13.7	46.0
At Least Half	7.1	98.9	30.2	35.1	9.6	1.6	8.5
All	2.0	91.9	6.2	6.8	0.8	0.1	0.6
Stacked and Weighted							
Fastbw							
W1	70.0	100	83.0	51.3	3.3	3.9	3.1
W2	64.2	100	79.4	46.1	2.4	3.4	2.4
W3	68.4	100	82.0	49.4	2.9	3.9	2.9
Backward							
W1	92.4	100	96.5	84.1	17.8	19.6	19.6
W2	91.2	100	96.3	82.8	15.7	16.7	17.6
W3	92.2	100	96.5	83.6	17.4	18.6	19.4
rpart							
W1	66.2	100	76.9	99.2	96.9	27.1	96.9
W2	65.8	100	76.1	98.8	96.6	25.8	96.4
W3	66.2	100	77.9	99.4	96.7	26.8	96.4
Imputation (RF)							
Fastbw	67.9	100	82.5	49.6	3.4	4.7	3.6
Backward	90.7	100	96.4	84.6	19.3	20.0	20.3
rpart	6.8	98.8	23.3	32.5	13.0	3.0	12.1

Table A.11: Scenario 10: Number of times a variable was chosen by a simulation into the Survival Model (MCAR and equal fractions of missing data (10% missing per variable) and sample size 100). Average complete case sample size is 72.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.3	1.1	2.5
Backward	100	100	100	100	13.8	15.3	16.6
rpart	88.6	100	99.9	82.1	0.1	0.0	0.1
Complete Cases							
Fastbw	100	100	100	100	1.3	1.9	2.3
Backward	100	100	100	100	14.2	16.2	16.7
rpart	47.6	100	57.0	43.9	0.2	0.0	0.1
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	4.9	7.9	6.2
At Least Half	100	100	100	100	1.3	1.6	3.6
All	100	100	100	100	0.1	0.1	1.2
Backward							
At Least Once	100	100	100	100	29.4	42.7	34.0
At Least Half	100	100	100	100	14.9	18.5	17.2
All	100	100	100	100	5.7	5.2	7.9
rpart							
At Least Once	97.1	100	100	94.9	0.9	0.1	0.9
At Least Half	90.6	100	99.9	86.0	0.1	0.0	0.1
All	73.0	100	99.1	60.3	0.0	0.0	0.1
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	1.5	1.3	2.9
W2	100	100	100	100	0.4	0.8	2.2
W3	100	100	100	100	0.9	1.1	2.8
Backward							
W1	100	100	100	100	14.0	16.8	16.4
W2	100	100	100	100	11.4	14.6	13.9
W3	100	100	100	100	13.3	16.2	15.6
rpart							
W1	90.3	100	99.9	88.4	0.2	0.0	0.1
W2	89.8	100	99.9	88.2	0.2	0.0	0.1
W3	90.1	100	99.9	88.8	0.2	0.0	0.1
Imputation (RF)							
Fastbw	100	100	100	100	1.1	2.0	3.0
Backward	100	100	100	100	14.3	17.6	16.5
rpart	89.9	100	99.8	81.5	0.2	0.1	0.1

Table A.12: Scenario 11: Number of times a variable was chosen by a simulation into the Survival Model (MNAR and equal fractions of missing data (10% missing per variable) and sample size 1000). Average complete case sample size is 859.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.7	2.4	2.4
Backward	100	100	100	100	16.0	14.7	16.5
rpart	89.3	100	99.9	83.7	0.2	0.0	0.2
Complete Cases							
Fastbw	100	100	100	100	1.7	2.4	2.4
Backward	100	100	100	100	15.0	17.0	16.0
rpart	11.6	100	13.0	7.0	0.0	0.0	0.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	8.3	10.9	9.7
At Least Half	100	100	100	100	1.7	2.1	2.5
All	100	100	100	100	0.2	0.2	0.6
Backward							
At Least Once	100	100	100	100	38.6	47.7	38.6
At Least Half	100	100	100	100	17.4	19.2	18.5
All	100	100	100	100	4.8	3.6	6.7
rpart							
At Least Once	97.1	100	99.9	96.8	2.0	0.0	1.6
At Least Half	90.2	100	99.9	88.1	0.1	0.0	0.1
All	67.9	100	96.4	55.3	0.0	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	1.5	1.9	2.0
W2	100	100	100	100	0.6	0.8	0.8
W3	100	100	100	100	1.3	1.7	1.9
Backward							
W1	100	100	100	100	15.8	17.1	16.6
W2	100	100	100	100	11.2	11.9	12.7
W3	100	100	100	100	14.4	16.2	15.5
rpart							
W1	85.9	100	99.7	87.8	0.1	0.0	0.0
W2	89.4	100	99.7	87.6	0.1	0.0	0.0
W3	89.6	100	99.7	87.5	0.1	0.0	0.0
Imputation (RF)							
Fastbw	100	100	100	100	2.0	3.5	3.0
Backward	100	100	100	100	18.3	19.7	19.2
rpart	88.0	100	99.7	84.0	0.1	0.0	0.1

Table A.13: Scenario 12: Number of times a variable was chosen by a simulation into the Survival Model (MNAR and equal fractions of missing data (20% missing per variable) and sample size 1000). Average complete case sample size is 818.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	2.0	1.9	1.7
Backward	100	100	100	100	13.2	17.0	14.3
rpart	89.2	100	100	83.1	0.3	0.0	0.5
Complete Cases							
Fastbw	100	100	100	99.6	1.9	2.5	1.5
Backward	100	100	100	100	15.2	15.8	16.0
rpart	16.6	100	11.9	9.1	0.0	0.0	0.0
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	13.2	26.4	11.0
At Least Half	100	100	100	100	2.6	3.6	2.4
All	100	100	100	100	0.4	0.0	0.6
Backward							
At Least Once	100	100	100	100	47.6	76.6	46.4
At Least Half	100	100	100	100	18.9	23.4	17.8
All	100	100	100	100	4.0	3.3	4.1
rpart							
At Least Once	98.1	100	100	96.9	3.3	0.6	1.6
At Least Half	90.4	100	99.7	83.7	0.3	0.0	0.2
All	62.9	100	93.6	47.1	0.0	0.0	0.0
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	2.0	3.1	1.9
W2	100	100	100	100	0.5	0.9	0.7
W3	100	100	100	100	1.8	2.2	1.6
Backward							
W1	100	100	100	100	16.6	19.6	16.3
W2	100	100	100	100	9.4	13.1	9.5
W3	100	100	100	100	14.9	17.6	14.8
rpart							
W1	88.2	100	99.7	81.6	0.1	0.0	0.2
W2	87.9	100	99.7	81.5	0.1	0.0	0.2
W3	88.4	100	99.7	81.0	0.1	0.0	0.1
Imputation (RF)							
Fastbw	100	100	100	100	3.5	5.2	3.8
Backward	100	100	100	100	19.4	25.0	20.1
rpart	86.5	100	99.0	80.3	0.5	0.0	0.3

Table A.14: Scenario 13: Number of times a variable was chosen by a simulation into the Survival Model (MNAR and equal fractions of missing data (30% missing per variable) and sample size 1000). Average complete case sample size is 498.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	100	100	100	100	1.8	1.7	2.0
Backward	100	100	100	100	16.1	16.9	14.6
rpart	84.1	100	100	82.8	1.3	0.0	1.2
Complete Cases							
Fastbw	100	100	100	100	2.0	1.8	1.5
Backward	100	100	100	100	15.8	15.0	15.6
rpart	47.7	100	57.7	43.0	1.0	0.0	1.1
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	100	100	100	100	6.7	8.5	6.4
At Least Half	100	100	100	100	2.2	1.7	1.8
All	100	100	100	100	0.8	0.3	0.6
Backward							
At Least Once	100	100	100	100	32.6	41.6	30.7
At Least Half	100	100	100	100	17.0	16.0	17.4
All	100	100	100	100	7.3	4.1	8.1
rpart							
At Least Once	95.2	100	100	97.0	7.4	0.6	8.2
At Least Half	85.8	100	100	87.7	0.8	0.2	1.4
All	65.6	100	96.8	57.0	0.0	0.0	0.1
Stacked and Weighted							
Fastbw							
W1	100	100	100	100	2.1	1.3	2.1
W2	100	100	100	100	1.4	0.7	1.2
W3	100	100	100	100	1.8	1.3	1.6
Backward							
W1	100	100	100	100	15.9	14.6	16.2
W2	100	100	100	100	14.3	12.9	14.3
W3	100	100	100	100	15.3	14.4	15.9
rpart							
W1	86.8	100	99.9	92.2	1.0	0.2	1.9
W2	86.7	100	99.8	92.1	1.0	0.2	1.7
W3	86.3	100	100	92.3	0.9	0.2	1.9
Imputation (RF)							
Fastbw	100	100	100	100	2.2	2.3	1.9
Backward	100	100	100	100	17.6	17.9	17.2
rpart	83.2	100	99.6	84.0	1.1	0.2	2.3

Table A.15: Scenario 14: Number of times a variable was chosen by a simulation into the Survival Model (MNAR and equal fractions of missing data (10% missing per variable) and sample size 700). Average complete case sample size is 603.

Method	Bilateral	LN Status	Metastasis	Size	Age	ER Status	Noise
Full (No Missing)							
Fastbw	70.4	100	88.4	53.5	4.0	3.8	3.0
Backward	91.8	100	98.3	89.4	19.1	19.1	17.3
rpart	5.8	99.0	31.3	35.5	12.8	3.2	13.0
Complete Cases							
Fastbw	64.9	100	82.5	45.6	2.9	3.6	3.7
Backward	86.4	100	97.3	84.6	18.7	18.8	19.7
rpart	6.4	95.2	19.6	30.6	10.3	1.3	9.5
Imputed (MICE)							
Voting							
Fastbw							
At Least Once	80.0	100	93.2	67.0	8.4	11.7	8.7
At Least Half	70.9	100	85.9	53.3	4.1	5.2	4.6
All	55.5	100	65.5	37.1	1.8	1.1	1.5
Backward							
At Least Once	93.3	100	99.2	92.3	31.3	41.5	29.4
At Least Half	91.4	100	97.9	88.9	19.8	20.4	17.7
All	84.0	100	88.6	78.2	9.3	8.1	9.7
rpart							
At Least Once	17.0	99.9	58.9	70.5	43.0	13.2	43.6
At Least Half	6.5	99.2	32.4	34.5	9.8	1.8	10.1
All	1.6	94.1	8.5	11.1	0.8	0.6	1.6
Stacked and Weighted							
Fastbw							
W1	69.9	100	83.4	53.2	3.9	5.0	4.1
W2	64.5	100	80.7	47.3	3.5	3.9	3.3
W3	68.8	100	83.0	52.0	3.8	4.7	4.0
Backward							
W1	91.8	100	97.7	89.0	19.2	19.2	18.1
W2	91.4	100	97.2	87.2	16.6	17.5	16.0
W3	91.6	100	97.7	88.6	18.6	18.9	17.3
rpart							
W1	65.1	99.9	79.9	98.7	96.0	29.8	97.0
W2	64.6	99.9	79.3	98.4	94.9	27.2	96.1
W3	65.2	99.9	80.4	98.8	95.6	29.5	96.8
Imputation (RF)							
Fastbw	69.7	100	84.3	52.8	3.8	5.0	3.9
Backward	90.3	100	97.9	88.6	18.6	21.0	18.2
rpart	6.4	99.5	28.0	37.6	13.5	3.4	13.9

Table A.16: Scenario 15: Number of times a variable was chosen by a simulation into the Survival Model (MNAR and equal fractions of missing data (10% missing per variable) and sample size 100). Average complete case sample size is 87.

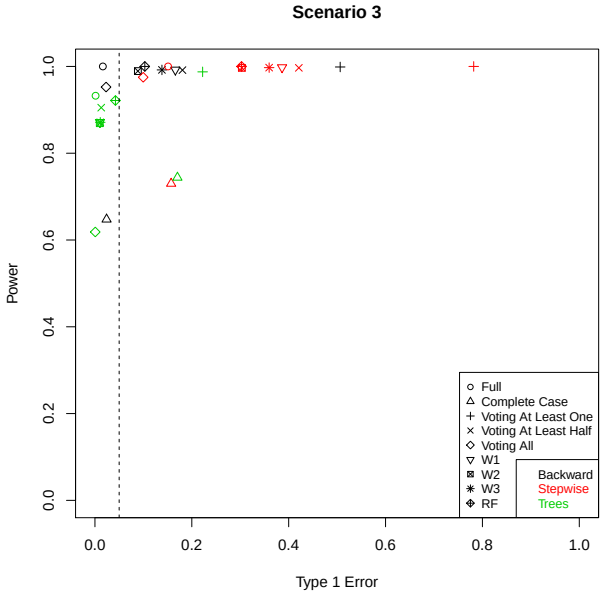


Figure A.6: Power and type 1 error for scenario three, MCAR, sample size 1000 and 10% missing in each variable.

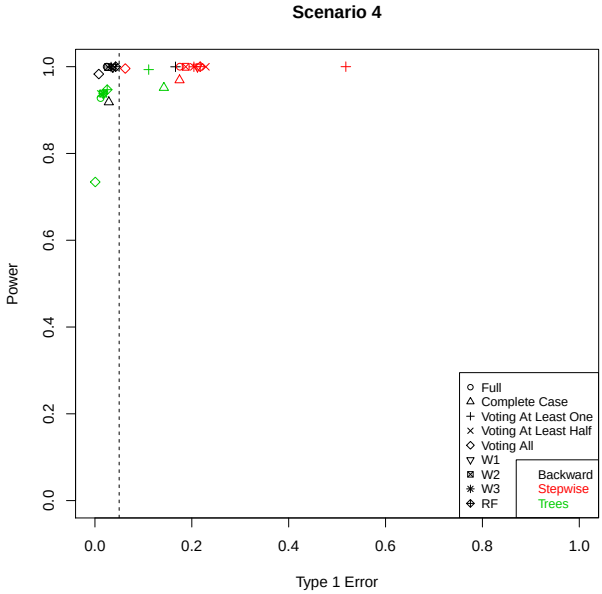


Figure A.7: Power and type 1 error for scenario four, MAR, sample size 700 and 10% missing in each variable.

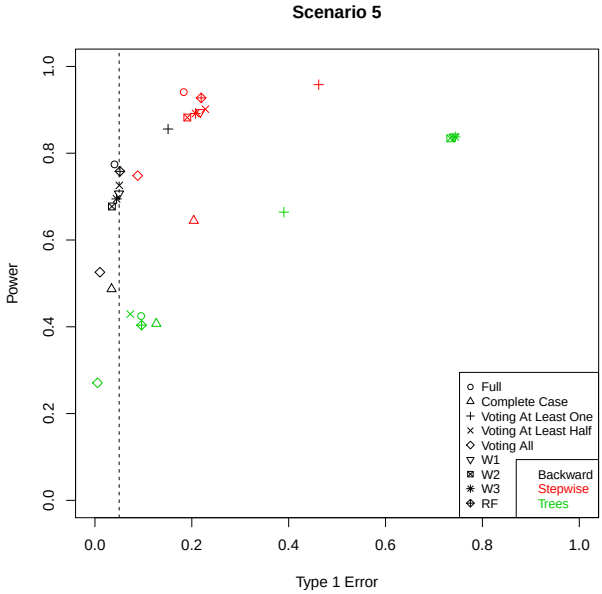


Figure A.8: Power and type 1 error for scenario five, MAR, sample size 100 and 10% missing in each variable.

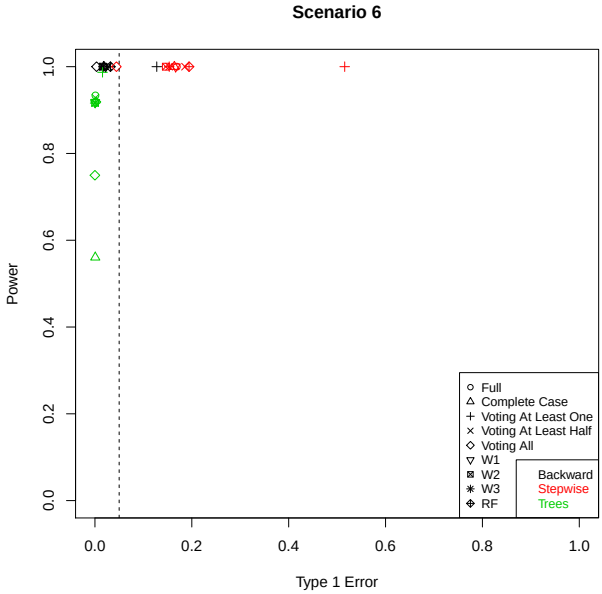


Figure A.9: Power and type 1 error for scenario six, MCAR, sample size 1000 and 10% missing in each variable.

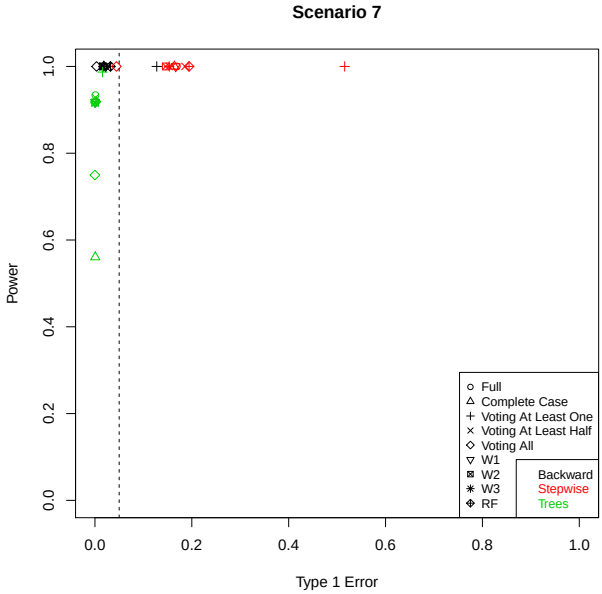


Figure A.10: Power and type 1 error for scenario seven, MCAR, sample size 1000 and 20% missing in each variable.

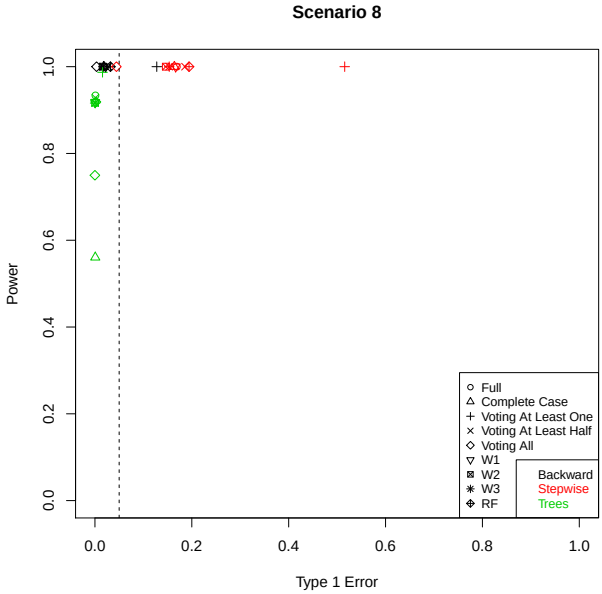


Figure A.11: Power and type 1 error for scenario eight, MCAR, sample size 1000 and 30% missing in each variable.

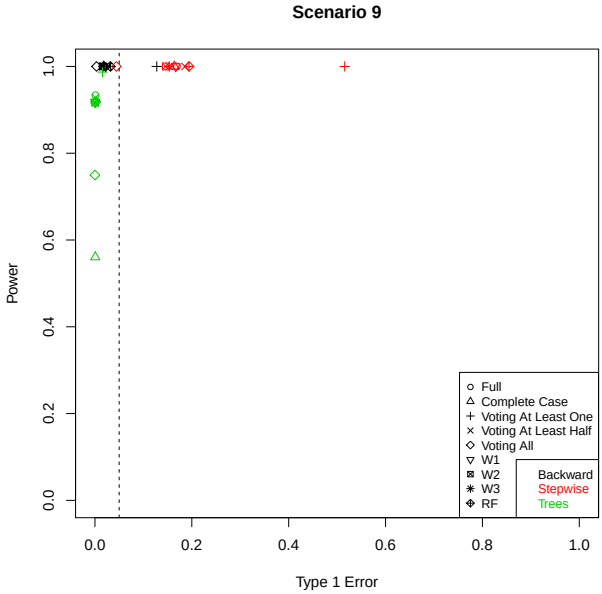


Figure A.12: Power and type 1 error for scenario nine, MCAR, sample size 700 and 10% missing in each variable.

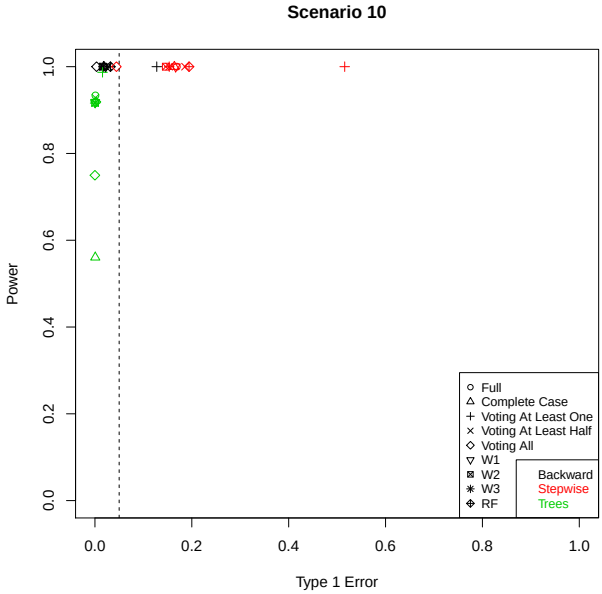


Figure A.13: Power and type 1 error for scenario ten, MCAR, sample size 100 and 10% missing in each variable.

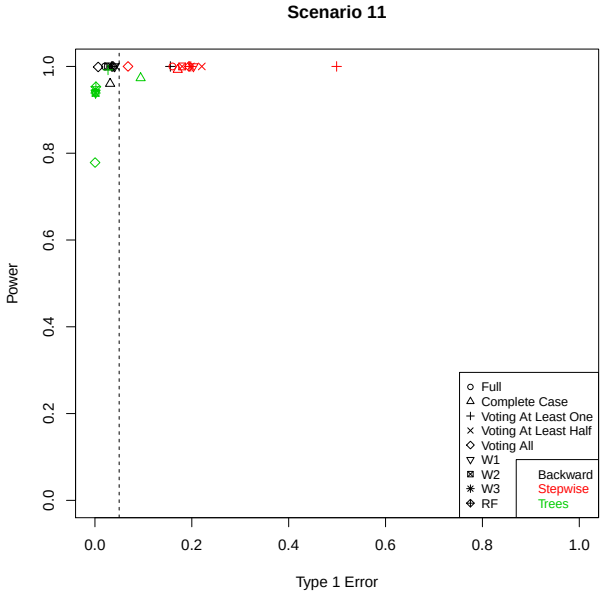


Figure A.14: Power and type 1 error for scenario eleven, MNAR, sample size 1000 and 10% missing in each variable.

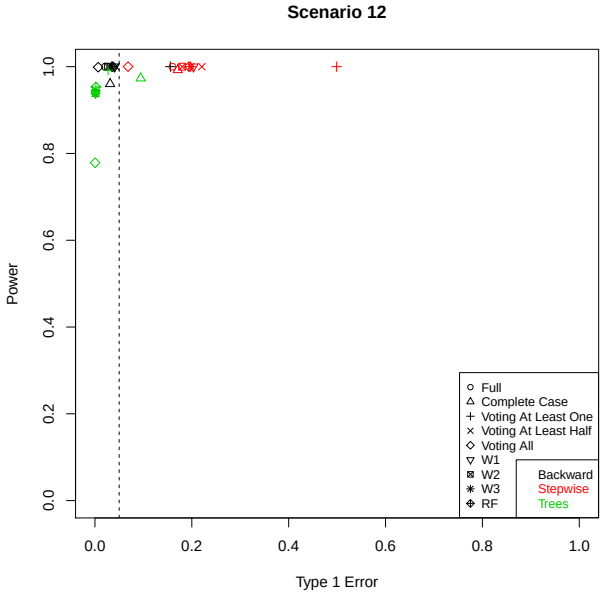


Figure A.15: Power and type 1 error for scenario twelve, MNAR, sample size 1000 and 20% missing in each variable.

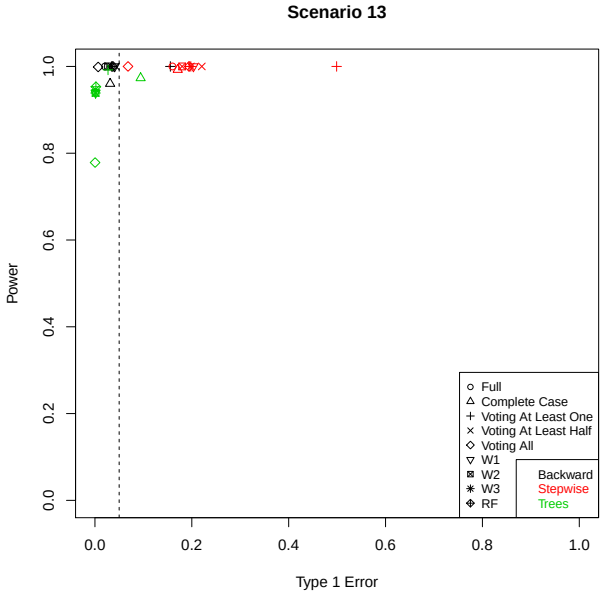


Figure A.16: Power and type 1 error for scenario thirteen, MNAR, sample size 1000 and 30% missing in each variable.

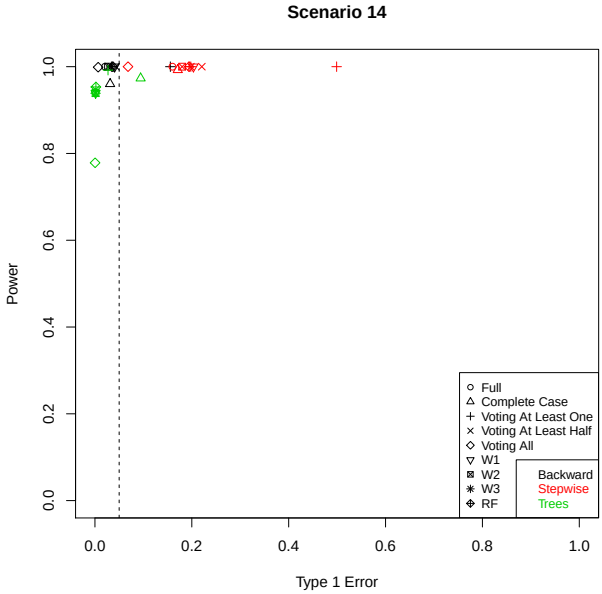


Figure A.17: Power and type 1 error for scenario fourteen, MNAR, sample size 700 and 10% missing in each variable.

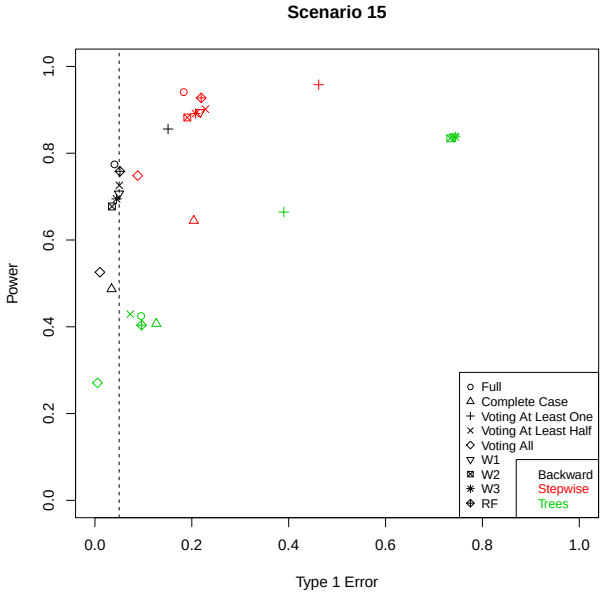
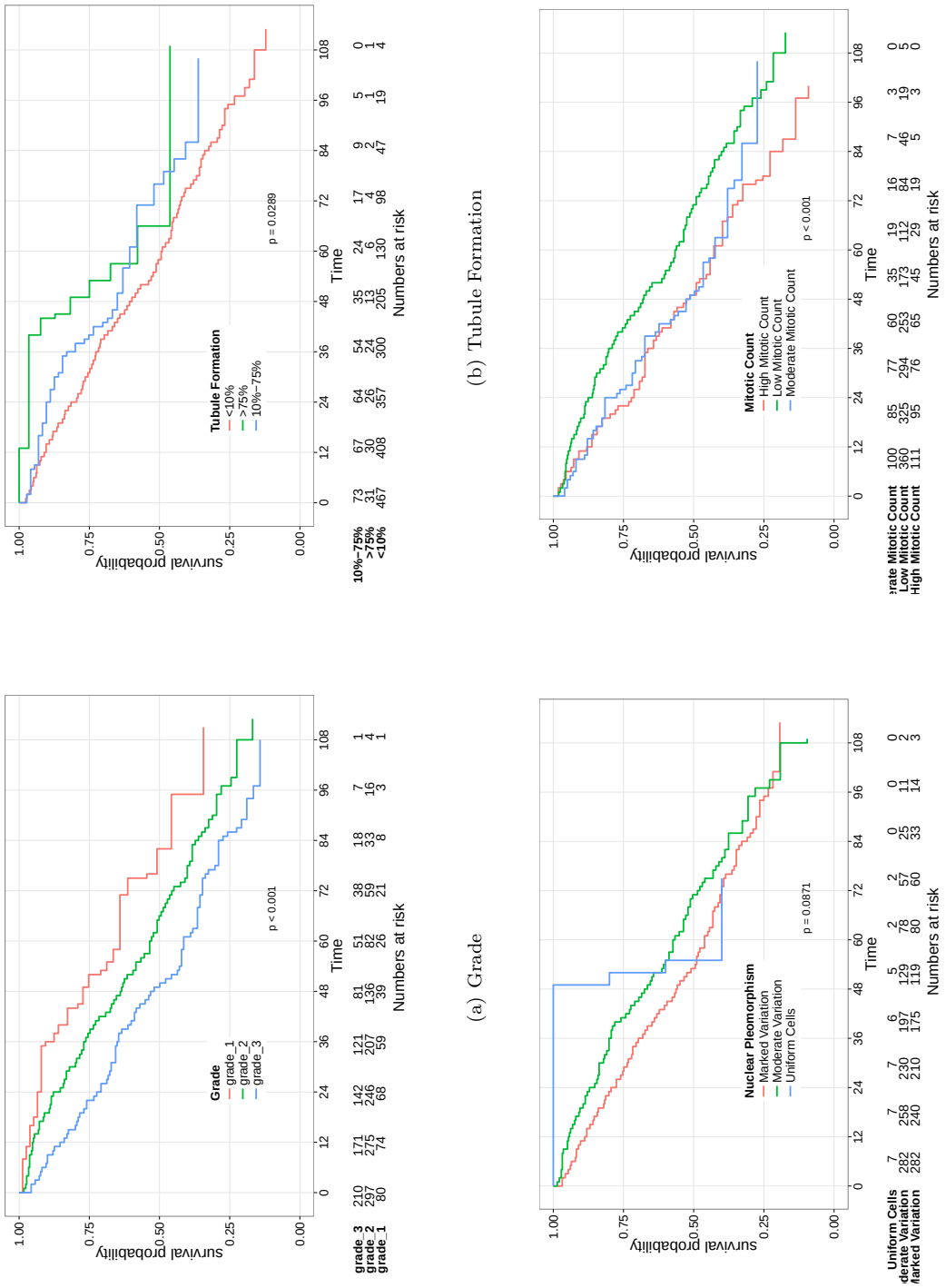


Figure A.18: Power and type 1 error for scenario fifteen, MNAR, sample size 100 and 10% missing in each variable.

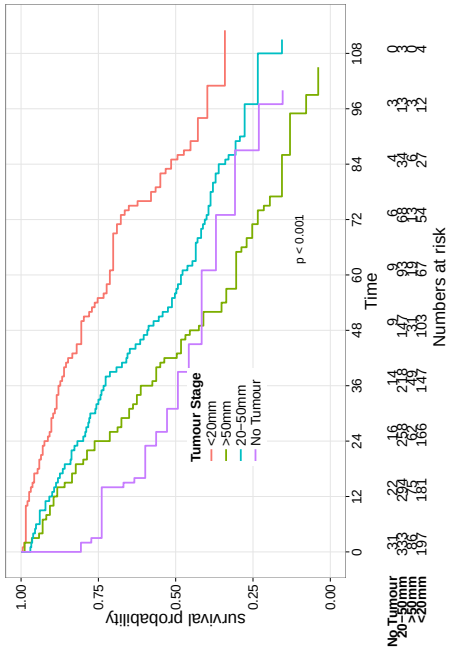
A.3 Kaplan Meier Estimates for DFS and OS



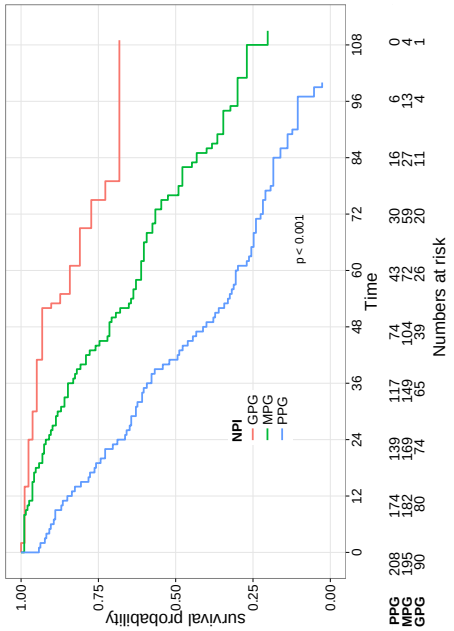
(a) Grade (b) Tubule Formation

(c) Nuclear Pleomorphism (d) Mitotic Count

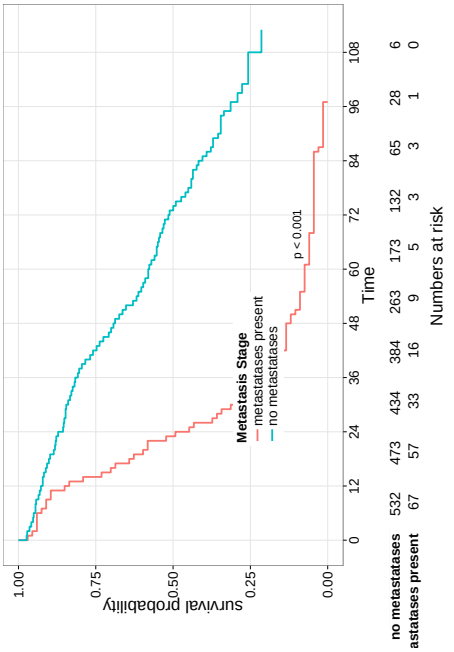
Figure A.19: Kaplan Meier estimates for various routinely assessed predictors for Disease Free Survival.



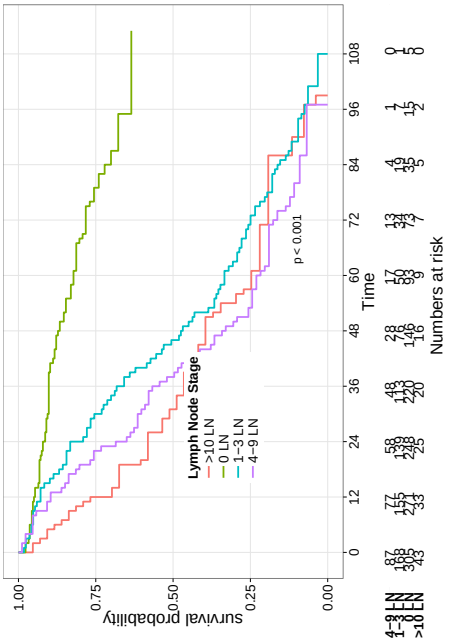
(a) NPI



(b) Tumour Stage

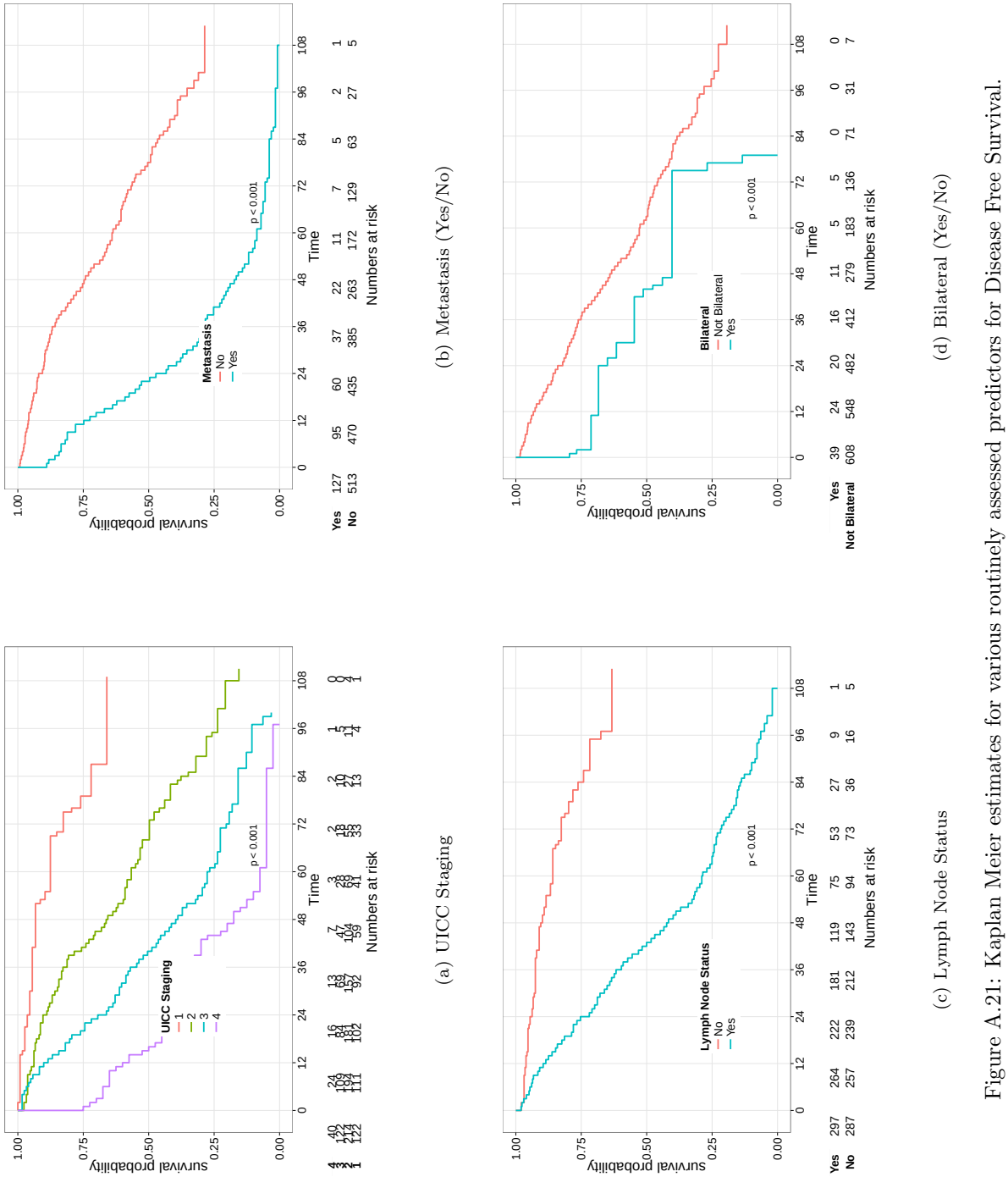


(c) Lymph Node Stage



(d) Metastasis Stage

Figure A.20: Kaplan Meier estimates for various routinely assessed predictors for Disease Free Survival.



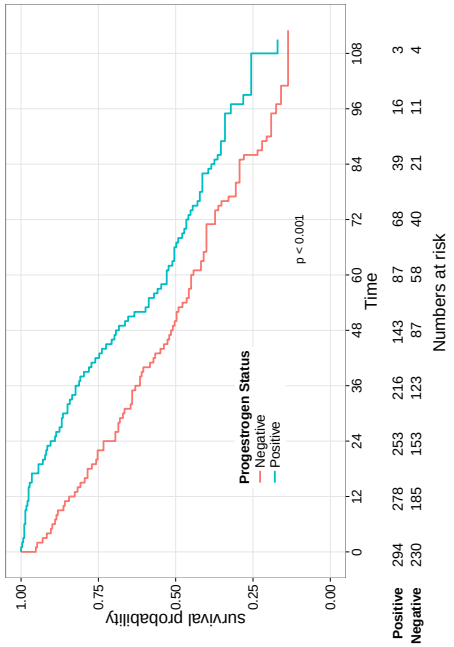
(a) UICC Staging

(b) Metastasis (Yes/No)

(c) Lymph Node Status

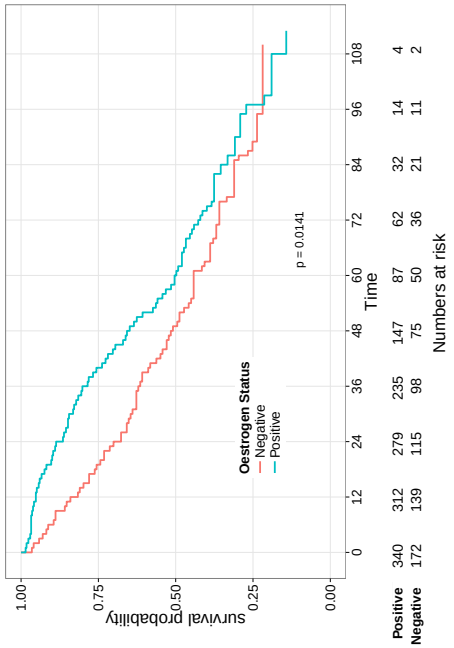
(d) Bilateral (Yes/No)

Figure A.21: Kaplan Meier estimates for various routinely assessed predictors for Disease Free Survival.



(a) Oestrogen Status

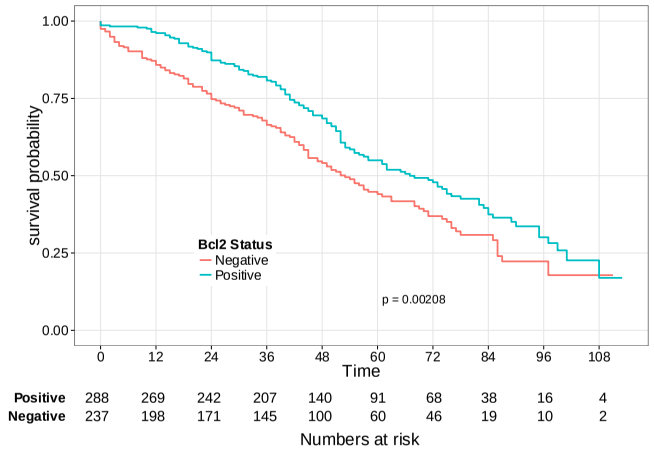
(b) Progesterone Status



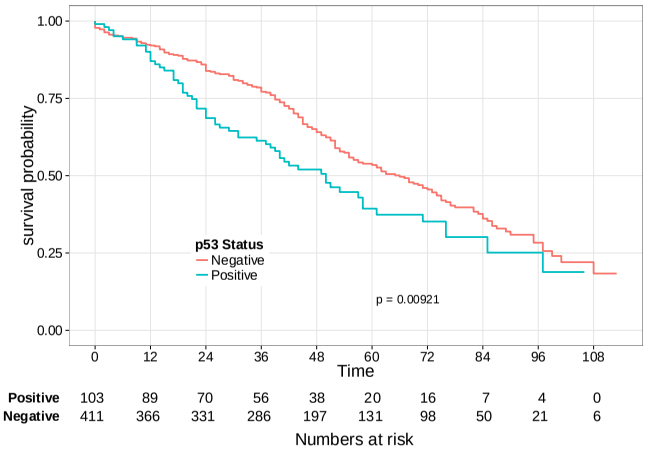
(c) Her2 Status

(d) Ki67 Status

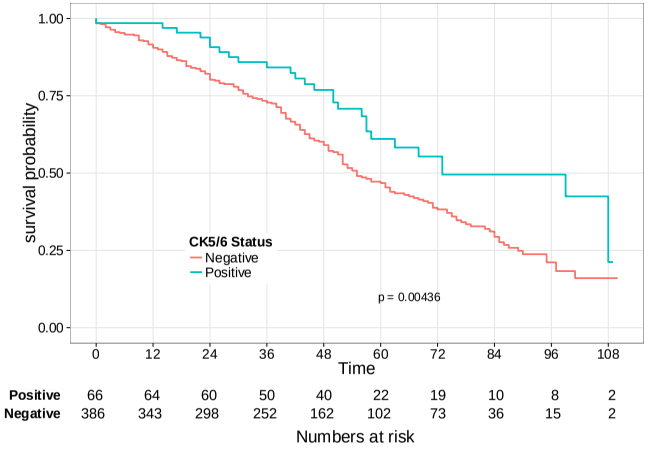
Figure A.22: Kaplan Meier estimates for various biomarkers for Disease Free Survival.



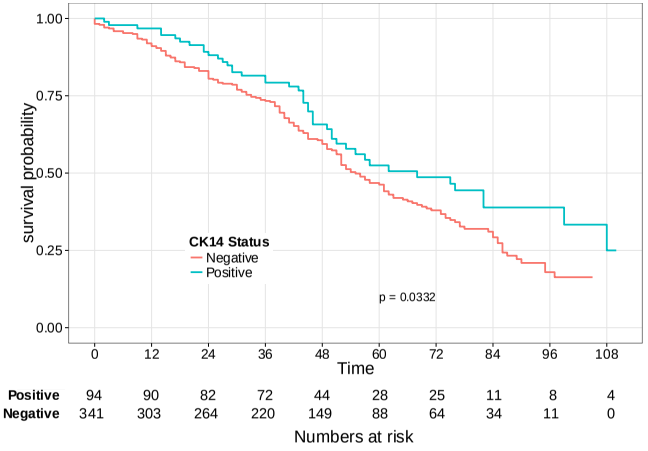
(a) Bcl2 Status



(b) P53 Status

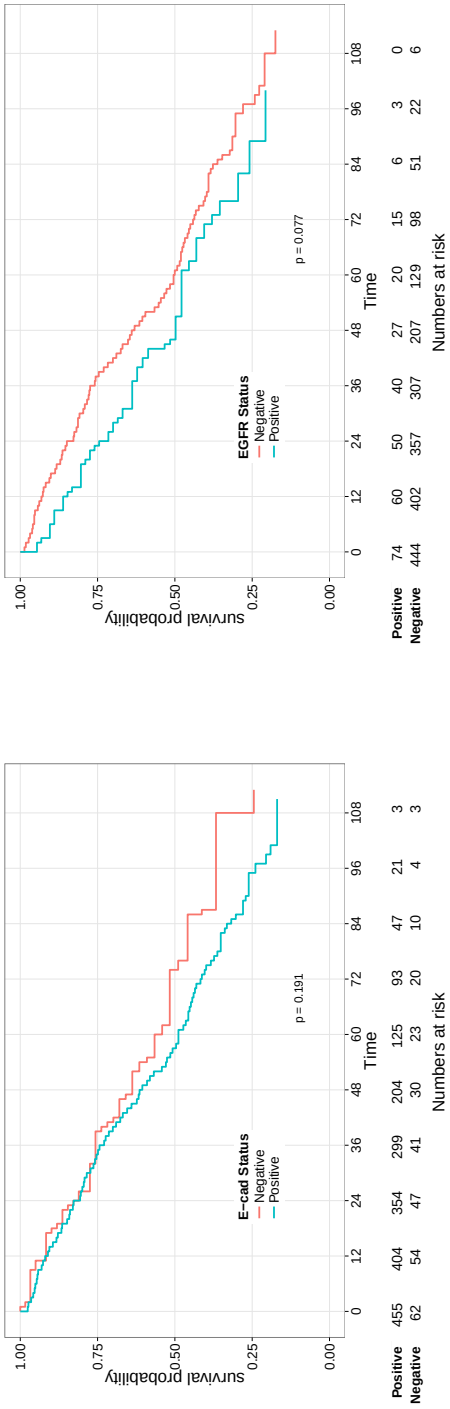


(c) CK5/6 Status



(d) CK14 Status

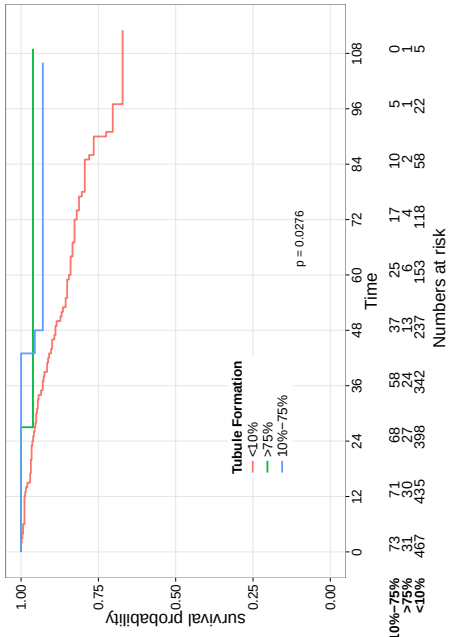
Figure A.23: Kaplan Meier estimates for various biomarkers for Disease Free Survival.



(a) E-cad Status

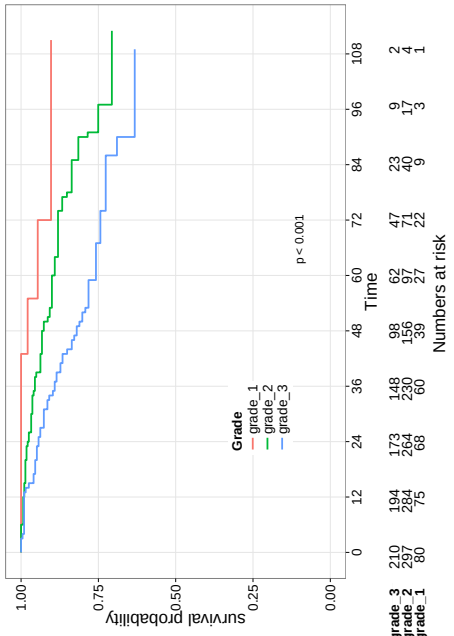
(b) EGFR Status

Figure A.24: Kaplan Meier estimates for various biomarkers for Disease Free Survival.

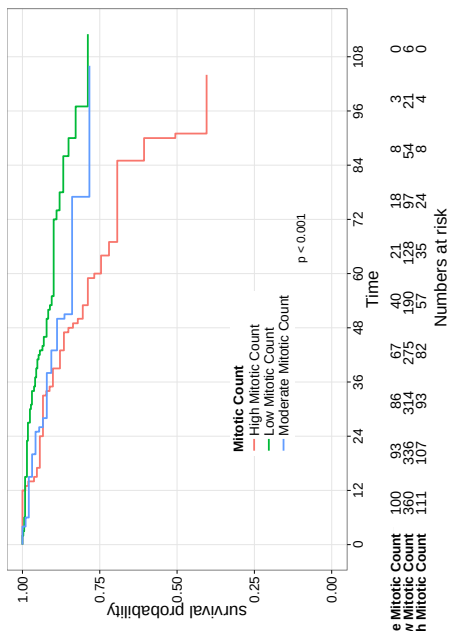


(a) Grade

(b) Tubule Formation

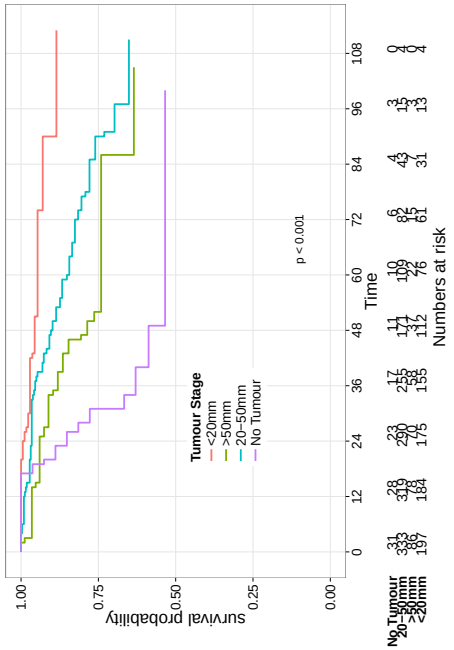


(c) Nuclear Pleomorphism

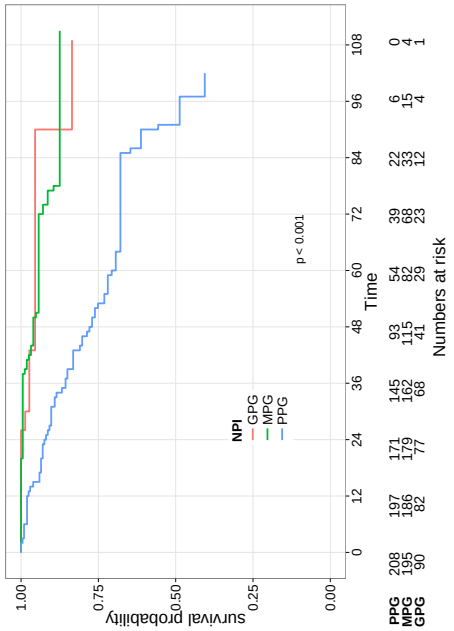


(d) Mitotic Count

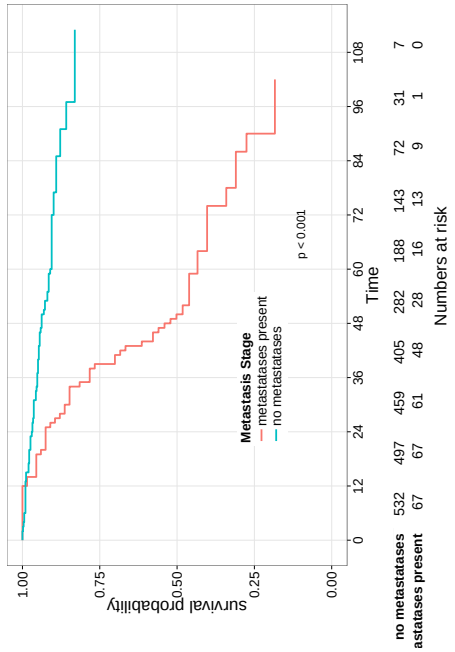
Figure A.25: Kaplan Meier estimates for various routinely assessed predictors for Overall Survival.



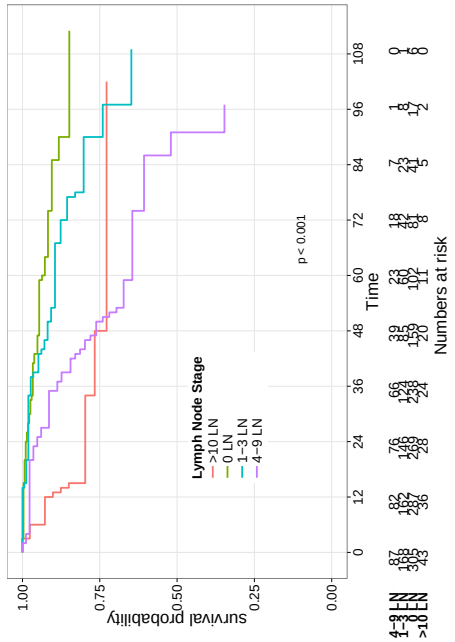
(b) Tumour Stage



(a) NPI

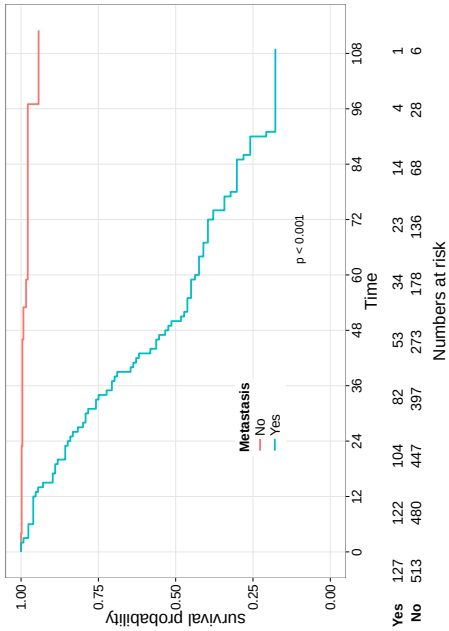


(d) Metastasis Stage

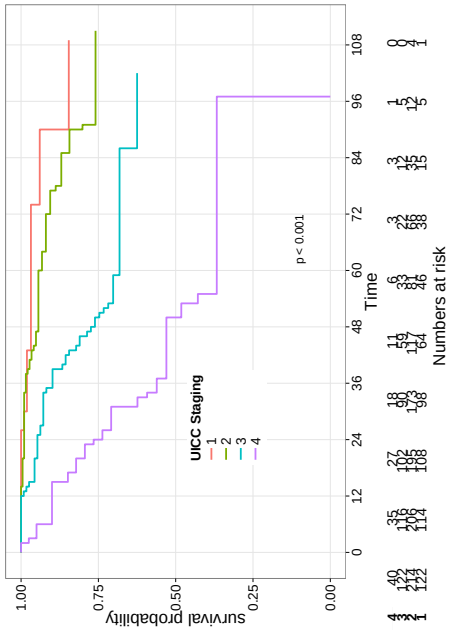


(c) Lymph Node Stage

Figure A.26: Kaplan Meier estimates for various routinely assessed predictors for Overall Survival.



(b) Metastasis (Yes/No)

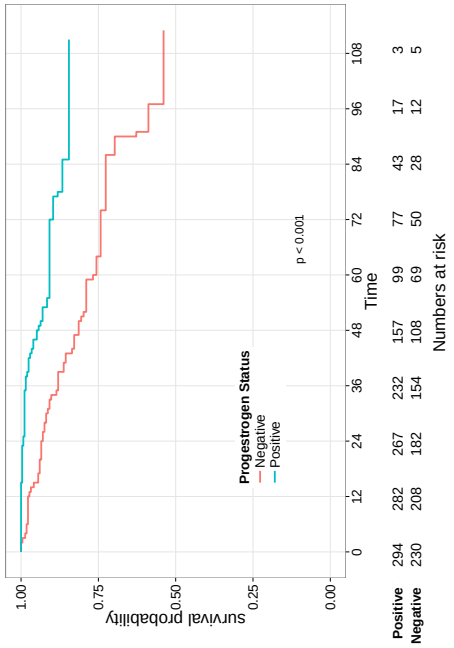


(a) UICC Staging

(d) Bilateral (Yes/No)

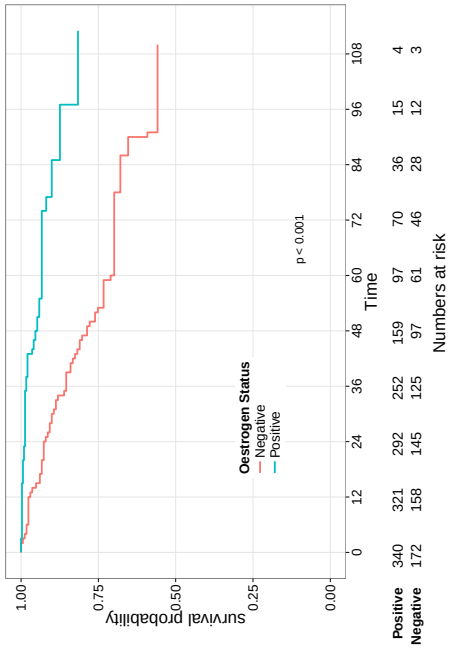
(c) Lymph Node Status

Figure A.27: Kaplan Meier estimates for various routinely assessed predictors for Overall Survival.



(a) Oestrogen Status

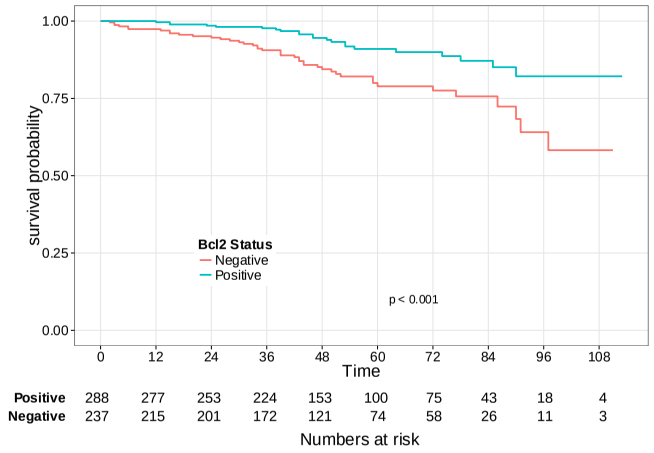
(b) Progesterone Status



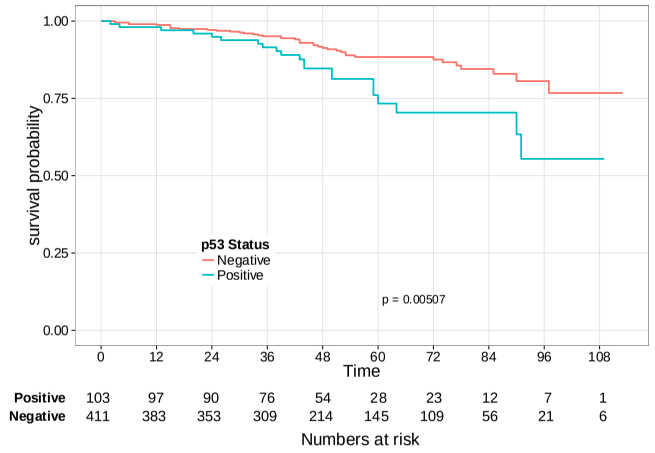
(c) Her2 Status

(d) Ki67 Status

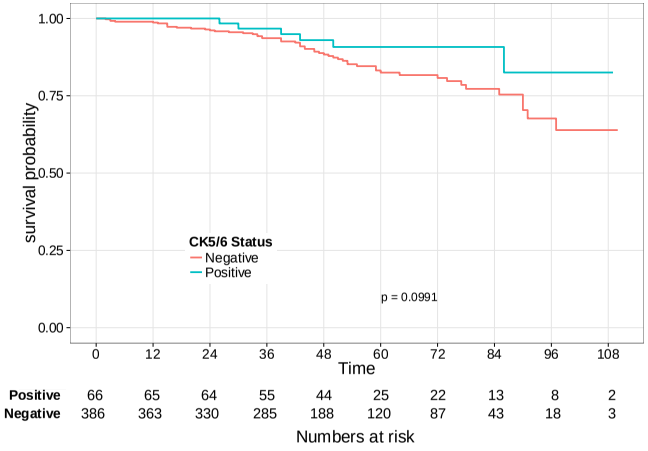
Figure A.28: Kaplan Meier estimates for various biomarkers for Overall Survival.



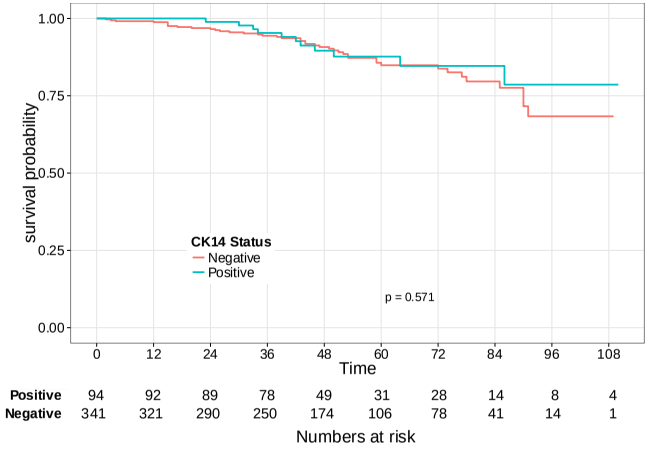
(a) Bcl2 Status



(b) P53 Status

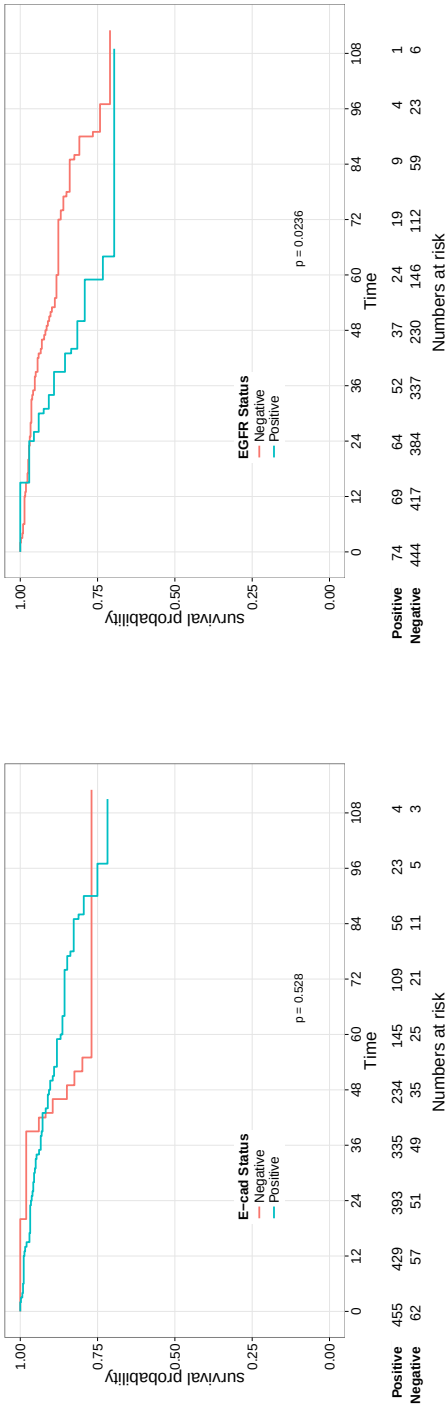


(c) CK5/6 Status



(d) CK14 Status

Figure A.29: Kaplan Meier estimates for various biomarkers for Overall Survival.



(a) E-cad Status

(b) EGFR Status

Figure A.30: Kaplan Meier estimates for various biomarkers for Overall Survival.

A.4 Checking the Proportional Hazards Assumptions

To examine the proportional hazards assumptions, the Schoenfeld residuals were calculated for each of the predictors separately and tested for correlation with time. Also plots of the smoothed trends in the residuals are given in `Figures A.31 and A.32`. The plot function for `cox.zph` objects uses restricted cubic splines to smooth the relationship.

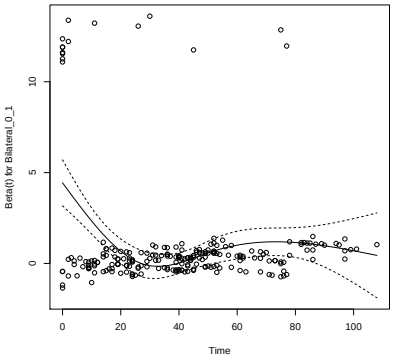
For the DFS model, Bilateral, Lymph Node status, Metastasis and UICC significantly changes over time `Table A.17`. A graphical examination of the trends (`Figure A.31`) doesn't find anything interesting. We are going to ignore the possible increase/decrease in effects over time. If the assumption is violated, a more accurate model could be created including an interaction of each of the predictors with time. For the OS model, there is no predictor which changes significantly over time and the global test has a p-value of 0.191.

	rho	chisq	p-value
Bilateral	-0.1589	5.752	0.017
Mitotic Count	-0.0604	0.894	0.344
Lymph Node Status	0.1854	8.371	0.004
Metastasis	-0.1748	7.106	0.008
UICC	-0.1557	5.714	0.017
GLOBAL	NA	24.099	< 0.001

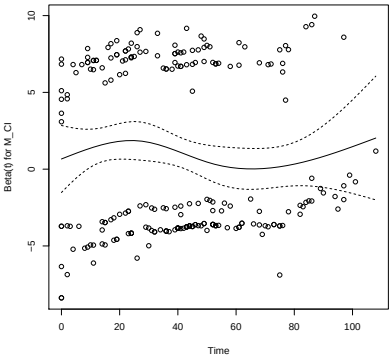
Table A.17: Checking the PH assumption by testing the correlations of the Schoenfeld residuals for each predictor with time for the DFS final model.

	rho	chisq	p-value
Mitotic Count	0.0709	0.342	0.559
Metastasis	-0.1613	1.654	0.198
UICC	-0.1429	1.351	0.245
GLOBAL	NA	4.755	0.191

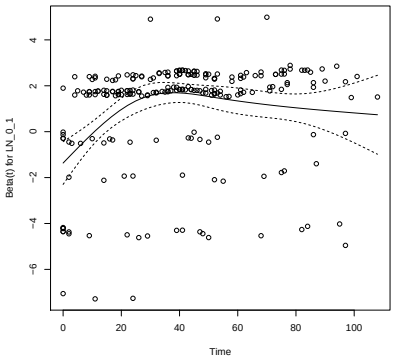
Table A.18: Checking the PH assumption by testing the correlations of the Schoenfeld residuals for each predictor with time for the OS final model.



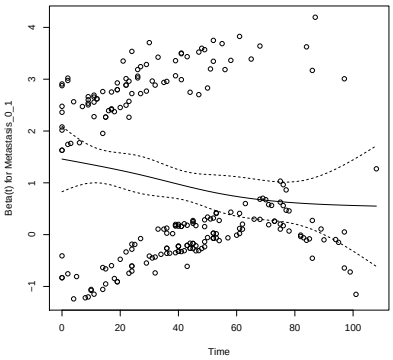
(a) Bilateral



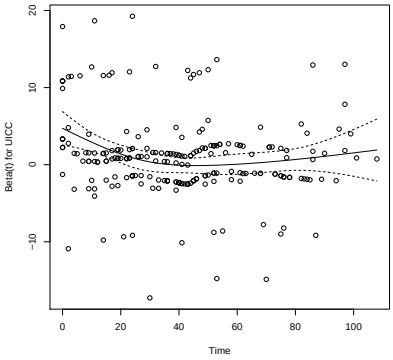
(b) Mitotic Count



(c) Lymph Node Status

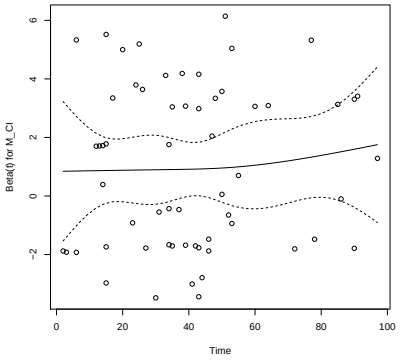


(d) Metastasis

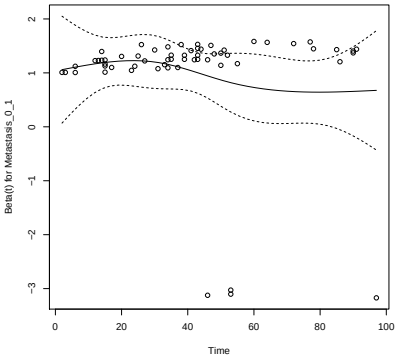


(e) UICC Staging

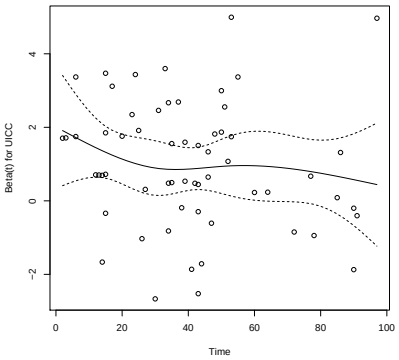
Figure A.31: Raw and spline smoothed scaled Schoenfeld residuals for each of the individual predictors.



(a) Mitotic Count



(b) Metastasis



(c) UICC Staging

Figure A.32: Raw and spline smoothed scaled Schoenfeld residuals for each of the individual predictors.

Bibliography

K. H. Allison, P.L. Kandalaft, C.M. Sitlani, S.M. Dintzis, and A.M. Gown. Routine pathologic parameters can predict oncotype dx recurrence scores in subsets of er positive patients: who does not always need testing? *Breast Cancer Res Treat.*, 131:413–424, 2011.

D.G Altman and G. H. Lyman. Methodological challenges in the evaluation of prognostic factors in breast cancer. *Breast Cancer Research and Treatment*, 52:289–303, 1998.

D.G Altman and P. Royston. What do we mean by validating a prognostic model? *Statistics in Medicine*, 19:453–473, 2000.

J. Auerbach, M. Kim, and S. Fineberg. Can features evaluated in the routine pathological assessment of lymph node-negative estrogen receptor-positive stage i or ii invasive breast cancer be used to predict the oncotype dx recurrence score. *Arch Pathol Lab Med*, 122(4):731–736, 2010.

R. A. Berk. *Statistical Learning From a Regression Perspective*. Springer, 2008.

R.W. Blamey, B. Hornmark-Stenstam, G. Ball, M. Blichert-Toft, L. Cataliotti, A. Fourquet, J. Gee, K. Holli, R. Jakesz, M. Kerin, R. Mansel, R. Nicholson, T. Pienkowski, S. Pinder, M. Sundquist, M. van de Vijver, and I. Ellis. Oncopool - a european database for 16,944 cases of breast cancer. *European Journal of Cancer*, 46:56–71, 2010.

L. Breiman. Random forests. *Machine Learning*, 45:5–32, 2001.

L. Breiman, J. Friedman, C.J. Stone, and R.A. Olshen. *Classification and Regression Trees*. Chapman and Hall, 1984.

Bibliography

K. Bull and D.J. Spiegelhalter. Tutorial in biostatistics: Survival analysis in observational studies. *Statistics in Medicine*, 16(9):1041–1074, 1997.

D.R. Cox. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 34(2):187–220, 1972.

C.W. Elston and I.O. Ellis. Pathological prognostic factors in breast cancer. i. the value of histological grade in breast cancer: experience from a large study with long-term follow-up. *Histopathology*, 41:151–152, 2002.

M. B. Flanagan, D.J. Dabbs, A.M. Brufsky, S. Beriwal, and R. Bhargava. Histopathologic variables predict oncotype dx recurrence score. *Modern Pathology*, 21(10):1255–1261, 2008.

J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1): 1–22, 2010.

J. Geradts, S.M. Bean, R.C. Bentley, and W.T. Barry. The oncotype dx recurrence score is correlated with a composite index including routinely reported pathobiologic features. *Cancer Invest.*, 28(9):969–977, 2010.

M. Greenwood. The natural duration of cancer. reports on public health and medical subjects. *London: His Majesty's Stationery Office*, 33:1–26, 1926.

Philippe Grosjean. *SciViews-R: A GUI API for R*. UMONS, MONS, Belgium, 2012. URL `http://www.sciviews.org/SciViews-R`.

B. Haibe-Kains, C. Desmedt, C. Sotiriou, and G. Bontempi. A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all? *Bioinformatics*, 19:2200–2208, 2008.

F. E. Harrell Jr. Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15:361–387, 1996.

F. E. Harrell Jr. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression and Survival Analysis*. Springer, 2001.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, second edition, 2008.

T. Hesterberg, D.S. Moore, S. Monaghan, A. Clipson, R. Epstein, and B.A. Craig. Bootstrap methods and permutation tests. 2007.

T. Hothorn, K. Hornik, and A. Zeileis. Unbiased recursive partitioning: A conditional inference framework. *Journal of computational and Graphical Statistics*, 15:651–674, 2006.

H. Ingoldsby, M. Webber, D. Wall, C Scarrott, J. Newell, and G. Callagy. Prediction of oncotype dx and tailorx risk categories using histopathological and immunohistochemical markers by classification and regression tree (cart) analysis. *The Breast*, 2013.

H. Ishwaran, U.B. Kogalur, E.H. Blackstone, and M.S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

K. J. M. Janssen, A.R. Donders, F.E. Harrell Jr, Y. Vergouwe, Q. Chen, D.E. Grobbee, and Moons K.G. Missing covariate data in medical research: To impute is better than to ignore. *Journal of Clinical Epidemiology*, 63:721–727, 2010.

E. L. Kaplan and P. Meier. Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481, 1958.

J. Karvanen and F. E. Harrell Jr. Visualizing covariates in proportional hazards model. *Statistics in Medicine*, 28:957–966, 2009.

S. Khan, D. Wall, et al. mir-379 acts as a tumour suppressor and regulates cyclin b1 in breast cancer. *PloS One*, 2013.

D.G. Kleinbaum and M. Klein. *Survival Analysis: A Self Learning Text*. Springer, 2005.

W.Y. Loh. Classification and regression trees. *Data Mining and Knowledge Discovery*, 1(1):14–23, 2011.

D. Machin, Y. B. Cheung, and M. Parmar. *Survival Analysis: A Practical Approach*. Wiley, 2006.

A.M. McDermott, D. Wall, et al. Identification and validation of oncologic mirna biomarkers for luminal a breast cancer. *PloS One*, 2013.

P. G. Neville. Decision trees for predictive modeling. *SAS Institute Inc*, 1999.

J. Newell, J.W. Kay, and T.C. Aitchison. Survival ratio plots with permutation envelopes in survival data problems. *Computers in Biology and Medicine*, 36(5):526–541, 2006.

C. Paul, Mason W.M., D. McCaffrey, and S.A. Fox. A cautionary case study of approaches to the treatment of missing data. *Statistical Methods and Applications*, 17:351–372, 2008.

P. Royston. The lognormal distribution as a model for survival time in cancer, with emphasis on prognostic factors. *Statistica Neerlandica*, 59, 2001.

P. Royston and I.R. White. Multiple imputation by chained equations (mice): Implementation in stata. *Journal of Statistical Software*, 45(4), 2011.

D.B. Rubin. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, 1987.

N. Simon, J. Friedman, T. Hastie, and R. Tibshirani. Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39 (5), 2011.

R. Simon and D.G. Altman. Statistical aspects of prognostic factor studies in oncology. *Br. J. Cancer*, 69:979–985, 1994.

L.H. Sobin, M.K. Gospodorowicz, and Wittekind C. *TNM Classification of Malignant Tumours (UICC International Union Against Cancer)*. Wiley-Blackwell, 7th edition, 2009.

C. Springer and W. P. Kegelmeyer. Feature selection via decision tree surrogate splits. In *ICPR'08*, pages 1–5, 2008.

E.W. Steyerberg. *Clinical Prediction Models*. Springer, 2009.

E.W. Steyerberg, M.J.C. Eijkemans, F.E. Harrell Jr, and J.D.F. Habbema. Prognostic modeling with logistic regression analysis: In search of a sensible strategy in small data sets. *Med Decis Making*, 21:45–56, 2001.

C. Strobl, T. Hothorn, and A. Zeileis. Party on! *R Journal*, 2009.

Salford Systems. Cart: Tree-structured non-parametric data analysis. 2001.

J.M.G Taylor, D.P. Ankerst, and R.R. Andridge. Validation of biomarker-based risk prediction models. *American Association Cancer Research*, 14(19):5977–5983, 2012.

J.M Teno. Prediction of survival of older hospitalized patients: The help survival model. *J Am Geriatr Soc*, 48(5):16–24, 2000.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of Royal Statistical Society: Series B*, 58(1):267–288, 1996.

R. Van Buuren and K. Oudshoorn. Multivariate imputation by chained equations. *TNO Prevention and Health*, 2000.

S. van Buuren and K. Groothuis-Oudshoorn. Mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 45(3), 2011.

H. Van Houwelingen and H. Putter. *Dynamical Prediction in Clinical Survival Analysis*. Chapman and Hall, 2011.

R. Van Oirbeek and E. Lesaffre. An application of harrell's c-index to ph frailty models. *Statistics in Medicine*, 2010.

D. Wall, G. Callagy, H. Ingoldsby, M.J. Kerin, C. Scarrott, , and J. Newell. Identifying underlying structure in classification and regression trees using surrogate splits. *International Workshop on Statistical Modelling*, 2012.

D. Wall, G. Callagy, H. Ingoldsby, M.J. Kerin, and J. Newell. Variable selection techniques for multiply imputed data. *Annual Conference of the Internation Society of Clinical Biostatisticians*, 2013.

P.S. Waters, D. Wall, et al. Relationship between circulating and tissue micrornas in a murine model of breast cancer. *PloS One In Press*, 7(11), 2012.

Bibliography

A. Wienke. *Frailty Models in Survival Analysis*. Chapman and Hall, 2010.

A.M. Wood, I.R. White, and P. Royston. How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, 27, 2008.