UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Dynamic prediction of long-term survival in patients with early-stage breast cancer

Ana Sofia da Silva Azevedo

Mestrado em Bioestatística

Dissertação orientada por:
Prof.ª Doutora Lisete Maria Ribeiro Sousa
Mestre Susana Luísa Augusto Esteves

2019

To everyone who has had - or is battling - cancer.

Sente esta brisa comigo
Dá-me a mão. Suavemente
Sem força, nem vigor
Toca-me apenas a mão. Suavemente
Sentemo-nos no cimo do monte
De mão dada. Suavemente
Sintamos então o ar que nos afaga
Com os dedos tocando-se. Suavemente
Observa até onde o mundo se estende
Toca-me o dedo. Suavemente
Toma o pulso do ar. Vê como é livre
Sente-me a pele. Suavemente
Nada esforces. Fecha a vista, se ela se cansa
Beija-me o pensamento. Suavemente
Ouve a liberdade do vento
Abraça-me o pensamento. Suavemente
Ele toca e vai. Vê como é livre
Toca-me apenas o pensamento. Suavemente
Vê como voa desprendido de tudo
Torna apenas o teu pensamento no meu. Suavemente
Voemos nós também livres e desprendidos
O pensamento deixa-o ir. Suave e livremente

Rodrigo Rufino

Acknowledgments

Sem nenhuma ordem específica: À minha mãe, por ser o meu símbolo de coragem e determinação. A que me deu vida e amor. A que sempre limpou as minhas lágrimas, e suportou as minhas quedas. Ao meu pai, que me deu todo o acesso à educação, permitindo que tudo isto fosse possível. Que sempre acreditou em mim e me fez procurar ser continuamente melhor. Ao Rodrigo por, ao crescer comigo, me tornar uma pessoa mais compreensiva, conscienciosa e inteligente. Ao Pedro Messias, pela insistência em que eu continuasse o trabalho, sem nunca me deixar desistir. Por acreditar que tenho talento. A todos os meus colegas da Exigo, em especial à Valeska e à Joana. À Valeska, por me ter ensinado o verdadeiro significado de resiliência. Por me tornar uma profissional mais competente. Por acreditar em mim e nas minhas capacidades. E, claro, por todas as gargalhadas no local de trabalho, essenciais para o meu bem-estar. À Joana, pela amiga fantástica que se tornou em tão pouco tempo. Por, em todos os momentos, conseguir verbalizar aquilo que penso, tendo-se tornado um pilar esssencial no meu dia-a-dia. Ao Tiago Marques, por despoletar e perpetuar o meu interesse na investigação. Por me dar a mão sempre que preciso e nunca me negar ajuda. O meu profundo agradecimento às minhas orientadoras. À professora Lisete, por toda a paciência que demonstrou nos momentos em que lhe desabafava os meus dilemas com a dissertação. Pela disponibilidade, compreensão e apoio prestado. Por ter permitido que eu fizesse investigação exactamente naquilo que ambicionava. À Susana, por me transmitir bastante conhecimento clínico e estatístico. Por ter sido essencial em todas as partes do trabalho. Por ter sempre mostrado disponibilidade em me receber no IPO. À Dr^{a}. Margarida por ter feito parte do processo e ter tornado o estudo possível. Um agradecimento especial ao Registo Oncológico Nacional, em especial à Dr^{a} Ana Miranda e à Dr^{a} Alexandra Mayer, pela disponibilização dos dados.

Obrigada a todos,

Sofia Azevedo,
Lisboa, Maio de 2019

Resumo

O cancro da mama é uma doença heterogénea, com prognóstico bastante variável que requer diversos cuidados. A avaliação do prognóstico, reportado como probabilidade de sobrevivência e/ou recaída, faz-se habitualmente no momento do diagnóstico da doença atendendo a critérios clínico-patológicos. No entanto, para doentes que já sobreviveram um determinado período de tempo, essa probabilidade poderá ser diferente e o seu prognóstico poderá ser descrito de forma mais exata através de métodos de predição dinâmica. Desta forma, medidas de predição dinâmica poderão facultar estimativas de sobrevivência mais corretas, sendo também de grande valor prático para médicos e investigadores. Para um médico, por exemplo, as medidas de predição dinâmica podem ser um grande auxílio no desenvolvimento de um plano de acompanhamento para um doente com determinadas características, na medida em que alguns doentes podem precisar de um tratamento mais intensivo, enquanto que noutros as consultas de rotina ou a realização de exames pode ser mais espaçada. É também muito importante que os pacientes tenham conhecimento do seu prognóstico atual e, portanto, a avaliação do risco necessita de ter em conta o tempo já sobrevivido até então. De facto, manter uma quantificação mais realista do seu prognóstico a longo prazo poderá ser benéfico a nível psicológico e emocional. Atualmente, no cancro da mama, não existem dados atualizados da evolução de estimativas de sobrevivência em função do tempo decorrido sem doença, sendo a pouca informação existente referente sobretudo a coortes mais antigas.

Este estudo é retrospetivo e unicêntrico, e inclui 4620 mulheres com cancro da mama em estadio I, II ou III, diagnosticadas e tratadas no Instituto Português de Oncologia de Lisboa Francisco Gentil de Janeiro de 2006 a Dezembro de 2011, identificadas através do Registo Oncológico Nacional. O objetivo principal foi o de desenvolver métodos de predição dinâmica em doentes com cancro da mama de forma a avaliar como os fatores de prognóstico da doença evoluem ao longo do tempo. As variáveis de interesse incluíram a idade, estadio da doença, grau histológico e subtipo imunohistoquímico, considerando o recetor hormonal (HR) e o *status* do recetor 2 do fator de crescimento epidérmico humano (HER2) (HR+/HER2-, HR+/HER2+, HR-/HER2+, HR-/HER2-). Estas variáveis foram selecionadas com base na sua significância no prognóstico inicial, de acordo com a literatura existente. A sobrevivência global foi definida como o tempo, em dias, desde diagnóstico até morte por qualquer causa. Já a sobrevivência livre de doença foi definida como o tempo, em dias, desde cirurgia até à recidiva do cancro da mama ou morte por qualquer causa. Numa primeira fase, avaliou-se a sobrevivência global e a

sobrevivência livre de doença, condicionais ao tempo vivido sem doença, através do estimador de Kaplan-Meier. A sobrevivência global condicional foi definida como a probabilidade de um paciente sobreviver mais 2 ou 5 anos, condicional a estar vivo e sem recidiva aos 0, 1, 2, 3, 4 e 5 anos após diagnóstico. Já a sobrevivência livre de doença condicional foi estabelecida como a probabilidade de um paciente sobreviver sem recidiva por mais 2 e 5 anos, condicional a estar vivo e sem recidiva aos 0, 1, 2, 3, 4 e 5 anos após cirurgia. Numa segunda fase, avaliou-se a significância a longo prazo de fatores de prognóstico que são relevantes ao diagnóstico e averiguou-se como é que estes variam ao longo do tempo.

Os resultados deste estudo mostraram que, na ausência de covariáveis, a sobrevivência global condicional e sobrevivência livre de doença condicional ao tempo vivido sem doença se mantêm razoavelmente constantes ao longo do tempo, isto é, a probabilidade de sobreviver (livre de doença ou não) por mais 2 e 5 anos é semelhante para um indivíduo que sobreviveu livre de doença 0 ou 5 anos após diagnóstico ou cirurgia. No entanto, na presença de covariáveis, verificou-se que, para indivíduos com estadio III e alto grau histológico ao diagnóstico, a probabilidade de sobreviver livre de doença tende a aumentar gradualmente à medida que mais tempo passa além do diagnóstico ou cirurgia, assemelhando-se a um indivíduo de melhor prognóstico ao diagnóstico. Não obstante, o mesmo se reflete em indíviduos com subtipo imunohistoquímico HR-/HER2-, cuja sobrevivência livre de doença também aumenta com o aumento do tempo desde cirurgia. Para estes indivíduos, 4 anos após cirurgia, a sua probabilidade de sobreviver livre de doença é idêntica à de indivíduos com subtipo imunohistoquímico HR+/HER2- (o grupo com melhor prognóstico no início do estudo). Observam-se ainda ganhos notáveis nas estimativas de sobrevivência dinâmica, comparativamente a estimativas tradicionais, quando estratificamos pacientes por um determinado factor de prognóstico. A título de exemplo, em pacientes com HR-/HER2-, a probabilidade de um paciente sobreviver livre de doença por mais 2 anos, dado que já sobreviveu livre de doença 3 anos após cirurgia, é de 0.91. No entanto, ao considerar uma estimativa estática da probabilidade de sobreviver livre de doença 5 anos, observada apenas no momento de cirurgia e não tendo em conta o tempo já vivido sem doença, esta reduz-se para 0.71. Desta forma, os dados sugerem que, após completar 4 anos após a cirurgia, um paciente com HR-/HER2- poderia mudar para um plano de vigilância similar ao de pacientes com HR+/HER2-. Estes resultados aproximam-se dos obtidos com a metodologia de *landmarking*. Numa primeira fase da abordagem por *landmarking*, foram ajustados modelos de Cox que incluem os efeitos principais das variáveis de interesse, considerando uma janela temporal de 2 e de 5 anos. Estes modelos foram ajustados a cada 3 meses até perfazer 5 anos desde cirurgia, resultando num total de 21 modelos *landmark*, para cada janela temporal. Considerando estes modelos ajustados separadamente para cada ponto no tempo, verificou-se, através de métodos gráficos, que em indivíduos com estadio II ou III e subtipos imunohistoquímicos HR-/HER2+ e HR-/HER2-, o risco de morte ou recidiva tende a diminuir com o tempo de forma linear, considerando uma janela temporal de 2 anos. Já em indíviduos com envolvimento ganglionar positivo ou com moderado/alto grau histológico, o risco de morte ou recidiva parece variar de forma quadrática. Considerando uma janela temporal de 5 anos, verificou-se também que indivíduos com estadio III e subtipos imunohistoquímicos HR-/HER2+ e HR-/HER2- apresentam um decréscimo linear no risco de recidiva e/ou morte. No entanto, o risco parece ser constante ao longo do tempo consoante o grau histológico do tumor ou o envolvimento ganglionar.

Para o modelo que constitui a agregação de todos os modelos *landmark*, verificou-se, através do procedimento de seleção de variáveis por eliminação *backward* e considerando uma janela temporal de 2 anos, que a interação com o tempo das variáveis de prognóstico correspondentes ao estadio da doença, envolvimento ganglionar e subtipo imunohistoquímico mostraram ter uma influência significativa no risco de morte ou recidiva. Aumentando a janela temporal para 5 anos, as interações com o tempo que se mostraram significativas no risco de morte ou recidiva reduziram-se apenas aos fatores de prognóstico correspondentes ao estadio da doença e ao grupo imunohistoquímico. Os dois modelos foram avaliados relativamente à sua capacidade preditiva, através de medidas que quantificam a discriminação e a calibração. Em ambos os modelos, tanto as medidas de discriminação como as de calibração, apresentam valores razoáveis.

Este pode ser o primeiro estudo português a atribuir explicitamente probabilidades de sobrevivência, aplicando modelos de sobrevivência condicional no contexto do cancro da mama. A adoção de sobrevivência condicional poderá ajudar os médicos a prever melhor a sobrevivência dos pacientes, ajustar o programa de vigilância e monitorização e conduzir uma discussão mais informada com os mesmos. Serão necessários mais estudos com acompanhamento a longo prazo para confirmar os nossos resultados. Se confirmados, estes são bastante relevantes para informar e aconselhar os pacientes sobre a natureza dinâmica do seu prognóstico a longo prazo e devem ser considerados nos planos de vigilância dos pacientes.

Palavras-Chave: Predição dinâmica, sobrevivência condicional, *landmarking*, cancro da mama, análise de sobrevivência

Abstract

Cancer estimates are typically reported in terms of survival from time of diagnosis. However, for patients surviving past a given duration from diagnosis, subsequent prognosis can be quite different from the one observed at the time of diagnosis. Given the heterogeneity of breast cancer, a more accurate quantification of prognosis for long-term survivors should be provided. The purpose of this study was to investigate the long-term effect of prognostic factors of breast cancer. Data variables included age, disease stage, tumour grade, axillary lymph node status and immunohistochemistry subgroups considering hormone receptor and human epidermal growth factor receptor 2 (HER2) status. Using data from 4620 patients diagnosed and treated in Instituto Português de Oncologia de Lisboa Francisco Gentil between january 2006 and december 2011 we analysed the overall survival and disease-free survival of patients with early-stage breast cancer conditional on time lived without disease for each covariate, through conditional survival techniques. Thus, we assessed time-varying effects of such covariates using the a novel approach: landmarking. Notable gains in conditional survival estimates were found in patients with negative hormone receptors and negative HER2 status. As time goes by, survival estimates for such patients tend to be equal to survival estimates of patients with better prognosis at baseline. For this reason, data suggests that, after completing 4 years after surgery, an HR-/HER2- patient could switch to a surveillance plan similar to HR+/HER2- patients (the group with better prognosis at baseline). Fitting a proportional baselines landmark supermodel allowed to verify a decrease over time of the prognostic significance of immunohistochemistry groups and stage at diagnosis. Models fitted were evaluated with respect to their predictive accuracy, through measures assessing discrimination and calibration. Further studies with long-term follow-up are needed to confirm our results. If confirmed, these findings are relevant to inform and counsel patients regarding the dynamic nature of their prognosis over time and should be considered in surveillance plans.

Keywords: Dynamic prediction, conditional survival, landmarking, breast cancer, survival analysis

Contents

List of tables

List of figures

1 Introduction

Cancer survival statistics are of great interest to patients, who understandably wish to have some information on the estimated prognosis for their condition, and to the clinicians providing direct care to patients. From a global public health perspective, cancer survival statistics also yield important information not only to identify policy approaches associated with best outcomes but also to inform programs, policies, and practices to address the needs of cancer survivors. Breast cancer is a heterogeneous disease whose diagnosis, prognosis, and treatment depend on multiple factors rather than a single characteristic. Therefore, a tailored treatment and follow-up are determined by several factors that significantly affect the disease-free survival of breast cancer patients (Paik *et al.*, 2017). In breast cancer patients, the risk of relapse and expected survival are usually estimated at the time of diagnosis based on clinical and pathological factors. Although risk stratification at diagnosis is central for the decision on the initial therapeutic approach, these survival estimates may not provide accurate information on long-term prognosis. This has been demonstrated in other types of cancer, where the probability that a cancer patient will survive for an additional period increases as patient lives longer (conditional survival) (Janssen-Heijnen *et al.*, 2010). Conditional survival analysis attempts to better understand the patient survival and long-term prognostic factors over the course of the disease.

In breast cancer there is limited up-to-date data on the evolution of the relapse and/or survival probability as a function of disease-free elapsed time. The existing information mostly refers to old cohorts which have not been treated with current standard treatments and does not discriminate conditional survival in biological subgroups that present different clinical and prognostic behaviour. Understanding more about conditional survival after diagnosis and treatment in breast cancer patients may help the clinician develop the plan for the follow-up period. For example, certain patients may need more intensive management during the follow-up period, whereas other patients may not need routine checks or the intervals between examinations can be lengthened (Paik *et al.*, 2017). In these circumstances, the baseline prognostic estimates become less relevant as time increases and dynamic prediction methods present more relevant measures to predict the future course of the disease, conditional on the current history. This emphasizes the importance of this work which may contribute to get further insights into the long-term prognosis and recurrence pattern of early-stage breast cancer survivors. In

addition, we hope to identify some relevant long-term prognostic factors, which would allow the identification of high and low-risk groups that might benefit from differentiated surveillance protocols.

1.1 Breast cancer

Non-communicable diseases, also known as chronic diseases, are growing in the world, due to an increased lifetime, prolonged exposure to risk factors, and life style changes (Ghoncheh *et al.*, 2016). It is indubitable that cancer is one of the most important diseases and a major cause of mortality worldwide (Boutayeb & Boutayeb, 2005). It is also expected that in the next two decades the number of new cancers will rise by about 70% (WHO, 2015).
One of the most common types of cancers is breast cancer. To better understand breast cancer, it helps to understand how a cancer develops. At the cellular level, the development of cancer is viewed as a multistep process involving mutation and selection for cells with progressively increasing capacity for proliferation, survival, invasion, and metastasis. The first step in the process is thought to be the result of a genetic alteration leading to abnormal proliferation of a single cell. Cell proliferation then leads to the outgrowth of a population of clonally derived tumour cells (Cooper & Hausman, 2007). Therefore, breast cancer is an uncontrolled growth of breast cells.

According to its physiological function, the structure of the female breast (shown in Figure 1.1) can be divided into three essential components: lobules, ducts and connective tissue (Zimmerman, 2004). The lobules contain glandular structures that, in the presence of adequate hormonal stimuli, produce breast milk; ducts are the channels that connect the glandular structures and lobules and transport their secretion to the nipple. The remainder of the breast is made up of fatty, connective, and lymphatic tissues. Most breast cancers begin either in the lobules, or in the ducts.

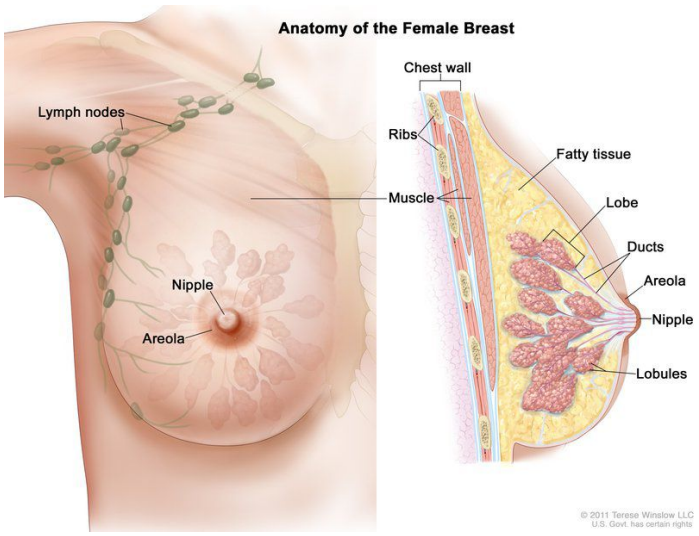

Figure 1.1: Anatomy of female breast (National Breast Cancer Foundation)

In order to determine treatment options and prognosis it is crucial to determine the stage of the disease. Different stages varies from 0 to IV, with stage 0 being the earliest stage (carcinoma *in situ*), and the IV stage the most aggressive type corresponding to a widespread disease. Hence, an increase in the staging is associated with decreased survival rates. The definition of the stage is usually based on the TNM system, overseen by the American Joint Committee on Cancer (AJCC) (Sobin *et al.*, 2009), and considers tumour size (T), involvement of axillary lymph nodes (N) and presence of distant metastasis in any other part of the body (M). The latest revision ($8th$ edition) to this system outlines a new prognostic staging system that relies not only on the anatomic extent of disease, but also on prognostic biomarkers, such as the human epidermal growth factor receptor 2 (HER2), estrogen receptor (ER) and progesterone receptor (PR). Estrogen and progesterone are hormones that can stimulate the growth of breast cancer cells (Louie & Sevigny, 2017). On the other hand, HER2 is a protein that belongs to the HER family of membrane receptors that are important mediators of cell growth, survival and differentiation (Chia *et al.*, 2008; Ross *et al.*, 2009; Wolff *et al.*, 2018).

Breast cancer is characterized by its heterogeneity in both its etiology and pathology with some women having good prognosis whereas others experience a highly aggressive clinical course. Hence, it is not a single disease, as it comprises many biologically different entities with distinct pathological roles, clinical implications and treatment responsiveness (Spitale *et al.*, 2009; Iwamoto & Pusztai, 2010).

1.1.1 Epidemiology

Breast cancer is a public health problem with a high incidence and a high mortality rate. It is the most commonly occurring malign neoplasm in women with an estimated 1.7 million new cases diagnosed in 2012 and 522.000 deaths (Ferlay *et al.*, 2015).

It is well recognized that while the breast cancer incidence is higher in more developed countries, the mortality due to breast cancer is higher in women from poorer countries (Tao *et al.*, 2015), where patients do not receive suitable care (Vahabi *et al.*, 2015). According to Ghoncheh, Pournamdar & Salehiniya (2016), the highest incidence rates were 91.6 for Northern America and 91.1 for Western Europe, per 100.000 person-years. In contrast, the lowest incidence rates were in Middle Africa and Eastern Asia (26.8 and 27 per 100.000 person-years, respectively). With regard to mortality, the highest rate was 17 per 100.000 person-years in Africa whereas the lowest is found in Eastern Asia.

Incidence rates of breast cancer are expected to further increase in less developed countries due to longer life expectancy coupled with the adoption of a more westernized lifestyle, less physical activity, and delays in childbearing. As reported by Tao *et al.* (2015), the international incidence of female breast cancer will probably reach approximately 3.2 million new cases per year by 2050.

Given the high rates of this disease, there is a particular interest in treating and preventing breast cancer. Consequently, more clinical trials and different treatments strategies are available with regard to breast cancer (Dieterich *et al.*, 2014).

1.1.2 Risk factors

There are many risk factors related to breast cancer, closely associated with the lifestyles and reproductive characteristics inherent in modern and westernized life. Note that there are 5-10% of breast cancers diagnosed with genetic and hereditary characteristics that require an earlier and careful monitorization of the family members.

Although one cannot change some breast cancer risk factors (e.g. family history and aging), some risk factors can be controlled. Thus, modifiable and lifestyle-associated risk factors are important to consider when developing a strategy for breast cancer prevention. The most common risk factors are listed below (Liga Portuguesa Contra o Cancro & Cancer.Net):

- **Age**: The possibility of having breast cancer increases with age; a woman over 60 has an increased risk. Also, breast cancer is less common before menopause;
- **Family history**: Women with one first-degree relatives who have been diagnosed with breast cancer before 55 years old have a higher risk of developing the disease. Nevertheless, fewer than 15% of women with breast cancer have a family member with this disease;
- **Personal history:** a woman who has already had breast cancer (in one breast), has a higher risk of having this disease in the other breast;
- **Race and ethnicity:** Breast cancer occurs more often in Caucasian women compared to Latina, Asian, or African-American women;
- **Obesity:** Overweight women have a higher risk of being diagnosed with breast cancer compared to women with a healthy weight, especially after menopause;
- **Pregnancy history:** Women who have not had a full-term pregnancy or that had their first child after 30 years of age have a higher risk of breast cancer compared to women who gave birth before 30 years old;
- **Breastfeeding history:** Breastfeeding can lower breast cancer risk especially if a woman breastfeed for longer than 1 year;
- **Menstrual history:** Women who started menstruating younger than the age of 12 have a higher risk of breast cancer later in life due to breasts forming earlier;
- **Using hormone replacement:** Women who take hormone therapy for menopause for 5 years or more after menopause also appear to be more likely to develop breast cancer;
- **Radiation therapy:** Women who have had chest radiation therapy are at increased risk for breast cancer.

1.1.3 Diagnosis, prevention and treatment

Early breast cancer is often asymptomatic and is usually diagnosed following an abnormal mammogram or by physical examination (Dipiro *et al.*, 2014). However, when symptomatic, signs and symptoms can include a lump in the breast, a lump or swelling in the armpit, change in

shape, size or texture of the breast, and a change in the nipple. Once breast cancer is suspected, a series of tests are performed to diagnose the patient. Many of these tests are also used to determine the stage of disease and include assessment of lymph node involvement and hormone receptor and HER2 status.

- **Lymph Node Assessment:** Breast cancer cells can spread to the axillary lymph nodes via the lymphatic system. To determine if the lymph nodes contain cancer, a lymph node pathology assessment is required. If the nodes contain cancer, this is known as lymph node-positive disease. Otherwise, it is known as lymph node-negative disease. The number and location of nodes containing cancer is used to determine the stage of the disease (Sobin *et al.*, 2009). Patients with lymph node-positive disease are a subgroup considered to be at high risk of recurrence, compared with patients with lymph node-negative disease (Cianfrocca & Goldstein, 2004)

- **Hormone Receptor Status Testing:** Hormone receptor (HR) testing is conducted at the time of initial diagnosis or when there are signs of disease recurrence. A tumour biopsy is taken to determine the presence or absence of HR, through the evaluation of estrogen receptor (ER) and progesterone receptor (PR). If HR are found, the tumour is hormone receptor-positive (HR+) otherwise is hormone receptor-negative (HR-). Tumours can also have a combination of positive and negative receptors (e.g., ER+/PR-). Patients with HR- disease are considered to be at high risk of recurrence and tend to relapse earlier than patients with HR+ disease (Strasser-Weippl *et al.*, 2015).

Early detection and appropriate diagnosis are critical to achieve a favourable breast cancer outcome, with mammography currently being the standard of care in breast screening. Mammogram screening can reduce breast cancer mortality by 20 - 30% in women over 50 years old in high-income countries when the screening coverage is over 70% (IARC, 2008). Furthermore, it has also been associated with less disabling treatments and better quality of life after treatment (Gastrin *et al.*, 1994; Alexander *et al.*, 1999). In Europe, the mortality rate due to breast cancer had a reduction of 19% between 1989 and 2006 as a result to the implementation of preventive strategies and a greater effectiveness of therapy (Moss *et al.*, 2012). Bastos *et al.* (2007) established that an increase in early detection of the disease and better access to more effective treatments can lead to a lower mortality.

There are several ways to treat breast cancer, depending on its type and stage. Local treatments treat the tumour without affecting the rest of the body and include surgery and radiation therapy. The most common surgery is mastectomy, where the entire breast is removed (in most cases, it is also removed the axillary lymph nodes). On the other hand, systemic treatments occur when drugs are used to treat breast cancer. These drugs are considered systemic therapies because they can reach cancer cells almost anywhere in the body. Depending on the type of breast cancer, different types of drug treatments might be used, including chemotherapy and hormone therapy (American Cancer Society). Treatments such as surgery, chemotherapy, or radiation may reduce the mass of the tumour, but metastatic cells may remain in lymph nodes and eventually resume their rampage travelling to other distant parts of the body (Zimmerman, 2004).

1.1.4 Prognostic and predictive factors

Breast cancer is a heterogeneous disease with variations in its clinical behaviour manifestations, with the biological nature of the disease and clinical outcomes being closely interlinked. Management of the breast cancer patient is a carefully planned exercise using a variety of factors which are associated with longer or shorter survival (prognostic factors), and/or can aid selection of relevant systemic therapy (predictive factors) (Clark, 1995).

Prognostic factors

Key prognostic factors of breast cancer are lymph node status, tumour size and tumour grade. Other prognostic factors include presence or absence of lymphovascular invasion, age, and ethnicity (Cianfrocca & Goldstein, 2004; Clark, 1995). Lymph node status and tumour size are associated with the ability of the cancer to spread beyond the primary tumour (Carter, Allen & Hension, 1989), whilst tumour grade and HR status are more associated with tumour cell proliferation (growth in primary tumour size) (Cianfrocca & Goldstein, 2004).

- **Axillary lymph node status:** The presence or absence of axillary node involvement is one of the most important prognostic factors for patients with breast cancer. In lymph node-positive disease, the risk of recurrence is sufficiently significant to warrant adjuvant systemic therapy (Veronesi *et al.*, 1993; Cianfrocca & Goldstein, 2004).

- **Tumour size and grade:** Tumour size correlates with the presence and number of involved axillary lymph nodes. Elston, Ellis & Pinder (1999) mentioned that is a time-dependent prognostic factor and can influence the outcome; patients with smaller tumours were shown to have a better long-term survival rate than those with larger tumours. According to the Nottingham grading system (also called the Elston-Ellis modification of the Scarff-Bloom-Rihardson grading system), histological tumour grade is based on the following features: tubule formation (how much of the tumour tissue has normal breast duct structures); nuclear grade (an evaluation of the size and shape of the nucleus in the tumour cells) and mitotic rate (how many dividing cells are present, which is a measure of how fast the tumour cells are growing and dividing). Each of the categories gets a score between 1 and 3; a score of 1 means the cells and tumour tissue look the most like normal cells and tissue, and a score of 3 means the cells and tissue look the most abnormal. The scores for the three categories are then added, yielding a total score. According to the total score obtained, the tumour can be categorized in three grades: low grade or well differentiated; intermediate grade or moderately differentiated and high grade or poorly differentiated. It is established that high grade tumours are associated with decreased survival rates and correlated with poor prognostics (Elston & Ellis, 1991; Ellis *et al.*, 1992; Rakha *et al.*, 2010). It represents the morphological assessment of tumour biological characteristics and has been shown to be able to generate important information related to the clinical behaviour of breast cancers (Rakha *et al.*, 2010). There is compelling evidence to suggest that histological grade can accurately predict tumour behaviour, particularly in earlier small tumours, more than other time-dependent prognostic factors such as tumour size (Rakha *et al.*, 2010).

Predictive factors

A predictive factor is any measurement that is associated with response or lack of response to a particular therapy. Key predictive factors of breast cancer include HR status and HER2/neu (Clark, 1995).

- **Hormone Receptor Status:** The presence or absence of ER and/or PR in breast cancer is both prognostic and predictive. In HER2+ breast cancer, the presence of ER and/or PR (known as HR+ disease) may be a predictor for a more indolent, slower growing tumour with longer times to disease recurrence (Cianfrocca & Goldstein, 2004). Patients with HR- disease are known to be a subgroup at high risk of recurrence within the HER2+ breast cancer population, and tend to relapse earlier than patients with HR+ disease (Strasser-Weippl *et al.*, 2015). Hence, HR+ breast cancers have a better prognosis because these tumours tend to be lower grade and have less aggressive phenotypes (Louie & Sevigny, 2017).

- **HER2/neu Status:** HER2/neu is a member of the transmembraneous HER family and is overexpressed in 15%-20% of tumours, mainly owing to amplification of the HER2/neu gene. This is strongly correlated with aggressive tumour type, down-regulation of HR++ and induced proliferation, with consequent decrease in overall survival (Stickeler, 2011). Breast cancers that overexpress HER2 are affected by abnormal HER2 signaling, and this is associated with increased tumour aggressiveness, high rates of recurrence and increased mortality (Ménard *et al.*, 2001; Brown *et al.*, 2008; Curigliano *et al.*, 2009).

1.1.5 Immunohistochemistry subtypes

It is well-known that the conventional histological classification system is indispensable for the accurate histological diagnosis of breast cancer. However, it does not always provide sufficient information to evaluate the tumours' individual biological characteristics and it is not useful for treatment selection given that tumours with the same histological subtypes can have very different biological trajectories (Yanagawa, 2012). Thus, determining the status of ER and PR receptors, HER2 amplification and Ki-67[1] antigen expression is practical and valuable for estimating the patient prognosis (Raica, 2009).

The St. Gallen International Expert Consensus proposed a new intrinsic biological classification system based on the expression of the ER, PR, HER2 and Ki-67. Approximations of molecular subtypes have been identified using routinely evaluated biological markers, including the presence or absence of HR (HR+/HR-) and excess levels HER2 and/or extra copies of the HER2 gene (HER2+/HER2-). The four main immunohistochemistry (IHC) subtypes are: HR+/HER2-; HR+/HER2+; HR-/HER2+; HR-/HER2-. In some works, these subgroups have been shown to be related with the biological subgroups Luminal A, Luminal B, HER2+ and triple negative, respectively. Figure 1.2 describes IHC subtypes.

[1]Ki-67 is a protein that is strictly associated with cell proliferation.

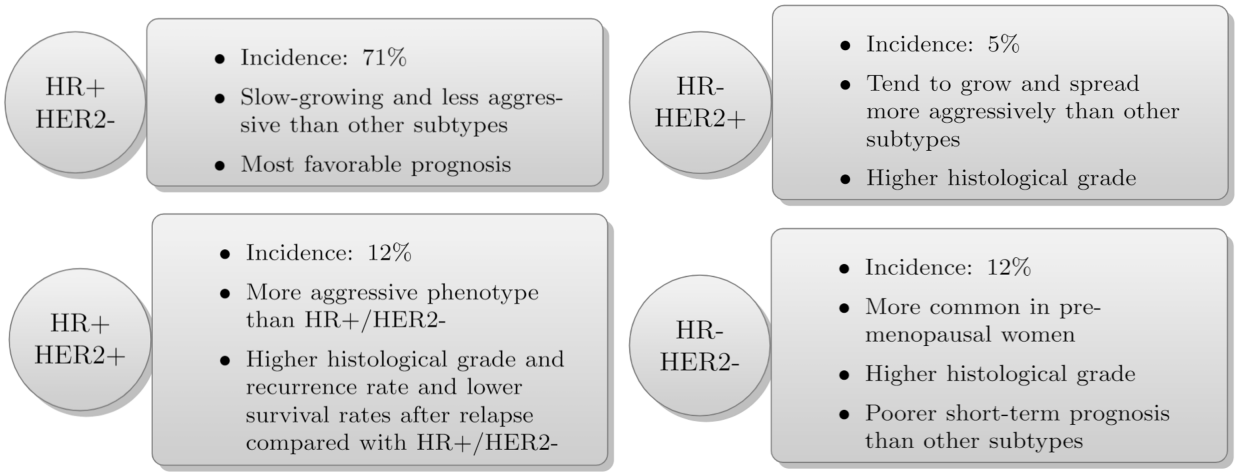

Figure 1.2: Intrinsic subtypes of breast cancer

The most common subtype is HR+/HER2-, representing more than 70% of all cases. It is also the subtype with a better prognosis, being typically a tumour of slow and less aggressive growth, responding better to the first therapeutic lines with anti-hormonal drugs. HER2+ subtypes have less favourable prognosis, and are usually more aggressive than HR+/HER2-. However, due to the development of drugs targeting the HER2 receptor, the prognosis of these patients has considerably improved in the past years. Patients with HR-/HER2- have poorer prognosis comparing with other subtypes as a result of insufficient targeted therapies for these tumours.

1.2 Objectives and outline

In this study we focused on developing dynamic prediction methods of patients with early-stage breast cancer, in order to gain insights regarding the long-term impact of prognostic factors of the disease. This general purpose can be divided into two broad objectives. In the first place, we aimed at evaluating the overall survival and disease-free survival, conditional on the time lived without disease. Secondly, we assessed the long-term prognostic significance of relevant prognostic factors at diagnosis and evaluated how their effects change with time.

This thesis is organized as follows. We begin with the above introduction, reflecting the importance of this project and reasons for its relevance in today's scientific overview, also describing breast cancer chronic disease, essentially for the full understanding of the study. The following chapter will introduce and describe the collected data (Chapter 2). Next, in Chapter 3, we describe the statistical methods used in which models' analyses and inferences are presented, focusing on dynamic prediction approaches. In Chapter 4 we present the results of all analyses implemented. To sum up, the methodology used and the results obtained are discussed in Chapter 5, regarding statistical and epidemiological studies in this research field. A brief conclusion is presented in Chapter 6.

2

Study Design and Description

This study is a non-interventional, cohort, single-center study with retrospective data collection using data from patients diagnosed and/or treated in Instituto Português de Oncologia de Lisboa Francisco Gentil (IPOLFG), a tertiatry cancer center located in Lisbon, Portugal.

Data for this study derived from the Portuguese population-based oncology registry in the National Cancer Registry.

2.1 Study population

Patients potentially eligible for the study were identified from the National Cancer Registry database using the following selection criteria:

- Malignant neoplasm of the breast (International Classification of Diseases for Oncology (ICD) codes C50.0 to C50.9);
- Diagnosis between January 2006 and December 2011;
- Stage of the disease at diagnosis different from IV;
- Female gender;
- Diagnosed and/or treated in IPOLFG.

Data extraction was performed in May 2018 and were identified a total of 5273 patients. Among the patients assessed for eligibility, were excluded those who presented at least one of the following:

- Histological diagnosis other than invasive carcinoma of the breast with a different illness trajectory and/or different standard treatment approach;
- Stage 0 or IV at diagnosis as the main interest was to study patients in early-stage of breast cancer;
- Insufficient information to meet the study objectives (e.g. unknown date of surgery);
- Internal inconsistency data (e.g. patients with recurrence reported prior to initial surgery).

Eligibility criteria led to exclusion of 653 patients, leading to a sample of 4620 individuals. (Figure 2.1).

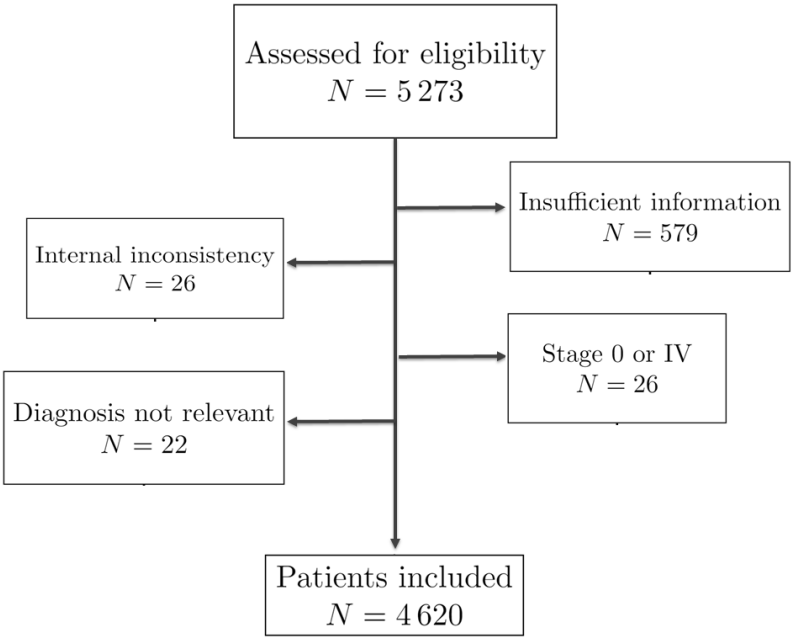

Figure 2.1: Flow chart of patients exclusion

2.2 Variables description and outcomes

Retrieved information from the National Cancer Registry database included the following variables.

- **Date of birth**;
- **Date of diagnosis:** Date of entry of the first biological product for cytological or histological examination into the laboratory;
- **Topographical code:** Topographical code according to ICD-O-3, which describes the anatomical site of origin (or organ system) of the tumour;
- **Morphology:** Morphological code according to ICD-O-3, which describes the cell type of the tumour together with the behaviour (malignant or benign);
- **Tumour grade:** Nottingham grading system (also called the Elston-Ellis modification of the Scarff-Bloom-Richardson grading system) for breast cancer at diagnosis;
- **Breast cancer staging:** TNM classification for breast cancer as per AJCC at diagnosis;
- **cT:** Clinical classification of primary tumour (T);
- **cN:** Clinical classification of regional lymph nodes (N);
- **cM:** Clinical classification of distant metastasis (M);

- **pT:** Pathologic classification of primary tumour (T);
- **pN:** Pathologic classification of regional lymph nodes (N);
- **pM:** Pathologic classification of distant metastasis (M);
- **Estrogen receptor status**: IHC results from breast cancer tissue on ER status at diagnosis, either expressed as a percentage, a score or as positive/negative qualitative result;
- **Progesterone receptor status:** IHC results from breast cancer tissue on PR status at diagnosis, either expressed as a percentage, a score or as positive/negative qualitative result;
- **HER2 status:** HER2 status evaluated at diagnosis by IHC and/or FISH test (Fluorescence In Situ Hybridization). Results expressed as positive (if IHC 3+ showing HER2 protein overexpression or HER2 gene amplification detected by FISH) or negative (if IHC 0 or 1+ or no HER2 gene amplification detected by FISH);
- **Type of event:** Events of disease relapse reported during follow-up. Coded as 9 if none reported; 2 if local relapse; 3 if distant metastasis;
- **Date of recurrence:** Date of diagnosis of disease relapse;
- **Vital status:** Vital status in last follow-up update. Coded as 0 if patient alive and 1 if dead;
- **Date of last follow-up:** Date of last follow-up update or date of death;
- **Date of surgery 1:** Date of first surgery reported;
- **Surgical procedure 1:** Description of the surgical procedures performed in surgery 1;
- **Date of surgery 2:** Date of second surgery reported;
- **Surgical procedure 2:** Description of the surgical procedures performed in surgery 2;
- **Date of surgery 3:** Date of third surgery reported;
- **Surgical procedure 3:** Description of the surgical procedures performed in surgery 3;
- **Radiotherapy 1:** Description of the type of first radiotherapy reported, if any. Classified as: radiotherapy, chemoradiotherapy, adjuvant, neoadjuvant and palliative;
- **Start date of radiotherapy 1:** Start date of first radiotherapy;
- **Chemotherapy 1:** Description of the type of first chemotherapy reported, if any. Classified as: chemoherapy, adjuvant, neoadjuvant and palliative;
- **Start date of chemotherapy 1:** Start date of first chemotherapy;
- **Hormone therapy 1:** If applicable, description of the drug used in the first hormone therapy for breast cancer;
- **Start date of hormonetherapy 1:** Start date of first hormone therapy.

ER and PR status were merged into a single variable, HR, which has been coded as negative if both ER and PR negative and positive if at least one of ER or PR was positive. HR and HER2 receptor status variables were then merged into one single variable representing the IHC subtypes. Variables of interest considered for statistical analyses were coded and are listed below.

- Age at diagnosis: Continuous variable. Measured in years, calculated from date of diagnosis of breast cancer and date of birth;
- Disease stage at diagnosis: Categorical variable with three possible values: 0 = Stage I, 1 = Stage II, and 2 = Stage III;
- Tumour grade: Categorical variable with three possible values: 0 = Well differentiated tumours (low grade), 1 = Moderately differentiated tumours (intermediate grade), and 2 = Poorly differentiated/undifferentiated tumours (high grade);
- Lymph node status: Categorical variable with two possible values: 0 = negative, and 1 = positive if a patient has metastasis in regional lymph nodes at presentation. Classification was based on cN and/or pN;
- IHC group: Categorical variable with four possible values: 0 = HR+/HER2-, 1 = HR+/HER2+, 2 = HR-/HER2+, and 3 = HR-/HER2-.

The two outcomes of interest are:

- Overall survival (OS), defined as the time, in days, from breast cancer diagnosis to death from any cause. It represents the difference between date of last follow-up/death and the date of diagnosis;
- Disease-free survival (DFS), defined as the time, in days, from surgery to recurrence of breast cancer or death from any cause. It represents the difference between date of recurrence or last follow-up/death (if the patient had no recurrence) and the date of surgery;

2.3 Data quality check

In order to validate the data, internal consistency was verified in what regards to date of birth, diagnosis, surgery, recurrence and last follow-up/death. Clinical validation was also performed through a full revision of the disease stage taking into account the clinical and pathological TNM classifications and any neoadjuvant treatments. Moreover, a quality check was conducted in order to assess the completeness and accuracy of the reported information concerning disease recurrence. This was done by reviewing a random sample of the 505 patients that were reported dead without disease recurrence. In this revision inaccuracy was identified in only 3% of the revised patients which was considered acceptable for the study. For the variables with the highest number of missing data, clinical notes from these patients were individually reviewed thus allowing crucial information recovery.

2.4 Ethical aspects

In compliance with the legal and regulatory requirements, this study was reviewed by the Institutional Ethics Committee and Research Council and authorized by the Administration Board of IPOLFG. Compliance with confidentiality requirements was ensured by data anonymization in the database provided for statistical analysis.

3
Statistical Background

3.1 Some insights on survival analysis

Survival analysis is one of the primary statistical methods for analysing data on time to an event, where the dependent variable or response is the time until the occurrence of a well-defined event. In survival analysis, subjects are usually followed over a specified time period and the focus is on the time at which the event of interest occurs. In medical research, the time origin is often the time of recruitment into a clinical trial or study. Although the event of interest can be the death of the patient, recurrence of symptoms or any other particular event, the event of interest is usually death and the time since the time origin until the event of interest is named survival time.

One of the reasons why standard statistical procedures do not apply to this type of data is that survival times are generally not symmetrically distributed. In fact, survival times tend to be positively skewed, therefore it is inadequate to assume that data of this type have a normal distribution. This difficulty could be resolved by first transforming the data to give a more symmetric distribution, for example by taking logarithms. However, a more satisfactory approach is to adopt an alternative distributional model for the original data (Collett, 2015). The most important feature of survival data that renders standard methods inappropriate is the existence of censored observations.

Censoring and truncation

In longitudinal studies exact survival time is only known for those individuals who show the event of interest during the follow-up period. For others (for instance, those who are disease-free at the end of the observation period or those that were lost) all we can say is that they did not show the event of interest during the follow-up period. The survival times of these individuals are called censored observations. An attractive feature of survival analysis is that we are able to include the data contributed by censored observations right up until they are removed from the risk set. There are types of censoring, such as right censoring, left censoring, and interval censoring.

Right censoring happens when the event of interest has not been observed for an individual when the study ends. This may be because the event of interest occurs after the end of the

study, or the patient may have been lost to follow-up. Opposite to right censoring, left censoring happens when the real survival time of an individual is less than the observed time and occurs less frequently than right censoring. A more general type of censoring occurs when individuals are known to have experienced an event within an interval of time. Such interval censoring occurs when patients in a clinical trial or longitudinal study have periodic follow-up and the patients' event time is only known to fall in a certain interval of time. Individuals censored at time t must be representative of all individuals that survived until t. At any time, individuals can not be selectively censored, either because their risk of death is high or low, i.e., censoring is not related to the event of interest. This is known as non informative censoring, which is a necessary condition for the validity of the methods typically used in survival analysis.

A second feature which may be present in some survival studies is truncation. Truncation of survival data occurs when only those individuals whose event time lies within a certain observational window are observed. An individual whose event time is not in that interval is not observed and no information on this subject is available to the investigator. This is in contrast to censoring, where there is at least partial information on each subject. Because we are only aware of individuals with event times in the observational window, the inference for truncated data is restricted to conditional estimation. Truncation can be categorized in left and right truncation. Left truncation occurs when subjects enter a study at a particular age (not necessarily the origin for the event of interest) and are followed from this delayed entry time until the event occurs or until the subject is censored. Right truncation occurs when only individuals who have experienced the event of interest are observable. Generally we deal with right censoring and sometimes left truncation.

Survival, hazard and cumulative hazard functions

In summarising survival data, there are three functions of central interest, namely the *survival function*, the *hazard function*, and the *cumulative hazard function*. These functions are therefore defined in this section.

The actual survival time of an individual, t, can be regarded as the observed value of variable T, that can take any non-negative value. Recall that T is the time until some specified event and this event may be death, the appearance of a tumour, the development of some disease, recurrence of a disease, remission after some treatment, and so forth. The different values that T can take have a *probability distribution*, and therefore we call T the *random variable* representing the survival time. Supposing that this random variable has a probability distribution with underlying *probability density function* $f(t)$, the distribution of T is given by

$$F(t) = P(T < t) = \int_0^t f(u)du, \tag{3.1}$$

and represents the probability that the survival time is less than some value t. This function is also called the *cumulative incidence function*. The basic quantity employed to describe time-to-event phenomena is the survival function, $S(t)$, and is defined as the probability of an individual surviving beyond time t (experiencing the event after time t), and so from Equation (3.1),

$$S(t) = P(T \geq t) = 1 - F(t) \tag{3.2}$$

The *hazard function* measures the instantaneous risk of dying right after time t given the individual is still alive at time t. It also represents the risk of the event of interest occur at time t. More formally, the hazard function is defined as:

$$h(t) = \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \tag{3.3}$$

The function $h(t)$ is also referred to as the hazard rate, the instantaneous death rate, the intensity rate or the force of mortality and has the following properties:

$$h(t) \geq 0; \qquad \int_0^\infty h(t)dt = \infty \tag{3.4}$$

The expression in (3.3) can be written as:

$$\begin{aligned} h(t) &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt | T \geq t)}{dt} \\ &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt)}{dt P(T \geq t)} \\ &= \lim_{dt \to 0} \frac{P(t \leq T < t + dt)}{dt S(t)} \\ &= \frac{f(t)}{S(t)} \end{aligned} \tag{3.5}$$

If now we integrate (3.5) and introduce the condition $S(0) = 1$ (since the event is sure not to have occurred by duration 0):

$$S(t) = \exp\left[- \int_0^t h(u)du \right] \tag{3.6}$$

The integral in curly brackets is called the *cumulative hazard function* and is denoted by:

$$H(t) = \int_0^t h(u)du \tag{3.7}$$

This function measures the risk of occurrence of the event until instant t. According to (3.6):

$$S(t) = \exp[-H(t)]$$

$$H(t) = -\log S(t) \tag{3.8}$$

$H(t)$, can be defined as the cumulative risk of an event occurring by time t. If the event is death, then $H(t)$ summarises the risk of death up to time t, given that death has not occurred before t. It can also be interpreted as the expected number of events that occur in the interval from the time origin to t.

3.2 Non-parametric inference

3.2.1 Kaplan-Meier estimator

Suppose first that we have a single sample of survival times, where none of the observations are censored. The survival function $S(t)$, defined in (3.2), is the probability that an individual survives for a time greater than or equal to t. This function can be estimated by the *empirical survival function*, given by

$$\widehat{S}(t) = \frac{\text{Number of individuals with survival times} \geq t}{\text{Number of individuals in the dataset}} \tag{3.9}$$

However, this method cannot be used when there are censored observations. Kaplan & Meier (1958) proposed a non-parametric estimator of the survival function in the presence of censored observations, also knows as Product-Limit estimator.

Denote $t_{(1)}, \ldots t_{(k)}$ the k ordered times where the *deaths* occurred in a sample of size n ($k \leq n$), d_i the number of *deaths* occurred in $t_{(i)}$ and r_i the number of individuals at risk at $t_{(i)}$. The Kaplan-Meier estimator takes the form:

$$\widehat{S}(t) = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{r_i}\right), \tag{3.10}$$

with $\widehat{S}(t) = 1$ when $0 \leq t < t_{(1)}$. If the largest observation is not censored $\widehat{S}(t) = 0$ for $t \geq t_{(k)}$. However, if the largest recorded observation t^* is censored, then $\widehat{S}(t)$ will never reach 0 and it is considered that the estimate is defined only until that time. The estimate $\widehat{S}(t)$ is a step function with jumps at the event times. The size of these jumps depends not only on the number of events observed at each event time $t_{(i)}$, but also on the pattern of the censored observations prior to $t_{(i)}$. Breslow & Crowley (1974) and Meier (1975) proved that $\widehat{S}(t)$ is a consistent estimator of $S(t)$, under certain conditions of regularity, and is asymptotically normally distributed. One can also be considered as a non-parametric maximum likelihood estimator of $S(t)$. The variance of the Kaplan-Meier estimator is estimated by Greenwood's formula:

$$\widehat{\text{var}}\{\widehat{S}(t)\} = [\widehat{S}(t)^2] \sum_{t_i \leq t} \frac{d_i}{r_i(r_i - d_i)} \tag{3.11}$$

In large samples the Kaplan-Meier estimator evaluated at a given time t, is approximately normally distributed so that a standard $100(1-\alpha)\%$ confidence interval for the survival function at t takes the form

$$\left[\widehat{S}(t) \pm z_{1-\alpha/2} \times \sqrt{\widehat{\text{var}}[\widehat{S}(t)]}\right] \tag{3.12}$$

where z_α is the α quantile of the $N(0,1)$ distribution.

3.2.2 Estimation of percentiles

Since the distribution of survival times tends to be positively skewed, the median is the preferred measure of the location of the distribution. Once the survival function has been estimated, it is straightforward to obtain an estimate of the *median survival time*. This is the time beyond which 50% of the individuals in the population under study are expected to survive, and is given by that value $\chi_{0.50}$ which is such that $S(\chi_{0.50}) = 0.5$. Because the non-parametric estimates of $S(t)$ are step-functions, it will not usually be possible to realise an estimated survival time that makes the survival function exactly equal to 0.5. Instead, the estimated median survival time, $\widehat{\chi}_{0.50}$, is defined to be the smallest observed survival time for which the value of the estimated survival function is less than 0.5. In mathematical terms:

$$\widehat{\chi}_{0.50} = min\{t_i : \widehat{S}(t_i) \leq 0.5\} \tag{3.13}$$

where t_i is the ith ordered *death* time, $i = 1, 2, \ldots, r$. It may also be convenient estimate another percentile of probability p:

$$\widehat{\chi}_p = min\{t_i : \widehat{S}(t_i) \leq 1 - p\} \tag{3.14}$$

3.2.3 Comparison of two groups of survival data

Kaplan-Meier estimator also allows the estimation of survival curves for different groups, in accordance with the categories of each variable. Thus, for each variable, the survival curve is estimated separately for each group, making it possible to assess whether this variable has influence on the survival time. After estimation of survival curves it is important to test whether there are significant differences between them. Based on two samples of m and n individuals from two populations with survival function $S_1(t)$ and $S_2(t)$ respectively, we intend to test the hypothesis:

$$H_0 : S_1(t) = S_2(t) \ vs. \ H_1 : S_1(t) \neq S_2(t)$$

There are a number of methods that can be used to quantify the extent of between-group differences. Non-parametric procedure considered in this study is named log-rank test.

3.2.3.1 Log-rank test

In order to construct the log-rank test, we begin by considering separately each death time in two groups of survival data. Suppose that there are k distinct *death* times, denoted $t_1 < t_2 < \ldots < t_k$, regarding $m + n$ individuals, and that at time t_j, there are d_{1j} individuals in Group 1 and d_{2j} individuals in Group 2 die, for $j = 1, 2, \ldots, k$. Suppose further that there are n_{1j} individuals at risk of death in the first group just before time t_j, and that there are n_{2j} at risk in the second group. Consequently, at time t_j, there are $d_j = d_{1j} + d_{2j}$ *deaths* in total out of $n_j = n_{1j} + n_{2j}$ individuals at risk. The relevant information at each time t_j can be summarised in a 2×2 contingency table (Table 3.1).

Table 3.1: Number of deaths at the jth death time in each of two groups of individuals.

Group	No of deaths at t_j	No of survivors beyond t_j	No of individuals at risk at t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{1j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

If the marginal totals in Table 3.1 are regarded as fixed, and the null hypothesis that survival is independent of group is true, the four entries in this table are solely determined by the value of d_{1j} , the number of deaths at t_j in Group 1. We can therefore regard d_{1j} as a random variable, which can take any value in the range from 0 to the minimum of d_j and n_{1j} . In fact, d_{1j} has a distribution known as the hypergeometric distribution, according to which the probability that the random variable associated with the number of deaths in the first group takes the value d_{1j} is

$$\frac{\binom{d_j}{d_{1j}}\binom{n_j-d_j}{n_{1j}-d_{1j}}}{\binom{n_j}{n_{1j}}} \tag{3.15}$$

The conditional mean of d_{1j} is $e_{ij} = \frac{n_{1j}d_j}{n_j}$, which represents the expected number of deaths in t_j, in Group 1. Conditional variance of d_{1j} is:

$$v_{1j} = \frac{n_{1j}n_{2j}d_j(n_j - d_j)}{n_j^2(n_j - 1)} \tag{3.16}$$

To obtain an overall measure of the deviation of the observed values of d_{1j} in relation to the expected values, we consider:

$$U = \sum_{j=1}^{k}(d_{1j} - e_{1j}) \tag{3.17}$$

where $\sum d_{1j} - \sum_{e_{1j}}$ is the difference between the total number of deaths observed and expected in group 1. This statistic will have zero mean, since $E(d_{1j}) = e_{1j}$. Moreover, since the death times are independent of one another, the variance of U is simply the sum of the variances of the d_{1j}. This method of combining information over a number of 2×2 tables was proposed by Mantel & Haenszel (1959) and is:

$$Q = \frac{U^2}{Var(U)} \tag{3.18}$$

which, under H_0, have asymptotic distribution χ_1^2. The statistic Q summarises the extent to which the observed survival times in the two groups of data deviate from those expected under the null hypothesis of no group differences. The larger the value of this statistic, the greater the evidence against the null hypothesis.

3.2.4 Comparison of three or more groups of survival data

So far, we presented statistical methods to test survival curves for two groups. However, log-tank test can be extended to enable three or more groups of survival data to be compared. Suppose that the survival distribution of g groups of survival data are to be compared, for $g \geq 2$. We then define analogues of the $U-$statistics for comparing the observed number of deaths in groups $1, 2, \ldots, g-1$ with their expected values. In as obvious extension of the notation used before, we obtain

$$U = \sum_{j=1}^{k} (d_{rj} - e_{rj}), \tag{3.19}$$

with $r = 1, 2, \ldots, g-1$ and $e_{rj} = \frac{n_{rj} d_j}{n_j}$.

The $(r, r^{'})$ element of the covariance matrix is given by:

$$V_{rr'} = \sum_{j=1}^{k} \frac{n_{rj} d_j (n_j - d_j)}{n_j (n_k - 1)} \left(\delta_{rr'} - \frac{n_{r'j}}{n_j} \right) \tag{3.20}$$

for r, $r^{'} = 1, 2, \ldots, g-1$ and $\delta_{rr'}$ is

$$\delta_{rr'} = \begin{cases} 1, & \text{if } r = r^{'}, \\ 0, & \text{otherwise.} \end{cases} \tag{3.21}$$

These terms are then assembled in the form of a variance-covariance matrix $\boldsymbol{V}$, which is a symmetric matrix that has the variances of the U down the diagonal, and covariance terms in the off-diagonals. For example, in the comparison of three groups of survival data, this matrix would be given by

$$V = \begin{pmatrix} V_{11} & V_{12} \\ V_{12} & V_{22} \end{pmatrix} \tag{3.22}$$

where V_{11} and V_{22} are the variances of U_1 and U_2, respectively, and V_{12} is their covariance.

Finally, in order to test the null hypothesis on no group differences, we make use of the result that the test statistics $U^{'} V^{-1} U$ has a chi-squared distribution with $(g-1)$ degrees of freedom, when the null hypothesis is true.

3.3 Cox Regression model

One downside of Kaplan-Meier estimator is that it is only capable of dealing with one explanatory variable at a time. When we want to consider the effect of several explanatory variables simultaneously, a regression model is the right approach for survival estimation. Cox regression model (Cox, 1972) is perhaps the most widely used regression model in medical research.

3.3.1 Formulation of the model

A Cox proportional hazards regression model is, as the name suggests, defined through the hazard function, which is required to be proportional between all individuals. This is done by assuming that the hazard consists of some arbitrary non-parametric function, usually referred to as the *baseline hazard*, multiplied by a constant that depends on a linear predictor for each individual. More concretely, the hazard of an individual is expressed as

$$h(t|\boldsymbol{x}) = h_0(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta}) \tag{3.23}$$

where $h_0(t)$ is the baseline hazard, $\boldsymbol{x}$ is the vector of the covariates of an individual and $\boldsymbol{\beta}$ is the vector of coefficients of the explanatory variables in the model.

The baseline hazard function describes how the risk of death changes over time at baseline levels of the covariate, and the exponential expression describes how the hazard varies in response to the explanatory variables. Cox model is often called a proportional hazards model because if we look at two individuals with covariate vectors $\boldsymbol{x_1}$ and $\boldsymbol{x_2}$, the ratio of their hazard rates is

$$\frac{h(t|\boldsymbol{x_1})}{h(t|\boldsymbol{x_2})} = \frac{h_0(t)\exp(x_{11}\beta_1 + \ldots + x_{1p}\beta_p)}{h_0(t)\exp(x_{21}\beta_1 + \ldots + x_{2p}\beta_p)} = \exp\left(\sum_{j=1}^{p}(x_{1j} - x_{2j})\beta_j\right) \tag{3.24}$$

which is a constant. Therefore, the model presupposes proportional hazards assuming that the effect of covariates does not change over time. The quantity (3.24) is called the relative risk (hazard ratio) of an individual with risk factor $\boldsymbol{x_1}$ having the event as compared to an individual with risk factor $\boldsymbol{x_2}$. In addition, the exponential form of relative risk ensures that the risk estimates are non-negative, which makes the Cox model very appealing. As mentioned above we do assume that all individuals share a common baseline hazard $h_0(t)$, but we do not make any assumption regarding the nature of the hazard function itself. As so, the Cox proportional hazards model is a semi-parametric model.

Adequacy of a fitted model needs to be assessed after a model has been estimated. Diagnostic procedures for model checking are known as essential parts of a modelling process and a residuals analysis should be performed. In survival analysis, especially when we build a Cox's proportional hazards model, few types of residuals can be considered for different purposes. Several useful diagnostic tools which are based on residuals are:

- Cox-Snell residuals: to evaluate the overall fit of the final model;
- Schoenfeld residuals: for checking the proportional hazards assumption for a covariate;

- Martingale residuals: to determine the functional form that should be used for a given covariate;
- Deviance residuals: for detection of poorly predicted observations.

3.3.2 Parameters interpretation

In fact, usually, $\exp(\beta_j)$ is preferred over β_j , since $\exp(\beta_j)$ provides a straightforward interpretation regarding the risk of death. $\exp(\beta_j)$ represents the relative risk of occurrence of the event of interest for two individuals that differ in one unit in the values of the covariate x_j, with the values of the remaining covariates being equal. Consider a binary covariate defined by $x = 0$ if the individual belongs to group 1 and $x = 1$ if the individual belongs to group 2. When the individual belongs to group 1 then $h(t|x = 0) = h_0(t)$ and when the individual belongs to group 2 then $h(t|x = 1) = h_0(t)\exp(\beta)$.

- If $\beta < 0 \Leftrightarrow \exp(\beta) < 1$, patients in group 2 have better prognosis than patients in group 1;
- If $\beta > 0 \Leftrightarrow \exp(\beta) > 1$, patients in group 1 have better prognosis than patients in group 2;
- If $\beta = 0 \Leftrightarrow \exp(\beta) = 1$, patients in groups 1 and 2 have a similar prognosis.

In the case of a numeric covariate:

$$\exp(\beta) = \frac{h(t|x = j + 1)}{h(t|x = j)} \tag{3.25}$$

For instance, if x corresponds to the age of a patient, $\exp(\beta)$ represents the risk of death of a patient with a certain age compared with a patient one year younger. The hazard ratio for a patient aged 50 relative to one aged 49 is the same as that for an individual aged 80 relative to one aged 79. Therefore, the hazard ratio does not depend on the actual value of the covariate.

3.3.3 Partial likelihood function

Due to the semi-parametric nature of the hazard specification in the Cox regression model, it is impossible to use ordinary likelihood methods. Instead one has to resort to a partial likelihood for estimation and inference. Inference on the vector of unknown parameters, $\boldsymbol{\beta}$, is based on the partial likelihood function (Cox, 1975).

Assuming that there are n individuals in the study and it was observed k different lifetimes $t_{(1)} < \ldots < t_{(k)}, k < n$. The set of individuals who are at risk at time $t_{(i)}$ will be denoted by $R(t_{(i)})$, so that $R(t_{(i)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(i)}$ and is called the risk set.

The likelihood function for the proportional hazards model is given by

$$L(\beta) = \prod_{i=1}^{k} \left\{ \frac{\exp(\boldsymbol{x}_{(i)}^{\top}\boldsymbol{\beta})}{\sum_{l \in R(t_{(i)})} \exp(\boldsymbol{x}_{\boldsymbol{l}}^{\top}\boldsymbol{\beta})} \right\} \tag{3.26}$$

where $\boldsymbol{x}_{(i)}$ is the vector of covariates associated with the individual who dies at the ith ordered death time, $t_{(i)}$. The summation in the denominator of this likelihood function is the sum of the values of $\exp(x^{\top}\boldsymbol{\beta})$ over all individuals who are at risk at time $t_{(i)}$. The product is taken over the individuals for whom death times have been recorded. Individuals for whom the survival times are censored do not contribute to the numerator of the likelihood function but they do enter into the summation over certain risk sets.

Moreover, the likelihood function depends only on the ranking of the death times. This likelihood function can be seen as a partial likelihood since this function does not depend on the baseline hazard function and allows inference on β, without any restriction regarding the form of $h_0(.)$. At each time t, only the information about the individuals at risk is considered. This formulation is similar to the non-parametric methods but allows an estimation of the effect of the covariates on the survival time. Under certain regularity conditions, the maximum partial likelihood estimator of β is consistent, normally asymptotic with mean value β and covariance matrix given by $I(\beta)^{-1}$, where $I(\beta)$ is the Fisher information matrix:

$$-\left[E\left(\frac{\partial^2 \log L}{\partial\beta_j \partial\beta_k}\right)\right]_{p\times p} \tag{3.27}$$

In case of simultaneous deaths or when the data is not recorded properly yielding equal values, the function is not appropriate. In this situation, for the n individuals in the study suppose that the distinct death times were observed $t_1 < t_2 < \ldots < t_k$. Denote d_i as the number of deaths occurred at time t_i and $\boldsymbol{x}_{ij}$ the vector of variables associated to individual j, that dies in t_i, $j = 1, \ldots, d_i, i = 1, \ldots, k$. If d_i is small, compared with the number of individuals in the risk set R_i, then the partial likelihood function can be approximated by the function, proposed by Peto & Peto (1972) and Breslow (1974).

$$L(\beta) = \prod_{i=1}^{k} \frac{\exp(\boldsymbol{s}_i^{\top}\boldsymbol{\beta})}{[\sum_{l \in R_i} \exp(\boldsymbol{x}_l^{\top}\boldsymbol{\beta})]^{d_i}} \tag{3.28}$$

where $\boldsymbol{s}_i = \sum_{j=1}^{d_i} \boldsymbol{x}_{ij}$, for $i = 1, \ldots, k$. This is the likelihood usually implemented in software packages. If the observations do not have ties, the function (3.28) reduces to the partial likelihood (3.26) (Collett, 2015).

3.3.4 Confidence intervals and hypothesis tests

Since the regression parameter estimators are asymptotically distributed according to a Gaussian distribution, it is easy to calculate asymptotic normal intervals and to use Wald tests.

The 95% confidence interval of the regression parameter β_j is: $[\hat{\beta}_j \pm 1.96 \times \hat{\sigma}_j]$, where $\hat{\beta}_j$ is the estimator of the parameter β_j and $\hat{\sigma}_j$ is the standard deviation of $\hat{\beta}_j$. In general, it is more interesting to provide the confidence interval of the hazard ratio. Since the hazard ratio is $\exp(\beta_j)$, its 95% confidence interval can be calculated by:

$$[\exp(\hat{\beta}_j - 1.96 \times \hat{\sigma}_j); \exp(\hat{\beta}_j + 1.96 \times \hat{\sigma}_j)] \tag{3.29}$$

As previously discussed, β_j represents the effect of the covariate x_j on the survival of the individual. To evaluate the existence of evidence that the covariate significantly influences the survival time, one can test:

$$H_0 : \beta_j = 0 \ vs. \ H_1 : \beta_j \neq 0$$

using the Wald test, where the test statistic $\hat{\beta}_j^2/\text{var}(\hat{\beta}_j)$ has, under H_0, an asymptotic χ_1^2 distribution. Similarly, one can use the test statistic $\hat{\beta}_j/\sqrt{\text{var}(\hat{\beta}_j)}$ which has, under H_0, an asymptotic $N(0,1)$ distribution. The null hypothesis tested is that the covariate x_j does not have a significant influence, in the presence of the remaining variables, in the survival. However, the estimates $\hat{\beta}$ are not all independent which difficult the interpretation of the results. Therefore, it is preferred to compare alternative models.

3.3.5 Estimation of cumulative hazards and survival probabilities

For various reasons, we may be interested in the estimated cumulative hazard under the assumption of the Cox proportional hazard model for a given covariate vector. One approach here is to use the estimator

$$\widehat{H}(t|\boldsymbol{x}) = \widehat{H}_0(t) \exp(\boldsymbol{x}^\top \hat{\boldsymbol{\beta}}) \tag{3.30}$$

where $\widehat{H}_0(t)$ is the Breslow estimator

$$\widehat{H}_0(t) = \sum_{t_i \leq t} \frac{d_i}{\sum_{le R_i} \exp(\boldsymbol{x}_l^\top \widehat{\boldsymbol{\beta}})} \tag{3.31}$$

We can also obtain an estimator of the survival function by transforming the cumulative hazard estimator, i.e

$$\widehat{S}(t|\boldsymbol{x}) = \exp(-\widehat{H}(t|\boldsymbol{x})) \tag{3.32}$$

3.3.6 Variables selection procedures

In a regression analysis we intend to construct a model that fits our data and identifies the explanatory variables significantly associated with the outcome of interest. It is important to note that, in survival studies, the contribution of clinicians is crucial for building models where clinically relevant explanatory variables that have not revealed statistical significance can be included. We then want to evaluate whether each explanatory variable has significance influence in the survival of an individual. A classical variable selection method is the stepwise regression using p-value as a criterion for inclusion or deletion of covariates. It combines forward selection and backward elimination methods, allowing variables to be added or dropped at various steps according to different pre-specified p-values for entry to or stay in the model (Klein, 2014). In this study we used two variable selection procedures: a forward and backward selection.

Forward selection, which involves starting with no variables in the model, testing the addition of each variable using a chosen model fit criterion - in our case the Wald test, adding the variable (if any) whose inclusion gives the most statistically significant improvement of the fit - in our case the most significant p-value, and repeating this process until none improves the model to a statistically significant extent.

Backward elimination, which involves starting with all candidate variables, testing the deletion of each variable using a chosen model fit criterion, deleting the variable (if any) whose loss gives the most statistically insignificant deterioration of the model fit, and repeating this process until no further variables can be deleted without a statistically significant loss of fit.

3.3.7 Check the proportional hazards assumption

As we have seen, Cox regression model relies on a fundamental assumption, the proportionality of the hazards, implying that the factors investigated have a constant impact on the hazard - or risk - over time, i.e., the model assumes that each covariate has a multiplicative effect in the hazard function that is constant over time. Violation of this assumption can result in misleading effect estimates and significant effect in the early (or late) follow-up period may be missed (Bellera *et al.*, 2010). Checking the proportionality of the hazards should be an integral part of a survival analysis by a Cox model.

Many approaches for assessing the proportional hazards assumption are available, including both graphical methods and statistical testes. Although graphical approaches involve a moderate degree of subjectivity in interpretation, they present a visual form of screening for non-proportionality which can provide insight into the temporality and the extent of non-proportionality that is otherwise difficult to obtain using statistical methods. Statistical tests typically screen for the lack of fit of a Cox model. Specifically, Gramsch and Therneau (1994) have shown that many of these statistical tests are essentially tests for a non-zero slope in generalized linear regression models of the Schoenfeld residuals (Schoenfeld, 1982) as a function of event time. Correlation tests of Schoenfeld residuals and event time (or log of the event time) or Kaplan-Meier survival curve estimates are among the most frequently used approaches for assessing the proportional hazards assumption.

3.3.7.1 Schoenfeld residuals

This type of residuals was proposed by Schoenfeld (1982). For the ith individual, the Schoenfeld residual corresponding to the covariate x_j, $j = 1, \ldots, p$, is expressed by:

$$r_{ji} = \delta_i\{x_{ji} - \hat{a}_{ji}\} \tag{3.33}$$

where $\delta_i = 1$ if t_i is a non-censored observation and $\delta_i = 0$ otherwise and

$$\hat{a}_{ji} = \frac{\sum_{l \in R_i} x_{jl} \exp(\boldsymbol{x}_l^\top \hat{\boldsymbol{\beta}})}{\sum_{l \in R_i} \exp(\boldsymbol{x}_l^\top \hat{\boldsymbol{\beta}})} \tag{3.34}$$

For an individual whose survival time was censored, residuals are always zero, usually indicated as missing values to distinguish them form residuals genuinely identical to zero. For an individual whose death was observed at t_i, the residual is the difference between x_j, corresponding to the ith individual, and a weighted average of the values of that variable for all individuals at risk at t_i (Collett, 2015).

Grambsch & Therneau (1994) proposed a version of these residuals which is more effective in detecting departures from the assumed model, named scaled Shoenfeld residuals. Let $\boldsymbol{r}_i = (r_{1i}, r_{2i}, \ldots, r_{pi})^\top$ be the vector of Schoenfeld residuals associated to the ith individual. The scaled Schoenfeled residuals, r^*_{ji} are expressed by:

$$\boldsymbol{r_i^*} = k \times \text{var}(\hat{\boldsymbol{\beta}})\boldsymbol{r_i} \tag{3.35}$$

where k is the number of observed deaths among n individuals and $\text{var}(\hat{\beta})$ is the covariance matrix of the parameter estimates in the fitted Cox regression model. Plotting the scaled Schoenfeld residuals against the survival times allows to verify if the residuals are equally distributed over time, check the adequacy of the proportional hazards model and therefore see how the effect of a covariate may change over time. Unusual patterns indicate that the proportional hazards model is inadequate. Besides the visual analysis, it is possible to test the existence of linear correlation between the time and the residuals. Under the null hypothesis, of correlation coefficient equal to zero, the test statistic has a χ^2_1 distribution. If the null hypothesis is not rejected, the assumption of proportionality of the hazards is sustained. The test for each covariate is based on a regression:

$$\beta_k(t) = \beta_k + \theta_k U_k(t), k = 1, \ldots, p \tag{3.36}$$

where θ_k is the variation in time parameter. The null hypothesis is that $\theta_k = 0$.

3.3.7.2 Graphical methods

For time-fixed variables that have a small number of levels, a simple graphical test of the assumption can be made by looking at the survival curves. If proportional hazards hold, then the log survival curves should steadily drifts apart. As so, after obtain the Kaplan-Meier estimate of the survival function for each group of individuals we should plot $\log[-\log \hat{S}_m(t)]$ as a function of the log survival time, for $m = 1, \ldots, M$ over different (combinations of the) M categories of variables being investigated. If the hazards are proportional, the stratum specific log-minus-log plots should exhibit constant differences, that is be approximately parallel. These visual methods are simple to implement but have limitations. When the covariate has more than two levels, Kaplan-Meier plots are not useful for discerning non-proportionality because the graphs become to cluttered (Therneau, 2000). Similarly, although the proportional hazards assumption may not be violated, the log-minus-log curves are rarely perfectly parallel in practice, and tend to become sparse at longer time points, and thus less precise. It is not possible to quantify how close to parallel is close enough, and thus how proportional the hazards are. The decision to accept the proportional hazards hypothesis often depends on whether these curves cross each other. As a result, the decision to accept the proportional hazards hypothesis can be subjective and conservative (Schemper, 1992), since one must have strong evidence (crossing lines) to conclude that the proportional hazards assumption is violated. Thus, when the covariate has many levels or is continuous, the Kaplan-Meier plot is not useful for discerning either the fact or the pattern of non-proportionality hazards (Therneau & Grambsch, 2000).

3.3.8 Strategies for non-proportional hazards

When the Schoenfeld residual plot or other diagnostic technique gives strong evidence of non-proportionality for one or more covariates, numerous approaches are possible. Several are particularly simple and can be done in the context of the Cox model itself, using available software. For more details see Therneau & Grambsch (2000).

1. Stratification: covariates with non-proportional effects may be incorporated into the model as stratification factors rather than regressors;

2. Partition the time axis: the proportional hazards assumption may hold at least approximately over short time periods;

3. Model non-proportionality by time-dependent covariates;

4. Use a different model: an accelerated failure time or additive hazards model might be more appropriate for the data.

All these techniques for dealing with non-proportional hazards are very well documented. However, our primary interest is not to explore these strategies but to evaluate how the effects of covariates may vary over time, and therefore methods of dynamic prediction are investigated.

3.4 Dynamic prediction

Many prediction models have been developed in medicine with the aim of providing predictions from diagnosis or the start of treatment for patients with a certain disease. It is unquestionable that these models are essential to inform patients about their prognosis and to guide clinicians in making treatment decisions. However, the information given by these models is not enough as it does not reflect how prognosis changes over time, and it can only be regarded as a "static" prediction. As so, predicting the risk of an event based on individual information has become more important, especially in the monitoring, screening and management of chronic diseases. Obtaining prediction probabilities using not only baseline information, but also at later points in time is called "dynamic prediction". Dynamic prediction can be more formally defined as the making of a prediction at a certain moment in time, given all the history of events and covariates up until that moment. More precisely, we want to continuously make predictions of an individual surviving a given period ahead in time from a certain time point, i.e, we want to be able to predict for instance 2 year survival for a patient, not only at the time of diagnosis, but at several points during the follow up of a patient.

In short the idea behind dynamic prediction is, for a pre-specified time point (usually denoted by s), to construct a dataset consisting of only subjects at risk at s, i.e, those still alive and under follow-up. After selection of these subjects, a prediction window w should be fixed and right-censored imposed at time $s + w$ (Figure 3.1).

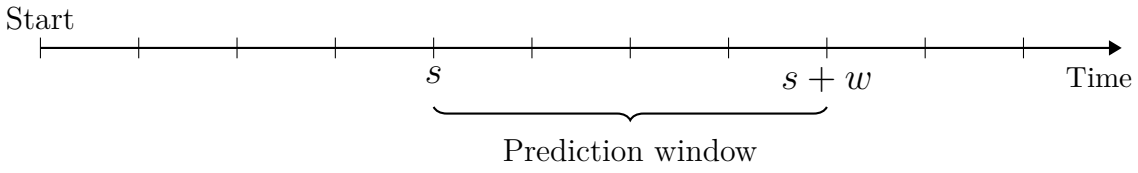

Figure 3.1: Mechanism to compute dynamic prediction methods

It is of great interest for clinicians to have an accurate prognosis tool at their disposal that will inform them about the future prospect of a patient in order to optimize medical care, adapt medical decisions (for instance, by changing the treatment and the frequency of the follow-up visits) and carefully monitor the disease. Likewise, from the statistical analysis viewpoint, the challenge is to utilize a technique capable of updating estimates of survival probabilities for a new patient as additional information is recorded. Conditional survival is the simplest form of a dynamic prediction. A more complex approach is landmarking method. With these models one not only can incorporate time-dependent information into risk prediction but also efficiently make predictions at a series of predetermined time points.

3.4.1 Conditional survival

Conditional survival (CS) is based on the concept of conditional probability and accounts for the fact that hazard rates can change over time. Since survival probability changes for patients who survive a given period of time, prognosis is more accurately described when using conditional survival measure. As so, for patients surviving past a given period, subsequent prognosis can be quite different from prognosis at the time of diagnosis. Such a patient, still alive after s years, would be much more interested in having information on the conditional probability of surviving further w years, given that she/he already survived s numbers of years. It can also be interpreted as the probability of surviving t years (with $t = s + w$) given that it has already survived s years. CS can therefore be a more accurate measure of survival probability for many surviving cancer patients.

In general, CS is the probability of surviving t years, given that the person has already survived s years (Hieke *et al.*, 2015), or $CS(t|s)$, and can be expressed as:

$$CS(t|s) = \frac{S(t)}{S(s)} \tag{3.37}$$

by using the definition of a conditional probability. Thus, for example, $CS(7|2)$ denotes the probability of surviving 7 years, given that the patient is still alive at $s = 2$ years (or the probability of surviving further 5 years given that it is alive at $s = 2$ years). Usually, s is called the prediction time or, more precisely, the time at which the prediction is made. Of course, CS can also be more specifically determined by using additional information on the patient's baseline characteristics. When a survival curve has a changing hazard rate over time, this will be reflected as a change in CS as more time elapses from the time of diagnosis. In principle, CS can be calculated considering usual Kaplan-Meier estimates $\widehat{S}(t)$ and is given by:

$$\widehat{CS}(t|s) = \frac{\widehat{S}(t)}{\widehat{S}(s)} \tag{3.38}$$

Thus, for instance, if we want to estimate the 5-year CS for a patient who has already survived 2 years from diagnosis, $CS(7|2)$, we simply divide the Kaplan-Meier estimator at $t = 7$ by the Kaplan-Meier estimator at $t = 2$. This leads to the same estimates as an approach in which we compute CS probabilities by restricting ourselves to all patients who are alive at s years and not censored before s and calculate the Kaplan-Meier estimator, so-called the conditional Kaplan-Meier estimator, for each of these restricted samples with s years, as a new time origins. Both approaches to estimate CS probabilities, $\widehat{CS}(t|s)$, provide identical results. Independently of the underlying approach used, CS can additionally be estimated in strata defined, for instance, by baseline patient and tumour characteristics.

3.4.2 Landmark models: a novel approach

The method of landmarking was first introduced by Anderson *et al.* (1983) as a way to properly handling the time-dependent covariate "tumour response" in survival models. The practice was to take "tumour response" as a fixed covariate in the Cox model. Since it takes time before "tumour response" can be assessed, this creates a substantial immortal time bias in favour of "tumour response." The remedy proposed in the paper is to take a fixed time point (t_{LM}), define "tumour response" as "response before t_{LM}," and use that in a Cox model for survival after t_{LM}. That approach circumvents the computational complications of fitting a time-dependent covariate.

As mentioned before, Cox regression model has been the most widely used in survival analysis. A key underlying assumption of this model is that the hazards are assumed to be proportional between individuals, or alternatively that the effects of the covariates are assumed to be constant in time. However, this assumption can be violated in some cases and the Cox model could be extended allowing time varying effects. One such extension of the Cox regression model is known as landmarking. The idea to use landmarking for dynamic predictions involves considering Cox regression models that are local in time. One considers the sequence of these local models, where each individual model belongs to a subset of the follow-up range. This sequence is termed by van Houwelingen & Putter (2011) a *sliding landmark model*. The purpose of landmarking is to create models that are better suited to make dynamic survival predictions than the Cox model when the proportional hazards assumption fails to hold. van Houwelingen (2007) was the first to suggest using landmarking for dynamic prediction. In this work, landmarking plays an essential role for two reasons: i) it keeps the models as transparent as possible; ii) it leads to robust predictions that are not sensitive to unchecked assumptions. Applying the concept explored in Figure 3.1 to the landmark concept, the general idea is to construct, for a landmark time point (t_{LM}), a landmark dataset by imposing left-truncation at t_{LM} and right-censoring at $t_{hor} = t_{LM} + w$. A Cox model is then applied to link covariates with prediction of survival at w time units after t_{LM}.

3.4.2.1 Robustness of Cox regression

We will go over some theoretical results that are the underpinnings of the landmarking technique for computing dynamic survival predictions in settings with time-varying effects. These results are taken from van Houwelingen (2007). Detailed derivations can be found in Appendix A.

The model given in equation 3.23 can be violated for many reasons, either because the effect of each component of $\boldsymbol{x}$ might not be linear or the effect of the covariates might vary with time. The latter model is often considered when the follow up is relatively long and when there are biological reasons which make the effect change over time. The time-varying effect model is denoted by

$$h(t|x) = h_0(t) \exp(\boldsymbol{x}^\top \boldsymbol{\beta}(t)) \tag{3.39}$$

i.e, there is a time-dependent effect of the covariates. Once this model has been fitted, the

survival function can be estimated through the cumulative hazard:

$$H(t|\boldsymbol{x}) \approx H_0(t)\exp(\boldsymbol{x}^\top \bar{\boldsymbol{\beta}}(t)) \tag{3.40}$$

where $\bar{\boldsymbol{\beta}}(t)$ is defined by

$$\bar{\boldsymbol{\beta}}(t) = \frac{\int_0^t h_0(s)\boldsymbol{\beta}(s)ds}{\int_0^t h_0(s)ds} \tag{3.41}$$

Derivation of (3.40) can be found in Appendix A. By definition, $\boldsymbol{\beta}(t)$ gives the instantaneous effect of covariate at time t conditioned on being at risk at t. In line with the landmarking idea explained above, this weighted average of $\boldsymbol{\beta}(s)$ over the interval $[t_{LM}; t_{hor}]$ is needed to obtain a reasonable proportional hazards model instead of the average over the whole follow-up range $[0; t_{hor}]$.

As we have seen, the main clinical interest is in survival up to a certain horizon t_{hor}, $S(t_{hor}|\boldsymbol{x})$. Estimates can be obtained through the model in (3.23), even if there is any reason to assume that the proportional hazard assumption is violated. This approximation works very well if $\boldsymbol{\beta}(t)$ does not vary too much over time, the covariate effects are not excessively large and if the follow up is not too long. Therefore, when we apply the simple Cox model (3.23) to the data up to t_{hor} in a situation where the true model is given by equation (3.39), the limiting value is approximately given by:

$$\tilde{\boldsymbol{\beta}}_{Cox} = \frac{\int_0^{t_{hor}} S(t)C(t)h(t)\text{var}(\boldsymbol{X}|T=t)\boldsymbol{\beta}(t)dt}{\int_0^{t_{hor}} S(t)C(t)h(t)\text{var}(\boldsymbol{X}|T=t)dt} \tag{3.42}$$

Here, $S(t)$, $C(t)$ and $h(t)$ are the marginal survival, censoring and hazard function, respectively and $\text{var}(\boldsymbol{X}|T=t)$ is the weighted covariance matrix of $\boldsymbol{X}$ in the risk set at time t. The approximation is valid under the condition that at each t, $\tilde{\boldsymbol{\beta}}_{Cox}$ does not differ to much from the true $\boldsymbol{\beta}(t)$, which is equivalent to requiring that $\boldsymbol{\beta}(t)$ does not vary too much over the interval $[0; t_{hor}]$. A simplification of the approximation is obtained if it can be assumed that $\text{var}(\boldsymbol{X}|T=t)$ is constant over the interval. This will be true if the effects of the covariates are minor and/or t_{hor} is not too far away. Under those conditions:

$$\tilde{\boldsymbol{\beta}}_{Cox} \approx \frac{\int_0^{t_{hor}} S(t)C(t)h(t)\boldsymbol{\beta}(t)dt}{\int_0^{t_{hor}} S(t)C(t)h(t)dt} \tag{3.43}$$

If the t_{hor} is small indeed, $C(t) \approx 1, S(t) \approx 1$ and $h(t) \propto h_0(t)$. The implication is that:

$$\tilde{\boldsymbol{\beta}}_{Cox} \approx \bar{\boldsymbol{\beta}}(t_{hor}) \tag{3.44}$$

Derivation of (3.44) can be found in Appendix A. Finally, it can be shown that under the

same conditions, the Breslow estimator of the baseline hazard in the Cox model converges to

$$h_{Cox,0}(t) \approx h_0(t) \exp(E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}(t) - \bar{\boldsymbol{\beta}}(t_{hor}))) \tag{3.45}$$

The corresponding cumulative hazard is:

$$H_{Cox,0}(t_{hor}) = \int_0^{t_{hor}} h_{Cox,0}(t)dt \approx H_0(t_{hor}) \tag{3.46}$$

and hence

$$H_{Cox}(t_{hor}|\boldsymbol{x}) \approx H(t_{hor}|\boldsymbol{x}) \tag{3.47}$$

Derivation of (3.47) can be found in Appendix A. So, the Cox model gives (approximately) correct predictions of surviving up to t_{hor} even though there might truly be a time-varying effect of the covariates, provided that $S(t)$ and $C(t)$ stay close to 1 and $\boldsymbol{\beta}(t)$ does not vary too much.

3.4.2.2 Sliding landmark

We have seen so far that the estimates from a Cox model might give a reasonable prediction of survival up to some t_{hor}, even if the assumption of proportional hazards fail. However, in this case, the Cox model may not be a good choice when it comes to making dynamic predictions as the Cox model does not capture dynamic differences. Instead we may assume that we are in the misspecification situation presented before, and rather use weighted averages of $\boldsymbol{\beta}(t)$ computed over the intervals $[t_{LM}; t_{hor}]$, in place of the average over the whole follow-up range $[0; t_{hor}]$. van Houwelingen & Putter (2011) call this a *sliding landmark model*.

In general terms, the procedure to obtain dynamic predictions using landmarking can be done by selecting all the individuals at risk at t_{LM}, and using the information available at that specific time to make a prediction. For each t_{LM} a prediction window of width w is defined and predictions are made from time point t_{LM} for a fixed horizon $t_{hor} = t_{LM} + w$. Such sectioned datasets are called *landmark datasets*.

Define $\boldsymbol{x}$ to be the vector of time-fixed and time-dependent covariates (for the time-dependent covariates the current value at t_{LM} should be taken). In order to obtain such a prediction the sliding landmark model is defined as the simple Cox model:

$$h(t|\boldsymbol{x}, t_{LM}) = h_0(t|t_{LM}) \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_{LM}) \tag{3.48}$$

for $t_{LM} \leq t \leq t_{hor}$. This model applies for all individuals at risk at t_{LM} and ignores any event after t_{hor}.

As discussed previously, the motivation for the sliding landmark model is rooted in the problem of giving dynamic survival predictions for a given individual. By this we mean predicting the probability of an individual surviving, say 5 years from some point in time given that the

individual has survived up to then. After having obtained estimates of the regression coefficient ($\hat{\boldsymbol{\beta}}_{LM}$) and the cumulative baseline hazard ($\hat{H}_0(t_{hor}|t_{LM})$) the estimate of the corresponding conditional cumulative hazard for an individual with covariate vector $\boldsymbol{x}_0$ is then:

$$\widehat{H}(t_{hor}|t_{LM}, \boldsymbol{x_0}) = \exp(\boldsymbol{x_0}^\top \hat{\boldsymbol{\beta}}_{LM})\widehat{H}_0(t_{hor}|t_{LM}). \tag{3.49}$$

Therefore the estimate of the conditional survival function used to predict survival for an individual with covariate vector $\boldsymbol{x}_0$ is:

$$\widehat{S}(t_{hor}|t_{LM}, \boldsymbol{x_0}) = \exp(-\widehat{H}(t_{hor}|t_{LM}, \boldsymbol{x_0})) \tag{3.50}$$

It should be stressed that it is not claimed that the proportional hazards model is correct for all $t_{LM} \leq t \leq t_{LM} + w$. The only claim is that it is a very convenient and useful way to obtain a dynamic prediction without having to fit a model with complicated time-varying effects.

3.4.2.3 Landmark supermodels

The approach sketched so far requires a separate Cox model to be fitted at each time point t_{LM} for which a prediction is required. Notwithstanding, this is not very practical and hard to communicate to clinical users. van Houwelingen (2007) presents an approach to obtain a prediction model that can be applied over a range of prediction times and is based on the following construction of a "super prediction dataset":

- Fix the prediction window w based on clinical knowledge;
- Select a set of uniformly spaced landmark prediction time points $\{s_1, \ldots, s_K\}$ based on clinical knowledge. Note that van Houwelingen & Putter (2011) suggest a grid between 20 and 100 points;
- Create a prediction dataset for each landmark time point $t_{LM} = s_k$ by left truncation and right administrative censoring at end of the prediction window $(s + w)$;
- Stack all stratified data frames vertically into a single super prediction dataset. This dataset is used to fit a *landmark supermodel*. Note that passing from one stratum to the next one corresponds to sliding the window over the range of time points.

More specifically, the general idea of the landmark supermodels is to select not just one, but several landmark time points $\{s_1, \ldots, s_K\}$. For each of these, a landmark dataset is created, as described above, by imposing left-truncation and right-censoring. The k datasets are then stacked into a super landmark dataset. This is similar to longitudinal survival data, where a subject can contribute with several observations. As outlined above, a selection of the set of prediction time points implicitly defines a weighting of the prediction time points in the model to be developed.

The first thing necessary to obtain one single supermodel is that the regression coefficients β_{LM} depend on s in a smooth way and to model that in a linear way. This means that:

$$h(t|\boldsymbol{x}, s) = h_0(t|s) \exp(\boldsymbol{x}^\top \boldsymbol{\beta}_{LM}(s)) \tag{3.51}$$

for $s \leq t \leq t_{hor}$, where $h_0(t|s)$ is the unspecified baseline hazard and $\beta_{LM}(s)$ is an arbitrarily defined smooth function (e.g. polynomial, spline) of the landmark time s. In practice we posit a linear model of $\beta_{LM}(s)$ on s.

$$\boldsymbol{\beta}_{LM}(s) = \sum_{j=1}^{m_b} \theta_j f_j(s) \tag{3.52}$$

with a set of m_b basis functions $\{f_1(s), f_2(s), \ldots, f_{m_b}(s)\}$ and a vector $\boldsymbol{\theta}$ of parameters. Considering a dataset with observations $(t_i, \delta_i, \boldsymbol{x_i})$ (observation time, event indicator, vector of covariates, respectively) for $i = 1, \ldots, n$, in which there are no ties concerning the event times, a dataset for landmarking at time s can be created by selecting all individuals with $t_i \geq s$. The corresponding partial log-likelihood is given by

$$\mathrm{pl}_s(\boldsymbol{\beta}_{LM}(s)) = \sum_{t_i \geq s} \delta_i \left[\boldsymbol{x}_i^\top \boldsymbol{\beta}_{LM}(s) - \ln \left(\sum_{t_j > t_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_{LM}(s)) \right) \right] \tag{3.53}$$

We can create multiple landmarking datasets for different values of s. Individual i will have a record in all datasets in which $s \leq t_i$. As we have seen, these datasets can be merged into one big stacked dataset, where the dependence of $\boldsymbol{\beta}_{LM}(s)$ can be investigated by fitting a Cox model in this big dataset with stratification on landmark points s. This leads to a pseudo-partial likelihood that is equal to $\sum_s \mathrm{pl}_s(\boldsymbol{\beta}_{LM}(s))$. Here, $\psi(s)$ is an indicator function that takes on value 1 in the specified window $[s; t_{hor}]$ and 0 otherwise. This function is mainly introduced to simplify the notation. The practical procedure outlined above with a very fine grid of landmark points is equivalent to maximizing the integrated partial log-likelihood:

$$\begin{aligned}
\mathrm{ipl}(\boldsymbol{\beta}_{LM}) &= \int_0^{t_{hor}} \mathrm{pl}_s(\boldsymbol{\beta_{LM}}(s)) \psi(s) ds = \\
&= \int_0^{t_{hor}} \sum_{t_i \geq s} \delta_i \left[\boldsymbol{x}_i^\top \boldsymbol{\beta}_{LM}(s) - \ln \left(\sum_{t_i \geq t_j} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_{LM}(s)) \right) \right] \psi(s) ds = \\
&= \sum_{i=1}^{n} \delta_i \left[\boldsymbol{x}_i^\top \int_0^{t_i} \boldsymbol{\beta}_{LM}(s) \psi(s) ds - \int_0^{t_i} \ln \left(\sum_{t_j \geq t_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_{LM}(s)) \right) \psi(s) ds \right]
\end{aligned} \tag{3.54}$$

The estimating equations for the parameters $\boldsymbol{\theta}$ in (3.52), when maximizing (3.54), are similar to the standard estimation equations for the Cox model and can be conceived as the weighted sum of contributions of the separate risk sets. The longitudinal nature of the new dataset gives rise to the possibility of dependence between observations, because we can have repeated observations on the same subject. To account for this possible dependence one can use the robust sandwich estimator of Lin & Wei (1989) that takes into account the correlation between risk sets contributions to the estimating equations induced by estimating the common regression parameters.

The stratified supermodel (by fitting (3.52) and maximizing (3.54)) conveniently estimates smooth landmark-dependent covariate effects $\beta_{LM}(s)$. However, it provides separate estimated baseline hazards at the event time t_i for each landmark stratum under the following expression:

$$\hat{h}_0(t_i|t_{LM} = s) = \frac{1}{\sum_{t_j \geq t_i} \exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_{LM}(s))} \tag{3.55}$$

To address the issue of separate baseline hazards for each landmark stratum, van Houwelingen (2007) proposed another model called the proportional baselines landmark supermodel. The premise is to model a common baseline hazard through a multiplicative dependence of two components: the set of landmark-specific baseline hazards and a smooth function of the landmark time s. Formally, $h_0(t|s) \equiv h_0(t)\exp(\boldsymbol{\gamma}(s))$. Therefore, the hazard function follows the form:

$$h(t|\boldsymbol{x}, s) = h_0(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s)) \tag{3.56}$$

for $s \leq t \leq t_{hor}$. In practice, we fit the gamma function via a linear model:

$$\boldsymbol{\gamma}(s) = \sum_{j=1}^{m_h} \eta_j g_j(s) \tag{3.57}$$

with a set of m_h basis functions $\{g_1(s), g_2(s), \ldots, g_{m_h}(s)\}$ and a vector of η parameters. In essence, when analysing a stacked dataset we can obtain such a model-based estimate by employing an unstratified analysis and adding a landmark term to the model. To be more precise, an individual i gets a record for each $s \leq t_i$ specifying that he enters the study at time s and has observation time t_i. In this approach the risk sets are much higher. Let $n_{is} = \#(s : s \leq t_i)$. Then, each individual at risk at t_i has n_{is} copies in that risk set and the individual with an event ($\delta_i = 1$) at t_i in the original dataset gets n_{is} tied events in the stacked data. In this stacked dataset Breslow's partial log-likelihood for tied events can be used to estimate the parameters. A more formal way is to start the integrated Poisson-type full log-likelihood:

$$\begin{aligned} &\text{il}(\boldsymbol{\beta}_{LM}, \boldsymbol{\gamma}, h_0) = \\ &= \int_0^{t_{hor}} \sum_{i=1}^{n} \left[-\exp(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s)) \sum_{s \leq t_j \leq t_i} h_0(t_j) + \delta_i(\boldsymbol{x}_i^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s) + \ln(h_0(t_i))) \right] \psi(s)ds \end{aligned} \tag{3.58}$$

Maximizing with respect to the baseline hazard at the event times leads to a slightly different version of the integrated partial log-likelihood, namely:

$$\begin{aligned} \text{ipl}^*(\boldsymbol{\beta}_{LM}, \boldsymbol{\gamma}) &= \sum_{t_i} \int_0^{t_i} \left[\boldsymbol{x}_i^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s) - \ln\left(\sum_{t_j \geq t_i} \int_0^{t_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s))\psi(s)ds \right) \right] \psi(s)ds = \\ &= \sum_{t_i} \left[\boldsymbol{x}_i^\top \int_0^{t_i} \boldsymbol{\beta}_{LM}(s)\psi(s)ds + \int_0^{t_i} \boldsymbol{\gamma}(s)\psi(s)ds \quad - \right. \\ &\left. - \int_0^{t_i} \psi(s) \ln\left(\sum_{t_j \geq t_i} \int_0^{t_i} \exp(\boldsymbol{x}_j^\top \boldsymbol{\beta}_{LM}(s) + \boldsymbol{\gamma}(s))\psi(s)ds \right) \right] \end{aligned} \tag{3.59}$$

The corresponding common baseline hazards will be estimated via the following formula:

$$\hat{h}_0^*(t_i) = \frac{\int_0^{t_i} \psi(s)ds}{\sum_{j:t_j \geq t_i} \int_0^{t_i} \exp(\boldsymbol{x}_j^\top \hat{\boldsymbol{\beta}}_{LM} + \hat{\boldsymbol{\gamma}}(s))\psi(s)ds} \tag{3.60}$$

Note that the estimated hazard $\hat{h}_0^*(t_i)$ no longer depends on s. Let $\hat{H}_0^*(t) = \sum_{t_j \leq t} \hat{h}_0^*(t_i)$ be the corresponding cumulative hazard, then the simple predictive landmark model is given by

$$\widehat{S}_{LM}(t|\boldsymbol{x}, s) = \exp(-\exp(x^\top \hat{\boldsymbol{\beta}}_{LM}(s) + \hat{\boldsymbol{\gamma}}(s))(\hat{H}_0^*(t) - \hat{H}_0^*(s-))) \tag{3.61}$$

3.5 Measures to assess predictive performance

Harrell *et al.* (1996) identify three distinct objectives for the use of measures of predictive capacity, including:

- To quantify the utility of a predictor or model to be used for prediction or for screening to identify subjects at increased risk of a disease or clinical outcome;
- To check a given model for overfitting or lack of fit;
- To rank competing methods or competing models.

The evaluation of the predictive capacity of a model can be decomposed in the evaluation of the calibration and discrimination. Calibration refers to the extent of bias. On the other hand, discrimination measures a predictor's ability to separate patients with different responses. To evaluate the predictive performance of the above methods, we focused on calibration and discrimination. The Brier Score was calculated since it assesses both discrimination and calibration. We assessed discrimination by calculating Harrel's c-index.

3.5.1 Brier Score

The Brier Score was implemented as defined in Graf *et al.* (1999). This measure evaluates the discrepancy between predicted and observed values in certain times t^*. In fact, the Brier score for a survival time that depends on covariates X is defined as the mean quadratic difference between survival status observed at a given time t^* and the expected probability of survival beyond this time according to the predictions of the model, as presented in (3.62). The generalization of the Brier score for censored data is presented in (3.63). Smaller values (closer to 0) of this measure indicate better forecasts.

Considering that, for each patient, we observe $\widetilde{T}_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$, where T_i represents the time to the event of interest and C_i the (hypothetical) time under observation $(i = 1, \ldots, n)$. C is distributed according to $G(t) = P(C > t)$. For a fixed time point t^*, the contributions to the Brier score can be split up into three categories:

1. $\widetilde{T}_i \leq t^*$ and $\delta_i = 1$;

2. $\widetilde{T}_i > t^*$ ($\delta_i = 1 \quad$ or $\quad \delta_i = 0$);

3. $\widetilde{T}_i \leq t^*$ and $\delta_i = 0$.

For the uncensored observations of category 1 the event occurred before t^*, and the event status at t^* is equal to $I(T_i > t^*) = 0$. Thus, the contribution to the Brier Score is $(0-\widehat{S}(t^*|X_i))^2$. In category 2 the observed event status at t* is equal to 1 since all of these patients are known to be event-free at t^*; the resulting contribution to the Brier score is $(1 - \widehat{S}(t^*|X_i))^2$. For the censored observations of category 3 the censoring occurred before t* so that the event status at t^* is unknown; thus their contribution to the Brier score cannot be calculated. To compensate for the loss of information due to censoring, the individual contributions have to be reweighted: observations in category 1 get the weight $1/\widehat{G}(\widetilde{T}_i)$ those of category 2 get the weight $\widehat{G}(t^*)$ and observations of category 3 get weight zero.

$$BS(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left(I(T_i > t^*) - \widehat{S}(t^*|X_i)\right)^2 \tag{3.62}$$

$$BS^c(t^*) = \frac{1}{n}\sum_{i=1}^{n}\left\{\left(0 - \widehat{S}(t^*|X_i)\right)^2 I(\widetilde{T}_i > t^*, \delta_i = 1)\left(1/\widehat{G}(\widetilde{T})\right) + \left(1 - \widehat{S}(t^*|X_i)\right)^2 I(\widetilde{T}_i > t^*)\left(1/\widehat{G}(t^*)\right)\right\} \tag{3.63}$$

where $I(T > t^*)$ and $I(\widetilde{T} > t^*) \in \{0, 1\}$ are the observed event status, $\widehat{S}(t^*|X_i)$ is the estimated probability that the event does not occur and $\widehat{G}(t)$ denotes the Kaplan-Meier estimate of the censoring distribution G.

3.5.2 Harrell's c-index

Statistics that summarise the agreement or concordance between the ranks of observed and predicted survival times are useful in assessing the predictive ability of a model. These statistics summarise the potential of a fitted model to discriminate between individuals, by separating those with longer survival times from those with shorter times. As for measures of explained variation, these statistics take values between 0 and 1, corresponding respectively to perfect discordance and perfect concordance. Values around 0.5 are obtained when a model has no predictive ability, and models with a reasonable degree of predictive ability would lead to a value greater than 0.7.

A particular measure of concordance is the c-statistic described by Harrell *et al.* (1996). This statistic is an estimate of the probability that, for any two individuals, the one with the shortest survival time is the one with the greatest hazard of death. To calculate this statistic, consider all possible pairs of survival times, where either both members of the pair have died, or where one member of the pair dies before the censored survival time of the other. Pairs in which both individuals have censored survival times, or where the survival time of one individual exceeds the censored survival time of the other, are not included. If in a pair where both individuals have died, the model-based predicted survival time is greater for the individual who lived longer, the two individuals are said to be concordant. In a proportional hazards model, an individual in a pair who is predicted to have the greatest survival time will be the

one with the lower hazard of death at a given time, the higher estimated survivor function at a given time, or the lower value of the risk score. For pairs where just one individual dies, and one individual has a time that is censored after the survival time of the other member of the pair, the individual with the censored time has survived longer than the other, and so it can be determined whether the two members of such a pair are concordant. The c-statistic is obtained by dividing the number of concordant pairs by the number of all possible pairs being considered (Collett, 2015).

4

Analysis of Breast Cancer Data

This study consists of 4620 female patients diagnosed with early-stage breast cancer between January 2006 and December 2011. Analysis called for a sequential approach to construct a dynamic prediction model. Upon performing a descriptive analysis, we develop the traditional survival regression models and progress to the conditional survival approach followed by the landmark method of van Houwelingen, and then examine its predictive ability. Statistical analysis were performed using R, version 3.4.1. Since the lack of availability of software to perform dynamic predictions, in-house script was developed, using our own functions. We only made use of one useful function to create landmark datasets (*cutLM*) presented in the *dynpred* package.

4.1 A closer look

4.1.1 Description of the data

A descriptive analysis was performed to characterize the cohort in what regards the clinical and demographic characteristics. The baseline characteristics of the patients are found in Table 4.1. Median age at diagnosis was 59 years (range: 23-95). The majority of patients had the disease in the earliest stage (41.1%) and only 18.9% of the cases corresponded to stage III. More than a half of the patients had intermediate histological grade (57.7%) and no axillary node involvement (58.8%). The most common immunohistochemistry (IHC) subtype was HR+/HER2- (69.2%), followed by HR+/HER2+ and HR-/HER2- (10% each) and HR-/HER2+ (5%).

Among the 4620 patients, 13.4% had a recurrence, and 20% died (all causes). In cancer statistical analysis, survival estimates are usually computed at 2 or 5 years after diagnosis or surgery. Figure 4.1 shows the Kaplan-Meier estimate of the survival function for overall survival (OS) and disease-free survival (DFS). Considering OS, the probability of a patient surviving beyond 2 and 5 years after diagnosis is $\hat{S}(2) = 0.964$ and $\hat{S}(5) = 0.879$, respectively. This means that, for instance, without prior information of the potential factors that can influence time to death, any patient has an estimated probability of 0.879 of surviving for more than 5 years after diagnosis. In what regards DFS, the probability of surviving without recurrence for more than 2 or 5 years after surgery is $\hat{S}(2) = 0.924$ and $\hat{S}(5) = 0.824$.

Table 4.1: Baseline clinical and demographic characteristics

Category	N = 4620
Age, Median (range)	59 (23,95)
Stage, N (%)	
I	1898 (41.1)
II	1714 (37.1)
III	872 (18.9)
Missing	136 (2.9)
Tumour grade, N (%)	
Low	727 (15.7)
Moderate	2666 (57.7)
High	871 (18.9)
Missing	356 (7.7)
Lymph node status, N (%)	
Negative	2715 (58.8)
Positive	1787 (38.7)
Missing	118 (2.5)
Immunohistochemistry subtype, N (%)	
HR+/HER2-	3196 (69.2)
HR+/HER2+	470 (10.2)
HR-/HER2+	245 (5.3)
HR-/HER2-	442 (9.5)
Missing	267 (5.8)

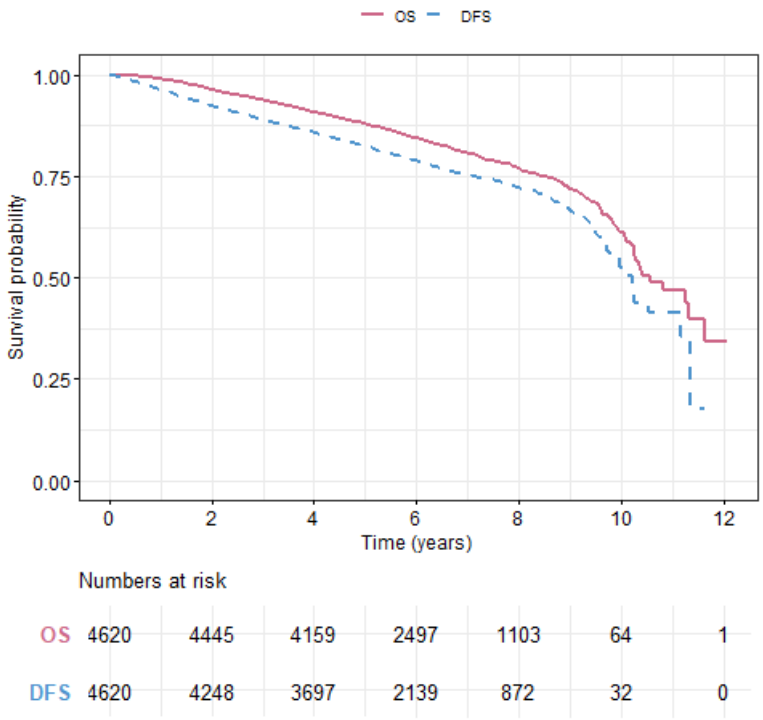

Figure 4.1: Kaplan-Meier estimate of the survival function for OS and DFS

Through the estimation of quantiles for both OS and DFS (Table 4.2), we see that up until 8.5 years, 25% of population died. The median time to death was 10.6 years and the median time to recurrence or death was 10.2 years. At the end of the study, more than 25% of the patients were alive.

Table 4.2: Quantile estimation for OS and DFS with 95% confidence interval (CI)

	OS	DFS
$\hat{\chi}_{0.25}$ [CI 95%]	8.5 [8.0;8.9]	7.1 [6.7;7.7]
$\hat{\chi}_{0.5}$ [CI 95%]	10.6 [10.2;-]	10.2 [9.9;-]
$\hat{\chi}_{0.75}$ [CI 95%]	-	11.4 [11.2;-]

Although we have prior knowledge about the factors that can influence time to death or time to recurrence or death, it seems reasonable to make such validation and therefore confirm that the clinical factors with known prognostic impact have the expected behaviour in our cohort. In Figure 4.2, Figure 4.3, Figure 4.4 and Figure 4.5 we display Kaplan-Meier estimate of the survival function for either OS and DFS stratified by disease stage, tumour grade, lymph node status and IHC subtype, respectively. Subsequently, we computed log-rank tests in order to evaluate possible differences in survival estimates between groups.

As we expected, patients with disease stage III, high grade tumour, lymph node-positive disease and belonging to HR-/HER2- group showed significantly poor prognostic for both outcomes investigated. Although 2-year OS estimates are similar irrespectively of the disease stage (stage I: 0.98; stage III: 0.92), 5-year estimates vary substantially, from 0.94 for patients with stage I to 0.73 for patients with stage III (Figure 4.2). A similar pattern is observed for DFS estimates but the disparity between groups is higher. For instance, 5-year DFS for patients with stage I is 0.92 against 0.62 for patients with stage III.

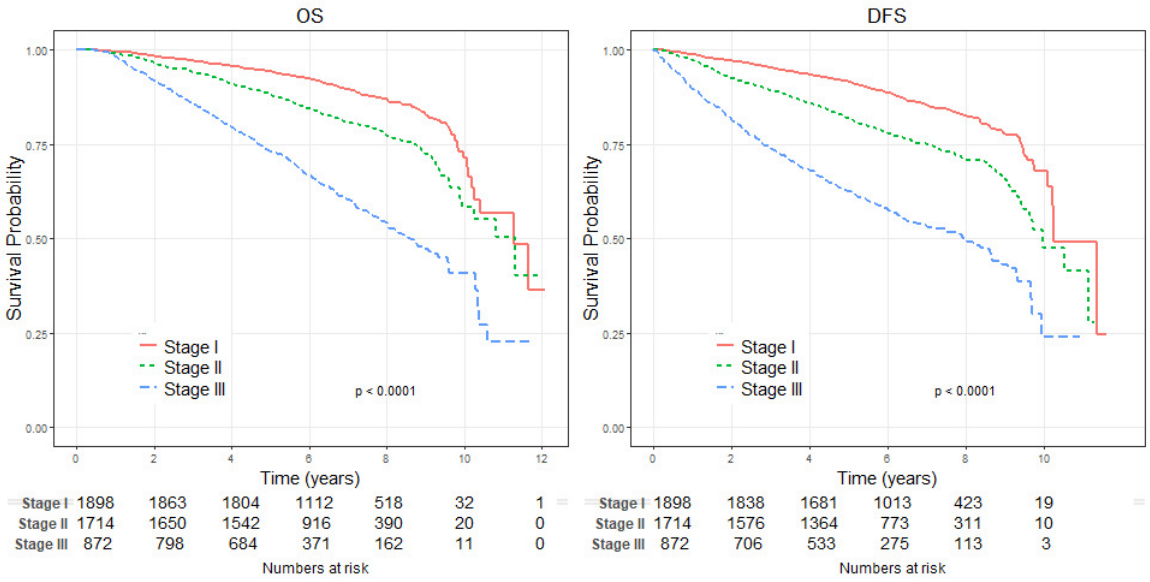

Figure 4.2: Kaplan-Meier estimate of the survival function stratified by disease stage

Regarding the histological grade of the tumour, from Figure 4.3, we see that although estimated 2-year and 5-year survival probabilities are different between groups, this difference is not as strong as it was according to the disease stage. For instance, 2-year DFS estimate is 0.89 for patients with high grade tumours against 0.96 for low grade patients.

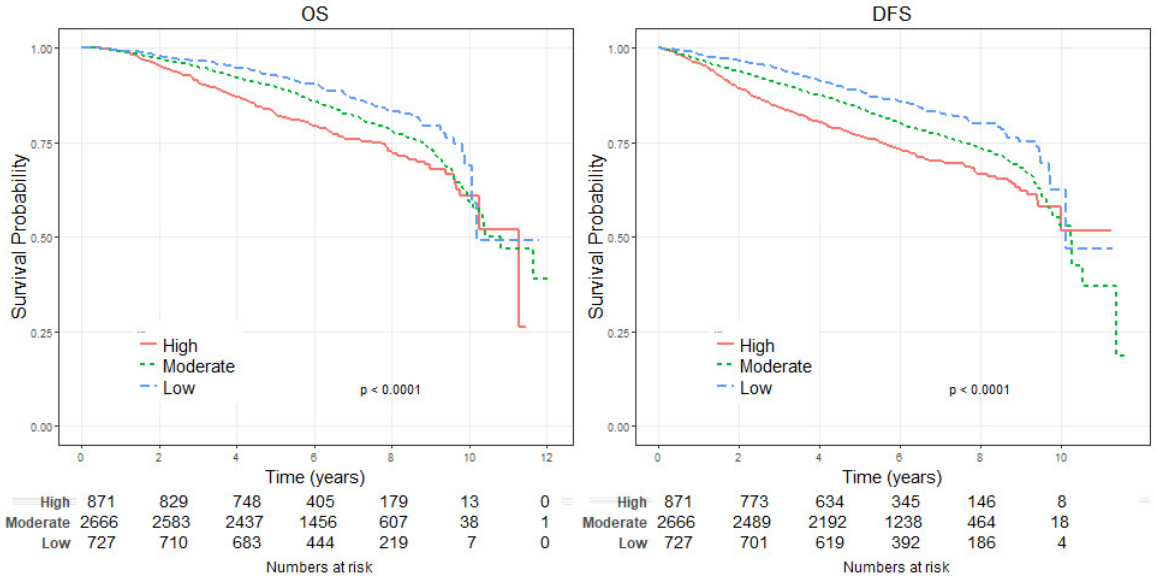

Figure 4.3: Kaplan-Meier estimate of the survival function stratified by tumour grade

From Figure 4.4 we see that patients with lymph node negative disease have an estimated 2-year and 5-year OS of about 0.98 and 0.91, respectively. Survival probability estimates were consistently lower in patients with lymph node positive disease. For instance, 2-year and 5-year OS for such patients is 0.94 and 0.82, respectively. Regarding DFS, we observe that differences in survival probability estimates between groups are more evident than in OS.

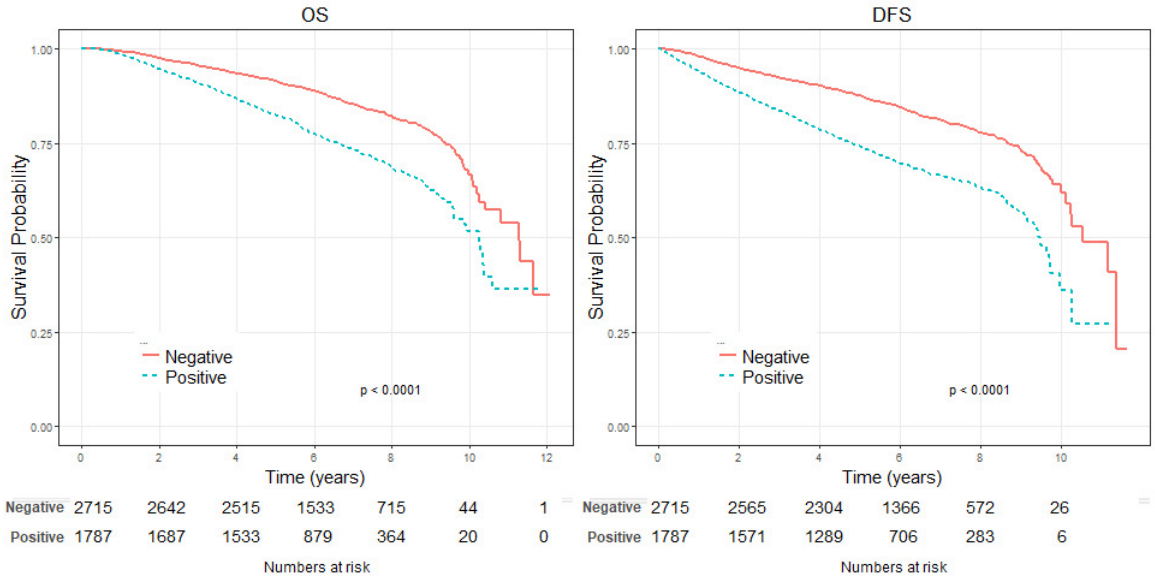

Figure 4.4: Kaplan-Meier estimate of the survival function stratified by lymph node status

Finally, considering the IHC subtype, we see that OS and DFS estimates for HR+/HER2- and HR+/HER2+ groups, remained very close until the end of the study. The same pattern is observed for HR-/HER2+ and HR-/HER2- subtypes. For instance, 5-year DFS estimate for patients with HR+/HER2-, HR+/HER2+, HR-/HER2+ and HR-/HER2- is 0.85, 0.82, 0.69 and 0.71, respectively. As so, 2-year and 5-year OS and DFS estimates are quite similar according to HR status.

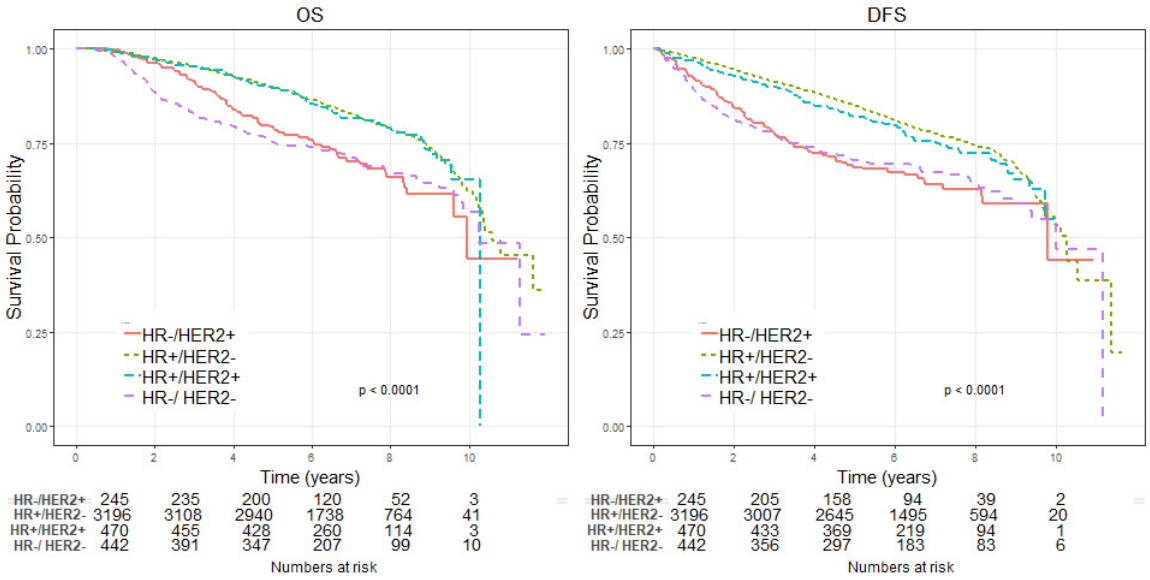

Figure 4.5: Kaplan-Meier estimate of the survival function stratified by immunohistochemistry subtype

In summarising a survival data set the most important information is given by an estimate of the survival function, but it is also relevant to show an estimate of the censoring function. As the results of survival analysis apply to the time frame in which most of the individuals were observed, it is important to quantify follow-up. Figure 4.6 displays the Kaplan-Meier estimates of the censoring distribution, sometimes also called the reverse Kaplan-Meier curves, for both OS and DFS. The reverse Kaplan-Meier survival curve is constructed by reversing 'censor' and 'event' of the standard Kaplan-Meier curve. The advantage of this curve is that it describes the extent as well as the timing of loss to follow-up occurred during the study follow-up. If this curve remained closed to 1 until later in the study, then one can infer nearly complete early follow-up therefore more reliable survival estimates at earlier times than later. The median values of the reverse Kaplan-Meier survival curves referring to the OS and DFS censoring distributions were 6.81 and 6.51 years, respectively. These median values correspond to the follow-up time after diagnosis (OS) or after surgery (DFS) in which 50% of the living patients were censored. Both censoring curves (Figure 4.6) show an initial plateau with values next to one during the first four years of follow-up (indicating virtually no censoring), followed by a linear decrease over time until a near zero value is reached at ten years of follow-up. This should be interpreted taking into account the methodological aspects of the study, namely that patients were included in the cohort between January 2006 and December 2011 (dates of diagnosis according to inclusion criteria) and that the follow-up data in the database was updated for all living patients until 2016 (although there were a few cases with posterior follow-up updates). Indeed this is in line

with what we see in the OS curve, with almost 100% of uncensored patients during the first four years of follow-up, which corresponds to the minimum follow-up time of living patients in our study, calculated as the difference between the 2016 cut-off date and the date of inclusion of the last patients in December 2011. After this plateau period, both OS and DFS curves show a linear decrease with time up to ten years, which roughly corresponds to the maximum follow-up time of living patients in our study (calculated as 2016 minus 2006). Given that the inclusion rate of patients was evenly distributed over the 2006-2011 period, this pattern of linear decrease over time is consistent with administrative censoring. Thus we can reasonably exclude a significant bias due to informative censoring in the data.

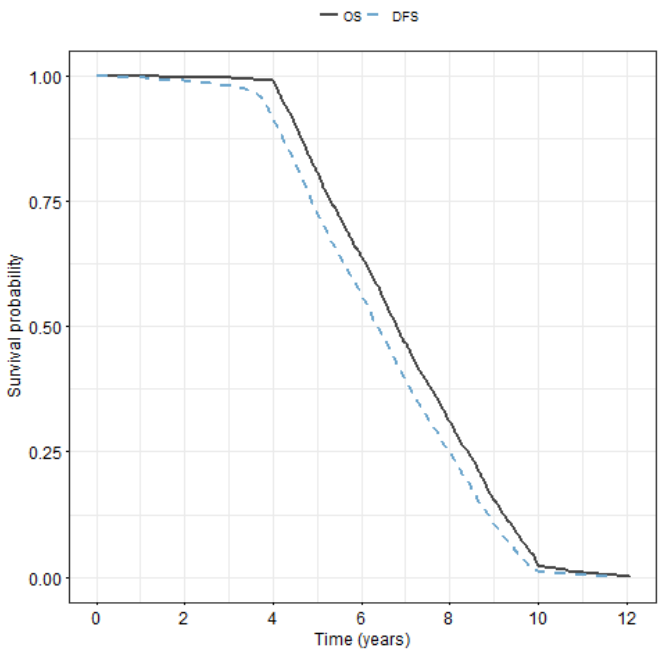

Figure 4.6: Reverse Kaplan-Meier estimates for OS and DFS

4.1.2 Exploratory analysis by Cox models

One purpose of this study is to estimate dynamic predictions of survival that allow the use of dynamic information without creating very complicated models and procedures. Such models cannot be defined without proper exploratory analysis of the data on which they should be based. Traditional survival analyses by the Cox model are a prerequisite to gain insight into the relevance of covariates and the way they are related to survival.

Some results of exploratory analysis by Cox models with regard to DFS can be found in Appendix B. We started by fitting univariable Cox models for each prognostic factor and age (Table B.1). Next, we fitted a multivariable Cox model (Table B.2) to describe how variables jointly impact on survival. We see that, in the presence of disease stage and IHC subtype variables, histological grade of the tumour and lymph node status lose their significance observed in separate models. For each covariate of the correspondent univariable Cox model, tests for a zero slope of the scaled Schoenfeld residuals were performed. The corresponding p-values, as well as the p-value associated with a global test of non-proportionality are reported

in Table 4.3. Considering the separate Cox models for each prognostic factor we see there is a strong evidence of non-proportionality in all cases ($p < 0.0001$). Regarding the multivariable Cox model, the global test suggested also strong evidence of non-proportionality (Table 4.4). Variables that deemed most likely to contribute to non-proportionality were the stage of the disease and the IHC group. These findings suggest a non constant hazard ratio for these variables. Schoenfeld residuals plotted over time for the multivariable model can be found in Figure B.1. Checking the validity of the proportionality assumption for the time-fixed covariates revealed that the effect of the covariates varies considerably over time. Strategies for dealing with non-proportional hazards can be applied but this is not pursued here since the emphasis of this study is on the use of conditional survival and landmarking for dynamic prediction purposes.

Table 4.3: Test for proportionality of the hazards for the univariables Cox models

	rho	chisq	p
Age	0.119	22.2	<0.0001
Disease stage			
Stage II	-0.079	6.84	0.008
Stage III	-0.158	26.47	<0.0001
Global	NA	26.77	<0.0001
Tumour grade			
High	-0.107	11.06	<0.0001
Intermediate	-0.037	1.34	0.247
Global	NA	14.75	<0.0001
Lymph node status			
Positive	-0.076	6.34	0.011
IHC subtype			
HR+/HER2+	-0.038	1.51	0.218
HR-/HER2+	-0.114	13.63	<0.0001
HR-/HER2-	-0.204	43.36	<0.0001
Global	NA	51.25	<0.0001

Table 4.4: Test for proportionality of the hazards for the multivariable Cox model

	rho	chisq	p
Age	0.114	15.29	<0.0001
Stage II	-0.064	3.96	0.04
Stage III	-0.104	10.38	0.0012
Tumour high	0.008	0.06	0.80
Tumour intermediate	-0.003	0.008	0.92
Lymph node status positive	0.031	0.921	0.33
HR-/HER2+	-0.095	8.095	0.004
HR+/HER2+	-0.031	0.877	0.348
HR-/HER2-	-0.189	34.181	<0.0001
Global	NA	77.00	<0.0001

4.2 Conditional survival displayed as a function of prediction time

Conditional survival is indubitably a very important and reliable measure which reflects how prognosis changes over time. We assessed 2-year and 5-year conditional overall survival (COS) and conditional disease-free survival (CDFS), denoted as:

- COS[1] - probability of a patient surviving t years, given that is alive and disease-free s years after diagnosis;
- CDFS - probability of a patient surviving and being disease-free t years, given that is alive and disease-free s years after surgery.

Prediction time points s were defined at $s = 0, 1, 2, 3, 4, 5$. After the construction of a dataset consisting of only individuals at risk at each prediction time point, a prediction window w was fixed, and right-censoring was imposed at $t = s + w$. In addition, to estimate 2-year and 5-year COS and CDFS, two prediction windows were considered: $w = 2$ and $w = 5$. It should be stressed that in COS, individuals at risk are those alive and disease-free s years after diagnosis, while in CDFS it corresponds to those alive and disease-free s years after surgery. Therefore, among individuals at risk at time s, time t can de defined as $t = s + 2$ and $t = s + 5$. This concept is illustrated more clearly in Figure 4.7 where it is shown that time t varies proportionally regardless of the fixed window w. The dark grey bar represents the time considering a window of 2 years whereas the light grey corresponds to a window of 5 years. It is important to note that in COS, time t corresponds to time since diagnosis until death, while in CDFS it corresponds to time since surgery until recurrence or death.

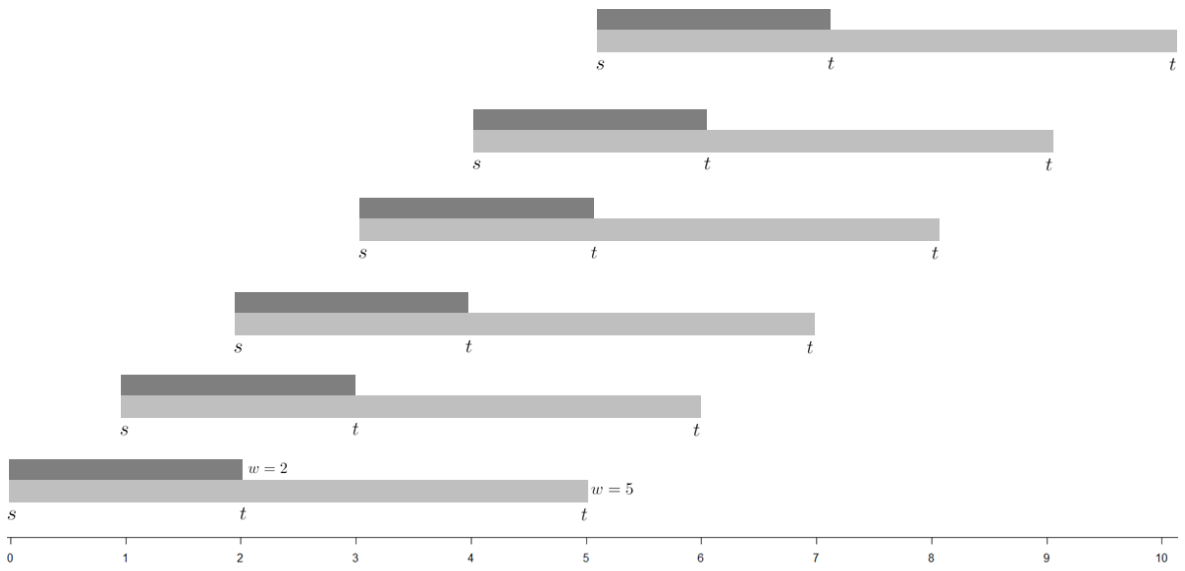

Figure 4.7: Dynamic prediction mechanism computed in this study

[1]The same approach was used by Zamboni (2010).

4.2.1 Exploratory analysis

Two important requirements for a suitable conditional survival analysis are a sufficiently large dataset and an almost complete follow-up. Our cohort consisted of 4620 patients diagnosed with breast cancer, which may be considered fairly large but, as expected, subgroups resulting from stratification, for example, by stage categories, are much smaller. Therefore, a limitation of conditional survival estimated in strata is that resulting subgroups become too small to produce sensible estimates for later prediction times. In Table 4.5 we give the overall number of patients at risk at various time points for both COS and CDFS and in Appendix C, Figure C.9 we present it, in a graphical form, stratified by disease stage, tumour grade, lymph node status and IHC subtype. For instance, the number of patients at risk at time $s = 2$ in the COS framework is the number of patients alive and disease-free 2 years after diagnosis. We see that these numbers dramatically decrease over time because patients who die, have a disease recurrence or are censored are no longer considered in the risk set.

In order to carry out a more complete inspection, we also present in Table 4.6 the number of events within 2 and 5 years for COS and CDFS after each time point s, among those alive and disease-free at each s years after diagnosis ou surgery. An example of interpretation is that, in the dataset for COS analysis comprising the individuals at risk (alive and disease-free) at $s = 2$ years after diagnosis, the number of events (deaths) observed in the time window of 2 and 5 years are 144 and 492, respectively. We also provide such analyses stratified by disease stage, tumour grade, lymph node status and IHC subgroup, in Appendix C, Figure C.10. It should be pointed out that in some subgroups the stratified number of individuals at risk and the number of events within 2 and 5 years are too small, particularly from $s = 4$ onwards, which demonstrates the lack of robustness in COS and CDFS from that time point. As so, and despite presenting full results, for the sake of accuracy in this work we will only interpret 5-year COS and CDFS estimates until $s = 4$ years.

Table 4.5: Number of individuals at risk at each time point s for both COS and CDFS

s	0	1	2	3	4	5
COS	4620	4536	4328	4153	3976	3093
CDFS	4620	4445	4248	4082	3697	2849

Table 4.6: Number of events within 2 and 5 years after each time point

	Within two years						Within five years					
	0	1	2	3	4	5	0	1	2	3	4	5
COS	165	218	144	172	157	135	543	585	492	421	342	264
CDFS	352	431	372	335	291	225	783	814	691	576	463	352

4.2.2 Overall conditional survival

Now that a complete inspection of the pre-requisites to perform a conditional survival analysis has been done, one can proceed with the technique. Figure 4.8 presents a valuable illustration of what conditional survival symbolizes, in which Kaplan-Meier estimates of COS($t|s$) and CDFS($t|s$) curves are displayed for $s = 0, 1, 2, 3, 4, 5$. Note that, the lower curve, COS($t|0$) and CDFS($t|0$), coincides with the OS and DFS curves, respectively, shown in Figure 4.1, whereas the upper curve represent COS($t|5$) and CDFS($t|5$). COS($t|5$) provides the conditional probability of surviving t years, given that a patient is alive and disease-free $s = 5$ years after diagnosis, whilst CDFS($t|5$) specifies the conditional probability of surviving being disease-free t years given that a patient is alive and disease-free $s = 5$ years after surgery. Figure 4.8 presents the data restricted to all patients alive and disease-free s years after diagnosis (on the left side) and s years after surgery (on the right side), in order to provide a global view of how the conditional survival performs over time. For instance, if we want to estimate COS(3|1) we have to look at the beginning of the curve that represents all individuals alive and disease-free at $s = 1$ years after diagnosis (golden curve) and then look the estimates at time after diagnosis = 3 years. From the set of curves, we can see, for example, that the estimated 2-year COS and CDFS probabilities, highlighted by dots, are almost identical and of about 0.97 and 0.92 respectively, for all prediction time points s. This constant pattern is also verified considering a time window of 5 years.

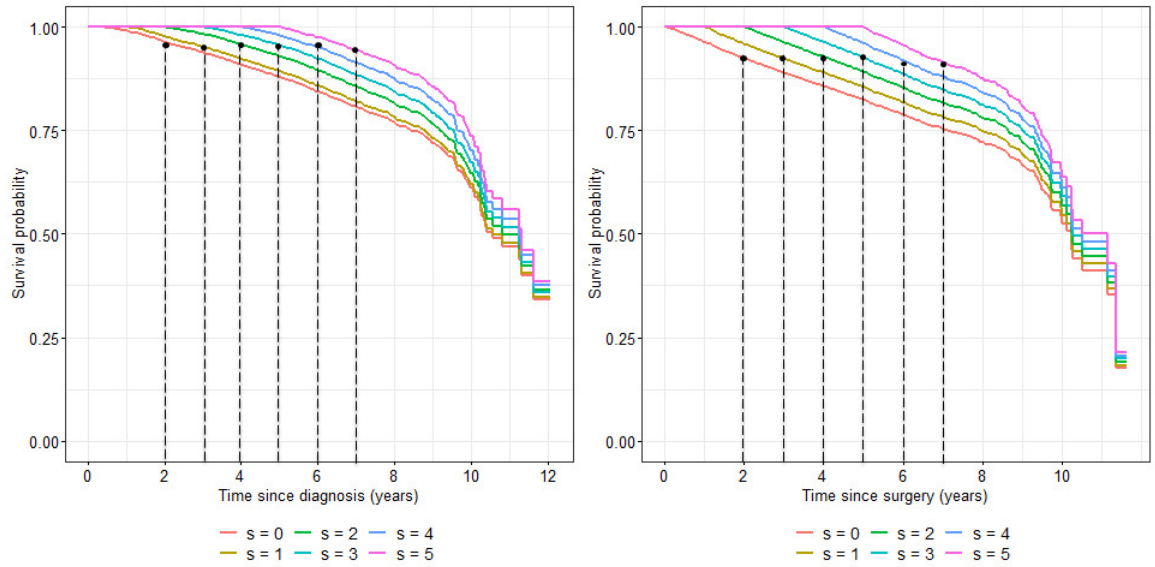

Figure 4.8: COS(t|s) and CDFS(t|s) in the whole cohort for prediction times $s = 0, 1, 2, 3, 4, 5$

Instead of showing a set of survival curves as done in Figure 4.8 we can display a specific time point estimate of COS and CDFS, making it easier to identify possible variations. This means that, for example, we would plot COS(s+2|s), the conditional probability of surviving further 2 years given that a patient is alive and disease-free s years after diagnosis. This can in principle be done for every prediction time s, but in this clinical context it is sufficient to do that considering prediction time points equally spaced by one year. In Figure 4.9, we show the overall estimated 2-year and 5-year COS and CDFS and Table 4.7 present such estimates in a tabular form, together with the respective 95% confidence intervals. Figures 4.10-4.13, present

the 2-year and 5-year COS and CDFS estimates stratified for each prognostic factor. For reasons of clarity, we did not include confidence intervals in the plots. However, we believe that their width has to be taken into account particularly when considering the higher prediction times ($s = 4$ and $s = 5$) and the time window of 5 years. For that reason, in Appendix C, Table C.1 and Table C.2 we present the COS and CDFS estimated probabilities, stratified by prognostic factor, in a tabular form, together with the 95% confidence intervals.

From Figure 4.9 we see that 2-year COS and CDFS are almost constant, and of about 96% and 92% respectively, regardless the prediction time s. The trend regarding the 5-year COS and CDFS also seems to be uniform, although with a slight decrease until $s = 4$ years, which can be justified through the cohort ageing. Furthermore, we recognize a marked decrease at $s = 5$ years. However, COS(10|5) and CDFS(10|5) estimates present wider confidence intervals, as we would expect. We believe that when considering a time windows of 5 years, estimates at $s = 5$ years are not robust and should not be regarded.

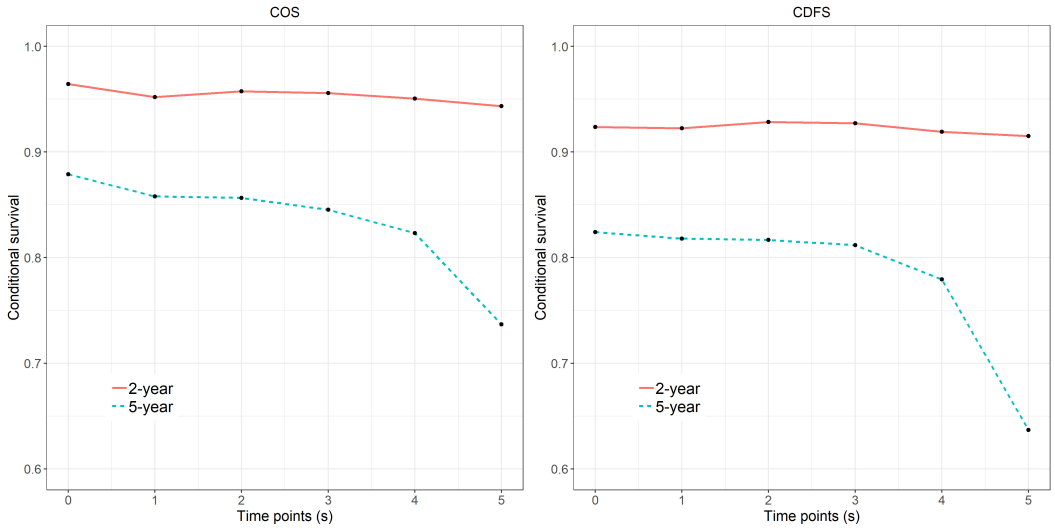

Figure 4.9: 2-year and 5-year COS and CDFS estimates

Table 4.7: 2-year and 5-year COS and CDFS estimates with 95% confidence intervals

	t=s+2					
	0	1	2	3	4	5
COS	0.96 (0.96;0.97)	0.95 (0.95;0.96)	0.96 (0.95;0.96)	0.96 (0.95;0.96)	0.95 (0.94;0.96)	0.94 (0.93;0.95)
CDFS	0.92 (0.92;0.93)	0.92 (0.92;0.93)	0.93 (0.92;0.94)	0.93 (0.92;0.94)	0.92 (0.91;0.93)	0.92 (0.90;0.93)
	t=s+5					
	0	1	2	3	4	5
COS	0.88 (0.87;0.89)	0.86 (0.85;0.87)	0.86 (0.84;0.87)	0.85 (0.83;0.86)	0.82 (0.80;0.84)	0.74 (0.69;0.78)
CDFS	0.82 (0.81;0.84)	0.82 (0.81;083)	0.82 (0.80;0.83)	0.81 (0.79;0.83)	0.78 (0.76;0.80)	0.64 (0.57;0.71)

4.2.3 Conditional survival for each prognostic factor

We will now analyse the conditional survival for each prognostic factor. Although it is typical to examine the big picture of conditional survival, evaluate its pattern and see how it changes over time, the greater interest in this project was to inspect if the survival in different groups of each prognostic factor seem to approximate in some way, i.e., investigate whether the features associated with poor prognosis at the time $s = 0$ maintain their prognostic relevance as more time elapses from diagnosis or surgery. This can be suggested whenever there is a clear approximation of the conditional survival curves (or estimates) between groups with poor and good prognostic baseline features, over time. In order to explore such approximation, we present in Appendix C Kaplan-Meier estimates of the survival function considering all individuals at risk at each prediction time point s, for both COS and CDFS, stratified by prognostic factor.

When looking at 2-year and 5-year COS and CDFS in the three categories of disease stage (Figure 4.10), we see that patients with stage I have a fairly constant 2-year and 5-year COS of about 0.97 and 0.92 respectively, whereas patients with stage III have a lower, but also constant 2-year and 5-year COS of about 0.91 and 0.70, respectively. We also observe a modest approximation in 2-year and 5-year CDFS between patients with stage III and stage I, as more time elapses from surgery. A same deduction can be made when looking at Kaplan-Meier estimates of the survival function considering datasets at $s = 0, 1, 2, 3, 4, 5$, presented in Appendix C, Figures C.1 and C.2. From this, we can see that there is an evident approximation of Kaplan-Meier estimates of the survival function for the different groups, as time goes by. Regarding CDFS estimates we see that, after $s = 4$ years there is once more a remarkable drop in the survival estimate: for stage III, $\text{CDFS}(9|4) = 0.63$, while $\text{CDFS}(10|5) = 0.39$. The latter estimate presents a wide 95% confidence interval of about $[0.21; 0.70]$ which corroborate the lack of robustness inherent to the prediction at $s = 5$ years.

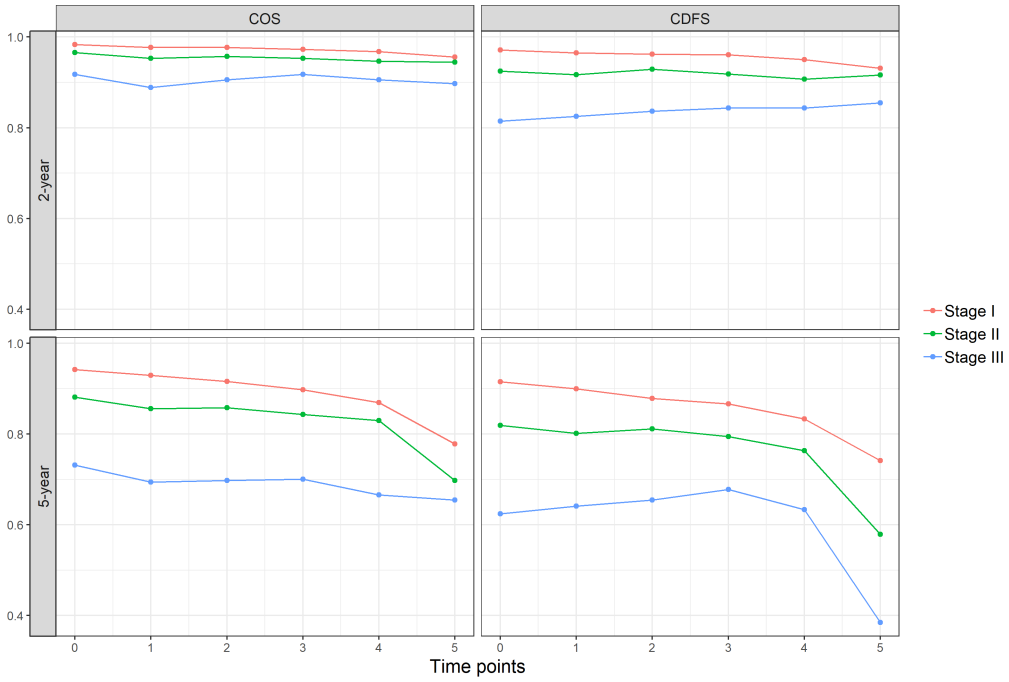

Figure 4.10: 2-year and 5-year COS and CDFS estimates according to disease stage

Considering the histological grade of the tumour, results indicate that, within each group, the probability of surviving further 2 years is almost the same regardless if survival is estimated at $s = 0$ or $s = 5$. These findings also extend themselves for CDFS estimates. For instance, for low tumour grade, the probability of surviving being disease-free further 2 years given that the patient is alive and disease-free 1 year and 5 years after surgery is 0.96 and 0.93, respectively, which demonstrates the slight difference on the survival as more time elapses since surgery. In 5-year COS and CDFS estimates, the gap between the three groups at $s = 0$ is more evident and there are more variations in estimates over the prediction time points s. For such time window, there is an approximation of the survival estimates for patients with high tumour grade with patients that present a moderate or low histological grade.

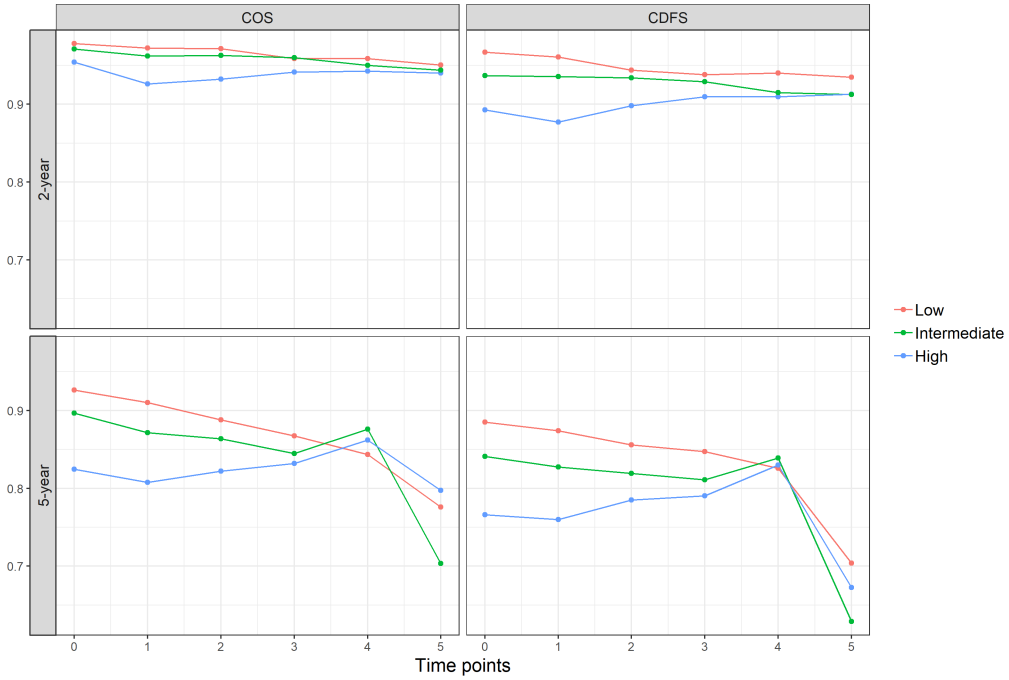

Figure 4.11: 2-year and 5-year COS and CDFS estimates according to tumour grade

Findings regarding the lymph node status have nothing new to offer. From Figure 4.12 we observe that 2-year COS and CDFS are almost constant regardless the prediction time s, being roughly similar for both lymph node status. Difference between both groups is more apparent in the 5-year window. Furthermore there is no evident approximation between both groups over time in 2-year or 5-year COS and CDFS in contrast with what we observe when looking at Kaplan-Meier estimates of the survival function obtained for $s = 0, 1, 2, 3, 4, 5$ (Appendix C, Figures C.5 and C.6).

Interesting comparisons can be made about the IHC subtypes behaviour, in which conditional survival estimates for patients with poor prognosis at $s = 0$ are fully distant from the ones observed at $s = 5$ (Figure 4.13). For instance, 2-year CDFS for HR-/HER2- is 0.81 at $s = 0$ and 0.94 at $s = 4$ years after surgery. We also note that although 2-year and 5-year COS and CDFS for HR+/HER2- and HR+/HER2+ is almost constant over time s, HR-/HER2+ and HR-/HER2- subtypes tend to approach them. Further, 2-year and 5-year COS and CDFS show a trend for a gradual increase in HR- groups, and a gradual decrease in HR+ groups. Additionally,

these findings suggest that in spite of the large differences found regarding survival estimates in IHC subgroups at $s = 0$, the survival of a patient with poorer prognosis detected at $s = 0$ tends to be similar to the survival of a patient with a favourable prognosis, as time goes by. A same conclusion can be provided when looking at Figures C.7 and C.8 in which we see a clear approximation of the survival curves between groups.

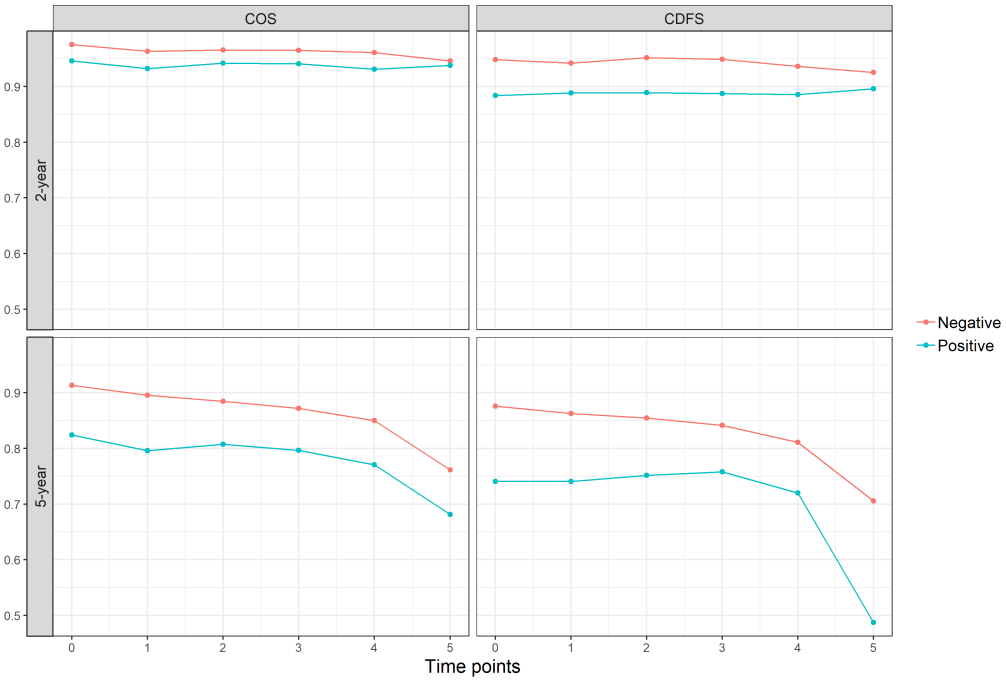

Figure 4.12: 2-year and 5-year COS and CDFS estimates according to lymph node status

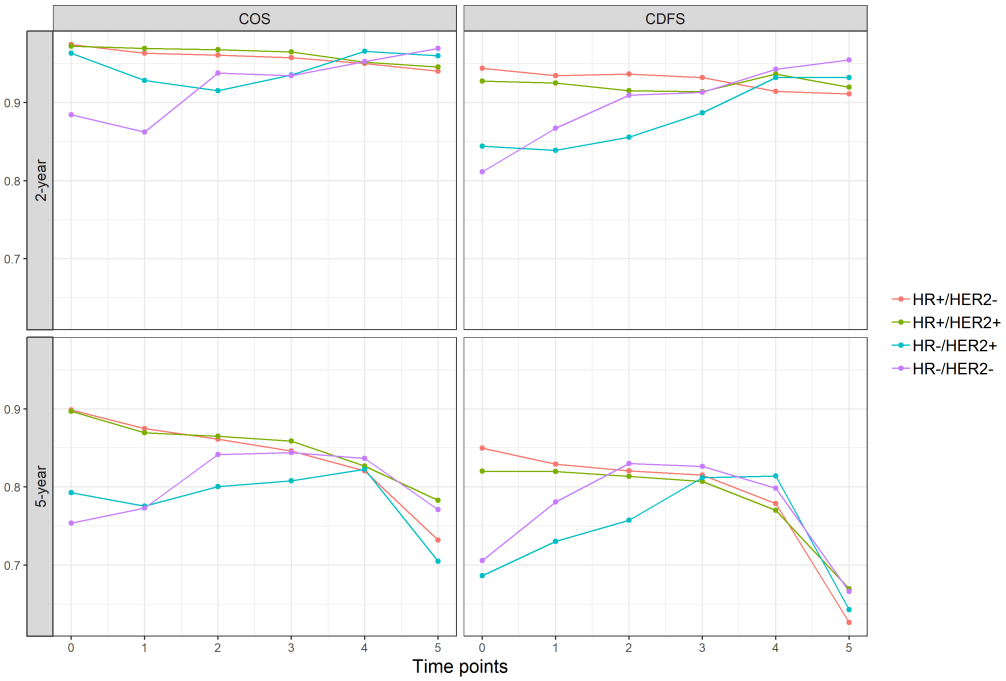

Figure 4.13: 2-year and 5-year COS and CDFS estimates according to immunohistochemistry subtype

4.2.4 Static vs. dynamic estimates

As we previously mentioned, traditional survival estimates are usually reported at the time of diagnosis or surgery, based on OS and DFS estimates. Even so, such estimates are not updated as time passes, and can only be regarded as 'static' estimates. This section aims to compare static and dynamic estimates, and to confirm that dynamic prediction measures, such as conditional survival, can be a more robust tool when it comes to give an estimated probability of survival.

Previously we noted that, without considering covariates, patients who survive being disease-free 3 years after surgery had a 0.93 probability of remaining disease-free additionally 2 years. However, the static 5-year DFS predicted at the time of surgery was 0.82. This gives a first insight of the huge differences between static and dynamic estimates. In fact, it is observed that dynamic prediction measures are more complete measures that produce higher estimates of survival.

There are even greater differences when we compute survival estimates stratified by prognostic factor. As we observed from Figure 4.2, 5-year OS predicted at the time of diagnosis for patients with stage III is 0.73. Of course, this also corresponds to the (conditional) probability of surviving further 5 years given that the patient is alive and disease-free 0 years after diagnosis (COS(5|0)). On the other hand, if we take into account years already passed without disease, different estimates are obtained. For instance, for a patient with stage III, the probability of surviving 5 years, given that it is alive and disease-free 3 years after diagnosis is 0.92 − a much greater probability than the previous one.

Another important prognostic factor that requires our major attention is the IHC subtype. For instance, from Figure 4.5 we observe that 5-year DFS predicted at time of surgery for patients with HR-/HER2- is 0.71. However, if we make use of the dynamic information of the time that the patient has already survived without recurrence, such estimates are greatly improved. For a patient with HR-/HER2- that has already survived being disease-free 3 years after surgery, the probability of surviving being disease-free for 5 years is 0.91.

In fact, such comparisons can be made for all the prognostic factors studied. Conditional survival approach that take into account dynamic information will always provide better estimates than the static ones that are obtained considering only information at the time of diagnosis or surgery. Conditional survival estimates are therefore more reliable, credible and realistic.

4.3 Prediction by landmarking

The next step for the development of dynamic prediction in breast cancer is to fit a proportional baselines landmark supermodel in order to assess time-varying effects of the following covariates: age, disease stage, tumour grade, lymph node status and IHC subtype. To compute this, the landmark time points s were established at every third month between 0 and 5 years after surgery, resulting in 21 points equally spaced. The time horizon is defined as $t_{hor} = s + 2$ and $t_{hor} = s + 5$. A prediction model for 2-year and 5-year DFS at a specific time point is constructed by selecting the individuals at risk (i.e., alive and disease-free and under follow-up) at that time point and incorporating the values of any covariate at that respective time point in a Cox proportional hazards model. More precisely, at each landmark point a simple Cox model was fitted on $(s, s + 2)$ and $(s, s + 5)$. This is called a landmark model. Landmark prediction models at different time points may be combined into a single supermodel through a single super dataset with a landmark column (LM) indexing the landmark points. Landmark supermodels quickly become too large and difficult to test.

4.3.1 Exploratory analysis

Figure 4.14 summarises the number of individuals in each landmark data set, and the number of events (death or recurrence) within 2 and 5 years of each landmark time point. The greatest decrease in the number of individuals at risk is observed at approximately 3.5 years after surgery. At latter prediction time points, the number of events is very reduced.

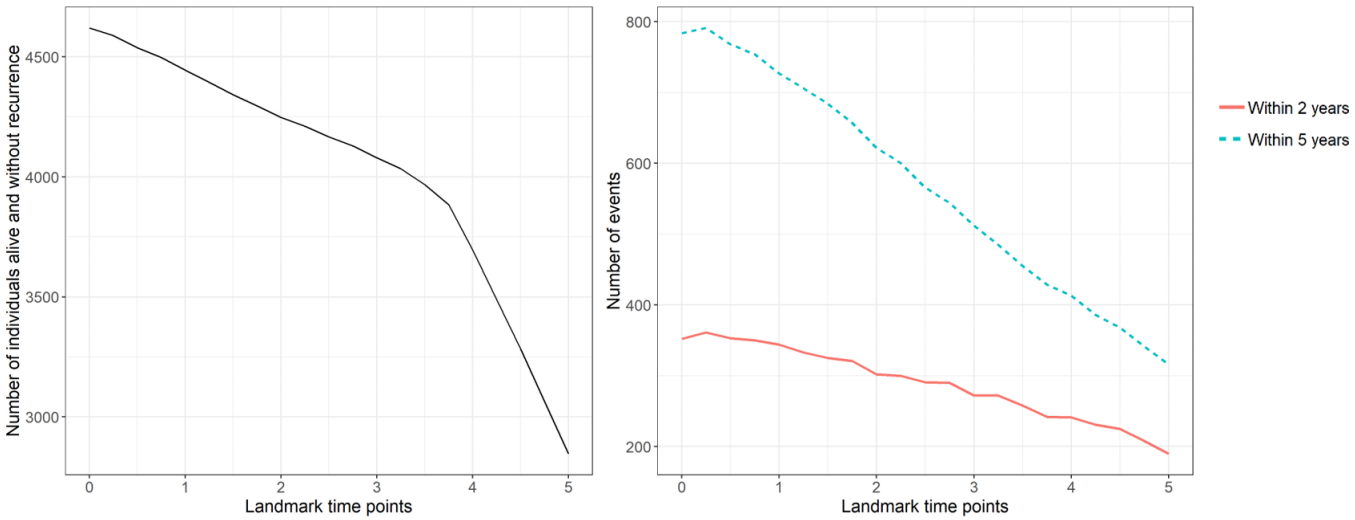

Figure 4.14: Overview of number of individuals in each landmark data set. On the left: Number of individuals alive and disease-free at each landmark time point during the study period. On the right: Number of deaths or recurrences within 2 and 5 years after each landmark time point, among those alive and disease-free at each landmark time point

It is also of great interest to explore how the risk of death and/or recurrence vary over time, in the presence of all prognostic factors which is within the context of a real-life situation. As such, we estimate a landmark model for each of the 21 landmark time points s. Each landmark model is a multivariable Cox model with all variables of interest. Figure 4.15 shows regression coefficients with 95% confidence intervals for each of the covariates at each landmark time point, considering both windows previously defined. Recall that exponentiate the regression coefficient in a context of Cox model, will give the hazard of one group compared to the reference. The time-varying regression coefficients can be undoubtedly detected. Considering a window of 2 years, disease stage and IHC subtype covariates seem to present a linear effect on risk whilst tumour grade and lymph node status come out with a quadratic one. Regarding the 5 years window, the regression coefficients for tumour grade and lymph node status categories seem to be reasonably stable over the (landmark) time, with a linear variation according to the disease stage and IHC subtype. We also observe that the unfavourable group of each covariate analysed presents a greater variation in the regression coefficients (for example, stage III and HR-/HER2-), indicating that for the groups with worse prognosis at $s = 0$ years, their risk of death and/or recurrence will decrease considerably over time, when compared to the group with the best prognosis.

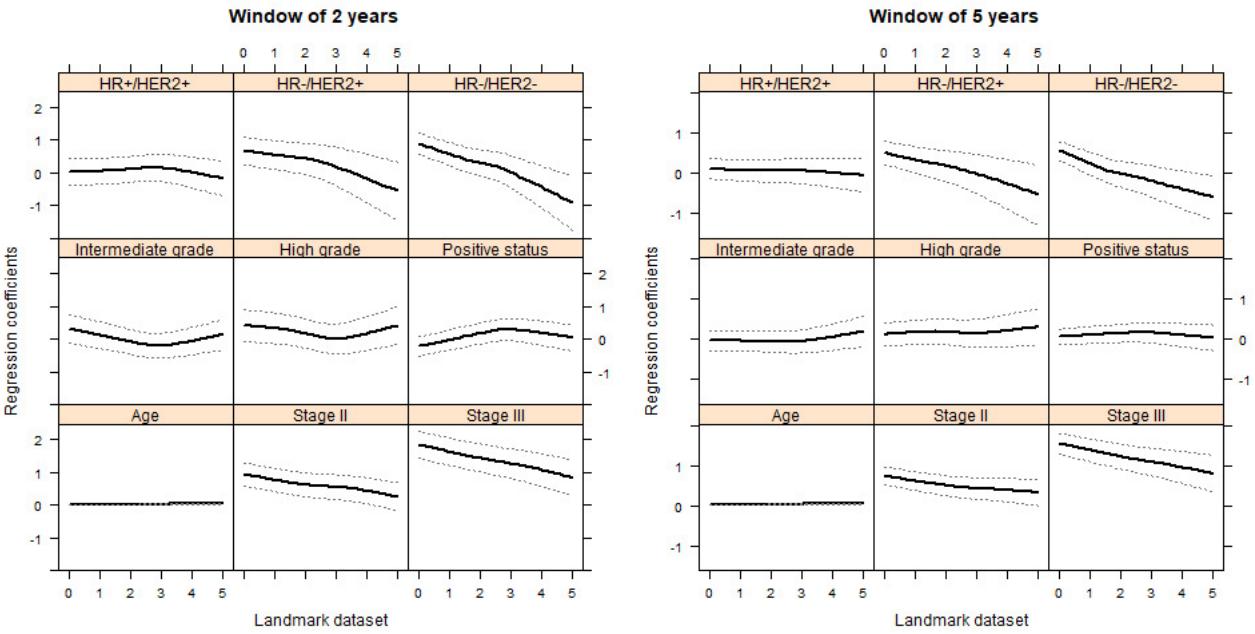

Figure 4.15: Regression coefficients with 95% confidence intervals for the separate landmark analysis

4.3.2 Model building

We now want to construct two landmark supermodels, one for each time window. The objective of a landmark supermodel is to obtain a parsimonious model combining possibly different effects over (landmark) time of covariates.

Table 4.8 shows the result of a backward selection procedure using Wald tests, based on robust standard errors, where we started from a full model with all the main effects, linear interactions of landmark time with all covariates, and quadratic interactions of landmark time with tumour grade and lymph node status covariates, considering a time window of 2 years. In Table 4.9 we again present the results of a backward selection procedure using Wald tests, in which only the main effects and linear interactions with landmark time and all covariates were included, considering a window of 5 years. Keep in mind that in the latter model, none quadratic interaction was tested since we did not observe any quadratic effect from Figure 4.15 considering a window of 5 years. It should also be noted that the main effects of the covariates were included, irrespective of statistical significance. Linear and quadratic interactions of landmark time with covariates were considered significant at the 0.10 level. A forward variable selection procedure was also implemented. Regarding a window of 2 years, this procedure lead to a different final model from the one that was obtained through a backward elimination. Considering a window of 5 years, both variable selection procedures led to the same model. A backward elimination was preferred as it starts with the assumed unbiased global model (Heinze *et al.*, 2017). Details of the both variable selection procedures can be found in Appendix D.

From Table 4.8 we observed that linear interactions of landmark time with age, disease stage, IHC subtype and lymph node status were retained. Plus, a quadratic interaction of landmark time with lymph node status was also kept in the model. In the landmark supermodel regarding a window of 5 years (Table 4.9) only (linear) interactions of landmark time with age, disease stage and IHC subtype were retained. This findings are reasonably in accordance with the results of the separate landmark models of Figure 4.15.

As we know, in a Cox model, a positive regression coefficient for an explanatory variable means that the hazard is higher, and thus the prognosis worse. Conversely, a negative regression coefficient implies a better prognosis for patients with higher values of that variable. An interesting observation regarding the landmark supermodels is that, when considering the main effect of the disease stage and the IHC subtype, regression coefficients are positive. This means that immediately after surgery ($s = 0$) and considering a window of 2 years, the risk of event is 6.25 times higher ($\exp(1.833)$) in stage III patients compared to patients in stage I. However, when considering the significant negative coefficient concerning the interaction with the landmark time, it becomes evident that the worst prognosis associated with stage III decreases with s, i.e., it decreases over time. These findings clearly show that when taking into account the landmark time, the risk of recurrence and/or death is completely different from the one that is obtained by ignoring the dynamic information.

Table 4.8: Landmark supermodel with proportional baseline hazards for death and/or recurrence, based on a spaced set of landmark time points from 0 to 5 with distance 0.25 considering a window of 2 years

Covariate	B	SE	p-value
Age	0.206	0.068	0.002
Age $\times s$	0.163	0.053	0.002
Stage			
I			
II	0.917	0.187	<0.0001
III	1.833	0.219	<0.0001
Stage $\times s$			
II	-0.235	0.133	0.077
III	-0.368	0.157	0.019
Tumour grade			
Low			
Intermediate	-0.020	0.117	0.861
High	0.198	0.143	0.1635
Lymph node status			
Negative			
Positive	-0.286	0.177	0.107
Lymph node status $\times s$			
Positive	0.706	0.274	0.009
Lymph node status $\times s^2$			
Positive	-0.228	0.106	0.003
IHC subtype			
HR+/HER2-			
HR+/HER2+	0.121	0.205	0.554
HR-/HER2+	0.777	0.216	<0.0001
HR-/HER2-	0.996	0.175	<0.0001
IHC subtype $\times s$			
HR+/HER2+	-0.041	0.145	0.779
HR-/HER2+	-0.402	0.171	0.019
HR-/HER2-	-0.674	0.148	<0.0001
$\gamma(s)$			
s	0.013	0.173	0.943
s^2	0.073	0.058	0.211

Table 4.9: Landmark supermodel with proportional baseline hazards for death and/or recurrence, based on a spaced set of landmark time points from 0 to 5 with distance 0.25 considering a window of 5 years

Covariate	B	SE	p-value
Age	0.333	0.049	<0.0001
Age $\times s$	0.308	0.086	<0.0001
Stage			
I			
II	0.708	0.126	<0.0001
III	1.53	0.148	<0.0001
Stage $\times s$			
II	-0.414	0.175	0.017
III	-0.676	0.192	<0.0001
Tumour grade			
Low			
Intermediate	-0.027	0.116	0.818
High	0.186	0.143	0.194
Lymph node status			
Negative			
Positive	0.134	0.099	0.213
IHC subtype			
HR+/HER2-			
HR+/HER2+	0.111	0.144	0.443
HR-/HER2+	0.539	0.173	0.002
HR-/HER2-	0.564	0.146	<0.0001
IHC subtype $\times s$			
HR+/HER2+	-0.125	0.236	0.595
HR-/HER2+	-0.933	0.335	0.005
HR-/HER2-	-1.262	0.269	<0.0001
$\gamma(s)$			
s	0.600	0.145	<0.0001
s^2	-0.176	0.043	<0.0001

4.3.3 Model assessment

One major challenge in evaluating prediction performances within the context of dynamic prediction is how to summarise calibration and discrimination measures over the landmark time points and time horizons. One solution is to display the evolution of such measures over the landmark time points for a fixed time window. We compared the predictive performances of different models in terms of both calibration and discrimination using the Brier score. Another measure used to assess discrimination was C-index. Brier scores and c-indexes were calculated separately for each landmark time point for prediction of 2 and 5-year survival. To compute this, we divided the data into a training-plus-validation set − a 2/3 of the random sample stratified by landmark time and a 'holdout' set with the remaining 1/3.

From Figure 4.16 and regarding a time window of 2 years, we see that for each landmark time point, the Brier score is very low and vary between 0.011 and 0.031 which demonstrates that the predictive ability of the models is very good. Considering a window of 5 years, Brier scores are slightly higher. As a discrimination measure we make use of Harrel's c-index. Recall that a c-index of 1 indicates that the model can perfectly discriminate between patients, while with a c-index of 0.5, the prediction is as good as chance. From Figure 4.17 we observe that c-indexes are almost constant and of about 0.7 for both models considered, showing that the models have a reasonably good predictive ability.

In the landmark supermodel concerning a window of 2 years, for instance, coefficients with regard to the linear interaction between age, disease stage, IHC subtype, and lymph node status covariates with landmark time s and quadratic interaction between lymph node status and landmark time s, were statistically significant and therefore retained. However, one important thing to note is that, when evaluating performances for each landmark time point, each landmark model does not include such interactions terms (since in each landmark model, time s is fixed). Hence, when evaluating prediction performances, the landmark models are not just simply a replica of the landmark supermodel divided into the 21 time points considered. Actually, since each landmark model does not include the interactions with time s, they only present the main effects as covariates. Evaluating predictive performances of landmark supermodels through the 21 landmark models, that don't comprise any interaction term is limited and may lead to biased estimates. However, information available on the literature regarding landmark supermodels assessment is pretty scarce as well as the developed software so far. In fact, although such an evaluation of model performances is technically improper due to the conditional nature of the estimates, we believe that the estimates provide a preliminary assessment of calibration and discriminative ability.

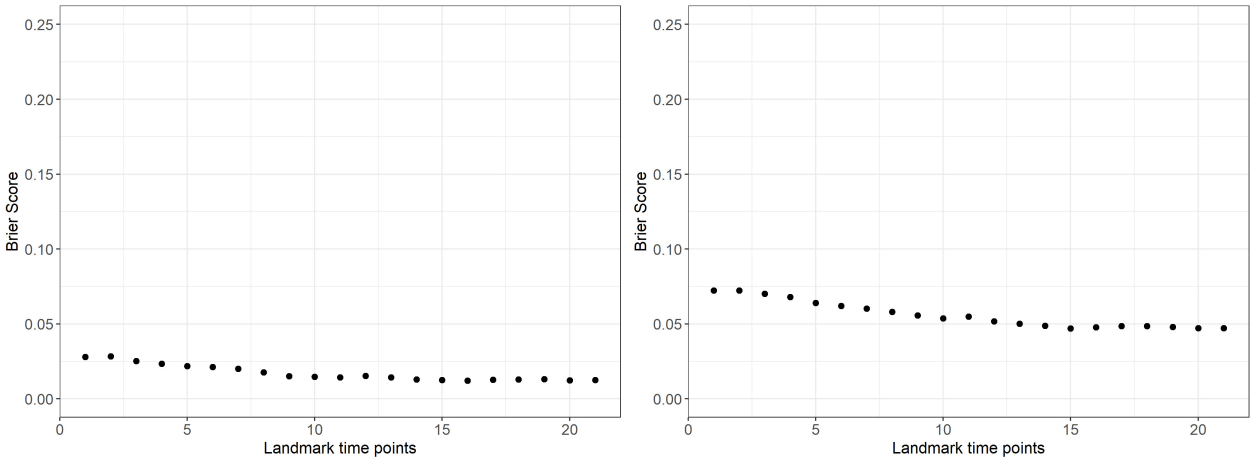

Figure 4.16: Brier score for each landmark model for a prediction at 2-year survival (left) and 5-year survival (right)

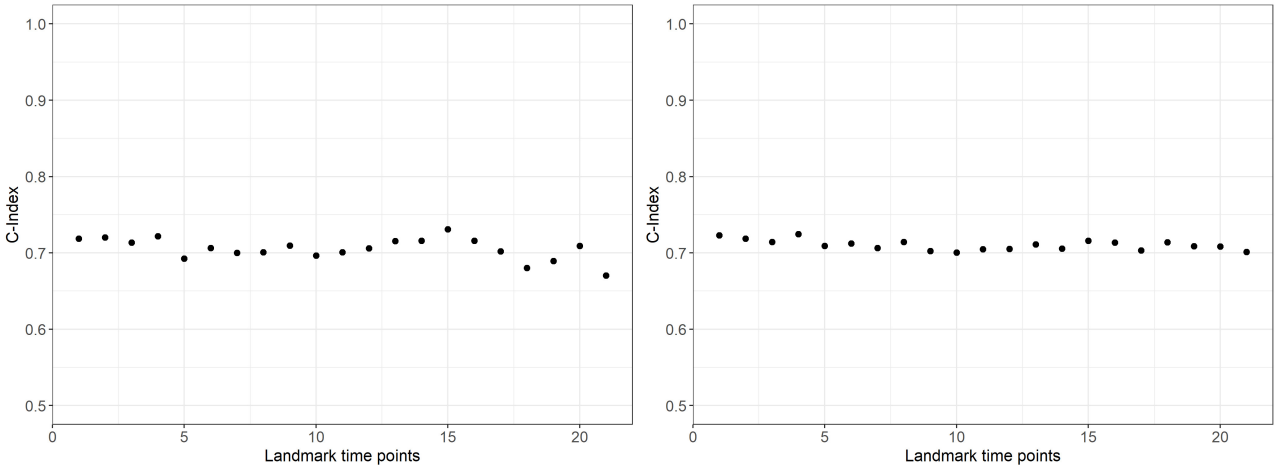

Figure 4.17: C-index for each landmark model for a prediction at 2-year survival (left) and 5-year survival (right)

5
Discussion

The main goal of this study was to develop dynamic prediction methods in patients with early-stage breast cancer, in order to provide further information on the prognosis of such patients, especially how it evolves over time. Actually, prognosis of breast cancer patients is usually evaluated in light of the prognostic factors at the time of diagnosis. However, the prognosis of some patients is not completely homogeneous, and the hazard function is not constant over time. Although diagnostic predictions are useful in guiding the selection of the treatment approach, they might lose their accuracy in the long run once a patient passes some predicted point. In this work, two approaches to compute dynamic predictions were used. First, we used conditional survival methods to evaluate overall survival (OS) and disease-free survival (DFS), conditional on the time lived without disease. Secondly, we used a landmarking approach to assess the long-term significance of prognostic factors.

We strongly believe that conditional survival statistics provide a more accurate prognostic estimate than traditional survival estimates that are based on OS and DFS and measured at the time of diagnosis. Although we cannot formally test a related hypothesis, the close examination of predictions enables us to support our claim. For example, we noted that patients who survive being disease-free 3 years after surgery had an 93% chance of remaining disease-free additionally 2 years, whereas the static 5-year DFS predicted at the time of surgery was 82% − a difference of 11%. Notable gains in dynamic survival estimates can be observed when we stratify patients, for instance, by immunohistochemistry (IHC) subtype. Hence, for patients with HR-/HER2-, the probability of remaining disease-free an additionally two years, given that it has already survived disease-free 3 years after surgery is 0.91, whereas when considering a static 5-year DFS estimate at the time of surgery is only 0.71.

Paik *et al.* (2017) analysed 3-year conditional disease-free survival (CDFS) in patients with breast cancer. In this study, overall 3-year CDFS presented a gradual decrease over time. Hence, for patients with Luminal A (HR+/HER2-) and Luminal B (HR+/HER2+) subtype, 3-year CDFS decreased continuously, whereas for patients with HER2+ (HR-/HER2+) and triple negative (HR-/HER2-) subtype, 3-year CDFS tended to increase continuously before year 4 and decrease at year 5. In fact, this is in accordance with our findings. In our cohort, 2-year and 5-year CDFS also presented a gradual decrease over time. Furthermore, our

results obtained by computing a subgroup analyses stratified by IHC subtype are very similar. We observe a decrease over time in 2-year and 5-year CDFS for patients with HR+/HER2- and HR+/HER2+ subtype and an increase for patients with HR-/HER2+ and HR-/HER2- subtype, in line with what was found previously. Maaren *et al.* (2018) analysed 10-year conditional overall survival (COS) for patients with breast cancer and again showed that differences between breast cancer IHC subtypes became smaller. In agreement with what we found, the OS of patients with triple negative subtype (HR-/HER2-) was similar to the OS of patients with Luminal A (HR+/HER2-) subtype at diagnosis, as time goes by. In fact, it is usually recognized that patients with poor prognostic features at the time of diagnosis or surgery present greater increases in conditional survival, when compared with those without these features. Kim *et al.* (2015) demonstrated that patients with gastric cancer at higher risk at baseline showed the greatest increases in conditional survival over time. Our current analysis indicate that this is also true for breast cancer. More concretely, our results showed that conditional survival increased with increasing time of survival from diagnosis and/or surgery in breast cancer patients with higher stage of disease, high histological grade and HR-/HER2- IHC subtype. Such findings are reasonably in accordance with what was found with the landmarking approach. Through a backward elimination procedure based on Wald tests, age, disease stage and IHC subtypes were the variables that presented statistically significant interaction with time, when considering a window of 5 years. When we decrease the time window for 2 years, we observed that not only age, disease stage and IHC subtype presented statistically significant interaction with time, but also the lymph node status prognostic factor. However, when evaluating conditional survival, lymph node status seems to be constant over time. Performance of both landmark supermodels were evaluated using the landmark models fitted for each landmark time and reported a small range of Brier Score and reasonably good c-index estimates.

We have defined conditional survival in a straightforward manner by using the fact that the patient is alive and disease-free at prediction time s as the conditioning event. This approach is used, for example, by Zamboni (2010) when determining conditional survival for patients with colon cancer. Modern statistical methodology (parametric, nonparametric, and regression models) can be used to estimate and analyse conditional survival. In this study, we have presented the simplest approach, through Kaplan-Meier estimates (in the whole patient cohort or in strata defined by baseline characteristics). Analysing conditional survival based on Kaplan-Meier estimates have the major advantage of not requiring any additional data, unjustified assumptions, or specialized methods. Conditional survival probabilities could also be derived, for example, from conditional versions of Cox regression models. We would like to emphasize that, as usual in the analysis of survival data, the application of regression models is based on specific assumptions. For the Cox regression model, this is the proportional hazards assumption, which should be routinely checked. However, within the context of conditional survival, fitting a model at each prediction time point relaxes the proportional hazards assumption and allows the effect of covariates to vary with time.

We also used the landmarking method to analyse dynamic predictions. In this context, landmarking is useful in the presence of covariates when either their values or their effects change over time. The clear advantages of landmarking are simplicity and transparency. It is easy to see what is happening, especially when time-varying covariates are categorical

because the resulting analysis is a relatively simple group comparison. When dealing with time-dependent covariates, a time-dependent Cox regression analysis will typically be more efficient and an additional advantage of a time-dependent Cox regression is that no subjective and arbitrary choices are needed for the landmark time points. However, landmarking also presents an important role in this context, in which complex modelling of time-dependent covariates is avoided. Disadvantages of landmarking are the need for a choice of landmark time point, and also a loss of power, especially for later landmark time points, because subjects with an event before the landmark time point are excluded from analysis. An alternative approach uses multi-state models. This approach typically consists of defining the different states in the model, estimating transition intensities between the states, incorporating covariates and therefore compute dynamic prediction probabilities of interest. It has some gains in comparison with landmarking: well-developed theory and existence of various software. However, when the Markov assumption is not met, dynamic prediction probabilities through the multistate approach could not be accurate. In such cases landmarking comes with a number of advantages since it avoids models for the transition hazards and uses sparser models. Another huge advantage of landmarking is its robustness against Cox proportional hazards assumption. When assumptions of a multistate model are violated, prediction probabilities may be irrelevant and misleading. Moreover, while in landmarking approach it is easy to incorporate any information about the patients history, multistate models can only be used to obtain predictions given the current state in the model. Comparison of both methodologies can be found in van Houwelingen & Putter (2008) and Parast, Cheng & Cai (2011).

Conditional survival constitutes the simplest form of dynamic prediction and is perhaps of great interest to patients, clinicians and researchers. It is of great importance to patients to know about their current prognosis, and therefore an accurate risk assessment that accounts for time already survived should be provided. In fact, holding a more realistic quantification of their prognosis over time may be of a large benefit both psychologically and emotionally. However, this information should be passed onto patients in a very clear way. In practical terms, 2-year and 5-year conditional survival is a simply understandable measure that can be used to convey a patient's current risk profile. Besides patients, clinicians (e.g. surgeons and oncologists) can also make use of the advantages of the conditional survival measure. Adoption of such measure can help them to better predict survival, make the most appropriate treatment decisions, and conduct a more fully informed discussion with patients in light of their survival expectancy or prognosis. For instance, a more evidence-based approach can be implemented to improve surveillance plans based on the changing risk of the patient. Frequency of follow-up visits are often reduced after 2 or 3 years. However, this is done without evidence to support the practice. Determination of the most favourable testing frequency and duration should be based on a dynamic risk assessment rather than a static one usually performed many time ago. For instance, conditional disease-free survival data presented here indicates that the risk of recurrence or death for patients with HR-/HER2- breast cancer who survive without disease more than 4 years from surgery is comparable to the patients with HR+/HER2- disease. For this reason, these data suggest that, after completing 4 years after surgery, an HR-/HER2- patient could switch to a surveillance plan similar to HR+/HER2- patients (the group with better prognosis at baseline). Baade, Youlden & Chambers (2011) recommended that knowledge on conditional survival estimates should be incorporated in routine statistical

reporting, as these estimates provide more accurate information for patients who survived several years after their diagnosis.

There are some limitations to our study. Although we enrolled relatively large numbers of patients, which is one strength of this study, selection bias might be occurred, due to the study's retrospective nature. Hence, it was not possible to obtain information regarding socioeconomic or clinical variables (e.g. presence of comorbidities) that might be confounding factors or related with other prognostic factors. Another reason why selection bias might be occurred is that we only used data from a single center (IPOLFG), even though the breast cancer being treated in other hospitals. Despite the treatment set being part of the recorded variables, the information contained was very limited so we did not attempt a subanalysis adjusted by treatment. Furthermore, for some prediction time points, the sample size for some subgroups is small, as reflected in the larger confidence intervals found in those subgroups. Such challenge does not allow us to make more definite generalizations regarding the observed differences between some subgroups. Despite the limited follow-up of the cohort (7 years), data collection reflects current practices in oncology, particularly in what regards treatment of HER2+ patients with trastuzumab, making this study very relevant and appealing.

In summary, we have developed novel approaches for dynamic prediction of survival for women with breast cancer. Our finding should be confirmed using a population-based sample, including additional relevant prognostic variables and using a longer follow-up. Altogether, this should give further information on late recurrences and its relationship with prognostic subgroups. Considering the statistical methodology applied, strategies to evaluate predictive performances of landmark supermodels should be developed. Besides, further work involves constructing of a predictive model. From this, individual dynamic predictions can be derived, for instance, by showing the trajectory, or the probability of survival, of a patient with specific features at diagnosis.

6
Conclusion

We showed that conditional survival improves over time for patients with stage III, high histological grade and HR-/HER2- subtype. In contrast, this pattern was not observed within the worst prognostic subgroups defined according to lymph node status. One important result is that the estimated survival of a patient with HR-/HER2- tended to be similar to that of a patient with a favourable prognosis (HR+/HER2- or HR+/HER2-), as more time elapses. Results obtained with conditional survival approach were fairly confirmed with landmark analysis which showed a decrease over time of the prognostic significance of IHC groups and stage at diagnosis but not for the other evaluated variables, in a time window of 5 years. The adoption of conditional survival will help clinicians to better predict survival, make the most appropriate treatment and surveillance decisions and conduct a more fully informed discussion with patients in light of their survival expectancy and prognosis. Updated prognosis is particularly relevant to chronic diseases, as clinicians are asked to repeatedly assess the survival for a given patient. In fact, most patients experience a fear of recurrence, which is one reason why individual tailored follow-up plans are needed instead of traditional ones. Identify whether some patients might need continuous close surveillance whereas, in other patients, it might be possible to extend the intervals between surveillance tests is very important. As such, the results of our study suggests minor changes in guidelines for breast cancer surveillance. Hence, we encourage that the surveillance plan of a patient with HR-/HER2- subtype who has survived disease-free 4 years after surgery to be similar to the surveillance plan of a patient with HR+/HER2-, which is the group of better prognosis at baseline. Regarding novelty, our study may be the first to explicitly assign probabilities and apply conditional survival modelling of clinical outcomes within the context of breast cancer, in Portugal. Although the limitations of landmarking method, it appears to be a promissive statistical method for analysing dynamic changes in conditional survival. Notwithstanding the feasibility of our study, further external validation with longer follow-up is necessary to enable implementation in clinical practice.

References

Alexander F., Anderson, T., Brown, H., Forrest A., Hepburn W., Kirkpatrick, A., Muir, B., Prescott, R. & Smith, A. (1999). 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. *Lancet*, 353(9168), 1903-1908. doi: 10.1016/s0140-6736(98)07413-3.

American Cancer Society. (n.d). *Treating Breast Cancer*. Retrieved from: https://www.cancer.org/cancer/breast-cancer/treatment.html.

Anderson, J., Cain, K. & Gelber, R. (1983). Analysis of survival by tumor response. *Journal of Clinical Oncology*, 1(11), 710-719. doi:10.1200/JCO.1983.1.11.710.

Baade, P., Youlden, D. & Chambers, S. (2011). When do I know I am cured? Using conditional estimates to provide better information about cancer survival prospects. *The Medical Journal of Australia*, 194(2), 73-77.

Bastos, J., Barros, H. & Lunet, N. (2007). Breast cancer mortality trend in Portugal (1955-2002). *Acta Médica Portuguesa*, 20(2), 139-144.

Bellera, C., MacGrogan, G., Debled, M., de Lara, C., Brouste, V. & Mathoulin-Pélissier, S. (2010). Variables with time-varying effects and the Cox model: some statistical concepts illustrated with a prognostic factor study in breast cancer. *BMC Medical Research Methodology*, 10(20). doi: 10.1186/1471-2288-10-20.

Boutayeb, A. & Boutayeb, S. (2005). The burden of non communicable diseases in developing countries. *International Journal for Equity in Health 2005*, 4(2). doi:10.1186/1475-9276-4-2.

Breslow, N. (1974). Covariance analysis of censored survival data. *Biometrics*, 30(1), 89-99. doi:10.2307/2529620.

Breslow, N. & Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics*, 2(3), 437-453.

Brown, M., Tsodikov, A., Bauer, K., Parise, C. & Caggiano, V. (2008). The role of human epidermal growth factor receptor 2 in the survival of women with estrogen and progesterone receptor-negative, invasive breast cancer: the California Cancer Registry, 1999-2004. *Cancer*, 112(4), 737-747. doi: 10.1002/cncr.23243.

Cancer.Net. (2018). *Breast Cancer: Risk Factors and Prevention.* Retrieved from: https://www.cancer.net/cancer-types/breast-cancer/risk-factors-and-prevention.

Carter, C., Allen, C. & Henson, D. (1989). Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases. *Cancer*, 63(1), 181-187. doi: 10.1097/00006534-198911000-00055.

Chia, S., Norris, B., Speers, C., Cheang, M., Gilks, B., Gown, A., Hutsman, D., Olivotto, I., Nielsen, T. & Gelmon, K. (2008). Human epidermal growth factor receptor 2 overexpression as a prognostic factor in a large tissue microarray series of node-negative breast cancers. *Journal of Clinical Oncology*, 26(35), 5697-5704. doi: 10.1200/JCO.2007.15.8659.

Cianfrocca, M. & Goldstein, L. (2004). Prognostic and predictive factors in early-stage breast cancer. *The Oncologist*, 9(6), 606-616. doi: 10.1634/theoncologist.9-6-606.

Clark, G. (1995). Prognostic and predictive factors for breast cancer. *Breast Cancer*, 2(2), 79-89. doi: 10.1007/BF02966945.

Collett, D. (2015). *Modelling Survival Data in Medical Research.* (3rd ed.). Bristol: Chapman and Hall.

Cooper, G. & Hausman, R. (2007). *The Cell: A Molecular Approach.* (4th ed.). Washington, D.C.: ASM Press.

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

Cox, D. (1975). Partial likelihood. *Biometrika*, 62(2), 269-276. doi:10.1093/biomet/62.2.269.

Curigliano, G., Viale, G., Bagnardi, V., Fumagalli, L., Locatelli, M., Rotmensz, N., Ghisini, R., Colleoni, M., Munzone, E., Veronesi, P., Zurrida, S. Nolè, F. & Goldhirsch, A. (2009). Clinical relevance of HER2 overexpression/amplification in patients with small tumor size and node-negative breast cancer.*Jornal of Clinical Oncology*, 27(34), 5693-5699. doi: 10.1200/JCO.2009.22.0962.

Dieterich, M., Stubert, J., Reimer, T., Erickson, N. & Berling, A. (2014). Influence of lifestyle factors on breast cancer risk. *BreastCare*, 9(6), 407-414. doi: 10.1159/000369571.

Dipiro, J., Talbert, R., Yee, G., Matzke, G., Wells, B. & Posey, M. (2014). *Pharmacotherapy: a Pathophysiologic Approach.* (9th ed.). New York: McGraw-Hill Medical.

Ellis, I., Galea, M., Broughton, N., Locker, A., Blamey, R. & Elston, C. (1992). Pathological prognostic factors in breast cancer. II. Histological type. Relationship with survival in a large study with long-term follow-up. *Histopathology*, 20(6), 479-489. doi: 10.1111/j.1365-2559.1992.tb01032.x.

Elston, C. & Ellis, I. (1991). 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*, 19(5), 403-410.

Elston, C., Ellis, I. & Pinder, S. (1998). Prognostic factors in invasive carcinoma of the breast. *Clinical Oncology*, 10(1), 14-17. doi: 10.1016/S0936-6555(98)80105-2.

Ferlay, J., Soerjomataram, I., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D. & Bray, F. (2015). Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International Journal of Cancer*, 136(5), E359-E836. doi: 10.1002/ijc.29210.

Gastrin, F., Millher, A., Aronson, K., Wall, C., Hakama, M., Louhivuori, K. & Pukkala, E. (1994). Incidence and mortality from breast cancer in the Mama Program for Breast Screening in Finland, 1973-1986. *Cancer*, 73(8), 2168-2174.

Ghoncheh, M., Pournamdar, Z. & Salehiniya, H. (2016). Incidence and mortality and epidemiology of breast cancer in the world. *Asian Pacific Journal of Cancer Prevention*, 17(S3), 43-46. doi: 10.7314/APJCP.2016.17.S3.43.

Graf, E., Schomoor, C., Sauerbrei, W. & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18), 2529-2545.

Grambsch, P. & Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3), 515-526. doi: 10.2307/2337547.

Harrell, F., Lee, K. & Mark, D. (1996). Multivariable prognostic models: Issues in developing models, evaluating assumptions and adequacy, and measuring and reducing errors. *Statistics in Medicine*, 15(4), 361-387.

Heinze, G., Wallisch, C. & Dunkler, D. (2017). Variable selection - A review and recommendations for the practicing statistician. *Biometrical Journal*, 60(3), 431-449. doi: 10.1002/bimj.201700067.

Heike, S., Kleber, M., König, C., Engelhardt, M. & Schumacher, M. (2015). Conditional survival: a useful concept to provide information on how prognosis evolves over time. *Clinical Cancer Research*, 21(7), 1530-1536. doi:10.1158/1078-0432.CCR-14-2154.

Iwamoto, T. & Pusztai, L. (2010). Predicting prognosis of breast cancer with gene signatures: are we lost in a sea of data? *Genome Medicine*, 2(81). doi: 10.1186/gm2021.

Janssen-Heijnen, M., Gondos, A., Bray, F., Hakulinen, T., Brewster, D., Brenner, H. & Coebergh, J. (2010). Clinical relevance of conditional survival of cancer patients in Europe: age-specific analyses of 13 cancers. *Journal of Clinical Oncology*, 20(15), 2520-2528. doi: 10.1200/JCO.2009.25.969.

Kaplan, E. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282), 457-481.

Kim, Y., Ejaz, A., Spolverato, G., Squires, M., Poultsides, G. & Fields, R. (2015). Conditional survival after surgical resection of gastric cancer: a multi-institutional analysis of the US gastric cancer collaborative. *Annals of Surgical Oncology*, 22(2), 557-564. doi: 10.1245/s10434-014-4116-5.

Klein, J., van Houwelingen, H., Ibrahim, J. & Scheike, T. (2014). *Handbook of Survival Analysis*. (1st ed.). Boca Raton: CRC Press.

Liga Portuguesa Contra o Cancro. (n.d). *Factores de risco*. Retrieved from: https://www.ligacontracancro.pt/cancro-da-mama-factores-de-risco/.

Lin, D. & Wei, L. (1989). The robust inference for the cox proportional hazards model. *Journal of the American Statistical Association*, 84(408), 1074-1078. doi:10.2307/2290085.

Louie, M. & Sevigny, M. (2017). Steroid hormone receptors as prognostic markers in breast cancer. *American Journal of Cancer Research*, 7(8), 1617-1636.

Maaren, M., Strobbe, L., Smidt, M., Moossdorff, M., Poortmans, P. & Siesling, S. (2018). Ten-year conditional recurrence risks and overall and relative survival for breast cancer patients in the Netherlands: Taking account of event-free years. *European Journal of Cancer*, 102, 82-94. doi: 10.1016/j.ejca.2018.07.124.

Mantel, N. & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22(4), 719-748.

Meier, P. (1975). Estimation of a distribution function from incomplete observations. In: *Perspectives in Probability and Statistics* (pp 67-87). London: Academic Press.

Ménard, S., Stefania, F., Castiglioni, F., Agresti, R. & Balsari, A. (2001). HER2 as a prognostic factor in breast cancer. *Oncology*, 61(2), 67-72. doi: 10.1159/000055404.

Moss, R., Hansen, C., Sanders, R., Hawley, S. & Steigbigel, R. (2012). A phase II study of DAS181, a novel host directed antiviral for the treatment of influenza infection. *The Journal of Infectious Diseases*, 206(12), 1844-1851. doi: 10.1093/infdis/jis622.

National Breast Cancer Foundation. (n.d). *Breast anatomy and how cancer starts.* Retrieved from: https://nbcf.org.au/about-national-breast-cancer-foundation/about-breast-cancer/what-you-need-to-know/breast-anatomy-cancer-starts/.

Paik, H., Lee, S., Ryu, J., Park, S., Kim, I., Bae, S., Yu, J., Lee, J., Kim, S. & Nam, S. (2017). Conditional disease-free survival among patients with breast cancer. *Medicine*, 96(1). doi: 10.1097/MD.0000000000005746.

Parast, L., Cheng, S. & Cai, T. (2011). Incorporating short-term outcome information to predict long-term survival with discrete markers. *Biometrical Journal*, 53(2), 294-307. doi: 10.1002/bimj.201000150.

Peto, R. & Peto, J. (1972). Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society*, 135(2), 185-207. doi: 10.2307/2344317.

Raica, M., Jung, I., Cîmpean, A. Suciu, C. & Muresan, A. (2009). From conventional pathologic diagnosis to the molecular classification of breast carcinoma: are we ready for the change?. *Romanian Journal of Morphology and Embryology*, 50(1), 5-13.

Rakha, E., Reis-Filho, J., Baehner, F., Dabbs, D., Decker, T., Eusebi, V., Fox, S., Ichihara, S., Jacquemier, J., Lakhani, S., Palacios, J., Richardson, A., Schnitt, St., Schmitts, F., Tan, P. Tse, G., Badve, S. & Ellis, I. (2010). Breast cancer prognostic classification in the molecular era: the role of histological grade. *Breast Cancer Research*, 12(4). doi: 10.1186/bcr2607.

Ross, J., Slodkowska, E., Symmans, W., Pusztai, L., Ravdin, P. & Hortobagyi, G. (2009). The HER-2 receptor and breast cancer: ten years of targeted anti-HER-2 therapy and personalized medicine. *The Oncologist*, 14, 320-368. doi: 10.1634/theoncologist.2008-0230.

Schoenfeld, D. (1982). Partial residuals for the proportional hazards regression model. *Biometrika*, 69(1), 239-241. doi: 10.2307/2335876.

Sobin, L., Gospodarowicz, M. & Wittekin, C. (2009). *TNM Classification of Malignant Tumours.* (7th edition). England: Wiley-Blackwell.

Spitale, A., Mazzola, P., Soldini, D., Mazzucchelli, L. & Bordoni, A. (2009). Breast cancer classification according to immunohistochemical markers: clinicopathological features and short-term survival analysis in a population-based study from the South of Switzerland *Annals of Oncology*, 20(4), 628-635. doi: 10.1093/annonc/mdn675.

Stickeler, E. (2011). Prognostic and predictive markers for treatment decisions in early breast cancer. *Breast Care*, 6(3), 193-198. doi: 10.1159/000329471.

Strasser-Weippl, K., Horick, N., Smith, I., O'Shaughnessy, J., Ejlertsen, B., Boyle, F., Buzdar, A., Fumoleau, P., Gradishar, W., Martin, M., Moy, B., Piccat-Gebhart, M., Pritchard, K., Lindquist, D., Rappold, E., Finkelstein, D. & Goss, P. (2015). Long-term hazard of recurrence in HER2+ breast cancer patients untreated with anti-HER2 therapy. *Breast Cancer Research*, 17(56). doi: 10.1186/s13058-015-0568-1.

Tao, Z., Shi, A., Lu, C., Song, T., Zhang, Z. & Zhao, J. (2015). Breast cancer: epidemiology and etiology. *Cell Biochemistry and Biophysics*, 72(2), 333-338. doi: 10.1007/s12013-014-0459-6.

Therneau, T. & Grambsch, P. (2000). *Modelling Survival Data: Extending the Cox Model*. (1st ed.). New York: Springer.

Vahabi, M., Lofters, A., Kumar, M. & Glazier, R. (2015). Breast cancer screening disparities among urban immigrants: a population-based study in Ontario, Canada. *BMC Public Health*, 15(679). doi 10.1186/s12889-015-2050-5.

van Houwelingen, H. (2007). Dynamic prediction by landmarking in event history analysis. *Scandinavian Journal of Statistics*, 34(1), 70-84. doi: 10.1111/j.1467-9469.2006.00529.x.

van Houwelingen, H. & Putter, H. (2008). Dynamic predicting by landmarking as an alternative for multi-state modeling: an application to acute lymphoid leukemia data. *Lifetime Data Analysis*, 14(4), 447-463. doi: 10.1007/s10985-008-9099-8.

van Houwelingen, H. & Putter, H. (2011). *Dynamic Prediction in Clinical Survival Analysis*. (1st ed.) . Boca Raton: CRC Press.

Veronesi, U., Luini, A., Del Vecchio, M., Greco, M., Galimberti, V., Merson, M., Rilke, F., Sacchini, V. & Saccozzi, R. (1993). Radiotherapy after breast-preserving surgery in women with localized cancer of the breast. *New England Journal of Medicine*, 328(22), 1587-1591. doi: 10.1056/NEJM199306033282202.

WHO. (2015). *Cancer*. Retrieved from: https://www.afro.who.int/health-topics/cancer.

Wolff, A., Hammond, M., Allison, K., Harvey, B., Mangu, P., Barlett, J., Bilous, M., Ellis, I., Fitzgibbons, P., Hanna, W., Jenkins, R., Press, M., Spears, P., Vance, G., Viale, G., McShane, L. & Dowsett, M. (2018). Human Epidermal Growth Factor Receptor 2 Testing in Breast Cancer: American Society of Clinical Oncology/College of American Pathologists Clinical Practice Guideline Focused Update. *Archives of Pathology & Laboratory Medicine*, 142(11), 1364-1382. doi: 10.5858/arpa.2018-0902-SA.

Xu, R. & O'Quigley, J. (2001). Estimating average regression effect under nonproportional hazards. *Biostatistics*, 1(4), 423-439.

Yanagawa, M., Ikemot, K., Kawauchi, S., Furuya, T., Yamamoto, S., Oka, M., Oga, A., Nagashima, Y. & Sasaki, K. (2012). Luminal A and luminal B (HER2 negative) subtypes of breast cancer consist of a mixture of tumors with different genotype. *BMC Research Notes*, 5(376). doi: 10.1186/1756-0500-5-376

Zamboni, B., Yothers, G., Choi, M., Fuller, C., Dignam, J., Raich, P., Thomas, C., O'Connell, M., Wolmark, N. & Wan, S. (2010). Conditional survival and the choice of conditioning set for patients with colon cancer: an analysis of NSABP trials C-03 through C-07. *Journal of Clinical Oncology*, 28(15), 2544-2548. doi: 10.1200/JCO.2009.23.0573.

Zimmerman, B. (2004). *Understanding Breast Cancer Genetics*. (Doctoral dissertation). University Press of Mississippi, USA.

A

Derivation of results from section 3.4.2.1

In this chapter we will provide some derivations of the results presented in section 3.4.2.1 of this thesis. In all of what follows, we assume that we have right censored survival data, where the hazard function can be described as

$$h(t) = h_0(t)\exp(\boldsymbol{x}^\top \boldsymbol{\beta})$$

We will assume that the covariates are centered, and that the survival and censoring times are independent given the covariates.

A.1 Derivation of equation 3.40

For the following we impose the condition that

$$\int_0^t h_0(s)(\boldsymbol{x}^\top(\boldsymbol{\beta}(s) - \bar{\boldsymbol{\beta}}(t))^2 ds$$

is small, where

$$\bar{\boldsymbol{\beta}}(t) = \frac{\int_0^t h_0(s)\boldsymbol{\beta}(s)ds}{H_0(t)}$$

which requires that $\boldsymbol{x}\ \boldsymbol{\beta}(s)$ is small and does not vary to much. Using a Taylor expansion of $\exp(\boldsymbol{x}^\top\boldsymbol{\beta}(s))$ around the point $\exp(\boldsymbol{x}^\top\boldsymbol{\beta}(s))$, we can write it as

$$\exp(\boldsymbol{x}^\top\boldsymbol{\beta}(s)) = \exp(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t)) + \exp(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t))\boldsymbol{x}^\top(\boldsymbol{\beta}(s) - \bar{\boldsymbol{\beta}}(t)) + \frac{e^c}{2}(\boldsymbol{x}^\top(\boldsymbol{\beta}(s) - \bar{\boldsymbol{\beta}}(t)))^2$$

where c lies between $\boldsymbol{x}\bar{\boldsymbol{\beta}}(t)$ and $\boldsymbol{x}\boldsymbol{\beta}(\boldsymbol{s})$. Multiplying both sides of this expression with $h_0(s)$ and integrating from 0 to t, we see that

$$\int_0^t h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}(s))ds = \exp(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t))\int_0^t h_0(s)ds$$
$$+\exp(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t))\left(\int_0^t h_0(s)\boldsymbol{x}^\top\boldsymbol{\beta}(s)ds - \boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t)\int_0^t h_0(s)ds\right)$$
$$+\frac{e^c}{2}\int_0^t h_0(s)(\boldsymbol{x}^\top(\boldsymbol{\beta}(s)-\bar{\boldsymbol{\beta}}(t))^2ds$$

Since

$$\exp(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t))\left(\int_0^t h_0(s)\boldsymbol{x}^\top\boldsymbol{\beta}(s)ds - \boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t)\int_0^t h_0(s)ds\right) = 0$$

and

$$\int_0^t h_0(s)(\boldsymbol{x}^\top(\boldsymbol{\beta}(s)-\bar{\boldsymbol{\beta}}(t)))^2ds$$

is small, we have that

$$H(t|\boldsymbol{x}) = \int_0^t h_0(t)\exp\left(\boldsymbol{x}^\top\boldsymbol{\beta}(s)\right) \approx H_0(t)\exp\left(\boldsymbol{x}^\top\bar{\boldsymbol{\beta}}(t)\right)$$

A.2 Derivation of equation 3.44

If we fit a Cox proportional hazard model with administrative censoring at some horizon t_{hor}, when the hazard can be described as:

$$h(t) = h_0(t)\exp(\boldsymbol{x}^\top\boldsymbol{\beta}),$$

then (van Houwelingen & Putter, 2011) the estimate converges to a limiting value approximately given by

$$\tilde{\boldsymbol{\beta}}_{Cox} \approx \left(\int_0^{t_{hor}} S(t)C(t)h(t)\text{var}(\boldsymbol{X}|T=t)dt\right)^{-1}$$
$$\cdot\int_0^{t_{hor}} S(t)C(t)h(t)\text{var}(\boldsymbol{X}|T=t)\boldsymbol{\beta}(t)dt$$

given that the true coefficients $\boldsymbol{\beta}(t)$ do not vary too much over time. Here $S(t)$, $C(t)$ and $h(t)$ are the marginal, censoring and hazard functions, respectively. $\text{var}(\boldsymbol{X}|T=t)$ is defined as the limiting value of

$$\frac{S^{(2)}(\boldsymbol{\beta}(t),t)}{S^{(0)}(\boldsymbol{\beta}(t),t)} - \left(\frac{S^{(1)}(\boldsymbol{\beta}(t),t)}{S^{(0)}(\boldsymbol{\beta}(t),t)}\right)\left(\frac{S^{(1)}(\boldsymbol{\beta}(t),t)}{S^{(0)}(\boldsymbol{\beta}(t),t)}\right)^\top$$

where

$$S^{(0)}(\boldsymbol{\beta}(t),t)=\sum_{i=1}^{n}Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}(t))$$

$$S^{(1)}(\boldsymbol{\beta}(t),t)=\sum_{i=1}^{n}Y_i(t)\boldsymbol{x}_i\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}(t))$$

and

$$S^{(2)}(\boldsymbol{\beta}(t),t)=\sum_{i=1}^{n}Y_i(t)\boldsymbol{x}_i\boldsymbol{x}_i^\top\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta}(t))$$

By limiting value, we here mean the value which the expression above, as it were, approaches when the number of observations increases. Under the conditions that t_{hor}, and the effects of the covariates are small, $\mathrm{var}(\boldsymbol{X}|T=t)$ is approximately constant over the interval $[0,t_{hor}]$. Thus, under these conditions, we have that

$$\tilde{\boldsymbol{\beta}}_{Cox}\approx\frac{\int_0^{t_{hor}}S(s)C(s)h(s)\boldsymbol{\beta}(s)ds}{\int_0^{t_{hor}}S(s)C(s)h(s)ds}$$

Furthermore, if $C(t)\approx 1$, $S(t)\approx 1$ and $h(t)\propto h_0(t)$, then by (3.41) we have that

$$\tilde{\boldsymbol{\beta}}_{Cox}\approx\bar{\boldsymbol{\beta}}(t_{hor})$$

A.3 Derivation of equation 3.47

We will now argue that under some conditions, $H_{Cox}(t_{hor}|x)\approx H(t_{hor}|x)$. The most important of these conditions is that $\boldsymbol{x}^\top\boldsymbol{\beta}(t)$ does not vary too much. First we observe that for the Breslow estimator of the baseline hazard, we have

$$\frac{d\hat{H}_0(\boldsymbol{\beta}(t),t)}{d\hat{H}_0(\boldsymbol{\beta}(t),t)}=\frac{S^{(0)}(\boldsymbol{\beta},t)}{S^{(0)}(\boldsymbol{\beta}(t),t)}$$

for arbitrary $\boldsymbol{\beta}$, where

$$\hat{H}_0(\boldsymbol{\beta},t)=\sum_{t_i\le t}\frac{d_i}{\sum_{\ell\in R_i}\exp(\boldsymbol{x}_\ell^\top\boldsymbol{\beta})}=\sum_{t_i\le t}\frac{d_i}{S^{(0)}(\boldsymbol{\beta},t_i)}$$

By defining

$$\pi_i(\boldsymbol{\beta},t)=\frac{Y_i(t)\exp(\boldsymbol{x}_i^\top\boldsymbol{\beta})}{\sum_{j=1}^{n}Y_j\exp(\boldsymbol{x}_j^\top\boldsymbol{\beta})}$$

and writing

$$\frac{S^{(0)}(\boldsymbol{\beta},t)}{S^{(0)}(\boldsymbol{\beta}(t),t)}=\sum_{i=1}^{n}\exp(\boldsymbol{x}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t)))\pi_i(\boldsymbol{\beta},t)\tag{A.1}$$

we see that the Theorem I of Xu & O'Quigley (2001), this converges in probability to

$$E(\exp(\boldsymbol{X}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t)))|T=t)$$

given that we have random censoring. By making a Taylor expansion of

$$\exp(\boldsymbol{X}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t)))$$

around $E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t))$, and then taking the expectation conditioned on that $T=t$, on both sides, one can see that

$$E(\exp(\boldsymbol{X}^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t)))|T=t) \approx \exp\Big(E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}-\boldsymbol{\beta}(t))\Big)$$

provided that $\text{var}(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}(\boldsymbol{\beta}-\boldsymbol{\beta}(t))-\boldsymbol{\beta}(t))$ is small. Thus we get that

$$\frac{d\hat{H}_0(\tilde{\boldsymbol{\beta}}_{Cox},t)}{d\hat{H}_0(\boldsymbol{\beta}(t),t)} \approx \exp\Big(E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}_{Cox})\Big),$$

and therefore

$$h_{0,Cox}(t) \approx h_0(t)\exp\Big(E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}_{Cox})\Big)$$

From this it can be argued that:

$$\begin{aligned}
H_{Cox}(t_{hor}|\boldsymbol{x}) &= \exp(\boldsymbol{x}^\top\tilde{\boldsymbol{\beta}})\int_0^{t_{hor}} h_{Cox,0}(t)dt \\
&\approx \exp(\boldsymbol{x}^\top\tilde{\boldsymbol{\beta}})\int_0^{t_{hor}} h_0(t)\exp\Big(E(\boldsymbol{X}|T=t)^\top(\boldsymbol{\beta}(t)-\tilde{\boldsymbol{\beta}}_{Cox})\Big)dt \\
&= \int_0^{t_{hor}} \exp\Big(\boldsymbol{x}^\top\boldsymbol{\beta}(t)+(\boldsymbol{x}-E(\boldsymbol{X}|T=t))^\top(\tilde{\boldsymbol{\beta}}_{Cox}-\boldsymbol{\beta}(t))\Big)dt \\
&\approx \int_0^{t_{hor}} h_0(t)\exp\Big(\boldsymbol{x}^\top\boldsymbol{\beta}(t)\Big) \\
&= H(t_{hor}|\boldsymbol{x})
\end{aligned}$$

B
Cox models results

B.1 Univariable models

Table B.1: Regression parameter estimates from the univariable Cox proportional hazards model

Covariate	Hazard ratio	p-value
Age	1.02	<0.001
Stage		
I	1	
II	1.87	<0.001
III	4.18	<0.001
Tumour grade		
Low	1	
Intermediate	1.39	<0.001
High	1.89	<0.001
Lymph node status		
Negative	1	
Positive	2.02	<0.001
IHC group		
HR+/HER2-	1	
HR+/HER2+	1.12	0.25
HR-/HER2+	1.86	<0.001
HR-/HER2-	1.71	<0.001

B.2 Multivariable model

Table B.2: Regression parameter estimates from the multivariable Cox proportional hazards model

Covariate	Hazard ratio	p-value
Age	1.03	<0.001
Stage		
I	1	
II	1.7	<0.001
III	3.77	<0.001
Tumour grade		
Low	1	
Intermediate	1.03	0.75
High	1.17	0.21
Lymph node status		
Negative	1	
Positive	1.11	0.21
IHC group		
HR+/HER2-	1	
HR+/HER2+	1.09	0.41
HR-/HER2+	1.34	0.03
HR-/HER2-	1.46	<0.001

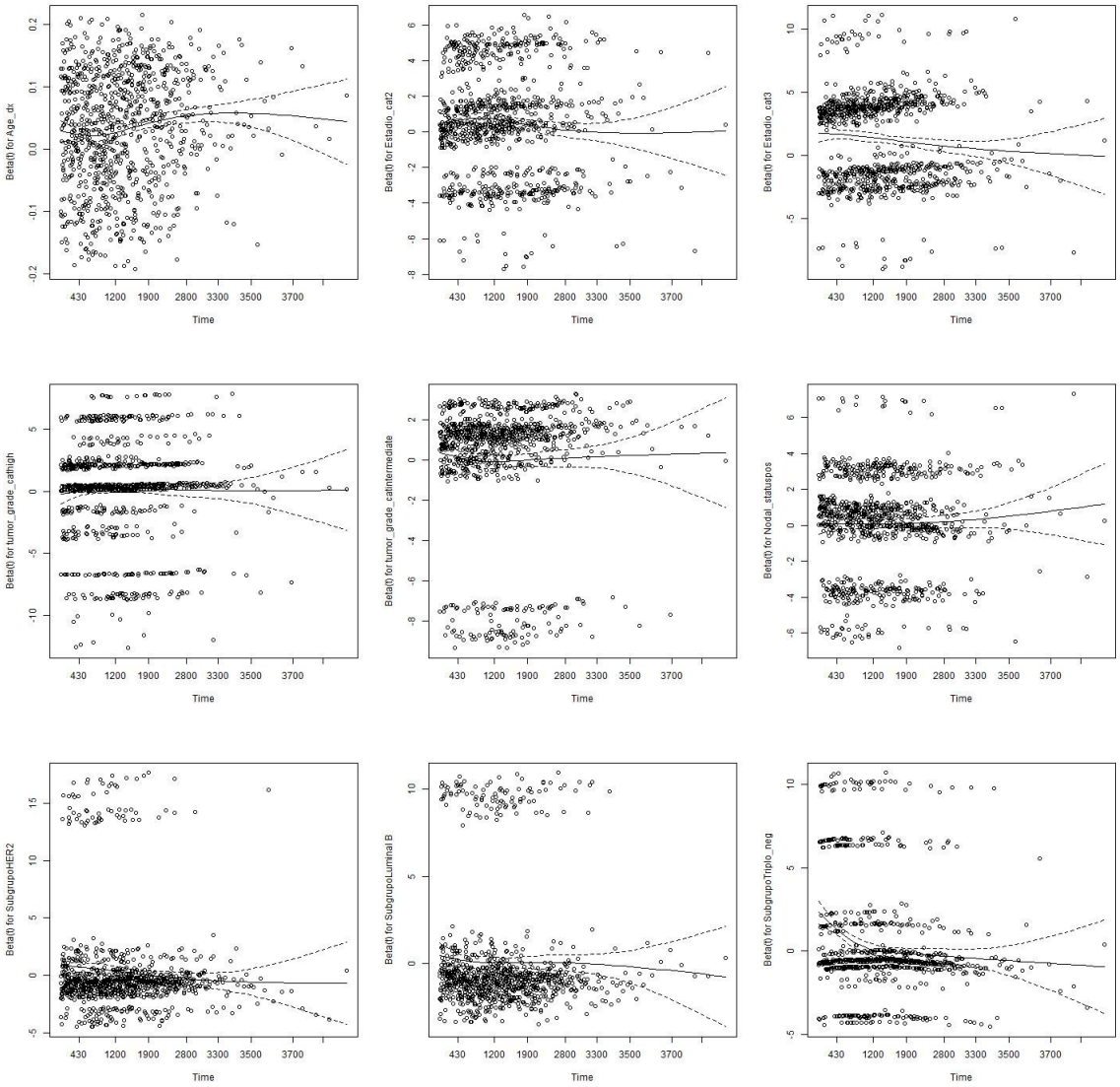

Figure B.1: Plot of the Schoenfeld residuals for the multivariable Cox proportional hazards model

C

Conditional overall survival and conditional disease-free survival estimates by prognostic factor

Table C.1: Conditional overall-survival estimates with the correspondent 95% CI for each prediction time point s, for both $w = 2$ and $w = 5$, stratified by prognostic factor

t=s+2												
Category	0		1		2		3		4		5	
Disease stage												
I	0.98	[0.98;0.99]	0.98	[0.97;0.98]	0.98	[0.97;0.98]	0.97	[0.96;0.98]	0.97	[0.96;0.98]	0.96	[0.94;0.97]
II	0.97	[0.96;0.97]	0.95	[0.95;0.96]	0.96	[0.95;0.97]	0.95	[0.94;0.96]	0.95	[0.96;0.96]	0.94	[0.96;0.96]
III	0.92	[0.89;0.94]	0.89	[0.87;0.91]	0.91	[0.88;0.93]	0.92	[0.89;0.94]	0.91	[0.88;0.93]	0.89	[0.86;0.93]
Tumour grade												
Low	0.98	[0.97;0.99]	0.97	[0.96;0.98]	0.97	[0.96;0.98]	0.95	[0.94;0.98]	0.96	[0.94;0.98]	0.95	[0.93;0.97]
Moderate	0.97	[0.96;0.98]	0.96	[0.95;0.97]	0.96	[0.96;0.97]	0.96	[0.95;0.97]	0.95	[0.94;0.96]	0.94	[0.93;0.96]
High	0.95	[0.94;0.97]	0.93	[0.91;0.94]	0.93	[0.91;0.95]	0.94	[0.92;0.96]	0.94	[0.92;0.96]	0.94	[0.92;0.96]
Lymph node status												
Negative	0.98	[0.97;0.98]	0.96	[0.96;0.97]	0.97	[0.96;0.97]	0.96	[0.96;0.97]	0.96	[0.95;0.97]	0.95	[0.93;0.96]
Positive	0.95	[0.94;0.96]	0.93	[0.92;0.94]	0.94	[0.93;0.95]	0.94	[0.93;0.95]	0.93	[0.92;0.95]	0.94	[0.92;0.95]
Immunohistochemistry subtype												
HR+/HER2-	0.97	[0.97;0.98]	0.96	[0.96;0.97]	0.96	[0.95;0.97]	0.96	[0.95;0.97]	0.95	[0.94;0.96]	0.94	[0.93;0.95]
HR+/HER2+	0.97	[0.97;0.98]	0.97	[0.95;0.99]	0.97	[0.95;0.98]	0.96	[0.95;0.98]	0.95	[0.93;0.98]	0.95	[0.92;0.98]
HR-/HER2+	0.96	[0.94;0.99]	0.93	[0.89;0.96]	0.92	[0.88;0.95]	0.94	[0.90;0.97]	0.97	[0.94;0.99]	0.96	[0.92;0.99]
HR-/HER2-	0.88	[0.86;0.91]	0.86	[0.83;0.89]	0.94	[0.91;0.96]	0.93	[0.91;0.96]	0.95	[0.93;0.98]	0.97	[0.95;0.99]
t=s+5												
Category	0		1		2		3		4		5	
Disease stage												
I	0.94	[0.93;0.95]	0.93	[0.92;0.94]	0.92	[0.90;0.93]	0.89	[0.88;0.92]	0.87	[0.84;0.89]	0.78	[0.72;0.85]
II	0.88	[0.87;0.89]	0.86	[0.84;0.87]	0.86	[0.84;0.88]	0.84	[0.82;0.87]	0.83	[0.79;0.86]	0.69	[0.61;0.79]
III	0.73	[0.70;0.76]	0.69	[0.66;0.73]	0.69	[0.66;0.74]	0.70	[0.66;0.75]	0.67	[0.61;0.73]	0.65	[0.57;0.76]
Tumour grade												
Low	0.93	[0.91;0.95]	0.91	[0.89;0.93]	0.89	[0.86;0.92]	0.87	[0.84;0.90]	0.84	[0.80;0.88]	0.78	[0.68;0.89]
Moderate	0.89	[0.89;0.91]	0.87	[0.86;0.89]	0.86	[0.85;0.88]	0.84	[0.83;0.86]	0.83	[0.80;0.88]	0.70	[0.64;0.78]
High	0.82	[0.79;0.95]	0.81	[0.78;0.84]	0.82	[0.79;0.85]	0.83	[0.79;0.87]	0.81	[0.76;0.86]	0.79	[0.72;0.88]
Lymph node status												
Negative	0.91	[0.90;0.92]	0.89	[0.88;0.91]	0.88	[0.87;0.89]	0.87	[0.85;0.89]	0.85	[0.83;0.87]	0.76	[0.71;0.82]
Positive	0.82	[0.81;0.84]	0.79	[0.78;0.82]	0.81	[0.79;0.83]	0.79	[0.77;0.82]	0.78	[0.73;0.81]	0.68	[0.56;0.78]
Immunohistochemistry subtype												
HR+/HER2+	0.89	[0.89;0.91]	0.87	[0.86;0.89]	0.86	[0.85;0.88]	0.85	[0.83;0.86]	0.82	[0.79;0.85]	0.73	[0.68;0.79]
HR+/HER2-	0.89	[0.87;0.93]	0.87	[0.84;0.90]	0.86	[0.83;0.90]	0.86	[0.82;0.90]	0.83	[0.76;0.89]	0.78	[0.69;0.89]
HR-/HER2+	0.79	[0.74;0.85]	0.78	[0.72;0.83]	0.80	[0.74;0.86]	0.81	[0.74;0.89]	0.82	[0.74;0.92]	0.71	[0.45;1.00]
HR-/HER2-	0.75	[0.71;0.79]	0.77	[0.77;0.81]	0.84	[0.80;0.88]	0.84	[0.79;0.89]	0.84	[0.78;0.90]	0.77	[0.66;0.89]

Table C.2: Conditional disease-free survival estimates with the correspondent 95% CI for each prediction time point s, for both $w = 2$ and $w = 5$, stratified by prognostic factor

t=s+2												
Category	0		1		2		3		4		5	
Disease stage												
I	0.97	[0.96;0.98]	0.96	[0.96;0.97]	0.96	[0.95;0.97]	0.96	[0.95;0.97]	0.95	[0.94;0.96]	0.93	[0.92;0.95]
II	0.92	[0.91;0.94]	0.92	[0.90;0.93]	0.93	[0.92;0.94]	0.92	[0.90;0.93]	0.91	[0.89;0.92]	0.92	[0.89;0.94]
III	0.81	[0.78;0.84]	0.82	[0.79;0.85]	0.84	[0.81;0.86]	0.84	[0.81;0.87]	0.84	[0.81;0.88]	0.85	[0.82;0.89]
Tumour grade												
Low	0.97	[0.95;0.98]	0.96	[0.95;0.98]	0.94	[0.93;0.96]	0.94	[0.92;0.96]	0.94	[0.92;0.96]	0.93	[0.91;0.96]
Moderate	0.94	[0.93;0.95]	0.94	[0.93;0.94]	0.93	[0.92;0.94]	0.93	[0.92;0.94]	0.91	[0.90;0.93]	0.91	[0.89;0.93]
High	0.89	[0.87;0.91]	0.88	[0.86;0.89]	0.89	[0.88;0.92]	0.91	[0.89;0.93]	0.91	[0.88;0.94]	0.91	[0.88;0.94]
Lymph node status												
Negative	0.95	[0.94;0.96]	0.94	[0.93;0.95]	0.95	[0.94;0.96]	0.95	[0.94;0.96]	0.94	[0.93;0.95]	0.93	[0.91;0.94]
Positive	0.88	[0.87;0.89]	0.89	[0.87;0.90]	0.89	[0.87;0.90]	0.89	[0.87;0.90]	0.89	[0.87;0.91]	0.89	[0.87;0.92]
Immunohistochemistry subtype												
HR+/HER2+	0.94	[0.94;0.95]	0.93	[0.93;0.94]	0.94	[0.93;0.95]	0.93	[0.92;0.94]	0.91	[0.90;0.93]	0.91	[0.89;0.93]
HR+/HER2-	0.93	[0.90;0.95]	0.93	[0.90;0.95]	0.92	[0.89;0.94]	0.91	[0.89;0.94]	0.94	[0.91;0.96]	0.92	[0.88;0.96]
HR-/HER2+	0.84	[0.80;0.89]	0.84	[0.79;0.89]	0.86	[0.81;0.91]	0.89	[0.84;0.94]	0.93	[0.89;0.98]	0.93	[0.88;0.99]
HR-/HER2-	0.81	[0.78;0.85]	0.87	[0.83;0.90]	0.91	[0.88;0.94]	0.91	[0.88;0.95]	0.94	[0.82;0.98]	0.95	[0.92;0.99]
t=s+5												
Category	0		1		2		3		4		5	
Disease stage												
I	0.92	[0.90;0.93]	0.89	[0.88;0.81]	0.88	[0.86;0.89]	0.87	[0.65;0.89]	0.83	[0.80;0.86]	0.74	[0.68;0.81]
II	0.82	[0.80;0.84]	0.80	[0.78;0.82]	0.81	[0.79;0.83]	0.79	[0.77;0.82]	0.76	[0.72;0.81]	0.58	[0.46;0.74]
III	0.62	[0.59;0.66]	0.64	[0.61;0.68]	0.65	[0.61;0.69]	0.68	[0.63;0.73]	0.63	[0.57;0.71]	0.39	[0.21;0.70]
Tumour grade												
Low	0.89	[0.86;0.91]	0.87	[0.85;0.90]	0.86	[0.83;0.89]	0.85	[0.81;0.88]	0.83	[0.78;0.88]	0.70	[0.58;0.85]
Moderate	0.84	[0.83;0.86]	0.83	[0.81;0.84]	0.82	[0.80;0.84]	0.81	[0.79;0.83]	0.78	[0.75;0.81]	0.63	[0.55;0.72]
High	0.77	[0.74;0.79]	0.76	[0.73;0.79]	0.79	[0.75;0.92]	0.79	[0.75;0.83]	0.78	[0.72;0.84]	0.67	[0.52;0.87]
Lymph node status												
Negative	0.88	[0.86;0.89]	0.86	[0.85;0.88]	0.85	[0.84;0.87]	0.84	[0.82;0.86]	0.81	[0.78;0.84]	0.71	[0.63;0.78]
Positive	0.74	[0.72;0.76]	0.74	[0.72;0.76]	0.75	[0.73;0.78]	0.76	[0.73;0.79]	0.72	[0.68;0.77]	0.49	[0.36;0.66]
Immunohistochemistry subtype												
HR+/HER2+	0.85	[0.84;0.86]	0.83	[0.81;0.84]	0.82	[0.80;0.84]	0.82	[0.79;0.83]	0.78	[0.75;0.81]	0.63	[0.55;0.71]
HR+/HER2-	0.82	[0.79;0.86]	0.82	[0.78;0.86]	0.81	[0.77;0.86]	0.81	[0.76;0.86]	0.77	[0.69;0.85]	0.67	[0.50;0.89]
HR-/HER2+	0.68	[0.63;0.75]	0.73	[0.67;0.79]	0.76	[0.69;0.83]	0.81	[0.74;0.89]	0.81	[0.72;0.92]	0.64	[0.36;1.00]
HR-/HER2-	0.71	[0.66;0.75]	0.78	[0.74;0.82]	0.83	[0.79;0.88]	0.83	[0.77;0.88]	0.79	[0.73;0.88]	0.67	[0.48;0.93]

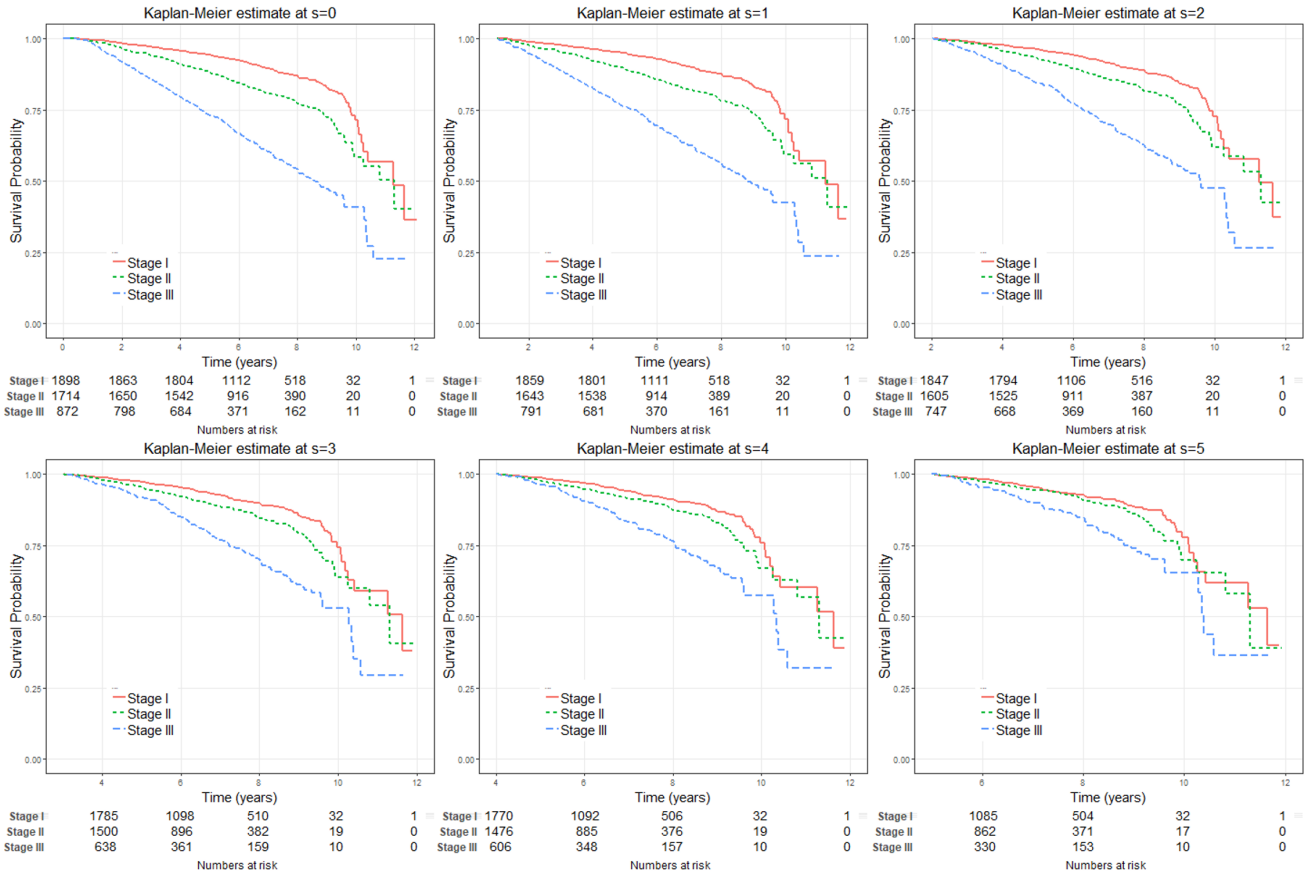

Figure C.1: Kaplan-Meier estimates of the survival function stratified by disease stage considering all individuals alive and disease-free s years after diagnosis.

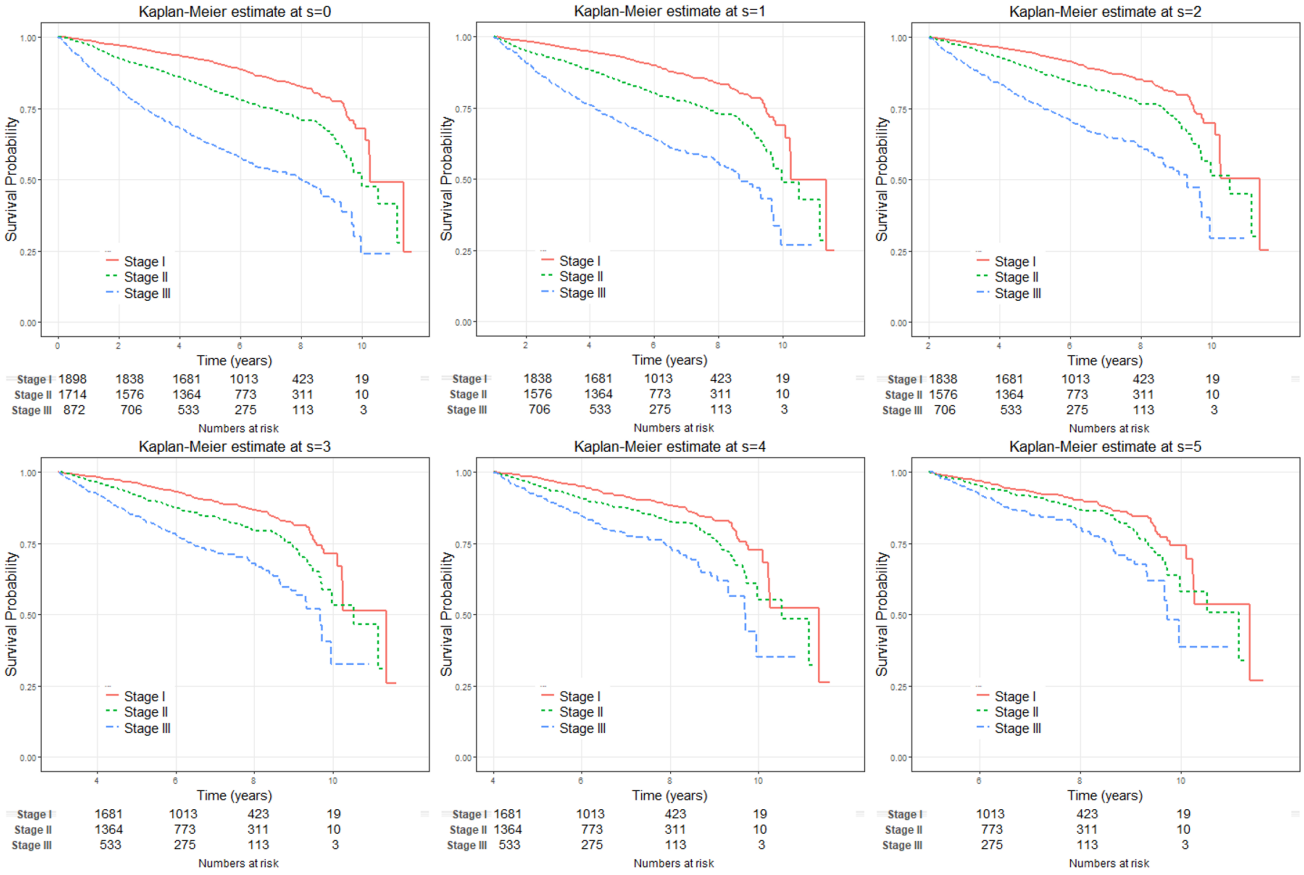

Figure C.2: Kaplan-Meier estimates of the survival function stratified by disease stage considering all individuals alive and disease-free s years after surgery.

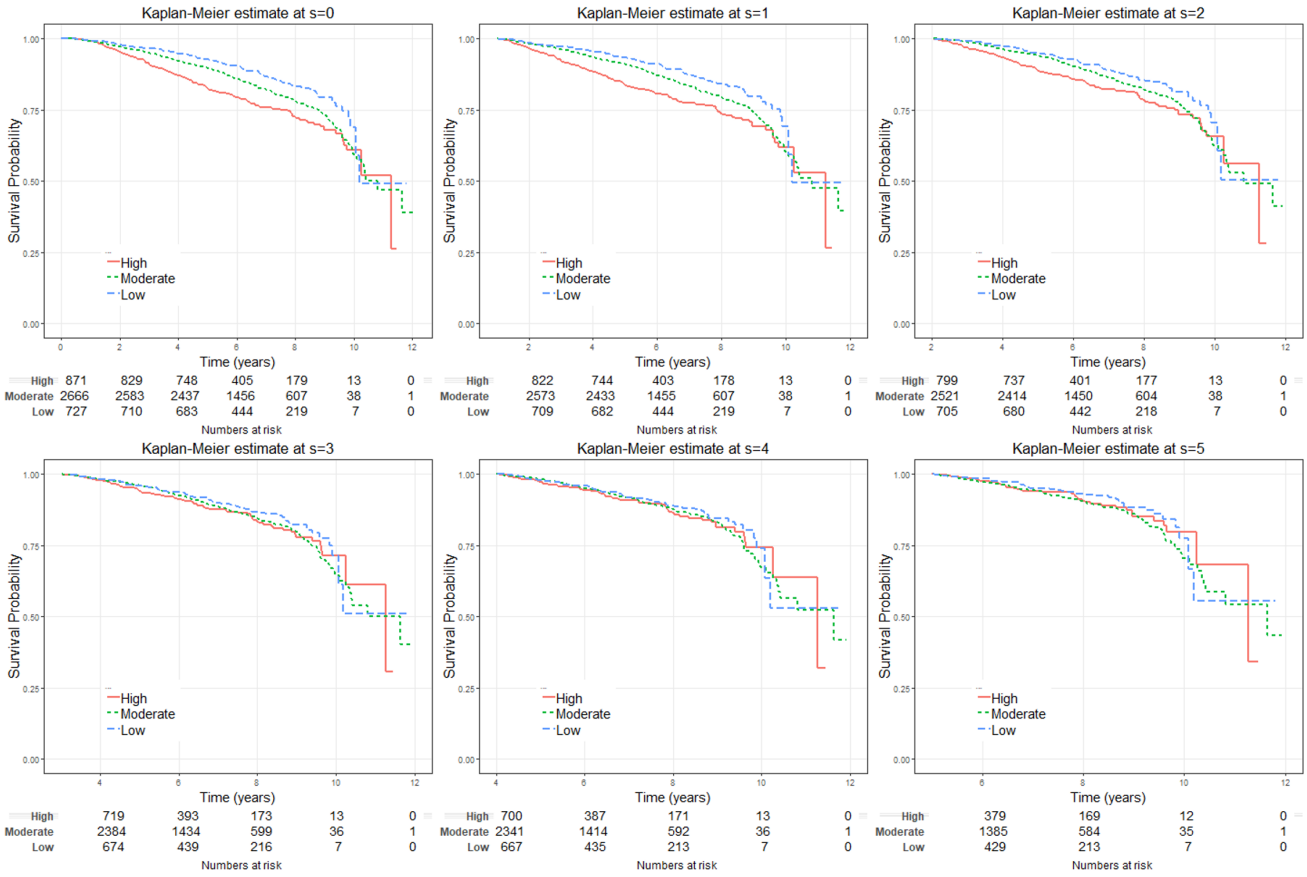

Figure C.3: Kaplan-Meier estimates of the survival function stratified by tumour grade considering all individuals alive and disease-free s years after diagnosis.

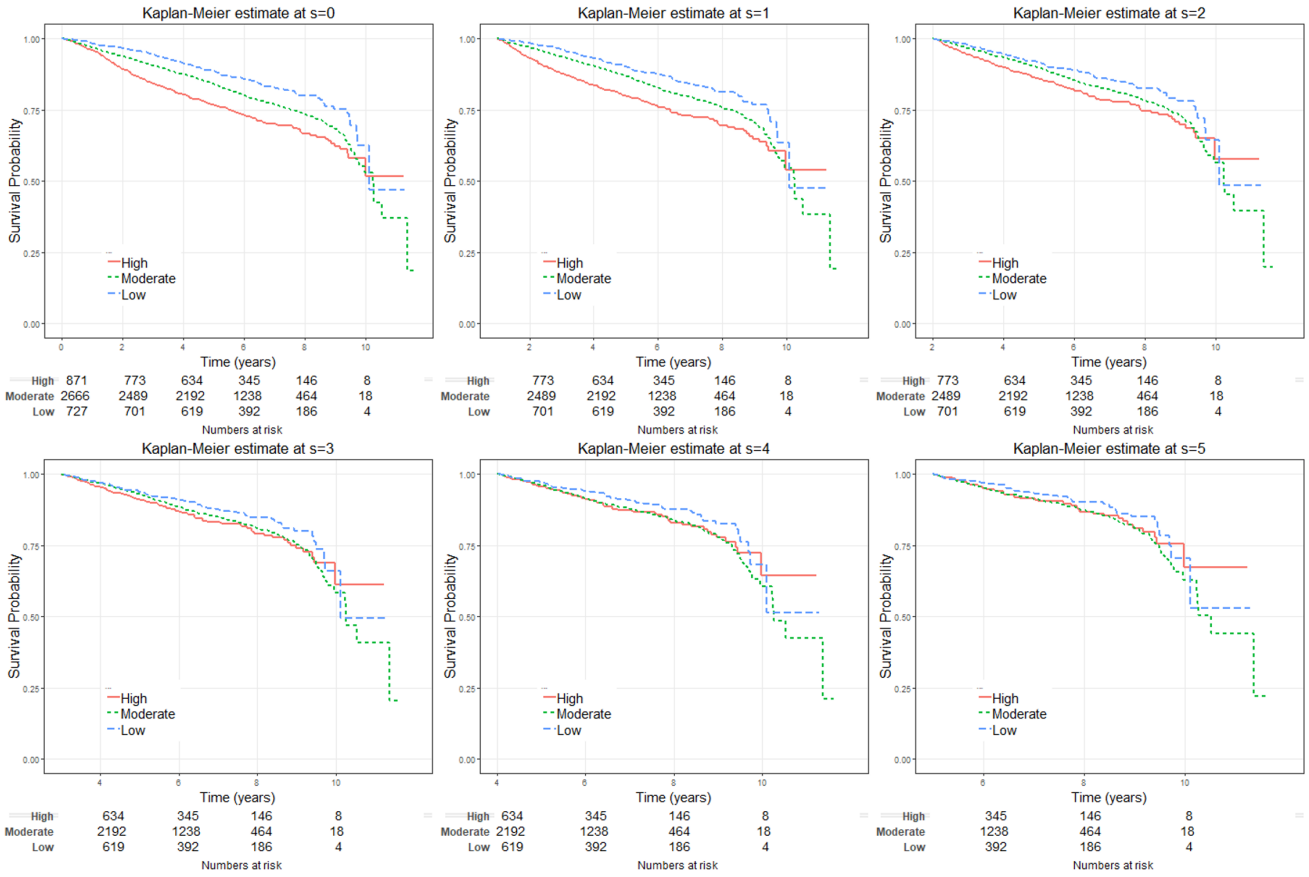

Figure C.4: Kaplan-Meier estimates of the survival function stratified by tumour grade considering all individuals alive and disease-free s years after surgery.

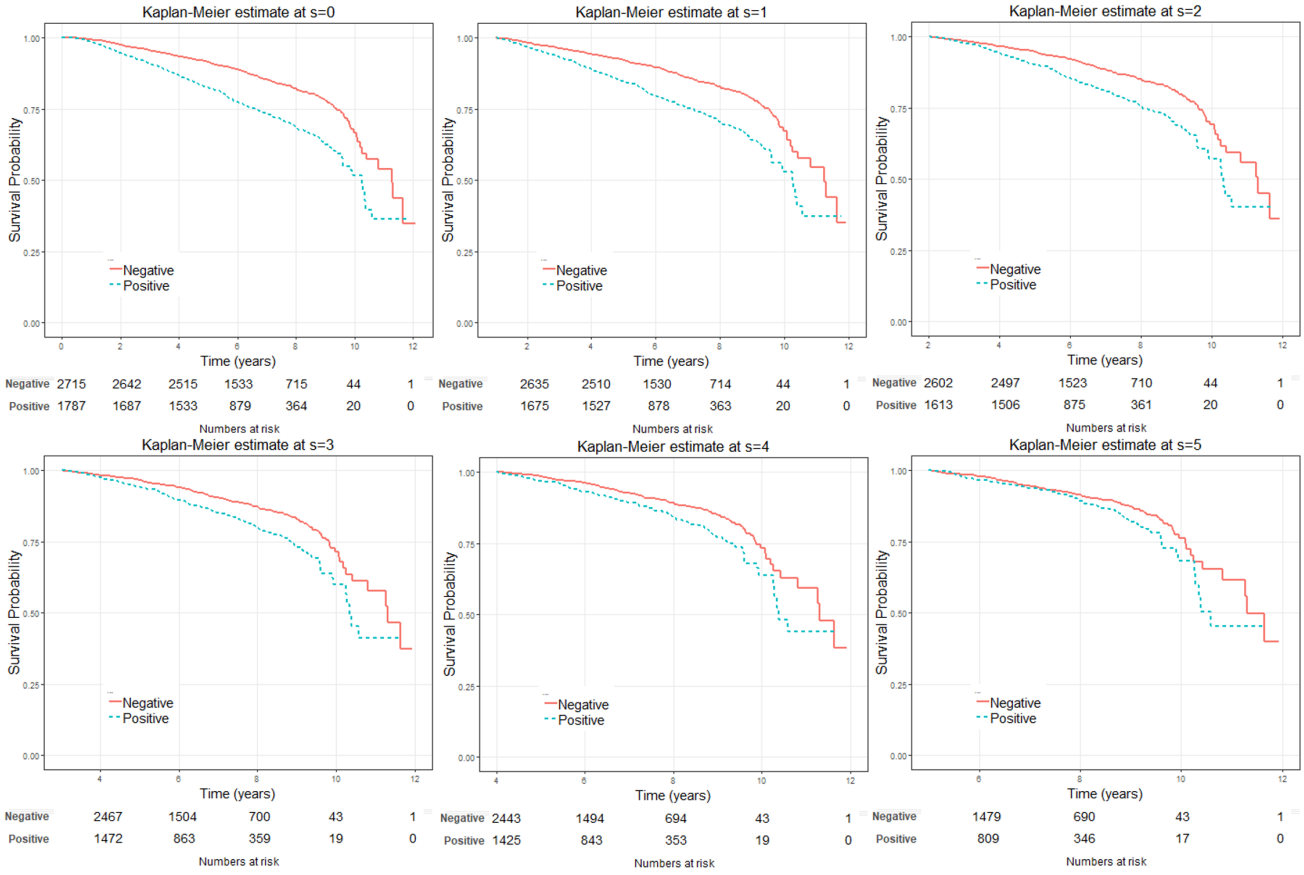

Figure C.5: Kaplan-Meier estimates of the survival function stratified by lymph node status considering all individuals alive and disease-free s years after diagnosis.

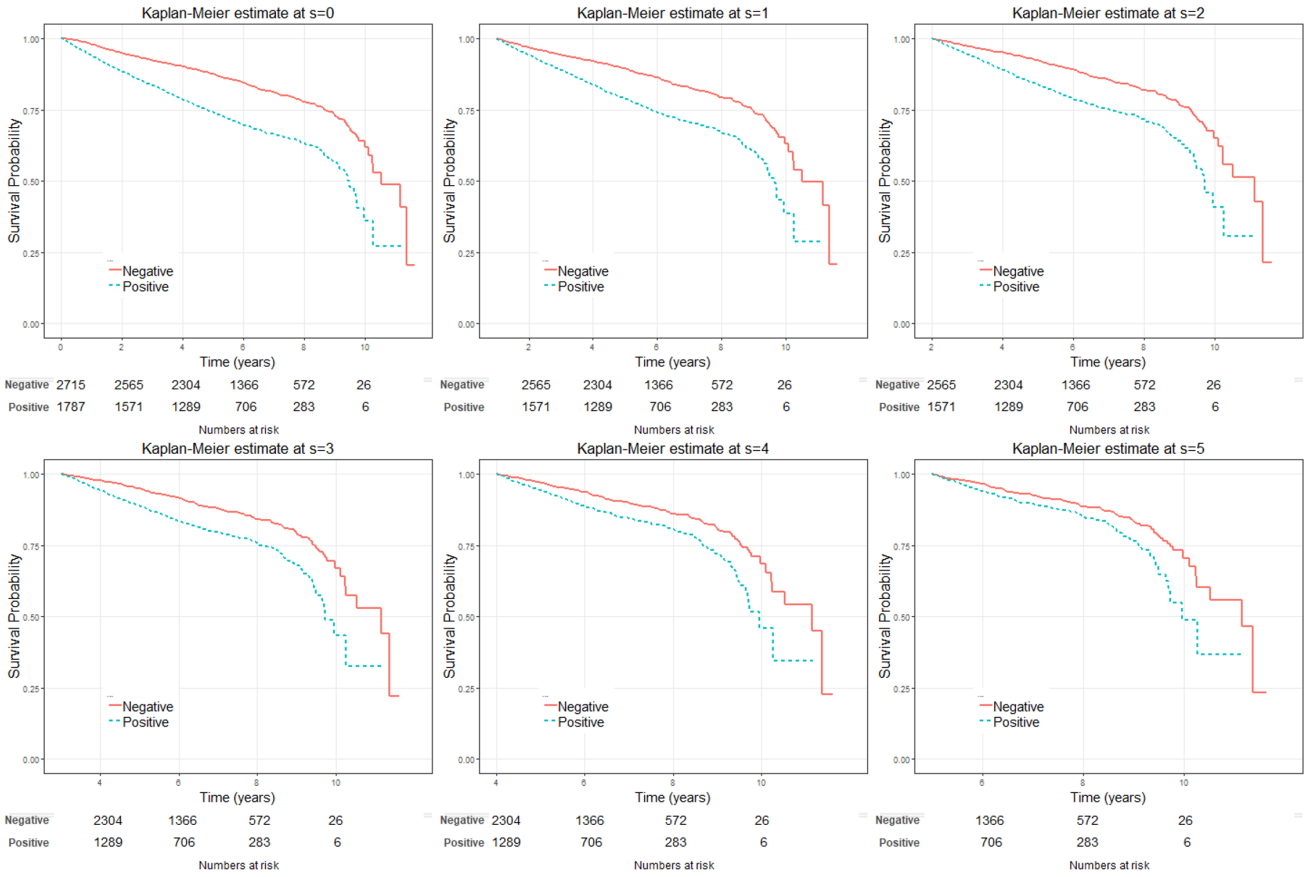

Figure C.6: Kaplan-Meier estimates of the survival function stratified by lymph node status considering all individuals alive and disease-free s years after surgery.

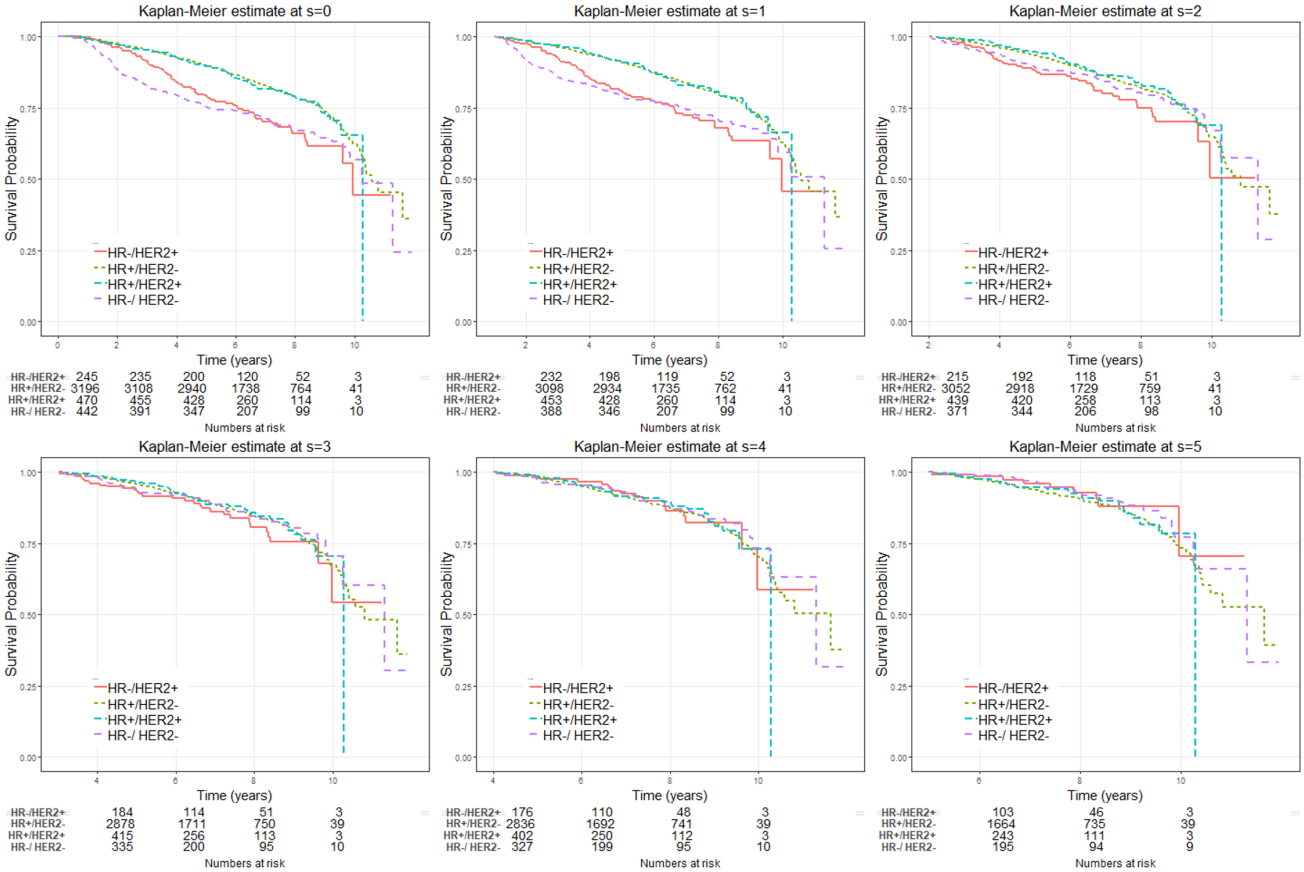

Figure C.7: Kaplan-Meier estimates of the survival function stratified by IHC subtype considering all individuals alive and disease-free s years after diagnosis.

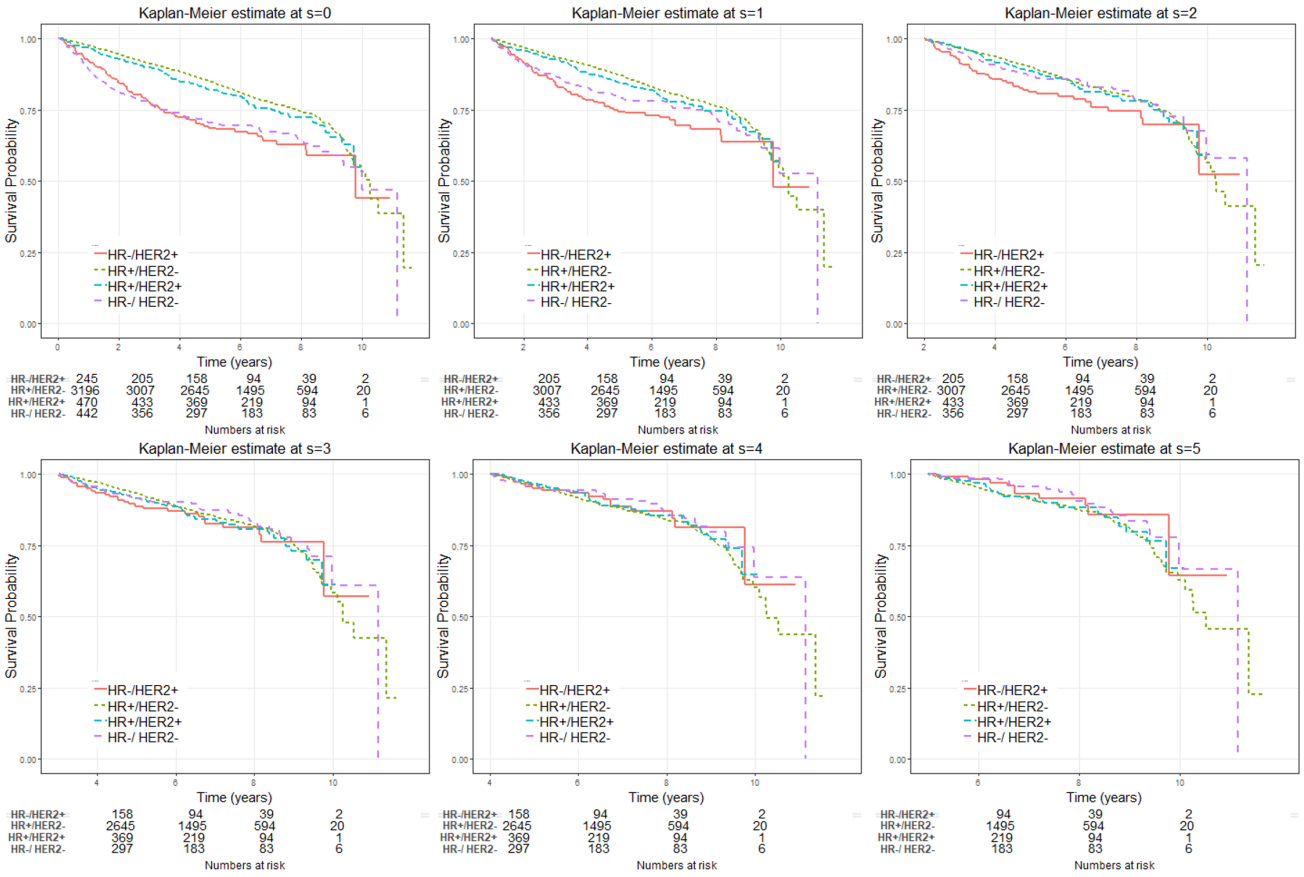

Figure C.8: Kaplan-Meier estimates of the survival function stratified by IHC considering all individuals alive and disease-free s years after surgery.

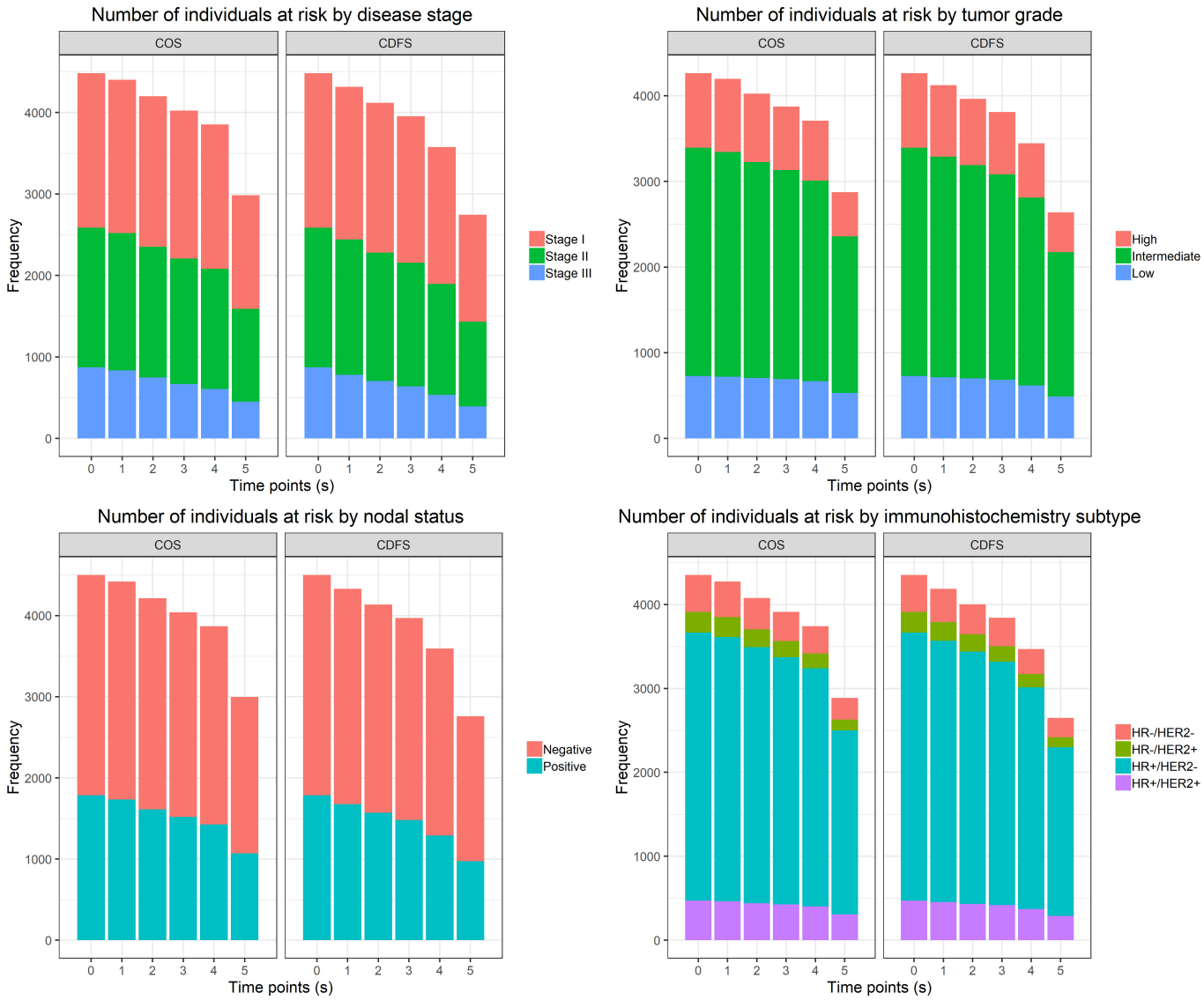

Figure C.9: Number of individuals at risk at each prediction time point s for conditional overall survival and conditional disease-free survival, stratified by prognostic factor

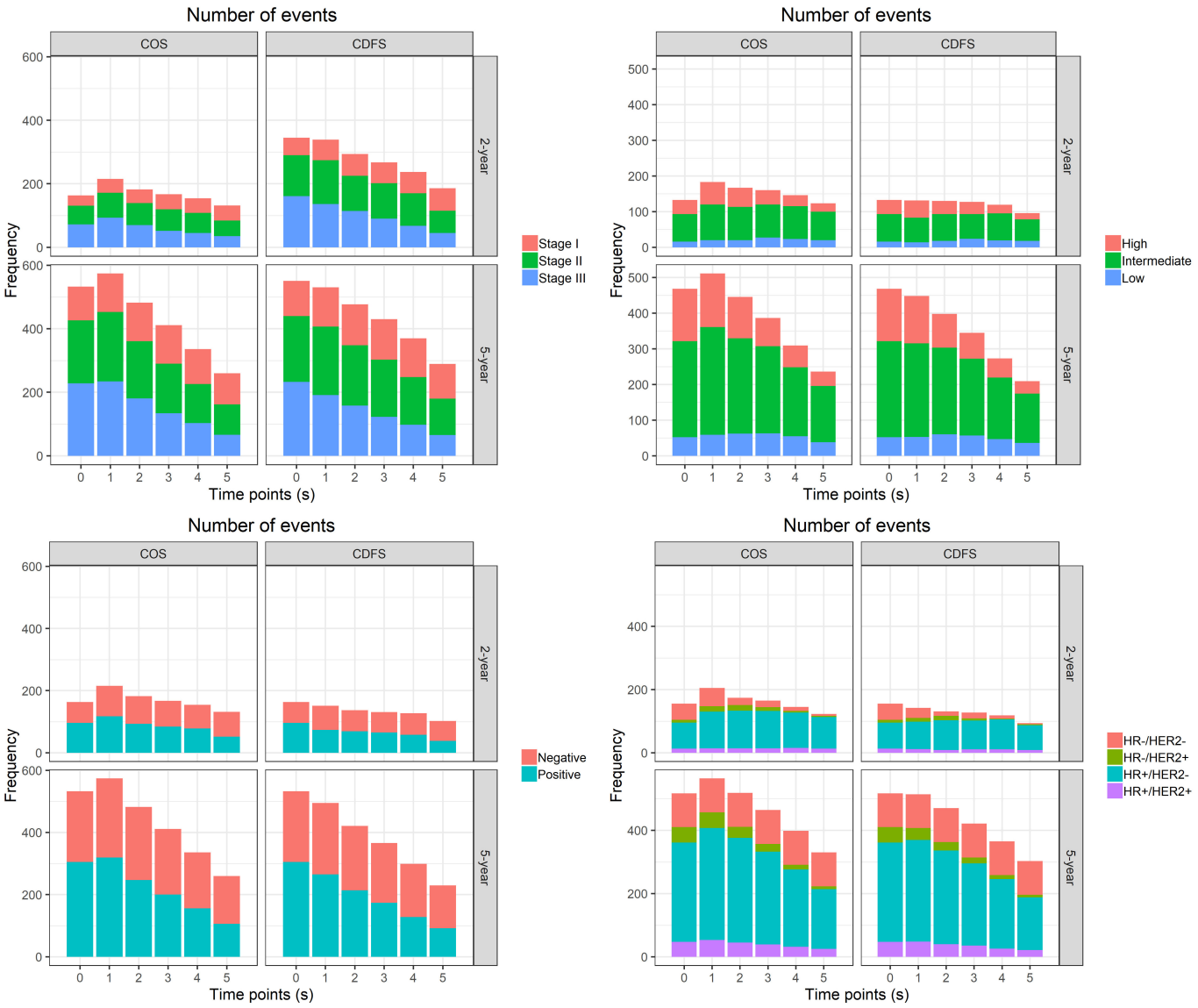

Figure C.10: Number of events within two and five years among those individuals at risk at s for conditional overall survival and conditional disease-free survival, stratified by prognostic factor

D

Variable selection procedures

D.1 Backward selection

Landmark supermodel considering a window of 2 years

- Step 1: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$, tumour grade $\times s$, lymph node status $\times s$, IHC subtype $\times s$, tumour grade $\times s^2$ and lymph node status $\times s^2$. The interaction that presented the highest p-value was tumour grade $\times s^2$ (0.166) and it was removed.

- Step 2: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$, tumour grade $\times s$, lymph node status $\times s$, IHC subtype $\times s$ and lymph node status $\times s^2$. The interaction that presented the highest p-value was tumour grade $\times s$ (0.604) and it was removed.

- Step 3: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$, lymph node status $\times s$, IHC subtype $\times s$ and lymph node status $\times s^2$. All interactions presented statistical significance, considering a 0.10 level. This is the final model.

Landmark supermodel considering a window of 5 years

- Step 1: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$, tumour grade $\times s$, lymph node status $\times s$, IHC subtype $\times s$. The interaction that presented the highest p-value was tumour grade $\times s$ (0.891) and it was removed.

- Step 2: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$, lymph node status $\times s$ and IHC subtype $\times s$. The interaction that presented the highest p-value was lymph node status $\times s$ (0.855) and it was removed.

- Step 3: A model containing the following variables was fitted: age, disease stage, tumour grade, lymph node status, IHC subtype, age $\times s$, disease stage $\times s$ and IHC subtype $\times s$. All interactions presented statistical significance, considering a 0.10 level. This is the final model.

D.2 Forward selection

Landmark supermodel considering a window of 2 years

- Step 1: Seven models were fitted. Each model contain all the main effects. Interaction terms are then added. The model that presented the most significant interaction was the one that comprises IHC subtype $\times s$ (model 5) with a p-value < 0.0001. The covariates of each model fitted are the following:

  1. age, disease stage, tumour grade, lymph node status, IHC subtype and age $\times s$
  2. age, disease stage, tumour grade, lymph node status, IHC subtype and disease stage $\times s$
  3. age, disease stage, tumour grade, lymph node status, IHC subtype and tumour grade $\times s$
  4. age, disease stage, tumour grade, lymph node status, IHC subtype and lymph node status $\times s$
  5. age, disease stage, tumour grade, lymph node status, IHC subtype and IHC $\times s$
  6. age, disease stage, tumour grade, lymph node status, IHC subtype and tumour grade $\times s^2$
  7. age, disease stage, tumour grade, lymph node status, IHC subtype and lymph node status $\times s^2$

- Step 2: Six models were fitted. Each model contain all the main effect and IHC subtype $\times s$. Remaining interaction terms are then added. The model that presented the most significant interaction was the one that comprises age $\times s$ (model 1) with a p-value $= 0.003$. The covariates of each model fitted are the following:

  1. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and age $\times s$
  2. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and disease stage $\times s$
  3. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and tumour grade $\times s$
  4. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and lymph node status $\times s$
  5. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and tumour grade $\times s^2$

6. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and lymph node status $\times s^2$

- Step 3: Five models were fitted. Each model contain all the main effect, IHC subtype $\times s$ and age $\times s$. Remaining interaction terms are then added. None interaction term had statistical significance. In the end, the final model has the following covariates: age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and age $\times s$.

Landmark supermodel considering a window of 5 years

- Step 1: Five models were fitted. Each model contain all the main effects. Interaction terms are then added. The model that presented the most significant interaction was the one that comprises IHC subtype $\times s$ (model 5) with a p-value < 0.0001. The covariates of each model fitted are the following:

  1. age, disease stage, tumour grade, lymph node status, IHC subtype and age $\times s$
  2. age, disease stage, tumour grade, lymph node status, IHC subtype and disease stage $\times s$
  3. age, disease stage, tumour grade, lymph node status, IHC subtype and tumour grade $\times s$
  4. age, disease stage, tumour grade, lymph node status, IHC subtype and lymph node status $\times s$
  5. age, disease stage, tumour grade, lymph node status, IHC subtype and IHC $\times s$

- Step 2: Four models were fitted. Each model contain all the main effect and IHC subtype $\times s$. Remaining interaction terms are then added. The model that presented the most significant interaction was the one that comprises age $\times s$ (model 1) with a p-value $= 0.003$. The covariates of each model fitted are the following:

  1. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and age $\times s$
  2. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and disease stage $\times s$
  3. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and tumour grade $\times s$
  4. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and lymph node status $\times s$
  5. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and tumour grade $\times s^2$
  6. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$ and lymph node status $\times s^2$

- Step 3: Three models were fitted. Each model contain all the main effect, IHC subtype $\times s$ and age $\times s$. Remaining interaction terms are then added. The model that presented the most significant interaction was the one that comprises disease stage $\times s$ (model 1) with a p-value $= 0.002$. The covariates of each model fitted are the following:

1. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$, age $\times s$ and disease stage $\times s$
2. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$, age $\times s$ and tumour grade $\times s$
3. age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$, age $\times s$ and lymph node status $\times s$

- Step 4: Two models were fitted. Each model contain all the main effect, IHC subtype $\times s$, age $\times s$ and disease stage $\times s$. Remaining interaction terms are then added. None interaction term had statistical significance. In the end, the final model has the following covariates: age, disease stage, tumour grade, lymph node status, IHC subtype, IHC subtype $\times s$, age $\times s$ and disease stage $\times s$