FCT
Fundação para a Ciência e a Tecnologia
MINISTÉRIO DA EDUCAÇÃO E CIÊNCIA

UMinho|2015
Ana Isabel Coelho Borges
Joint Modelling of Longitudinal and Survival Data on Breast Cancer

Universidade do Minho
Escola de Ciências

Ana Isabel Coelho Borges

Joint Modelling of Longitudinal and Survival Data on Breast Cancer

novembro de 2015

Universidade do Minho
Escola de Ciências

Ana Isabel Coelho Borges

Joint Modelling of Longitudinal and Survival Data on Breast Cancer

Tese de Doutoramento em Ciências
Especialidade em Matemática

Trabalho efectuado sob a orientação da
Professora Doutora Inês Pereira Silva Cunha Sousa

novembro de 2015

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração da presente tese. Confirmo que em todo o trabalho conducente à sua elaboração não recorri à prática de plágio ou a qualquer forma de falsificação de resultados. Mais declaro que tomei conhecimento integral do Código de Conduta Ética da Universidade do Minho

Universidade do Minho, 27 /11 /2015

Nome: Ana Isabel Coelho Borges
Assinatura: ______________________________

Acknowledgements

The PhD, while being an individual process, apparently lonely, always involves a range of participants behind the curtain.

In these four years, rich in scientific and personal development, I have to begin by thanking my supervisor - Professor Inês Sousa - for having done more than supervising. She guided me, believed in me, taught me and, with her (never-ending!) energy and scientific proactivity, became a reference, an example for me. And to top it entered me in a project that has given me so much pleasure performing continuously for four years.

I must thank Minho's University, in particular the school of Sciences, for having welcomed me so well and provided all the conditions so that I could have done my work. Also, to my fellow colleagues, especially to Lisandra Rocha, since they made me company and provided good times.

The Polytechnic Institute of Porto (IPP), especially the President Rosário Gamboa, for allowing to conduct a PhD in excellent condition. To ESTGF, the presidency and my colleagues, for the encouragement and help in completing the PhD.

To FTC for financial support, by assigning an individual doctoral fellowship SFRH/BD/74166/2010.

On a personal level I was fortunate to have multiple pillars that accompanied me and supported me during these four years. It was essential, since

starting and finishing a PhD with pregnancies, and some sad misfortunes as well, required assistance in physical and psychological terms.

I want to start by thanking my mother Teresa, the main responsible for my fondness for mathematics, for being an example of strength, a catalyst for my adventurous spirit in professional terms and a wonderful grandmother.

To my brothers: my male doppelganger João and my hero Rui. Because they are the ones who tirelessly do not let me accommodate in any situation, always aspiring to be more and better.

To Marco, for his patience and invaluable help, undoubtedly one of the most important pillars of my life. My son Francis for teaching me so much during these four years, and for being a proof that we are able to create formidable creatures. To Margarida, that is yet to be born, for allowing me to conclude the PhD with (some) quietness.

I could not end without thanking all my friends.

To Francisco

and Margarida

Abstract

Joint Modelling of Longitudinal and Survival Data on Breast Cancer

The main motivation of this work is to contribute to the understanding of the progression of breast cancer, within the Portuguese population, using a statistical model with more complex assumptions than the traditional analysis. We aim to infer which risk factors affect the survival of Braga's Hospital patients, diagnosed with breast tumour. While analysing risk factors that affect two tumour markers used on the surveillance of disease progression the Carcinoma Antigen 15-3 (CA 15-3) and the Carcinoembryonic antigen (CEA). We take into account the expected association between the progressions in time of each tumour markers with patient's survival.

A data set of 540 patients, along with 50 variables, was collected from medical records of the Hospital. For survival analysis it was applied the proportional hazards Cox model and the Royston-Parmar flexible parametric model. General linear mixed effects longitudinal models were used to study the progression of both tumour markers values. After conducting survival and longitudinal separate analysis, a joint modelling of these two processes to infer on the association of these was conducted, adopting the methodology of random effects.

Results from joint models, showed that the longitudinal CA-15-3 and CEA values were significantly associated with the survival probability of these patients. We conclude that independent analysis brings up bias parameter estimates and an assumption of association between the two processes in a joint model of breast cancer data is necessary.

Keywords: Breast Cancer, Survival analysis, Longitudinal analysis, Joint modelling.

Resumo

Modelação Conjunta de dados Longitudinais e de Sobrevivência em Cancro da Mama

A principal motivação deste trabalho é contribuir para a compreensão da progressão do cancro da mama, na população Portuguesa, usando um modelo estatístico com pressupostos mais complexos do que a análise tradicional. O objetivo é inferir sobre fatores de risco que afetam a sobrevivência de pacientes do Hospital de Braga, diagnosticados com cancro de mama. Analisando, ainda, fatores de risco que afetam a progressão no tempo de dois marcadores tumorais usados na vigilância da doença - o Carcinoma Antigénio 15-3 (CA15-3) e o Antigénio carcino embrionário (CEA). Temos em conta a associao esperada entre a progressão de cada marcador tumoral com a sobrevivência dos pacientes.

Um conjunto de dados de 540 pacientes, englobando 50 variáveis, foi recolhido dos registos médicos do Hospital. Para análise de sobrevivência recorreu-se ao modelo de riscos proporcionais de Cox e ao modelo paramétrico flexível Royston-Parmar. Modelos lineares generalizados de efeitos mistos foram utilizados para estudar a progressão longitudinal de ambos os valores de marcadores tumorais. Após análises de sobrevivência e longitudinais separadas, foi realizada uma modelacção conjunta destes dois processos para inferir sobre a associacção destes, adotando a metodologia de efeitos aleatórios. Resultados indicaram que as progressões longitudinais de CA-15-3 e CEA estão significativamente associadas com a probabilidade de sobrevivência destes pacientes. Conclui-se que a análise independente devolve estimativas dos parâmetros enviesadas sendo necessário considerar a associacção entre os dois processos num modelo conjunto de dados do cancro da mama.

Palavras-Chave: Cancro da Mama, Análise de Sobrevivência, análise Longitudinal, Modelos Conjuntos.

Contents

Glossary of Terms

AIC	Akaike Information Criterion
AJCC	American Joint Committee on Cancer
CA15-3	Carcinoma Antigen 15-3
CEA	Carcinoembryonic Antigen
CI	Condent Interval
CPHM	Cox Proportional Hazards Model
ER	Estrogen Receptor
FRPM	Flexible Royston-Parmar Survival Model
GEE	Generalized Estimating Equations
GF	Generalized F Distribution
GG	Generalized Gamma Distribution
HER-2 neu	Human Epidermal Growth Factor Receptor 2 Oncogene
HR	Hazard Ratio
Ki.67	Antibody Immunostaining Index
OLS	Ordinary Least Square
PR	Progesterone Receptor
REE	Random Effects, Exponential Serial Correlation and Measurement Error
REG	Random Effects, Gaussian Serial Correlation and Measurement Error
RORENO	North Regional Cancer Registry
TN	Triple Negative
TNBC	Triple Negative Breast Cancers

List of Figures

List of Tables

1

Introduction

This work proposes to develop statistical methods in the scope of biostatistics to study breast cancer within the senology unit of Braga's Hospital, located in Portugal. Biostatistics is a science that develops statistical methodologies motivated by questions and scientific problems within the areas of medicine, epidemiology, public health and biology.

The area of joint models has been receiving a lot of attention in the last years. By joint modelling, in a statistical context, we refer to model simultaneously the joint distribution of two response variables with different distributions.

In this work, we are interested in two response variables commonly measured in health sciences: a longitudinal variable, which is a variable repeatedly measured in time for different subjects; and a time from an origin up to an event of interest. The former known as longitudinal data and the latter as survival, or time-to-event data, in the statistical methodology literature.

Particularly, we analyse two longitudinal variables: the Carcinoma Antigen 15-3 (CA15-3) and the Carcinoembryonic antigen (CEA). Both consisting in a large roll of measures obtained in all blood test of each patients diagnosed with breast cancer, from the senology unit of Braga's Hospital. Which means that we have, for each patient, several CEA and CA15-3 values and the respective moments of their measurements. Also, we are interested in studying the time from breast cancer diagnose until the death from breast cancer.

Survival models are widely used in many medical cancer studies (Porta *et al.* (1991), Alexanian & B. Barlogie (1998), Porta *et al.* (1998), Yuste *et al.* (2003), Law *et al.* (2008)), and in particular, many medical studies regarding breast cancer survival (Gómez *et al.* (1996), Barnett *et al.* (2008), Sparano *et al.* (2009), Kotsopoulos *et al.* (2010), Gramling *et al.* (2010), Rosso *et al.* (2010), Bastiaannet *et al.* (2011), Ardoino *et al.* (2012), Macià *et al.* (2012), Alsaker *et al.* (2013), Crozier et al. (2013), Nechuta *et al.* (2013)). As Ardoino *et al.* (2012) explain, analysis of the hazard rate in medical research provides useful information about disease dynamics as it allows the generation of biological hypotheses and patient follow-up planning. Clarifying, survival models are able to identify prognostic factors, and establishing relationships between the survival probability function and explanatory variables.

Contrariwise, there are not many cancer studies regarding longitudinal data. In fact, most of the epidemiological work on breast cancer focus on survival, because studies of that nature only aggregated data is necessary. Studies that rely on variables such as tumour markers for example, appear in the literature of biological experiments rather than in medical studies, as longitudinal studies are more expensive to develop. Nonetheless, as Chiang *et al.* (2014) point out, longitudinal measurements contain information that cannot be represented by single values or characteristics, and it would be beneficial to develop a model that takes into account all longitudinal measurements to give a clear indication of patient-specific tumour markers profiles for the analyses of an association with the patient health progression.

Hence, the importance of this work is on combining the survival and longitudinal perspective at individual level data. In many medical studies, it is possible that the progression of disease might be associated with relapses times (women getting worst are more likely to have another tumour earlier) or even with death due to the disease. In a setting where the longitudinal observations may be correlated with survival, joint models of longitudinal and time to event processes have been increasingly proposed, to recover information from this potentially informative censoring. Joint modelling of longitudinal data and time to event data is already recognized in the medical community as a powerful tool. In this context, we emphasise the study of Chiang *et al.* (2014) where they joint model the progression of tumour marker CA-125 and the overall survival of patients diagnosed with ovarian cancer.

Oncological diseases are the second highest cause of death in Portugal, and they have a big social impact in patients and their families (Pinheiro *et al.*, 2003). In Europe breast cancer is the tumour with highest incidence in women (Boyle & Ferlay, 2004). In Portugal there are not many published studies on breast cancer over the years.

However, the publication of 2003 by Pinheiro *et al.*, refers that since 1995 mortality, due to breast cancer, has been decreasing in Portugal. They argue that this improvement, is a consequence of earlier diagnostic and better quality of treatment. The Strategic National Health Plan refers, in the recommendations for breast cancer, the importance of "medicine based on evidence" to establish guidelines and specify protocols, for a good medicine in practice.

Therefore, it is of great importance the continuous investment on statistical and epidemiological studies in oncological diseases for understanding the progression of the disease in Portugal.

Currently, only a few studies characterizing the complexity of this cancer on Portuguese population exist, and are mainly conducted by population-based cancer registries, as for example RORENO (North Regional Cancer Registry). Although of great importance, as they quantify the incidence and prevalence of the disease in the population (dos Santos Silva, 1999), the information they gather can be somehow scarce in terms of tumour specificity or even, for example, in terms of information about the real cause of death. Thus, data taken directly from the hospital's Senology Unit allows to investigate cancer-specific survival estimates and the prognostic value of clinical factors unrecorded in some population registries (Macià *et al.*, 2012). Therefore, it is of great importance the continuous investment on statistical models in oncological diseases for understanding the progression of the disease in Portugal.

So far, the most recent publication we have access to is a publication of Cancer Registries Northern Region (RORENO) presenting results of survival of cancer patients diagnosed with malignant tumours in the period 2007-2008 and residents to diagnostic date in the area of influence of RORENO, which includes Braga, the area of interest of this work.

This work has four main objectives: (i) produce an epidemiological study on breast cancer in a specific senology unit in Portugal; (ii) develop statistical models for breast cancer in Portugal, regarding longitudinal progression of tumour markers and survival; (iii) develop a statistical model on joint modelling of longitudinal tumour markers process and time to death for breast cancer, and (iv) compare results of different models in order to conclude on the importance of joint models.

For this study we collected data directly from the medical records of each patient diagnosed with breast cancer, listed in the computer system of Braga's Hospital. As there is no computer application, or database, available that resumed all patients' breast cancer related information, it was necessary to read and analyse all the clinical reports from each patients (which includes, as well, numerous blood and histological reports), to verify and complete every information written by the doctors, as we were confronted with some missing, and even contradictory information. It was, in fact, a time-consuming and labour-intensive process, with a duration of almost a year. Which reflects the need of a software for this purpose.

Nonetheless, we were able to obtain a group of 50 variables able to describe subject and tumour specificities.

The motivation for the study is described in the second chapter, presenting global, European and national breast cancer statistics along with a descriptive analysis of the data collected.

The following chapter presents the survival analysis preformed in order to understand what the possible risk factors were for death from breast cancer, for these patients. Describing, initially, the methodology employed, along with a chronological overview of existing models in the literature, the main results and a short discussion of these.

The fourth chapter focuses on longitudinal analysis of tumour markers CA15-3 and CEA, implemented in order to identify risk factors related to the increase of its values. Beginning with a description of the methodology adopted in this work, following with the presentation of the main results of each longitudinal models and ending with a summary discussion of these.

Since it was found that the evolution of the values of the two tumour markers may be associated with the probability of survival for these patients we performed a joint analysis of survival data and longitudinal data, with the purpose of creating a joint model that incorporated both factors risk for death from breast cancer and for the progression of tumour markers. Chapter 5 explains of the main methodologies applied in this field along with a detailed description of the one employed in this particular study, ending with the presentation of the main results followed by a discussion of these. Finally, the final chapter provides a discussion of the main result along with an indication of future work to consider.

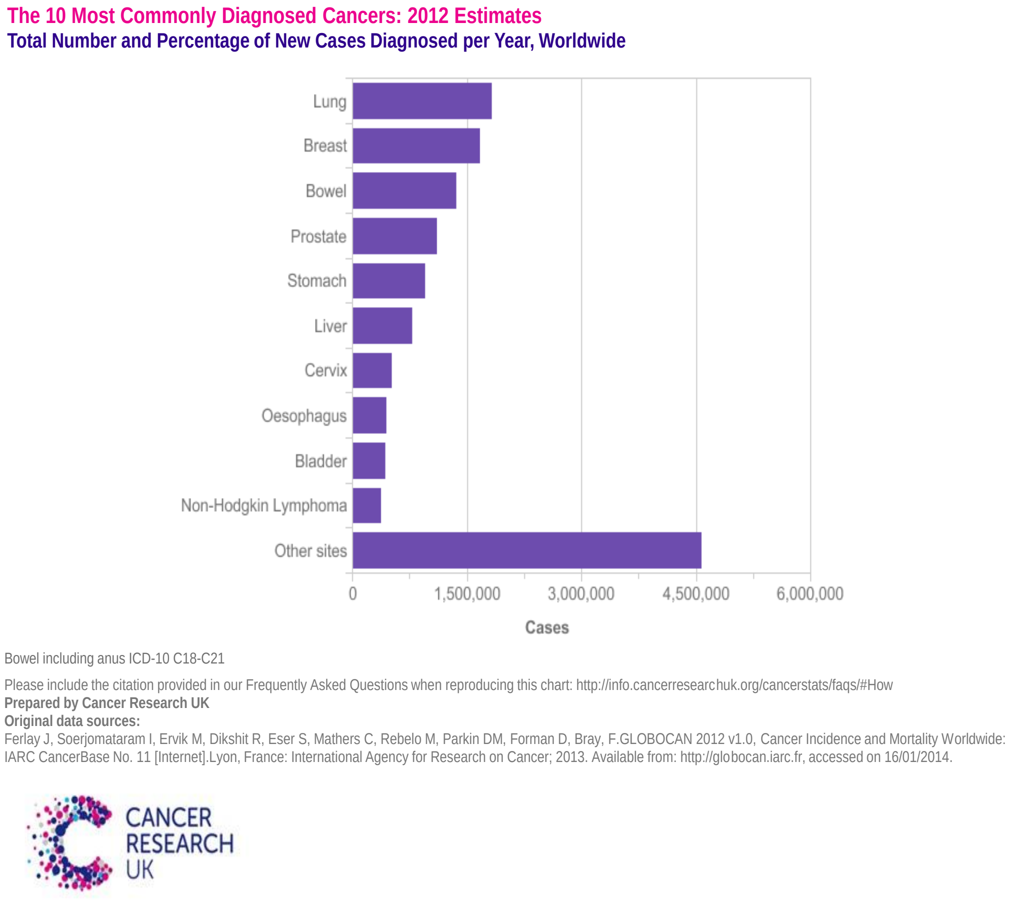
2

Breast Cancer at a Senology Unit

2.1 Breast Cancer Epidemiology

Breast cancer is a malignant tumour, affecting mostly women, summary characterized as an abnormal cell proliferation starting on the breast. It can invade surrounding tissues or spread to distant areas of the body. This is an extremely complex disease, characterized by a progression multistep process caused by interactions of both environmental and genetic factors.

As Ferlay *et al.* (2013b) report, breast cancer is the second most common cancer in the world and the most frequent cancer among women. They estimated that 1.67 million new cancer cases were diagnosed in 2012 (25% of all cancers). It is, actually, the most common cancer in women both in more and less developed regions with slightly more cases in less developed (883 000 cases) than in more developed (794 000 cases) regions. Figure 2.1 presents the distribution of the number of new cases diagnosed per year, worldwide, of the 10 most commonly diagnosed cancers, in 2012.



The 10 Most Commonly Diagnosed Cancers: 2012 Estimates

Total Number and Percentage of New Cases Diagnosed per Year, Worldwide

Bowel including anus ICD-10 C18-C21

Please include the citation provided in our Frequently Asked Questions when reproducing this chart: http://info.cancerresearchuk.org/cancerstats/faqs/#How

Prepared by Cancer Research UK

Original data sources:

Ferlay J, Soerjomataram I, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F.GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet].Lyon, France: International Agency for Research on Cancer; 2013. Available from: http://globocan.iarc.fr, accessed on 16/01/2014.

Figure 2.1: The 10 Most Commonly Diagnosed Cancer: 2012 Estimates. Ferlay *et al.* (2013b), available from: http://globocan.iarc.fr

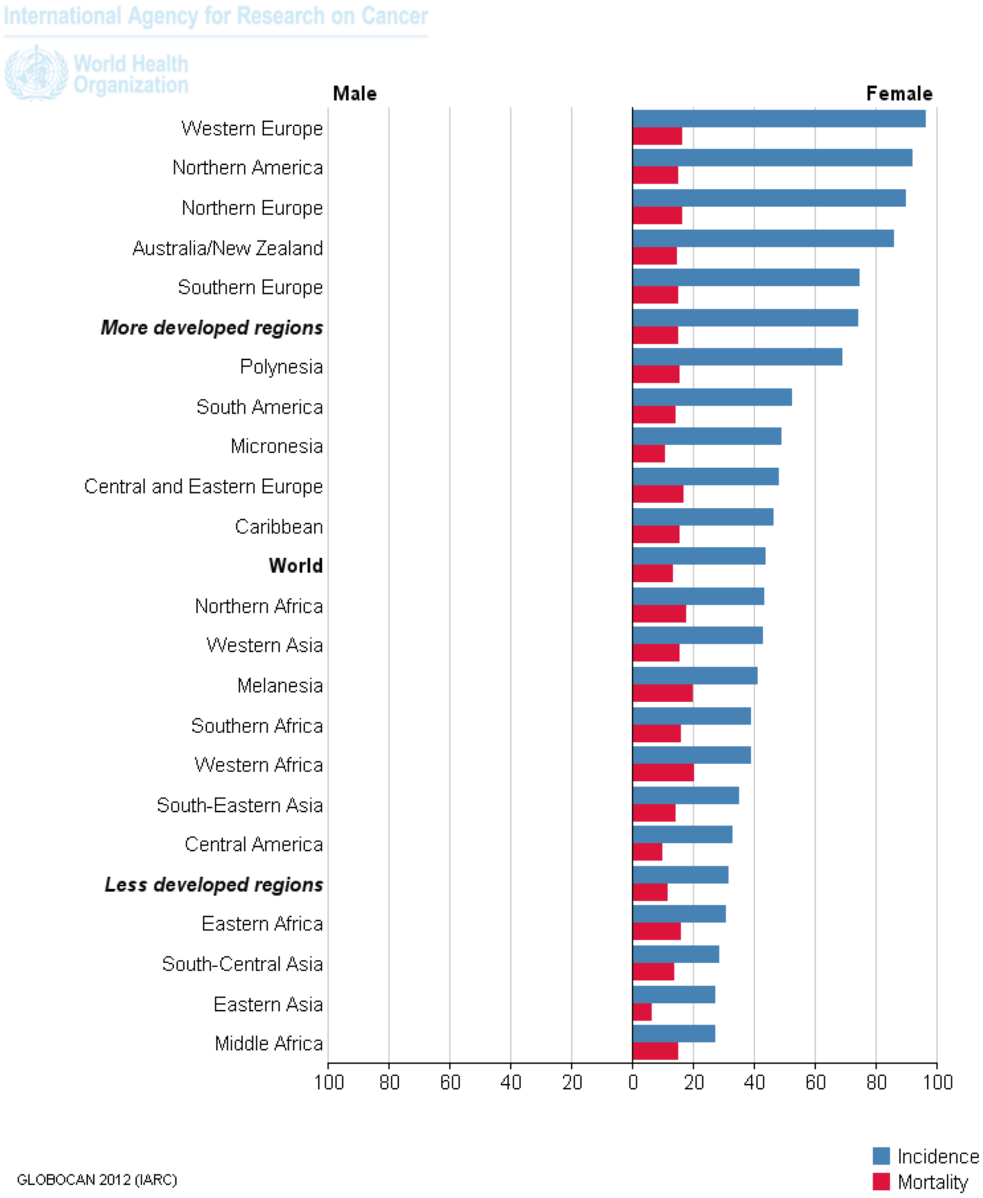


Figure 2.2: Incidence and Mortality of Female Breast Cancer Worldwide. Ferlay *et al.* (2013b), available from: http://globocan.iarc.fr

It is possible to note a great geographical variation worldwide associated with the incidence and mortality of female breast cancer, as shown in Figure 2.2, being the incidence higher for Western Europe, but mortality higher for Melanesia and Western Africa.

In fact, Ferlay *et al.* (2013b) explain that breast cancer ranks as the fifth cause of death from cancer overall (522 000 deaths) and while it is the most frequent cause of cancer death in women in less developed regions (324,000 deaths, 14.3% of total), it is now the second cause of cancer death in more developed regions (198,000 deaths, 15.4%) after lung cancer.

According to the 2008 National Register Oncology document, where all the information on all types of cancer diagnosed in 2008 in Portugal is documented, about one-third of the tumours diagnosed in women corresponded to breast cancer (30.2%) in 2008 and it was reported as the most frequent cancer among women.

Similarly, the European Cancer Observatory reported, in the EUROCARE-5 study (Angelis *et al.*, 2014), breast cancer as the cancer with higher estimated incidence for women in Portugal, as it is also reported in Figure 2.3.

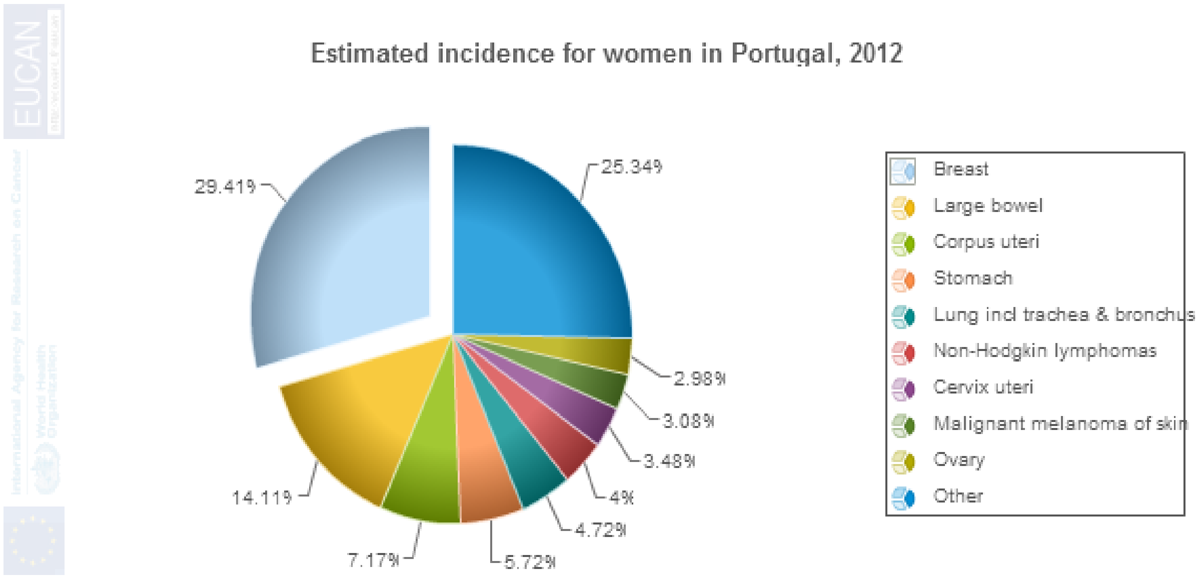


Figure 2.3: Incidence of Female Breast Cancer in Portugal in 2012. Ferlay *et al.* (2013a), available from: Available at: http://eco.iarc.fr/eucan/Country.aspx?ISOCountryCd=620

They also estimated an incidence rate per 100 000 women for breast Cancer in

Portuguese women in 2012 of 85.6%, and the mortality rate per 100 000 women due to this type of cancer of 18.4%, both values lower than the European average (92.8% and 23.1% respectively).

Furthermore, it is reported that in Portugal the estimated 1-year prevalence from breast cancer in women in 2012 is 12.06%, the estimated 3-year prevalence is 34.11% and 5-year prevalence is 53.83%. Lower that the European average prevalence for these periods is, respectively, 12.17%, 34.24% and 53.58% (Angelis *et al.*, 2014).

The 2008 National Register Oncology document reports a geographically diversity in the incidence rate (standardized for the European population) in Portugal, observed in Figure 2.4), being higher for districts nearby Lisbon.

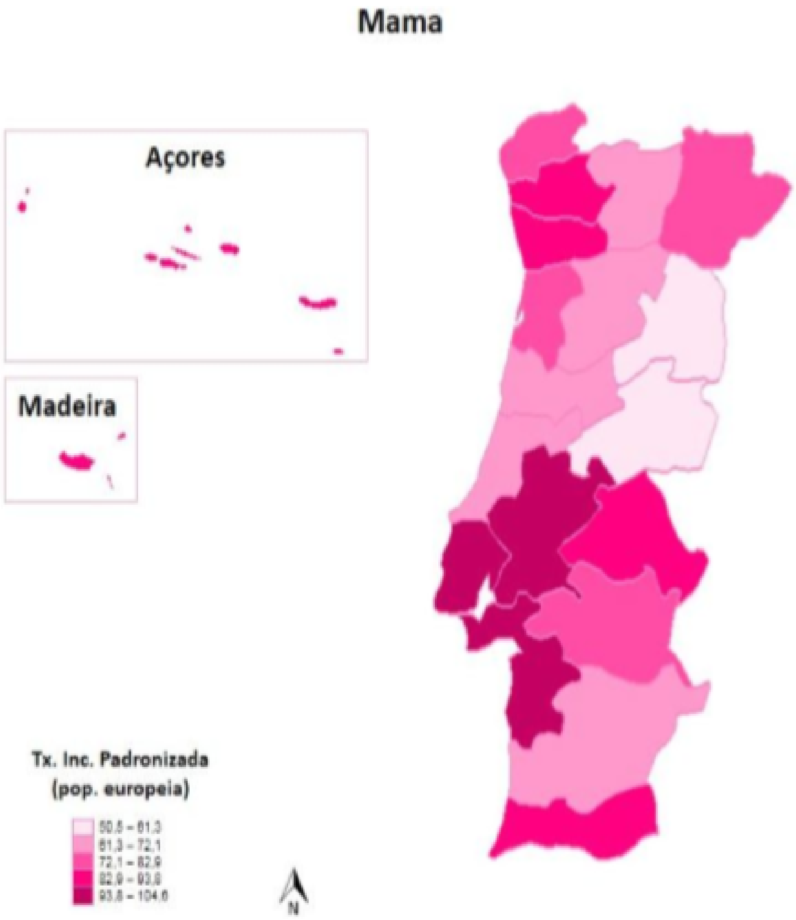


Figure 2.4: Geographical distribution of Breast Cancer in women, diagnosed in 2008 (2008 National Register Oncology, page. 36)

Rodrigues (2011) explains that mortality, due to breast cancer, has increased in most countries since the 1950s to the 1980s, particularly in the southern countries and Eastern Europe. Since the 1990s values have stabilized, situation first observed in England, Wales and the Netherlands, and subsequently in most other European countries. However, this result has been almost exclusively limited to the age to 50 years.

Particularizing, Bastos *et al.* (2007) indicate that breast cancer mortality rates have also declined in the Portuguese population in the 1990s with an estimated annual percent change of -2% per year.

Angelis *et al.* (2014) reported a five-year age-standardised relative survival for adult cancer patients diagnosed in 2000-2007 in Portugal of 83.3% (IC[82.4;84.2]). And informed that for most countries, 5-year survival for breast cancer (women only), for that period, was fairly close to the European mean (81.8%, IC[81.6;82.0]). They point out that in all northern and central European countries, and also Italy, Spain, and Portugal, survival was greater than 80%, for the period between 2000 and 2007. For all regions, survival peaked at 4554 years, and fell with age thereafter. Also, survival for the whole of Europe increased over time: from 78.4% (IC[78.1;78.8]) in 19992001 to 82.4% (IC[82.2;82.7]) in 20052007.

The most recent study, conducted by the North Region Cancer Registry of Portugal (RORENO), entitled "Global survival, patients diagnosed in 2007-08", considered all patients aged over 15 years, diagnosed with malignant tumours in the period from 2007 to 2008 and resident at the time of diagnosis in the area of influence of RORENO - where Braga's Hospital is located. They compared the value reported by the study of Angelis *et al.* (2014) with the value of 87% (IC[84.6;88.8]) they obtained for the standardized 5-year survival. Meaning that the 5-year survival for the north of Portugal was greater in the period of 2007-2008 compared to the national a five-year survival for the period of 2000-2007.

The referred RORENO's study also compared the results obtained with the ones they reported in a previous period analysed (2005/2006), where differences with statistical significance for breast cancer survival estimates emerged, resulting in an improvement of survival.

2.2 Braga's Hospital

Braga's Hospital is situated in the east of the city of Braga, located in northern Portugal. Today it serves a direct area of about 275,000 clients.

Braga's Hospital area of influence includes the districts of Braga and Viana do Castelo and serves as 1st line Hospital the municipalities of Braga, Braga, Póvoa de Varzim, Terras de Bouro, Vieira do Minho and Vila Verde and has 2nd line hospital for the remaining area. Being a Central Hospital, it covers a population, according to the 2011 census, of 1 081 641 inhabitants. The female population over 15 years of 1st line municipalities is 128 859.

In 2008, a Senology unit at the Braga's Hospital was created. Currently, it operates an average of 130 new cases of breast cancer per year.

2.3 Descript Analysis of Breast Cancer Data from Braga's Hospital

Data were collected directly from the medical records of each patient, listed in the computer system of Braga's Hospital Glintt HS. We therefore have access to baseline and clinical history of each patient (a roll of information such as diagnosis; pre-surgery, post-surgery, group meetings; follow-up and medical exams). The authorization for collect and use of senology data was approved by the Ethical Committee of Braga's Hospital.

Since there is no computer application, or database, that resumed all the patients information we had to read and analyse all the clinical reports from each patients, including blood and histological reports, to verify and complete every information written by the doctors, as we were confronted with some missing, and even contradictory information. It was, in fact, a time-consuming and labour-intensive process, with a duration of almost a year. Which reflects the need of a software for this purpose.

Data were collected on all patients diagnosed with breast cancer from 2008 until 2012, in the hospital, and for all the patients in follow-up at the hospital at the date of 1st January 2008.

We collected information on 596 patients, 56 patients were excluded since they corresponded to, at least, one of the following exclusion criteria's:

1. No diagnostic, treatment or follow-up information;
2. Male gender;
3. Benign breast neoplasm.

Therefore 540 patients were eligible for the present analysis. As 19 patients presented bilateral breast cancer that should be treated as independent cases, on doctor's advice, which translates into a total number of 559 cases analysed of female patients diagnosed with a malignant tumour.

It is noteworthy that, despite the Senology unit of Braga's Hospital being created only in 2008, there were patients being followed up by that date, but with a diagnostic date before 2008. There are patients with diagnostic dates as far as 1993. These specific cases sums into 186 cases. However, we do not have any information about patients who have died before 2008. It is a relevant and important information to consider when performing, for example, the survival analysis where it is necessary to perform a left truncation of the data.

The total number of deaths is 74, however the total number of deaths from breast cancer is only 55.

From the information gathered in all the medical reports we were able to collect more than 50 variables. They can be grouped into two categories: (i) explanatory variables at individual level, which are a group of demographic characteristics, prognostic and etiologic factors (i.e. a cause or origin of a disease) resumed by Rodrigues (2011), based on the work of Trichopoulos *et al.* (2008), as for example: age, menopause, age at first full term pregnancy, family history of breast cancer, etc.; and (ii) explanatory variables at tumour level, that include characteristics of the tumour, some of them important prognostic factors already reported in the literature and resumed by Fitzgibbons *et al.* (2000) and Cianfrocca & Goldstein (9), such as TNM stage, histological type of tumour, hormonal receptors or vascular or lymphatic invasion, among others. Table 2.1 gives an exhaustive list of these variables.

Table 2.1: Variables collected from the Senology Unit of Braga's Hospital.

Explanatory variables at individual level	**Explanatory variables at tumour level**
Date of Birth	Histologic Type of tumour
Marital status	Degree of differentiation - Bloom-Richardson
District of Residence	Presence of Carcinoma Associated
County of Residence	Images of lymphatic invasion
Parish of Residence	Images of venous invasion
Profession	Estrogen receptors expression
Qualifications	Progesterone Receptor Expression
Age at Menarche	HER-2/neu
Births (number of children)	Ki-67 proliferative index
Age at first full term pregnancy	Neoadjuvant treatment
Breastfeeding	Surgical Treatment
Duration of Breastfeeding	Sentinel lymph Biopsy
Menopause	Result of sentinel lymph Biopsy
Type of Menopause	Axillary dissection
Age at Menopause	Results of axillary dissection
Family history of breast cancer	Treatment received after surgery
Degree of parentage	Breast intervened
Therapeutic Hormone Replacement	Breast Quadrant
Oral Hormonal Contraceptive	Tumour Stage (TNM)
Age at Diagnosis	Size of primary tumour
	Degree of spread to regional lymph nodes
	Presence of distant metastasis
	Recurrence
	Type of Recurrence
	CEA values
	CEA dates
	CA 15-3 values
	CA 15-3 dates
	Dates of follow up consultations or death
	Vital status

As expected, with Braga's Hospital being a public hospital of Braga, the vast majority of patients live in Braga (96.4%). There are a few patients living in Porto (only 4 cases) and Viana do Castelo (16 cases).

More than half of the cases (64.19%) are related to patients with a very low education level (with only 4 grade or less) in contrast to the 9.77% of cases of patients with a higher level of education, as can be seen in Figure 2.5.

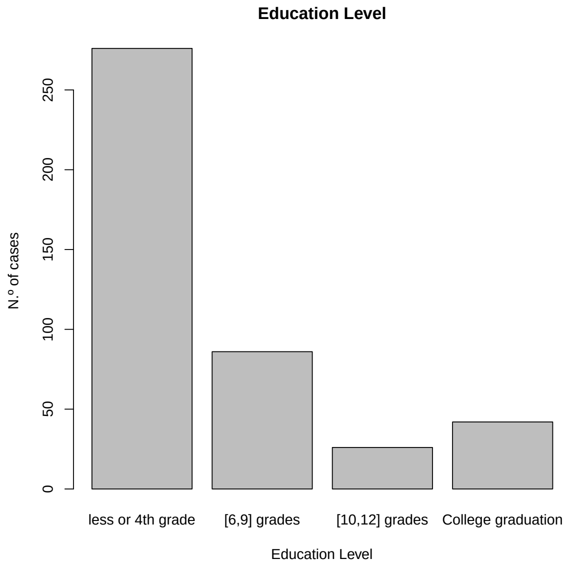


Figure 2.5: Frequency Distribution Graph of Braga's Hospital Patient's Education Level.

Regarding to breast cancer etiological factors summarized, as already mentioned, by Rodrigues (2011), we will briefly describe the information collected for the following factors: age at diagnosis, menarche, parity (number of children), age at 1st full term pregnancy, lactation, age at menopause, type of menopause, hormone replacement therapy, oral contraceptives, family history of cancer and the presence of cancer in the other breast (bilateral cancer).

The age at diagnosis of the patients diagnosed with malignant breast tumour varies between 20 and 92 years, with an average of 58 years. The age distribution is illustrated in Figure 2.6. Note that increasing age is documented as an etiologic factor however, as Cianfrocca & Goldstein (9) indicate, there are two major studies that demonstrated a poor prognosis for patients with less than 35 years, even after adjusting for other prognostic factors. This means that age at diagnosis is a possible prognostic factor to be taken into account.

Actually, the work of Leclère *et al.* (2013), where they describe the trends in incidence of breast cancer in women under 40 from 1990 to 2008, using pooled European data, show that in Portugal the overall incidence rate increased linearly during that period by 2.68% (CI = [1.97; 3.40]). They explain that the increase can be due to a rise in risk factors and/or changes in diagnosis and surveillance practices. Nonetheless, they could not clearly distinguish between those two non-exclusive explanations.

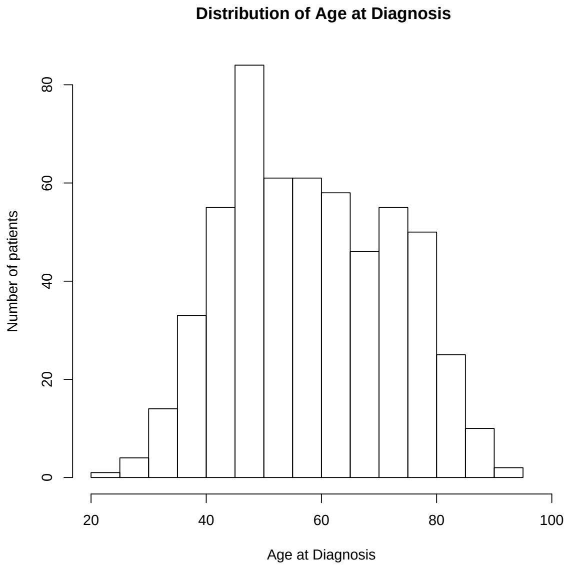


Figure 2.6: Frequency Distribution Graph of Braga's Hospital Patient's Age at Diagnosis.

Menarche at an early age (under 12 years) is considered an etiological factor. In our sample, this represents 32.4%. Figure 2.7 shows the distribution of ages at menarche for cases analysed, where there is a greater concentration of data between 11 and 14 years. However, we accounted for 42.58% of missing information for this variable, that is, 238 cases.

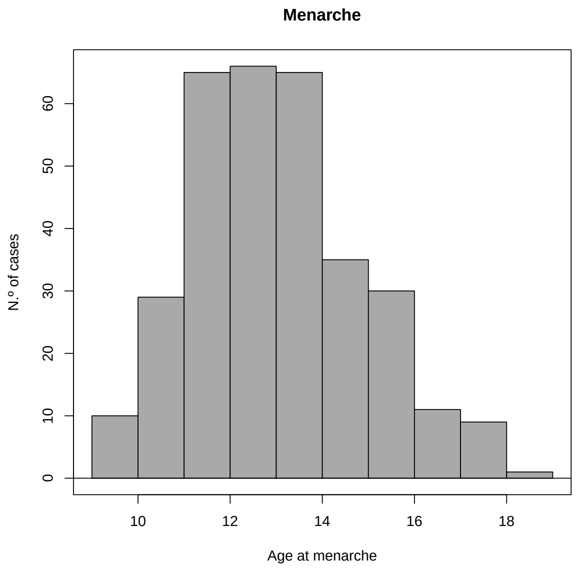


Figure 2.7: Frequency Distribution Graph of Braga's Hospital Patient's Age at Menarche.

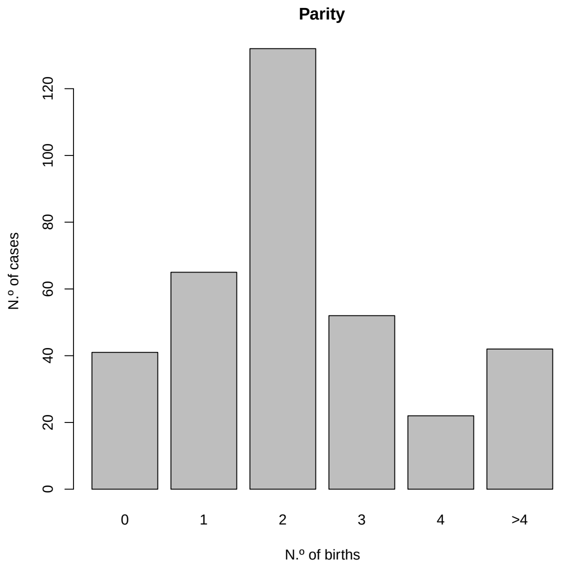


Figure 2.8: Frequency Distribution Graph of Braga's Hospital Patient's Parity.

Low parity is also reported to be an etiologic factor. In this particular study only 11.58% had no children, 18.36% had only one child, 37.29% had 2 children, 14.69% had 3 children and 18.07% had four or more children. Figure 2.8 shows a plot with the distribution on the number of cases, by number of births. Similarly for this variable we registered a high number of missing cases (205 cases, corresponding to 36.67%).

In our study the distribution of age at first full-term pregnancy behaves as shown in Figure 2.9, ranging between 15 and 40 years, being 25 the median age and 25.3 the mean age. Only 4.46% of the cases were patients with age at first full-term pregnancy age 35 or above.

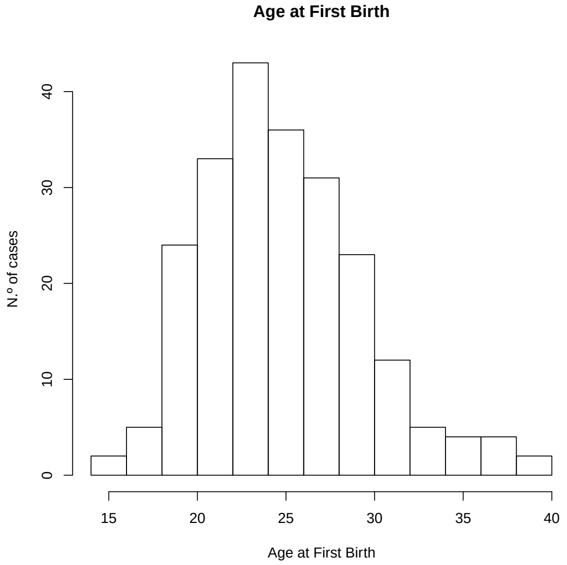


Figure 2.9: Frequency Distribution Graph of Braga's Hospital Patient's Age at First full-term pregnancy.

For cases where information was available relating to breastfeeding, we accounted for 93.72% that breastfeed and only 6.28% that did not breastfeed. Note that we do not have information for 60.11% of cases, that is, 336 cases.

Results show that 63.75% of the cases correspond to patients already in menopause (we have no information for only 3.76% of cases). The type of menopause - natural or artificial - it is an etiological factor, however, in our sample, only 9.25% of cases are artificial menopause. In our database age at menopause is distributed as shown in Figure 2.10, ranging between 33 and 58 years. Age-late menopause (after 55 years) also represents an etiological factor, but only represents 4.79% of the cases in our study.

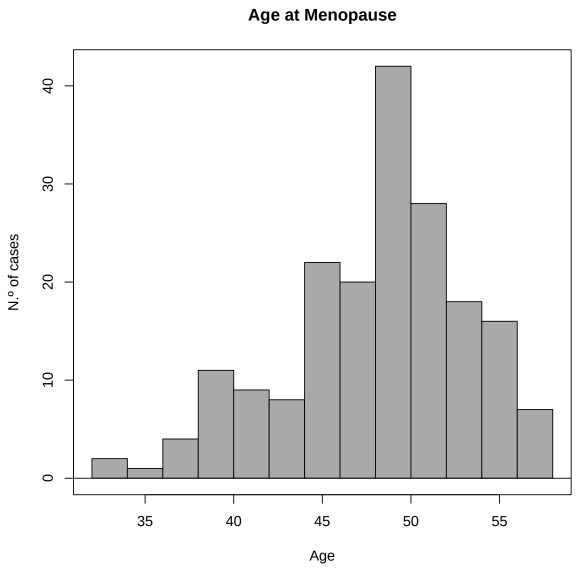


Figure 2.10: Frequency Distribution Graph of Braga's Hospital Patient's Age at Menopause.

Both variables submission to hormone replacement therapy and the recent use of oral contraceptives are reported as etiological factors (Rodrigues, 2011). However, the percentage of missing cases (missing information) in our study is very high: 83.01% and 67.08%, for hormone replacement therapy and use of oral contraceptives respectively. In cases where information exists, 27.37%, underwent hormone replacement therapy and 61.96% used oral contraceptives.

With regard to family history of breast cancer (in 1-degree relatives), a recognized etiologic factor in breast cancer (Rodrigues, 2011), in our study 57.63% of cases, for which this information is existing, have a family history of breast cancer. Of the cases with family history of breast cancer, 26.87% reported a family member of 1 degree with breast cancer. The distribution by degree of relationship is illustrated in Figure 2.11. It is noteworthy to point out that the percentage of missing information for this variable, 78.89%, is extremely high.

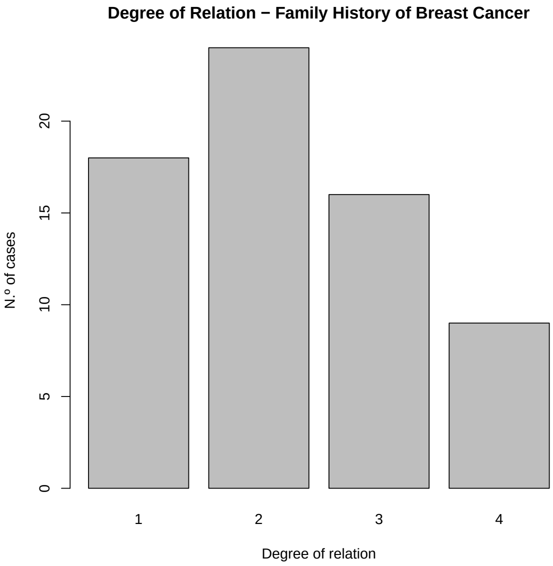


Figure 2.11: Frequency Distribution Graph of Braga's Hospital Patient's Family History of Breast Cancer Degree of Relation.

We accounted for 38 cases with bilaterally cancer cases (i.e., 19 patients with malignant tumours in both breasts, not necessarily at the same time). Note that patients with bilateral cancer were considered as independent cases under medical advice.

Overall, only 81 cases (15.5%) of recurrence were detected, and 11 were local recurrences (i.e. tumour returns to the same, or right near, the original tumour site), 64 distant recurrences (i.e., the tumour has spread to other organs), and 6 cases showed the same two types of recurrence - local and distant. Of these 6, 5 eventually died from breast cancer, 41 of the cases of distant recurrence returned as deaths from breast cancer and, from the cases with local recurrence, we accounted for 3 deaths from breast cancer. Of the 478 cases without recurrence only 6 translated into deaths from breast cancer. But as indicated by the medical consultant a death from breast cancer always present in the moment before death a recurrence of breast cancer. What makes ambiguous the classification "no recurrence" of the latter 6 cases referred. However, we ensure that at least 89.9% of deaths in our sample breast cancer are related to tumour recurrence.

The usual surgical treatment for breast cancer treatment comprises mastectomy with or without axillary lymph node dissection (which includes the modified radical mastectomy and radical mastectomy) or breast-conserving surgery (including lumpectomy and quadrantectomy), with or without emptying axillary. Overall, 97.5% of patients underwent surgical treatment, where 57.42% underwent mastectomy and the remaining 40.07% breast-conserving surgery. Of the 14 (2.5%) who did not undergo surgery (because they refused it or because they only underwent neoadjuvant treatment), 4 eventually die from breast cancer. Results show that 12% of the cases underwent neoadjuvant treatment.

It is noteworthy that 74.6% of deaths from breast cancer were patients who underwent mastectomy. However, this high percentage may be associated prior to the fact that now this type of surgery is associated with a more aggressive type of tumour and larger, than with the surgical procedure itself (Abraúl *et al.*, 2011).

Comparing the cases diagnosed before the creation of Senology unit with cases diagnosed after its creation, is important to note that 77.4% of cases diagnosed before the creation underwent mastectomy, this value decreased to 47.5% after the creation of Senology unit. But the reason for this difference may be due to current guidelines outlined in the document "National Recommendations for Diagnosis and Treatment of Breast Cancer", that indicate that the treatment of breast cancer has undergone tremendous development in recent years, with the tendency to use less invasive approaches. The breast-conservative surgery, where only part of the breast tissue is removed during surgery (as opposed to the entire breast) is accepted as an alternative technique to mastectomy in most patients with ductal carcinoma in situ and invasive carcinoma in early stages. In fact, as will be shown below, it is exactly this type of tumours that are most frequently in our sample. The referred document also indicates that several prospective randomized studies have shown that conservative surgery provides long-term survival rates equivalent to mastectomy in patients with breast cancer in early stages.

The axillary dissection is a surgical procedure to remove all nodes from the armpit to study whether the cancer is at an early or advanced stage, which follows the mastectomy and may or may not follow the conservative surgery. According to the "National Recommendations for Diagnosis and Treatment of Breast Cancer", this procedure is carried out because of its excellent area control and the potential impact on the overall patients' survival. However, being the axillary dissection a procedure that affects the quality of life, limiting patients' lives, it was established the sentinel node biopsy as the technique recommended in patients with clinically negative axilla. Being the recommended axillary dissection only when identifying metastases in sentinel node. The introduction of this technique enables the axillary staging with less surgical morbidity without interfering in overall survival and local recurrence. In this data set from Braga's Hospital, 53% of patients underwent axillary dissection and 44% of patients underwent sentinel node search.

However, comparing the percentage of cases subjected to the above procedures before and after the creation of Braga's Senology unit, it is noted that, in this specific data set, before the existence of the unit 67.7% of patients underwent axillary dissection, in contrast to 46.4% of cases undergoing axillary dissection after the creation of the unit. Opposite to what occurs to the research of sentinel node, where only 1.6% of cases, in this dataset, diagnosed and treated, before the unit of creation, were submitted to this procedure, and after the creation of the unit underwent 65.4% of cases.

Chemotherapy, hormonal therapy and radiotherapy are adjuvant treatment (post-surgery). Although both chemotherapy and hormone therapy are also the usual neo-adjuvant therapies. In our study, more than half of the cases (67.09%) were submitted to chemotherapy, 67.89% of the cases were submitted to radiotherapy and a large percentage (88.35%) were submitted to hormonal therapy. Almost half of the cases (45.29%) were submitted to the three adjuvant treatments referred.

As de Oliveira & da Silva (2011) explain, the staging system adopted for breast cancer is the one proposed by the American Joint Committee on Cancer (AJCC). In this system, T is defined as the category that describes the extent of tumour size, N is the category that describes the nodal status and M describes the presence or absence of distant metastases. The classification of oncological disease in stages - from 0 to IV are grouping categories of TNM classification. It is very important the collection of this information since many of the pathological factors reflected in the TNM classification (tumour size, lymph node status, histological grade, histological type and tumour vascular invasion) are the pillars for physicians establish a prognosis.

Table 2.2, adapted from de Oliveira & da Silva (2011), summarizes the various categories of staging

Table 2.2: American Joint Committee on Cancer TNM staging system for breast cancer.

TNM staging system for breast cancer	
Primary tumour (T)	
TX	Primary tumour cannot be found.
T0	There is no evidence of the primary tumour.
Tis	Intraductal carcinoma, lobular carcinoma in situ, Pagetś disease of the nipple with no tumour invasion in normal breast tissue.
T1	Tumour does not exceed 2.0 cm in its largest dimension.
T2	Tumour greater than 2.0 cm but less than 5.0 cm in its largest dimension.
T3	Tumours greater than 5.0 cm in its largest dimension.
T4	Tumour of any size with direct extension to the chest wall or skin.
Pathological classification of regional lymph nodes (N)	
NX	Regional lymph nodes cannot be assessed (not removed for study or previously removed).
N0	No regional lymph node metastasis.
N1	Metastasis in 1-3 ipsilateral axillary lymph node(s), and/or in ipsilateral internal mammary nodes with microscopic metastasis detected by sentinel lymph node dissection but not clinically apparent.
N2	Metastasis in 4-9 ipsilateral axillary lymph nodes, or in clinically apparent ipsilateral internal mammary lymph node(s) in the absence of axillary lymph node metastasis.
N3	Metastasis in 10 or more ipsilateral axillary lymph nodes; or in ipsilateral infraclavicular lymph nodes; or in clinically apparent ipsilateral internal mammary lymph nodes in the presence of one or more positive axillary lymph nodes; or in more than 3 axillary lymph nodes with clinically negative, microscopic metastasis in internal mammary lymph nodes; or in ipsilateral supraclavicular lymph nodes.
Distant Metastasis (M)	
M0	No distant metastasis.
M1	Distant metastasis.
*In this study we used the pathological classification, commonly referred to as ”pN”, a classification based on axillary lymph node dissection with or without sentinel node biopsy, which maintains the same meaning for the categories.	

In this study, T, N and M categories were analysed individually. For the category T, were recorded 48.8% cases with T1 and 37.3% with T2. This means that the vast majority of cases correspond to lower classes in this category. Only 5.1% of patients had a tumour classified as in situ (Tis class). Only 10 cases (1.79%) had no information for this category. Figure 2.12 represents graphically these results.

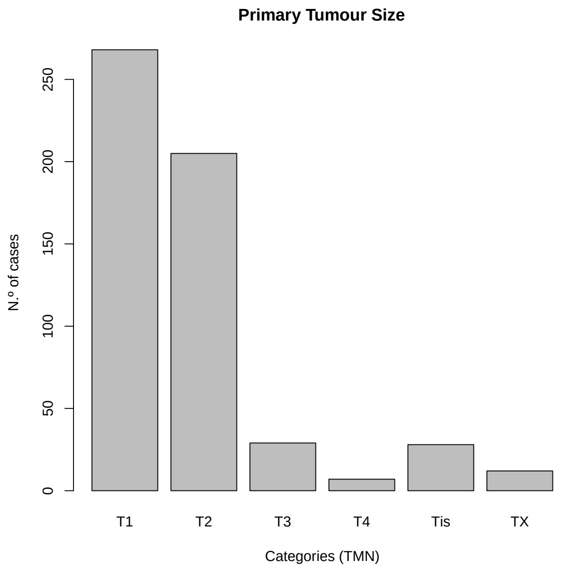


Figure 2.12: Frequency Distribution Graph of Braga's Hospital Patient's Primary Tumour Size Categories (TNM).

Also, N staging category revealed a greater concentration data for the lower classes - N0 and N1 -, representing 49.91% and 30.06% of cases, respectively. For this category we had no information for 20 cases (3.6%). Figure 2.13 represents the distribution of the cases by Regional Lymph Node Involvement.

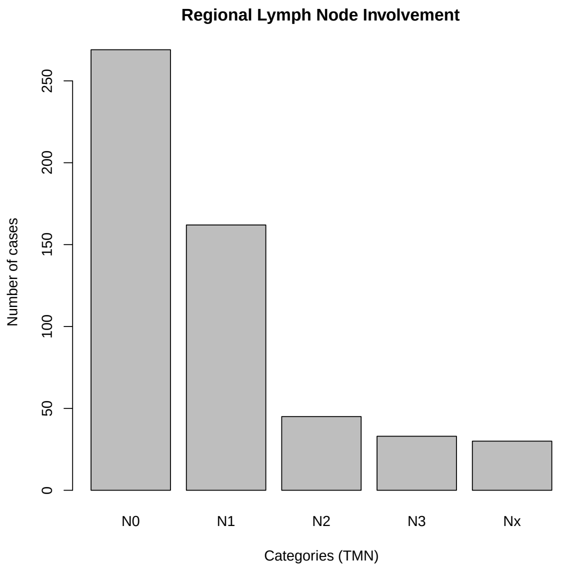


Figure 2.13: Frequency Distribution Graph of Braga's Hospital Patient's Regional Lymph Node Involvement Categories (TNM).

For M staging category, we only accounted for 4 cases (0.73%) in the M1 class (that is, who had distant metastases), with 15.2% of missing cases.

The classification of tumour on a stage is a combination of T, N and M categories. Table 2.3 shows how the classification in the different stages is done according to T, N and M categories. Most of the cases presented stage I (36.45%) and II (40.75%) tumours. Figure 2.14 shows the distribution of stages, where it is visible a higher concentration data for stages I and II. We only accounted for 4.29% of cases without any Stage information.

Table 2.3: American Joint Committee on Cancer Tumour Staging.

AJCC Staging groups		
Stage 0		Tis,N0,M0
Stage I	Stage IA	T1,N0,M0
	Stage IB	T0,N1mi,M0
		T1,N1mi,M0
Stage II	Stage IIA	T0,N1,M0
		T1,N1,M0
		T2,N0,M0
	Stage IIB	T2,N1,M0
		T3,N0,M0
Stage III	Stage IIIA	T0,N2,M0
		T1,N2,M0
		T2,N2,M0
		T3,N1,M0
		T3,N2,M0
	Stage IIIB	T4,N0,M0
		T4,N1,M0
		T4,N2,M0
	Stage IIIC	Any T,N3,M0
Stage IV		Any T, Any N, M1

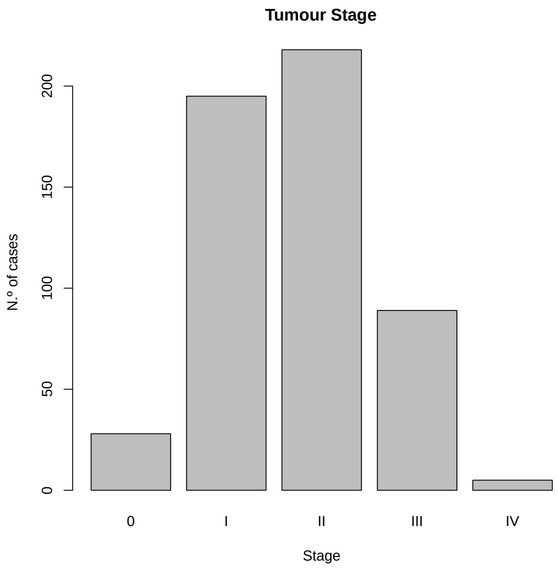


Figure 2.14: Frequency Distribution Graph of Braga's Hospital Patient's Tumour Stage.

For histological type of cancer, it prevails in our study the invasion type of carcinoma (90.38%) - breast cancers occurring in the ducts or the lobules (de Oliveira & da Silva, 2011) - in detriment of carcinoma in situ (9.04%) - in which histological lesions are identified malignant proliferation of neoplastic cells in the ductal phenotype wall and/or in the mammary ducts, without identifying the cancer tissue surrounding the breast ducts (Lopes, 2011). We noticed a large number of invasive ductal carcinoma (67.86%), and they also represent the majority of deaths from breast cancer (77.08%). Table 2.4 summarizes the distribution of carcinoma types under study.

Table 2.4: Distribution of histological type of cancer.

Histological type	**Number of patients**	%
Ductal *in situ*	40	7.56
Invasive ductal	359	67.86
Mixed invasive carcinoma	24	4.54
Lobular *in situ*	6	1.13
Lobular invasive	41	7.75
Other (invasive)	55	10.4
Others	4	0.76
NA's	30	5.37

Note that, in terms of prognosis, invasive ductal type of tumours (usually referred to as invasive ductal Not Otherwise Specified/NOS) belong to the poor prognosis group, such as a mixed invasors compared to invasive lobular or otherwise (de Oliveira & da Silva, 2011). The in situ types of carcinoma, will eventually have a more favorable prognosis than the invasors. In the study sample, from the cases of carcinoma in situ, only 3 resulted into death from breast cancer (6.25% of deaths from breast cancer).

It was assumed that the invasion of blood and lymphatics vessels of tumour emboli can be a prognostic information as powerful as Staging (Schoppmann *et al.*, 2004). The prognostic value of vascular invasion was confirmed in studies concerning survival and is considered one of the most important factors for predicting local recurrence after conservative surgery (Schoppmann *et al.*, 2004). The study of Schoppmann *et al.* (2004) concludes, in fact, that patients without lymph node involvement but with lymphovascular invasion have mortality rates higher than patients without lymphovascular invasion. In our sample, only 4.65% of the cases showed venous vascular invasion images and 17.53% presented lymphatic vascular invasion images.

The Bloom & Richardson (1957) classification divides the histological grade in three categories: grade I (well-differentiated carcinoma); grade II (moderately differentiated carcinoma) and grade III (poorly differentiated carcinoma). The histological grade, as a prognostic factor, has been widely studied and its impact on the survival of patients was confirmed - survival gets progressively worse as the histologic grade increases (Cianfrocca & Goldstein, 9). In our sample cases that present a histological grade II are predominant, representing 43.84% of the cases. In lower percentage are the cases with grade III (23.09%). Note that in our study the Gx rating means it was not possible to classify the histological grade of the tumour and quantifies only 2 cases analysed. 8.59% of the cases did not present any information regarding the histological grade of the tumour. Figure 2.15 represents the distribution of the cases by degree of differentiation.

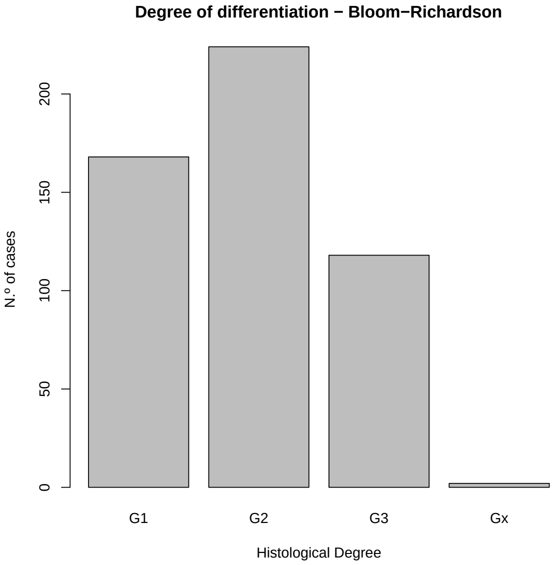


Figure 2.15: Frequency Distribution Graph of Braga's Hospital Patient's Tumour Degree of Differentiation.

The characterization of breast carcinoma with respect to expression of estrogen receptor (ER), progesterone (PR), and the human epidermal growth factor receptor 2 oncogene (HER-2 neu) is a usual procedure and required by doctors. A positive expression of ER and RP is usually related to a good prognostic. As Gobbi *et al.* (2008) concludes that tumours presenting an overexpressing HER2-neu are associated with a worse prognosis and have a predictive value in breast cancer.

Abraúl *et al.* (2011) clarify that a more accurate histological and molecular characterization of breast cancer, with the systematic use of molecular markers with predictive value of response to treatment, has allowed more appropriate and therapeutic individualization. As so, it is problematic the existence of missing values, in our study, for RE receptors, PR and Her-2 neu (13.77%, 17.89% and 25.75%, respectively).

Dawson *et al.* (2009) explain triple negative breast cancers (TNBC) are dened by the absence of estrogen, progesterone and HER2-neu expression. Although the prognostic significance of triple negative tumours remains unclear we decided to create a binary variable to indicate the cases with TNBC and the ones with non TNBC. As a result, in our study we were able to account for 4.38% of TNBC cases that represents 21.28% of the deaths from breast cancer.

Also the proliferative activity of cancer - particularly the Ki.67 antibody immunostaining index (high versus low) - is considered important in determining the prognosis (Fitzgibbons *et al.*, 2000). A high expression of Ki67 is associated with a worst prognostic as, for example, Viale *et al.* (2008) indicate. In our sample 56.7% of cases showed a high proliferative activity of Ki.67 antibody. However, for 43.83% of cases no information for this variable is available.

We will now present some descriptive of the longitudinal variables.

As Duffy *et al.* (2000) explain, Carcinoma Antigen 15-3 (CA15-3) is the most widely used serum marker in breast cancer. It consists in a large transmembrane glycoprotein which is frequently overexpressed and aberrantly glycosylated in cancer. Currently the main uses of CA15-3 are in preclinically detecting recurrent breast cancer and monitoring the treatment of patients with advanced breast cancer, although its clinical value is not validated in a high-level evidence study, as pointed out by the American Society of Clinical Oncology. Nevertheless, as Harris *et al.* (2007) point out, there are several well-designed studies that show that an increase in CA15-3 after primary and/or adjuvant therapy, can predict recurrence an average of 5 to 6 months before other symptoms or tests.

The tumour marker Carcinoembryonic antigen (CEA) is usually used for therapy monitoring in advanced disease, although recent reports, e.g. Guadagni *et al.* (2001), discourage its routine use because of low sensitivity. The authors conclude that its use should be considered as an inecient method of follow-up evaluation for breast cancer patients, and it provides no additional value when used in combination with CA15-3. Nevertheless, as Sturgeon *et al.* (2008) point out, on occasion, it can be informative when levels of CA15-3 remain below the cut-off point.

We only have information on tumour marker CA15-3 values for 534 patients. This translates into a total number of 552 cases analysed, since 18 cases presented bilateral breast cancer. The total number of deaths from breast cancer is 55. There were 5166 measurements of tumour marker CA15-3, with a number of observations per patient varying between 1 and 48 measurements. Being 8 the median number of measurements per person. Figure 2.16 shows the distribution of the number of measurements per patient on the CA15-3 tumour marker.

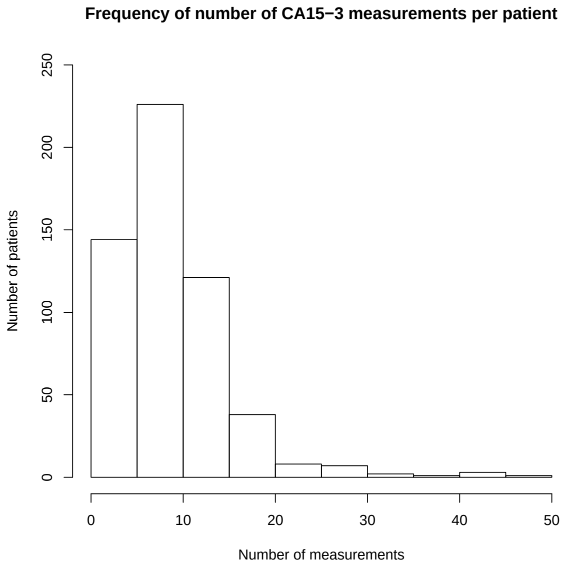


Figure 2.16: Histogram for the number of measurements per patients for tumour marker CA15-3.

We only have information on tumour marker CEA values for 532 patients. Since 19 patients had bilateral breast cancer it translates into a total number of 551 cases analysed. The total number of deaths from breast cancer is 54. There were 4166 measurements of tumour marker CEA, with a number of observations per patient varying between 1 and 23 measurements, as shown in Figure 2.17. The median number of measurements per person is 7.

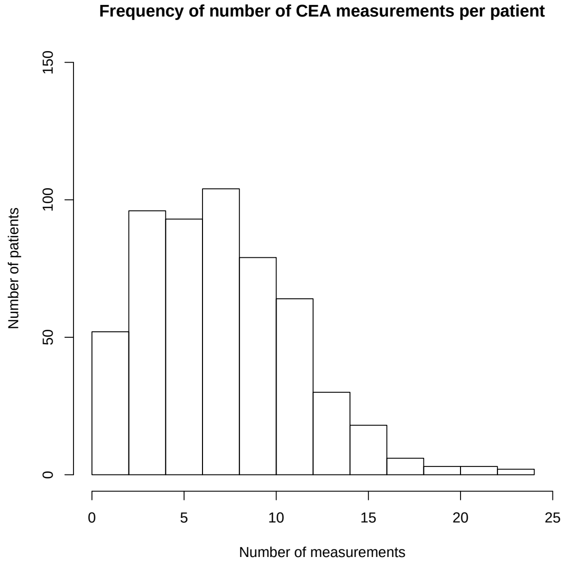


Figure 2.17: Histogram for the number of measurements per patients for tumour marker CEA.

3

Survival Analysis

Many medical or biological studies focus its analyses investigating time to a certain event, for example, time it takes for a patient to respond to a therapy, time until disease recurrence or relapse or time until death. When the response variable of interest is time from an origin to an event of interest predefined, survival statistical methods must be used (Cox, 1972), modelling the hazard of the event at each time point.

However, it is frequent in studies of this natures that data are collected over a finite period of time and hence the event of interest may not be observed for all the individuals under study. Which means that the time-to-event of those individuals who have not experienced the event is censored. This results in what is usually called censored data, and it contains important information regarding the time an individual has been in risk of failure. For this reason, it is necessary to accommodate those observations in survival models.

The combination of censoring and, also the heterogeneity in each individual's times until the event of interest that may exist, generates unusual difficulties in the analysis of this type of data that cannot be handled properly by the standard statistical methods. Which justifies the need for specific models, such as survival models, to deal with time-to-event data.

Survival models allows the investigator to be able to answer questions such as: What is the effect of a treatment on survival time? Which factors are associated with the time duration of a disease? What are the predictors for death or even for cancer recurrence, considering the time between relapses?

The work presented in this chapter is an extension of the work already published by Borges *et al.* (2014). This Chapter begins with the clarification of the methodology applied in the survival analysis of the patients from Braga's Hospital Senology unit diagnosed with breast cancer. It is introduced by a summary explanation of the main approaches adopted in studies of this nature, and also describing the models applied on the present analysis. Subsequently, the results from the survival analysis of the data set in study will be exposed, ending with a sub-chapter that resumes a discussion of the significant results.

The entire analysis was performed with open source statistical software R, in particular making use of *Survival* (Therneau, 2015) and *flexsurv* (Jackson, 2015) packages.

3.1 Definition of Concepts

We will present some notation and concepts to describe the distribution of time to the event for a given population.

Let T be the random variable representing the time until the event of interest.

It would be ideal if we could observe the exact failure times of a random sample of n individuals: $(t_1, ..., t_n)$, where for each individual $i = 1, ..., n$ the observation t_i is a realization of the random variable T_i with the same (common) distribution depending on parameters that we want to estimate.

However, for certain individuals we will not observe T, but rather a realization c_i, of a random variable C_i, representing the time censored. The only thing we can assure about the failure time, when we observe a censored time, is that failure time is certainly greater than the censored time, namely, that $T_i > C_i$.

What we got to observe then is the minimum of two values, which gives rise to a set of realizations $(t_1, c_2, c_3, ..., t_{12}, c_{13}, ..., c_{n-1}, t_n)$.

Hence, what is actually observed is a realization of the variable $S_i = min(T_i, C_i)$, time of survival, that is, the minimum between failure time and censored time.

To distinguish what is failure time and what is censored time we consider $\delta_i = I(T_i \leq C_i)$, an indicator that takes value 1 if the observed is the failure time, and zero otherwise. Thus, a set of survival data is a pair (s_i, δ_i).

The distribution of the random variable T of interest can be described in several ways, using previously known functions. For example, by the already known cumulative distribution function:

$$F(t) = P(T \leq t), \quad t \geq 0, \tag{3.1}$$

which is right continuous, i.e., $\lim_{u \to t^+} F(u) = F(t)$. When T is a survival time, $F(t)$ is the probability that a randomly selected subject from the population will experience the event before time t. If T is a continuous random variable, then it has a density function $f(t)$, which is related to $F(t)$ through the following equation:

$$f(t) = \frac{dF(t)}{dt}, \tag{3.2}$$

or even, through

$$F(t) = \int_0^t f(u)du. \tag{3.3}$$

Since we are interested in the time of an individual survive, i.e., probability of an individual surviving beyond time t, we are, hence, interested in the survival function defined as:

$$S(t) = P(T \geq t) = 1 - F(t). \tag{3.4}$$

The survival function $S(t)$ is a non-increasing function over time, taking the value 1 when $t = 0$, i.e., $S(0) = 1$. Being T a continuous random variable, it has the characteristic $S(\infty) = 0$, which means that everyone will experience, at some moment, the event. However, there are cases where it is possible to make $S(\infty) > 0$, which corresponds to the situation where there is a positive probability of not dying or not experiencing the event. For example, if the event of interest is the time from a response until the returning of the disease, and if there is a possible cure to the disease, then in this case $S(\infty) > 0$, where $S(\infty)$ is the proportion of subjects healed.

We can also establish the following relationships between these functions:

$$S(t) = \int_t^\infty f(u)du, \tag{3.5}$$

or even

$$f(t) = -\frac{dS(t)}{dt}, \tag{3.6}$$

The hazard rate is a convenient way to describe the distribution of time to event, as it has a very natural interpretation related to how the event occurs in the population. We will begin to define mortality rate, which is a discrete version of the hazard rate.

The mortality rate for a particular time t, where t is an integer usually measured in units of time (e.g. years, months, days, etc.), the proportion of people who experienced the event between time t and $(t+1)$ between the living individuals at time t, i.e.:

$$m(t) = P(t \leq T < t+1 \mid T \geq t). \tag{3.7}$$

The hazard rate, $h(t)$, is the limit of the mortality rate if the time interval is considered very short (rather than a unit). The hazard rate is the instantaneous rate of the event occurring at time t given that the individual has survived up to time t.

Specifically, the hazard rate is defined as:

$$h(t) = \lim_{\Delta t \to 0} P(t \leq T < t + \Delta t \mid T \geq t). \tag{3.8}$$

Hence, if Δt is very small, we will have:

$$P(t \leq T < t + \Delta t \mid T \geq t) \approx h(t)\Delta t. \tag{3.9}$$

This definition implies other relations:

$$h(t) = \frac{\lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t)}{P(T \geq t)} = \frac{f(t)}{S(t)} = -\frac{d\text{log}S(t)}{dt}. \tag{3.10}$$

By integrating both sides of this last equality we obtain the cumulative hazard function:

$$H(t) = \int_0^t h(u)du = -\log S(t). \tag{3.11}$$

Hence, we obtain:

$$S(t) = \exp\left(-H(t)\right) = \exp\left(-\int_0^t h(u)du\right), \tag{3.12}$$

and also:

$$f(t) = h(t)\exp\left(-\int_0^t h(u)du\right), \tag{3.13}$$

The likelihood function of θ, a vector of parameters, given the data set $Y = (T_i, \delta_i)$, $i = 1, ..., N$, where N is the size of a generic sample, is given by:

$$L\left(\theta \mid Y\right) = \prod_{i=1}^{N} f(t_i \mid \theta)^{\delta_i} S(t_i \mid \theta)^{1-\delta_i}. \tag{3.14}$$

From the relations in 3.10 and 3.13 we can rewrite 3.14 as it follows:

$$L\left(\theta \mid Y\right) = \prod_{i=1}^{N} h(t_i \mid \theta)^{\delta_i} \exp\left(-\int_0^{t_i} h(u \mid \theta)du\right). \tag{3.15}$$

3.2 Censoring and Truncation

When working with real data, especially in a study in which time marks such as date of beginning of the study (that could be the date of the tumour diagnose) and end of study (that could be the date of death) are extremely important, we need to take into account the existence of censoring and, even truncation. In this particularly study we detected the need to acknowledge left truncation and right censoring of the data.

We consider the beginning of the study as the date of the malignant tumour diagnosis and the end of the study the date 30/11/2014. Because subjects enter this study at different times and the terminal point of the study is predetermined by us, they have their own specific, fixed, censoring time. This form of censoring has been named generalized Type I censoring (David & Moeschberger, 1978).

Another type of censoring scheme that also arises in this type of data is interval censoring. This occurs when all that is known is that the event of interest occurs between two known times, in a time interval.

As previously mentioned, the Senology Unit of Braga's Hospital was only created in 2008, however we have information on patients who were diagnosed with malignant tumour before that time and alive at the Unit's creation. As we do not have any information on patients who were diagnosed before 2008 and died before that year, we have to consider a left truncation of the data in that year.

Klein & Moeschberger (2003) clarify that truncation of survival data occurs when only those individuals whose event time lies within a certain observational window (Y_L, Y_R) are observed. When Y_R is infinite then we have left truncation. An individual whose event time is not in this interval, is not observed and no information on this subject is available to the investigator. Also, they alert to the fact that because we are only aware of individuals with event times in the observational window, the inference for truncated data is restricted to conditional estimation.

Note that, in our specific case, patients alive in 2008 and diagnosed before will naturally be the "stronger" compared to others that had already died and were not observed.

Adapting Klein & Moeschberger (2003) notation, for a specic subject under study, we assume that there is a lifetime V and a fixed censoring time, C_r (C_r for "right" censoring time). The $V's$ are assumed to be independent and identically distributed with probability density function $f(v)$ and survival function $S(v)$. The exact lifetime V of an individual will be known if, and only if, V is less than or equal to C_r. If V is greater than C_r, the individual is a survivor, and his event time is censored at C_r.

The likelihoods for this two types of censoring may be written by incorporating the following components: for the right-censored observations $S(C_r)$; and for the left-truncated observations $\frac{f(x)}{S(Y_L)}$. Hence, in our study, the likelihood function may be constructed by putting together the component parts as:

$$L\left(\theta \mid Y\right) = \prod_{i \in D} \frac{f(x_i)}{S(Y_{Li})} \prod_{i \in R} S(C_r), \tag{3.16}$$

where D is the set of death times and R the set of right-censored observations.

3.3 Non-Parametric Survival Models

The Kaplan-Meier estimate (Kaplan & Meier, 1958), also called the Product-Limit estimator, is a nonparametric estimate of the survival function, $S(t)$ widely used in survival analysis. This estimate is a step function with jumps at observed event times, t_i. The Kaplan-Meier estimate of the survival function is given by:

$$\widehat{S}(t) = \begin{cases} 1 & \text{if } t < t_1 \\ \prod_{t_i \leq t} \left[1 - \frac{d_i}{Y_i}\right] & \text{if } t_1 \leq t \end{cases}, \tag{3.17}$$

where $0 < t_1 < t_2 < ... < t_D$, also the number of individuals with an observed event time t_i is d_i, and the value Y_i represents the number of individuals at risk at time t_i (individuals who die at time t_i or later).

The variance of the Product-Limit estimator is estimated by Greenwood's formula (Klein & Moeschberger, 2003), given by:

$$\widehat{V}[\widehat{S}(t)] = \widehat{S}(t)^2 \sum_{t_i \leq t} \frac{d_i}{Y_i(Y_i - d_i)}. \tag{3.18}$$

However, Kaplan-Meier estimator only applies to right censored type of data (Klein & Moeschberger, 2003). When the data is left truncated and right censored, the Kaplan-Meier estimator may have a relatively large variance for small t, where the risk sets are small, and also for large t. This early instability in the point estimator of the survival function may propagate throughout the entire curve. Lai & Ying (1991) suggested a solution to this problem by a slight modication of the Product-Limit estimators where deaths are ignored when the risk set is small. Their estimator is given by:

$$\widehat{S}(t) = \prod_{t_i \leq t} \left(1 - \frac{d_i}{Y_i} I[Y_i \geq cn^{\alpha}]\right), \tag{3.19}$$

where I is the indicator of a set A, n is the sample size, and $c > 0$, $0 < \alpha < 1$ are constants. This estimator is asymptotically equivalent to the usual Product-limit estimator.

Note that for right-censored data Y_i, is, in fact, the number of individuals on study at time 0 with a study time of at least t_i. For left-truncated data, Y_i can be redefined as the number of individuals who entered the study prior to time t_i and who have a study time of at least t_i, that is, Y_i is the number of individuals with $L_j < t_i T_j$.

To evaluate the statistical significance of differences among the Kaplan-Meier survival curves of each variable category, we made use of the G^ρ family of tests of Harrington & Fleming (1982), with weights on each death of $S(t)^\rho$, where $S(t)$ is the Kaplan-Meier estimate of survival, implemented in the statistical software used. Since we made $\rho = 0$, we actually made use of the log-rank or Mantel-Haenszel test where the null hypothesis being tested is: no difference between (true) survival curves.

It should be emphasized that a limitation of non-parametric methods is that they do not allow the examination of the effect of covariates on survival time.

3.4 Parametric Survival Models

As Cox *et al.* (2007) refer parametric models can be attractive because standard methods such as maximum likelihood are available for parameter estimation and testing and a complete description of the hazard function can be obtained.

In fact, such models play an important role in the survival analysis since many medical and biological statistical analyses are performed using parametric techniques. Such models are constructed from families with known probability distributions such as Exponential and Weibull. Thus, it was considered appropriate to explain, although briefly, these two families, in the survival context.

Exponential Model

The Exponential distribution is characterized for having constant hazard $\lambda(t) = \lambda$. Thus, the density function for the survival time T, is given by:

$$f(t) = \lambda \exp(-\lambda t), \tag{3.20}$$

and the respective survival function by:

$$S(t) = \exp(-\lambda t). \tag{3.21}$$

The corresponding likelihood function of λ, given the data set $Y = (T_i, \delta_i)$, $i = 1, ..., N$, where N is the size of a generic sample, can be written as:

$$L(\lambda \mid Y) = \prod_{i=1}^{N} (\lambda \exp(-\lambda t_i))^{\delta_i} S(t_i \mid \lambda)^{1-\delta_i} = \lambda^{\sum_{i=1}^{N} \delta_i} \exp\left(-\lambda \sum_{i=1}^{N} t_i\right). \tag{3.22}$$

Note that, regardless of how much time an individual has lived, the risk of failure is always the same. For processes where the risk changes with time, this model will not be appropriate.

Weibull Model

Considering that T is Weibull distributed with shape and scale parameters σ and α, respectively, the density function is given by:

$$f(t) = \alpha\sigma^{\alpha} t^{\alpha-1} \exp\left(-(\sigma t)^{\alpha}\right). \tag{3.23}$$

And the survival function as:

$$S(t) = \exp\left(-(\sigma t)^{\alpha}\right). \tag{3.24}$$

The respective hazard and cumulative hazard functions are given by:

$$h(t) = \sigma^{\alpha}\alpha t^{\alpha-1}, \tag{3.25}$$

and

$$H(t) = (\sigma t)^{\alpha} t. \tag{3.26}$$

The logarithm of the hazard function is a linear function of log time with constant $(\alpha \log \sigma + \log \alpha)$ and slope $(\alpha - 1)$. Thus, the hazard is rising if $\alpha > 1$, constant if $\alpha = 1$, and declining if $\alpha < 1$.

The corresponding likelihood function of (σ, α) can be written as:

$$L(\sigma, \alpha \mid Y) = \prod_{i=1}^{N} f(t_i \mid \sigma, \alpha)^{\delta_i} S(t_i \mid \sigma, \alpha)^{1-\delta_i} = \alpha^{\sum_{i=1}^{N} \delta_i} \exp\left\{\sum_{i=1}^{N} (\delta_i \log \sigma + (\alpha - 1)\delta_i \log(t_i) - \sigma t_i^{\alpha}\right\}. \tag{3.27}$$

In the field of fully parametric models, in the context of survival, is important to refer the work of Stacy (1962) in the generalization of the gamma (GG) distribution that allowed Ardoino *et al.* (2012) to use the *GG* model to investigate the shape of the hazard function in early breast cancer patients. As they refer its use may be particularly advantageous, as it allows testing different shapes of the hazard function. Later on, Cox *et al.* (2007) developed a taxonomy of the hazard functions of the *GG* family, which includes various special distributions and allows depiction of effects of exposures on hazard functions.

Also, the generalized F distribution (*GF*), a larger family of functions that contains the *GG* function, was applied in the survival context by Ciampi *et al.* (1986) and later Cox (2008) provided a characterization of the hazard functions of the *GF* model.

Nonetheless, in general, fully parametric model with a low number of parameters (such as Weibull, exponential, lognormal, loglogistic, among others) are considered insufficiently exible, especially for studying long follow-up times (Ardoino *et al.*, 2012).

3.5 Semi-Parametric Survival Models

There are also different approaches to the analysis of survival data which are not fully parametric, as for example Herndon & Harrell (1990) and Herndon & Harrell (1995) who incorporate cubic splines in the hazard model or Rosenberg (1995)that made use of B-splines in the hazard function estimation. Rossini & Tsiatis (1996) also propose a semiparametric proportional odds model for the analysis of current status data. Royston & Parmar (2002) propose a flexible parametric models for censored survival data, which will be explained in detail later on. Putter *et al.* (2005) made the more complete use of time-varying covariates.

More recently, Cadarso-Suarez *et al.* (2010), proposed an extension to these models, having a flexible hazard ratio curves for continuous predictors, with an application for breast cancer. And Zou *et al.* (2011), suggests an alternative to the Cox Proportional hazard model (that would be explained following), when the proportional hazard assumption is not satisfied, constructing a semiparametric accelerated failure time partial linear model and developing its estimation method based on the penalized spline (P-spline) and the rank estimation approach. Also, Aalen's additive model is an alternative approach when proportional hazard assumption is not verified (Aalen, 1989).

However, the most popular semi-parametric approach to model survival is the well-known Cox proportional hazards model (Cox, 1972), which will be detailed following.

3.5.1 Cox Proportional Hazards Model

Cox proportional hazards model (CPHM) is, doubtless, the most commonly used model in survival analysis.

The CPHM defines the probability of survival as a function of time t, for a vector of covariates x_i, as it follows:

$$S(t \mid x_i) = [S_0(t)]^{\exp(x_i^{'}\beta)}, \tag{3.28}$$

where $S_0(t)$ is the baseline survival function and β the vector of coefficients estimated.

The corresponding hazard function can be written as:

$$h(t \mid x_i) = h_0(t)\exp(x_i^{'}\beta), \tag{3.29}$$

where $h_0(t)$ is the baseline hazard function.

This model implies that the risk ratio between any two individuals with covariates vectors, respectively, x_1 and x_2, given by:

$$\frac{h(t \mid x_1)}{h(t \mid x_2)} = \frac{h_0(t)\exp(x_1^{'}\beta)}{h_0(t)\exp(x_2^{'}\beta)} = \exp\left\{(x_1 - x_2)^{'}\beta\right\}, \qquad \text{for all } t \geq 0, \tag{3.30}$$

is constant and not time dependent. Justifying the name of proportional hazards model.

Cox model is said to be a semi-parametric model because $h_0(t)$ is non-parametric, it is a baseline hazard function non specified, and $\exp(x_1^{'}\beta)$ is parametric.

Cox (1975) proposed an estimation of the primary parameters of interest, β, by maximization of a partial log-likelihood function. Which is none other but the likelihood of the data conditioning in times of failure.

To briefly describe the partial log-likelihood function proposed let's consider the failure and censor times ordered. Let t be a time where an individual has experienced the event, and $R(t)$ be the set of individuals at risk at time t. This function considers, for each individual i with failure time, the probability of the individual who failed in interval $[t, t + \delta t)$ actually be the individual i, as:

$$P\left\{i \text{ fails in } [t, t+\delta t) \mid \text{one failure occurred in } [t, t+\delta t), R(t)\right\} = \frac{\exp(x_1^{'}\beta)}{\sum_{k \in R(t_i)} \exp(x_k^{'}\beta)}. \tag{3.31}$$

Making the product for all individuals with failure time, comes:

$$PL(\beta) = \prod_{i=1}^{n} \left\{ \frac{\exp(x_1^{'}\beta)}{\sum_{k \in R(t_i)} \exp(x_k^{'}\beta)} \right\}. \tag{3.32}$$

Note that the partial likelihood function does not depend on $h_0(t)$, hence there is no need to assume a distribution of the latter.

Summarizing, the particularity of this model, which can be viewed as one of its advantages, is that the estimation of the coefficients does not require the formulation of the baseline cumulative survival function as it gets absorbed when the coefficients are estimates by the method of partial log likelihood.

3.5.2 Flexible Parametric Survival Models

Despite the vastly usage of CPHM, there are some constrains to the application of these models namely, the proportional hazard assumption and the non formulation of the baseline cumulative survival function, that can be of medical interest.

In this analysis, we made use of Flexible Parametric Survival Models, in particular a flexible Royston-Parmar survival model (FRPM) (Royston & Parmar, 2002), to estimate hazard ratios over time since diagnosis by a set of statistical significant covariates. And, also the well-known CPHM, only for the purpose of comparing estimates.

As Royston & Parmar (2002) stress out it is often of interest, in medical studies, the smoothly estimation of the baseline hazard function as it is directly related to the time-course of an illness. Also the incorporation of time varying regression coefficients on the CPHM, when the assumption of proportional hazards is violated, results in a complex practical interpretation of the coefficients and in the robustness of the resulting model. So it is of interest the use a more flexible model where the visualization of the hazard function is much easier.

For that we choose to work with the approach proposed by Royston & Parmar (2002) where they model the logarithm of the baseline cumulative hazard function as a "natural" cubic spline function of log time.

The FRPM comes from the family of functions that are based on transformation of the survival function by a link function $g(.)$:

$$g\left[S(t \mid x_i)\right] = g\left[S_0(t)\right] + x_i\beta. \tag{3.33}$$

Since we are interested in estimating hazard ratios, and as Royston & Parmar (2002) suggest, we use natural cubic splines to model $g[S_0(t)]$ within the Aranda-Ordaz family of link functions (Aranda-Ordaz, 1981):

$$g(x;\theta) = \log\left(\frac{x^{-\theta}-1}{\theta}\right). \tag{3.34}$$

Choosing to work with hazards scaling by making $\theta \to 0$, rather than with more general values of θ as they refer that the interpretation of covariate effects would turn out obscure.

Thus, the model transformation may be written as it follows:

$$g\left[S(t \mid x_i)\right] = \log\left[H(t \mid x_i)\right] = \eta_i = s(\log(t) \mid \gamma, k) + x_i\beta. \tag{3.35}$$

where $H(t \mid x_i)$ is the cumulative hazard function and s is a natural cubic spline acting on a log scale of time t, with adjustable parameter vector and k knots.

The choice of number of knots can be made using the minimum combination of Akaike Information Criterion (AIC) (Akaike, 1973).

The respective survival function can be formulated as:

$$S(t \mid x_i) = \exp\left(-\exp(\eta_i)\right). \tag{3.36}$$

And the hazard function as:

$$h(t \mid x_i) = \left(\frac{ds(\log(t) \mid \gamma, k_0^{'})}{dt}\right)\exp(\eta_i). \tag{3.37}$$

The contribution to the log likelihood for the *ith* individual for a flexible parametric model on the log cumulative hazard scale can be written as (Lambert & Royston, 2009):

$$\log L_i(\beta) = \delta_i\left(\log[s^{'}\log(t_i) \mid \gamma, k_0] + \eta_i\right) - \exp(\eta_i). \tag{3.38}$$

Making the product for all individuals with failure time, comes:

$$\log L(\beta) = \prod_{i=1}^{n}\left(\delta_i\left(\log[s^{'}\log(t_i) \mid \gamma, k_0] + \eta_i\right) - \exp(\eta_i)\right). \tag{3.39}$$

3.6 Model Diagnostic

Since both CPHM and FRPM rely on the assumption of proportional hazards we preformed a statistical test proposed by Grambsch & Therneau (1994) based on the Schoenfeld residual calculation, to assess the validity of that assumption.

Also we graphical represented the Cox-Snell residuals (Cox & Snell, 1968), a type of residuals for survival models, for assessing the overall goodness-of-fit of the CPHM.

If the model adjusted is correct it is well known that, if we make the probability integral transformation on the true death time T, the resulting random variable has a uniform distribution on the unit interval or that the random variable $U = H(T_j|x_j)$, where x_j is a vector of all fixed-time covariates, has an exponential distribution with hazard rate 1. Here, $H(T_j|x_j)$ is the true cumulative hazard rate for an individual with covariate vector x_j. If the estimates of the β's from the model are $b = (b_1, ..., b_p)$, then, the Cox-Snell residuals are defined as (Klein & Moeschberger, 2003):

$$r_{CS_i} = \widehat{H}_0(T_j) \exp\left(\sum_{k=1}^{p}(x_{ijk}, b_k)\right), \qquad j = 1, ..., n. \tag{3.40}$$

where $\widehat{H}_0(t)$ is the baseline hazard rate estimated.

If the model is correct and the $b's$ are close to the true values of β then, the $r'_{CS_i}s$ should be a censored sample from a unit exponential distribution.

A graphical assessment of goodness-of-fit is to compare the survival function of the unit exponential distribution $S_{\exp}(t) = \exp(-t)$, with the Kaplan-Meier estimate of the survival function of r_{CS_i} (Rizopoulos, 2012). Accounting for the fact that $r'_{CS_i}s$ are a censored sample from a unit exponential distribution.

Also, for model diagnose we made use of a graphical superimposing of the nonparametric survival curve (Kaplan-Meier) versus the FRPM curve versus the CPHM curve, for a certain combination of covariates values.

3.7 Breast Cancer Survival Analysis

3.7.1 Kaplan-Meier Analysis

The exploratory analysis presented in Chapter 2, reports, for the 540 patients eligible for the present analysis, a majority of early-stage breast cancer cases. We mean by early-stage breast cancer, cases whose values are on the first categories of tumour stage, size and grade. This can be related to an increased screening, leading to early detection of the tumor. It is an important result as it may translates in a high probability of survival of these patients.

Figure 3.1 presents the Kaplan-Meier survival curves for all patients considering only right censoring (dashed line) and considering right censoring and left truncation data (solid line). The graph reveals that the Kaplan-Meier estimate for 10-year survival of these patients is quite near of 80%. Also, it seems that for the entire observed follow up time the survival probability is above the 50%.

The importance of acknowledging left truncation is well expressed in the differences between the Kaplan-Meier estimates. As it is visible that ignoring left truncation leads to an overestimation of survival. It would be expected to observe this difference as the Kaplan-Meier allowing for left truncation is giving less weight to the survival times of patients still alive in January 2008.

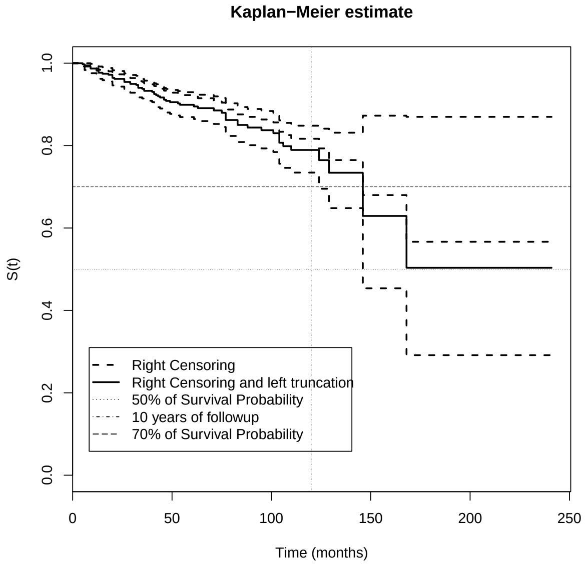


Figure 3.1: Kaplan-Meier curve estimates of Breast Cancer patients of Braga's Hospital.

In order to estimate the relative risk of death from breast cancer we proceeded initially with the calculation of the Kaplan-Meier estimates stratified by category, and compared these through their graphical representation.

A total of 16 variables turned out to have significant differences in their categories survival curves. Which means that they have a statistical significant effect on patient's risk of dying from breast cancer. Table 3.1 presents the results on significance of Log-rank test, under the null hypothesis that the survival curves of the different categories are all equal. In this table only the variables found significantly different are presented.

All the others do not reject the null hypothesis. Figures 3.2 through 3.20 present the Kaplan-Meier curves for the categories of each variable with significant effect on patients survival.

Table 3.1: Log-Rank Test Results for variables with significant effect on patients Survival on Breast Cancer

Log-rank Test Results	
Variable	**p-value**
Tumour Stage	< 0.001
Bilateral Breast Cancer	0.002
Breast Cancer Recurrence	< 0.001
Neoadjuvant Treatment	< 0.001
Type of Surgery	< 0.001
Venous Vascular Invasion	< 0.001
Sentinel Lymph Node Biopsy	0.02
Estrogen Receptor Expression	< 0.001
Progesterone Receptor Expression	< 0.001
Triple Negative	< 0.001
Degree of differentiation	< 0.001
Size of Primary Tumour	0.005
Regional Lymph Node Involvement	< 0.001
Age at Diagnosis (Categorized)	0.017
Images of lymphatic invasion	0.04
Hormonotherapy	< 0.001

From results presented in Table 3.1 and observing Figure 3.2, we can accept that tumour stage is a significant risk factor for the survival of these patients. Also, log-rank test results confirm that there is no significant difference between the survival rate between the first three categories of tumour Stage (0, I and II) (p-value = 0.851), and also between the last two categories of this covariate (III and IV) (p-value = 0.714). Hence we were able to group the first three categories into only one category (0/I/II) and to group the two last categories into one (III/IV), that according to log-rank test results the corresponding survival curves are significant different (p-value < 0.001). Figure 3.3 suggest that the cases diagnosed in stage III or IV have lower probability of survival compared to a tumour in stage 0, I or II, for these patients.

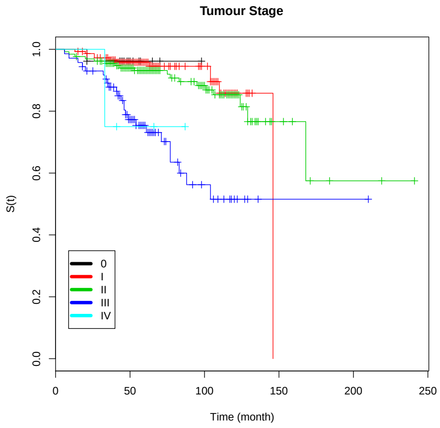


Figure 3.2: Kaplan-Meier curves for categories of variable tumour stage.

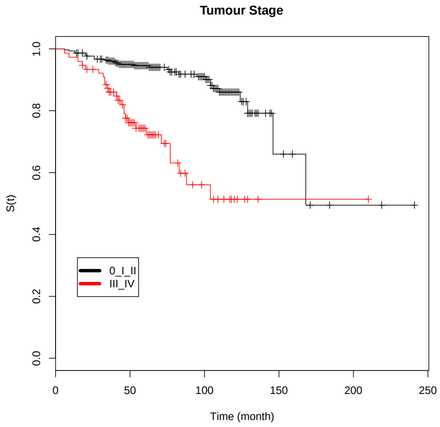


Figure 3.3: Kaplan-Meier curves for categories of variable tumour stage with different categorization.

Cases that presented a bilateral tumour (tumour in both breasts) have higher risk of dying from breast cancer compared to unilateral cases of tumour, as expected by doctors.

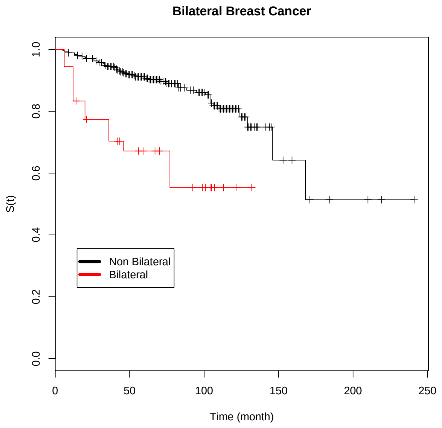


Figure 3.4: Kaplan-Meier curves for categories of variable bilateral breast cancer.

As expected, there is a relationship between breast cancer recurrence and a higher risk of dying from the disease, as can be perceived in Figure 3.5. Although, as mentioned in the previous descriptive analysis, not all cases of recurrence translated in cases of death from breast cancer, our medical consultant informed that cases of death from breast cancer developed in some moment a recurrence of breast cancer (even in the moment right before the time of death). That ambiguity was responsible for the non-incorporation of this variable in the survival model. Meaning that indicator of death and recurrence are equal.

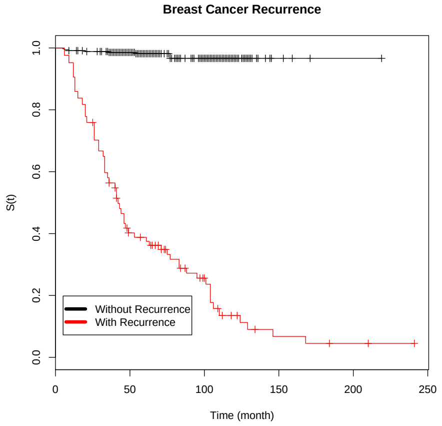


Figure 3.5: Kaplan-Meier curves for categories of variable breast cancer recurrence.

Patients that were submitted to neoadjuvant treatment, as shown in Figure 3.6, have a higher probability of dying from breast cancer, probably, as already mentioned in previous chapters, due to tumour characteristics. Since cases with a non-operable tumour (for example, for being too large) are considered ”worst” and are submitted to neoadjuvant treatment.

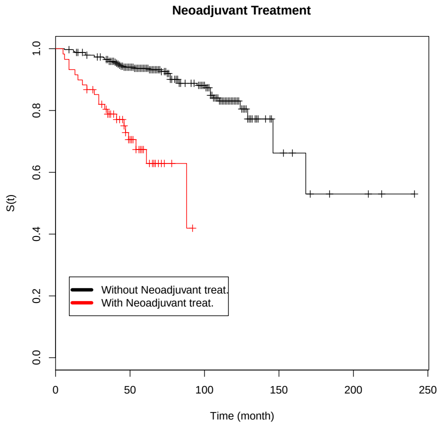


Figure 3.6: Kaplan-Meier curves for categories of variable neoadjuvant treatment.

Also, as can be observed in Figure 3.7 , patients that were not submitted to surgery have a higher risk of dying from breast cancer compared to those submitted to surgery. This result is expected as those cases not submitted to surgery are the ones considered ”worst” and of high risk to undergo a surgery. Those submitted to mastectomy presented a higher risk of dying from breast cancer compared do those submitted to conservative surgery. The few cases that were submitted to sentinel lymph node biopsy presented a lower risk of dying from breast cancer, as can be observed in Figure 3.8. This could be related to a better diagnostic and consequently more efficient treatment.

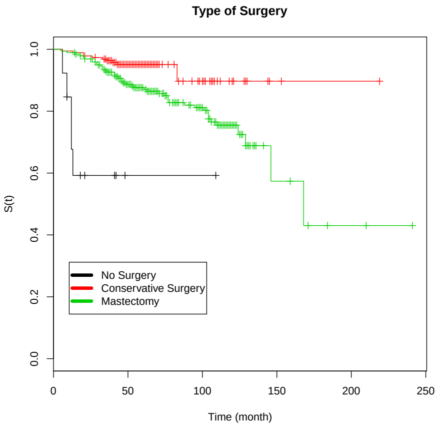


Figure 3.7: Kaplan-Meier curves for categories of variable type of surgery.

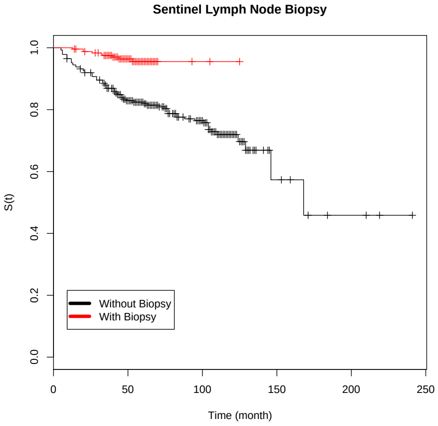


Figure 3.8: Kaplan-Meier curves for categories of variable sentinel lymph node biopsy.

Figures 3.9 and 3.10 clarify that cases that shown images of both venous vascular invasion and lymphatic invasion, respectively, also presented a higher probability of dying from breast cancer.

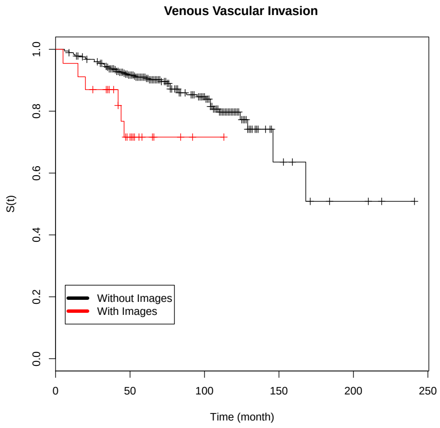


Figure 3.9: Kaplan-Meier curves for categories of variable images of venous vascular invasion.

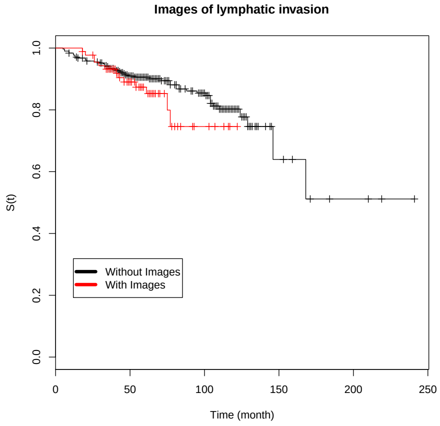


Figure 3.10: Kaplan-Meier curves for categories of variable images of lymphatic invasion.

Figures 3.11 and 3.12 clarify that a positive expression of both estrogen and progesterone receptors, respectively, are associated with a lower probability of dying from breast cancer.

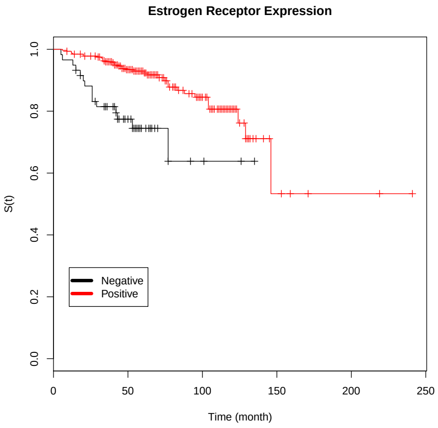


Figure 3.11: Kaplan-Meier curves for categories of variable estrogen receptor expression.

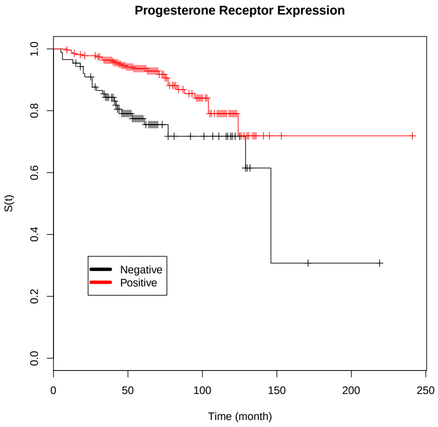


Figure 3.12: Kaplan-Meier curves for categories of variable progesterone receptor expression.

It turns out, as expected, that there is a significant difference on the survival rates, for the categories of the covariate Triple Negative (TN), as shown in Figure 3.13 and confirmed in the log rank test results exposed in Table 3.1, being, as expected the TN breast cancer cases the ones with lower survival rate.

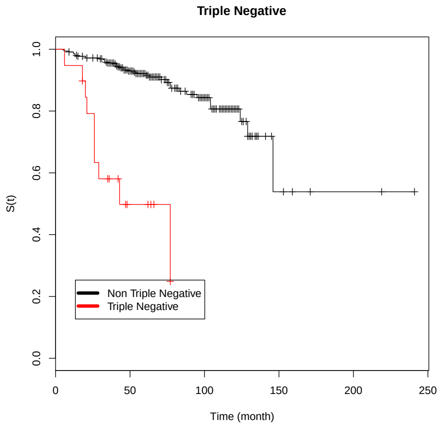


Figure 3.13: Kaplan-Meier curves for categories of variable triple negative.

Our results, as shown in Figure 3.14, point out to a significant difference between the survival rate of the three grade categories (G1, G2 and G3). The survival rate of patients with tumour grade G3 is lower comparing to the other two categories, and the cases with tumour diagnosed in grade G1 have a higher survival rate.

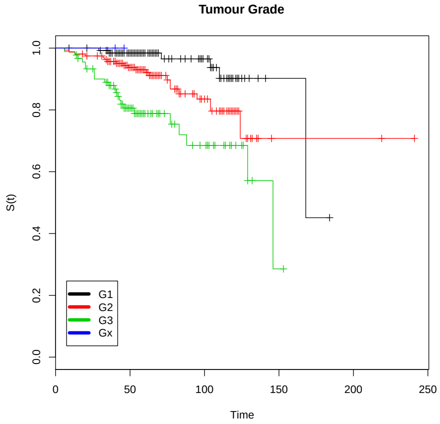


Figure 3.14: Kaplan-Meier curves for categories of variable tumour grade.

Higher categories of primary size tumour are, as expected, related to a higher risk of dying from breast cancer, as shown in Figure 3.15. Although, Tis and T1 type of tumour do not seem to differ in their probability of survival.

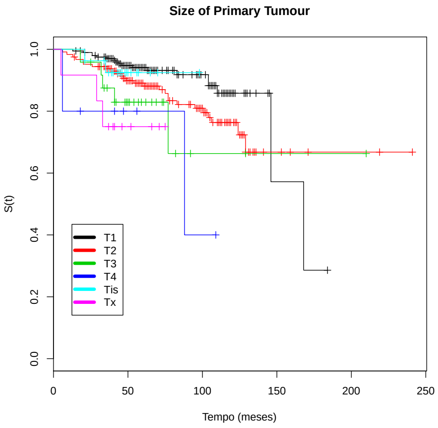


Figure 3.15: Kaplan-Meier curves for categories of variable size of primary tumour.

Figure 3.16 suggests that, for the covariate Regional Lymph Node involvement, patients with tumour with Nx, N0 or N1 degree of regional lymph node involvement have a significant higher survival probability than those with higher degree. We were able to access, through log-rank tests results, that there are no significant differences on the survival curves between the Nx, N0 and N1 degrees (p-value = 0.873) and also between the N2 and N3 categories (p-value = 0.754). Later, grouping Nx, N0 and N1 degrees in one category and N2 and N3 in one other category, we were able to access a statistical difference on survival curves between this two new categories given that the Log-rank test p-value obtained was lower than 0.0001. Figure 3.17 elucidates the referred difference between those two categories.

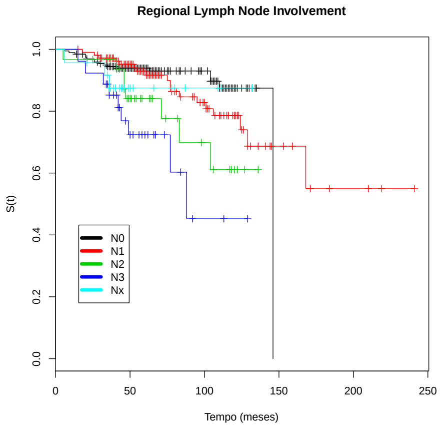


Figure 3.16: Kaplan-Meier curves for categories of variable regional lymph node involvement.

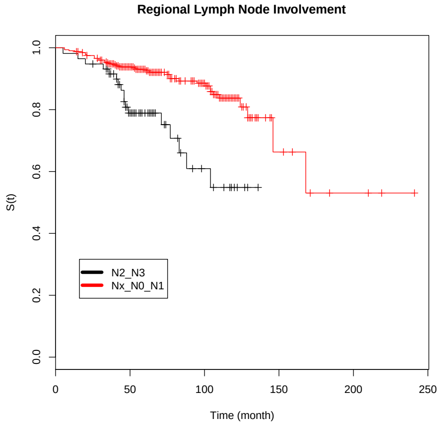


Figure 3.17: Kaplan-Meier curves for categories of variable regional lymph node involvement with different categorization.

Although age at diagnose treated as a continuous variable was not statistically significant in terms of survival, a different result emerge when it was categorized into two categories: patients younger than 44 years and patients older than 44 years (or equal). We verified that the lower category (younger women) has a lower survival probability than the other category, and this difference is statistically significant (p-value = 0.017). Figure 3.19 shows the difference between survival curves of both categories. Note that we firstly categorized age at diagnosis as it is done in studies reported by the North

Region Cancer Registry of Portugal (RORENO), as shown in Figure 3.18.

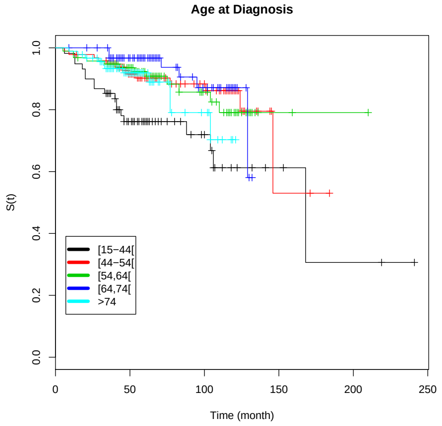


Figure 3.18: Kaplan-Meier curves for categories of variable age at diagnosis.

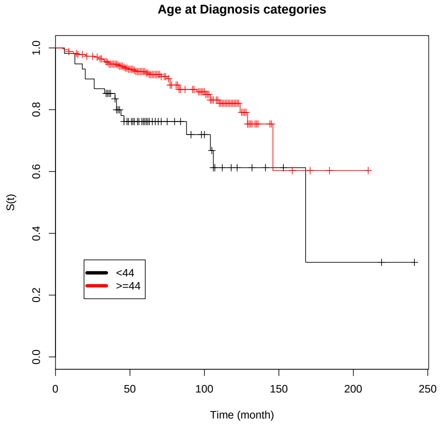


Figure 3.19: Kaplan-Meier curves for categories of variable age at diagnosis with different categorization.

As Figure 3.20 shows, cases submitted to adjuvant hormonotherapy presented a higher probability of survival.

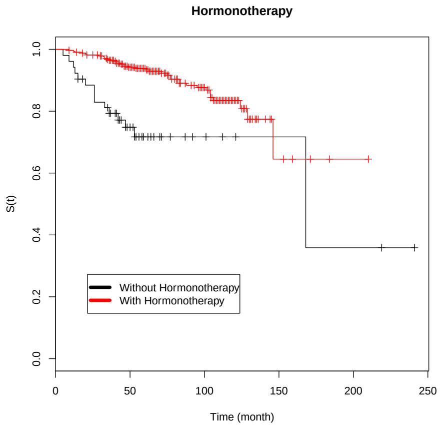


Figure 3.20: Kaplan-Meier curves for categories of variable hormonotherapy.

3.7.2 Analysis with Semi-Parametric Models

For the analysis of the breast cancer data in study, we considered two semi-parametric models: (i) the vastly used Cox proportional hazards model (CPHM) and (ii) the flexible Royston-Parmar survival model (FRPM) that allows, contrariwise to CPHM, the formulation of the baseline cumulative survival function.

The CPHM survival function is given by equation 3.28, described in the last section, however to access the importance of acknowledging left truncation, along with right censoring, of data in the model we compared estimates obtain in two moments: (i) considering only right censoring, hence the partial likelihood function use do estimate the model parameters is simply given by equation 3.32; and (ii) considering both left truncation and right censoring of data, and so we incorporated the correspondent components in equation 3.32 and indicated in equation 3.16.

In the present analysis, using the statistically significant variables obtained in log-rank tests, presented in the previous section, we adjusted several survival models with multiple covariates to estimate the joint effect of several independent variables on survival of patients.

However, we did not consider, in our models, the variables related to treatment, as treatment decisions by the medical team are made upon tumour characteristics (such as, for example, size, staging, grade). As it would result into a strong correlation between treatment type of variables and variables related to tumour characteristics.

The final model (adjusted) was estimated by "step-wise backwards", starting with the saturated model with all the significant variables, and then eliminating one-by-one the variables with lower significance, with a limit of 0.08 for the p-value, for inclusion of the variable. The choice of the number of knots for the FRPM was done by varying between 0 and 4 the number of knots, as suggested by Royston & Parmar (2002), and choosing the model that presented the minimum combination of Akaike Information Criterion (AIC) ((Akaike, 1973)).

After adjusting several survival models with multiple covariates, beginning, as already mentioned, with the saturated model, we ended up with a CPHM and a zero knots FRPM that incorporate five covariates with highly significant effect on the survival of these patients: triple negative (yes versus no); age at diagnosis (< 44 versus ≥ 44); tumour stage (0/I/II versus III/IV), tumour grade (G1 versus G2 versus G3) and regional lypmh node involvement degree (Nx/N0/N1 versus N2/N3).

Note that, as Royston & Parmar (2002) explain, by convention the zero knot model means that no internal and no boundary knots are specified and so, the baseline distribution is none other but the Weibull.

Table 3.2 compares the estimates obtained (and respective 95% confident intervals (CI)) when adjusting to the FRPM with the ones obtained when adjusting to the CPHM. As one can see, the estimates values are identical for both models. Since the main difference between both models relies in the fact that the FRPM allows the formulation of the baseline cumulative survival function, which can be of medical interest, we choose to refer, in the present text, to parameters estimates obtain with this model.

There are a few differences in the results when we ignore or allow left truncation in the modelling process. For example, the covariate lymph node involvement does not have a significant effect in the patient's survival if we ignore left truncation. However, there are differences in the 95% confidence intervals for the estimates, i.e., in the standard errors of the parameters estimates. Specifically, the confidence intervals for the parameters estimates obtained when left truncation is acknowledge are wider, which is the same as saying that the standard errors are bigger. This is expected since the model considering left truncation gives distinctive weighs to the patients diagnosed before 2008.

Analysing the hazard ratio (HR) estimates presented in table 3.2, as expected, the hazard of dying from breast cancer is significantly higher in women with triple negative breast cancer (FRPM HR = 6.86). Also it seems that the hazard ratio for the ones older or equal than 44 years at the time of diagnose is lower (FRPM HR = 0.38). The risk of dying from breast cancer is 7.00 times higher for the cases with tumour stage III or IV compared to cases with tumour on stage 0, I or II. Also, the risk of dying from a tumour with G2 type of grade is 4.72 times higher than G1 type grade of tumour, and the risk increases for the ones with G3 type of grade (FRPM HR = 6.99). Finally, those who present a regional lymph node involvement of N2 or N3 degree have a higher risk (FRPM HR = 2.24) of dying from breast cancer.

Table 3.2: Hazard Ratios for each significant effect covariates for Cox proportional Hazard Model and for Flexible Royston-Parmar Model.

	Only Right Censoring		**Right Censoring and Left truncation**	
	CPHM	**FRPM (0 knots)**	**CPHM**	**FRPM (0 knots)**
Covariates	**HR 95% CI**	**HR 95% CI**	**HR 95% CI**	**HR 95% CI**
Triple Negative (with)	8.07 [3.27; 19.96]	7.10 [2.80; 17.99]	7.82 [2.81; 21.98]	6.86 [2.51; 18.7]
Age at Diagnosis (≥ 44)	0.39 [0.20; 0.73]	0.41 [0.22; 0.78]	0.37 [0.18; 0.71]	0.38 [0.20; 0.73]
Tumour Stage (III or IV)	3.89 [2.09; 7.25]	3.94 [2.13; 7.24]	7.58 [2.95; 19.45]	7.00 [2.80; 17.5]
Tumour Grade (G2)	4.51 [1.04; 19.47]	3.86 [0.96; 15.33]	5.05 [1.16; 21.94]	4.72 [1.09; 20.50]
Tumour Grade (G3)	6.14 [1.37; 27.65]	5.75 [1.38; 24.05]	6.39 [1.39; 29.47]	6.99 [1.53; 31.90]
Lymph Node Involvement (N2 or N3)	—	—	2.39 [0.86; 6.62]	2.24 [0.81; 6.21]

Figure 3.21 presents the plot of the Kaplan Meier estimates, the FRPM and the CPHM for a subject with following characteristics, most frequent observed in the data under study: non triple negative; age at diagnosis $\geq$ 44, tumour stage 0 or I or II, tumour Grade G2 and Regional Lymph nodes involvement degree Nx, N0 or N1. This graph was built in order to graphically asses that both models, the FRPM and the CPHM, fit well the data.

Both models curves remain within the 95% confidence interval of the Kaplan-Meier curve, which confirms a good fit to the data of both models. Also, CPHM and FRPM survival curves are similar which, along with the almost identical parameters estimates presented in table 3.2, confirms that results obtained from both models are identical.

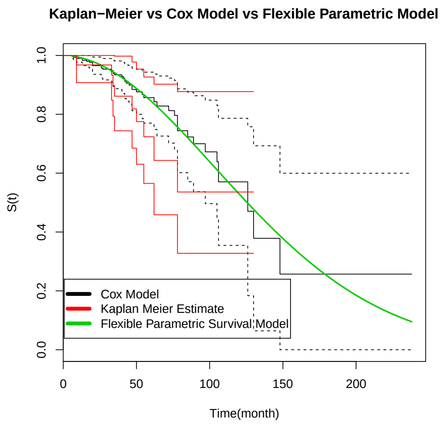


Figure 3.21: Cox Proportional Hazards Curve versus Kaplan Meier curve versus Flexible Royston-Parmar Survival Curve for Combination of covariates: Non Triple Negative; Age at diagnosis ≥ 44, Tumour Stage 0 or I or II and Tumour Grade G2 and Regional Lymph nodes involvement degree Nx, N0 or N1.

As shown in Table 3.3, the proportional hazards assumption for these variables, important for both models, was not violated.

Table 3.3: P-values obtained from testing the proportional hazards assumption

Variable	**p-value**
Triple Negative	0.3885
Age at diagnosis (≥ 44)	0.0621
Tumour Stage (III or IV)	0.5721
Tumour Grade (G2)	0.1727
Tumour Grade (G3)	0.4238
Lymph Node Metastasis (N2 or N3)	0.3312

From the graphical representation of Cox-Snell residuals shown in Figure 3.22, where the black solid line denotes the Kaplan-Meier estimate of the survival function of the residuals (with the dashed lines corresponding the 95% pointwise confidence intervals), and the grey solid line, the survival function of the unit exponential distribution, suggests a good fit of the CPHM to the data.

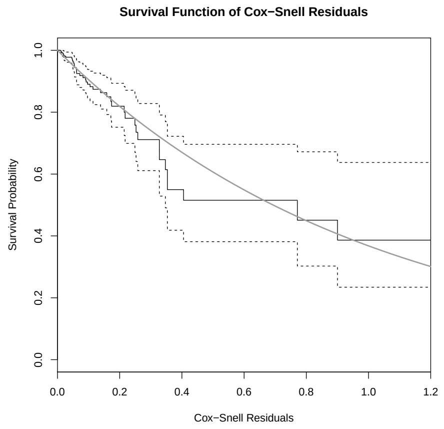


Figure 3.22: Superposition of Kaplan-Meier estimates of the survival function of the residuals and survival function of the unit exponential distribution.

3.8 Discussion

Although many of 50 variables obtained are recognized in literature as potential prognostic factors only five are shown to have a statistically significant effect on survival of the 540 patients of the Unity of Senology of Braga's Hospital when adjusted to other prognostic factors, namely: triple negative (yes versus no); age at diagnosis (< 44 versus ≥ 44); tumour stage (0 or I or II versus III or IV), tumour grade (G1 versus G2 versus G3) and regional lymph node involvement degree (Nx or N0 or N1 versus N2 or N3).

The resulting adjusted FRPM, with 0 knots, and CPHM reveal that the patients with triple negative type of tumour, younger than 44 years at diagnose, with tumour in stage III or IV, grade G3 and regional lymph node metastasis degree N2 or N3, are the ones with higher risk of dying from breast cancer.

It was found also that ignoring a left truncation of the breast cancer data under study leads to an overestimation of survival estimates.

The estimates obtained by adjusting two different semi-parametric models - the CPHM and the FRPM - were identical. Fact sustained by a graphical comparison of both survival curves. The main difference between both models relies in the fact that the latter allows the formulation of the baseline cumulative survival function, which can be of medical interest. Hence, the choice between both CPHM and FRPM depends on the investigator's interest of having the baseline cumulative survival function specified or not.

To compare the results obtained with previous studies we calculated the survival rate at 1, 3 and 5 years for the adjusted FRPM for the combination of variable values most frequent in the dataset in study: non triple negative; age at diagnosis ≥ 44, tumour stage 0 or I or II, tumour Grade G2 and Regional Lymph nodes involvement degree Nx, N0 or N1. Table 3.4 presents these results and the last results reported by the North Region Cancer Registry of Portugal (RORENO) for the districts of North of Portugal and also for Braga's district itself, which points out to an improvement of the survival rate for the Unity of Senology of Braga's Hospital.

Table 3.4: Comparison between study results and relative survival estimates (%) for breast cancer for the North Districts of Portugal and for Braga's district of residence, by RORENO

Years after diagnosis	**Braga's Hospital FRMP**	**RORENO 2000/2001**	**RORENO 2000/2001 Braga**
1 year	98.2%	96%	96 %
3 years	88%	86%	87%
5 years	81%	78%	79%
Number of cases	540	2320	470

Bering in mind that this is a comparison between a cause-specific survival estimated (adjusted to risk factors) and an overall relative estimation. And also, different time intervals of disease diagnose were taken into account.

Nevertheless, it resulted that the survival estimates are better for Braga's Hospital breast cancer data, compared to the north region of Portugal and to Braga's region.

4

Longitudinal Analysis

Longitudinal data is usually characterized as response variables that are measured repeatedly through time for a group of individuals.

It is important to use longitudinal methods when studying medical or biological data, as they can distinguish, for example, changes over time within individuals from differences among individuals in their baseline levels (Diggle *et al.*, 2002). The main characteristic of longitudinal models is that they model both the dependence among the response on the explanatory variables and the autocorrelation among the responses. Ignoring correlation in longitudinal data could lead to incorrect inferences about the regression coefficients, inefficient estimates of the coefficients and, also, sub-optimal protection against biases causes by missing data (Diggle *et al.*, 2002).

In this particular study, the response variables, values of CA15-3 and CEA tumour markers, were analysed making use of longitudinal models as defined in Diggle *et al.* (2002), where different correlation structures were tested.

The work presented in this chapter is an extension of the work published by Borges *et al.* (2015). Firstly it will be presented the methodology applied, introduced by a summary explanation of the main approaches to analyze longitudinal data, describing the longitudinal models used to infer about the factors that affect the progression of both CA15-3 and CEA tumour markers values. Subsequently, the results from the longitudinal analysis of these responses will be demonstrated ending with a discussion of both longitudinal analysis.

The entire analysis was performed using R software, in particular making use of both *nlme* (Pinheiro *et al.*, 2015) and *JoineR* (Philipson *et al.*, 2012) packages.

4.1 Definition of Concepts

As Diggle *et al.* (2002) explains, there are three main approaches when dealing with longitudinal data: i) the marginal model approach; ii) the random effects model and the iii) transition model approach.

The marginal model approach consists in modelling the marginal mean as in a cross-sectional study. This approach has the advantage of separately modelling the mean and covariance. Although, as Fitzmaurice *et al.* (2008) explain, marginal models account for the consequences of the correlation among the repeated measures, they do not provide any explanation for its potential source.

The random effects model assumes the existence of correlation among repeated responses because the regression coefficients vary across individuals. This is the approach adopted in the present study, which will be detailed later.

The transition model approach focuses on the conditional expectation of the response variable given past outcomes. This type of models combine the assumptions about the dependence of the response variable on the covariates among repeated values of the first, into a single equation.

Diggle *et al.* (2002) and Fitzmaurice *et al.* (2008), present an interesting framework for the analysis of longitudinal data, explaining that we can make use of several types of models, depending on the nature of the response variable (for example, continuous Gaussian or non-Gaussian, binary, count) or even the type of data set. Those include: (i) mixed-effects models, (ii) generalized estimating equations (GEE) models, (iii) transitional models, (iv) Bayesian models or (v) nonparametric/semiparametric models.

The linear mixed-effects models, primary proposed by Laird & Ware (1982), are a flexible class of models that incorporate both fixed effects and random effects. In practical terms, fixed effects explain individual-specificity and random effects incorporate the between individual variation and the within-individual correlation in longitudinal data. In this models, the mean response is modelled as a combination of population characteristics (fixed effects) assumed to be shared by all individuals, and subject-specific effects (random effects) that are unique to a particular individual. As Fitzmaurice *et al.* (2008) explain this models have two desirable features: first, there are fewer restrictions on the design matrices for the fixed and random effects; second, the model parameters could be estimated efficiently via likelihood based methods. The main restriction on the application of these type of models rely on assumptions of multivariate normality.

As Pinheiro & Bates (2000) point out the increasing popularity of mixed-effects models is explained by the flexibility they offer in modeling the within-group correlation often present in grouped data, by handling both data with measures equally or unequally separated in time, in a unified framework, and by the availability of reliable and efficient software for fitting them.

The GEE models were firstly introduced by Liang & Zeger (1986), and their primary characteristic is that the mean structure and the correlation structure are modelled separately without distributional assumptions for the data. They consist in a semi-parametric approach and an extension of generalized linear models. These models account for correlation of responses within subject and it can be used when the response variable is not normally distributed.

In transitional models the within-individual correlation is modelled via Markov structures. They are appealing due to the sequential nature of longitudinal data. Is this type of models the conditional distribution of each response is expressed as an explicit function of the past responses and covariates. In general, transition models have been developed for repeated measures that are equally separated in time. Although its use is advantageous when the response variable in non-Gaussian, they are more difficult to apply when there are missing data and non-equidistant intervals between measurement occasions (Fitzmaurice *et al.*, 2008).

In longitudinal studies each patient is denoted by the subscript $i = 1, ..., n$. Repeated measurements for each patient i, at corresponding time t_{ij}, are denoted by Y_{ij}, where $j = 1, ..., m_i$. Let $N = \sum_{i=1}^{n} m_i$, be the total number of measurements in the data set. In some studies m_i can be equal to all subjects, i.e., for all patients measurements were taken at the same time, which means that $m_i = m$. In that particular case we say that it is a balanced study. If m_i's are different for all subjects it is called an unbalanced study.

The construction of a longitudinal model is based on the ordinary least square (OLS) model to explain the mean progression of a given response variable through time, given by:

$$Y_{ij} = \mu_{ij} + \varepsilon_{ij}, \tag{4.1}$$

where $E[Y_{ij}] = \mu_{ij}$ and ε_{ij} are N independent and identically distributed (i.i.d.) realizations of $N(0, \xi^2)$.

However, since the OLS model assumes independence between any two measurements, from the same or different subject, it is important to consider different models, in the context of longitudinal analysis, that take into account the correlation that usually exists in the measurements of the same subject. What longitudinal models purpose is to decompose the variability in ε_{ij}, in variability between subjects and variability within subjects, through time.

In fact, in many medical studies it is important to consider not only random effects (variability between subjects) but also a possible variability within subjects as it may have important medical implications. Liang & Zeger (1986) alert that treating the correlation as a nuisance may be less appropriate when the time course of the outcome for each subject is of primary interest or when the correlation itself has scientic relevance.

4.2 Longitudinal Model

The general longitudinal model described as in Diggle *et al.* (2002) is:

$$Y_{ij} = \mu_{ij} + \mathbf{d}_{ij}^{'}\mathbf{U}_i + W_i(t_{ij}) + Z_{ij}. \tag{4.2}$$

Where $\mathbf{U}_i$ are n i.i.d. realizations of $MVN(0, \Sigma)$, representing the random effects at individual level, and $\mathbf{d}_{ij}^{'}$ is a vector of covariates for the random efects. $W_i(t_{ij})$ is a continuous time Gaussian Process with $E[W_i(t_{ij})] = 0$ and $Var[W_i(t_{ij})] = \sigma^2$, representing the variability within subjects, where the correlation between two measurements of an individual is described by: $corr(W_i(t_{ij}), W_i(t_{ik})) = \rho(t_{ij}, t_{ik})$.
Finally, Z_{ij} are N i.i.d. realizations of $N(0, \tau^2)$, representing the measurement error (variability non specied).

Since $W_i(t_{ij})$ is assumed a stationary process we have $\rho(t_{ij}, t_{ik}) = \rho(|t_{ij} - t_{ik}|)$.

We can have different definitions for the function $\rho(|t_{ij} - t_{ik}|)$. That is, if we consider the correlation among $W_i(t_{ij})$, let say between $W_i(t)$ and $W_i(t-u)$, determined by the autocorrelation function $\rho(u)$, we will have for a longitudinal model that accounts for an exponential correlation structure within individuals:

$$\rho(u) = \exp\left(-\frac{1}{\phi}|u|\right), \tag{4.3}$$

and for a longitudinal model that accounts for a Gaussian correlation structure within individuals:

$$\rho(u) = \exp\left(-\frac{1}{\phi}u^2\right), \tag{4.4}$$

where ϕ is the range parameter that specifies the rate at which the correlation stables.

To model the fixed term of the longitudinal model, μ_{ij}, we can consider a model with a changing point δ on the effect of time on the response variable. In practice, the changing point is the moment where there is an alteration on the slope of the linear response variable's progression, on average. Considering δ the changing point, we have $E[Y_{ij}] = \mu_{ij}$ with:

$$\mu_{ij} = \left\{ \begin{array}{ll} X_{ij}\beta + \alpha_1 t_{ij}, & \text{if } t_{ij} < \delta \\ X_{ij}\beta + \alpha_2(t_{ij} - \delta) & \text{if } t_{ij} \geq \delta \end{array} \right. , \tag{4.5}$$

where X_{ij} represents the vector of covariates, β the vector of unknown regression coefficients, α_1 and α_2 the coefficients representing the slope before and after the changing point, respectively.

For simplicity, let's consider the complete set of N measurements, y, as a realization of a multivariate Gaussian random vector Y with $Y \sim MVN(X\beta, \eta^2 V)$. Where $\mathbf{X}$ is an $N \times p$ matrix of all values of the p explanatory variables. And $\eta^2 V$ as a block-diagonal matrix with non-zero $n \times n$ blocks $\eta^2 V_0$, each representing the variance matrix for the vector of measurements on a single subjects.

For parameter estimation we adopted the maximum likelihood method, whose associated log likelihood function for observed data y, under Gaussian assumption, is given by:

$$L(\beta, \eta^2, V_0) = -\frac{1}{2}\left\{nm\log(\eta^2) + m\log(|V_0|) + \frac{1}{\eta^2}(y - X\beta)^{'}V^{-1}(y - X\beta)\right\}. \quad (4.6)$$

The maximum likelihood estimator for β, given V_0 is the weighted least-squared estimator (Diggle *et al.*, 2002), given by:

$$\hat{\beta}(V_0) = \left(X^{'}V^{-1}X\right)^{-1} X^{'}V^{-1}y. \quad (4.7)$$

Substituting 4.7 in 4.6 we will have:

$$L(\hat{\beta}(V_0), \eta^2, V_0) = -\frac{1}{2}\left\{nm\log(\eta^2) + m\log(|V_0|) + \frac{1}{\eta^2}RSS(V_0)\right\}, \quad (4.8)$$

where,

$$RSS(V_0) = (y - X\hat{\beta}(V_0))^{'}V^{-1}(y - X\hat{\beta}(V_0)). \quad (4.9)$$

The maximum likelihood estimator for η^2, for fixed V_0, is obtained differentiating 4.8 with respect to η^2, resulting:

$$\hat{\eta}^2(V_0) = \frac{RSS(V_0)}{nm}. \quad (4.10)$$

The reduced log likelihood for V_0, apart from a constant term, is obtained substituting 4.7 and 4.10 into 4.6, as:

$$L_r(V_0) = -\frac{1}{2}m\left\{n\log RSS(V_0) + log(|V_0|)\right\}. \quad (4.11)$$

Maximizing 4.11 we obtain the maximum likelihood estimator of V_0. Finally, substituting $\hat{V_0}$ in 4.7 and 4.10 we obtain the maximum likelihood estimators $\hat{\beta}$ and $\hat{\eta}^2$, respectively.

4.3 Missing Data

One important aspect, when dealing with real data, is acknowledging the missing data mechanism behind the lack of measurements of the response variable that can, and usually, arise.

When dealing with unbalanced data, such as the one analysed in this study, is difficult to access if the lack of a measurement can actually be defined as missing data. Nevertheless, we should take into account the different classifications of missing data mechanism in order to incorporate it in the longitudinal model, if necessary.

Adopting Diggle *et al.* (2002) notation, let's consider $Y^* = (Y^o, Y^m)$ where Y^* is the complete set of measurements with no missing values, Y^o the measurements actually obtained (observed) and Y^m the measurements that would have been available had they not been missing. Also, let R denote a set of indicator random variables, denoting which elements of Y^* fall into Y^o , and which ones fall into Y^m. A probability model for the missing value mechanism defines the probability distribution of R conditional on Y^*.

Little & Rubin (2000) classify the missing data mechanism into three different categories:

- Completely random if R is independent of both Y^o and Y^m
- Random if R is independent of Y^m
- Informative if R is dependent on Y^m

Since we are dealing with unbalance type of data and, also, the missing data detected in our data base does not seem to be related to the progression of the disease, we consider the first type of mechanism - the completely random mechanism. Which can be ignorable when adopting the likelihood function as the basis for inference (Little & Rubin, 2000).

For a detailed description of the random and informative mechanisms and the incorporation of these in the longitudinal model we suggest the work of Little & Rubin (2000) and Diggle *et al.* (2002).

4.4 Variogram for Longitudinal Data

To model the correlation structure for each model we analysed the empirical variogram of OLS residuals from the saturated model for the mean response (Diggle *et al.*, 2002).

The variogram (Diggle *et al.*, 2002) of a stochastic process $Y(t)$ is given by:

$$V(u) = \frac{1}{2}\text{Var}\left\{Y(t) - Y(t-u)\right\}, \ u \geq 0. \tag{4.12}$$

For a stationary process, the autocorrelation function, $\rho(u)$, and the variance of $Y(t)$, σ^2, are related by:

$$\gamma(u) = \sigma^2\{1 - \rho(u)\}. \tag{4.13}$$

The estimation of the empirical variogram is based on the calculation of the observed half-squared-differences between pair of residuals, $\nu_{ij} = \frac{1}{2}(r_{ij} - r_{ik})^2$, and the corresponding time-differences, $u_{ijk} = t_{ij} - t_{ik}$, where $r_{ij} = Y_{ij} - \mu_{ij}$, and $j < k = 1, ..., m_i$.

The autocorrelation function at any lag u is estimated from the sample variogram by:

$$\hat{\rho}(u) = 1 - \frac{\hat{\gamma}(u)}{\hat{\sigma}^2}, \tag{4.14}$$

where $\hat{\gamma}(u)$ is the average of all the ν_{ij} corresponding to that particular value of u, and $\hat{\sigma}^2$ is the estimated process variance.

4.5 Model Diagnostic

Diagnostic checking of the longitudinal fitted model can be made by superimposing the fitted mean response profiles on a time-plot of the average observed response within each combination of covariates categories and time, as suggested by Diggle *et al.* (2002). However, this task can be rather difficult if we are dealing with a model that incorporates a great number of covariates.

One other form of model fitting diagnose is to superimpose the fitted variogram on a plot of the empirical variogram. Which will allow the validation of the correlation structure by graphical comparison.

Another procedure to additionally access which correlation structure best describe data variability, within a longitudinal model, with a predetermined fixed part, is by comparison of the maximized log likelihood values.

Plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals can, also, be used for graphical validation of the assumptions of the constant variance and normal distribution of Z_{ij}, respectively.

4.6 Breast Cancer Longitudinal Analysis

The same covariates used in the survival model, previous adjusted (Chapter 3), were tested in the longitudinal model fitted. The reference time used was time, in months, since diagnose of breast cancer until the date of the blood test. According to the usual medical procedures, physicians stay alert to a possible recurrence of breast cancer for patients that present values of tumour markers above a certain reference value. We have considered the reference value of 37 U/ml for CA15-3 tumour marker, and 5.0 ng/mL for the CEA tumour marker, as this were the values referred in the blood tests analysed.

The normality assumption of both response variables CA15-3 and CEA was first tested preforming a Shapiro-Wilk normality test (Shapiro & Wilk, 1965), both with a returned p-value < 0.0001 and therefore was rejected. It was used a log-transformation of both the tumour markers CA15-3 and CEA values. It is, in fact, a usual transformation in many biological cancer markers studies, as, for example, in studies of Goldhirsch *et al.* (1988), Crump *et al.* (2000), J. Louhimo *et al.* (2002) or even Liu *et al.* (2014).

Keene (1995) gives examples that evidences the log transformed analysis as more supportive of a treatment effect than the untransformed analysis. Also, he mentions that variables such as biochemical measurements typically show a skewed distribution, which can often be made symmetric using a log transformation. Noting, as well, that has been argued that the theoretical justification for using this transformation for most scientific observations is probably better than that for using no transformation at all.

A longitudinal model was also fitted only to the subset of patients that died, using the reference time, in month, from date of death until blood tests, in order to search for more significant results to explain the progression of both tumour marker values through time related to the death from breast cancer.

For both analysis, we began with an exploratory analysis and point estimation by modelling a OLS model with the variables that had shown significant effect on patients' survival, given by equation 4.1.

We then conducted a backward elimination to delete variables not significant, with a limit of 0.08 for the p-value, for inclusion of the variable, until the mean structure was well defined with only significant covariates.

Observing graphical representation of the empirical variogram, patterns suggested the existence of variability between subjects, and a possible variability within subjects. Hence, maintaining the same mean structure we compared two nested models with different covariance structures, such as: (i) random effects, exponential serial correlation and measurement error (REE), and (ii) random effects, Gaussian serial correlation and measurement error (REG).

The entire analysis was performed using R software, in particular making use of both *nlme* (Pinheiro *et al.*, 2015) and *JoineR* (Philipson *et al.*, 2012) packages.

4.6.1 Results of Tumour Marker CA15-3 Longitudinal Analysis

As reported in Chapter 1, there is only available information on tumour marker CA15-3 values for 534 of the 540 patients eligible for analysis. A total of 5166 measurements of tumour marker CA15-3 represent all the available measurements of those patients from the time of diagnosis of breast cancer to the end of study. The total number of deaths from breast cancer is 55.

Time Since Diagnose

A spaghetti plot, in longitudinal study, represents the individual progressions in time of the response variable.

Figure 4.1 presents the progression in time, since diagnostic, of the CA15-3 values for each patient with grey lines, against the reference value, and the non-parametric smooth spline line with dash line, indicating the average trend of progression.

The smooth spline suggests that, on average, the marker progression increases at a slow rate, staying below the reference value, until about 10 years (120 months) after breast cancer diagnose. After that point the increase seems to be at a higher rate. This fact may point out the existence of a changing point in its progression in time, after the referred cutting point. Thus, a linear parametric modeling approach, testing for changing points, given by equation 4.5 appears to be reasonable.

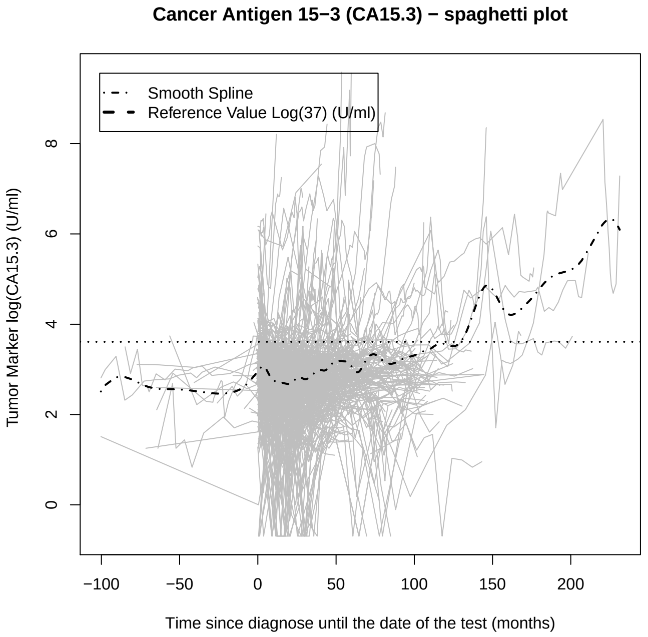


Figure 4.1: Spaghetti plot for tumour marker CA15-3 values.

However, after fitting several saturated parametric models considering several changing points' values, its existence was not signicant in the mean time trend of the tumour progression. Which could be due to the high increase at the end with only the contribution of 3 (or 4) of the patients.

With the motivation to understand which of the explanatory variables with significant effect on patients survival would affect the progression of the tumour marker CA15-3, and also how it would affect, we fitted single longitudinal models with single explanatory variable, using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables, that will be presented next.

In table 4.1 we present the results of the single models, with single variables that were significant in the survival analysis (Chapter 3), in the longitudinal analysis (when considering various variables in the longitudinal model) and any other variables (not present in survival and longitudinal models) that turned out significant by itself. Note that the number of observation in each model is different due to the different number of missing data in each variable.

Results show that variable triple negative, although with significant effect on the survival of the patients, does not have a significant effect on its own on the mean progression of the tumour marker CA15-3 (p-value = 0.7471).

Age at diagnosis treated whether as categorized (Survival) or as continuous (Longitudinal), by itself, does not present a significant effect on the mean progression of the tumour marker. As expected, since they are reported as possible prognostic factors (Chapter 1), an increase of tumour values progression is related to higher categories of variables such as tumour stage, tumour degree, regional lymph node involvement and primary tumour size, as informed in Table 4.1.

The presence of venous vascular invasion, on its own, is responsible for an increase of the starting point of the mean progression of the log values of the tumour marker of 0.7413, comparing to the cases with no venous vascular invasion. The same occurs to the cases that presented lymphatic invasion, although the increment is only of 0.1751.

Table 4.1: Estimated Parameters Values for Single CA15-3 Longitudinal Models

Variable	β	SE	p-value	DF	ν^2	σ^2	ϕ	τ^2	Loglike
Time	0.003	0.001	< 0.001	4498	0.428	0.033	0.934	0.401	-4251.094
Triple negative (Yes)	-0.058	0.179	0.747	471	0.444	0.033	0.915	0.414	-3585.843
Age at diag. (≥ 44)	-0.164	0.092	0.074	548	0.428	0.033	0.969	0.410	-4260.855
Age at diag. (continuous)	0.001	0.002	0.675	548	0.432	0.033	0.967	0.410	-4262,368
Tumour stage (III or IV)	0.352	0.086	< 0.001	525	0.398	0.034	1.0195	0.4108	-4079.871
Tumour degree (G2)	0.186	0.079	0.019	500	0.431	0.025	0.983	0.415	-3773.657
(G3)	0.253	0.094	0.007						
Regional Lymph nodes (N1)	0.206	0.074	0.005	527	0.3943	0.0338	1.0076	0.4081	-4120.342
(N2)	0.351	0.120	0.004						
(N3)	0.530	0.140	< 0.001						
(Nx)	0.065	0.159	0. 681						
Venous vascular invasion (Yes)	0.741	0.157	< 0.001	548	0.413	0.033	0.964	0.409	-4251,446
ki.67(low)	-0.209	0.093	0.026	309	0.488	0.029	0.876	0.424	-2083,901
Recurrence (with)	0.936	0,082	< 0.001	548	0.313	0.034	0.999	0,417	-4204,607
Lymphatic invasion (Yes)	0.175	0.086	0.042	548	0.427	0.033	0.970	0.410	-4260,388
Primary tumour size (T2)	0.157	0.070	0.025	534	0.403	0.034	0.991	0.034	-4184,692
(T3)	0.431	0.147	0.004						
(T4)	0.730	0.319	0.022						
(Tis)	-0.334	0.168	0.047						
(Tx)	0.324	0.226	0.151						
Neoadjuvant treatment (with)	0.462	0.101	< 0.001	548	0.411	0.033	0.977	0.412	-4252.045
lymph node Biopsy (with)	-0.310	0.065	< 0.001	548	0.409	0.033	0.971	0.410	-4251.318

Recurrence of breast cancer is, as expected, related to a low survival probability (Chapter 3), and that result is concordant with the significant effect of this variable on the mean progression of this tumour marker. Cases that presented recurrence of breast cancer had an increment of 0.936 on the starting point of the mean progression of the log values of the tumour marker.

A low ki.67 expression appears to be related to a mean progression starting -0.2085 times lower than the mean progression of cases that presented a higher expression of that biomarker.

An interesting result is that the cases submitted to neoadjuvant treatment are related to an increase of the starting point of the mean progression of the tumour compared to those not submitted to it. That increment is of 0.4623 on the log values of the tumour marker. That could be due to the fact that cases of large primary tumour are usually submitted to that kind of treatment in order to decrease its size for posterior surgery.

Cases where it was performed a biopsy of sentinel lymph node present a lower starting point (-0.3095) of the mean progression, compared to those were it was not preformed any surgery of that type. As already referred in Chapter 1, this type of surgery is adopted recently and it applies to tumour in earlier stages, so that patients are spared of being submitted to axillary dissection (a rather aggressive surgery), hence, this result may be camouflage by the tumour characteristics rather than the surgery itself.

That is one of the reasons why all variables related to type of treatment (neoadjuvant, surgery or adjuvant) were excluded from the posterior analysis that will be presented. One other reason is that those variables are due to medical decision taken upon the tumour characteristics and thus correlation with variables related to tumour characteristics may occur.

A longitudinal analysis starting with all explanatory variables was performed. Different correlation structures for the longitudinal model given by equation 4.2 were tested, based on the analysis of the empirical variogram, presented in Figure 4.2. Based on the structure of the empirical variogram, we present the results, comparing two models, both with a random intercept an individual level, and with a serial correlation structure: one with Exponential structure (REE), given by 4.3, and the second with a Gaussian structure (REG), given by 4.4.

Table 4.2 presents the estimated parameters (Est) of the fitted longitudinal models, and compares the estimates to those obtained by tting the simple OLS model, and the respective log Likelihood values. It is noticeable that, although the estimates values are similar for the two models, there are slightly variations on the significance of these (p-value).

Table 4.2: Estimated Parameters Values for General Linear Model and Longitudinal Models – CA15-3 Tumour Marker

	REE Model		**REG Model**		**OLS Model**	
	Est	**p-value**	**Est**	**p-value**	**Est**	**p-value**
Intercept	2.226	< 0.001	2.215	< 0.001	2.287	< 0.001
Time (month)	0.003	0.006	0.004	< 0.001	0.004	< 0.001
Age at diagnosis	0.007	0.018	0.007	0.018	0.007	< 0.001
Tumour Stage (III or IV)	0.277	0.035	0.243	0.063	0.408	< 0.001
Venous Vascular Invasion (Yes)	0.610	0.003	0.623	0.002	0.812	< 0.001
Ki.67 (low)	-0.168	0.067	-0.163	0.076	-0.226	< 0.001
ν^2	0.426		0.5011		.	
σ^2	0.031		0.066		.	
ϕ	10.552		6.119		.	
τ^2	0.403		0.284		.	
ξ^2	.		.		1.084	
Log Likelihood	-1953.611		-2472.825		-3447.758	

Since REE longitudinal model presented the higher Log likelihood value we selected this as the model to describe the progression of the tumour marker CA15-3 values in time.

The fixed part of the longitudinal model, which describes the mean progression of the marker, is composed by the following significant covariates on the intercept component of the model: time (in month), age at diagnosis, tumour stage (0/I/II versus III/IV), venous vascular invasion (Yes vs No), and Ki.67 expression (low vs high). The intercept component of the model, in this particular case, means that a patient with a tumour stage of 0, I or II, with no venous vascular invasion images and a high Ki.67 expression, at an earlier age of diagnosis will start the progression of the tumour marker with a value of 2.2258, on a logarithmic scale.

We can infer that age at diagnosis affects the log value of the marker at a rate of 0.0074 per year of age at diagnosis. Also, a patient with a tumour on stage III or IV implicates an increasing of the log value of the tumour marker by an increment of 0.2769, comparing to those with a tumour on stage 0, I or II. Also, a tumour that presents images of venous vascular invasion has an increment in the starting point of the marker value of 0.6097. On the other hand, a low expression of the biomarker Ki.67 decreases the starting point of the log value of the tumour marker by 0.1681. Also, there is an increase of 0.0033 per month of the tumour marker value.

The correlation structure that best represents the variability of the data is, in fact, the one that incorporates random effects at individual level with $\hat{\nu}^2 \approx 0.426$, an exponential correlation structure to describe the variability within patients with $\hat{\rho}(u) \approx \exp(-\frac{1}{10.552}.|u|)$ and $\hat{\sigma}^2 \approx 0.031$, and a measurement error with variance $\hat{\tau}^2 \approx 0.403$.

Figure 4.2 represents the empirical and both fitted variograms for the different models. The superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structures with the empirical variogram supports the choice of the REE model, since the REE curve (green line) is the one that best approaches the empirical curve.

Figures 4.3 and 4.4 present, the plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals, respectively. Observing this plots, the assumptions of the longitudinal model, concerning to the homogeneity of variances of the measurement errors and the normality of these do not seem likely to be rejected.

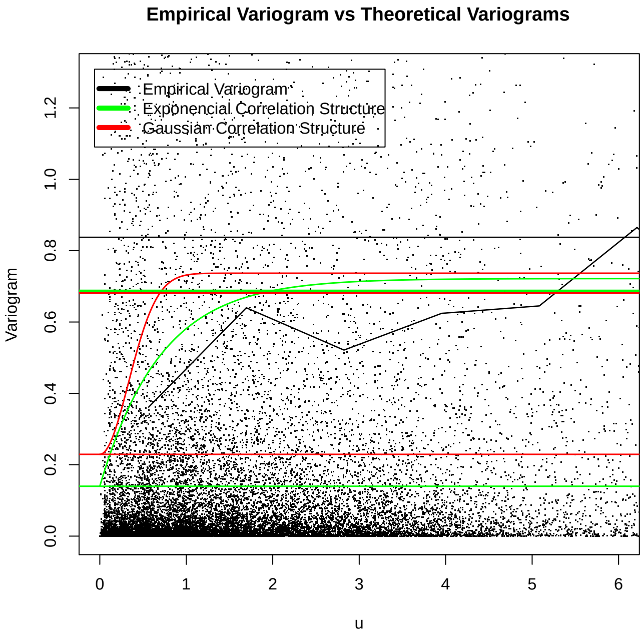


Figure 4.2: Superposition of empirial variogram and theoretical variograms - CA15-3 tumour marker.

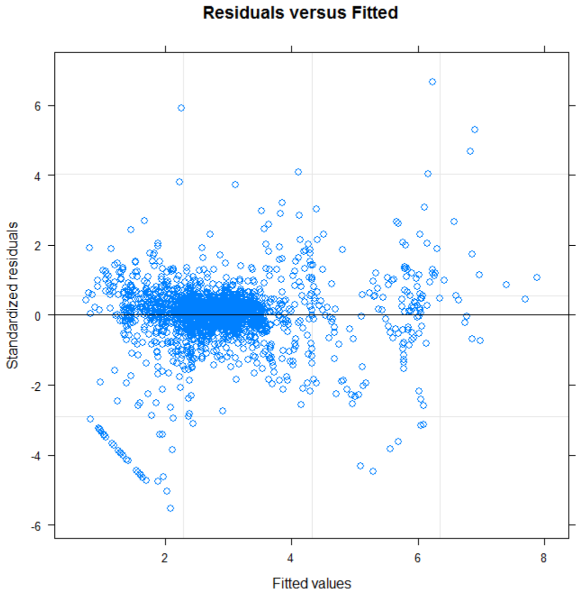


Figure 4.3: Residuals versus the Fitted Responses of CA15-3 values.

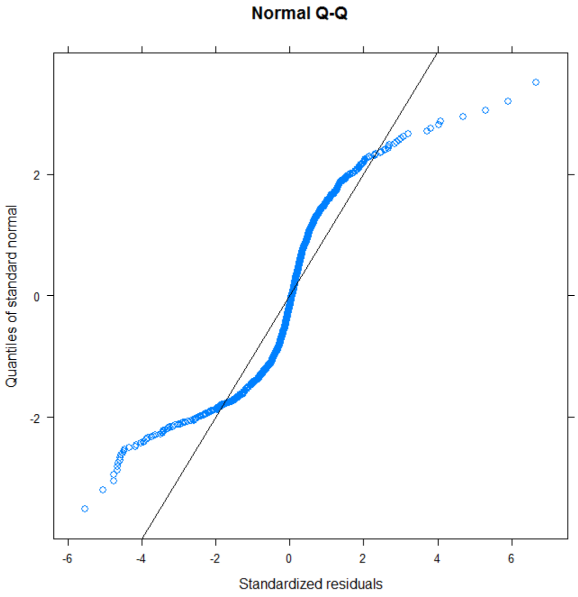


Figure 4.4: Q-Q plot of the CA15-3 longitudinal model.

Time Since Death

A similar longitudinal analysis was performed, but now using a different time scale, time since death until date of blood test, in order to search for more significant results to infer about the relation between the progression of the tumour marker values and the death from breast cancer. Note that, with this time scale the analysis is only performed with the subset of patients that died, and also, as we are analysing the marker values from date of blood tests until death, we are dealing with duration at a negative scale.

Figure 4.5 presents the progression in time, since date of death, of the CA15-3 values for each patient that died from breast cancer, with grey lines, against the reference value, and the non-parametric smooth spline line with dash line, indicating the average trend of progression.

The smooth spline of the spaghetti plot suggests that, on average, the marker progression increases at a slow rate until about 20 months before death from breast cancer. In fact, when fitting the saturated general linear model for the subset of patients who died from breast cancer, given by equation 4.5, we detected a changing point at

2 years before death (24 months). Which means that, for the patients that died from breast cancer it was detected an abrupt rise of CA15-3 values, above the reference value, 2 years before their death.

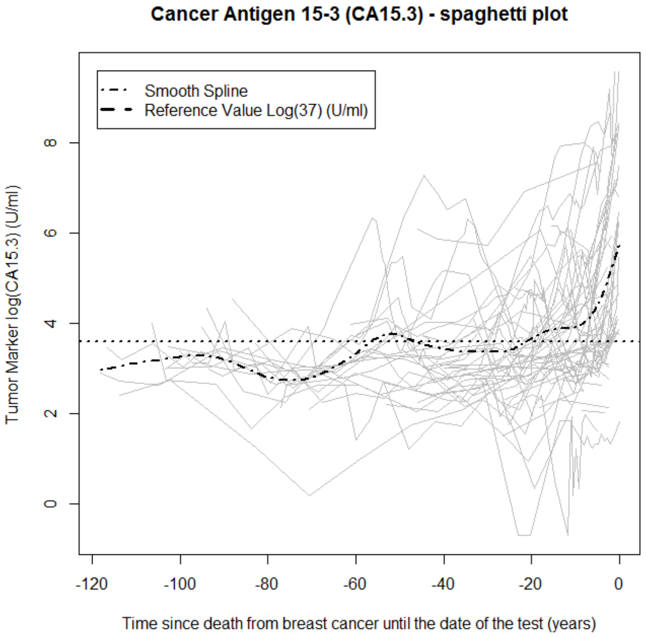


Figure 4.5: Spaghetti plot for tumour marker CA15-3 values - for the subset of patients that died from breast cancer.

Table 4.3 summarizes and compares the estimated parameters for the two longitudinal models fitted REE and REG - with those of the general linear model (OLS Model). Note that the variable ”Time further than 2 years before death” is, in fact, time before the changing point (2 years before death from breast cancer), in the negative scale considered and, equivalently, ”Time earlier than 2 years before death” is time after the changing point, in the negative scale considered.

Table 4.3: Estimated Parameters Values for General Linear Model and Longitudinal Models For Patients Who Died From Breast Cancer – CA15-3 Tumour Marker

	REE Model		**REG Model**		**OLS Model**	
	Est	**p-value**	**Est**	**p-value**	**Est**	**p-value**
Intercept	3.879	< 0.001	3.756	< 0.001	3.501	< 0.001
Time further than 2 years before death	-0.001	0.718	0.002	0.504	0.006	0.0853
Time earlier than 2 years before death)	0.015	0.006	0.018	< 0.001	0.015	< 0.001
Bilateral (Yes)	-0.684	0.072	-0.583	0.0867	-0.714	< 0.001
Venous Vascular Invasion (Yes)	1.367	0.005	1.312	0.002	1.474	< 0.001
Estrogen Receptor (Positive)	0.654	0.052	0.627	0.031	0.923	< 0.001
ν^2	0.00000003		0.435		.	
σ^2	0.040		0.080		.	
ϕ	23.247		6.588		.	
τ^2	1.826		1.149		.	
ξ^2	.		.		1.828	
Log Likelihood	-713.339		-719.435		-1232.524	

Similar to results obtained considering time since diagnose until date of blood test, we selected the REE longitudinal model to describe the progression of the tumour marker CA15-3 values in time, since it was the one with higher Log likelihood value.

Results show that time after the changing point, that is before two years prior to death has a significant effect (p-value = 0.006) on the mean progression of the tumour marker values, related to an increase of 0.0153 per month of the CA15-3 values. However, before changing point the effect of time is not significant (p-value=0.7175) in the mean progression of the tumour marker values.

The mean progression of the marker, when considering the reference time as time since death of breast cancer until date of blood test, is composed by the following significant covariates on the intercept component of the model: bilateral (yes versus no), images of venous vascular invasion (yes versus no) and estrogen receptor expression(positive versus negative). The intercept component of the model, in this particular case, means that a patient with no bilateral tumour, with no venous vascular invasion images and a negative estrogen receptor expression will start the progression of the tumour marker with a value of 3.879, on a logarithmic scale.

As expected, the presence of venous vascular invasion has an increasing effect on the average CA15-3 linear progression in time, as it is related to a worst prognostic case in the previous survival analysis (Chapter 3).

Contradictory results are the decreasing effect of a bilateral type of tumour and the increasing effect of a positive estrogen. The mentioned covariates have a statistical significant effect on the intercept component of the model.

Bilateral cancer cases have a decrease of 0.6839 on the intercept component, and a case with venous vascular invasion a decrease of 1.3667 compared to those with no venous vascular invasion. A positive estrogen receptor expression has an increasing effect on the intercept component by 0.6535, compared to those with a negative expression.

For this subset, the correlation structure that best represent the variability of the data is the structure that incorporates random effects at individual level with $\hat{\nu}^2 \approx 3 \times 10^{-8}$, an exponential correlation structure to describe the variability within patients with $\hat{\rho}(u) \approx \exp(-\frac{1}{23.247}.|u|)$ and $\hat{\sigma}^2 \approx 0.040$, and a measurement error with variance $\hat{\tau}^2 \approx 1.826$. The superposition of the theoretical variogram of both exponential and exponential correlation structures with the empirical variogram, presented in Figure 4.6, validates the choice of an exponential correlation structure, since it is the one that best approximates the empirical variogram curve.

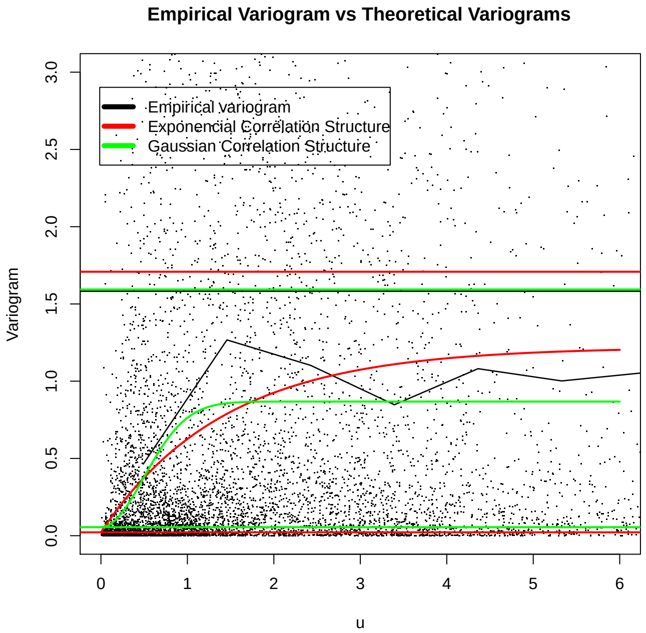


Figure 4.6: Superposition of empirial variogram and theoretical variograms, for patients that died from breast cancer - CA15-3 tumour marker.

Figures 4.7 and 4.8 present, the plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals, respectively. This plots suggest that the assumptions of the longitudinal model, concerning to the homogeneity of variances of the measurement errors and the normality of these do not seem likely to be rejected.

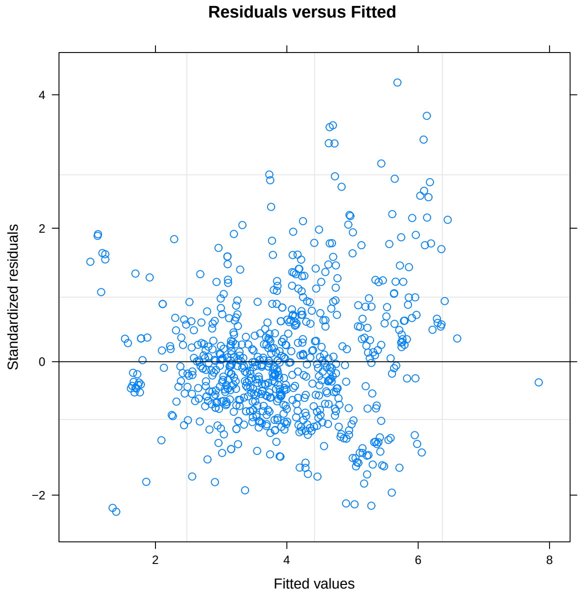


Figure 4.7: Residuals versus the Fitted Responses of CA15-3 values - for the subset of patients that died from breast cancer.

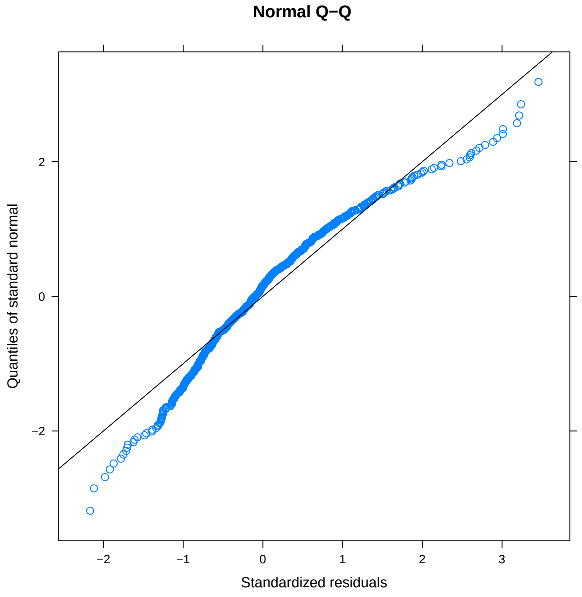


Figure 4.8: Q-Q plot of the CA15-3 longitudinal model - for the subset of patients that died from breast cancer.

4.6.2 Results of Tumour Marker CEA Longitudinal Analysis

As reported in Chapter 1, there is only available information on tumour marker CEA values for 532 patients. This translates into a total number of 551 cases analysed, since 19 cases presented bilateral breast cancer. In total there are 4166 measurements of tumour marker CEA. Also, the total number of deaths from breast cancer is 54.

Time Since Diagnose

The spaghetti plot in Figure 4.9, presents the progression of the CEA values, since tumour diagnose, for each patient with grey line, against the reference, and the non-parametric smooth spline line, indicating the average trend of progression. The smooth spline suggests that, on average, the marker progression stays below the reference value with a non accentuated slope in its increase. However, it is possible to see that there are individuals with values above the reference value of log (5.0)ng/mL. Nevertheless a linear modeling approach appears to be reasonable. Also, it does not point out to the existence of a changing point in its progression in time.

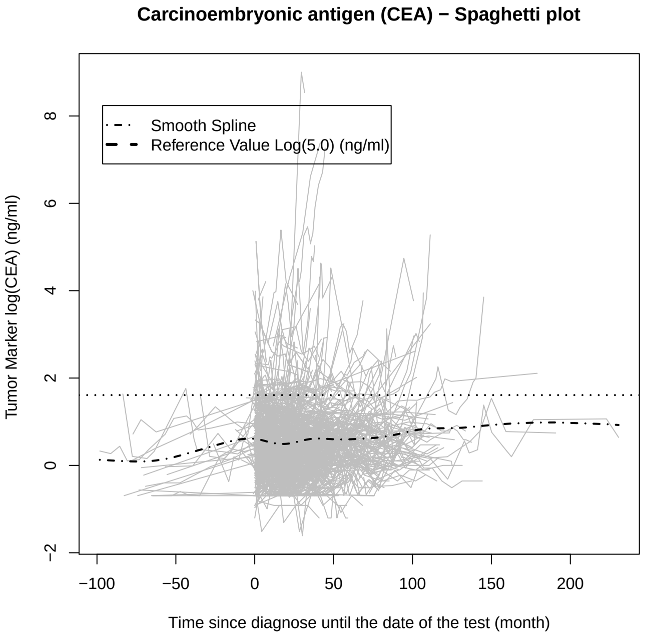


Figure 4.9: Spaghetti plot for tumour marker CEA values.

In fact, after fitting several saturated parametric models considering several changing points' values, given by equation 4.5, its existence was not significant in the mean time trend of the tumour progression.

With the motivation to understand which of the explanatory variables with significant effect on patients survival would affect the progression of the tumour marker CEA, and also how it would affect, we fitted single longitudinal models with single explanatory variable, using the same correlation structure as the one obtained in the selected longitudinal model with multiple explanatory variables, that will be presented following.

Table 4.4 presents the results of the single longitudinal models considering each one of the variables in the final survival and longitudinal models, and other variables that end up significant by itself. Note that the number of observation in each model is different due to the different number of missing data in each variable.

Table 4.4: Estimated Parameters Values for Single CEA Longitudinal Models

Variable	β	SE	p-value	DF	ν^2	σ^2	ϕ	τ^2	Loglike
Time	0.004	0.001	< 0.001	3504	0.147	0.041	64.871	0.655	-2707,571
Triple negative (Yes)	-0.243	0.183	0.185	469	0.064	0.032	5.417	0.731	-2227.894
Age at diag. (≥ 44)	0.211	0.095	0.026	546	0.000	0.040	6.450	0.789	-2736.615
Age at diag. (continuous)	0.012	0.002	< 0.001	546	0.000	0.040	6.234	0.767	-2726.959
Tumour stage (III or IV)	0.301	0.083	< 0.001	527	0.000	0.039	5.819	0.691	-2562.447
Tumour degree (G1)	0.025	0.079	0.752	498	0.000	0.038	6.759	0.748	-2365.831
(G3)	0.160	0.089	0.073						
Regional Lymph nodes (N1)	0.154	0.072	0.033	525	0.000	0.040	5.828	0.689	-2630.571
(N2)	0.214	0.118	0.069						
(N3)	0.511	0.140	< 0.001						
(Nx)	0.116	0.154	0.453						
Venous vascular invasion (Yes)	0.317	0.166	0.057	546	0.000	0.040	6.479	0.792	-2737.273
Recurrence (with)	0.786	0.090	< 0.001	546	0.000	0.040	5.829	0.722	-2703.008
Primary tumour size (T2)	0.070	0.071	0.325	532	0.000	0.040	6.212	0.7470	-2653.722
(T3)	0.342	0.152	0.025						
(T4)	1.006	0.331	0.003						
(Tis)	-0.081	0.170	0.634						
(Tx)	0.174	0.229	0.449						
Neoadjuvant treatment (with)	0.377	0.107	< 0.001	546	0.000	0.040	6.391	0.783	-2732.881
Surgery (with)	-1.710	0.242	< 0.001	546	0.000	0.040	6.067	0.746	-2714.994
Menopause (Yes)	0.245	0.071	< 0.001	525	0.000	0.041	6.559	0.779	-2590.513

Results show that variable triple negative, although with significant effect on the survival of the patients, does not have a significant effect on its own on the mean progression of the tumour marker CEA (p-value = 0.185).

Contrary to results obtained for the tumour marker CA15-3 (Section 4.6.1), age at diagnosis treated whether as categorized (Survival) or as continuous (Longitudinal) variable, by itself, has a significant effect on the mean progression of the tumour marker. Both have an increasing effect of the starting point of the mean progression of the log values of the tumour marker. For age at diagnose categorized, the value 0.211 means that there is a rise of that amount in the starting point for any patient with 44 years or more at the time of diagnose. The value 0.018, for age at diagnose continuous, means that there is a rise of that amount per year of age.

As expected, since they are reported as possible prognostic factors (Chapter 1), an increase of tumour values progression is related to higher categories of variables such as tumour stage, regional lymph node involvement and primary tumour size. Although in this two last variables we can only consider a significant difference between the baseline category (N0 for regional lymph node and T1 for primary size tumour) and the higher categories (N3 and T3 or T4, respectively), since the p-values registered are lower than 0.05.

For tumour degree variable, it was only obtained a marginal significant difference between the cases that presented a G3 degree and the cases that presented a G2 degree (the baseline category), corresponding to an increase of the starting point of the mean progression of the log values of the tumour marker of 0.16.

The presence of venous vascular invasion, on its own, is responsible for an increase of the starting point of the mean progression of the log values of the tumour marker of 0.3166, comparing to the cases with no venous vascular invasion (although the p-value of 0.057 is considered marginally significant).

Lymphatic invasion and, also, estrogen receptor expression have no significant effect on the tumour marker progression, by themselves.

Recurrence of breast cancer is, as expected, related to a low survival probability as shown in previous chapters, and that result is concordant with the significant effect of this variable on the mean progression of this tumour marker. Cases that presented recurrence of breast cancer had an increment of 0.7857 on the starting point of the mean progression of the log values of the tumour marker, compared to non-recurrence cases.

Cases submitted to surgery present a lower starting point (-1.7097) of the mean progression of the tumour marker, compared to those were it was not preformed any surgery. A result similar to the one obtained in the survival analysis (Chapter 3), where cases submitted to surgery presented a higher survival probability.

Cases of menopausal patients have an increment of 0.2452 on the starting point of the mean progression of the log values of the tumour marker, compared to pre-menopausal patient's cases.

A similar result obtained for the CA15-3 longitudinal analysis (Chapter 4.6.1) is that the cases submitted to neoadjuvant treatment are related to an increase of the starting point of the mean progression of the tumour compared to those not submitted to it. That increment is of 0.3769 on the log values of the tumour. Again, that could be due to the fact that cases of large primary tumour are usually submitted to that kind of treatment in order to decrease its size for posterior surgery.

It is important to underline that all variables related to type of treatment (neoadjuvant, surgery or adjuvant) were excluded from the posterior analysis that will be presented. One other reason is that those variables are due to medical decision taken upon the tumour characteristics and thus correlation with variables related to tumour characteristics may occur.

Then, a longitudinal model was fitted, starting with all explanatory variables, and using backward stepwise for variables selection. As for CA15-3, different correlation structures for the longitudinal model given by equation 4.2 were tested, based on the analysis of the empirical variogram presented in Figure 4.10. Based on the structure of the empirical variogram, we present the results, comparing two models, both with a random intercept an individual level, and with a serial correlation structure: one with Exponential structure (REE) and the second with a Gaussian structure (REG).

Table 4.5 presents the estimated parameters of the two fitted longitudinal models REE and REG, and compares the estimates to those obtained by fitting the simple OLS model and the respective log Likelihood values.

Table 4.5: Estimated Parameters Values for General Linear Model and Longitudinal Models – CEA Tumour Marker

	REE Model		**REG** Model		**OLS** Model	
	Est	**p-value**	**Est**	**p-value**	**Est**	**p-value**
Intercept	-0.280	0.067	-0.303	0.046	-0.045	0.525
Time (month)	0.005	< 0.001	0.005	< 0.001	0.003	< 0.001
Age at diagnosis	0.013	< 0.001	0.013	< 0.001	0.011	< 0.001
Tumour Stage (III or IV)	0.284	0.002	0.285	0.001	0.267	< 0.001
Tumour Degree (Gx or G1 or G2)	-0.134	0.089	-0.133	0.093	-0.263	< 0.001
Primary Tumour Size						
(T4)	1.057	0.002	1.049	0.003	0.132	< 0.001
(Tis)	0.045	0.795	0.057	0.742	0.050	0.627
ν^2	0.263		0.413		.	
σ^2	0.037		0.068		.	
ϕ	36.596		15.272		.	
τ^2	0.368		0.184		.	
ξ^2	.		.		0.694	
Log Likelihood	-2218.830		-2302.393		-4457.261	

The longitudinal model with random effects plus exponential correlation structure was selected as the one that best describes the mean progression of the tumour, at a logarithmic scale, since it is the one with the lower Log Likelihood value (-2218.83). As expected all estimates for all covariates and between REE and REG models are quite similar, and the differences between the p-values are rather small. Hence, the choice of the adequate model relayed also on the graphical analysis of the correlation structure, as it will be described later on.

The fixed part of the longitudinal model, which describes the mean progression of the marker, is composed by the following significant covariates on the intercept component of the model: age at diagnosis, tumour stage (0/I/II versus III/IV), tumour Degree (G3 versus Gx or G1 or G2) and primary tumour size (Tx/T1/T2/T3 versus T4 versus Tis). The intercept component of the model, in this particular case, means that a patient at an earlier age of diagnosis, with a tumour stage of 0, I or II, G3 tumour degree and a T1, T2 or T3 primary tumour size will start the progression of the tumour marker with a value of 0.756 ng/mL ($\exp(-0.278)$), although this value is marginally significant for presenting a p-value of 0.067, meaning that the progression of the tumour marker could start with a value of 1 ($\exp(0)$).

Time, in month, has a significant effect on the progression of CEA values, and is responsible for an increase of 0.0045 per month of the tumour marker log values.

Age at diagnosis affects the log value of the marker at a rate of 0.0132 per year of age at diagnosis.

A tumour degree different from G3 (the highest category of this variable) has a decreasing effect on the starting log value of the tumour mean progression by an increment of 0.1341.

A patient with a tumour on stage III or IV implicates an increasing of the log value of the tumour marker by an increment of 0.2835, comparing to those with a tumour on stage 0, I or II. Also, a tumour that presented a T4 primary tumour size implicates an increase in the starting point of the marker value by an increment of 1.0572.

The correlation structure chosen to represent the variability of the data is the one that incorporates random effects at individual level with $\hat{\nu}^2 \approx 0.263$, an exponential correlation structure to describe the variability within patients with $\hat{\rho}(u) \approx \exp(-\frac{1}{36.596}.|u|)$ and $\hat{\sigma}^2 \approx 0.037$, and a measurement error with variance $\hat{\tau}^2 \approx 0.368$.

Figure 4.10 that represents the empirical and both fitted variograms for the different models. This superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structures with the empirical variogram supports the choice of the REE model, since the REE curve (green line) is the one that best approaches the empirical curve.

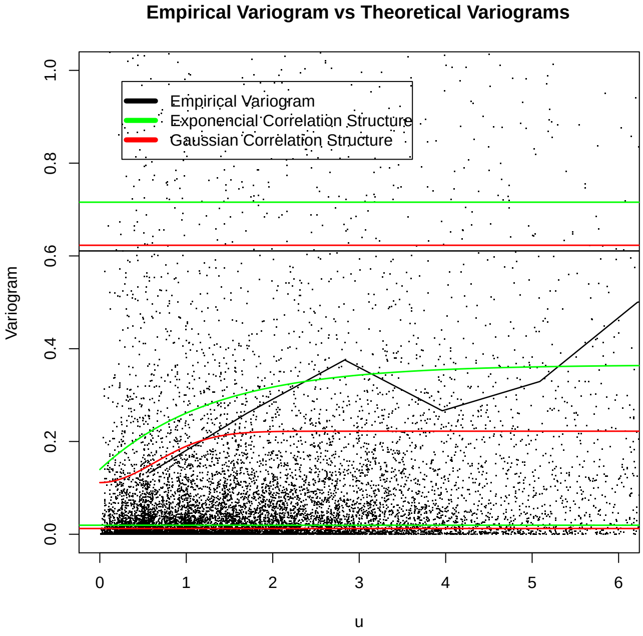


Figure 4.10: Superposition of empirial variogram and theoretical variogram - CEA tumour marker

Figures 4.11 and 4.12 present, the plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals, respectively. Observing this plots, the assumptions of the longitudinal model, concerning to the homogeneity of variances of the measurement errors and the normality of these do not seem likely to be rejected.

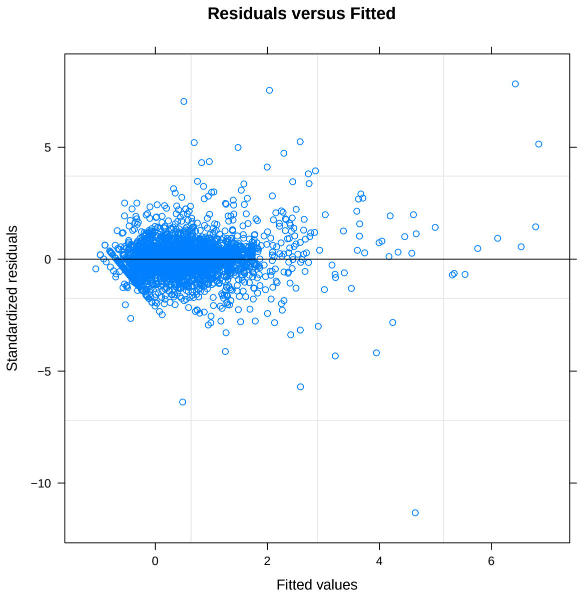


Figure 4.11: Residuals versus the Fitted Responses of CEA values.

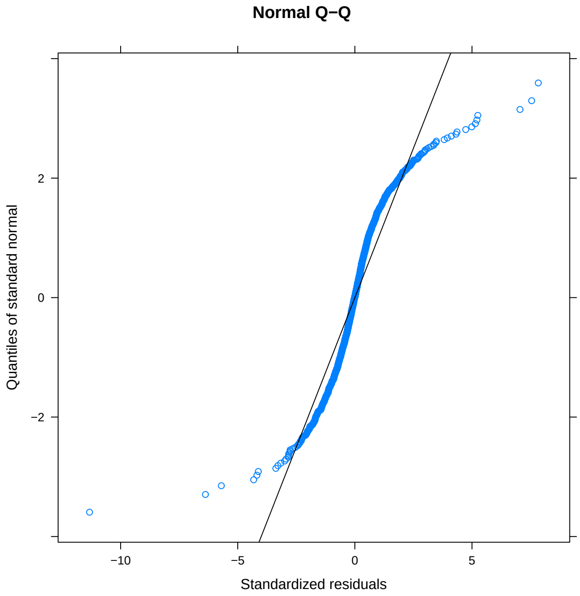


Figure 4.12: Q-Q plot of the CEA longitudinal model.

Time Since Death

A similar longitudinal analysis, as for the CA15-3 tumour marker, was performed, but now with a different time scale time since death until date of blood test in order to search for more significant results to infer about the relation between the progression of the tumour marker values and the death from breast cancer. Note that, again, with this time scale the analysis is only performed with the subset of patients that died, and also, as we are analysing the marker values from date of blood tests until death, we are dealing with duration at a negative scale. However, these patients are not necessarily the same as for CA15.3 as the number of death from breast cancer are different for both tumour markers, as already mentioned.

Figure 4.13 presents the progression in time, since date of death, of the CEA values for each patient that died from breast cancer, with grey lines, against the reference value, and the non-parametric smooth spline line with dash line, indicating the average trend of progression.

Similarly to results obtained to tumour marker CA15-3 longitudinal analysis, the smooth spline of the spaghetti plot suggests that, on average, the marker progression increases at a slow rate until about 20 months before death from breast cancer. In fact, when fitting the saturated general linear model for the subset of patients who died from breast cancer, given by equation 4.5, we detected a changing point at 2 years before death (24 months). Which means that, for the patients that died from breast cancer it was detected an abrupt rise of CA15-3 values, above the reference value, 2 years before their death.

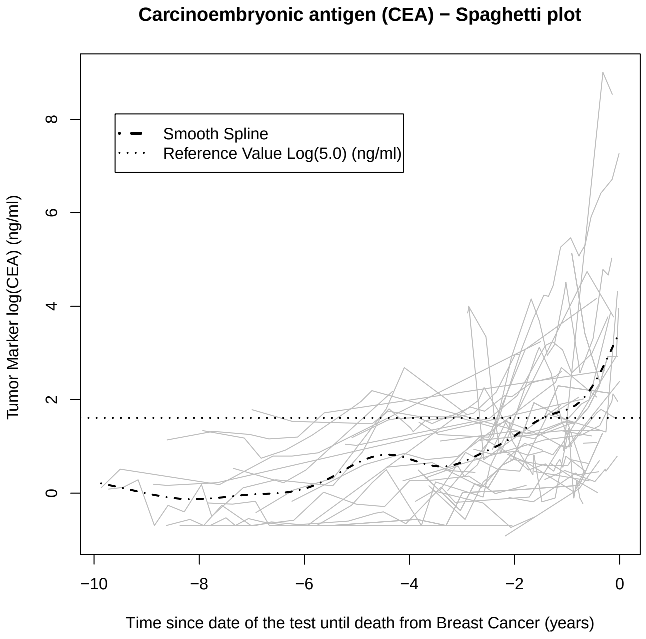


Figure 4.13: Spaghetti plot for tumour marker CEA values - for the subset of patients that died from breast cancer.

Table 4.6 summarizes and compares the estimated parameters for the two longitudinal models fitted REE and REG - with those of the general linear model (OLS Model). Note that the variable ”Time further than 2 years before death” is, in fact, time before the changing point (2 years before death from breast cancer), in the negative scale considered and, equivalently, ”Time earlier than 2 years before death” is time after the changing point, in the negative scale considered.

Table 4.6: Estimated Parameters Values for General Linear Model and Longitudinal Models For Patients Who Died From Breast Cancer – CEA Tumour Marker

	REE Model		**REG Model**		**OLS Model**	
	Est	**p-value**	**Est**	**p-value**	**Est**	**p-value**
Intercept	0.980	0.002	0.925	0.002	1.100	< 0.001
Time further than 2 years before death	0.016	< 0.001	0.016	< 0.001	0.019	< 0.001
Time earlier than 2 years before death	0.050	< 0.001	0.053	< 0.001	0.062	0.085
Bilateral (Yes)	-0.697	0.034	-0.615	0.075	-0.972	< 0.001
Venous Vascular Invasion (Yes)	1.008	0.010	0.948	0.016	1.166	< 0.001
Lymphatic Invasion (Yes)	-0.585	0.029	-0.608	0.026	-0.827	< 0.001
Estrogen Receptor (Positive)	0.924	0.001	0.964	0.001	0.947	< 0.001
ν^2	0.00000042		0.525		.	
σ^2	0.002		0.091		.	
ϕ	19.958		7.451		.	
τ^2	1.425		0.772		.	
ξ^2	.		.		1.428	
Log Likelihood	-502.163		-523.639		-724.714	

Results show that both time after the changing point, that is before two years prior to death, and before changing point have a significant effect (p-value < 0.001) on the mean progression of the tumour marker CEA values. Although both are related to an increase of the CEA values in time, that increment per month is higher after the changing point (0.050) than before the changing point (0.016).

As expected, the presence of venous vascular invasion has an increasing effect on the average CEA linear progression in time, as it is related to a worst prognostic case in the previous survival analysis (Chapter 3).

Bilateral type of tumour and a positive estrogen have a statistical significant effect on the intercept component of the model (2.66 ng/mL). Bilateral cancer cases have a decrease of 0.697 on the intercept component, and a case with lymphatic invasion a decrease of 0.585 compared to those with no lymphatic invasion. A positive estrogen receptor expression has an increasing effect on the intercept component by 0.924, compared to those with a negative expression.

A case that presents images of vascular invasion increases of the start value of the tumour marker by an increment of 1.008, comparing to those that do not present any image.

For this subset, the correlation structure chosen to represent the variability of the data is the one that incorporates random effects at individual level with $\hat{\nu}^2 \approx 4.2 \times 10^{-7}$, an exponential correlation structure to describe the variability within patients with $\hat{\rho}(u) \approx \exp(-\frac{1}{19.958}.|u|)$ and $\hat{\sigma}^2 \approx 0.002$, and a measurement error with variance $\hat{\tau}^2 \approx 1.425$.

Figure 4.14 represents the empirical and both fitted variograms for the different models. The superposition of the theoretical fitted variogram of both exponential and Gaussian correlation structures with the empirical variogram supports the choice of the REE model, since the REE curve (green line) is the one that approaches the empirical curve.

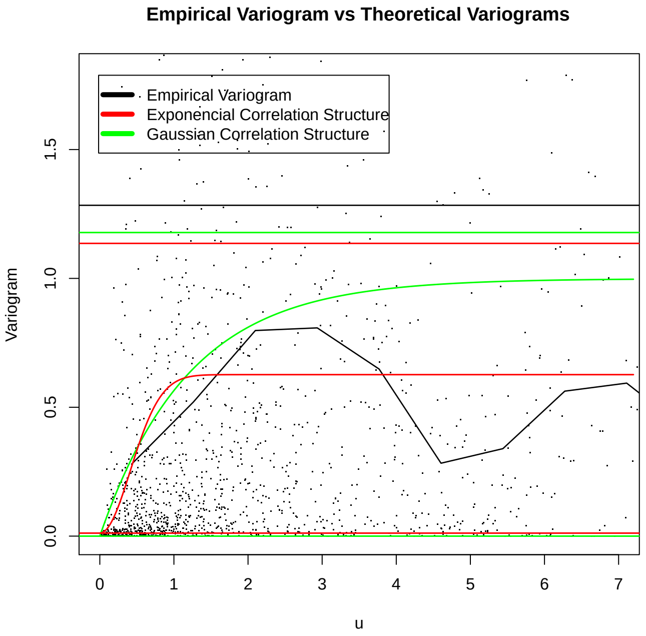


Figure 4.14: Superposition of empirial variogram and theoretical variograms, for patients who died from breast cancer - CEA tumour marker

Figures 4.15 and 4.16 present, the plot of the subject specific residuals versus the fitted responses and a Q-Q plot of the subject-specific residuals, respectively. Observing this plots, the assumptions of the longitudinal model, concerning to the homogeneity of variances of the measurement errors and the normality of these do not seem likely to be rejected.

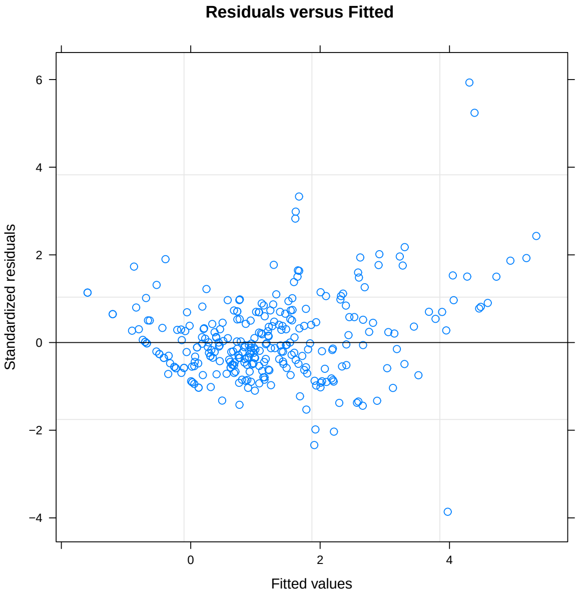


Figure 4.15: Residuals versus the Fitted Responses of CEA values.

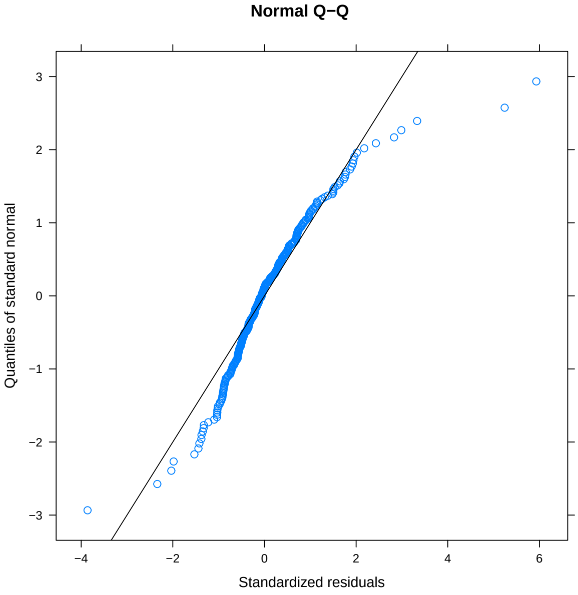


Figure 4.16: Q-Q plot of the CEA longitudinal model.

4.7 Discussion

According to medical procedures of Braga's Hospital Senology unit, an abrupt rise in values of CA15-3 or CEA tumour markers, over a reference value, is an alert sign to a possible recurrence of breast cancer.

With the previous statistical analysis presented in section 4.6 we can conclude that an abrupt rise in values of those two tumour markers is related to death from breast cancer. This can be seen by the results of the longitudinal model assuming time since death.

The resulting model fitted to explain the variation of the tumour marker CA15-3 values, confirms that the factors that have a statistically significant effect on the linear progression of the tumour marker are: age at diagnosis, tumour stage (III/IV versus 0/I/II), venous vascular invasion (Yes versus No), and Ki.67 antibody immunostaining index (high versus low). As expected, a III or IV tumour stage, the presence of venous vascular invasion, high age at diagnosis and high ki.67 index have an increasing effect on the average tumour marker progression in time, as they are related to a worst prognostic case.

However, there are other variables that, by themselves (i.e., fitting a longitudinal model considering only that specific variable alone) have a statistical significant effect on the mean progression of the tumour marker, namely: regional lymph nodes invasion, recurrence, lymphatic invasion, primary tumour size, neoadjuvant treatment and biopsy of sentinel lymph node. This can be explained by the fact that when using only a single variable the model captures variability by that only variable.

When analysing all patients that were diagnosed with breast cancer, in our study, the only variables that have a statistically significant effect on the linear progression of the tumour marker CEA are: tumour stage (III/IV versus 0/I/II), primary tumour size (Tx/T1/T2/T3 versus T4 versus Tis), age at diagnosis and tumour degree (Gx/G1/G2 versus G3). As expected, a III or IV tumour stage, a T4 type of tumour, a G3 type of tumour and age at diagnosis have an increasing effect on the average tumour marker progression in time, as they are related to a worst prognostic case (Chapter 1).

Also for CEA tumour marker, there are other variables that, by themselves have a statistical significant effect on the mean progression of this tumour marker, namely: regional lymph nodes invasion, recurrence, lymphatic invasion, surgery, menopause, neoadjuvant treatment and venous vascular invasion. Again, this results can be related by the fact that when using only a single variable the model captures variability by that only variable.

For both tumour markers, it was detected a changing point on the linear progression of the tumour markers for the subset of patients that died from breast cancer two years before the death. This means that, at that point, there is a change on the rate of progression of the tumour markers.

The risk factors for the progression of CA15-3 tumour marker, for the subset of patients that died from breast cancer, i.e. considering time since death until date of blood test, are: bilateral tumour (Yes versus No), venous vascular invasion (Yes versus No) and estrogen receptor expression (positive versus negative). As expected, the presence of venous vascular invasion has an increasing effect on the average CA15-3 linear progression in time, as it is related to a worst prognostic case in the previous survival analysis (Chapter 3). Unexpectedly, a bilateral type of tumour has a decreasing effect and a positive estrogen has an increasing effect. These two last results contradict the results from the previous survival analyses (Chapter 3) since bilateral cases are related to lower survival probability and, a positive estrogen expression is related to a higher probability of survival.

For CEA tumour marker, the risk factors for the progression of the marker, for the subset of patients that died from breast cancer, i.e. considering time since death until date of blood test, are: bilateral (Yes versus No), lymphatic invasion (Yes versus No), venous vascular invasion (Yes versus No) and estrogen receptor expression (positive versus negative). As expected, the presence of venous vascular invasion has an increasing effect on the average CEA linear progression in time, as it is related to a worst prognostic case in the previous survival analysis (Chapter 3).

A bilateral type of tumour and the presence of lymphatic invasion have a decreasing effect. A positive estrogen has an increasing effect. These three results contradict the results from the previous survival analyses (Chapter 3) since bilateral cases and lymphatic invasion are related to lower survival probability and, a positive estrogen is related to a higher probability of survival.

For all models fitted, the fact that the estimated variance of the measurement error is quite lower than the estimated variance of the OLS model, means that the fitted REE longitudinal model explains the variability of the data mainly by means of variability between patients and within patients assigning a very low value for measurement error (or white noise as usually mentioned in literature).

5

Joint Modelling of Longitudinal and Survival data

In the previous longitudinal analysis (Chapter 4), we verified that observing the subset of patients that died from breast cancer, and using as reference time the one since death until date of blood tests, the results of the time effect, as so of the explanatory variables, are different from those when considering all patients in our data base. Showing that there is a relation between survival and longitudinal measures of both tumour marker CA15-3 and CEA.

In scientific problems where the longitudinal observations may be correlated with time-to-event, joint models have been increasingly proposed, to recover information from these potentially informative censorings (Wulfsohn & Tsiatis (1997), Henderson *et al.* (2000) and Diggle *et al.* (2008)).

In fact, in medical data, such as the one presented in this study, where survival is collected simultaneously alongside the repeated measures data (CA15-3 and CEA tumour markers values), it is typically found that longitudinal data have an association to the time-to-event process of subjects (McCrink *et al.*, 2013).

Joint models for longitudinal and survival data are statistical models that allow us to understand two processes of interest simultaneously, longitudinal and survival (also known as time-to-event process), given that there is an association between them

(Sousa, 2011).

This Chapter begins with the clarification of the methodology applied in the joint modelling of the survival and longitudinal processes, introduced by a summary explanation of the main approaches reported in studies of this nature, and also describing the construction of the joint model applied on the present analysis. Subsequently, the results from the joint analysis of these two processes to the data set in study will be demonstrated, ending with a section that resumes a discussion of the significant results.

5.1 Methodology

As already explain in the previous chapter, longitudinal data is originated when different individuals are measured repeatedly over time for some response variable of interest. Time-to-event data deals with times from a reference point until the occurrence of an event of interest. Generally, longitudinal models describe the process underneath the observed data allowing for different sources of variability in the data (Diggle *et al.*, 2002) whereas classic survival models describe the hazard of an event at any given time as a function of explanatory variables (Cox, 1972).

The need for considering joint modelling these two processes arises, as Sousa (2011) explains, when the interest is on studying the association between the response variable and the survival mechanism, allowing also the estimation of parameters that represent their association.

Depending on the nature of the statistical problem and the scientific questions to be answered, joint modelling of longitudinal and time-to-event data can have different focus. It can focus on the inference on the survival model parameters, improving the inference by allowing for correlation in the longitudinal measurements; or it can have a primary interest on the longitudinal trajectory that might be associated with an event pattern, or, even, it can focus on the association between the measurements and time-to-event assuming that both repeated measurements and event time depend on a unobserved random effect.

Joint modelling of longitudinal data and survival data is already recognized in the medical community as a powerful tool. Ibrahim *et al.* (2010) present a joint modelling for cancer clinical trials and gives a general discussion for designing and analysing clinical trials data using joint models. Deslandes & Chevret (2010), propose a joint model of multiple longitudinal responses and a competing risk dropout process to patients in an intensive care unit.

The analysis presented here focusses on the detection of risk factors that affect the survival of the patients from Braga's Hospital and also, that affect the progression of tumour markers CA15-3 and CEA values for these patients. As an abrupt rise on these markers values above a certain reference value is an alarm sign for a possible tumour recurrence or even imminent death, the survival process and the longitudinal process of both tumour markers are obviously associated. As McCrink *et al.* (2013) point out, when there is an association between the two processes that is not appropriately taken into account, unnecessary bias can be introduced into the model, affecting the results and therefore the inferences obtained. For that, it is pertinent the construction of two joint models to study the association between the survival of the patients and the longitudinal progression of the CA15-3 tumour marker values, and the association between the survival of the patients and the longitudinal progression of the CEA tumour marker values.

In the statistical community, joint modelling has been widely studied in the biostatistics setting. One of the first joint models proposed was by Wulfsohn & Tsiatis (1997), followed by a generalization by Henderson *et al.* (2000). Diggle *et al.* (2008) propose a more empirically interpretable model. Recent work includes, for example, Ding & Wang (2008), Nathoo & Dean (2008), Ye *et al.* (2008), Rizopoulos *et al.* (2009), Albert & Shih (2010), Jacqmin-Gadda *et al.* (2010), Wu *et al.* (2010) and Huang *et al.* (2011).

In the context of joint modelling, Tsiatis & Davidian (2004) clarify that random effect models (Laird & Ware, 1982) are usually used to model the marker process and the individual random effects are included in the survival model. They explain that the two main approaches to jointly model this processes are: a two-stage approach, developed by Self & Pawitan (1992) (Tsiatis *et al.* (1995), Bycott & Taylor (1998), Dafni & Tsiatis (1998)) and a likelihood based approach (Faucett & Thomas (1996), Wulfsohn & Tsiatis (1997), Henderson *et al.* (2000), Wang & Taylor (2001) and Xu & Zeger (2001)).

Briefly, the two-stage method consists in fitting, primarily, a linear mixed effect model to the longitudinal covariate data, and estimate the missing or mismeasured covariates based on the fitted model; the second stage consists in fitting the survival model separately, with the missing or unobserved true covariate values substituted by their estimates from the first stage as if they were observed values and then proceed with the usual survival analysis (Wu *et al.*, 2012).

As Wu *et al.* (2012) alert in their review of joint modelling, that although this methodology is appealing for its simplicity and easy implementation with existing software, they may lead to biased inference in the estimation of the longitudinal covariate model parameters because the truncations resulted from the events are not incorporated. Also, the uncertainty of the estimation in the first stage is not incorporated in the second stage of the survival model estimation. Thus, standard errors of the parameter estimates of the survival model may be underestimated. As well, all information in the longitudinal process and the survival process is not fully combined in each model fitting to produce the most efficient estimates.

Among others alternative methods for these two approaches, there is one that should be mentioned here, often referred in literature as a naive method (alongside with the two-stage method): the time-dependent model. An extension of the Cox proportional hazard model, described in Chapter 3, that allows the incorporation of time-varying covariates introduced by Cox (1972). However, as McCrink *et al.* (2013) clarify, many authors such as Henderson *et al.* (2000), Sweeting & Thompson (2011b) and Sweeting & Thompson (2011a) verify, if a longitudinal covariate has been measured with error (as it is usual), the time-dependent Cox model will not be able to take into account this error and thus can introduce bias. Furthermore, one of the Cox model assumption is that the covariates are external, not related to the failure mechanism, which is an unrealistic assumption for the longitudinal process. In addition, and being a characteristic of the presented dataset, it is important to refer that Cox model is not applicable to unbalance data, as it requires that the value of each covariate is known at every failure time for all subjects.

The likelihood of a joint model is based on the product of the likelihood for both longitudinal data and survival data and it produces valid and the most efficient inference if the assumed models are correct. Even though its formulation requires intensive computation, it is, currently, the methodology widely adopted as it alleviates the potential bias caused by the two-stage approach and the time-dependent Cox model. For that reason it is the methodology adopted in this study.

As Diggle *et al.* (2008) clarify, the principle of likelihood approaches in joint modelling is the specication of the joint distribution of Y, a vector of repeated measurements, and a single event time F, which they denote by $[Y, F]$. There are three types of likelihood approaches that differ in their parameterization of the joint likelihood of longitudinal and survival processes: the selection models; the pattern-mixture models and the random effects models (McCrink *et al.* (2013) and Sousa (2011)). Their differences can be perceived in the following equations, where U represents the latent random effect that links the longitudinal and survival processes:

$$\text{Selection models: } [Y, F, U] = [U][Y|U][F|Y]. \tag{5.1}$$

$$\text{Pattern-mixture models: } [Y, F, U] = [U][F|U][Y|F]. \tag{5.2}$$

$$\text{Random effects models: } [Y, F, U] = [U][Y|U_1][F|U_2], \text{ where } U = (U_1, U_2). \tag{5.3}$$

As the main interest of the present analysis lies in the association between the longitudinal process and survival process, this study adopts the random effects approach that will be detailed following. Sousa (2011) gives an extended literature review on joint modelling, describing as well the selection models and the pattern-mixture models.

In the random effects approach to joint modelling, given by equation 5.3, the observable outcomes Y and F are assumed to be conditionally independent given a latent variable U, hence the joint distribution of Y and F takes the form:

$$[Y, F] = \int_U [U][Y, F|U]dU = \int_U [U][Y|U][F|U]dU. \tag{5.4}$$

In random effects joint models the association between longitudinal measurements and time-to-event is completely determine by the correlation structure between the two random effects U_1 and U_2.

One widely used method to estimate the parameters of the joint model, adopted in the present work, is the maximum likelihood (ML) that maximizes de log likelihood of the joint distribution given by (McCrink *et al.*, 2013):

$$\prod_{i=1}^{N} \int f(Y_i|U_i, \theta) f(F_i, \delta_i|U_i, \theta) f(U_i|\theta) dU_i, \tag{5.5}$$

where $f(Y_i|U_i, \theta)$, $f(F_i, \delta_i|U_i, \theta)$ and $f(U_i|\theta)$ are the densities for the longitudinal process, survival process, and random effects, respectively, and δ is the event indicator, equaling one if the event occurred and zero otherwise.

As Sousa (2011) explains, these type of models are also called shared parameter models, because the longitudinal response and missing mechanisms (in our particular case, the time-to-event mechanism) are modelled by sharing random effects. Also, in these models it is assumed that both survival and longitudinal process dependent on an underlying disease or illness progression, dened by a random effect, rather than to the actual outcome.

Historically contextualizing, after a first approach on joint modelling by Schluchter (1992), Faucett & Thomas (1996) proposed one of the first random effects models with proportional hazards for the event time. They consider the joint analysis of longitudinal measurements and survival time as the joint distribution of two models, covariate tracking model and disease risk model. The first models the longitudinal response as a linear mixed model with subject specific random effects, intercept and slope, as in a linear growth curve model, and the second allows a Cox proportional hazards model for the disease risk, with the same random effects as in the longitudinal model, assuming that these describe the true latent process (Sousa, 2011).

Following, Wulfsohn & Tsiatis (1997) propose the estimation of the parameters using all the information available at each time point, by maximising the full likelihood of the joint distribution. As Sousa (2011) clarifies, although, the models by Faucett and Thomas and by Wulfsohn and Tsiatis are the same, they use different approaches for the parameter estimation. Faucett and Thomas follow a Markov chain Monte Carlo approach, with a Gibbs sampling whereas Wulfsohn and Tsiatis use an Expectation Maximization algorithm for the estimation. Henderson *et al.* (2000) gives an extension of random effect joint model originally developed by Wulfsohn & Tsiatis (1997), including also the possibility of a serial correlation structure shared.

Usually the repeated measurements are modeled by a linear mixed effects model, already explained in Chapter 4, with random effects at individual level and the model for the survival outcome is a Cox proportional hazards model with log-Gaussian frailty. The stochastic dependence is, then, captured by allowing the Gaussian random effects of the linear model to be correlated with the frailty term of the Cox proportional hazards model.

A typical form for the latent variable U, that will be discussed later on, is the random intercept and slope effect given by (Henderson *et al.*, 2000):

$$U_i(t) = U_{0i} + U_{i1} t_{ij}, \tag{5.6}$$

where U_{0i} and U_{i1} represent the random intercept and slope effects, respectively, for subject i.

The present joint analysis adopts of random effect methodology developed by Wulfsohn & Tsiatis (1997) with the extension of Henderson *et al.* (2000), making use of two packages of R software: *JM*, developed by Rizopoulos (2010), and *joineR*, developed by Philipson *et al.* (2012). McCrink *et al.* (2013) provide, in their review of advances in joint modelling, a clear comparison of the usage of these two packages, alongside with the comparison of recent alternative methods and various formulations of the joint model built.

However, concerning this particular analysis there are two major constrains in the software used: although the random effect joint model allows to deal with right censoring and left truncated data, the two packages used are not yet able to deal with left truncated data. Hence, we will only consider a Cox model that incorporates right censoring in the survival process. For that, it is only relevant to compare estimate results for the survival process with the model, previously obtain in Chapter 3, which considers only the right censoring mechanism. The other constrain regards the fact that these two packages are not able to incorporate a serial correlation, between two measures taken in different times for the same subject, in order to account for the existence of a possible variability within subjects. In our specific case it was show, in Chapter 4, the existence of an exponential correlation structure that represents the within subject variability that should not be ignored, to both tumour markers longitudinal processes. Hence, it was decide, in order to allow for correlation between two measures taken in two different times to depend on time, to consider a longitudinal random intercept and slope, as in equation 5.6.

The choice between the two packages depends on the focus of the research being implemented (McCrink *et al.*, 2013). The joint model implemented in the *JM* package focuses on the survival process and how it is influenced by a longitudinal time-dependent covariate that is measured with error. The *joineR* package implements a joint model where the focus is on both processes with the purpose of inferring about the strength of the link between both processes. Since both objectives are of interest for the present analysis, we made use of both packages.

Although the literature on joint models is vast, this two software are the only ones that can rapidly adjust this type of models.

Maintaining the notation of the previous chapters, both model implementations use a linear mixed effects model, to represent the longitudinal process given by:

$$\text{Longitudinal Process: } Y_{ij} = \mu_{ij} + U_i + Z_{ij} = X_1\beta + U_{0i} + U_{1i}t_{ij} + Z_{ij}, \tag{5.7}$$

where X_1 is the design matrix for the fixed covariates, with corresponding regression parameters β. In the present analysis the fixed covariates considered are the ones that shown significant effect on the independent longitudinal analysis, presented on Chapter 4. The latent variable given by (U_{0i}, U_{1i}) is a realization of a $MVN(0, \Sigma)$ where $\Sigma = \begin{pmatrix} \nu_1^2 & \nu_{12} \\ \nu_{12} & \nu_2^2 \end{pmatrix}$ is the variance/covariance matrix of the random effects. Z_{ij} are Normal i.i.d. realizations of $N(0, \tau^2)$, representing the measurement error (variability non specified).

However, the main difference between *joineR* and *JM* packages relies on the way the incorporation of the latent random variable in the survival process is made. *JM* incorporates the precise estimate of the longitudinal unobserved process, $m_i(t) = \hat{\mu}_{ij} + \hat{U}_i$, into the survival process as it follows:

$$\text{JM Survival Process: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2 + \alpha m_i(t)\}, \tag{5.8}$$

where X_{2i} represents the baseline covariates, in this particular study, the covariates with significant effect on the previous independent survival analysis (Chapter 3), and β_2 the vector of corresponding regression parameters. The parameter α represents the effect of the true longitudinal response on the survival process, in this study the longitudinal responses are the log value of tumour marker CA15-3 and, also, the log value of tumour marker CEA.

On the other hand, *joineR* incorporates the longitudinal random effects into the survival model through the random effects, as it follows:

$$\text{joineR Survival Process: } h_i(t) = h_0(t)\exp\{X_{2i}\beta_2 + \gamma_0 U_{0i} + \gamma_1 U_{1i}t\}, \tag{5.9}$$

where γ_0 and γ_1 represent the effect of the longitudinal random intercept and slope on the survival process, that is, the association parameters.

Note that, actually both approaches differ only on the survival sub-model, since the longitudinal sub-model used in both software is the same. In fact, the difference between the two survival sub-models may lead to different estimate values of β_2 and, in the limit, even to differences on the effects of the covariates in the individuals' survival (i.e., covariates may have a statistical significant effect on subject's survival within the joint model fitted with *joineR* but no significant effect within the model fitted with *JM*). This can be easily perceived by the main difference between equations 5.8 and 5.9 that is the incorporation of the effect of $m_i(t) = \hat{\mu}_{ij} + \hat{U}_i$, which will lead to an effect of the longitudinal covariates in the survival process, i.e., in the estimates of the survival covariates, within the joint model fitted with *JM* package. On the other hand, estimates of β_2 within the joint model fitted by *joineR* package will only incorporate the effect of the covariates of the survival process.

To synthesize, using the model implemented by *JM* package we will be able to determinate the factors that influence the change in CA15-3 values, as so in CEA values, on the patient's and what effect that change has on their survival. *joineR* package is appropriate to use in our analysis since we are interested in determining the effect of the initial response (values of CA15-3 and also CEA), γ_0, and change in the response over time, γ_1, on a patient's survival. As there seems, from results of previous Chapter, to exist a link between the longitudinal processes of CA15-3 and CEA and patient's survival.

For model diagnose we will use graphical representation of the subject specific residuals versus the fitted responses for the validity of the mixed effects model, already explained in detail in Chapter 4. This plot is used to assess the assumption of constant variance of Z_{ij}. Also, a Q-Q plot of the subject-specific residuals is represented to verify the assumption of normality of Z_{ij}. To validate the survival process fitted we made use of the graphical representation of Cox-Snell residuals (Cox & Snell, 1968), already explained in detail in Chapter 3.

5.2 Breast Cancer Joint Analysis

5.2.1 Tumour Marker CA15-3

The joint model of the longitudinal process of tumour marker CA15-3 values and the survival process, as described previously, was fitted using the two packages, *joineR* and *JM* described in the previous section.

Analogously to the longitudinal analysis presented in Chapter 4, it was used a log-transformation of the tumour marker CA15-3 values in the longitudinal process within both joint models adjusted.

Table 5.1 presents the results from joint modelling the longitudinal process of CA15-3 tumour marker values (on logarithmic scale, as in Chapter 4 analysis) and the survival process, where the event is death from breast cancer and the reference time, for the longitudinal process, is time (in months) from tumour diagnose to death from tumour. It is shown the estimates obtained for the significant covariates considered in the previous separate longitudinal analysis (Chapter 4) and survival analysis (Chapter 3), as well its standard errors (SE) and respective p-values.

Table 5.1: Estimated parameters values for Joint Model CA15-3 tumour marker

	JoineR Package			JM Package		
Covariates	**Estimates**	**SE**	**p-value**	**Estimates**	**SE**	**p-value**
LONGITUDINAL PROCESS						
Intercept	2.212	0.170	< 0.0001	2.225	0.097	< 0.0001
Time (month)	0.006	0.001	< 0.0001	0.006	0.001	< 0.0001
Age at diagnosis	0.007	0.002	0.003	0.005	0.002	< 0.001
Tumour Stage (III or IV)	0.054	0.099	0.584	0.134	0.078	0.088
Venous Vascular Invasion (Yes)	0.753	0.269	0.005	0.642	0.112	< 0.0001
Ki.67 (low)	-0.136	0.062	0.090	-0.009	0.071	0.900
SURVIVAL PROCESS						
Triple Negative (with)	1.878	0.408	< 0.0001	2.303	0.718	0.001
Age at diagnosis (≥ 44)	-0.070	0.371	0.850	-0.636	0.522	0.223
Tumour Grade (III or IV)	2.874	0.608	< 0.0001	3.557	0.706	< 0.0001
Tumour Grade (G2)	1.332	0.738	0.071	2.451	0.910	0.007
(G3)	1.513	0.738	0.040	2.804	0.859	0.001
Lymph Node (Nx or N0 or N1)	2.729	0.830	0.001	3.998	0.632	< 0.0001
LATENT ASSOCIATION						
Association parameter(s)	γ_0 1.136	0.182	< 0.0001			
	γ_1 1.099	0.103	< 0.0001	α 1.194	0.143	< 0.0001
ν_1^2	0.2301			0.4970		
ν_2^2	1.6×10^{-7}			0.00038		
τ^2	0.0636			0.2566		
Loglikelihood	-2302.3			-2230.6		

It is important to highlight that in the present joint analysis only patients with information of CA15-3 values available were included. Which means that only 534 patients were eligible for the present analysis, whereas 540 patients were included in the independent survival analysis presented in Chapter 3. This translates into different values for the independent survival model parameters estimated when considering only the 534 patients with information on CA15-3 values. Thus, a direct comparison of estimates obtained in the survival process within the joint models with the ones obtained in the individual survival analysis must take into account that results are not based in the exactly same dataset. However, a comparison between the covariates that have effect on the survival process of both joint modelling and independent survival analysis, presented in Table 3.2, is pertinent.

Moreover, it is essential to recall that, as already explained in the previous section, both joint models fitted with the two packages only account for right censoring, as this are not yet able to fit random joint models to left truncated data.

Furthermore, it should be noted that the differences between the estimates obtained for the survival process for both packages is justified by the way these incorporate the effect of the longitudinal variable in the survival process, as explained in the previous section.

Covariates such as triple negative (yes versus no), Tumour stage (0 or I or II versus III or IV) and lymph node involvement (Nx or N0 or N1 versus N2 or N3) maintain their significant effect on patients' survival, as in the individual survival model. Even though lymph node involvement had only shown significant effect when considering left truncation in the survival model, as can be observed in Table 3.2.

Tumour grade (Gx or G1 versus G2 versus G3) has a borderline significant effect on patient's survival within the joint model adjusted with *joineR* package, but a clearly significant effect within the joint model adjusted with *JM*. Contrariwise to results from the independent survival analysis presented in Table 3.2, age at diagnosis (categorized as $<$ 44 years old versus $\geq$ 44 years old) does not show a significant effect in patients survival within both joint models adjusted.

The differences between the correlation structure of the linear mixed effect considered for the longitudinal process within the joint model and the one considered in the individual analysis presented in Table 4.3, may lead to some differences in the covariates estimates. Hence, the comparison will concern on the covariates effect on the linear mean progression of CA15-3 log value.

Comparing with the results of the individual longitudinal analysis, covariates such as age at diagnosis, venous vascular invasion (yes versus no) maintain their significant effect on initial values of the linear mean progression of CA15-3 log value. However, ki.67 index (low vs high) and tumour stage (0 or I or II versus III or IV) do not present a significant effect on the longitudinal process for both joint models, contrariwise to results presented in Table 4.3.

For the joint model obtained with *JM* package, the significance of the association parameter $\hat{\alpha}$ (p-value < 0.0001) in the survival process highlights the importance for a joint modelling of the present data. This significant association verifies the relationship between the longitudinal and survival processes. Likewise, within the joint model adjusted with joineR package, both association parameters, $\hat{\gamma_0}$ and $\hat{\gamma_1}$, are significant (p-value < 0.0001) in the survival process. Which confirms, in fact, the relevance for a joint model of the present data.

With regard to results obtain for *JM*'s joint model, through determination of hazard ratio (HR), it is apparent that individuals with higher CA15-3 log value tend to have a worst survival (HR $= \exp(1.1942) = 3.3$).

Within estimates obtained for the joint model adjusted with *joineR* package, through calculation of hazard ratios, $\hat{\gamma_0}$ indicates that individuals with initial values of CA15-3 log values that are higher than the population average tend to have worst survival (HR $= \exp(1.136) \approx 3.11$). Likewise, $\hat{\gamma_1}$ indicates that individuals who have a greater increase in CA15-3 log values linear mean progression tend to have worst survival (HR$=exp(1.099) \approx 3$). This was expected as high values of CA15-3 corresponds to a deterioration of disease.

Next, we provide diagnostics for the joint model fitted. In fact, the diagnostics given are for both longitudinal and survival sub-models.

In Figure 5.1 we present the graphical representation of Cox-Snell residuals, where the black solid line denotes the Kaplan-Meier estimate of the survival function of the residuals (with the dashed lines corresponding the 95% pointwise condence intervals), and the grey solid line, the survival function of the unit exponential distribution. This graphic suggests a good fit of the survival process within the joint model.

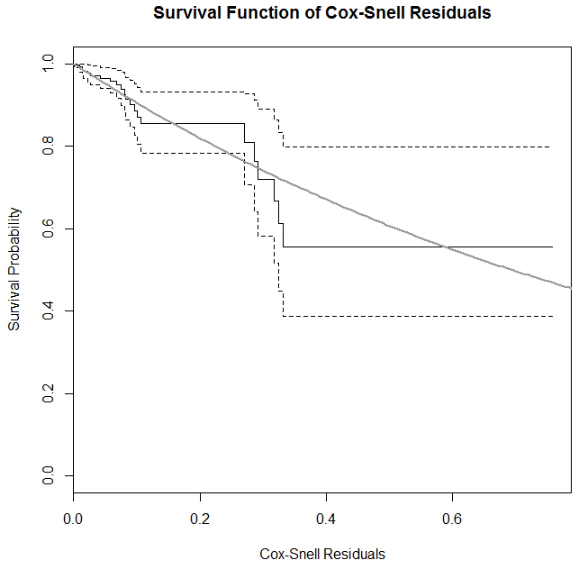


Figure 5.1: Superposition of Kaplan-Meier estimates of the survival function of the residuals and survival function of the unit exponential distribution - CA15-3 Tumour Marker.

In Figure 5.2 is represented the subject specific residual plot for the validity of the linear mixed effects model, which seems to validate the assumption of constant variance of Z_{ij}. And in Figure 5.3 is represented de Q-Q plot of the subject-specific residuals in which the assumption of normality of the Z_{ij}'s appears not to be violated.

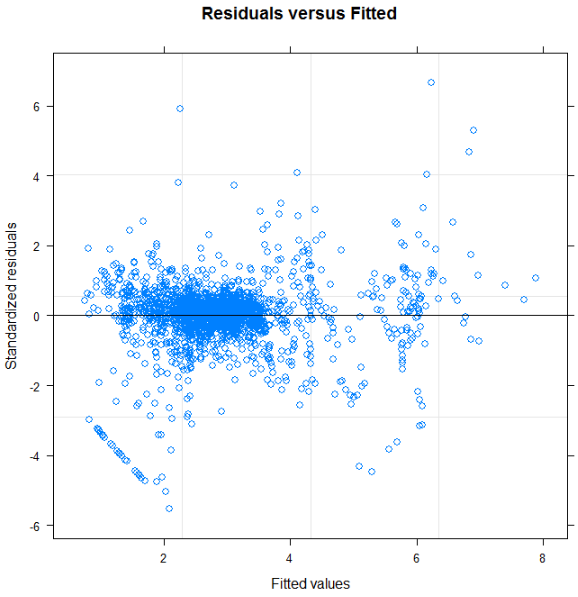


Figure 5.2: Subject-specific residual plot - CA15-3 Tumour Marker.

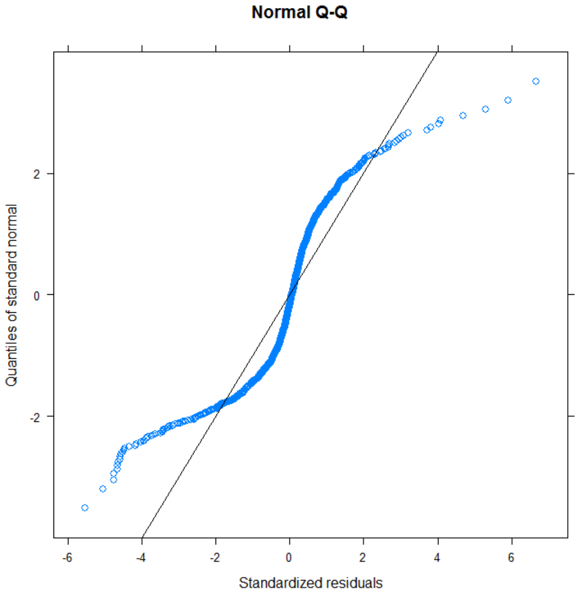


Figure 5.3: Subject-specific residual plot - CA15-3 Tumour Marker.

5.2.2 Tumour Marker CEA

Similarly as the joint analysis for tumour marker CA15-3, the joint model of the longitudinal process of tumour marker CEA values and the survival process was fitted using the two packages, *joineR* and *JM* described in the previous section.

Just as the longitudinal analysis presented in Chapter 4, it was used a log-transformation of the tumour marker CEA values in the longitudinal process within both joint models adjusted.

Table 5.2 presents the results from joint modelling the longitudinal process of CEA tumour marker values and the survival process, where the event is death from breast cancer and the reference time is time (in months) from tumour diagnose to death from tumour. It is shown the estimates obtained for the significant covariates considered in the previous separate longitudinal analysis (Chapter 4) and survival analysis (Chapter 3), as well its standard errors (SE) and respective p-values.

Table 5.2: Estimated parameters values for Joint Model CEA tumour marker

	JoineR Package			**JM Package**		
Covariates	**Estimates**	**SE**	**p-value**	**Estimates**	**SE**	**p-value**
LONGITUDINAL PROCESS						
Intercept	-0.297	0.118	0.012	-0.517	0.090	< 0.0001
Time (month)	0.004	0.001	< 0.0001	0.006	0.0004	< 0.0001
Age at diagnosis	0.011	0.001	< 0.0001	0.013	0.001	< 0.001
Tumour Stage						
(III or IV)	0.203	0.066	0.002	0.242	0.045	< 0.001
Degree						
(Gx or G1 or G2)	-0.012	0.062	0.846	-0.008	0.038	0.827
Estrogene Receptor						
(Positive)	0.151	0.064	0.019	0.190	0.059	0.001
Tumour Size						
(T4)	1.052	0.366	0.004	0.979	0.130	< 0.001
(Tis)	0.0120	0.224	0.930	0.139	0.171	0.418
SURVIVAL PROCESS						
Triple Negative (with)	3.163	0.576	< 0.0001	3.243	0.523	< 0.0001
Age at diagnosis						
($\geq$ 44)	-0.416	0.382	0.113	-0.503	0.347	0.147
Tumour stage						
(III or IV)	1.797	0.499	< 0.0001	1.259	0.367	0.0006
Tumour Degree						
(G3)	-0.464	0.456	0.317	-0.386	0.415	0.353
(Gx or G1)	-0.734	0.566	0.353	-0.986	0.625	0.115
Lymph Node						
(Nx or N0 or N1)	0.529	0.566	0.351	0.654	0.378	0.083
LATENT ASSOCIATION						
Association	γ_0 0.959	0.175	< 0.0001			
parameter(s)	γ_1 0.929	0.172	< 0.0001	α 1.048	0.115	< 0.0001
ν_1^2	0.2235			0.4757		
ν_2^2	1.94×10^{-7}			0.0003		
τ^2	0.0117			0.1320		
Loglikelihood	-2254.97			-2308.11		

Also in this analysis, it is relevant to highlight that only patients with information of CEA values available were included. Which means that only 532 patients were eligible for the present analysis, whereas 540 patients were included in the independent survival analysis presented in Chapter 4. Again, this translates into different values for the independent survival model parameters estimated when considering only the 532 patients with information on CEA values. Thus, a direct comparison of estimates obtained in the survival process within the joint models with the ones obtained in the individual survival analysis, presented in Table 3.2, must take into account that results are not based in the exactly same data set. However, a comparison between the covariates that have effect on the survival process of both joint modelling and independent survival analysis is pertinent.

Once again, it should be noted that the differences between the estimates obtained for the survival process for both packages is justified by the way these incorporate the effect of the longitudinal variable in the survival process, as explained in section 5.1.

For both joint models adjusted, covariates such as triple negative (yes versus no), Tumour stage (0 or I or II versus III or IV) maintain their significant effect on patients' survival, as in the survival model presented in Table 3.2. However lymph node involvement (Nx or N0 or N1 versus N2 or N3), tumour grade (Gx or G1 versus G2 versus G3) and age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) do not show a significant effect in patients survival within both joint models adjusted, contrariwise to the results from the independent survival analysis.

Also for the joint modelling of CEA longitudinal process and survival process, the differences between the correlation structure of the linear mixed effect considered for the longitudinal process within the joint model and the one considered in the individual analysis presented in Table 4.5, may lead to differences in the covariates estimates. Hence, a comparison merely between covariates effect on the linear mean progression of CEA log value may be appropriate.

All covariates maintain their significant effect on initial values of the linear mean progression of CEA log value, as in the individual longitudinal model presented in Table 4.5, except tumour degree (Gx or G1 or G2 versus G3). However, it is important to note that the tumour degree effect was marginally significant (p-value of 0.067) in the individual longitudinal analysis.

Analogous to results obtained for CA15-3 tumour marker, for the present joint model obtained with *JM* package, the significance of the association parameter $\hat{\alpha}$ (p-value$<$ 0.0001) in the survival process highlights the importance for a joint model for the present data. This significant association verifies the relationship between the longitudinal and survival processes. Likewise, within the joint model adjusted with *joineR* package, both association parameters, $\hat{\gamma_0}$ and $\hat{\gamma_1}$, are significant (p-value < 0.0001) in the survival process. Which confirms, in fact, the relevance for a joint model of the present data.

With regard to results obtain for *JM*'s joint model, through determination of hazard ratio (HR), it is apparent that individuals with higher CEA log value tend to have a worst survival (HR $= \exp(1.048) \approx 2.85$).

Within estimates obtained for the joint model adjusted with *joineR* package, through calculation of hazard ratios, $\hat{\gamma_0}$ indicates that individuals with initial values of CEA log values that are higher than the population average tend to have worst survival (HR $= exp(0.959) \approx 2.61$). Likewise, $\hat{\gamma_1}$ indicates that individuals who have a greater increase in CEA log values linear mean progression tend to have worst survival (HR$=exp(0.929) \approx 2.53$). This was expected as high values of CEA corresponds to a deterioration of disease.

Next, we provide diagnostics for the joint model fitted. In fact, the diagnostics given are for both longitudinal and survival sub-models.

In Figure 5.4 we present the graphical representation of Cox-Snell residuals, where the black solid line denotes the Kaplan-Meier estimate of the survival function of the residuals (with the dashed lines corresponding the 95% pointwise condence intervals), and the grey solid line, the survival function of the unit exponential distribution. The Cox-Snell residuals plot suggests a go fit of the survival process within the joint model.

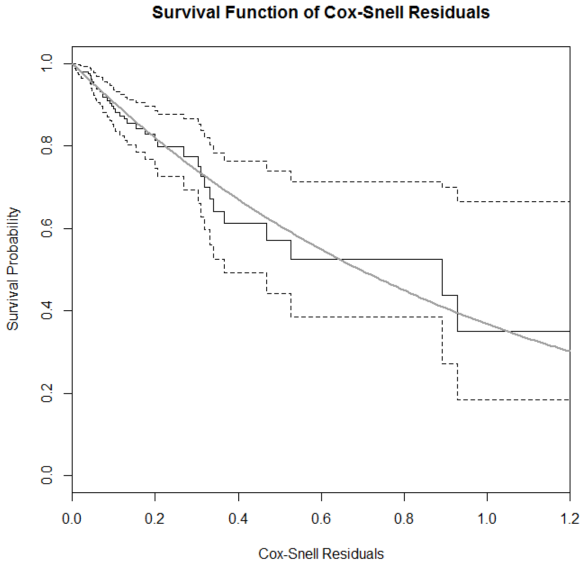


Figure 5.4: Superposition of Kaplan-Meier estimates of the survival function of the residuals and survival function of the unit exponential distribution - CEA Tumour Marker.

In Figure 5.5 is represented the subject specific residual plot for the validity of the linear mixed effects model, which seems to validate the assumption of constant variance of Z_{ij}. And in Figure 5.6 is represented de Q-Q plot of the subject-specic residuals in which the assumption of normality of the Z_{ij}'s appears not to be violated.

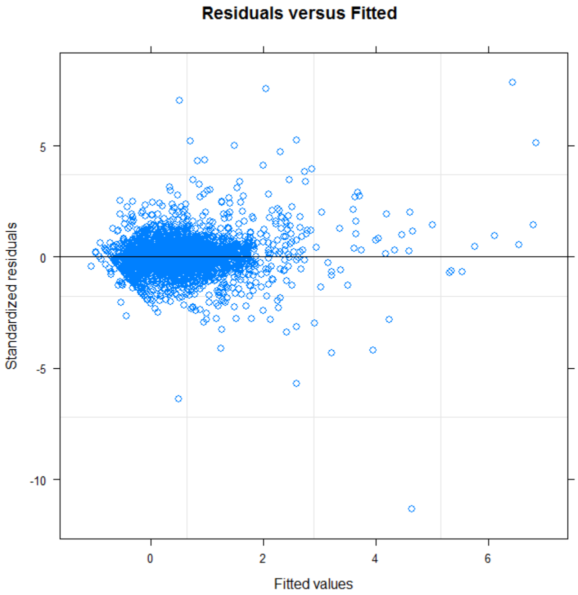


Figure 5.5: Subject-specific Residual Plot - CEA Tumour Marker.

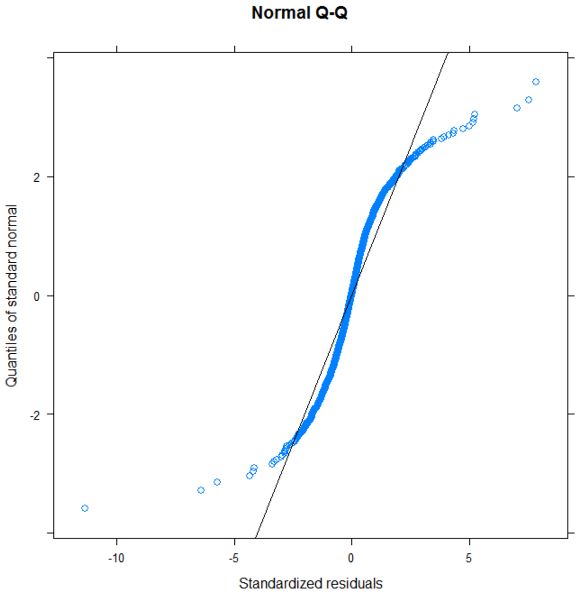


Figure 5.6: Subject-specific residual plot - CEA Tumour Marker.

5.3 Discussion

The importance of adopting a joint model approach to deal with longitudinal and survival data arises from the fact that separate analysis based on the longitudinal model and the survival model may lead to biased estimates or even to inefficient results.

Being the main goal of the present analysis to understand the joint evolution of repeated measurements of two tumour markers CA15-3 and CEA and a time-to-event process at the level of an individual subject, the shared random effect methodology was adopted.

Results showed that the longitudinal CA 15-3 and CEA values were significantly associated with the survival probability of these patients. Actually, for both tumour markers, it was apparent that individuals with higher values of the two tumour markers tend to have a worst survival.

This goes also in the same direction of the results in Chapter 4 of the individual longitudinal analysis, where we could see a decrease of tumour markers values at a higher rate, at around 2 years before death from breast cancer.

As expected, differences were found in parameter estimates for the longitudinal and survival process, compared to those obtained in the respective individual analyses. Resulting, also, that covariates with significant effects on both individual processes would cease to have significant effect in the joint model.

For the joint model with tumour marker CA15-3, in the process of survival age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) is no longer a risk factor in the probability of survival of these patients, and in the longitudinal process covariates tumour stage (0 or I or II versus III or IV) and ki.67 (low versus high) ceased to have significant effect on the average progression of CA15-3 values.

For the joint model with tumour marker CEA, in the process of survival covariates lymph node involvement (Nx or N0 or N1 versus N2 or N3), tumour grade (Gx or G1 versus G2 versus G3) and age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) cease to show a significant effect in patients survival. Also, all covariates maintain their significant effect on initial values of the linear mean progression of CEA log value, except tumour degree (Gx or G1 or G2 versus G3).

For both tumour markers, in the joint models obtained with *JM* package the statistical significance of the association parameter in the survival process highlights the importance for a joint modelling of the present data. This significant association validates the relationship between the longitudinal and survival processes. Equally, within the joint model adjusted with *joineR* package, both association parameters are significant, for both joint models, in the survival process. Which confirms, in fact, the relevance for a joint model of the present data, for both tumour markers CA15-3 and CEA.

From the exposed it is clear that independent analysis brings up bias parameter estimates and an assumption of association between the two processes in a joint model is necessary.

6

Conclusion and Further Work

The main motivation of the present work is to contribute to the understanding of the progression of a complex disease that currently affect a great percentage of women – the breast cancer – using a statistical model with more complex assumptions than the traditional analysis. As well, to contribute to the characterization of this particular disease on the Portuguese population. Since the number of studies on this subject carried out in Portugal, specifically focusing on the Portuguese population, is greatly reduced, it is extremely important the continued investment in statistical studies in oncological diseases to understand the evolution of the disease in Portugal.

This work aims, particularly, to infer which risk factors affect the survival of patients diagnosed with malignant breast tumour, and how do they affect. While analysing, also, risk factors that affect two important tumour markers utilized by doctors on the surveillance of disease progression the Carcinoma Antigen 15-3 (CA15-3) and the Carcinoembryonic antigen (CEA). Moreover, this analysis should take into consideration the expected association between the progressions in time of each of the referred tumour markers with the patient's survival.

A data set of 540 patients, eligible for the present analysis, was collected from medical records of Braga's Hospital, located in north of Portugal. Statistical methods for survival analysis and longitudinal analysis were used to study a roll of more than 50 variables that combines important and vastly reported prognostic factors with others whose effect is not recognised.

These include survival analysis methods, where the event of interest is death from breast cancer and the time is measured from the diagnosis of malignant tumour to the time at which the event occurs. All patients that did not observe the event were censored ate the end of the observational study at 30/11/2013, if not censored before. In survival analysis it was applied the proportional hazards Cox model (Cox, 1972) and the Royston-Parmar flexible parametric model (FRPM) (Royston & Parmar, 2002).

Correspondingly, longitudinal analysis methods were used to understand the progression of the values of tumour markers CEA and CA15-3. Longitudinal models with various correlation structures defined by Diggle *et al.* (2002) were considered in this particular analysis.

However, as these two processes may be associated, since CA15-3 and CEA values above a certain reference value are a warning sign for a possible tumour recurrence, which may lead to death from breast cancer, it is important to model these two processes together. Ignoring a possible association could lead, as several authors point out, to a biased estimation of parameters of interest. Hence, after conducting survival and longitudinal separate analysis, a joint modelling of these two processes to infer the combination of these was conducted.

Although many of 50 variables collected are recognized in literature as potential prognostic factors only 16 variables shown to have, individually, a statistically significant effect on these patients probability of surviving, namely: tumour stage, bilateral breast cancer, breast cancer recurrence, neoadjuvant treatment, type of surgery, venous vascular invasion, sentinel lymph node biopsy, estrogen receptor expression, progesterone receptor expression, triple negative, degree of differentiation, size of primary tumour, regional lymph node involvement, age at diagnosis (categorized), images of lymphatic invasion and hormonotherapy.

When adjusting a survival model incorporating multiple risk factors only the following five covariates shown to have a statistically significant effect on survival of those patients: triple negative (yes versus no); age at diagnosis (< 44 versus ≥ 44); tumour stage (0/I/II versus III/IV), tumour grade (G1 versus G2 versus G3) and regional lymph node metastasis degree (Nx/N0/N1 versus N2/N3). The resulting adjusted FRPM, with 0 knots, reveals that the patients with triple negative type of tumour, younger than 44 years at diagnose, with tumour in stage III or IV, grade G3 and regional lymph node metastasis degree N2 or N3, are the ones with higher risk of dying from breast cancer.

Moreover, besides right censoring alto left truncation for the survival analysis was considered given the characteristics of these data. It found that ignoring left truncation of this data in the survival model, leads to an overestimation of survival estimates.

The resulting model fitted to explain the variation of the tumour marker CA15-3 values, confirms that the factors that have a statistically significant effect on the linear progression of the tumour marker are: age at diagnosis, tumour stage (III/IV versus 0/I/II), venous vascular invasion (Yes versus No), and ki.67 antibody immunostaining index (high versus low). As expected, a III or IV tumour stage, the presence of venous vascular invasion, high age at diagnosis and high ki.67 index have an increasing effect on the average tumour marker progression in time, as they are related to a worst prognostic case.

When analysing all patients that were diagnosed with breast cancer, in our study, the only variables that have a statistically significant effect on the linear progression of the tumour marker CEA are: tumour stage (III/IV versus 0/I/II), primary tumour size (Tx/T1/T2/T3 versus T4 versus Tis), age at diagnosis and tumour degree (Gx/G1/G2 versus G3). As expected, a III or IV tumour stage, a T4 type of tumour, a G3 type of tumour and age at diagnosis have an increasing effect on the average tumour marker progression in time, as they are related to a worst prognostic case (Chapter 1).

However, modelling together these two processes, adopting the methodology of random effects (Wulfsohn & Tsiatis, 1997), shows that the association between the longitudinal processes and survival is significant and it is essential its recognition. The importance of adopting the joint modelling methodology to deal with longitudinal and survival data arises from the fact that separate analysis based on the longitudinal model and the survival model may lead to biased estimates or even to inefficient results.

Being the main goal of the present analysis to understand the joint evolution of repeated measurements of two tumour markers CA15-3 and CEA and a time-to-event process at the level of an individual subject, the random effect methodology was adopted.

Results showed that the longitudinal CA-125 and CEA values were significantly associated with the survival probability of these patients. Actually, for both tumour markers, it was apparent that individuals with higher values of the two tumour markers tend to have a worst survival.

As expected, there were found differences in parameter estimates for the longitudinal and survival process compared to those obtained in the respective individual analyses. Resulting, also, that covariates with significant effects on both individual processes would cease to have significant effect in the joint model. This shows how misleading ignoring the association between the two processes could be.

For the joint model with tumour marker CA15-3, in the process of survival age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) is no longer a risk factor in the probability of survival of these patients, and in the longitudinal process covariates tumour stage (0 or I or II versus III or IV) and ki.67 (low versus high) ceased to have significant effect on the average progression of CA15-3 values.

For the joint model with tumour marker CEA, in the process of survival covariates lymph node involvement (Nx or N0 or N1 versus N2 or N3), tumour grade (Gx or G1 versus G2 versus G3) and age at diagnosis (categorized as < 44 years old versus ≥ 44 years old) cease to show a significant effect in patients survival. Also, all covariates maintain their significant effect on initial values of the linear mean progression of CEA log value, except tumour degree (Gx or G1 or G2 versus G3).

From the exposed it is clear that independent analysis brings up bias parameter estimates and an assumption of association between the two processes in a joint model is necessary.

Although the existence of breast cancer epidemiological studies, which include survival analysis, carried out, for example, by cancer registers centres (as RORENO), the information they gather can be somehow scarce in terms of tumour and patient specificity or even in terms of information on the actual cause of death. Thus, dealing with data as the one presented in this work, obtained directly from a senology unit of a Hospital, enables to investigate specific estimates of cancer survival and the prognostic value of certain clinical factors that usually cannot be collected by the population registers. In fact, an analysis of this nature would not be possible without access to specific data at patient and tumour level. The exposed confirms the need to invest in similar studies to the one presented, within senology units of hospitals in order to have access to all clinical and personal patient information.

At the moment there are only two packages that can easily fit these models, only available in R. However, these can only fit restricted models with random effects. For example, longitudinal models with stochastic correlation structure is not possible. This is mainly due to computation complexity.

A possible extension of this model is to consider a joint multivariate model that could deal with the variables CEA and CA15-3 and, also, with the survival process, accounting for left truncation, simultaneously.

Future work includes the creation of a joint longitudinal survival-cure model of this data (or of a possible extension of this dataset), adopting, for example, the methodology proposed in Kim *et al.* (2013) work, where they present a semiparametric joint models to analyse longitudinal measurements and survival data with a cure fraction. Also, as the inclusion of left truncation data in the joint models presented was not possible the development of software that can incorporate left truncation in joint modelling of longitudinal and survival processes is aspired. As it is the construction of individual survival prediction models that incorporates longitudinal measurements of tumour markers.

References

Aalen, O. (1989). A linear regression model for the analysis of life times. *Statistics in Medicine*, **8**, 907–925. 47

Abraúl, E., Raimundo, D. & Frutuoso, C. (2011). *Manual de Ginecologia*, vol. II, chap. 38, 289–314. Permanyer Portugal. 23, 32

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. 50, 65

Albert, P. & Shih, J. (2010). On estimating the relationship between longitudinal measurements and time-to-event data using a simple two-stage procedure. *Biometrics*, **66**, 983–987. 119

Alexanian, R. & B. Barlogie, D.D. (1998). Prognosis of asymptomatic multiple myeloma. *Archives of Internal Medicine*, **148**, 1963–1965. 2

Alsaker, M., Opdahl, S., Romundstad, P. & Vatten, L. (2013). Association of time since last birth, age at rst birth and parity with breast cancer survival among parous women: A register-based study from norway. *International Journal of Cancer*, **132**, 174–181. 2

Angelis, R.D., Sant, M., Coleman, M., Francisci, S., Baili, P., Pierannunzio, D., Trama, A., Visser, O., Brenner, H., Ardanaz, E., Bielska-Lasota, M., Engholm, G., Nennecke, A., Siesling, S., Berrino, F. & Group, R.C.E..W. (2014). Cancer survival in europe 1999-2007 by country and age: results of eurocare-5 - a population based study. *Lancet Oncology*, **15**, 23–34. 10, 11, 13

Aranda-Ordaz, F. (1981). On two families of transformations to additivity for binary response data. *Biometrika*, **68**, 357–363. 49

Ardoino, I., Biganzoli, E., Bajdik, C., Lisboa, P., Boracchi, P. & Ambrogi, F. (2012). Flexible parametric modelling of the hazard function in breast cancer studies. *Journal of Applied Statistics*, **39**, 1409–1421. 2, 46

Barnett, G., Shah, M., Redman, K., Easton, D., Ponder, B. & Pharoah, P. (2008). Risk factors for the incidence of breast cancer: do they affect survival from the disease? *Journal of Clinical Oncology*, **26**, 3310–6. 2

Bastiaannet, E., Liefers, G., Craen, A., Kuppen, P., Water, W., Portielje, J., Geest, L., Janssen-Heijnen, M., Dekkers, O., Velde, C. & Westendorp, R. (2011). Breast cancer in elderly compared to younger patients in the netherlands: stage at diagnosis, treatment and survival in 127 805 unselected patients. *Breast Cancer Research and Treatment*, **124**, 801–7. 2

Bastos, J., Barros, H. & Lunet, N. (2007). Evolução da mortalidade por cancro da mama em portugal (1955-2002). *Acta Médica Portuguesa*, **20**, 139–44. 12

Bloom, H. & Richardson, W. (1957). Histological grading and prognosis in breast cancer; a study of 1409 cases of which 359 have been followed for 15 years. *British Journal of Cancer*, **11**, 359–77. 31

Borges, A., Sousa, I. & Castro, L. (2014). *Theoretical and Applied Issues in Statistics and Demography*, chap. 3, 209–220. Athens: International Society for the Advancement of Science and Technology (ISAST). 38

Borges, A., Sousa, I. & Castro, L. (2015). Longitudinal analysis of tumor marker cea of breast cancer patients from braga's hospital. *REVSTAT Statistical Journal*, **13**, 63–78. 73

Boyle, P. & Ferlay, J. (2004). Cancer incidence and mortality in europe. *Annals of Oncology*, **16**, 481–8. 3

Bycott, P. & Taylor, J. (1998). A comparison of smoothing techniques for cd4 data measured with error in a time dependent cox proportional hazards model. *REVSTAT*, **17**, 2061–2077. 120

Cadarso-Suarez, C., Meira-Machado, L., Kneib, T. & Gude, F. (2010). Flexible hazard ratio curves for continuous predictors in multi-state models: an application to breast cancer data. *Statistical Modelling*, **10**, 291–314. 46

Chiang, A., Chen, J., Chung, Y., Huang, H., Liou, W. & Chang, C. (2014). A longitudinal analysis with ca-125 to predict overall survival in patients with ovarian cancer. *Clinical Cancer Research*, **25**, 51–7. 2, 3

Ciampi, A., Hogg, S. & Kates, L. (1986). Regression analysis of censored survival data with the generalized f family - an alternative to the proportional hazards model. *Statistics in Medicine*, **5**, 85–96. 46

Cianfrocca, M. & Goldstein, L. (9). Prognostic and predictive factors in early-stage breast cancer. *The Oncologist*, **6**. 15, 17, 31

Cox, C. (2008). The generalized f distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, **27**, 4301–4312. 46

Cox, C., Chu, H., Schneider, M. & noz, A.M. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in Medicine*, **26**, 4352–4374. 44, 46

Cox, D. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 187–220. 37, 47, 118, 120, 142

Cox, D. (1975). Partial likelihood. *Biometrika*, **62**, 269–276. 47

Cox, D. & Snell, E. (1968). A general definition of residuals. *Journal of the Royal Statistical Society. Series B (Methodological)*, **30**, 248–275. 50, 126

Crump, C., McIntosh, M., Urban, N., Anderson, G. & Karlan, B. (2000). Ovarian cancer tumor marker behavior in asymptomatic healthy women: implications for screening. *Cancer Epidemiology, Biomarkers & Prevention*, **9**, 1107–11. 82

Dafni, U. & Tsiatis, A. (1998). Evaluating surrogate markers of clinical outcome when measured with error. *Biometrics*, **54**, 1445–1462. 120

David, H. & Moeschberger, M. (1978). *The Theory of Competing Risks*. 41

Dawson, S., Provenzano, E. & Caldas, C. (2009). Triple negative breast cancers: clinical and prognostic implications. *European Journal of Cancer*, **45**, 27–40. 33

de Oliveira, C.F. & da Silva, T.S. (2011). *Manual de Ginecologia*, vol. II, chap. 37, 247–288. Permanyer Portugal. 24, 25, 30, 31

Deslandes, E. & Chevret, S. (2010). Joint modeling of multivariate longitudinal data and the dropout process in a competing risk setting: application to icu data. *BMC Medical Research Methodology*, **10**. 119

Diggle, P., Heagerty, P., Liang, K. & Zeger, S. (2002). *Analysis of Longitudinal Data*. Wiley, 2nd edn. 73, 74, 77, 78, 79, 80, 81, 118, 142

Diggle, P., Sousa, I. & Chetwynd, A. (2008). Joint modelling of repeated measurements and time-to-event outcomes: The fourth armitage lecture. *Statistics in Medicine*, **27**, 2981–2998. 117, 119, 121

Ding, J. & Wang, J. (2008). Modeling longitudinal data with nonparametric multiplicative random effects jointly with survival data. *Biometrics*, **64**, 546–556. 119

dos Santos Silva, I. (1999). *Cancer epidemiology. Principles and methods*. Lyon: International Agency for Research on Cancer. 3

Duffy, M., Shering, S., Sherry, F., McDermott, E. & OHiggins, N. (2000). Ca 15-3: a prognostic marker in breast cancer. *The International Journal of Biological Markers*, **15**, 330–333. 33

Faucett, C. & Thomas, D. (1996). Simultaneously modelling censored survival data and repeatedly measured covariates: a gibbs sampling approach. *Statistics in Medicine*, **15**, 1663–1685. 120, 122

Ferlay, J., Soerjomataram, I., Ervik, M., Dikshit, R., Eser, S., Mathers, C., Rebelo, M., Parkin, D., Forman, D. & Bray, F. (2013a). Globocan 2012 v1.0, cancer incidence and mortality worldwide: Iarc cancerbase no. 11 [internet]. *Lyon, France: International Agency for Research on Cancer*. xiii, 10

Ferlay, J., Steliarova-Foucher, E., Lortet-Tieulent, J., Rosso, S., Coebergh, J., Comber, H., Forman, D. & Bray, F. (2013b). Cancer incidence and mortality patterns in europe: estimates for 40 countries in 2012. *European Journal of Cancer*, **49**, 1374–403. xiii, 7, 8, 9, 10

Fitzgibbons, P., Page, D., Weaver, D., Thor, A., Allred, D. & Clark, G. (2000). Prognostic factors in breast cancer, college of american pathologists consensus statement 1999. *Archives of Pathology & Laboratory Medicine*, **124**, 966–978. 15, 33

Fitzmaurice, G., Davidian, M., Molenberghs, G. & Verbeke, G. (2008). *Longitudinal Data Analysis*. CRC Press. 74, 75, 76

Gobbi, H., Rocha, R. & Nunes, C. (2008). Predictive factors of breast cancer evaluated by immunohistochemistry. *Jornal Brasileiro de Patologia e Medicina Laboratorial*, **44**, 131–140. 32

Goldhirsch, A., Berger, E., Mller, O., Maibach, R., Misteli, S., K, K.B., Roesler, H. & Brunner, K. (1988). Ovarian cancer and tumor markers: Sialic acid, galactosyltransferase and ca-125. *Oncology*, **45**, 281–286. 82

Gómez, G., Porta, M., Griful, E., Maguire, A., Calle, M., Malats, N., Fernandez, E., Piol, J. & Gall, M. (1996). Modelling breast cancer survival and the symptom-to-treatment interval. *Journal of Epidemiology and Biostatistics*, **1**, 175–182. 2

Grambsch, P. & Therneau, T. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–26. 50

Gramling, R., Lash, T., Rothman, K., Cabral, H., Silliman, R., Roberts, M., Stefanick, M., Harrigan, R., Bertoia, M. & Eaton, C. (2010). Family history of later-onset breast cancer, breast healthy behavior and invasive breast cancer among postmenopausal women: a cohort study breast cancer research. *Breast Cancer Research*, **12:R82**, 1–7. 2

Guadagni, F., Ferroni, P., Carlini, S., Mariotti, S., Spila, A., Aloe, S., D'Alessandro, R., Carone, M., Cicchetti, A., Ricciotti, A., Venturo, I., Perri, P., Filippo, F.D., Cognetti, F., Botti, C. & Roselli, M. (2001). A re-evaluation of carcinoembryonic antigen (cea) as a serum marker for breast cancer: A prospective longitudinal study. *Clinical Cancer Research*, **7**, 2357–2362. 33

Harrington, D. & Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika*, **69**, 553–566. 44

Harris, L., Fritsche, H., Mennel, R., Norton, L., Ravdin, P., Taube, S., Somereld, M., Hayes, D. & Bast, R.J. (2007). American society of clinical oncology 2007 update of recommendations for the use of tumor markers in breast cancer. *Journal of Clinical Oncology*, **25**, 5287–5312. 33

Henderson, R., Diggle, P. & Dobson, A. (2000). Joint modelling of longitudinal measurements and event time data. *Biostatistics*, **1**, 465–480. 117, 119, 120, 123

Herndon, J. & Harrell, F. (1990). The restricted cubic spline hazard model. *Communications in Statistics - Theory and Methods*, **19**, 639–663. 46

Herndon, J. & Harrell, F. (1995). The restricted cubic spline as baseline hazard in the proportional hazard model with step function time-dependent covariables. *Statistics in Medicine*, **14**, 2119–2129. 46

Huang, Y., Dagne, G. & Wu, L. (2011). Bayesian inference on joint models of hiv dynamics for time-to-event and longitudinal data with skewness and covariate measurement errors. *Statistics in Medicine*, **30**, 2930–2946. 119

Ibrahim, J., Chu, H. & Chen, L. (2010). Basic concepts and methods for joint models of longitudinal and survival data. *Journal of Clinical Oncology*, **28**, 2796–801. 119

J. Louhimo, a.M.C.H., Alfthan, H., Stenman, U., Jrvinen, H. & Haglund, C. (2002). Serum hcgβ, ca 72-4 and cea are independent prognostic factors in colorectal cancer. *International Journal of Cancer*, **101**, 545–548. 82

Jackson, C. (2015). flexsurv: A platform for parametric survival modelling in r. version 0.6. 38

Jacqmin-Gadda, H., Proust-Lima, C., Taylor, J. & Commenges, D. (2010). Score test for conditional independence between longitudinal outcome and time to event given the classes in the joint latent class model. *Biometrics*, **66**, 11–19. 119

Kaplan, E. & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481. 43

Keene, O. (1995). The log transformation is special. *Statistics in Medicine*, **14**, 811–9. 82

Kim, S., Zeng, D., Li, Y. & Spiegelman, D. (2013). Joint modeling of longitudinal and cure-survival data. *Journal of Statistical Theory and Practice*, **7**, 324–344. 145

Klein, J. & Moeschberger, M. (2003). *SURVIVAL ANALYSIS Techniques for Censored and Truncated Data*. Statistics for Biology and Health, second edition edn. 42, 43, 51

Kotsopoulos, J., Chen, W., Gates, M., Tworoger, S., Hankinson, S. & Rosner, B. (2010). Risk factors for ductal and lobular breast cancer: results from the nurses' health study. *Breast Cancer Research*, **12:R106**, 1–11. 2

Lai, T. & Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *Annals of Statistics*, **19**, 417–442. 43

Laird, N. & Ware, J. (1982). Randomeects models for longitudinal data. *Biometrics*, **38**, 963–974. 75, 119

Lambert, P. & Royston, P. (2009). Further development of flexible parametric models for survival analysis. *The Stata Journal*, **9**. 50

Law, M., Young, R., Babb, J., Peccerelli, N., Chheang, S., Gruber, M., Miller, D., Golnos, J., Zagzag, D. & Johnson, G. (2008). Gliomas: predicting time to progression or survival with cerebral blood volume measurements at dynamic susceptibility weighted contrast-enhanced perfusion mr imaging. *Radiology*, **247**, 490–498. 2

Leclère, B., Molinié, F., Trétarre, B., Stracci, F., Daubisse-Marliac, L. & Colonna, M. (2013). Trends in incidence of breast cancer among women under 40 in seven european countries: A grell cooperative study. *Cancer Epidemiology*, **37**, 544–549. 17

Liang, K. & Zeger, S. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22. 75, 76

Little, R. & Rubin, D. (2000). *Statistical Analysis with Missing Data*. New York: Wiley, 2nd edn. 79, 80

Liu, F., Du, F. & X, X.C. (2014). Multiple tumor marker protein chip detection system in diagnosis of pancreatic cancer. *World Journal of Surgical Oncology*, **12**, 333. 82

Lopes, C. (2011). *Manual de Ginecologia*, vol. II, chap. 36, 221–246. Permanyer Portugal. 30

Macià, F., Porta, M., Murta-Nascimento, C., Servitja, S., Guxens, M., Burón, A., Tusquets, I., Albanell, J. & Castells, X. (2012). Factors affecting 5- and 10-year survival of women with breast cancer: An analysis based on a public general hospital in barcelona. *Cancer Epidemiology*, **36**, 554–9. 2, 3

McCrink, L., Marshall, A. & Cairns, K. (2013). Advances in joint modelling: A review of recent developments with application to the survival of end stage renal disease patients. *International Statistical Review*, **81**, 249–269. 117, 119, 120, 121, 122, 124

Nathoo, F. & Dean, C. (2008). Spatial multistate transitional models for longitudinal event data. *Biometrics*, **64**, 271–279. 119

Nechuta, S., Lu, W., Zheng, Y., Cai, H., Bao, P., Gu, K., Zheng, W. & Shu, X. (2013). Comorbidities and breast cancer survival: a report from the shanghai breast cancer survival study. *Breast Cancer Research and Treatment*, **139**, 227–235. 2

Philipson, P., Sousa, I., Diggle, P., Williamson, P., Kolamunnage-Dona, R., Henderson, R. & Team, R.C. (2012). Joiner: Joint modelling of repeated measurements and time-to-event data. 74, 82, 124

Pinheiro, J. & Bates, D. (2000). *Mixed-effects Models in S and S-Plus*. Springer-Verlag. 75

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. & Team, R.C. (2015). nlme: Linear and nonlinear mixed effects models. 74, 82

Pinheiro, P., Tyczynski, J., Bray, F., Amado, J., Matos, E. & Parkin, D. (2003). Cancer incidence and mortality in portugal. *European Journal of Cancer*, **39**, 2507–20. 3

Porta, M., Gallén, M., Malats, N. & PLanas, J. (1991). The influence of diagnostic delay upon cancer survival. an analysis of five tumour sites. *Journal of Epidemiology and Community Health*, **45**, 225–230. 2

Porta, M., Malats, N., Morell, E., Gomez, G., Gallen, M., Macia, F., Casamitjana, M. & Fabregat, X. (1998). Decreased survival of patients with lung cancer admitted to a teaching hospital through the emergency department in barcelona, spain. *Journal of Epidemiology and Community Health*, **52**, 137–138. 2

Putter, H., Sasako, M., Hartgrink, H., van de Velde, C. & van Houwelingen, J. (2005). Long-term survival with nonproportional hazards: results from the dutch gastric cancer trial. *Statistics in Medicine*, **24**, 2807–2821. 46

Rizopoulos, D. (2010). Jm: An r package for the joint modelling of longitudinal and time-to-event data. *Journal of Statistical Software*, **35**, 1–33. 123

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data: With Applications in R*. Chapman & Hall Book. 51

RIZOPOULOS, D., VERBEKE, G. & LESAFFRE, E. (2009). Fully exponential laplace approximations for the joint modelling of survival and longitudinal data. *Journal of the Royal Statistical Society Series B*, **71**, 637–654. 119

RODRIGUES, V. (2011). *Manual de Ginecologia*, vol. II, chap. 34, 175–201. Permanyer Portugal. 12, 15, 17, 21

ROSENBERG, P. (1995). Hazard function estimation using b-splines. *Biometrics*, **51**, 874–887. 46

ROSSINI, A. & TSIATIS, A. (1996). A semiparametric proportional odds model for the analysis of current status data. *Journal of the American Statistical Association*, **91**, 713–721. 46

ROSSO, S., GONDOS, A., ZANETTI, R., BRAY, F., ZAKELJ, M., ZAGAR, T., SMAILYTE, G., PONTI, A., BREWSTER, D., VOOGD, A., CROCETTI, E. & BRENNER, H. (2010). Up-to-date estimates of breast cancer survival for the years 2000-2004 in 11 european countries: The role of screening and a comparison with data from the united states. *European Journal of Cancer*, **46**, 3351–7. 2

ROYSTON, P. & PARMAR, M. (2002). Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects. *Statistics in Medicine*, **21**, 2175–2197. 46, 48, 49, 65, 66, 142

SCHLUCHTER, M. (1992). Methods for the analysis of informatively censored longitudinal data. *Statistics in Medicine*, **11**, 1861–1870. 122

SCHOPPMANN, S., BAYER, G., AUMAYR, K., TAUCHER, S., GELEFF, S., RUDAS, M., KUBISTA, E., HAUSMANINGER, H., SAMONIGG, H., GNANT, M., JAKESZ, R. & HORVAT, R. (2004). Prognostic value of lymphangiogenesis and lymphovascular invasion in invasive breast cancer. *Annals of Surgery*, **240**, 306–12. 31

SELF, S. & PAWITAN, Y. (1992). Modelling a marker of disease progression and onset of disease. *In AIDS Epidemiology*, 231–255. 120

SHAPIRO, S. & WILK, M. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611. 81

SOUSA, I. (2011). A review on joint modelling of longitudinal measurements and time-to-event. *REVSTAT*, **9**, 57–81. 118, 121, 122, 123

SPARANO, J., MAKHSON, A., SEMIGLAZOV, V., TJULANDIN, S., BALASHOVA, O., BONDARENKO, I., BOGDANOVA, N., MANIKHAS, G., OLIYNYCHENKO, G., CHATIKHINE, V., ZHUANG, S., XIU, L., YUAN, Z. & RACKOFF, W. (2009). Pegylated liposomal doxorubicin plus docetaxel signicantly improves time to progression without additive cardiotoxicity compared with docetaxel monotherapy in patients with advanced breast cancer previously treated with neoadjuvant-adjuvant anthracycline therapy: results from a randomized phase iii study. *Journal of Clinical Oncology*, **27**, 4522–4529. 2

STACY, E. (1962). A generalization of the gamma distribution. *Annals of Mathematical Statistics*, **33**, 1187–1192. 45

Sturgeon, C., Duffy, M., Stenman, U., Lilja, H., Brunner, N., Chan, D., Babaian, R., Bast, R.J., Dowell, B. & Esteva, F. (2008). National academy of clinical biochemistry laboratory medicine practice guidelines for use of tumor markers in testicular, prostate, colorectal, breast and ovarian cancers. *Clinical Chemistry Journal*, **54**. 34

Sweeting, M. & Thompson, S. (2011a). Approximate nonparametric corrected-score method for joint modeling of survival and longitudinal data measured with error. *Biometrical Journal*, **53**, 557–577. 120

Sweeting, M. & Thompson, S. (2011b). Joint modelling of longitudinal and time-to-event data with application to predicting abdominal aortic aneurysm growth and rupture. *Biometrical Journal*, **53**, 750–763. 120

Therneau, T. (2015). A package for survival analysis in s. version 2.38. 38

Trichopoulos, D., Adami, H., Ebkom, A., Hsieh, C. & Lagiou, P. (2008). Early life events and conditions and breast cancer risk: from epidemioloy to etiology. *International Journal of Cancer*, **122**, 481–495. 15

Tsiatis, A. & Davidian, M. (2004). An overview of joint modelling of longitudinal and time-to-event data. *Statistica Sinica*, **14**, 809–834. 119

Tsiatis, A., DeGruttola, V. & Wulfsohn, M. (1995). Modelling the relationship of survival to longitudinal data measured with error: applications to survival cd4 counts in patients with aids. *Journal of the American Statistical Association*, **90**, 27–37. 120

Viale, G., Giobbie-Hurder, A., Regan, M., Coates, A., Mastropasqua, M., DellÓrto, P., Maiorano, E., MacGrogan, G., Braye, S., Ohlschlegel, C., Neven, P., Orosz, Z., Olszewski, W., Knox, F., Thurlimann, B., Price, K., Castiglione-Gertsch, M., Gelber, R., Gusterson, B. & 1-98., A.G.B.I.G.T. (2008). Prognostic and predictive value of centrally reviewed ki-67 labeling index in postmenopausal women with endocrine-responsive breast cancer: results from breast international group trial 1-98 comparing adjuvant tamoxifen with letrozole. *Journal of Clinical Oncology*, **26**, 5569–75. 33

Wang, Y. & Taylor, J. (2001). Jointly modelling longitudinal and event time data, with applications to aids studies. *Journal of the American Statistical Association*, **96**, 895–905. 120

Wu, L., Liu, W. & Hu, X. (2010). Joint inference on hiv viral dynamics and immune suppression in presence of measurement errors. *Biometrics*, **66**, 327–335. 119

Wu, L., Liu, W., Yi, G. & Huang, Y. (2012). Analysis of longitudinal and survival data: Joint modeling, inference methods, and issues. *Journal of Probability and Statistics*, **2012**, 17. 120

Wulfsohn, M. & Tsiatis, A. (1997). A joint model for survival and longitudinal data measured with error. *Biometrics*, **53**, 330–339. 117, 119, 120, 123, 143

Xu, J. & Zeger, S. (2001). Joint analysis of longitudinal data comprising repeated measures and times to events. *Journal Of The Royal Statistical Society Series C-Applied Statistics*, **50**, 375–387. 120

Ye, W., Lin, X. & Taylor, J. (2008). Semiparametric modeling of longitudinal measurements and time to-event dataa two-stage regression calibration approach. *Biometrics*, **64**, 1238–1246. 119

Yuste, A., Aparicio, J., Segura, A., López-Tendero, P., Gironés, R., Pérez-Fidalgo, J., Díaz, R. & Calderero, V. (2003). Analysis of clinical prognostic factors for survival and time to progression in patients with metastatic colorectal cancer treated with 5-uorouracil-based chemotherapy. *Clinical Colorectal Cancer*, **2**, 231–234. 2

Zou, Y., Zhang, J. & Qin, G. (2011). A semiparametric accelerated failure time partial linear model and its application to breast cancer. *Computational Statistics and Data Analysis*, **55**, 1479–1487. 46