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

Determinants of adherence to breast cancer screening in primary health care

Mestrado em Bioestatística

Ana Catarina da Costa Antunes dos Santos Sousa

Trabalho de Projeto orientado por:
Prof.ª Doutora Maria Helena Mouriño Silva Nunes (FCUL)
Mestre Paulo Jorge de Morais Zamith Nicola (IMPSP)

2016

“Persistence is the road to accomplishment.”

Charles Chaplin

ACKNOWLEDGMENTS

I would like to briefly thank all of those who contributed throughout my thesis journey:

Helena Mouriño – my supervisor –,

for all the strength, for all the support, for all the share and of course, for always being on my side.

Duarte Tavares,

for all the work hours, all the discussions, ideas and of course, all the support and help.

Unit of Epidemiology – Institute of Preventive Medicine and Public Health,

for your support and for allowing me to work on this project. A special thank you for Paulo Nicola for all the help and support.

Alastair Gilhooley,

a really big thank you for being much more than a boss.

All my family,

for all the support specially to my parents – Adelaide & António Sousa. Without you and without your unconditional support and persistency this thesis would not have been possible.

And last but not least, to you – Sandro Leitão –,

thank you for all your support, for always being there for me and for always making me fight and not letting me give up! You know…

ABSTRACT

Cancer is a major cause of suffering and death in the European Union. Every year around 3.2 million Europeans are diagnosed with cancer. In women, every year, there is about 331,000 cases and 90,000 deaths due to breast cancer. A burden that is expected to grow even further due to demographic trends in Europe.

But with regular and systematic examinations, using evidence-based screening tests followed by appropriate treatment, it is possible to reduce cancer mortality and improve the quality of life for ones that are suffering from cancer by detecting cancer at earlier stages, when it is more responsive to less aggressive treatment.

In December 2003, the European Council unanimously adopted a set of cancer screening fundamental principles as best practice in early detection of cancer. This Council recommended to all member states that they should screen for breast cancer every woman aged between 50 and 69 years old.

Although in Portugal, screening started in the 90's due to a pilot program where it was said that all women between 45 and 69 years old should be screened it was only in 2003 due to the European Council guideline that Portugal adopted screening with a mammogram on a biennial basis for women between 50 and 69 years old.

Mammogram screening is the only screening method that has proven to be effective. It can reduce breast cancer mortality by 20 – 30% in woman over 50 years old in high-income countries (when the screening coverage is over 70%) and is has also been associated with less disabling treatments and better quality of life after treatment.

The present work was developed in partnership with the *Unidade de Epidemiologia do Instituto de Medicina Preventiva e de Saúde Pública, Faculdade de Medicina da Universidade de Lisboa*, where the main goal is to identify the profile of women who have a longer screening delay between consecutive mammograms in primary health care units. To study the screening delay it used the Prentice-Williams-Peterson (PWP), an extension to the Cox regression model.

From the initial population (n=41,361) 1,926 women were included. All the significant variables prove to have a protective impact on the screening delay. Women who uses hormonal contraception have an 8.5% decrease on the delay when comparing with women who do not use hormonal contraception. Women with BMI in [25 ; 30[do screening mammograms 13.5% times with less delay when comparing to women with “normal” BMI ([18.5 ; 25[).While women with BMI ≥ 30 do screening mammograms 24.7% with less delay when comparing to women with “normal” BMI ([18.5 ; 25[).

Keywords: Breast Neoplasms, Early Detection of Cancer, Survival Analysis, Cox Regression Model

RESUMO

As doenças oncológicas são um dos principais problemas a nível mundial, sendo a segunda principal causa de morte em Portugal apenas atrás das doenças do aparelho circulatório.

O cancro foi a segunda causa de morte na União Europeia em 2006, seguindo as doenças cardiovasculares e tendo sido responsável por cerca de duas em cada dez mortes nas mulheres (23%) e três em cada dez nos homens (29%).

Apenas no ano de 2005 estima-se que tenham sido perdidos mais de 17 milhões de anos vida ajustados pelas incapacidades devido ao cancro na região europeia da Organização Mundial de Saúde. Segundo as estimativas da Organização Mundial de Saúde os novos casos de cancro no mundo aumentaram em 2012 e as projeções antecipam um aumento considerável para 19,3 milhões de novos casos por ano até 2025.

Devido ao envelhecimento da população espera-se que este número aumente se não forem tomadas medidas.

Os cancros da mama, colo do útero e colorretal são uma causa importante de morbilidade e mortalidade na União Europeia. Nas mulheres estes três cancros são responsáveis por cerca de um em cada dois (47%) novos casos de cancro e por uma em cada três (32%) mortes por cancro ao passo que nos homens o carcinoma colo-rectal é responsável por cerca de um em cada oito (13%) novos casos de cancro e uma em cada nove (11%) mortes o que vem aumentar a importância da implementação de programas de rastreio.

Em Portugal a situação é semelhantes com o cancro a ser a segunda cause de morte, seguindo as doenças cardiovasculares, sendo responsável por 21,1% dos óbitos.

Nas mulheres o cancro da mama é a primeira causa de morte por cancro sendo responsável por 15,9% das mortes.

Muitas vezes os médicos não conseguem explicar porque é que uma pessoa desenvolve cancro e outra não. No entanto, a investigação demonstra que determinados fatores de risco aumentam a probabilidade de uma pessoa vir a desenvolver cancro. Atualmente a tendência é de diminuição da mortalidade por cancro da mama e esta passa sobretudo pela prevenção primária. Este tipo de medida é a estratégia mais económica e eficaz no controlo do cancro e estima-se que cerca de um terço de todos os cancros possam ser evitados se forem alterados

ou evitados os principais fatores de risco como o tabagismo, o consumo de álcool, exposição à luz solar, radiação ionizante, determinados químicos e outras substâncias, alguns vírus e bactérias, dieta pobre, o escasso consumo de frutas e legumes, falta de atividade física ou excesso de peso. Contudo, ter um fator de risco ou vários não implica que a doença se desenvolva. Muitas mulheres com vários fatores de risco nunca desenvolveram a doença enquanto outras que desenvolveram a doença, aparentemente, não sofriam de qualquer fator de risco.

No entanto, também a prevenção secundária pode levar à diminuição da incidência de alguns tipos de cancro mediante deteção e tratamento das suas lesões precursoras.

O rastreio consiste na procura ativa de uma doença ou condição precursora de doença em indivíduos presumivelmente saudáveis em risco de desenvolver a doença, de modo a permitir terapêutica precoce.

Existem dois tipos de rastreio distintos, o populacional, no qual as pessoas em risco são convidadas a ser submetidas a rastreio, e o oportunista, que ocorre quando se aproveita para sugerir a indivíduos que recorrem aos Cuidados de Saúde Primários por outro motivo.

De um modo geral os programas de rastreio organizado são mais eficazes do que os rastreios oportunistas sendo mais económicos, mais fáceis de avaliar e, se necessário, mais fáceis de suspender.

Apesar de tudo, na União Europeia só menos de metade dos exames são efetuados no âmbito de programas populacionais que proporcional o enquadramento adequado para a implementação da garantia de qualidade exigida nos termos da recomendação do Conselho Europeu.

Rastreio é o processo seletivo para a deteção de formas precoces da doença em indivíduos assintomáticos, visando a melhoria do prognóstico da doença e a redução da mortalidade.

O rastreio oncológico pressupõe uma sequência de intervenções em tempo útil e de forma integrada desde a identificação da população alvo até à terapêutica e vigilância após tratamento para detetar o cancro com o objetivo de reduzir a mortalidade e, em alguns casos, a sua incidência.

Como o cancro é uma doença potencialmente letal o objetivo principal do rastreio oncológico é a redução da mortalidade por cancro e a avaliação da sua eficácia deve ser feita com base

nesta característica. No entanto, nessa avaliação é importante considerar outras consequências importantes como a utilização de recursos de saúde e o impacto na qualidade de vida.

A evidência atual é consensual sobre a utilidade de programas de rastreio do cancro em várias áreas, incluindo, no cancro da mama.

O rastreio oncológico também acarreta várias limitações. Logo à partida baseiam-se as decisões nos benefícios populacionais em detrimento dos benefícios individuais, realizando testes num grande número de indivíduos assintomáticos, dos quais a grande maioria, não tem a doença em cause e só uma pequena parte usufruirá de benefícios pela deteção precoce do cancro.

Outro problema deve-se à acuidade dos testes, mais concretamente, à existência de falsos positivos e falsos negativos. Maioritariamente as pessoas aceitam ser rastreadas pela segurança transmitida por um resultado negativo o que tornaria este o resultado ideal. No entanto, atendendo às características do teste utilizados, perante um resultado negativo, existe sempre possibilidade de se tratar de um falso negativo e a pessoa ter a doença em causa, o que leva a uma falsa sensação de segurança e possível atraso no diagnóstico e tratamento.

Para contrariar esta limitação tendem a usar-se testes com maior sensibilidade mas existe um aumento inerente do número de falsos positivos que por sua vez causam ansiedade, rotulagem do individuo e investigação adicional desnecessária com os riscos, custos e limitações associados.

Contudo, na Europa, a mortalidade relativa ao cancro da mama reduziu 19% entre 1989 e 2006 devido à implementação das medidas de prevenção estratégica e a uma maior eficácia na terapêutica.

Uma medida de prevenção é a mamografia de rastreio e em 2003, o Conselho da União Europeia recomendou a todos os estados membros que as mulheres com idades compreendidas entre os 50 e os 69 anos deveriam efetuar rastreio de dois em dois anos.

A mamografia de rastreio provou ser o método mais eficaz. Esta é muito importante pois consegue detetar o cancro da mama mesmo antes da sensação de caroço na apalpação.

O presente estudo foi desenvolvido em parceria com a Unidade de Epidemiologia do Instituto de Medicina Preventiva e de Saúde Pública, Faculdade de Medicina da Universidade de Lisboa e o seu principal objetivo é identificar o perfil das mulheres que apresentam um maior atraso relativamente ao rastreio entre mamografias consecutivas. Isto é, as mamografias

deveriam ser efetuadas de dois em dois anos, e o objetivo é identificar o perfil das mulheres que mais tempo deixam passar após a marca dos dois anos. Para traçar o perfil da mulher usaram as variáveis disponíveis e tiveram-se em conta os mais comuns fatores de risco de cancro da mama. O objetivo de identificar estas mulheres é perceber se existe algum fator comum explicativo que se possa introduzir no rastreio para que estas cumpram os dois anos.

Para modelar os tempos foi usado o modelo de Prentice-Williams-Peterson (PWP), uma extensão do Modelo de Regressão de Cox.

Da população inicial (n=41.361) 1.926 mulheres foram incluídas no estudo. Todas as variáveis significativas provaram ter um efeito protetor em relação ao tempo de não rastreio da mulher. Mulheres que usam contraceção hormonal apresentam um decréscimo no tempo de não rastreio de 8,5% quando comparadas com mulheres que não usam contraceção hormonal. Mulhers com índice de massa corporal dentro do intervalo [25;30[fazem mamografias de rastreio 13,5% com menos atraso quando comparadas com mulheres com índice de massa corporal considerado normal, [18.5;25[. Enquanto que mulheres com índice de massa corporal superior a 30kg/m2 apresentam um tempo de não rastreio inferior em 24,7% quando comparadas com mulheres com índice de massa corporal considerado normal.

Palavras-chave: Neoplasia da Mama, Deteção Precoce de Cancro, Análise de Sobrevivência, Modelo de Regressão de Cox

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Chapter 1

INTRODUCTION

1. INTRODUCTION

1.1 Breast cancer

Breast cancer is the most popular type of cancer within women, with approximately one million new cases every year. This cancer is also the women's secondary cause of death in the occidental world [1].

In Portugal, it is the most frequent cancer in women with 4,500 new cases every year. This means 11 new cases per day with a daily mortality rate of 4 women with this disease [2].

The incidence of breast cancer is increasing in the developing world due to increase life expectancy, increase urbanization and adoption of western lifestyles but this increase has been counteracted in developed countries due to early diagnostic strategies and increased therapeutic effectiveness [3].

Indeed, low effective early detection programs result in a high proportion of women presenting at the latter stages of the disease, that as well as the lack of adequate diagnosis and treatment facilities are the most responsible for low survival rates [3].

Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy – particularly in p53 genes [4]. So breast cancer is an uncontrolled growth of breast cells.

Normally, the cells replace themselves through an orderly process of cell growth – healthy new cells take over as the old ones die out. Over time, mutations can "turn on" certain genes and "turn off" others in a cell [5]. The changed cell gains the ability to keep dividing without control or order, producing more cells just like it and forming a tumour.

Breast cancer is caused by a genetic abnormality – a mistake in the genetic material. However, only 5 – 10% of cancers are due to an abnormality inherited from the parents. Instead, 85 – 90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process and lifestyle in general [5].

It is possible to take steps to help the body stay as healthy as possible, these steps may have some impact in the risk of getting breast cancer but they cannot eliminate the risk.

1.2 Risk factors

Although it is common sense that is not feasible to control whether one will have cancer, or not, we are starting to understand how it is possible to lower the risk. If it is possible to understand what to do to lower the risk then it is also possible to identify who has an increased risk. The most common risk factors are listed in Table 1 [1], [5].

Table 1. Risk factors.

Risk Factors	
Gender	Simply being a woman is the main risk factor for developing breast cancer. Men can develop breast cancer but it is very rare (0.1%).
Age	The risk of developing breast cancer increases as one gets older. About two out of three invasive breast cancers are found in women 55 years or older.
Family history	Women with one first-degree relatives who have been diagnosed with breast cancer before 55 years old have a higher risk of developing the disease. Nevertheless, fewer than 15% of women with breast cancer do have a family member with this disease.
Personal history of breast cancer	If somebody has been diagnosed with breast cancer, then it is be 3 to 4 times more likely to reoccur in the same or other breast.
Genetics	About 5 – 10% of breast cancers are thought to be hereditary caused by abnormal genes passed from parents to child.
Race and Ethnicity	White women are slightly more likely to develop breast cancer than African American, Hispanic and Asian women. Nonetheless African American women are more likely to develop more aggressive breast cancer at a younger age.
Overweight	Overweight women have a higher risk of being diagnosed with breast cancer compared to women with a healthy weight, especially after menopause.
Pregnancy history	Women who have not had a full-term pregnancy or that had their first child after 30 years of age have a higher risk of breast cancer compared to women who gave birth before 30 years old.
Breastfeeding history	Breastfeeding can lower breast cancer risk especially if a woman breastfeeds for longer than 1 year.
Menstrual history	Women who started menstruating younger than the age of 12 have a higher risk of breast cancer later in life due to breasts forming earlier, meaning they are ready to interact with hormones inside and outside the body sooner.
Breast changes	If a person was diagnosed with certain benign breast conditions, the risk of developing breast cancer is higher.
Using Hormone Replacement	Current or recent past users of HRT have a higher risk of being diagnosed with breast cancer.
Alcohol	Research consistently shows that drinking alcoholic beverages increases a woman's risk of hormone-receptor-positive breast cancer.
Unhealthy food	Diet is thought to be at least partially responsible for about 30% – 40% of all cancers. No food or diet can prevent women from getting breast cancer but some healthy choices can make the body healthier and boost the immune system.
Lack of physical activity	Research shows a link between exercising regularly at a moderate or intense level for 4 to 7 hours per week and a lower risk of breast cancer.
Light exposure at night	The results of several studies suggest that women who work at night have a higher risk of breast cancer compared to women who work during the day.

Some of the factors associated with breast cancer cannot be changed: gender, age, etc… but others can, by making some healthier choices. This is the case in regards to alcohol or tobacco use.

But risk factors do not tell the whole story. Having a risk factor, or even several, does not necessarily mean that the disease will trigger [1]. Most women who have one or more breast cancer risk factors never develop the disease, while many women with breast cancer have no apparent risk factors – other than being a woman and getting older. Even when a woman with risk factors develops breast cancer it is hard to know just how much these factors might have contributed to that individual case [1].

In Europe, the mortality rate due to breast cancer had a reduction of 19% between 1989 and 2006 due to the implementation of preventive strategies and a greater effectiveness on therapy [6].

One preventive strategy is screening mammograms and in 2003, the Council of the European Union recommended to all member states that they should screen every woman between 50 and 69 years of age. Despite the recommendation most of the European countries only do an opportunist screening [7].

In Portugal, the screening started in the 90's, due to a pilot program implemented by the Region of Centro under the European Program against cancer [8]. In this program all women between 45 and 69 years of age should be screened [9]. Later, through the guideline 2003/878/CE, the European Council recommended that the screening should be done with a mammogram on a biennial basis for women between 50 and 70 years old [10].

Mammogram screening is the only screening method that has proven to be effective. It can reduce breast cancer mortality by 20 – 30% in woman over 50 years old in high-income countries when the screening coverage is over 70% [4]. Is has also been associated with less disabling treatments and better quality of life after treatment [11], [12].

The breast cancer screening has been able to reduce the mortality rate by 15% in the world and more or less 30% in Portugal. Where nine in each ten women had done screening mammograms, according to the last Portuguese National Health survey 2005/2006 [13]. A study about breast cancer showed a lower mortality in Portugal due to the increase in early detection of the disease and better access to more effective treatments [14].

According to Harris [15], observational research could potentially help to modify existing screening programs in several ways:

- By showing ways in which it is possible to improve effectiveness by changing the way screening programs are implemented;
- By finding advances in treatment or personal factors that reduces the magnitude of screening effect to the point at which the benefits no longer outweigh the harms and costs;
- By finding that the screening program grows more effective over time.

There are reasonable arguments for any of these three possible future trends in breast cancer screening programs.

1.3 Main goal

The present study is part of the project "*Impact of the automatic call for breast cancer screening in primary health care*" developed by the *Unidade de Epidemiologia do Instituto de Medicina Preventiva e de Saúde Pública, Faculdade de Medicina da Universidade de Lisboa*.

The main objective of the present work is to identify the profile of women who have a longer screening delay between consecutive mammograms in primary health care units. The time period under consideration spans from January 2001 to January 2013. Due to the dynamic nature of this study, each woman within the database might have multiple mammograms. An illustrative example of the event times under consideration is displayed in Figure 1, for a hypothetical woman.

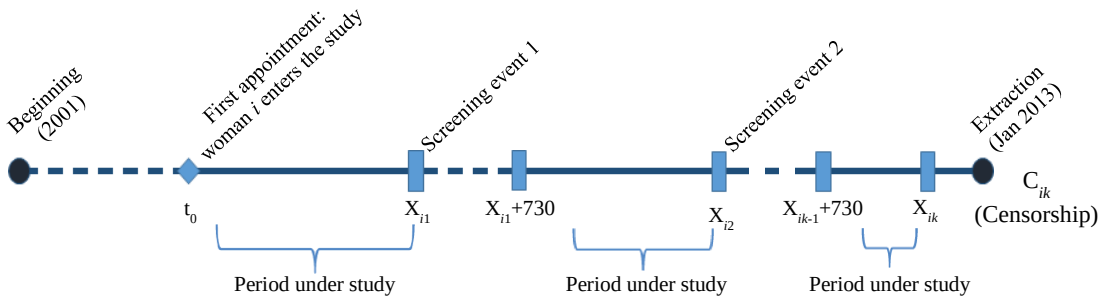


Figure 1. Mammograms' dynamics: Multiple event times for a hypothetical woman.

1.4 Overview

Chapter 1 gives an overview on breast cancer and risk factors. It also explains the main goal of this thesis.

Chapter 2 starts with an overview of the major project in which this thesis is included. The covariates and the response variable are also presented. Afterwards, we focus on theoretical issues about the survival analysis. More precisely, it describeds the basic Cox regression model. A brief review of the usual procedures to evaluate the proportional hazards assumption of the Cox regression model is provided. A summary of the most common residuals used in this context is also presented. The chapter ends with one of the newer areas of application of survival analysis: the use of the Cox regression model for describing multiple events per subject. It reviews three common models to accommodate the feature of the data sets: Andersen-Gill model, Wei-Lin-Weissfeld model and Prentice-Williams-Peterson model. In the present study, it applied the Prentice-Williams-Peterson model to describe mammogams' dynamics. Therefore, it also provideds the hazard function and the partial likelihood function for modelling the delay between consecutive mammogams.

Chapter 3 summarises the most relevant results from exploratory analyses to the data sets under consideration. Then, it applies the Prentice-Williams-Peterson model to measure the impact of the covariates under consideration to the response variable, that is, screening delay between consecutive mammograms. Assessing the fit of the estimated model was also carried out.

Finally, the main conclusions and discussion on these results are given in Chapter 4. Directions for future research are also provided.

Chapter 2

METHODOLOGY

2. METHODOLOGY

2.1 The project

The present work is part of the project "*Impact of the automatic call on the breast cancer screening in primary health care"*, which is an observational retrospective longitudinal cohort study of the primary health care units users in the larger and most populated region of Portugal – the area of *Lisboa e Vale do Tejo*. The major goal of the main project was to determine the effectiveness and clinical pathways on breast cancer screening performed in the primary care settings. The aim was to evaluate the impact of screening program in a global perspective, considering all stages, from the invitation process, to the screening, detection, referencing, diagnosis, treatment, follow-up and defined outcomes. This is because a screening program is much more than applying a screening technique to a vulnerable population.

This study aims at identifying the factors that lead to screening delay between mammograms. The time period under analysis spans from January 2001 to January 2013. It is worth stressing that there can be some biases, due to the institutional incentive for mammograms screening.

2.2 Study population

2.2.1 Database, covariates and sampled data

On 8th January 2013 individual clinical data was extracted from 10 primary health care units with Medicine One®, which is the software for the electronic medical records used by the doctors.

This extraction allowed having information regarding socio-demographic, clinical and comorbidities, history of mammograms and their results, history of invitations for breast cancer screening, utilization of primary health care – frequency of appointments.

So our study considers individual data, and not institutional, regional or any other form of aggregated data. Therefore, the impact of hospitals or other institutional practices to which suspected mammograms are referenced are not systematically introduced in the study.

This study uses a sample of women that were followed on the primary health care units discriminated in Table 2.

Table 2. Women's distribution by primary health care units and intervention area.

Primary health care unit	Number of women	Intervention Area
Amato Lusitano	3,095	Amadora
Cidadela	5,013	Cascais
Dafundo	4,424	Cruz Quebrada – Dafundo
FF-Mais	5,370	Fernão Ferro
Magnólia	5,100	Santo António dos Cavaleiros
Marginal	3,531	Estoril
Rodrigues Miguéis	4,798	Benfica
Tílias	3,609	São Domingos de Benfica
Tornada	2,905	Carvalhal / Tornada
Villa Longa	3,516	Vialonga

As it is possible to see in Table 2 the population in study is made up of 41,361 women but, of course, not every woman satisfied the criteria to be included in the study.

The first rule of inclusion, following the European Union preventive strategy, was to consider only women between 50 and 70 (exclusive) years of age. As the study began on the 1st of January 2001 all women were included, where the first doctor's appointment via the primary health care unit on or after that date, as illustrated in Figure 2.

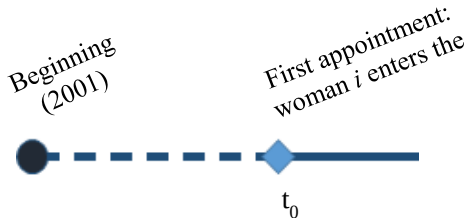


Figure 2. Beginning of the study vs. woman entering the study.

The second rule of inclusion was to consider women with no missing data.

The covariates focus on contraception, alcoholic habits, tobacco habits, the age as at the woman enters the study, the body mass index during the study, menarche's age and the

number of doctor's appointments between consecutive mammograms [3], [16], [17], [18], [19], [20], [21].

A summary of the covariates under study is presented in Table 3.

Table 3. Description of the covariates under study.

Variable	Type	Description
Age	Numerical	woman's age as at she enters the study
Menarche	Numerical	age when the woman started menstruating
BMI	Categorical	categorization of the average BMI during the study
Contraception	Categorical	hormonal vs. non-hormonal
Alcohol	Categorical	drinker vs. does not drinker
Tobacco	Categorical	smoker vs. does not smoker
Doctor's appointments	Numerical	number of appointment's between consecutive mammograms (at doctor's office, at home or by telephone)

By restricting the age to the time at which the woman enters the study, the number of women dropped from 41,361 to 22,830.

Therefore all women were included according to their age at the moment of the screening event [15], and where no data was missing from the body mass index (BMI) and menarche's age variables because they can be related with the appearance of cancer [21].

Right censoring occurs when a subject leaves the study because a certain event occurs, or the study ends before the event has occurred. For this particular project there were five censorship criteria [22]:

- Achieving 70 years old, as per the European Union preventive strategy and the inclusion criteria's
- The woman's enrolment in the primary health care unit ends, from that moment on there was no more information for that woman
- Diagnosis of breast cancer, because if cancer it is diagnosed the strategy changes
- Death
- End of the study on the 8th January 2013

And by restricting the missing data the number dropped again to 1,926 women, as represented in Figure 3.

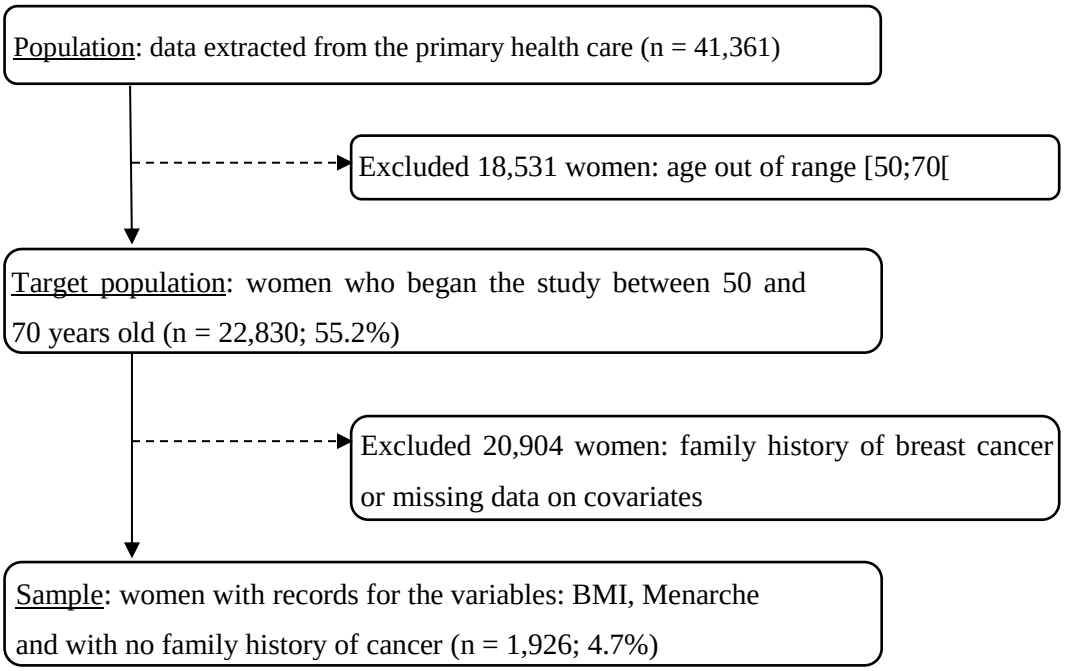


Figure 3. Study design: describing the sample data for modelling purposes.

In the end, by combining all the criteria the sample is composed of 1,926 women. This final sample will be used for our study and for modeling the screening delay between consecutive mammograms. The distribution of the sampled data through the primary health care units is displayed in Table 4.

Table 4. Percentage of sampled women by each Health Care Unit.

Primary Health Care Unit	Population size	Number of women in the sample	% of women
Amato Lusitano	3,095	57	1.8%
Cidadela	5,013	116	2.3%
Dafundo	4,424	369	8.3%
FF-Mais	5,370	313	5.8%
Magnólia	5,100	269	5.3%
Marginal	3,531	198	5.6%
Rodrigues Miguéis	4,798	63	1.3%
Tílias	3,609	56	1.6%
Tornada	2,905	119	4.1%
Villa Longa	3,516	366	10.4%

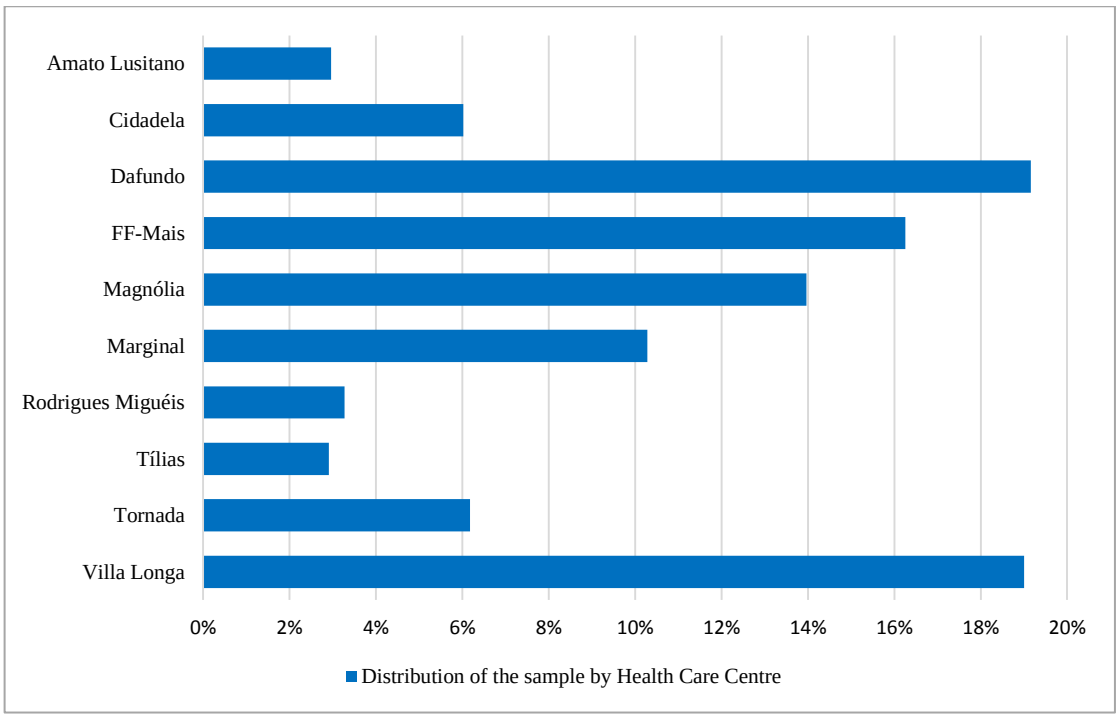


Figure 4. Percentage of women in the sample enrolled by Health Care Unit.

2.2.2 Response variable

As it was written above, the main goal is to understand the variables that have impact on the screening delay between consecutive mammograms.

Therefore the response variable is the delay between two consecutive mammograms, as illustrated in Figure 4. This delay begins two years and one day after the date when the mammogram took place and it ends on the previous day of the following mammogram.

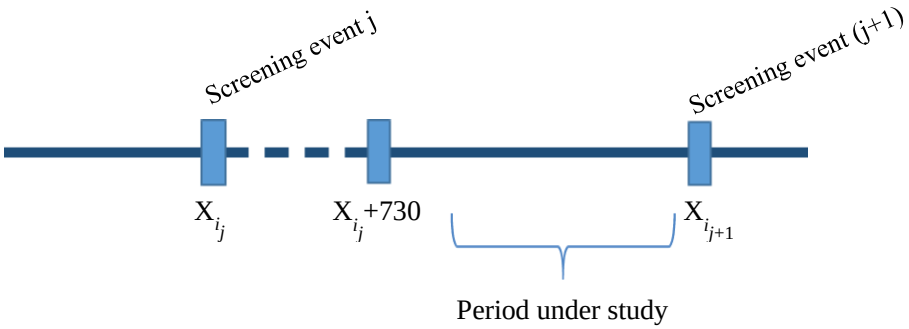


Figure 5. Definition of the screening delay.

Is expected that the delay between consecutive mammograms is low as this can be impacted by the doctor's opinion. In Health Economics this is called Agency relationship and it represents the role of a health professional in determining the patient's best interest and acting in a fashion consistent with it, as demonstrated in Figure 5. The patient is the principal and the health professional is the agent. In health care, the situation can become complicated by virtue of the facts that the professional has an important role in determining the demand for a service as well as its supply and, also that doctors are expected – in many systems – to act not only for the patient but also for the society [23].

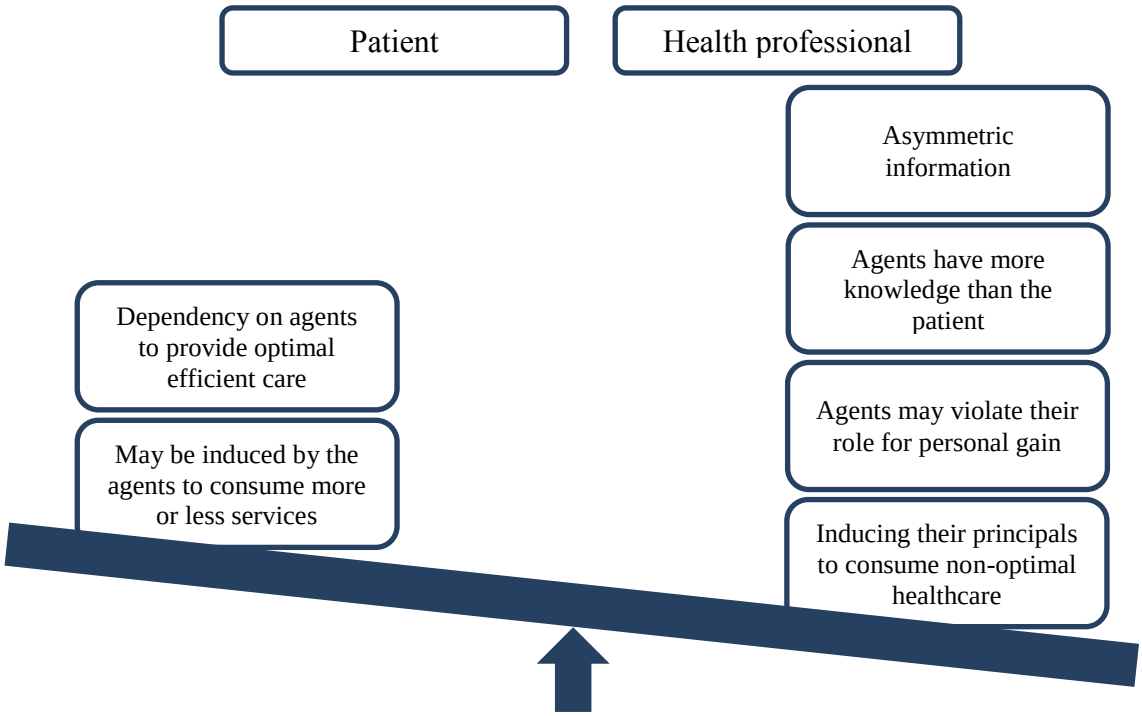


Figure 6. Agency relationship.

2.2.3 Ethical questions

From ongoing projects the *Unidade de Epidemiologia* already had the permission from the *Comissão de Ética da Faculdade de Medicina*, the *Administração Regional de Saúde de Lisboa e Vale do Tejo,IP* and the *Comissão Nacional de Protecção de Dados* for studying the impact of breast screening invitation in the primary health care units. This includes the ability to access to the Primary Care Electronic Medical Records and the Primary Care Referral

System, with close collaboration of the *Administração Regional de Saúde de Lisboa e Vale do Tejo*.

For this project, a new submission for the *Comissão de Ética da Faculdade de Medicina da Universidade de Lisboa*, *Comissão de Ética para a Saúde da Administração Regional de Saúde e Vale do Tejo, IP e Comissão Nacional de Protecção de Dados* was required and approved, as well as the permission to interlink the three databases – Primary Care Electronic Medical Record, Primary Care referral system and the South Regional Cancer Registries.

2.3 Cox Regression Model

A very common issue in medical research and longitudinal studies is to model the relationship between a set of independent variables and the survival, or time to loss or censoring, outcome. The main reason why this research question cannot be addressed by straight forward multiple regression techniques, is to do with censored or incomplete event times. In these cases the Cox proportional hazard model has become the most used procedure [24], [25], [26], [27].

Considering T_i the time to the event of interest for each subject *i*, $i = 1, \ldots, n$, where *n* is the sample size. Thereafter we assume that the random variables T_i are independent and identically distributed with *T*. The survival function is defined as $S(t) = P(T > t)$. The hazard function, at time *t*, is the instantaneous rate of failure at time *t*, that is,

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \le T < t + \Delta t \mid T \ge t)}{\Delta t}.$$

The cumulative hazard function takes the form, $\Lambda(t) = \int_0^t \lambda(s)ds, \quad t > 0.$

Suppose that C_i is the censoring time for subject *i*, $i = 1, \ldots, n$. The C_i's may be random variables or predetermined constants. We assume that each C_i is independent of the corresponding T_i. Let $X_i = \min\{T_i, C_i\}$ be the follow-up time, and $\delta_i = I(\{T_i \le C_i\})$ is the status 0/1 indicator which is 1 if T_i is observed, and 0 if the observation is censored. Therefore, the observed data corresponds to the pair (T_i, δ_i). Additionally, the data set might

include a vector-valued covariates for each individual *i*, denoted by $\mathbf{Z}_i$, $i = 1, \ldots, n$. The main goal is to estimate the hazard function or assess how the covariates affect it [24].

The Cox Regression model is the most popular procedure for modelling the time it takes for an event to occur, and it defines the hazard function for the individual *i*, given the covariates, as:

$$\lambda_i(t; \boldsymbol{Z}_i) = \lambda_0(t) e^{\beta_1 Z_{i1} + \cdots + \beta_p Z_{ip}} = \lambda_0(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_i}, i = 1, \ldots, n, \quad (1)$$

where $\lambda_0(t)$ is the unspecified baseline hazard function (that is, the hazard function for the model with no covariates, $\boldsymbol{Z}_i = \mathbf{0}, \forall_i$), which is a nonnegative function of time. $\mathbf{Z}_i$ is a *p*-dimensional vector of the observed covariates for subject *i*, and $\boldsymbol{\beta}$ is a *p*x1 vector of coefficients representing the effects of the covariates [24], [28].

The event rates cannot be negative and this feature explains why the exponential function takes here a crucial role. Therefore, the Cox Regression model is a loglinear model for the covariates. In fact, expression **(1)** can be rewritten as $\log\big(\lambda_i(t; \boldsymbol{Z}_i)\big) = \log\big(\lambda_0(t)\big) + \boldsymbol{\beta}' \boldsymbol{Z}_i$. Therefore,

$$\log\left(\frac{\lambda_i(t; \boldsymbol{Z}_i)}{\lambda_0(t)}\right) = \beta_1 Z_{i1} + \beta_2 Z_{i2} + \cdots + \beta_p Z_{ip}.$$

Also, model **(1)** is a semi-parametric model since the baseline hazard is non-parametric. The non-parametric term $\lambda_0(t)$ of the Cox Regression model makes the model flexible since no specific distribution is assumed for the baseline group.

From this model is possible to define the hazard ratio for two individuals *j* and *k* with fixed vectors of covariates, namely $\mathbf{Z}_j$ and $\mathbf{Z}_k$,

$$\frac{\lambda(t; \boldsymbol{Z}_j)}{\lambda(t; \boldsymbol{Z}_k)} = \frac{\lambda_0(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_j}}{\lambda_0(t) e^{\boldsymbol{\beta}' \boldsymbol{Z}_k}} = e^{\boldsymbol{\beta}' (\boldsymbol{Z}_j - \boldsymbol{Z}_k)},$$

which is a constant function over time. Therefore, this model is also called the Cox Proportional Hazard model.

In short, the Cox proportional hazard model can be distinguished in two parts: the first part is the underlying hazard function, called the baseline hazard, $\lambda_0(t)$, is the hazard function for the individual when all independent variables are equal to zero; the second part describes how the hazard function varies in response to the explanatory covariates [24].

Another advantage of the Cox Regression model is the easy interpretation of the regression parameters.

The Hazard Ratio (HR) is, by definition, the ratio of two hazard rates corresponding to the conditions described by two levels of a covariate. For illustrative purposes, suppose there is only one dichotomous covariate in the model (p=1), with levels 0 (z=0) and 1 (z=1). The HR of the individual with z=1 and the one with z=0, is given by:

$$HR = \frac{\lambda(t|z=1)}{\lambda(t|z=0)} = \frac{\lambda_0(t)e^{\beta}}{\lambda_0(t)} = e^{\beta} \Leftrightarrow \log HR = \beta.$$

Hence, an individual with z=1 is exp(β) times more likely to experience the event than an individual with z=0. Therefore, the parameter β quantifies the increase (or decrease) in the log-hazard ratio for a unit increase in the covariate. If $\beta > 0 \Rightarrow e^{\beta} > 1$, so the risk (or the hazard) of the event increases by (HR-1)% for an individual with z=1 compared to one with z=0. If $\beta < 0 \Rightarrow e^{\beta} < 1$, so the risk (or the hazard) of failure of the event decreases by (1-HR)% for an individual with z=1 compared to one with z=0.

In terms of the survival at time t, $S(t; z)$, it can be shown that $S(t; z) = e^{\left(-\Lambda_0(t)e^{\beta z}\right)}$.

Thus,

$$S(t; z = 1) = \left(e^{-\Lambda_0(t)}\right)^{e^{\beta}} = \left(S(t; z = 0)\right)^{e^{\beta}} = \left(S(t; z = 0)\right)^{HR}.$$

This means that if the $HR=e^{\beta}$ is:

- Equal to 1 (that is, β=0): the covariate does not have a significant meaning in the survival time, when comparing an individual with z=1 to one with z=0;
- Higher than 1 (that is, β>0): the covariate has a nonprotective effect (i.e., lower survival) when comparing an individual with z=1 to one with z=0;
- Less than 1 (that is, β<0): the covariate has a protective effect (i.e., higher survival) when comparing an individual with z=1 to one with z=0.

A word of caution should be added here. In this study, the survival time corresponds to the time interval between two consecutive mammograms. Hence, the interpretation of the parameters of the model (and of the survival function) is the opposite of the standard interpretation. More precisely, the less the survival time the less the time interval between two consecutive mammograms; and vice-versa.

If the covariate is continuous, the ratio of the hazards for any two individuals who differ in the value of the covariate by *k* units is:

$$HR = \frac{\lambda(t|z=x+k)}{\lambda(t|z=x)} = \frac{\lambda_0(t)e^{\beta(x+k)}}{\lambda_0(t)e^{\beta x}} = e^{\beta k} \Leftrightarrow \log(HR) = \beta k.$$

Hence, *k* unit change in the covariate leads to the log hazard ratio of βk.

This features of the Cox proportional hazards model also applies to a *p*-dimensional vector of covariates: the coefficient β_j, j=1,…,p, represents the increase in the log hazard ratio for a unit increase in the corresponding covariate, z_j, j=1,…,p, holding the values of the other covariates constant.

Inference on the vector of unknown parameters, **β**, is based on the partial likelihood function [28], [29]. In this framework, we cannot use the full likelihood due to many nuisance parameters involved [29]. The partial likelihood depends only on the ordering of the survival times, not the actual values [24].

Suppose there are *m* distinct survival times, $t_1 < t_2 < \ldots < t_m$, for a sample of *n* observations ($n \geq m$). For each time point t_j there is always a group of individuals at risk, denoted by $R(tj) = R_j$. $\mathbf{Z}_i$ is the covariate vector for the individual *i* (i = 1,…,n).

Cox, 1975, showed that the partial likelihood is given by [29]:

$$\mathrm{L}(\boldsymbol{\beta}) = \prod_{j=1}^{m} \frac{e^{\boldsymbol{\beta}' \mathbf{z}_j}}{\sum_{i \epsilon R_j} e^{\boldsymbol{\beta}' \mathbf{z}_i}}.$$

The partial likelihood is not, in general, a likelihood in the sense of being proportional to the probability of an observed data set. However, it can be treated as likelihood for purposes of asymptotic inference [24]. In fact, Cox, 1975, [29] proved that, under very broad conditions, the usual properties of the maximum likelihood estimates and tests based on the partial likelihood still hold. More precisely, if $\widehat{\boldsymbol{\beta}}$ is the maximum partial likelihood estimator of **β** (hereafter designated by maximum likelihood estimator) then $\widehat{\boldsymbol{\beta}}$ is consistent and asymptotically Normal distributed with mean **β** and variance-covariance matrix given by the inverse of the observed information matrix $I^{-1}\widehat{\boldsymbol{\beta}}$ [24]. The Newton-Raphson algorithm is used to solve the partial likelihood equation given by:

$$\hat{\beta}^{n+1} = \hat{\beta}^{n} + I^{-1}\left(\hat{\beta}^{n}\right) U\left(\hat{\beta}^{n}\right),$$

until converge, starting with the initial guess $\hat{\beta}^{0}$.

2.4 Assessment of fitting and residual analysis

Once the Cox model is fitted with survival data, regression diagnostics are necessary for verifying whether the statistical model fits the data appropriately or meets the proportional hazards assumption [30].

The Cox Regression model can fail in various ways. The functional form of the individual covariates may be misspecified. Furthermore, the regression coefficients may not be constant over time – violation of the proportional hazards assumption. Therefore, the proportional hazards assumption is vital to the interpretation and use of a fitted proportional hazards model.

Several ideas for residuals based on a Cox Regression model have been proposed on an ad-hoc basis, most with only limited success [24]. In the last decades, the theoretical basis for the Cox Regression model has been solidified by connecting it to the study of counting processes and Martingale theory. Therefore, the current and most successful methods are all based on counting process arguments, and in particular on the individual-specific counting process martingale that arises from this formulation [24]. Section 2.4.1 outlines the common procedures to check the proportional hazards assumption. Another graphical approach to attain this goal is based on the Schoenfeld residuals. This procedure will be postponed to Section 2.4.2.4, where we will introduce the Schoenfeld residuals. Section 2.4.2 provides a brief description of the most common residuals in the framework of the Cox regression.

2.4.1 Testing the proportional hazards

When modelling a Cox proportional hazard model a key assumption is the proportional hazards [24] – the assumption of a constant relationship between the dependent variable and the covariates which means that the hazard functions for two individuals at any point in time are proportional and independent of time. That is, the risk of an event of two individuals is the same no matter how long they survive [31].

There are several approaches for testing the proportional hazards assumption of the Cox model, ranging from simple graphical displays to sophisticated statistical tests. In this work, an overview of the most common procedures to detect nonproportionality is carried out. Section 2.4.1.1 describes a graphical procedure based on the Kaplan-Meier estimates. Nonproportionality evaluation by time-dependent covariates is described in Section 2.4.1.2.

2.4.1.1 Kaplan-Meier survival curves

For time-fixed variables that have a small number of levels, the simplest check of proportional hazards is to plot the survival curves for the different levels of the covariate. In Section 2.3, we showed that the survival function under the Cox model proportional hazards assumption is $S(t;z) = e^{-\Lambda_0(t)e^{\beta z}}$. Thus, $\log\big(-\log\big(S(t;z)\big)\big) = \log(\Lambda_0(t)) + \beta z$. Plot of the standard Kaplan-Meier estimates of survival at different levels of the covariate, and also on log-log scale, should be approximately parallel, if the proportional hazards is satisfied [24].

This method does not work well for continuous predictor or categorical predictors with many levels because the graph becomes too "cluttered".

Plot of Kaplan-Meier estimates of survival, whether transformed or not, have some limitations. Namely, the curves become sparse at longer time points; there is no reliable way to quantify how close to “parallel lines” is “close enough”; there is not a clear relationship between these plots and standard tests of proportional hazards [24].

2.4.1.2 Including time-dependent covariates in the Cox model

There are several approaches to formally test the proportional hazards assumption. Therneau and Grambsch [24] review of some of the most common methodologies. In the present work, we focus on the statistical test based on introducing time-dependent variables in the model under evaluation. For simplification purposes, we consider here that there is only one covariate in the Cox regression model.

Therneau and Grambsch [24] developed a methodology based on testing the assumption that the coefficients of the model are not time-dependent, i.e. do not change over time. This methodology relies on the creation of time-dependent covariates and then testing if the respective coefficients for interaction (first analysed separately, and then globally) are significantly different from zero.

An overview of the procedure [24]: Generate the time dependent covariates by creating interactions terms, and include them in the model. The choice of the function could be based on theoretical considerations; inspired by the smoothed residual plot, etc... Two of the most common interaction terms are the linear and log functions, leading to the following log hazard functions:

$$\log\frac{\lambda(t|z)}{\lambda_0(t)} = \beta(t)z = (a+bt)z = az + bzt;$$

$$\log\frac{\lambda(t|z)}{\lambda_0(t)} = \beta(t)z = (a+b\log t)z = az + bz\log t$$

If the time dependent covariates are significant, then those predictors do not satisfy the proportional assumption. In this case, we would need to test the null hypothesis $H_0: b = 0$, in either situations. If this hypothesis were rejected, then we would conclude that the proportional hazards assumption was not appropriated.

2.4.2 Residuals

In linear regression models, it is straightforward to compute residuals from the difference between the observed and the expected values for a continuous outcome variable. The lack of adequate information makes computation of regression residuals challenging in the Cox model, in turn prompting the development of a number of residual types for assessing the adequacy of the proportional hazards model [30]. Residuals may even play an important role in an attempt to assess globally goodness-of-fit.

The majority of the residuals that have been proposed in this context, mainly on an *ad hoc* basis, are of limited success. The current and most successful method is based on counting process arguments, and in particular on the individual-specific counting process martingale that arises from this formulation. Barlow and Prentice [32] provided the basic framework and further initial work was done by Therneau [24].

This section focus on four types of residuals that have been largely used in survival analysis:

- Martingale residuals are used to determine the functional form of a covariate and also the outliers.
- Deviance residuals are used to identify outliers.
- Score residuals are used to identify influential observations.
- Schoenfeld residuals are used to test the independence between residuals and time and hence is used to test the proportional hazards assumption.

It is worth stressing that in the counting process formulation, the pair of variables (T_i, δ_i), referred to in Section 2.3, is replaced by the pair of functions $(N_i(t), Y_i(t))$, where $N_i(t)$

represents the number of observed events in [0, *t*], for individual *i*; the indicator variable $Y_i(t)$ takes the value 1 if individual *i* is under observation and at risk at time t, and 0 otherwise.

Consider the following counting process martingale for the individual *i*,

$$M_i(t) = N_i(t) - \int_0^t Y_i(s)\lambda_i(s)ds, \quad (2)$$

where $\lambda_i(s)$ is the hazard function for individual *i*, $i = 1, \ldots, n$.

Denoting $E_i(t) = \int_0^t Y_i(s)\,\lambda_i(s)ds$, then expression **(2)** can be written as $N_i(t) = E_i(t) + M_i(t)$.

Therneau and Grambsch [24] defines the martingale residuals process as,

$$M_i(t) = N_i(t) - E_i(t) = N_i(t) - \int_0^t Y_i(s)e^{\hat{\boldsymbol{\beta}}\mathbf{Z}_i}\hat{\lambda}_0(s)ds = N_i(t) - \int_0^t Y_i(s)e^{\hat{\boldsymbol{\beta}}\mathbf{Z}_i}d\hat{\Lambda}_0(s),$$

where $\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate and $\hat{\Lambda}_0$ estimates the baseline cumulative hazard (for details see [24]).

2.4.2.1 Martingale residuals

The martingale residual, for the *i*th individual, is defined by $M_i = N_i - E_i$, $i=1,\ldots,n$, at the end of the study. Formally, $M_i = \widehat{M}_i(\infty) = N_i(\infty) - \widehat{E}_i(\infty)$. Hereafter, we will use M_i as a shorthand for $\widehat{M}_i(\infty)$. Martingale arguments are used to show that the counting process is suitably centered and scaled and is asymptotically normally distributed.

Therneau and Grambsch [24] mentioned four properties for the martingale residuals parallel to the familiar properties from an ordinary linear model:

i. $E(M_i) = 0$. The expected value of each residuals is 0, when evaluated at the true parameter vector $\boldsymbol{\beta}$;
ii. $\sum \widehat{M}_i = 0$. The observed residuals based on $\widehat{\boldsymbol{\beta}}$ must sum to 0;
iii. $Cov(M_i, M_j) = 0$. The residuals computed at the true parameter vector $\boldsymbol{\beta}$ are uncorrelated;
iv. $Cov(\widehat{M}_i, \widehat{M}_j) < 0$. The actual residuals are negatively correlated, as a consequence of condition (ii).

Also, martingale residuals have a heavily skewed distribution with zero mean and they range in the interval $(-\infty,1)$. For censored observations, martingale residuals are negative [31].

Martingale residuals cannot fulfill all the properties that the linear model residuals do. The overall distribution of the residuals may not be helpful in the global assessment of the fit. In the linear model, the sum of squared residuals provides an overall measure of the goodness-of-fit. In the Cox Regression model when comparing models with approximately the same number of parameters, the best model do not need to have the lowest sum of squares error [24].

An additional procedure for model validation for linear models is plotting the residuals against the fitted values. A model would be accepted if a reasonable structure less horizontal band of points is observed. This is however useless for martingale residuals since the fitted values and the martingale residuals are negatively correlated.

In the linear context, the errors are assumed to be normally distributed. If the model fits the data, the residuals should have approximately the Normal distribution. The most common diagnostic plot is the Normal QQ plot, where the ordered residuals are plotted against the expected Gaussian order statistics. If the data follow a Normal distribution, then the points on the QQ plot will roughly fall along a straight line. In the Cox model context, the analogue would be the comparison of the martingale residuals to a unit exponential distribution. Therneau and Grambsch [24] shows that this procedure is a hopeless endeavour. This has to do with the fact that the semiparametric estimators of the baseline cumulative hazard rescales the martingale residuals to be roughly exponential no matter how bad the model is.

Direct assessment of the residuals, as a measure of the difference between the observed and expected values, allow us to identify individuals that are poorly fitted by the model. On the other hand, plots of the martingale residuals against individual covariates should be linear if the proportional hazards model is appropriate.

2.4.2.2 Deviance residuals

The martingale residual is highly skewed, particularly for single-event survival data, with a long right-hand tale. The deviance residual is a normalizing transform. If all covariates are time-fixed, the deviance residual is given by:

$$d_i = sign(\widehat{M}_i)\sqrt{-\widehat{M}_i - N_i \log\left(\frac{N_i - \widehat{M}_i}{N_i}\right)},$$

Where sign(.) is the sign function which takes the following values: 1, if its argument is positive; 0, if its argument is zero; -1, if its argument is negative. The martingale residual for the *i*th subject is denoted by M_i.

The martingale residuals are must more symmetrically distributed around zero than the martingale residuals.

It can be shown that the deviance residuals are formally equivalent to the Pearson residuals of the generalized linear models [24]. The deviance residuals were designed to improve on the martingale residuals for revealing individual outliers, particularly in plotting applications. In practice it has not been as useful as anticipated [24].

The sum of squared deviance residuals cannot be used to compare models because there is no guarantee that this quantity decreases with improved model fit.

The deviance residuals are often used in assessing the goodness-of-fit of a proportional hazards model. Unusual patterns of plots of deviance residuals versus individual covariates indicate that the proportional hazards model is inadequate [31].

2.4.2.3 Score residuals

The score residuals quantify, for each individual, its contribute to the partial likelihood function. Instead of a single residual for each individual, there is a separate residual for each individual and each covariate. Therefore, the score residual for the *i*th individual on the *j*th covariate is given by:

$$U_{ij} = \int_0^{+\infty} \left[\boldsymbol{Z}_{ij}(s) - \bar{z}_j(\widehat{\boldsymbol{\beta}}, s)\right] dM_i(s),$$

where $\boldsymbol{Z}_{ij}(s)$ is the value of the *j*th covariate, for the *i*th subject, if this observation is still at risk at time *s*; $\bar{z}_j\left(\hat{\boldsymbol{\beta}}, s\right)$ is the weighted mean of the column vector of length *n*, the number of individuals, associated with the *j*th covariate over the observations still at risk at time *s*.

The Score residuals are useful for assessing individual influence: large values of the score residual imply large influence of the *i*th individual on the estimate of the parameter associated with $\boldsymbol{Z}_j$, that is, β_j. They are also useful for robust variance estimation.

The sum of the score residuals is equal to zero. Also, they are uncorrelated with one another.

2.4.2.4 Schoenfeld residuals

Tests and graphical diagnostics for the proportional hazards assumption, presented in section 2.4.1, may be based on the scaled Schoenfeld residuals. Schoenfeld proposed the first set of residuals applied with the Cox Regression model [33]. Likewise to the score residuals, instead of a single residual for each individual there is a separate residual for each individual and each covariate.

The Schoenfeld residuals are, for data without tied event times, given by:

$$r_k = \boldsymbol{Z}_{i(k)} - \bar{z}(\widehat{\beta}, t_{ik}),$$

where $\boldsymbol{Z}_{i(k)}$ is the covariate vector of the *i*th subject experiencing the *k*th event, at the time of that event; and $\bar{z}\left(\hat{\boldsymbol{\beta}}, t_k\right)$ is the weighted mean of the covariates, for the *i*th subject, over those at risk at the time of the *k*th event.

The Schoenfeld residuals are only defined at the uncensored survival times [31]. In fact, the Schoenfeld residuals are equal to zero for all censored subjects – the majority of the packages set the value of the Schoenfeld residuals to missing for subjects whose observed survival time is censored [34].

The graphs of Schoenfeld residuals against the survival time or a covariate can be used to check the adequacy of the proportional hazards model. Unusual patterns of plots of the Schoenfeld residuals versus survival time, or individual covariates, indicate that the proportional hazards model is inadequate. Extreme departures from the main cluster indicate possible outliers or potential stability problems [31].

2.5 Multiple events per subject

The Cox Regression model was presented in section 2.3 and in that model only one event can occur for each individual. The model is simple and easy to interpret, however information about multiple and competing events is lost. There are many multivariate extensions of the survival models, namely:

- Reoccuring events

  For each individual the same type of event can occur repeatedly
- Competing risks

  Only one event per individual yet the event can be of different types
- Multi-state models

  Several events and types can occur for each individual

Since this study concerns to reoccurring events for each subject, the focus will be on these models. There are several models proposed in the literature for modelling these events but it is important to distinguish between ordered and unordered data sets. Unordered datasets are not arranged in hierarchically order, e.g. family data – each family is a correlated group yet there is no constraint that one family member should die before another. Ordered data sets take in consideration the subject, i.e. each individual has the chance of experiencing the same event multiple times.

For ordered datasets, we have three approaches:

- Andersen-Gill (AG) Model
- Wei, Lin and Weissefeld (WLW) Model
- Prentice, Williams and Peterson (PWP) Model

Although these models have a common base there are some small differences, which will be explained next.

2.5.1 Symbols and notation

For the three models explained in detail below, it was considered the following notation, with $i=1,\ldots n$, $k=1,\ldots,K$:

- T_{ki} are the survival times for the individual i and event k
- C_{ki} are the censorship times for the individual i and event k

- $X_{ki} = \min(T_{ki}, C_{ki})$ is the observation time
- $\mathbf{z}_{ki}(t)$ is the covariate vector for the individual *i* and event *k*
- $\mathbf{z}_i(t)$ is the covariate vector for the individual *i*
- $G_{ki} = X_{ki} - X_{k-1,I}$ is the gap time with $X_{0i} = 0$
- $I(.)$ the variavel that indicates censorship with $I(E) = 1$ when E is true and $I(E) = 0$ when E is false. Then for the individual *i* and event *k*, $\delta_{ki} = I(T_{ki} \leq C_{ki})$.
- $\lambda_{ki}(t)$ is the risk function the the individual *i* and event *k*
- β is the unknown regression parameters vector
- β_k is the regression parameters vector for the event *k*

2.5.2 Andersen-Gill Model (AG)

This model – proposed by Andersen and Gill [35] – is considered the simplest of the three models cited above but it is the one that assumes more strict assumptions [24], for example the independence between times of two events from the same individual.

Although the AG model corresponds to an extension of the Cox Regression model, it is closest in spirit to a Poisson regression by assuming that the events follow a Poisson Process dependent in time.

In this model, it is assumed each individual counting process contains independent increments – with the number of events in nonoverlapping time intervals being independent.

For the individual *i*, *i*=1, …,*n*, the hazard function and the partial likelihood are displayed in Table 5.

Table 5. AG model.

AG Model	
Hazard function	$\lambda_{ki}(t) = \lambda_0(t)e^{\boldsymbol{\beta}'\mathbf{z}_{ki}(t)}, t \geq 0$
Partial likelihood	$L(\beta) = \prod_{i=1}^{n}\prod_{k\geq 1}\left(\frac{e^{\boldsymbol{\beta}'\mathbf{z}_{ki}(x_{ki})}}{\sum_{j=1}^{n}\sum_{l\geq 1} Y_{lj}(x_{ki})e^{\boldsymbol{\beta}'\mathbf{z}_{lj}(x_{ki})}}\right)^{\delta_{ki}}, Y_{lj}(t) = I(x_{l-1,j} < t \leq x_{lj})$

The independence between times means that, for the same individual, the occurrence of one event has nothing to do with the occurrence of past events.

2.5.3 Wei-Lin-Weissfeld Model (WLW)

Although this model is described as treating the events in an ordered way, the WLW model treats the data as unordered competitive risks [24].

In this model, the individual is assumed to be simultaneously at risk for all events and it is at risk for each event until this event occurs. The WLW model [36] estimates the treatment effect using independent models for each event and, therefore, the relationship structure between event times does not need to be known [24].

The WLW model is the only one that allows an individual to be at risk in more than one stratum. If *S* is the maximum events observed for the individual than there are *S* stratum for the *S* events.

The hazard function and the partial likelihood for both models are defined in Table 6.

Table 6. WLW model.

WLW Model	
Hazard function	$\lambda_{si}(t) = \lambda_{s0}(t)e^{\boldsymbol{\beta}'_s \boldsymbol{z}_{si}(t)}, t \geq 0$
Partial likelihood	$L_s(\beta) = \prod_{i=1}^{n} \prod_{s=1}^{S} \left(\frac{e^{\boldsymbol{\beta}' \boldsymbol{z}_{si}(x_{si})}}{\sum_{j=1}^{n} Y_{sj}(x_{si}) e^{\boldsymbol{\beta}' \boldsymbol{z}_{sj}(x_{si})}} \right)^{\delta_{si}}, Y_{sj}(t) = I(x_{sj} \geq t)$

2.5.4 Prentice-Williams-Peterson Model (PWP)

The model proposed by Prentice, Williams and Peterson is also known as the Conditional model because it assumes that the individual is not at risk of suffering the *j* event if the *j*-1 event has not happen yet [37].

There are two possible approaches depending on the processe used:

- PWP counting processes (PWP-CP) based on counting processes - the dataset follows the structure (date of enter the study; date of first event], (date of first event; date of second event], …, (date of j event; date of the last record]
- PWP gap time (PWP-GT) based on time intervals - on this after each event the time goes back to zero.

For instant *t*, $\lambda_{s0}(t)$ is the function for the time since the beginning of the study until *t*. The hazard function and the partial likelihood for both models are defined in Table 7.

Table 7. PWP model.

PWP-CP	
Hazard function	$\lambda_{si}(t) = \lambda_{s0}(t)e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(t)}$
Partial likelihood	$L(\beta)=\prod_{i=1}^{n}\prod_{s\geq 1}\left(\frac{e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(x_{si})}}{\sum_{j=1}^{n} Y_{sj}(x_{si})e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(x_{si})}}\right)^{\delta_{si}}, Y_{sj}(t) = I(x_{s-1,j} < t \leq x_{sj})$
PWP-GP	
Hazard function	$\lambda_{si}(t) = \lambda_{s0}(t - t_{i,s-1})e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(t)}$
Partial likelihood	$L(\beta)=\prod_{i=1}^{n}\prod_{s\geq 1}\left(\frac{e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(x_{(s-1,i)}+g_{si})}}{\sum_{j=1}^{n} Y_{sj}(x_{si})e^{\boldsymbol{\beta}_s' \boldsymbol{z}_{si}(x_{(s-1,i)}+g_{si})}}\right)^{\delta_{si}}, Y_{sj}(t) = I(g_{sj} > t)$

2.5.5 Comparing the models

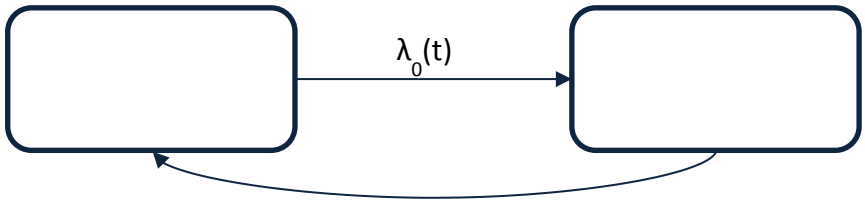
The key to represent any of the models is the structure of the dataset. Assuming a subject with events at times 10, 30 and 42, with no further follow up after the last event. For each model the intervals would be as defined in Table 8.

Table 8. Representation of a subject for the three models.

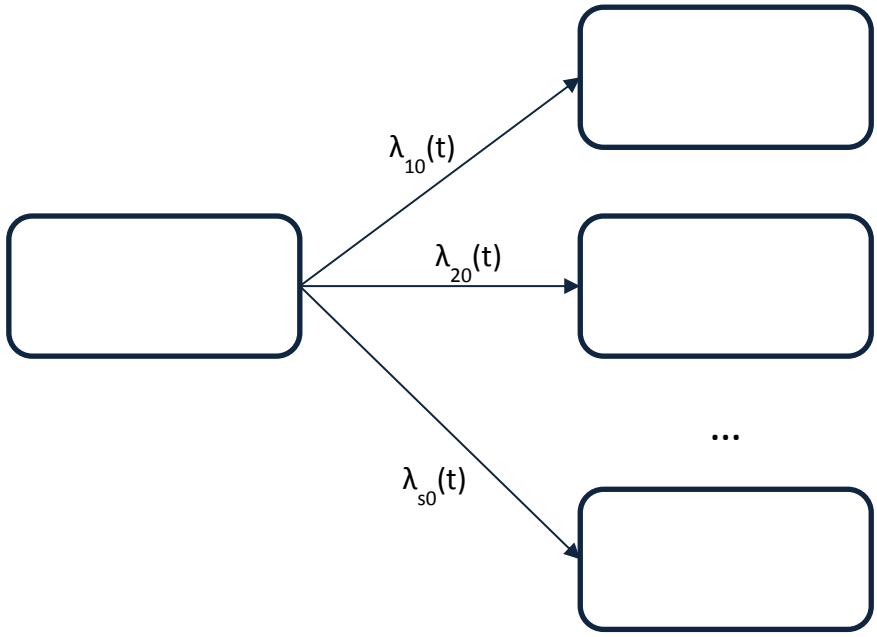
Model	Interval	Stratum
AG model	]0,10]	1
	]10,30[	1
	]30,42[	1
WLW model	]0,10]	1
	]0,30[	2
	]0,42[	3
PWP model	]0,10]	1
	]10,30[	2
	]30,42[	3

As illustrated in Figure 6, a good way of understanding the main differences is to check the schematic form for the three models. Possible transitions are represented by an arrow, with each distinct arrow corresponding to a separate stratum in the Cox Regression model [24]:

AG model



WLW model



PWP model

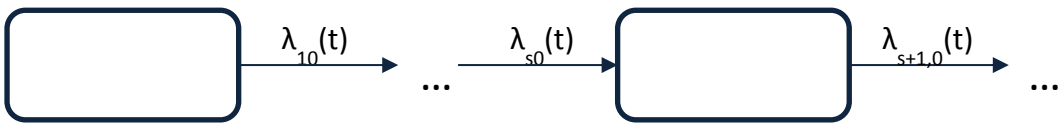


Figure 7. Models – schematic form.

For our project we will use the PWP model because, as it is possible to see in Figure 6, one woman cannot be at risk of the second screening delay if she never had the first screening delay.

Some considerations on the models:

- WLW model was initially proposed as an alternative to the PWP model
- WLW model works better with larger dimensions
- On the PWP model the dimension reduces as the number of events occur
- The WLW model has been criticized because there is no natural order between reoccurring events

To determine which model to use, it is best to take into account the purpose of the study:

- AG model, if the interest is in regards to the recurrence
- WLW model, if the interest is in regards to the time between the beginning of the observation and if the events can occur simultaneously
- PWP model, if the interest is in regards to the time intervals and the risk changes after each event

2.6 Modelling the delay between consecutive mammograms

For this project the hazard function and the partial likelihood follow a specific approach as it is designed for modelling the delay between consecutive mammograms. Table 9 shows the respective hazard function and partial likelihood.

Table 9. Model specific for the project.

Project	
Hazard function	$\lambda_{ij}(t)=\begin{cases}\lambda_{0j}(t-(t_{j-1}+730))e^{\mathbf{z}_{ij}'(t)\boldsymbol{\beta}}, & j\geq 2\\ \lambda_{01}(t-0)e^{\mathbf{z}_{i1}'(t)\boldsymbol{\beta}}, & j=1\end{cases}$
Partial likelihood	$L(\boldsymbol{\beta})=\prod_{i=1}^{n}\prod_{j=1}^{J}\left(\frac{e^{\mathbf{z}_{ij}'(g_{ij})\boldsymbol{\beta}}}{\sum_{k=1}^{n}Y_{kj}(g_{ij})e^{\mathbf{z}_{kj}'(g_{ij})\boldsymbol{\beta}}}\right)^{\delta_{ij}}$

The whole modeling procedure was performed by means of the package *Survival* [38], from the R software, version 3.0.3 [39].

Chapter 3

RESULTS

3. RESULTS

3.1 Descriptive analysis

As referred to in Section 2.2.1, the target population under study is composed of 22,830 women, and the sample data – used for modeling purposes - reflects 1,926 women (8.4% of the target population), with distribution by Health Care Units illustrated within Figure 7.

Table 10 shows a summary of the main socio-demographic and clinical characteristics, for the target population and the sample data. Considering the sample data, it is possible to conclude that 82.4% of the women are European of which 97.9% are Portuguese. It is also possible to verify that most of these women do not drink (93.5%) or smoke (83.9%), %), which is in accordance with the behavior of the target population. However, unemployed women have almost doubled when comparing the sample data to the target population (23.0% and 40.7%, respectively). In regard to the clinical characteristics, the sampled data behaves similarly to the target population, in what concerns to the Menarche and BMI (Table 10). However, there is an increase in the percentage of women who have used hormonal contraception in the sampled data compared to the target population (37.2% and 58.0%, respectively).

Table 10. Target population and sampled data: socio-demographic and clinical characteristics.

Socio-demographic characteristics		Sample		Target Population	
Family care unit		(n = 1,926)		(n = 22,830)	
	Amato Lusitano	57	3.0%	1,729	7.6%
	Cidadela	116	6.0%	2,634	11.5%
	Dafundo	369	19.2%	2,329	10.2%
	FF-Mais	313	16.3%	2,112	9.3%
	Magnólia	269	14.0%	1,923	8.4%
	Marginal	198	10.3%	2,910	12.7%
	Rodrigues Miguéis	63	3.3%	2,735	12.0%
	Tílias	56	2.9%	1,935	8.5%
	Tornada	119	6.2%	1,672	7.3%
	Villa Longa	366	19.0%	2,851	12.5%
Place of birth		(n = 1,926)		(n = 22,830)	
	European	1,587	82.4%	19,504	85.4%
	Portuguese	1,553	97.9%	18,964	97.2%
	Other	34	2.1%	540	2.8%
	African	249	12.9%	2,494	10.9%
	American	78	4.0%	651	2.9%
	Asian	12	0.6%	175	0.8%
	Oceania	0	0.0%	6	0.0%
Education level		(n = 536)		(n = 4,100)	
	Illiterate	21	3.9%	349	8.5%
	Primary school	34	6.4%	439	10.7%
	Middle school	333	62.4%	2,154	52.5%
	High school	109	20.4%	693	16.9%
	University	37	6.9%	465	11.3%
Professional status		(n = 431)		(n = 3,137)	
	Employed	331	77.0%	1,860	59.3%
	Unemployed	99	23.0%	1,277	40.7%
Age when entering the study		(n = 1,926)		(n = 22,830)	
	[50 - 55[	1,615	83.9%	9,675	42.4%
	[55 - 60[	196	10.2%	4,788	21.0%
	[60 - 65[	87	4.5%	4,496	19.7%
	[65 - 70[	28	1.5%	3,871	17.0%
Tobacco consumption		(n = 1,926)		(n = 22,830)	
	Yes	311	16.1%	1,646	7.2%
	No	1,615	83.9%	21,184	92.8%
Alcohol consumption		(n = 1,926)		(n = 22,830)	
	Yes	125	6.5%	915	4.0%
	No	1,801	93.5%	21,915	96.0%
Clinical characteristics					
Menarche		(n = 1,926)		(n = 4,382)	
	< 11	167	8.7%	362	8.3%
	[11 - 16[	1,643	85.3%	3,733	85.2%
	≥ 16	116	6.0%	287	6.5%
Contraception		(n = 1,926)		(n = 2,801)	
	Hormonal	1,118	58.0%	1,041	37.2%
	Non-hormonal	808	42.0%	1,760	62.8%
BMI		(n = 1,926)		(n = 17,399)	
	< 18.5	17	0.9%	164	0.9%
	[18.5 - 25[	525	27.3%	4,148	23.8%
	[25 - 30[	760	39.5%	6,507	37.4%
	≥ 30	624	32.4%	6,580	37.8%

By analysing the descriptive statistics in Table 10 it appears that the behaviour of the sample is in line with the target population. However, to compare the target population against the sample with a goal to study the delay in screening time between two consecutive mammograms, the decision was made to compare adherence rates and non-screening average time. This means that for the adherence rates we compared all women aged 51 versus women of the same age with no screening delay. The same analysis was also undertaken for women aged: 55, 60, 65 and 69. The delayed time represents the average days a screening mammogram was delayed. The results are presented in Table 11.

Table 11. Adherence rates to mammograms screening and delayed time.

		Age				
		51	55	60	65	69
Target Population	Adherence rate	62.3%	42.8%	44.9%	48.4%	62.7%
	Women (n)	2,931	1,881	1,776	1,559	1,407
	Average delayed time (days)	1,151	1,437	1,409	1,414	1,209
Sample	Adherence rate	80.6%	74.2%	71.7%	46.7%	100%
	Women (n)	459	89	53	15	3
	Average delayed time (days)	967	1,116	1,154	1,204	-

To better understand the distribution of the data – with emphasis on the skewness pattern – highlighting variability outside the first and the third quartiles and the candidates to outliers, the box plot diagrams were produced for both target population and sampled data. The lengths of time (in days) of the first, second and third screening delays were analysed. The results are displayed in Figures 8, 9 and 10, respectively. The fourth screening delay, and so on, were not analysed due to the lack of data.

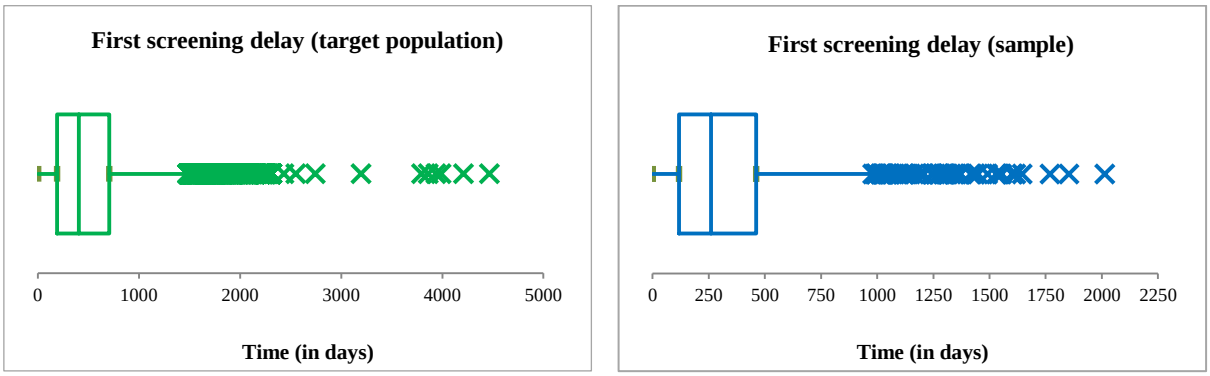


Figure 8. First screening delay for the target population and the sample.

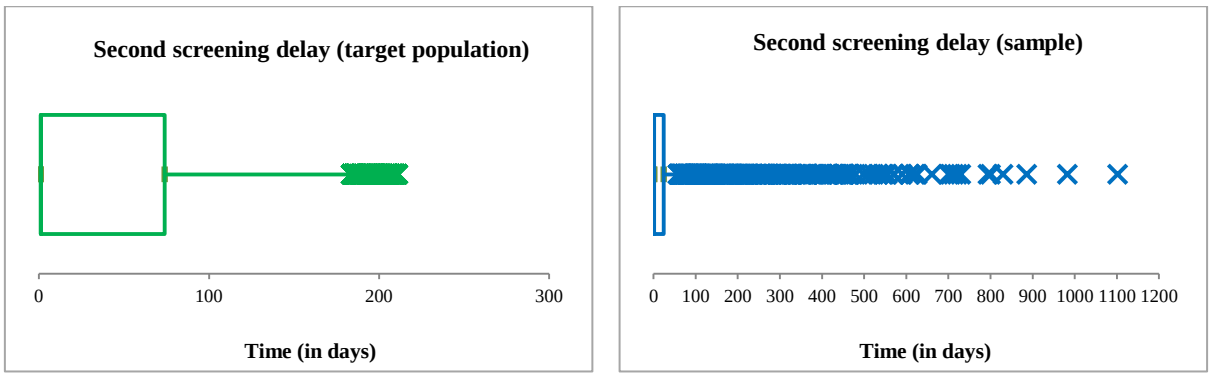


Figure 9. Second screening delay for the target population and the sample.

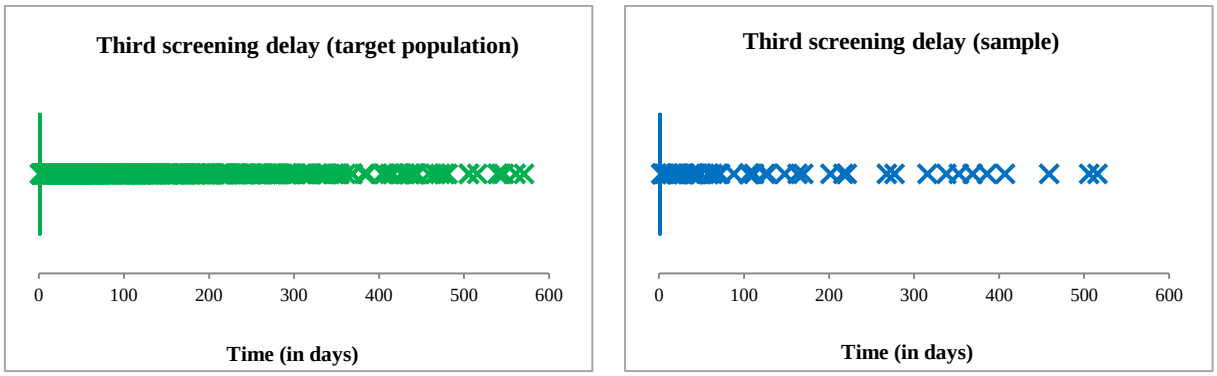


Figure 10. Third screening delay for the target population and the sample.

In line with Table 10 and Table 11 by comparing the box plots in Figures 8-10 it appears that both the target population and the sample have similar behaviour in terms of screening delay. Figure 8 shows that the underlying distributions of the target population and the sample are similar, with positive skewness. Also, there are a large number of outliers at the right side of both box-plots. This is certainly related with the non-Normal distribution of the data sets. Figures 9 and 10 are very difficult to interpret in terms of the shape of the underlying distributions. To better understand these graphics, some statistical measures are displayed in Table 12, namely: number of women in each screening delay; the first, second and third quartiles; minimum and maximum. It is possible to check that the strange patterns of the box-plots for the second and third screening delays - target population and sample – are due to the minimum, first quartile, median and third quartile being all equal to one day. An exception is made to the 3rd quartile of the second screening delay (74 days vs 24 days, respectively for the target population and the sample).

Table 12. Statistic measures, in days, for both target population and sample.

		number of women	1st quartile	Median	3rd quartile	min	max
Target Population	1st screening delay	22,831	188	406	706	1	4,469
	2nd screening delay	17,850	1	1	74	1	1,568
	3rd screening delay	7,057	1	1	1	1	647
Sample	1st screening delay	1,926	118	260	461	1	2,311
	2nd screening delay	1,591	1	1	24	1	1,102
	3rd screening delay	489	1	1	1	1	516

Besides the box plots, it were also produced the histograms for the three screening delays to try to understand its distribution. The analysis was carried out for both the target population and the sampled data. By analysing Figure 11, all three screening delays appear to have an exponential distribution. This empirical result is in accordance with the literature. In fact, the theoretical basis for the Cox model with multiple events per subject relies on the counting process formulation of the Cox model, which assumes the exponential distribution to model the time between consecutive events.

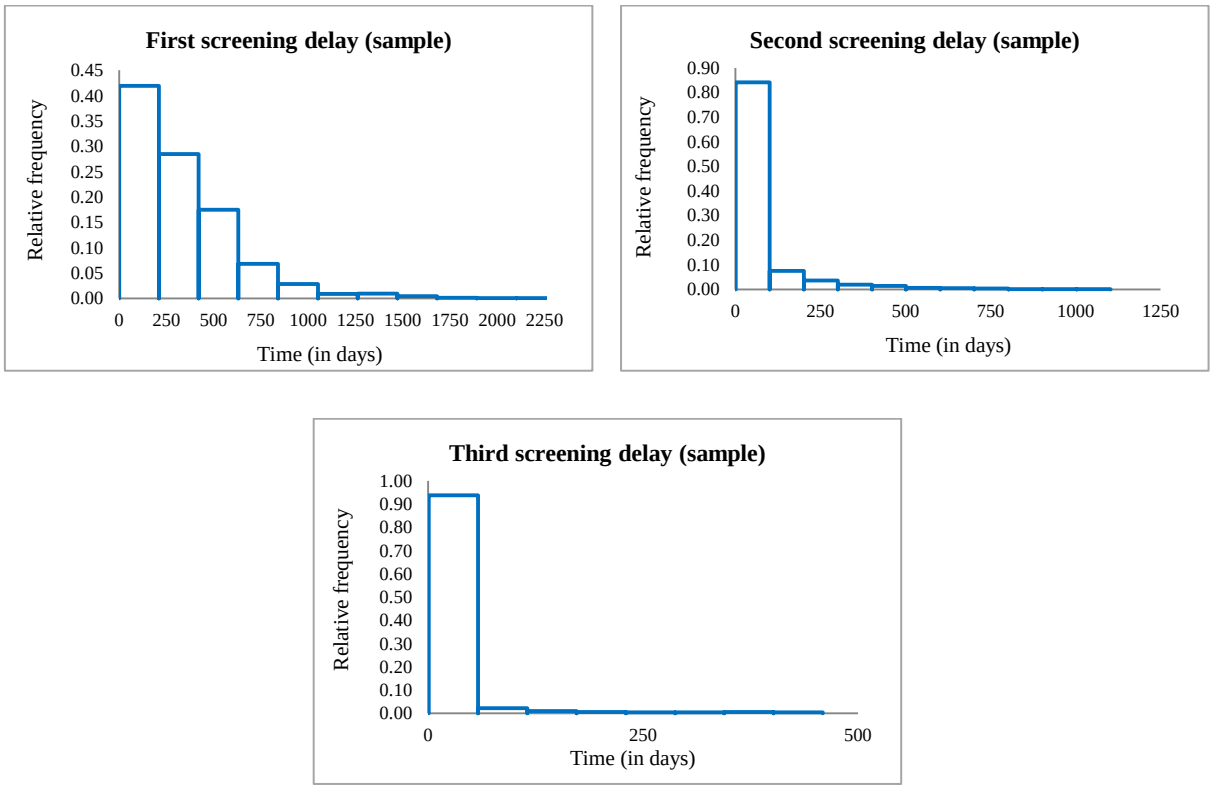


Figure 11. Histograms for first, second and third screening delay.

To try to understand the impact of each covariate in the screening delay, the Kaplan-Meier curves were produced an illustrated in Figure 12.

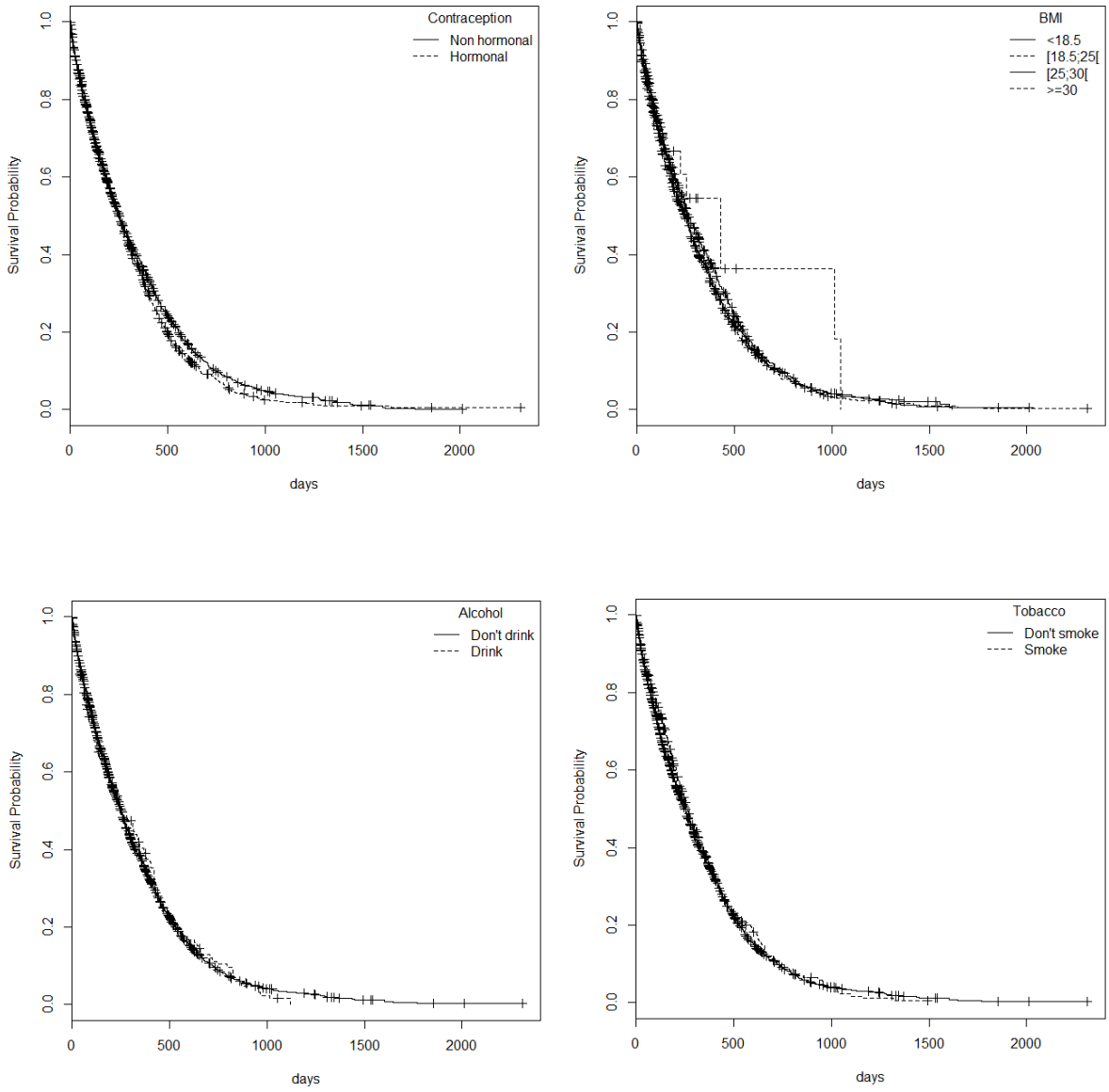


Figure 12. Kaplan-Meier curves for the nominal variables in study.

From Figure 12 it seems reasonable to assume that there is a similar pattern for the covariates contraception, alcohol and tobacco on the screening delay for both groups of women (non hormonal *versus* hormonal contraception, do not drink *versus* drinks and do not smoke *versus* smoke) – the lines for both group of women follows the same distribution. However, for body mass index it seems to be significant differences on the screening delay when the BMI ≥ 30 – for the women in this group the line does not follow the same distribution as the others (women with BMI < 18.5, [18.5;25[, [25;30[).

3.2 PWP Model

By using the package *Survival* [38], from the R software, version 3.0.3 [39], Table 13 includes the results for the univariate regression analysis with the PWP model.

Table 13. Univariate regression estimation of the parameters.

Variable	β	*p* value	HR	CI HR (90%)
Age as at enter the study	0.003	0.523	1.004	(0.994 ; 1.013)
Menarche	0.013	0.281	1.013	(0.993 ; 1.034)
Contraception - non hormonal			ref	
Contraception - hormonal	0.087	0.045	1.091	(1.014 ; 1.173)
BMI				
< 18.5	-0.316	0.279	0.729	(0.451 ; 1.178)
[18.5 ; 25[			ref	
[25 ; 30[	0.063	0.250	1.065	(0.973 ; 1.166)
≥ 30	0.097	0.088	1.101	(1.003 ; 1.209)
Tobacco consumption - no			ref	
Tobacco consumption - yes	-0.043	0.492	0.958	(0.865 ; 1.062)
Alcohol consumption – no			ref	
Alcohol consumption - yes	0.003	0.974	1.003	(0.870 ; 1.155)

*β>0 means a decrease on the time between two consecutive mammograms. This means that in this case β>0 as protective impact.

Analysing Table 13 it is possible to determine that all significant variables have a protective impact on the screening delay. This means that, for example, women with BMI ≥ 30 do screening mammograms 10.1% with less delay when comparing to women with "normal" BMI ([18.5 ; 25[). While women who use hormonal contraceptives have 9.1% decrease on the delay when comparing to women who do not use hormonal contraception.

For our sample the PWP model was the one used, as explained in Figure 6 Section 2.5.5, as a woman cannot be at risk of suffering the second screening delay if she never attended the first. This model was run with stratification by the number of doctor's appointments.

Table 14. PWP model estimation of the parameters.

Variable	β	*p* value	HR	CI HR (90%)
Age as at enter the study	0.013	0.046	1.013	(1.002 ; 1.026)
Menarche	0.024	0.077	1.025	(1.002 ; 1.048)
Contraception - non hormonal			ref	
Contraception - hormonal	0.081	0.081	1.085	(1.005 ; 1.171)
BMI				
< 18.5	-0.362	0.187	0.696	(0.445 ; 1.093)
[18.5 ; 25[				
[25 ; 30[	0.127	0.022	1.135	(1.036 ; 1.244)
≥ 30	0.220	0.000	1.247	(1.130 ; 1.375)
Tobacco consumption - no			ref	
Tobacco consumption - yes	-0.031	0.718	0.970	(0.844 ; 1.115)
Alcohol consumption - no			ref	
Alcohol consumption - yes	-0.004	0.953	0.993	(0.897 ; 1.107)

Analysing Table 14 it is possible to conclude that, with 10% as a significant level, all the significant variables have a protective impact on the screening delay. This means that, the variable "Age as at enter the study" provides a decrease of 1.3% in the delay, as the "Menarche" provides a decrease of 2.5% in the screening delay. Women who uses hormonal contraception have an 8.5% decrease in the delay when comparing with women who do not use. Women with BMI in [25 ; 30[do screening mammograms 13.5% times with less delay when comparing to women with "normal" BMI ([18.5 ; 25[). While women with BMI ≥ 30 do screening mammograms 24.7% with less delay when comparing to women with "normal" BMI ([18.5 ; 25[).

For the hazard ratios, using a confidence interval of 90%, it was produced the forest plot (that is the graphical display of the hazards ratio for each covariate) as illustrated in Figure 13 where it is possible to take the same conclusions as from Table 14.

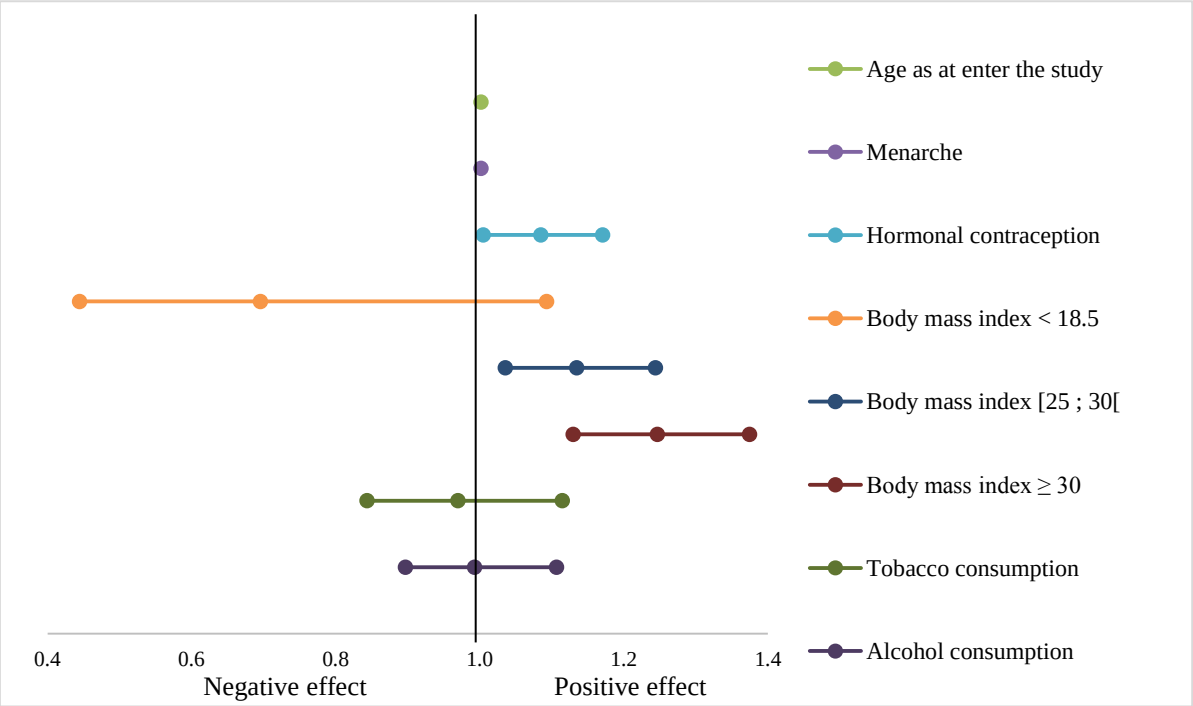


Figure 13. Forest plot for hazard ratios.

The forest plot is a graphic that quickly summarizes the hazard ratio data across multiple variables. If the line crosses the value 1, the hazard ratio is not significant and there is no clear advantage for either the positive or negative arm.

After the PWP model and to assess its adequacy, analysis on residuals was conducted. The Schoenfeld and the martingale residuals were the ones studied in this section.

The proportional hazards assumption was investigated through the Schoenfeld residuals.

By analysing the Schoenfeld residual graphics it is possible to check if the effect of each variable is constant over time. This means, if the proportional hazards are proportional as time goes by.

From Figure 11 to Figure 14 it is possible to analyse the Schoenfeld graphics for each covariate.

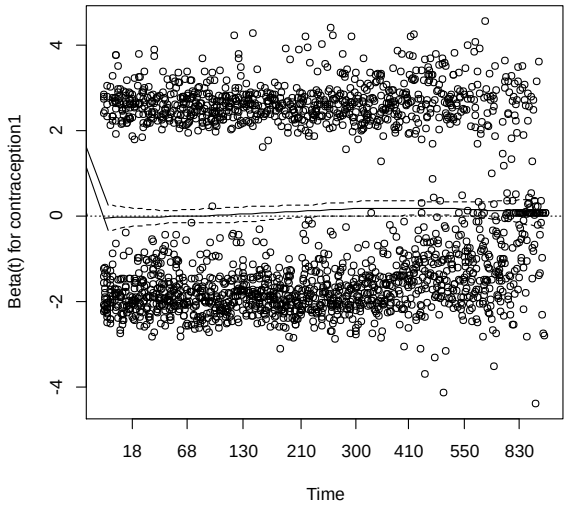


Figure 14. Schoenfeld residuals graphic for Contraception.

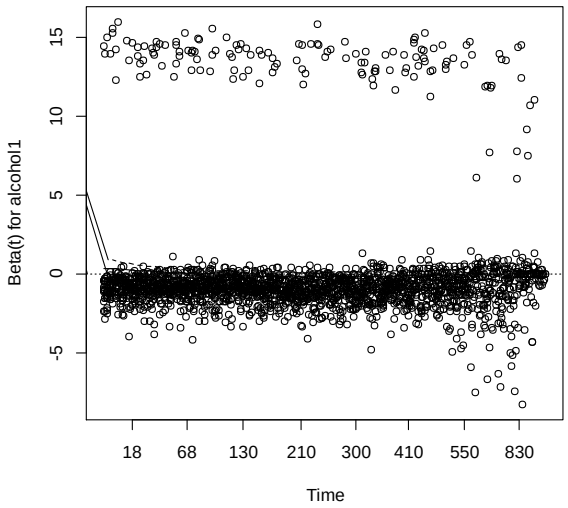


Figure 15. Schoenfeld residuals graphic for Alcohol.

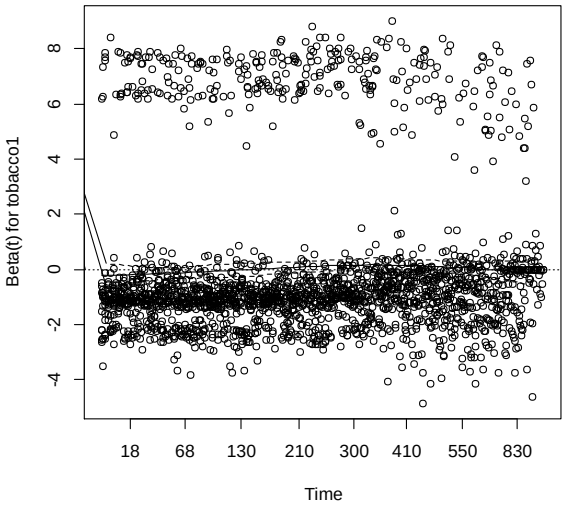


Figure 16. Schoenfeld residuals graphic for Tobacco.

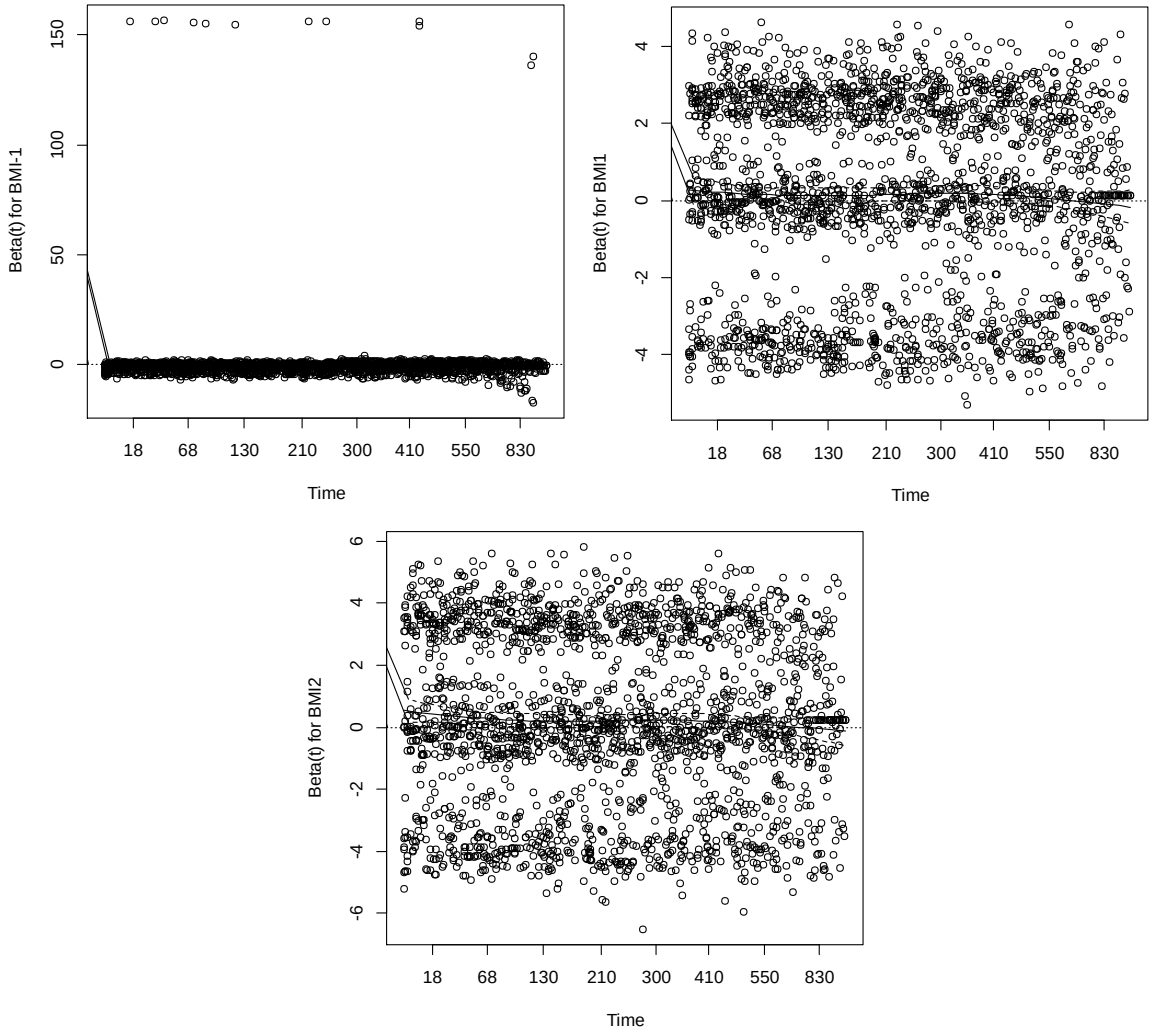


Figure 16. Schoenfeld residuals graphics for each BMI category.

By analysing the Schoenfeld residuals, from Figure 11 to Figure 14, it is possible to conclude that the proportional hazards are constant in time. This means that the effect of each covariate stays constant in time with, apparently, no oscillations.

Besides the graphical analysis on the Schoenfeld residuals, the proportional hazards were also tested.

Table 15. Proportional hazards assumption result.

Variable	Rho	*p* value	chisq
Age as at enter the study	-0.023	0.289	1.126
Menarche	-0.011	0.594	0.284
Contraception - non hormonal		ref	
Contraception - hormonal	0.031	0.151	2.062
BMI			
< 18.5	-0.008	0.720	0.129
[18.5 ; 25[		ref	
[25 ; 30[	-0.023	0.295	1.098
≥ 30	-0.041	0.054	3.712
Tobacco consumption - no		ref	
Tobacco consumption - yes	0.026	0.241	1.375
Alcohol consumption - no		ref	
Alcohol consumption - yes	-0.018	0.438	0.601
Global	N/A	0.241	10.36

Considering the *p*-values for each covariate in Table 14 it is possible to conclude that, using a significance level of 5%, the proportional hazards assumptions should not be rejected – all the *p*-values are bigger than 0.05. Even the *p*-value for the global model is larger than 0.05 (0.241).

The martingale residuals are used to determine the functional form of a covariate and also the outliers.

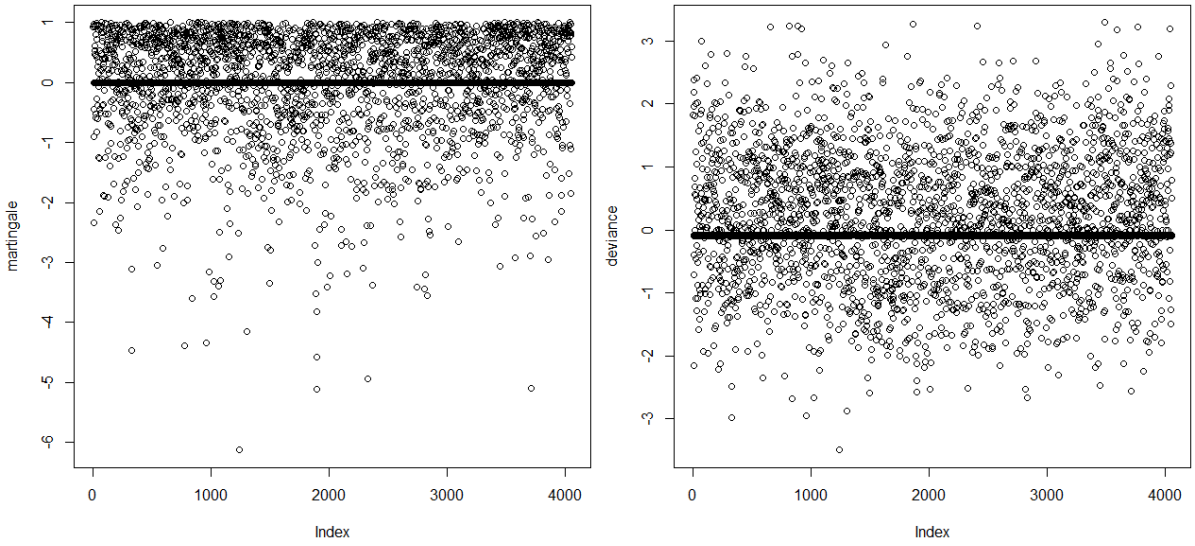


Figure 17. Martingale residuals *versus* index and deviance residuals *versus* index.

The graphic for the martingale residuals *versus* the index for each individual, Figure 15, does not present any pattern which corresponds to a good fit, once the data is evenly distributed above the zero line. This means that the plot seems to be uniformed. It is also possible to conclude that there are a few outliers as there are some individuals on very small values (-4, -5, -6). Regarding the outliers it is possible to say that the ones on the bottom of the graphic are the ones with a higher screening time delay.

Still analysing Figure 18 but now looking at the deviance residuals it is possible to say that its distribution is more symmetric near zero. This means that the information on both graphics is similar.

Chapter 4

DISCUSSION & CONCLUSION

4. DISCUSSION & CONCLUSION

The present work focus on registries since 1st January 2001 to 8th January 2013 from electronic health records data, to study the delay on the screening time between two consecutive mammograms.

The main challenge regarding this data comes from the fact that for electronic health records the data is introduced in a clinical view and not with the purpose of investigation. This means that the variables can have many missing information. For this project there were many variables available like ethnia, age of menopause, BiRads, etc., but it was not possible to include them in the model because of the number of missing observations [23].

Screening may be influenced by external assumptions like agency relationship in health care, when a health professional determines the patient's best interest and acting upon that [23]. In this project, this means that some doctors may insist on breast screening with their patients more than others do. This feature might have an impact on the time interval between mammograms, which can introduce a non-controllable bias in the whole analysis. There is no way to quantify the extent of such bias.

On the other hand, the agency relationship in the primary health care context implies that the main goal of physicians is to deliver and coordinate comprehensive care for patients. Achieving such goal requires availability, a broad spectrum of medical knowledge, effective use of local health care systems, and attention to the "big picture" and the details of a patient's life and health [40]. As a consequence, doctors are more likely to control screening mammograms for women who have a higher risk for breast cancer. This means that the covariates used in this study are not only the ones available in the clinical records but also those that lead to a higher risk for breast cancer. These are: age as at the women enters the study; age that the women had the menarche; type of contraception (hormonal versus non hormonal); body mass index; tobacco and alcohol consumption [3], [16], [17], [18], [19], [20], [21]. Due to the patient-physician relationship, the number of doctor's appointments and the number of previous mammograms were also included in the model.

Since the present project is about to study the delay on the screening time between two consecutive mammograms, there is the need to use mathematical techniques in the context of the Survival Analysis, as the case of the Cox regression model. As each woman can have more than one screening delay, it was necessary to use an extension of the Cox model: the

Prentice-Williams-Peterson (PWP) model. This model was used to describe the impact of the variables in study on the screening time delay between two consecutive mammograms. After a first iteration of the estimation process, both covariates "number of doctor's appointments" and "number of previous mammograms" revealed strong violations of the proportional hazards assumption. As a consequence, the "number of doctor's appointments" was incorporated into the model as a stratification variable. The main pitfall of this strategy has to do with the fact that the impact of this variable on screening delay cannot be measured. The choice of the covariate "number of doctor's appointments" instead of the "number of previous mammograms" is because the former variable is more reliable than the later. Also, the PWP model does not allow for two stratification variables.

Based on the Kaplan-Meier survival curves and on the residual analysis (namely, Schoenfeld and Martingale residuals) the proportional hazards where tested to check the assumption of the constant relationship between the dependent variable and the covariates. The functional relationships between the dependent variable and the covariates were also evaluated. The results showed that the proportional hazards are constant in time. This means that the effect of each covariate remains constant in time. Additionally, the linear relationships between the covariates and the dependent variable were confirmed.

It was possible to conclude that all the significant variables have a protective impact on the screening delay and this means that all the significant variables tend to reduce the delay between two consecutive mammograms.

Specifically, on a significant level of 10%, the variable "Age as at enter the study" provides a decrease of 1.3% in the delay; the "Menarche" provides a decrease of 2.5% in the screening delay. Women who use hormonal contraception have a decrease of 8.5% in the screening delay when comparing with women who do not use. Women with BMI in [25 ; 30[do screening mammograms 13.5% times with less delay when comparing to women with "normal" BMI ([18.5 ; 25[) while, comparing with the same group, women with BMI $\geq$ 30 do screening mammograms 24.7% with less delay.

These results are similar to the work developed by Katapodi in 2004 [41] where he states that as a woman gets older the risk perception gets higher which in this project refers to the fact that as the woman gets older the screening delay is smaller. There are more similar studies regarding the age as for example Bish (2005) [42] and Jones (2011) [43].

In regards to the body mass index, there are also studies that support the same as this one. Kasper (2015) [44] states that obese women can see their gap between two consecutive

mammograms to be smaller because they have a higher risk of developing breast cancer. There are more similar studies regarding the body mass index as Harvey (2004) [45] and Boyd (2005) [46] however there are also studies that have different results, Holm (2015) [47].

Regarding hormonal contraception there is a vast number of works in this area. Kumble (2002) [48] and Beaber (2012) [49] state that the usage of hormonal contraception shows evidence of a higher risk on developing breast cancer. The conclusion is similar in this project where the usage of hormonal contraception has a positive impact on smaller gaps in delays between two consecutive mammograms.

The electronic health records data under study contains at least two more variables that may have a clinical interest on screening delay, namely: number of doctor's appointments, the number of previous mammograms. However, it was not possible to quantify the impact of each of these covariates on the response variable, as stated above. This limitation is related with violation of the proportional hazards by these two covariates. Future research may rely on developing an extension of the PWP model, from a theoretical point a view, to accommodate the violation of the proportional hazards assumption of some covariates of the model. This type of developments will certainly be very relevant in Clinical Biostatistics since there is an increasing need to apply survival analysis techniques to data sets with multiple events per subject

REFERENCES

REFERENCES

[1] Cancer: American cancer society [Internet]. [cited 2nd October 2014]. Available from: http://www.cancer.org/

[2] Liga Portuguesa Contra o Cancro [Internet]. [cited 2nd October 2014]. Available from: http://www.ligacontracancro.pt/

[3] WHO: World Health Organization [Internet]. [cited 2nd October 2014]. Available from: http://www.who.int/en/

[4] IARC (2008). World cancer report 2008. Lyon, International Agency for Research on Cancer

[5] Breastcancer.org: Breast cancer information and awareness [Internet]. [cited 22nd February 2015]. Available from: http://www.breastcancer.org/

[6] Moss M, Nyström L, Jonsson H, Paci E, Lynge E, Njor S. The impact of mammographic screening on breast cancer mortality in Europe: a review of trend studies. J Med Screen 2012; 26-32

[7] Bastos J, Peleteiro B, Gouveia J, Coleman MP, Lunet N. The state of the art of cancer control in 30 European countries in 2008. Int J Cancer. 2010 Jun 1;126(11):2700-15

[8] Rodrigues V. Prevenção secundária do cancro da mama feminina. Em: Oliveira C., coord. Manual de Ginecologia. Lisboa. Permanyer Portugal 2009; 191-201

[9] Ministério da Saúde – Administração Regional de Saúde do Centro, IP. Relatório de Actividades, 2010

[10] LEX: Access to European Union law [Internet]. [cited 2nd December 2014]. Available from: http://eur-lex.europa.eu

[11] Gastrin G, Miller A, To T, Aronson K, Wall C, Hakama M, et al. Incidence and mortality from breast cancer in the Mama Program for Breast Screening in Finland 1973-1986. Cancer 1994; 73:2168-74

[12] Alexander F, Anderson T, Brown H, Forrest A, Hepburn W, Kirkpatrick A, et al. 14 years of follow-up from the Edinburgh randomised trial of breast-cancer screening. Lancet 1999; 353:1903-8

[13] INSA: Instituto Nacional de Saúde [Internet]. [cited 15th November 2014]. Available from: www.insa.pt

[14] Barros H, Bastos J, Lunet N. Evolução da Mortalidade por Cancro da Mama em Portugal (1955-2002). Acta Med Por 2007; 20: 139-144

[15] DGS: Direcção Geral de Saúde [Internet]. [cited 14th January 2015]. Available from: www.dgs.pt

[16] Borrayo E, Hines L, Byers T, Risendal B, Slattery M, Sweeney C, Baumgartner K, & Giuliano A. Characteristics associated with mammography screening among both Hispanic and Non-Hispanic white women. Journal of Women's Health 2009; 18(10): 1585-1594

[17] Decarli A, Vecchia C, Cislaghi C, Mezzanotte G, Marubini E. Descriptive epidemiology of gastric cancer in Italy. 2006

[18] Fair A, Wujcik D, Lin J, Grau A, Wilson V, Champion V, Zheng W, Egan K. Obesity, Gynecological Factors, and Abnormal Mammography Follow-Up in Minority and Medically Underserved Women. J Womens Health 2009; 18(7): 1033-1039

[19] Freitas-Júnior R, Gonzaga C, Freitas N, Matins E, Dardes R. Disparities in female breast cancer mortality rates in Brazil between 1980 and 2009. Clinics 2012; 67(7):731–737

[20] Gail M, Brinton L, Byar D, Corle D, Green S, Schairer C, Mulvihill J. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. J Natl Cancer Inst. 1989; 81(24): 1879-86

[21] Lagerlund M, Sontrop J, Zackrisson S. Psychosocial factors and attendance at a population-based mammography screening program in a cohort of Swedish women. BMC Womens Health 2014; 14-33

[22] Iezzoni LI, Risk Adjustment for Measuring Health Care Outcomes. Fourth Edition. Chicago: Health Administration Press; 2013

[23] Folland, S., Goodman, A. C., & Stano, M. The economics of health and health care (Vol. 6). New Jersey: Pearson Prentice Hall; 2007

[24] Therneau T, Grambsch P. Modelling Survival Data: Extending the Cox Model. New York, United States: Springer, 2000

[25] Ata N, Sözer M. Cox Regression Models with Nonproportional Hazards Applied to Lung Cancer Survival Data. Hacet J Math Stat. 2007; 36(2):157-67

[26] Ahmed F, Vos P, Holbert D. Modeling survival in colon cancer: a methodological review. Molecular Cancer 2007; 6-15

[27] Abadi A, Yavari P, Dehghani-Arani M, Alavi-Majd H, Ghasemi E, Amanpour F, Bajdik C. Cox Models Survival Analysis Based On Breast Cancer Treatments. Iranian Journal of Cancer Prevention 2014; 124:129

[28] Cox D. Regression Models and Life-Tables. Journal of the Royal Statistical Society 1972; Vol.34, No.2, 187-220

[29] Cox D. Partial Likelihood. Biometrika 1975; Vol.62 , No.2, 269-276

[30] Liu X. Parametric Regression Models of Survival Analysis, in Survival Analysis: Models and Applications. Chichester, UK: John Wiley & Sons, Ltd, 2012

[31] Lee E, Wang J. Statistical methods for survival analysis. Fourth Edition. New Jersey: Wiley & Sons; 2013

[32] Barlow W, Prentice R. Residuals for relative risk regression. Biometrika, 75:65-74, 1988

[33] Schoenfeld D. Partial Residuals for the Proportional Hazards Regression Model. Biometrika 1982; Vol.69, No.1, 239-241

[34] Hosmer D, Lemeshow S, May S. Applied Survival Analysis. Third Edition. New Jersey: Wiley & Sons; 2013

[35] Andersen P, Gill R. Cox's Regression Model for Counting Processes: A Large Sample Study. Ann. Statist. 1982; 1100-1120

[36] Wei J, Lin Y, Weissfeld L. Regression Analysis of Multivariate Incomplete Failure Time Data by Modeling Marginal Distribuitions. Journal of the American Statistical Association 1989; Vol.84, No.408, 1065-1073

[37] Prentice R, Williams B, Peterson A. On the regression analysis of multivariate failure time data. Biometrika 1981; 68(2): 373-379

[38] Therneau T. A Package for Survival Analysis in R. R package version.

[39] R Core Team (2014) R: A language and environment for statistical computing

[40] Dugdale D, Epstein R, Pantilat S. Time and the Patient-Physician Relationship, J Gen Intern Med. 1999; 14 (Suppl 1) S34:S40

[41] Katapodi M, Lee K, Facione N, Dodd M. FAANa - Predictors of perceived breast cancer risk and the relation between perceived risk and breast cancer screening: a meta-analytic review. Preventive Medicine 2004; 38: 388 – 402

[45] Bish A, Ramirez A, Burgess C, Hunter M. Understanding why women delay in seeking help for breast cancer symptoms. Journal of Psychosomatic Research 2005; 58: 321 – 326

[43] Jones S, Magee C, Barrie L, Iverson D, Gregory P, Hanks E, Nelson A, Nehill C, Zorbas H. Australian Women's Perceptions of Breast Cancer Risk Factors and the Risk of Developing Breast Cancer. Women's Health Issues 2011; 21-5: 353–360

[44] Braunwald F, Kasper H, Longo J. Harrison Principles of Internal Medicine: Volumes I and II. 19th Edition. Mc Graw Hill; 2015

[45] Harvey J, Bovbjerg V. Quantitave Assessment of Mammographic Breast Density: Relationship with Breast Cancer Risk. RSNA Radiology 2004; 230: 1

[46] Boyd N, Rommens J, Vogt K, Lee V, Hopper J, Yaffe M, Paterson A. Mammographic breast density as an intermediate phenotype for breast cancer. The Lancet Oncology 2005; 10: 798-808

[47] Holm J, Humphreys K, Li J, Ploner A, Cheddad A, Eriksson M, Törnberg S, Hall P, Czene K. Risk Factors and Tumor Characteristics of Interval Cancer by Mammographic Density. Journal of Clinical Oncology 2015; 10.1200/JCO.2014.58.9986

[48] Kumle M, Weiderpass E, Braaten T, Persson I, Adami H, Lund E. Use of Oral Contraceptives and Breast Cancer Risk - The Norwegian-Swedish Women's Lifestyle and Health Cohort Study. Cancer Epidemiology Biomarkers November 2002; 11: 1375

[49] Beaber, E. Use of oral contraceptives (OC) and breast cancer risk among young women by OC formulation and breast cancer subtype. University of Washington 2012

[50] Therneau T, Grambsch P. Extending the Cox Model – Technical report Number 58. Mayo Clinic, Rochester, Minnesota, 1996

[51] Carvalho M, Andreozzi V, Codeço C, Campos D, Barbosa M, Shimakura S. Análise de Sobrevivência – Teoria e aplicações em Saúde. 2011; 2ªed, Rio de Janeiro, Editora Fiocruz

[52] Kelly P, Lim L. Survival analysis for recurrent event data: an application to childhood infectious diseases. Stats Med 2000; 19(1): 13-33

[53] Marcus A, Crane L, Kaplan C, Reading A, Savage E, Gunning J, Bernstein G, Berek J. Improving adherence to screening follow-up among women with abnormal Pap smears: results from a large clinic-based trial of three intervention strategies. Med Care 1992; 30(3): 216-30

[54] Harris R, Yeatts J, Kinsinger L. Breast cancer screening for women ages 50 to 69 years a systematic review of observational evidence. Prev Med. 2011; 53(3): 108-14

[55] Wei L, Glidden D. An overview of statistical methods for multiple failure time data in clinical trials. Stat Med 1997; 16: 833-839

[56] Ghosh D. Methods for analysis of multiple events in the presence of death. Control Clin Trials 2000; 21: 115-126

[57] Lim H, Liu J, Melzer-Lange M. Comparison of methods for analyzing recurrent events data: application to the Emergency Department Visits of Pediatric Firearm Victims. Accid Anal Prev 2007; 39: 290-299

[58] Metcalfe C, Thompson S. Wei, Lin and Weissfeld's marginal analysis of multivariate failure time data: should it be applied to a recurrent events outcome?. Stat Methods Med Res 2007; 16: 103-122