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

Insights into tuberculosis: a survival analysis of time to recurrence

Patricia Soares

Projecto
Mestrado em Bioestatística

2014

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

Insights into tuberculosis: a survival analysis of time to recurrence

Patricia Soares

Projecto
Mestrado em Bioestatística

Projecto orientado por Professora Doutora Cristina Rocha and Doutor João Lopes

2014

Acknowledgement

My sincere gratitude to Professor Cristina and Dr. João, supervisors of this project, for all the availability and for interrupting your holidays to read this thesis carefully and rigorously. Your comments have enormously contributed to enrich this thesis. Thank you for all the patience and support to overcome all the doubts that have arisen. I learned a lot in this process and all of our meetings were extremely enlightening. It was a great pleasure to do this thesis under your supervision. I also take this opportunity to thank Professor Valeska who enlightened me about the utility of survival analysis and all the initial support, including the development of various ideas for the realization of this thesis. A thank you to all the teachers of this Masters that have always been available. A special thank you to Professor Lisete for the constant support and tireless in solving problems.

A special thank you to Dr. Jose Pereira-Leal, Dr. Gabriela Gomes and Dr. Carlos Penha-Gonçalves to give me the opportunity to work under the FCT project PFE-EA-FCT-PTDC/SAU-ESA/71208/06. Besides providing the data used in this thesis, the time spent at Instituto Gulbenkian de Ciência provide me valuable skills for this thesis and for my future career. Thank you to all the past and current members of the Bioinformatics Unit, Computational Biology Group and Collective Dynamics group, for all the brainstorming and constant challenge. I left Instituto Gulbenkian de Ciência with more scientific knowledge and with some amazing friends. I truly appreciate your constant support. It was an amazing place to be and learn and the time spent at the Institute will always have a special place in my heart.

This journey was easier thanks to my amazing classmates. We had a fantastic group and it was a pleasure to share this experience with all of you. To my close friends, for all the patience, support and for cheering me up when needed.

Last but definitely not least, I want to thank my parents, brother and boyfriend for having faith in me and encouraging me in every decision I made in life, always believing in the choices that led me to this place. Thank you for all the love and never-ending support even when I was not so easy.

Abtract

Recurrence in tuberculosis is a serious health problem. To understand recurrence it is necessary to understand the risk factors associated and how that information can impact future strategies.

To achieve this purpose, data from the SVIG-TB database was analyzed. Only patients diagnosed with their first episode of TB between 2002 and 2009, in Portugal, were included. The BCG vaccination was one of the most interesting variables to include in the analysis due to the long-lasting discussion about the protective effect of the vaccine. However, using this variable lead to a problem with the amount of missing data in the dataset.

Therefore, to avoid discarding the variable Vaccine and lose the information regarding patients with missing information several techniques of multiple imputation were used: Predictive Mean Matching, two different models of Random Forest and a model of Expectation-Maximization with Bootstrapping. To compare the results obtained, models were fitted to the complete dataset, the complete dataset without the Vaccine and a dataset imputed by mean imputation.

The model fitted to the complete dataset without the Vaccine discarded the variable HIV and included the variable regarding Residence Community presenting the most distinct results compared with the other models. Mean imputation, imputation via maximum likelihood and Random Forest **missForest** selected the same variables (Vaccine, Clinical Form, Situation, Alcohol, Prison, HIV, Diabetes and age). Random Forest **missForest** and imputation via maximum likelihood presented the most consistent results. These results suggests that the majority of the recurrence cases may be due to relapse since extrapulmonary TB, younger age and HIV are associated with relapse.

To conclude, inclusion of information about treatment noncompletion, drug resistance and genotyping data (to distinguish between relapse and reinfection) is essencial. Imputation should be implemented in case of missing information since it carries less assumptions that performing a complete case analysis.

Keywords: Tuberculosis; Recurrence; Multiple Imputation; Survival Analysis

Resumo

A recorrência em tuberculose, seja devido a recaída ou a reinfecção, é um grave problema de saúde pública. A proporção de pacientes multi resistentes ou extensivamente resistentes aos antibióticos é maior entre os casos recorrentes. Algumas medidas precisam de ser implementadas de maneira a reduzir a frequência de casos recorrentes. Para compreender a recorrência é necessário compreender os factores de risco associados, o seu papel e como essa informação pode afectar estratégias futuras. O objectivo desta tese é analisar casos recorrentes de tuberculose de maneira a identificar covariáveis que influenciam o tempo desde o fim do primeiro episódio até ao início do segundo episódio.

Para atingir este propósito, dados do SVIG-TB foram analisados. Apenas pacientes diagnosticados com o seu primeiro episódio entre 2002 e 2009, em Portugal, foram incluídos. A realização da vacina BCG era uma das variáveis mais interessantes para incluir devido a perdurável discussão sobre o efeito protector da vacina. No entanto, usar esta variável leva a um problema de dados omissos na base de dados. Os dados omissos variam entre um baixo valor de 0.03% para a Forma Clínica até um elevado valor de 59% para a Vacina. Uma possível explicação para a diferente quantidade de dados omissos poderá ser que cada centro de saúde pergunta questões ao paciente, inserindo essa informação no SVIG-TB. Provavelmente, alguns centros de saúde dão mais importância a algumas variáveis enquanto outros centros ignoram essas variáveis.

Portanto, para evitar descartar a variável Vacina e perder a informação referente aos pacientes com dados omissos várias técnicas de imputação múltipla foram usadas. Um modelo foi ajustado aos dados completos de forma a comparar os resultados com os resultados obtidos por modelos ajustados a base de dados com os valores imputados. Um modelo foi também ajustado aos dados completos sem a variável Vacina de modo a compreender as implicações de descartar uma variável com uma grande proporção de dados omissos. Uma base de dados "completa"foi obtida através de imputação por substituição da média de modo a comparar os resultados de imputação simples com imputação múltipla. Vários métodos de imputação múltipla foram usados: *Predictive Mean Matching*, dois modelos diferentes de *Random Forest* e um modelo de *Expectation-Maximization with Bootstrapping*. *Predictive Mean Matching* apresenta, no geral, resultados consistentes na literatura. Investigação recente apresenta vantagens em imputar dados com *Random Forest* e *Expectation-Maximization with Bootstrapping* imputa dados via a máxima verosimilhança.

Esta análise alerta para o perigo de descartar uma variável para efectuar uma análise de dados completos. De facto, este modelo rejeitou a variável HIV e incluíu a variável relacionada com a Residência Communitária apresentando os resultados mais distintos comparado com os restantes modelos. A variável Álcool e Diabetes não foi significativa no modelo ajustado aos dados completos, sendo a última excluída devido a um problema de separação completa. Imputação por substituição da média, imputação via máxima verosimilhança e *Random Forest* **missForest** seleccionaram as mesmas variáveis (Vacina, Forma Clínica, Situação, Álcool, Prisão, HIV, Diabetes e idade). O modelo ajustado aos dados imputado por *Predictive Mean Matching* apresentou resultados semelhantes para as estimativas comparando com os restantes modelos; no entanto, não incluíu Prisão e Vacina. As duas técnicas implementadas de *Random Forest* tiveram resultados semelhantes mas o package **mice** é extremamente lento. *Random Forest* **missForest** e imputação via máxima verosimilhança apresentaram os resultados mais consistentes.

Como esperado, as estimativas dos coefficientes e os erros padrão do modelo ajustado aos dados completos são maiores que as estimativas dos modelos ajustados aos dados imputados. Os valores de R^2 and C são mais elevados que nos modelos ajustados aos dados imputados no entanto estes valores não podem ser correctamente comparados uma vez que os modelos são baseados em conjuntos de dados diferentes. As estimativas dos coefficientes e erros padrão, entre imputação simples e múltipla imputação, são muito semelhantes, excepto para a variável Prisão que possui um ligeiro aumento da estimativa do coefficiente. De realçar que imputação simples não introduz variabilidade no modelo, ignorando que os valores não são todos verdadeiros.

Ambos os modelos ajustados aos dados imputados por *Expectation-Maximization with Bootstrapping* e por *Random Forest* **missForest** produzem resultados adequados. No entanto, não é possível seleccionar o "melhor"método de imputação. Cada base de dados deve ser tratada independentemente. Uma série de escolhas (como o número de imputações, o número de iteracções, o método ou métodos para imputar os dados, como incorporar interacções e não-linearidades, etc) devem ser consideradas e ser tratadas com cuidado uma vez que escolhas erradas levam a estimativas incorrectas.

Muitos estudos debatem sobre a eficácia da vaccina BCG embora a maior parte apenas ignore este problema e não adiciona informação sobre a vaccinação no modelo. A vaccina foi incluída neste estudo e verificou-se significativa em alguns modelos. O efeito estimado para a variável BCG varia de 56% (modelo ajustado aos dados imputados por *Expectation-Maximization with Bootstrapping*) até 80% (análise de casos completos). Este resultado indica que pessoas não vacinadas tem um risco maior de sofrer um novo episódio de tuberculose comparado com pessoas vacinadas. O risco de um episódio recorrente para um indivíduo com uma forma extrapulmonar é entre 1.86 vezes (modelo ajustado aos dados imputados por *Expectation-Maximization with Bootstrapping*) até 2.59 vezes (modelo ajustado aos dados completos) o de um indivíduo com uma forma pulmonar de tuberculose. O risco de recorrência para indivíduos que desistiram do tratamento é entre 3 vezes (modelo ajustado aos dados imputados por *Expectation-Maximization with Bootstrapping*) até 9 vezes (valor semelhante para os restantes modelos) o de indivíduos que completaram o tratamento. Um indivíduo alcoólico tem um aumento entre 70% a 80% no risco de recorrência que um indivíduo sem problemas alcoólicos. As estimativas do coefficiente para a variável Prisão são mais diversas. O risco de recorrência para alguém preso é entre 3 vezes (modelo ajustado aos dados imputados por

Expectation-Maximization with Bootstrapping) até 10 vezes (análise de dados completos) o de alguém que não está na prisão. O risco de um episódio recorrente para um indivíduo com HIV é entre 1.93 vezes (modelo ajustado aos dados imputados por *Expectation-Maximization with Bootstrapping*) até 3.29 vezes (modelo ajustado aos dados completos) o de um indivíduo sem HIV. Os resultados obtidos para as estimativas do coefficiente da variável Diabetes foram surpreendentes. A análise mostrou que indivíduos com Diabetes tem uma diminuição de 80% no risco comparado com indivíduos sem Diabetes. Em Portugal, a taxa de indivíduos não diagnosticados é de 43%, o que pode levar a uma subestimação do verdadeiro efeito da variável Diabetes. Um aumento de um ano de idade leva a uma diminuição de 1% a 2% no risco de recorrência.

No entanto, é provável que algumas destas variáveis estejam correlacionadas com variáveis que não foram medidas, em particular, com a aderência ao tratamento e o estado da doença ao iniciar o tratamento. Estes resultados sugerem que a maioria dos casos de recorrência podem ser devido a uma recaída uma vez que tuberculose extrapulmonar, idade jovem e HIV são factores associados a recaída.

Para concluir, estudos adicionais são necessários para confirmar estes resultados. Incluir informação acerca da aderência ao tratamento, da resistência aos antibióticos e dados de genotipagem (para distinguir entre recaída e reinfecção) é essencial. Em casos de dados omissos deve ser realizada uma imputação de dados uma vez que possui menos suposições que realizar uma análise de casos completos. Uma análise exaustiva deve ser realizada de modo a avaliar o método mais apropriado para a imputação.

Keywords: Tuberculose; Recorrência; Múltipla Imputação; Análise de Sobrevivência

Contents

List of Figures

List of Tables

Acronyms and Abbreviations

TB Tuberculosis

MTBC Mycobacterium tuberculosis Complex

Mtb Mycobacterium tuberculosis

NTM Nontuberculous Mycobacteria

HIV Human Immunodeficiency Virus

MDR Multi Drug Resistant

BCG Bacille Calmette-Guerin

BCE Before the Common/Current/Christian Era

WHO World Health Organization

EU European Union

DGS Direcção Geral de Saúde

SVIG-TB Sistema de Viligância Intrínseco ao Programa Nacional de Luta Contra a Tuberculose

CAT Centro Atendimento Toxicodependente

CAGE Cut Down, Annoyed, Guilty and Eye Opener (alcohol use disorders screening test)

MCAR Missing Completely At Random

MAR Missing At Random

MNAR Missing Not At Random

IQR Interquartile Range

CC Complete Case

SE Standard Error

ML Maximum Likelihood

EM Expectation-Maximization

EMB Expectation-Maximization with Bootstrapping

FCS Fully Conditional Specification

PMM Predictive Mean Matching

RF Random Forest

OOB error Out-of-bag error

CI Confidence Interval

KM Kaplan-Meier

LOWESS Locally Weighted Scatterplot Smoother

VIF Variance Inflation Factor

Chapter 1

Introduction

1.1 Some facts about tuberculosis

Human tuberculosis (TB) is an airborne infectious disease that may affect the lungs (pulmonary TB) or other parts of the body (extrapulmonary TB). The most commom form of TB is pulmonary but both forms can also co-exist. Tuberculosis can have a wide range of symptoms such as cough, chest pain, shortness of breath, fatigue, fever or weight loss. It is transmitted when people infected with pulmonary TB cough or sneeze.

The causative agents of TB are grouped in the *Mycobacterium tuberculosis complex* (MTBC): *Mycobacterium tuberculosis* (Mtb), *Mycobacterium bovis*, *Mycobacterium africanum*, *Mycobacterium canneti* and *Mycobacterium microti* (1). Mtb is responsible for most cases of TB and although it can affect animals, humans are the main hosts. *Mycobacterium bovis* was also a public health concern in the early twentieth century due to infection from the consumption of unpasteurized or unboiled milk, but with the introduction of pasteurized milk the number of cases infected with this species has decreased significantly (1; 2). *Mycobacterium africanum* represents more than half of the cases in Africa, being more common in HIV infected patients; however, it is not widespread (3; 4). *Mycobacterium canetti* is rare and seems to be limited to Africa (5). *Mycobacterium microti* natural hosts are small rodents and it is uncommon in humans, the few cases found were immunodeficient people. However, several studies claim that the prevalence may be underestimated (6; 7). Regarding other mycobacteria, *Mycobacterium leprae* and *Mycobacterium lepromatosis* are the pathogens that cause leprosy, while *Mycobacterium avium* and *Mycobacterium kansasii* are Nontuberculous mycobacteria (NTM) which cause pulmonary diseases that resemble TB. Although, *Mycobacterium avium* and *Mycobacterium kansasii* are not contagious, their prevalence is increasing (8).

Most tuberculosis infections are asymptomatic and latent. However, one in ten latent infections eventually progresses to active disease which, if left untreated, kills approximately 66% of infected people. An untreated person with active disease can infect 10 to 15 people per year, but if treated and diagnosed promptly this rate lowers considerably (9). If left untreated, tuberculosis can last for years and become a chronic disease. The progression towards an active disease seems to be related to the immune system. A depressed immune response at

the time of infection increases the risk for active disease, and, in the same way, for someone already infected the risk for reactivation increases when his immunity is low (1; 10). A detailed explanation of the infection process can be seen in figure 1.1.

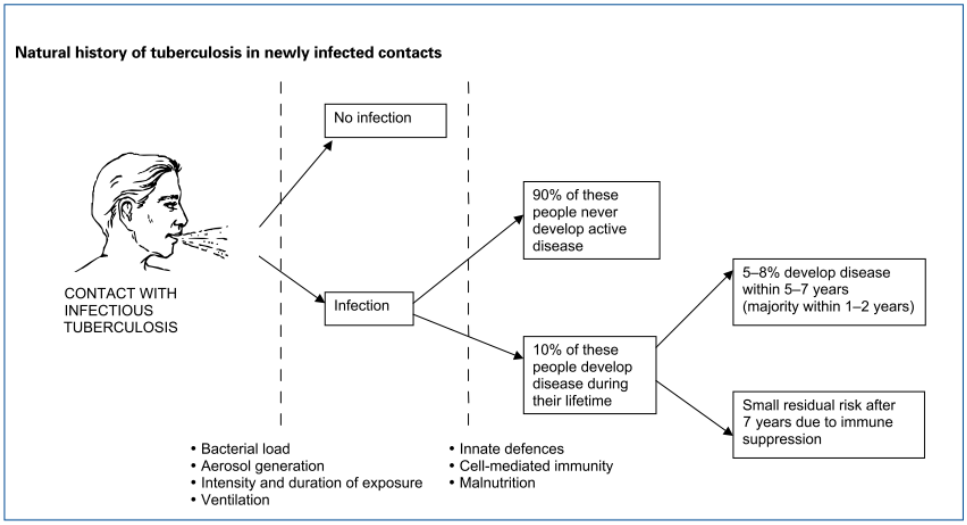

Figure 1.1: Natural history of tuberculosis in newly infected contacts (10)

There are several known risk factors for TB infection. Nowadays, the most important risk factor is HIV - instead of the usual 10% progress rate, 30% of those co-infected with HIV develop active disease. Other important risk factors are overcrowding, malnutrition, addiction to drugs or alcohol, smoking, high-risk ethnics, being children in contact with high-risk patients or health-care workers (1; 11; 12). In fact, individuals with prolonged or close contact with people with TB are at high risk of infection. (9)

At the moment, TB is typically diagnosed using chest X-ray and sputum cultures. Molecular tests are also used to diagnose TB, particularly, for multi drug resistant TB (MDR). The only vaccine available today is bacillus Calmette-Guerin (BCG), which although effective against disease in children, it confers inconsistent protection against pulmonary TB in adults. Nevertheless, new drugs and vaccines are being tested in clinical trials (12). New cases of TB are typically treated following a 6 month regimen of four first line drugs: isoniazid, rifampicin, ethambutol and pyrazinamide. In case of MDR TB (resistance to at least isoniazid and rifampicin) the treatment can last up to 20 months and requires the use of second line drugs such as kanamycin, capreomycin and amikacin (12).

1.2 Tuberculosis history

More than a decade ago, researchers believed that *Mycobacterium bovis* was the ancestor of Mtb. This hypothesis was based on the fact that Mtb is a human pathogen, whereas *Mycobacterium bovis* can infect both animals and humans. However, this theory has been disproved after the complete sequencing of their genomes. The results indicate that *Mycobac-*

terium bovis has a smaller genome due to numerous deletions [1] when compared with Mtb, thus it is likely that *Mycobacterium bovis* evolved from Mtb and not the other way around. The transition from humans to animals may be linked to animal domestication that occurred 13 000 years ago (13; 14).

"Phthisis", a term used in pulmonary TB that means consumption, appeared first in Greek literature. Hippocrates identified phthisis around 460 BCE, giving an accurate description of the characteristics of the disease: fever, colorless urine, cough, and loss of thirst and appetite. He also noted that it was almost always fatal (15; 16).

In the 18th-19th centuries, TB imposed a high burden to society because of the life conditions in cities due to the Industrial Revolution [17]. In Bristol, between 1790 and 1796, 683 out of 1571 deaths were due to TB. In Shropshire, between 1750 and 1759, the rate of death was one in six, ten years later, the rate of death was one in three.

In 1720, the English physician Benjamin Marten proposed in his publication *A New Theory of Consumption* that TB, known at the time as consumption, could be caused by "creatures" that could cause lesions and symptoms of the disease. Marten's writings displayed a good understanding of the disease for the time. However, his work was completely discarded and it took more than one century before Robert Koch demonstrate it to be true (16).

A description of tuberculosis meningitis was given by Robert Whytt, in 1768. And Percivall Pott correctly described the vertebral lesions of tuberculosis vertebral, also known as Pott's disease, in 1779. Around the same time, William Stark suggested that different forms of TB were in fact different manifestations of the disease. Unfortunately, Stark's observations were ignored after his death while studying scurvy.

In his 1819 book, *D'Auscultation Mediate*, Laennec, who is best remembered for his invention of the stethoscope, clearly elucidated the pathogenesis of TB and described most of the physical signs of pulmonary disease. Indeed, modern understanding of TB began with Laennec's treatise.

In 1854, Hermann Brehmer presented his doctoral thesis *Tuberculosis is a Curable Disease*, after having returned cured of TB from a trip to the Himalayan Mountains. At the same time, he built an institution in Gorbersdorf, Germany, where among trees and with good nutrition, patients were exposed to a healthier climate. This led to the introduction of sanatorium which was a significant mark in the history of TB (16).

In 1869, the French military doctor Jean Antoine Villemin demonstrated that the disease was contagious, after conducting an experiment in which tuberculous from human cadavers were injected into laboratory rabbits, who then became infected (16).

When Robert Koch incubated the bacteria, a breakthrough in tuberculosis happened. He named tuberculosis bacillus, after some inoculated laboratory rabbits died with symptoms of tuberculosis. This was the prove that the bacilus was the cause of tuberculosis. He made his result public at the Physiological Society of Berlin on 24 March 1882, in a famous lecture entitled *Über Tuberculose*, where he demonstrated that Mycobacterium was the single cause of tuberculosis in all of its forms. Since 1882, 24 March has been known as World Tuberculosis Day.

[1] Deletion is the loss of genetic material.

Koch's contributions to bacteriology were remarkable and he was awarded the Noble Prize in Medicine or Physiology in 1905 for his elucidation of the etiology of tuberculosis.

Another mark in the history of TB was the discovery of the X-ray, by Wilhelm Roentgen in 1895. This invention led to better diagnosis and, consequently, to a decline in TB incidence and mortality (15).

A more direct approach to tackle the public health challenges of TB was taken by Albert Calmette and his associate Camille Guérin. In the Pasteur Institute of Lille, where Calmette was the director, the researchers tried to use *Mycobacterium bovis* as a vaccine. As a result of their work, BCG was tested for the first time in 1921. The recipient was an infant born to a mother dying of pulmonary tuberculosis and placed in the care of a tuberculous grandmother. The child survived and did not develop tuberculosis.

In 1948, a campaign to control tuberculosis with the sponsorship of the UNICEF and the Danish Red Cross was undertaken. It started in Poland and spread to other European countries and ultimately to Ecuador. During the campaign, nearly 14 million people were vaccinated with BCG. This was the first disease control program undertaken by an agency of the World Health Organization (15).

After the introduction of chemotherapy the history of tuberculosis changed. In 1944, in the USA, Albert Schatz, Elizabeth Bugie, and Selman Waksman isolated streptomycin, the first antibiotic effective against Mtb. Since the amount of streptomycin available from the USA was limited, it was ethically acceptable for the control group to be treated without the drug and therefore the first randomized control trial was conducted, in 1946, by the United Kingdom Medical Research Council (17; 18). Isoniazid, the first oral mycobactericidal drug, was developed in 1952, followed by rifampicin in 1957 (15).

Despite many efforts to fight TB, hopes of erradicating the disease were dashed after the appearance of drug-resistant strains in the 1980s. The resurgence of tuberculosis resulted in the declaration of a global health emergency by the World Health Organization in 1993. Along with HIV and malaria, TB has been declared a global enemy (12; 19).

1.3 Tuberculosis worldwide

TB causes ill-health among millions of people each year and it ranks as the second leading cause of death from an infectious disease worldwide, after HIV. A global overview of this epidemic disease can be seen in figure 1.2.

In 2011, 8.7 million people were diagnosed with TB as new cases, 13% being co-infected with HIV. In that year, 1.4 million people died from TB, among them 430 000 were HIV-positive. Almost 80% of TB cases co-infected with HIV come from Africa. TB is also the third cause of death among women between 15-44 years in low-income countries and the fifth worldwide among women between 20-59 years. It is estimated that 2 billion people have latent TB and half million children are infected worldwide (12).

Figure 1.3 represents the proportion of MDR cases. It is estimated that 3.7% of new cases and 20% of previously treated cases are MDR. Country-wise, India, China, the Russian Federation

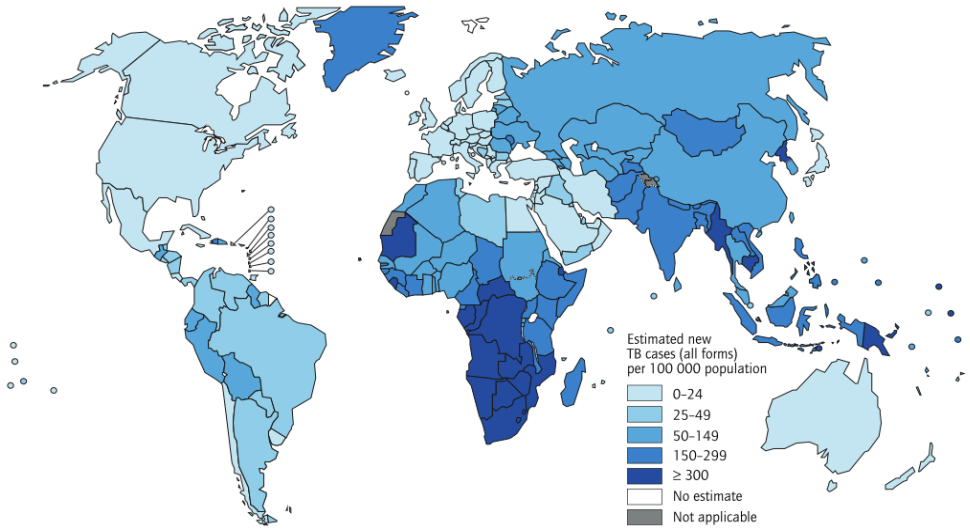

Figure 1.2: Estimated TB incidence rates, 2011 (12)

and South Africa are clearly hotspots for MDR, representing almost 60% of the world's cases of MDR. This figure is a clear representation of the importance of identifying recurrent cases.

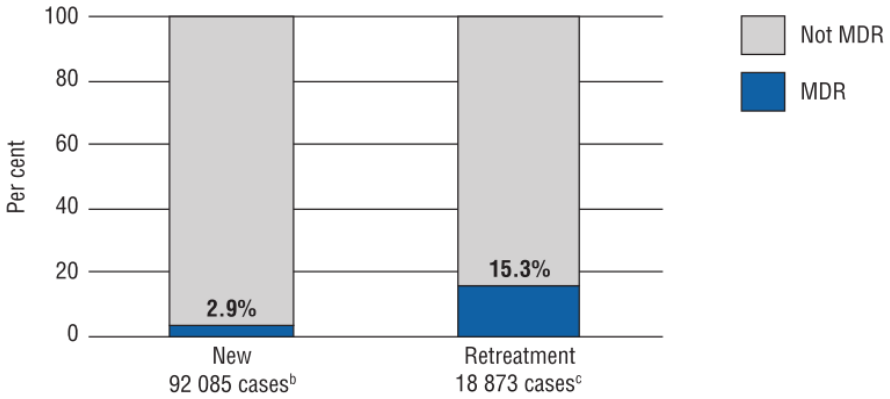

Figure 1.3: Weighted mean of MDR-TB in new and retreatment TB cases between 1994 and 2007. [b] data from 105 countries, [c] data from 94 countries (20)

In places where the transmission of TB is stable or increasing, the incidence of TB cases is higher among young adults and most cases are due to recent infection or reinfection. On the other hand, if transmission decreases, the incidence of TB is higher in old adults and the cases are attributable to reactivation of a latent infection. In countries with low incidence rates, such as Western Europe and North America, indigenous patients tend to be older whereas younger patients are mostly immigrants from high-incidence countries (21).

The STOP TB Strategy and the Global Plan to Stop TB, currently in execution, were launched in 2006 by the World Health Organization (WHO). These plans aim to address major TB problems, namely, weak health systems, the epidemics of HIV-associated TB and the emergence

of MDR. WHO set a few targets in the control of TB. One of the targets is, to reduce the prevalence of TB and the deaths due to TB by 50% compared with the baseline of 1990 by 2015. The aim for 2050 is to eliminate TB as a public health problem, defined as a worldwide incidence of TB of less than 1 case per million per year. However, the target for 2015 is unlikely to be met worldwide due to the epidemics of HIV-associated TB in Africa and the emergence of MDR tuberculosis in Eastern Europe.

1.4 Tuberculosis in Portugal

According to the national health report, in 2011, Portugal had 2388 diagnosed cases of TB, including new and recurrent cases. Of those, 2016 were from Portuguese patients and 372 (16.6%) were from foreigners. At the moment, Portugal is a medium-incidence country, with an annual decrease of 6.4% since 2002. Most of TB cases are concentrated in the metropolitan areas, such as Lisboa, Porto, Setúbal, Braga and Beja, as depicted in figure 1.4 (22).

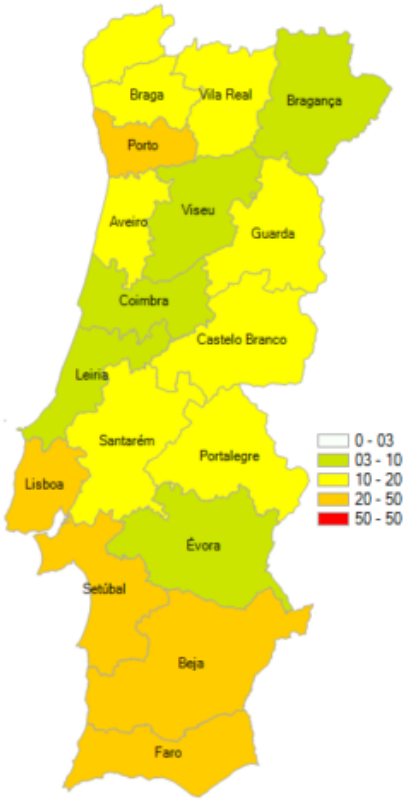

Figure 1.4: Geographic distribution of new cases of TB, in 2011 (22)

In 95% of the total number of cases, the diagnostic was made when symptomatic patients sought medical care and only 5% were detected through screening. Most of the infected patients had pulmonary disease (73%) and, among these, 8% also had lesions in other parts of the body. Among foreigners, the proportion of cases is stable since 2004, representing 16% of the total of cases. This is the lowest proportion of cases among foreigners, in the European

Union. However, a foreigner has still an higher risk (four times higher) than a local to acquire TB. Through the National report (22), it is possible to notice the decreased risk of tuberculosis in all age groups, more pronounced between 24 and 44 years. In 2010, 89% of TB cases were tested for HIV and 12% of them had a positive result. This rate is one of the highest in all EU. In fact, TB is the leading cause of death among people with HIV, being responsible for about 40% of deaths. In 2011, MDR represented approximately 1.7% of the cases of TB (1.3% among new cases and 8.2% in recurrent cases). This proportion is somewhat smaller than the average one across EU (22).

Following EU Health policies, the main goals for the Portuguese Directorate General for Health (DGS) are the detection of at least 70% of the cases and the cure of 85% of TB cases per year. At the moment, Portugal is one of the seven EU countries to reach the aimed detection rate with 87% (corresponding to the blue line in figure 1.5). Regarding the therapeutically success rate, Portugal is one of the three EU countries that consistently surpassed the 85% threshold despite a recent drop to 77% (green line in figure 1.5) (22).

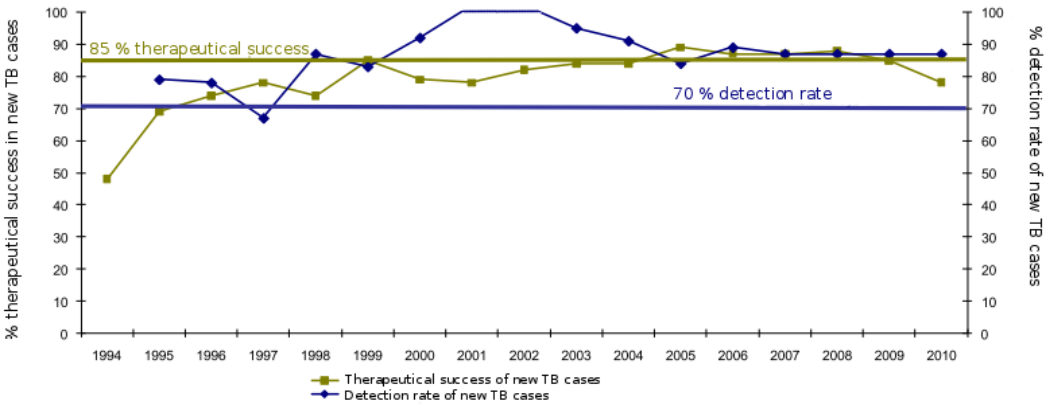

Figure 1.5: Rate of detection and therapeutically success (22)

1.5 Objectives of the thesis

The main goal of this project is to analyze recurrent cases of TB, in order to identify the main risk factors for time to reccurence. Understanding the risk factors can help to minimize the frequency of recurrence. Given the controversy surronding it, a secondary aim is to analyze the protective effect of the BCG. However, there is a large number of patients for whom there is no information about their BCG vaccination. Therefore, in order to consider this variable, multiple imputation methods are used.

Chapter 2

SVIG-TB Database

2.1 Database description

The TB data that were analyzed came from SVIG-TB, an informatic application from the "Programa Nacional de Luta Contra a Tuberculose", from the Portuguese National Health System.

Cases of TB disease, confirmed or probable, and cases of TB infection [1], detected in centros de saúde, hospitals, prisons or centers for addictions treatment (CATs) are mandatory notified to SVIG-TB (i.e. TB episode). In most cases the doctor fills in a form, by hand, with some of the most relevant clinical and socio-demographic information about the patient. Afterwards, these data are typed in and sent to SVIG-TB. This step, although important in the data collection, is also one of the main source of errors and represents a limitation in the quality of the data. WHO estimates that around 10% of the cases in Portugal are not detected and therefore not notified to the SVIG-TB (12).

2.2 Criteria for inclusion and exclusion

The statistical analysis was conducted using data from patients diagnosed between 1st of January 2002 and 31st December 2009. Although data from previous years are available they do not constitute a good sample and were discarded from the analysis. For the same reason, an upper limit was defined at the end of 2009. Data between 2002 and 2009 also have some missing patients, however these are not significant and therefore these years are a appropriate representation of the reality. Indeed, due to a slow update of the SVIG-TB database, the years after 2009 are incomplete, and, therefore, were discarded. The data was also selected so that only laboratory confirmed cases of TB [2] were used.

For the analysis of recurrent TB cases, patients without identification were removed from the

[1] TB disease refers to active TB while TB infection refers to latent infection, as explained in section 1.1.
[2] Disease caused by a bacteria from the MTBC.

analysis. Without the identification number it was impossible to link episodes to a patient. Recurrence is an episode occurring since the day the first treatment ended. Patients that died during the first episode were also discarded as they died before ending the treatment. Some outcomes, namely chronic TB and failure, were excluded since the number of patients with these outcome were very low. Individuals whose follow-up time is smaller than 12 months were also excluded from the analysis as they were considered not to have enough time for suffering a recurrent case of TB.

2.2.1 Limitations of the database

As seen in section 2.2, the choice on the time period was limited on the starting and ending dates of the collected data by SVIG-TB. Another serious limitation is the amount of missing data. In fact, important information such as being vaccinated with BCG has more than 50% of the data missing, and other relevant data, like radiology results, job of the patient, among others, have between 5% and 10% missing data.

A recurrent episode may be due to endogenous reactivation, also named as relaspe or exogenous reinfection. To differentiate between relapse and reinfection genotyping methods are needed. Relapse is defined when the two episodes of TB have identical or similar genotypes while episodes with different genotypes are defined as reinfection (12; 23) Ideally, it would be interesting to study the different risk factors for each case. However, molecular data on each TB episode is not available in the SVIG-TB dataset. Other important clinical data that was mostly absent from the used dataset was information on antibiotic resistance of the bacteria and on the socio-economic status of the patient, both of which can have an impact on TB infection. There is also a limitation with self-reported variables namely, use of drugs, alcohol, smoke etc. The patient could be not willing to share his addiction due to psychological factors. This could be influenced by the fear of his physician reaction or the inability to see his substance abuse as a problem. Whereas variables like date of symptoms consists of the memory of the patient to record the first day he had symptoms.

2.3 Variable description

The variables considered in this study were previously reported as risk factors for TB in the national and/or international literature. Some risk factors were not included since they were absent from the SVIG-TB database. The variables considered can be divided in two main categories: clinical variables (vaccine, clinical form of TB, time of diagnosis, radiology, previous TB treatment, outcome, treatment duration, HIV, diabetes mellitus and number of comorbidities) and socio-demographic variables (gender, age, country of origin, number of years in Portugal, smoking habits, alcoholism, unemployment, drug abuse, prison, job, community residence, homeless and transfered). Table 2.3.2 has a brief introduction of the variables considered.

2.3.1 Clinical variables

Vaccine is a binary variable representing whether the patient was vaccinated with BCG or not. The vaccine protects against severe forms of TB in children (TB meningitis and miliary TB), although it is not recommended for use in infants known to be infected with HIV (12). Its efficacy in preventing pulmonary TB in adults, in the other hand, is highly variable (12; 24). Another problem is in detecting BCG vaccination. Indeed, most studies refer only to the presence or absence of a BCG scar, despite that although the presence of a scar confirms vaccination, its absence does not confirm lack of it. In fact, some BCG-vaccinated children do not develop a detectable scar (25).

Clinical form is a binary variable indicating if the patient has a pulmonary form of TB or an extrapulmonary form. It is important to distinguish TB clinical forms since they can have very different characteristics, namely, regarding the efficacy of BCG and the infectiousness of the disease (12; 24).

Time to diagnosis is a continuous variable representing the time (in weeks) since the onset of symptoms until the diagnosis. A late diagnosis may worsen the disease and increase the risk of death. It can also enhance tuberculosis transmission in the community, since usually the transmission occurs between the onset of symptoms and initiation of treatment (19; 26). The recommended time until TB treatment (i.e. time to diagnosis and starting treatment) is 3 to 4 weeks, but in practice this is rarely the case (27; 28). Note that the onset of symptoms was determined by retrospective self-diagnose which is likely not to be very accurate.

Radiology is a categorical variable, with the values: "Normal", "With cavitated lesions" and "Without cavitated lesions". The term "Normal" indicates that the patient does not have any lesion; "With cavitated lesions" indicates that the patient have cavitated lesions; "Without cavitated lesions" indicates that the patient does not have cavitated lesions but have other type of lesions. Detecting cavitated lungs is important, since these lesions were frequently found to be significantly associated with recurrence risk (29; 30).

Previous TB treatment is a binary variable indicating if the TB case is the first infection of the patient or a recurrent episode.

Patient outcome is a categorical variable corresponding to the final state of the disease episode, which can be "death", "default" or "therapeutical success". The term "death" indicates the occurrence of death of the patient by any cause during treatment; "default" corresponds to the interruption of the treament by the patient for two or more consecutive months; "therapeutical success" is defined by WHO as a TB case with positive culture having the final outcome of "cured" (Patient with positive culture at the beginning of treatment and negative result at the end of treatment) or "completed treatment" (Patient who completed the treatment but does not have a negative culture at the end of treatment) (12), however, in the SVIG-TB dataset obtaining an initial positive culture of TB-agent was not necessary to consider "therapeutical success". Interestingly, previous studies showed that patients who defaulted treatment were associated with an higher risk of having a recurrent episode of TB (29; 30).

Duration of treatment is a continuous variable representing the time (in months) since the beginning of treatment and until the outcome occurred ("death", "default" or "therapeutical

success").

HIV is a binary variable indicating whether the patient has HIV or not. Having HIV is one of the most important risk factor for progression of Mtb infection to active disease, and several studies showed that HIV infection increases the risk of TB recurrence (29; 30).

Diabetes is a binary variable indicating whether the patient has diabetes mellitus or not. Several studies indicated that patients with diabetes have an increased risk of failing the treatment or of dying during TB treatment, and of relapse (31; 32).

Number of comorbidities is a discrete variable representing the number of diseases, other than TB, that the patient suffers from, such as liver disease, sarcoidosis, silicosis, collagen disease, chronic obstructive pulmonary disease, neoplasm or lymphomas.

2.3.2 Socio-demographic variables

Gender is a binary variable representing the gender of the patient. The relation male-female in TB patients for Portugal is 2 to 1, which is similar to the rest of the EU, with an increasing trend in males.

Age is a continuous variable representing the age of the patient (in years) at the time of diagnosis. Several studies suggest that older adults may have an higher risk of death or recurrent episodes (30; 33). However, linking old age with TB-related death can be difficult since natural death also increases with age.

Country of origin is a categorical variable with the following values: "Born in Portugal", "Born in a high risk country", "Born in a low risk country". According to WHO (34), when the proportion of foreign-born cases exceeds 70% of the total reported cases, no greater than 2% decrease in annual TB incidence can be expected. This is not the case of Portugal, which has one of the smallest rates of foreign-born cases, corresponding to a very small proportion of the total cases. Anyhow, some studies showed an association between the country of origin and TB recurrence. In fact, immigration from a high-burden country for TB has been shown to be a major risk-factor in developed countries (30; 35).

Number of years in Portugal is a continuous variable representing the number of years for which the patient has been living in Portugal.

Smoking habits is a binary variable representing whether the patient smokes or not. The database had a free field where the physician could write the number of cigarettes smoked per day or if the patient had stop smoking however, this information was very scarce. Hence, to assess if the patient had smoking habits or not, a field smoker/non-smoker was used. Being a smoker can be a particularly important factor in TB recurrence, since nicotine has been associated to a higher chance to progress from latent infection to active infection (36).

Alcoholism is a binary variable indicating if the patient has an alcohol dependence or not. Alcohol dependency is based on CAGE score. It is considered dependency if the patient has a need to consume alcohol in the morning or if two of the following three criteria are met: feel the need to quit alcohol; feel angry by receiving criticisms regarding alcohol; feel guilty to drink. There seems to be a substantial risk increase of TB infection among people who drink more than 40 g of alcohol per day, and/or have an alcohol use disorder (37).

Unemployment is a binary variable indicating whether the patient has been unemployed for more than 24 months. Unemployment is often associated with recurrence of TB, thus, this variable was included in the study (30).

Drug abuse is a binary variable indicating whether the patient is a drug addict or not. Dependency of drugs excludes the occasional use of drugs. Drug users remain a group at high risk of TB infection and drug using has been occasionally reported as associated with recurrence (29; 38).

Prison is a binary variable representing if the patient works in a prison or is an inmate. Occurrence of active TB in prisons is usually reported to be much higher than the average level reported for the corresponding general population. Overcrowding, late diagnosis, inadequate treatment of infectious cases, precarious hygiene conditions, low quality of food and stress, are all factors that favour transmission of TB (39; 40).

Health-care workers is a binary variable representing if the patient works/studies in a health care facility or not. Health-care workers, which are potentially exposed to TB on an every day basis, are usually considered a high risk group for TB infection and transmission.

Community residence is a binary variable representing whether the patient lives in a community residence or not. This variable was included in the analysis, since most of the people living in a community residence may have lower socio-economic status and there is a high degree of social interaction, both of which are known risk factors to increase TB transmission and infection.

Homeless is a binary variable indicating if the patient is homeless or not. Homelessness has been associated with TB infection.

Transferred is a binary variable representing whether the patient was transfered from another health institution or not. Both records are used to complement information, however only the second record (corresponding to the end of treatment) is kept. Information about the transference if save on this variable. This variable was included in the analysis since stopping the treatment, although temporarily, can be associated with a higher risk of infection.

Table 2.1: Coding of the variables

Name	Description	Coding
Vac	BCG vaccination	0 - Yes 1 - No
CliForm	Clinical form of TB	0 - Pulmonary form 1 - Extrapulmonary form
Symp	Time to diagnosis tuberculosis, in weeks	Continuous
Radio	Radiology	0 - Normal 1 - Without cavitation 2 - With cavitation
PrevTB	Previous TB treatment	0 - First infection 1 - Recurrent episode
Sit	Patient outcome	0 - Cured 1 - Death

Continue on next page

Table 2.1 – continued from previous page

Name	Description	Coding
		2 - Default
Treat	Duration of treatment, in months	Continuous
HIV	HIV	0 - No 1 - Yes
Diabetes	Diabetes	0 - No 1 - Yes
NumCo	Number of comorbidities	0 - No more diseases 1 - Has one disease +2 - Has two or more diseases
Sex	Gender	0 - Female 1 - Male
age	Age of the patient	Continuous
Origin	Country of birth	0 - Portuguese 1 - Low risk country 2 - High risk country
Arrival	Number of years in Portugal	Continuous
Smk	Smoker	0 - No 1 - Yes
Alc	Alcoholic	0 - No 1 - Yes
Unemp	Unemployed	0 - No 1 - Yes
Drugs	Drug dependent	0 - No 1 - Yes
Prison	Prison	0 - No 1 - Yes
Job	Health care worker	0 - No 1 - Yes
Commu	Community residence	0 - No 1 - Yes
Hmless	Homeless	0 - No 1 - Yes
Transf	Transfered from another institution	0 - No 1 - Yes

Chapter 3

Methodology

3.1 Understanding and handling missing data

The goal of any statistical analysis is to make valid inferences regarding a population of interest. A common problem of data analysis is the existence of missing data, which is unavoidable in epidemiological and clinical research. The way most studies deal with it is to discard the whole entries with missing information, performing a complete case analysis. In fact, most software packages automatically exclude individuals with missing observations and perform a complete case analysis (41). This is not always the most appropriate solution and can lead to inferences substantially different from those who be would obtained if no data had been missing, and to a reduction of the statistical power of the analysis. Two other methods which are frequently used are Mean Imputation (the mean of the observed values is used to replace a missing value) and Missing Indicator (the missing values are treated as a separate category), but generally they also lead to biased results (42; 43; 44).

Types of missing data

Firstly, it is important to understand why the data are missing. Rubin (45) classified the missing data mechanisms into three types: Missing Completely At Random (MCAR), Missing At Random (MAR) and Not Missing At Random (NMAR).These mechanisms describe relationships between measured variables and the probability of missing data.

When data are **MCAR**, missing cases are not different from non-missing cases. Thus, the probability of missingness of the value of a variable is unrelated to other variables and to their values, and the missing values are randomly distributed through the dataset. Examples of MCAR data are when a tube with a blood sample of an individual is accidentally broken or when a patient questionnaire is lost. In cases like these, individuals with missing values can be excluded from the analysis and valid inferences obtained, nevertheless, it will still result in loss of statistical power. The MCAR mechanism is expressed through the formula:

$$P(\boldsymbol{R}|\boldsymbol{Z}) = P(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}, \phi) = P(\boldsymbol{R}|\phi) \tag{3.1}$$

where $\boldsymbol{Z}$ is a vector of partially observed data, that is, $\boldsymbol{Z} = (\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis})$, $\boldsymbol{R}$ is a set of response indicators (i.e., $\boldsymbol{R}_j = 1$ if the jth element of $\boldsymbol{Z}$ is observed, and $R_j = 0$ if the jth element of $\boldsymbol{Z}$ is missing), indexed by parameters ϕ (41; 43; 45; 46; 47).

However, if the missingness of a variable depends on observed values then the missing-data mechanism is said to be **MAR**. The MAR assumption is valid if it can be assumed that the pattern of missing values is conditionally random, i.e. given the observed data, the probability distribution of $\boldsymbol{R}$ is independent of the missing data. The MAR mechanism is expressed through the formula:

$$P(\boldsymbol{R}|\boldsymbol{Z}) = P(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \phi) \tag{3.2}$$

But if the missingness depends on information that has not been recorded, the mechanism is classified as **NMAR** or nonignorable (41; 43; 45; 47):

$$P(\boldsymbol{R}|\boldsymbol{Z}) = P(\boldsymbol{R}|\boldsymbol{Z}^{obs}, \boldsymbol{Z}^{mis}, \phi) \neq P(\boldsymbol{R}|\boldsymbol{Z}^{\boldsymbol{obs}}, \phi) \tag{3.3}$$

The majority of studies deals with MAR data and the majority of newly developed software packages makes the assumption that the data are MAR (43). For instance, for patients with more severe symptoms, the priority is to perform more tests in order to check for more serious conditions, whereas patients with mild symptoms are less likely to go through additional tests and a greater emphasis is given to understand the patient history. Therefore, under these conditions, perform a complete case analysis will lead to biased estimates of regression coefficients and overestimation of precision since the subset of complete observations is highly selected (41). The same scenario is present when the proportion of missing data is large. Performing a complete case analysis or mean imputation will also lead to biased results since the sample of complete cases might not be a proper representation of the entire dataset (42; 44; 48).

Another important concept relates to whether the pattern of missing data is monotone or nonmonotone. The missingness is monotone when the data matrix can be arranged in a way to create a hierarchy, i.e., observing a variable Z_b for a subject implies that variable Z_a is observed for all $a < b$. A nonmonotone pattern happens when the variables are never observed simultaneously (46; 49; 50; 51). Figure 3.1 displays two hypothetical datasets. The majority of the multiple imputation software sorts the data into groups based on whether the variables are observed or missing.

3.1.1 Complete Case Analysis

Complete case (CC) analysis is also known by listwise deletion. CC consists of omitting cases with missing data and proceed with the analysis on what remains. However, using CC analysis is only reasonable, although inefficient, when the subjects with missing data are a random sample of all the cases, that is under the MCAR assumption. However, even under the MCAR assumption, one of problem of using CC is that the number of complete observations could be rather small if many variables are considered. In that case, the problem is that a small dataset leads to higher standard errors and wider confidence intervals (52; 53).

	Hypothetical Monotone				Hypothetical Non-monotone			
Pattern	Y	X_1	X_2	X_3	Y	X_1	X_2	X_3
1	Obs	Obs	Obs	Obs	Obs	Obs	Obs	Obs
2	Obs	Obs	Obs	M	Obs	Obs	Obs	M
3	Obs	Obs	M	M	Obs	Obs	M	M
4	Obs	M	M	M	Obs	Obs	M	Obs
5					Obs	M	Obs	Obs
6					Obs	M	M	Obs

Figure 3.1: Monotone and non-monotone patterns of missingness (Obs=observed, M=missing) (50)

If the data is MAR or NMAR, using CC will lead to biased estimates and loss of efficiency, power and precision (52).

3.1.2 Mean Imputation

Single imputation is another frequently used method. However, this method treats all values as observed values, ignoring that some values have been imputed, resulting in small standard errors (54).

Mean imputation is an unconditional single imputation method, since that for any variable, the missing value is replaced by the mean or median of the observed values. The term unconditional refers to the fact that information about the individual is not collected to impute the missing value. Mean imputation generally results in unbiased estimates but overestimation of the precision (43; 55).

3.1.3 Multiple Imputation

Multiple imputation is an effective method to deal with missing data and relatively easy to implement since it is now available in diverse statistical software. However, imputations are computionally intensive and need to be applied carefully to avoid distortion of estimates and standard errors. Multiple imputation should not be regarded as a routine technique, the modelling process should be done carefully and appropriately. A recent study (56) used multiple imputation to handle the missing data in the dataset. The results of this study suggested that cardiovascular risk was not associated to cholesterol. Using a complete case analysis, the authors found a clear association between cardiovascular risk and cholesterol. The reason for the lack of association between these variables in the imputed dataset was the omission of the cardiovascular disease outcome in the imputation model and the high proportion of missing values for HDL (High Density Lipoprotein) cholesterol. Therefore, it is important to be aware of the problems that an incorrect imputation may bring (54).

Understanding how multiple imputation works is not very difficult. Multiple imputation is implemented in three steps: imputing the data, analyzing the data and pooling the results, sometimes referred as the imputation phase, the analysis phase and the pooling phase, respectively. The first step consists in replacing the missing values by imputed values, thus creating multiple versions of "complete" datasets without missing values. Multiple imputation is based on a bayesian approach since the imputed values are sampled from their predictive distribution based on the observed data. The true values of the missing data are never known, therefore, in order to account for the uncertainty in predicting the missing values, some variability is included into the multiple imputed values. The analysis are run separately on each dataset since the estimates will differ in each imputation due to the variability introduced. The final step consists in combining (pooling) the results of multiply imputed datasets using "Rubin's rules" (49), which account for the variability between the imputed datasets and generates correct standard error estimates and coverage rates (54; 55; 57; 58). The estimate of each parameter (e.g. regression coefficent *b*) is simply the average of the parameter estimates β_m obtained over the *m* imputed datasets ($m = 1, .., M$):

$$\beta^* = \frac{1}{M} \sum_{m=1}^{M} \beta_m \tag{3.4}$$

In multiple imputation, the variance of the estimator is partitioned into the within imputation variance, which captures the sampling variability, and the between imputation variance, which captures the estimation variability due to missing data (59; 60). The within imputation variance, U_b, is the average of the squared standard error (SE) of the regression coefficient estimates over the *m* imputed datasets.

$$U_b = \sum \frac{SE_b^2}{m} \tag{3.5}$$

The between imputation variance, B_b, is the sample variance of the parameter estimates over the *m* imputed datasets.

$$B_b = \frac{1}{(m-1)\sum(b-\bar{b})^2} \tag{3.6}$$

These two variances are combined in order to provide a single variance, given by

$$T_b = U_b + \left[1 + \frac{1}{m}\right] B_b \tag{3.7}$$

Assumptions of the methods

The majority of algorithms for multiple imputation assumes that the data are normally distributed. Some debate exists around the best way to deal with skewed variables. Some authors (54; 61) claim that the inclusion of non-normally distributed variables may introduce bias. Therefore, it is recommended to transform such variables in order to approximate normality before imputation and then transform the imputed values back. However, recent studies

by Von Hippel (62; 63) argue against transforming skewed variables since by transforming the variable to meet the normality assumption, one also changes the distribution of the variable and the relationship between that variable and the others to impute. Von Hippel claims that the methods used to transform variables create more bias than imputing the skewed variable.

For a plausible imputation, it is crucial to include as much information as possible. Any variable that will be in the analysis model should also be in the imputation model. The dependent variable should also be included since it may contain information about the missing values of some covariables. Variables with interactions are more complicated to add to the imputation. Several solutions exists to deal with interactions, in case of an interaction between a binary variable and a continuous variable, the dataset is divided in two (one dataset for each category of the binary variable) and the imputation is done separately for each dataset. After the imputation, the datasets should be combined (61).

Number of iterations and imputations

Recent investigation (55; 64) suggests that a small number of iterations should be sufficient, usually between 5 to 20 iterations for each imputation. The idea is that after 10 iterations, the order in which variables were imputed no longer matters since the imputations have stabilized. Therefore, in this study a total of 10 iterations will be used. Although 10 iterations is not a large number it will produce accurate results and prevent loss of power.

The traditional approach to select the number of imputations (m) is based on the efficiency proposed by Rubin (49). The idea is that 2 to 10 imputations are enough since it generates confidence intervals and hypothesis tests close to their nominal coverage and significance levels. However, if the fraction of missing information (γ) is large, a larger number of imputations is needed. The fraction of missing information can be informally described as the quantity of information lost about each coefficient due to the missing data. The missing data are specific for each coefficient. In case of multivariate datasets, the fraction of cases with missing values is not equivalent to the fraction of missing information (57).

Graham et al. (59) compared the effect of different values of γ with different m. When m was low, the values of mean squared error (MSE) and SE increased, the power of statistical tests was reduced and the variability of estimates increased. Although statistical efficiency and power are important, one should also consider the reproducibility of the experiment, this means considering the Monte Carlo error of the results. The Monte Carlo error is defined as the standard deviation across repeated runs of the same imputation. With the increase of m the Monte Carlo error tends to zero (60).

The relative efficiency proposed by Rubin is calculated by:

$$\frac{1}{1+\frac{\gamma}{m}} \tag{3.8}$$

However, this formula assumes knowledge of γ which, most of the times, is unknown. Besides, Rubin's method does not address imputation variability (57).

A different approach suggested by Bodner et al. (57) suggests computing a estimated of γ:

$$\hat{\gamma_L} = 1 - \frac{n_L}{n} \tag{3.9}$$

Where $n_L = 2890$ and $n = 8364$, where n_L correspond to the number of complete cases without missing information and n correspond to the total number of observations, including missing observations, therefore:

$$\hat{\gamma} = 1 - \frac{2890}{8364} = 0.65 \tag{3.10}$$

This method is conservative since γ would be less than $\hat{\gamma}$ with this difference increasing with variables with missing values strongly related to other variables.

Several articles (57; 59; 60) suggests the application of a rule of thumb. *m* should be at least equal to the percentage of incomplete cases. For this study 70 imputations will be consider. The efficiency can be calculated with the estimate of γ :

$$\frac{1}{1 + \frac{0.65}{70}} = 0.99 \tag{3.11}$$

Guidelines for imputation

Although, creating multiple imputed datasets can be computationally demanding, researchers have much to gain with multiple imputation. With the recent investigation around this topic, it is no longer excusable to discard the missing values and use only a complete case analysis. To avoid pitfalls, Sterne et al. (54) mention some guidelines to follow when presenting papers or reports about missing data. These are useful suggestions and all will be followed through this thesis.

* Mention the number of missing values or the number of complete cases for each variable. Try to interpret why these values are missing.
* Compare the distributions of variables for individuals with complete and incomplete data.
* Describe the methods used to deal with missing data and the assumptions made, i.e. type of missing data, etc. Provide details of the imputation technique as well as details about the software used.
* Mention the number of imputations used and the reason for that number.
* List the variables used in the imputation and their characteristics, i.e. normally distributed, binary, interactions etc, and how these variables were handled.
* In studies with a large amount of missing data, compare observed and imputed values.
* When possible, include results from complete case analysis and compare with the results obtained with multiple imputation, discussing the differences (if any is present) and suggest ideas about these differences.
* Discuss the plausibility of MAR assumption and whether the variables included in the imputation follow this assumption.

3.1.3.1 Multivariate Imputation by Chained Equations (MICE)

MICE, also known as Fully Conditional Specification(FCS), is one of the approaches used to impute multivariate data. MICE is a flexible method, that does not rely on the assumption of multivariate normality. The multivariate imputation model is specified on a variable-by-variable basis by a set of conditional densities, one for each incomplete variable. This approach is adequate since it does not restrict conditional distributions to being normal.

Let $Y_j = (j = 1, ..., p)$ be one of p incomplete variables, where $Y = (Y_1, ...Y_p)$. Let $Y_{-j} = (Y_1, ..., Y_{j-1}, Y_{j+1}, ..., Y_p)$ denote the collection of $p - 1$ variables in Y except Y_j. Let the hypothetically complete data Y be a partially observed random sample from the p-variate multivariate distribution $P(Y|\theta)$. Assuming the multivariate distribution of Y is completely specified by θ, a vector of unknown parameters. The chained equations proposes to obtain a posterior distribution of θ by sampling iteratively from conditional distributions of the form:

$$P(Y_p|Y_{-p}, \theta_p) \tag{3.12}$$

The parameters θ_p are specific to the conditional density. The ith iteration consists of successive draws of the Gibbs sampler:

$$\begin{aligned} \theta_p^{*(i)} &\sim P(\theta_p|y_k^{obs}, y_1^{(i)}, ..., y_{p-1}^{(i)}) \\ y_p^{*(i)} &\sim P(y_k|y_k^{obs}, y_1^{(i)}, ..., y_k^{(i)}, \theta_k^{*(i)}) \end{aligned} \tag{3.13}$$

where $Y_j^{(i)} = Y_j^{obs}, Y_j^{*(i)}$ is the jth imputed variable at interation i. Iterations of equation (3.13) are executed *m* times to generate *m* imputations.

The Multivariate Imputation by Chained Equations was implemented through the package **mice** (61). Figure 3.2 illustrates the steps involved in multiple imputation: imputation, analysis and pooling. First, the variable with least missingness is imputed conditional on all variables with no missing values, the variable with the second least missingness is then imputed conditional on the variables without missing values and the imputed variable, until all variables have been imputed. This process is repeated for *x* iterations. Each of the *m* imputed datasets will have distinct imputed values. The magnitude of the difference reflects the uncertainty about the missing value. The second step consists in analysing the data to estimate the parameter of interest. After the analysis, the results are pooled using Rubin rules (49; 55), obtaining a single estimate for the parameter of interest.

Several problems may arise when dealing with multivariate data: variables could be correlated with each other, they may be of different types (some binary, continuous, ordinal, etc), relations between variables could be complex (for instance in cases of censoring data), etc. To address these issues, it is convenient to specify the imputation model separately for each column in the data. The name chained equations refers to the fact that the Gibbs sampler can be implemented as a concatenation of univariate procedures to impute the missing data (61).

Imputation models Van Buuren (61) specifies some rules in order to perform the most accurate imputation.

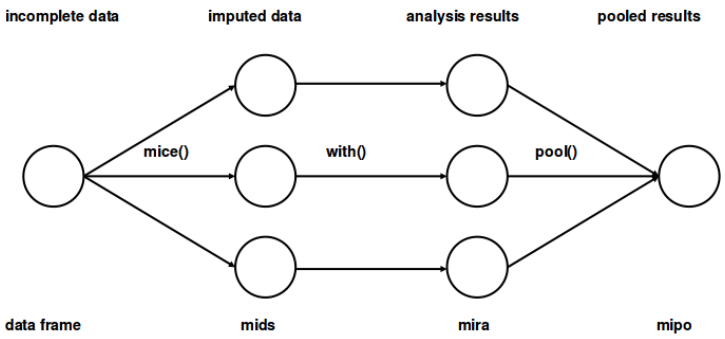

Figure 3.2: Schematic approach of the **mice** algorithm (61)

* Explore the data in order to understand what type of missing data is present, MCAR, MAR or MNAR. The package **mice** handles MAR and MNAR data.
* The second concern should be the form (structural part and error distribution) of the imputation. The form has to be specified for each column with missing data.
* When imputing missing values it is recommend to use as many variables as possible; in fact, adding 'auxiliary variables' related to the missingness strenghten the MAR assumption (55). **mice** has a function to easily specify the set of predictors to be used for each incomplete variable.
* The fourth point relates to the form of variables. If any of the variables needs a tranformation, if any interaction term is present, etc.
* Another aspect to consider is the order by which the variables are imputed. Within **mice** it is easy to change the order.
* An important consideration is the number of iterations that should be between 5 to 20.
* In what concerns the number of imputed datasets (m), setting m too small lead to low *p*-values.

For every distinct dataset, these points should be checked and carefully chosen to originate an accurate imputation model.

3.1.3.1.1 Predictive Mean Matching (PMM)

PMM is a semi-parametric imputation method, developed by Little (65). This method combines elements of regression, nearest-neighbour and hot deck imputation. The basic concept of PMM is to impute a missing value by matching its predictive mean to a nearest neighbor among the predictive means of the observed values, and to adopt the actual observed value.

A regression model is estimated where the dependent variable is the variable to impute and the remaining variables act as independent variables. A value for the dependent variable is estimated for individuals with missing data. The predicted value of the dependent variable is matched with the nearest fitted value, calculated through euclidean distance. Let y_i be a

value of a variable Y, with $i = 1, ..., n$ where only units $1, ..., n_{obs}$ are observed. Let $\hat{y}_i$ be a predicted value from a regression of Y on some explanatory variables. The euclidean distance between $\hat{y}_i$ and $\hat{y}_j$ is given by

$$D_{i,j} = |\hat{y}_i - \hat{y}_j| \tag{3.14}$$

and $y_{obs,j}$ is imputed for $y_{mis,i}$ (66). The observed value corresponding to the nearest fitted value is imputed. If more than one fitted value has equal distance to the minimum distance found, the value to impute is randomly chosen between those that were tied.

By imputing observed values, PMM only gets plausible values for the imputed data which is an advantage since the range and shape of distribution is preserved. This method is complicated when imputing a categorical variable with more than 2 levels, like Radio and Origin.

3.1.3.1.2 Random Forest (RF)

A decision tree is a tree with thresholding nodes (continuous variables) or categorical nodes (categorical data), which contains information about the attributes in the input vector. The information is used to follow a decision path. Random Forest is an extension of classification and regression trees, that uses ensembles of decision trees. A powerful advantage of RF is that it does not rely on distributional assumptions and works well with nonlinear relations and interactions. In fact, a previous study (67) compared the performance of several methods applied to datasets with interactions between variables and the imputation using **mice** with RF resulted in less biased paramater estimates.

To impute missing values, the algorithm inaccurately fills the missing values, then performs a forest run and computes proximities. In case of a continuous variable, the value is estimated as the mean over the observed values of the y_{th} variables weighted by the proximities between the n_{th} case and the observed value. In case of a categorical variable, the missing value is replaced by the most frequent value weighted by proximity. Iterate a new forest using the previously imputed values to impute new values and iterate.

If two trees are highly correlated in the forest, the forest error rate will increase. A tree with a high error rate is a low classifier. Increasing the strength of each individual tree will decrease the forest error. A low value of m reduces both the correlation and the strength. During each run, the out-of-bag (OOB) error is estimated. Each tree is constructed with a different bootstrap sample from the dataset. About one third of the cases are not used in the bootstrap sample. For each case not used in the construction of the k_{th} tree, put it down the k_{th} tree to get a classification. In that sense, a test set classification is obtained for each case in about one-third of the trees. At the end, let j be the class with most votes every time case n was OOB. The OOB error estimate corresponds to the proportion of times that j is not equal to the true class of n, averaged over all classes (68; 69).

Rather than taking random draws from a distribution, the package **missForest** aims to predict individual missing values. This may lead to biased parameter estimates (70). Whereas the method implemented in **mice** imputes values by randomly drawing values from independent normal distributions centered on conditional means (61).

The OOB error rate can be used to determine the number of trees. Although, a large number of trees will lead to better results this also lead to a huge computational time. Therefore, the number of trees was set to 25, since the OOB error seems to be quite small already (figure 3.3).

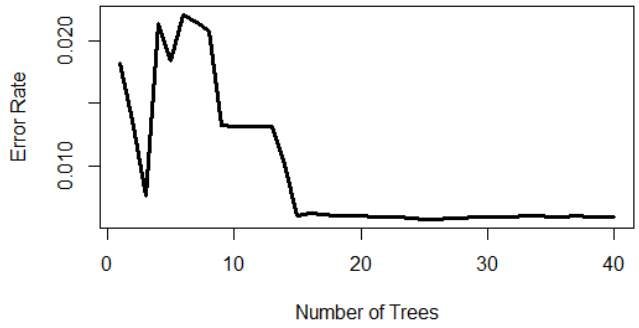

Figure 3.3: Variation of OOB error with the increase in the number of trees

3.1.3.2 Expectation-Maximization with Bootstrapping (EMB)

To implement the EMB algorithm, the package **Amelia II** was used. **Amelia II** uses the EM algorithm on multiple bootstrapped samples of the incomplete dataset to estimate values and replace the missing values by these estimated values. First, the EM algorithm is briefly explained.

Expectation-Maximization (EM)

The Expectation-Maximization (EM) algorithm, proposed by Dempster, Laird and Rubin (71), is an approach to the iterative computation of maximum likelihood (ML) estimates, useful in a variety of incomplete-data problems, either because the values were not reported, or due to the impossibility of direct observation of these values.

First, a distribution and the starting values for the mean and variance are assumed. Using these simulated values, an expected value of model likelihood is calculated, the likelihood is maximized, model parameters are estimated that maximize these expected values and the distribution is updated. The process is repeated until the values converges.

Let $\boldsymbol{X}$ be a set of observed data, $\boldsymbol{Z}$ a set of missing values and θ a vector of unknown parameters. The *log likelihood function* to estimate θ

$$\mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}) = P(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} P(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}) \tag{3.15}$$

The ML estimate of the unknown parameters is determined by the marginal likelihood of the observed data.

Each iteration consists of two steps: the E-step (Expectation-step) and the M-step (Maximization step). In the E-step, the missing data are estimated given the observed data and current estimate of the model parameters.

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = E_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^{(t)}}[\log \mathcal{L}(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z})] \tag{3.16}$$

In the M-step, the likelihood function is maximized under the assumption that the missing data are known using the estimate obtained in the E-step (formula 3.17). In summary, EM attempts to find the estimates $\hat{\theta}$ that maximizes the log probability $(x; \theta)$ of the observed data (72; 73).

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) \tag{3.17}$$

Expectation-Maximization with Bootstrapping (EMB)

Figure 3.4 represents how the EMB algorithm works. First non-parametric boostrap is applied to the incomplete dataset in order to obtain boostrap subsamples of size *n* (sample size = *n*) to be drawn of the incomplete dataset *m* times (in this case, $m = 5$). The EM algorithm is applied to each one of the bootstrap subsamples. The analysis is performed in each one of the imputed subsamples and the results are pooled using Rubin's rules.

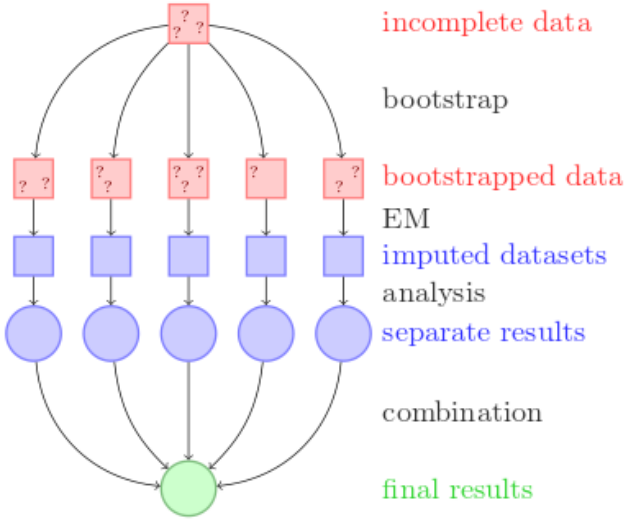

Figure 3.4: Schematic approach of the EMB algorithm (74)

The imputation model in **Amelia II** assumes that the complete data is multivariate normal, although this is often a crude approximation to the true distribution of the data. An interesting feature of **Amelia II** is that the multivariate normal model works well even when variables are discrete or non-normal. Like the majority of multiple imputation methods, **Amelia II** assumes that the data are MAR (74; 75).

3.1.4 Comparison between methods

To compare the models fitted to imputed datasets several features were compared. Among them the variables that each model selected, the direction of the estimate of the coefficients, the absolute value and standard error of the estimate of the coefficient, as well as the *p*-value associated with the corresponding test. The width of the CI was assessed for each model, since a smaller CI means that the estimate of the true effect is more precise. To validate each model, some measures were used. The R^2 of Nagelkerke represents the amount of variability explained by the covariates included in the model. Another measure used is the index of concordance, the C index, which is used to evaluate the power to discriminate between individuals with different responses. The time of execution for each imputation is also reported. The values of the statistic $-2\log\hat{\mathcal{L}}$ was used to compare models fitted to the same data.

3.2 Survival Analysis

3.2.1 Basic Concepts of Survival Analysis

Survival analysis is an area of Statistics taht was developed to analyse data representing times from a *time origin* until the ocurrence of the event of interest. In medical research, the *time origin* is often the time of recruitment into a clinical trial or study. Although, the event of interest can be the death of the patient, recurrence of symptoms or any other particular event, the event of interest is usually named *death* and the time since the *time origin* until the event of interest is named survival time.

One of the reasons why standard statistical procedures do not apply to this type of data is that survival times are generally not symmetrically distributed. In fact, survival times tend to be positively skewed, therefore it is inadequate to assume a normal distribution. However, the most important feature of survival data is the existence of censored observations (76; 77).

Censoring

Censoring is present when for some individuals the event of interest is not observed during the time of observation. Although these individuals only have partial information about their survival time, they should be included in the analysis, otherwise there will be loss of information.

Right censoring happens when the event of interest has not been observed for an individual when the study ends. This may be because the event of interest occurs after the end of the study, or the patient may have been lost to follow-up. The right-censored time is less than the actual, yet unknown, survival time. Right censoring can be classified into censoring type I, censoring type II and random censoring. Right censoring type I occurs when the observation times are fixed by the researcher. The number of *deaths* is random. However, if the number of *deaths* is not random, i.e., the study ends when r *deaths* are observed the study presents right censoring type II. r is a fixed number, smaller than the total number of individuals. The most common type of right censoring is the random censoring. In medical research, usually individuals enter the study randomly, according to the diagnosis date or another important

time origin. If the study ends at a fixed date, the time since the *time origin* until the end of the study is random.

Opposite to right censoring, left censoring happens when the real survival time is less than the observed time. Left censoring occurs less frequently than right censoring. A good example of left censoring is when the researcher intends to study the age that a child performs a given task. However, some children may already perform the task when entering the study, therefore the observations are left censored since the recorded value will be the age of the child at the beginning of the study. Another type of censoring is interval censoring. In this case, individuals are known to have experienced an event within a certain interval of time.

The individuals censored at time *t* must be representative of all individuals that survived until *t*. At any time, individuals can not be selectively censored, either because their risk of *death* is high or low, i.e., censoring is not related to the event of interest. This is known as non-informative censoring, necessary to validate the methods usually used in survival analysis. This hypothesis is usually true in case of censorship that occurs at the end of the study $\rightarrow$ random right censorship. Non-informative censorship is also known as independent censorship (76).

Survivor function and Hazard function

Let T be a positive continuous random variable that represents the survival time of an individual, i.e., the time until the ocurrence of the event of interest. The survivor function is defined as the probability that an individual survives beyond the time *t* (3.18). This function is continuous and monotonically decreasing.

$$S(t) = P(T > t), \qquad t \geq 0 \tag{3.18}$$

The probability density function is given by

$$f(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t)}{\Delta t} = -S'(t) \tag{3.19}$$

The distribution of T can also be characterized by the hazard function (3.20). This function describes the risk or hazard of *death* at some time t, and is obtained from the probability that an individual *dies* at time *t*, conditional on having survived to that time.

$$h(t) = \lim_{\Delta t \to 0^+} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \tag{3.20}$$

The function $h(t)$ is also referred to as the hazard rate, the instantaneous death rate, the intensity rate or the force of mortality. From equation (3.20), $h(t)\Delta t$ is the approximate probability that an individual *dies* in the interval $(t, t + \Delta t)$, conditional on that person having survived to time t.

The function $H(t)$, called integrated or cumulative hazard, is defined by

$$H(t) = \int_0^t h(u)du, \qquad t \geq 0 \tag{3.21}$$

These functions are related and can be obtained from each other. Some useful relationships are the following:

$$\begin{aligned} S(t) &= 1 - F(t) \\ h(t) &= -\frac{\partial \ln(S(t))}{\partial t} \\ h(t) &= \frac{f(t)}{S(t)} \\ H(t) &= -\ln(S(t)) \\ S(t) &= \exp(-H(t)) \end{aligned} \tag{3.22}$$

3.2.2 Non-parametric Inference

Kaplan-Meier Estimator

In the absence of censoring, the survivor function at a given time t is estimated by the proportion of individuals that survived beyond the time t, i.e., with observed survival times greater than t. This estimator is the empirical survivor function, given by

$$\hat{S}(t) = \frac{\text{Number of individuals with survival times} > t}{\text{Number of individuals in the dataset}} \tag{3.23}$$

However, this method cannot be used when there are censored observations. Kaplan and Meier (78) proposed a non-parametric estimator of the survivor function in the presence of censored observations. Denote $t_{(1)}, ..., t_{(r)}$ the r distinct times where the *deaths* occurred in a sample of size $n(r \leq n)$, d_i the number of *deaths* occurred in $t_{(i)}$ and n_i the number of individuals at risk at $t_{(i)}$. The Kaplan-Meier estimator of the survivor function is given by

$$\hat{S}(t) = \prod_{i:t_{(i)} \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{3.24}$$

with $\hat{S}(t) = 1$ when $0 \leq t < t_{(1)}$. If the largest observation is not censored $\hat{S}(t) = 0$ for $t \geq t_{(r)}$. However, if the largest recorded observation t^* is censored, then $\hat{S}(t)$ will never reach 0 and it is considered that the estimate is defined only until that time. When a *death* time and a censored survival time are registered with the same value, it is assumed that the *death* precedes the censoring.

The estimated variance of $\hat{S}(t)$ is known as Greenwood's formula:

$$\widehat{var}\hat{S}(t) = [\hat{S}(t)]^2 \sum_{i:t_i \leq t} \frac{d_i}{n_i(n_i - d_i)} \tag{3.25}$$

A confidence interval can be obtained for the survivor function at a given time t, based on that the estimator of the survivor function has asymptotic normal distribution with mean value $S(t)$ and estimated variance $\widehat{var}[\hat{S}(t)]$. A $100(1-\alpha)$ confidence interval for the survivor function at time t is given by

$$\hat{S}(t) \pm z_{1-\frac{\alpha}{2}} \sqrt{\widehat{var}[\hat{S}(t)]} \tag{3.26}$$

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

Given that the distribution of the survival time is generally positively skewed, the median is the usual summary measure of location. Once the survivor function has been estimated, an estimate of the median, $\hat{t}(50)$, is given by

$$\hat{t}(50) = min\{t_i : \hat{S}(t_i) \leq 0.5\} \tag{3.27}$$

being t_i the ith ordered *death* time, $i = 1, 2, ..., k$. The median estimate is the smaller observed survival time for which the estimate of the survivor function is smaller or equal to 0.5, i.e., the time beyond which 50% of the individuals in the population under study are expected to survive.

The Kaplan-Meier method allows to compare survival curves of different groups. The total of observations is divided in groups or strata (k groups), according to the covariables of interest and the survivor functions are estimated separately for each strata. It is usual to plot the estimates of the k survivor functions on the same axes. There are two possible explanations for an observed difference between two estimated survivor functions. One is that there is a real difference between the survival times of the two groups. The other is that there are no real differences and that the observed difference is only due to chance variation (76; 77; 79).

To compare the survivor functions some non-parametric tests are used, which are named rank tests, since the test statistic depends only on the order of the observations. Thus, evaluating if significant differences in survival exists between the groups. In this thesis, the log-rank test and Peto-Peto test were used since both were available in R. The null hypothesis of these tests is that there are no difference between groups

$$H_0 : S_1(t) = S_2(t) = ... = S_k(t) \quad \text{vs} \quad H_1 : \exists_{(i,j)} S_i(t) \neq S_j(t) \tag{3.28}$$

being k the number of strata.

Log-rank test

Based on the work of Mantel and Haenszel (80), Mantel (81) proposed a test designated by log-rank or Mantel-Haenszel test. Consider two groups (1 and 2) with sample sizes m and n, respectively. Let $t_1 < t_2 < ... < t_k$ be the k distinct *death* times regarding $m+n$ individuals. n_{ij} individuals are at risk in group i, (i=1,2), just before t_j, and at that instant d_{1j} individuals of group 1 and d_{2j} individuals of group 2 suffered the event, for $j = 1, 2, ..., k$. $d_j = d_{1j} + d_{2j}$

is the total number of events in the two groups from the total $n_j = n_{1j} + n_{2j}$ individuals at risk at time t_j. The information can be summarized in a 2x2 contingency table:

Group	Number of events in t_j	Number of survival beyond t_j	Number individuals at risk, in t_j
1	d_{1j}	$n_{1j} - d_{1j}$	n_{1j}
2	d_{2j}	$n_{2j} - d_{2j}$	n_{2j}
Total	d_j	$n_j - d_j$	n_j

Assuming the null hypothesis is true, the distribution of d_{1j}, conditional to the marginal values, is hypergeometric

$$p(d_{1j}|d_j, n_j) = \frac{\binom{d_j}{d_{1j}}\binom{n_j - d_j}{n_{1j} - d_{1j}}}{\binom{n_j}{n_{1j}}} \tag{3.29}$$

Under H_0, the mean value of d_{1j} is given by

$$e_{1j} = \frac{n_{1j} d_j}{n_j} \tag{3.30}$$

which represents the expected number of individuals who experience the event at t_j in group 1. The conditional variance of d_{1j} is given by

$$v_{1j} = \frac{n_{1j} n_{2j} d_j (n_j - d_j)}{n_j^2 (n_j - 1)} \tag{3.31}$$

To obtain a global measure of the deviation of the observed values of d_{1j} from their expected values consider the statistic:

$$U = \sum_{j=1}^{k} (d_{1j} - e_{1j}) \tag{3.32}$$

where U is the difference between the total observed number and expected number of events in group 1. This statistic presents $E(U) = 0$ since $E(d_{1j}) = e_{1j}$. Since the *death* times are independent, the variance of U is the sum of the variances of d_{1j}

$$var(U) = \sum_{j=1}^{k} v_{1j} \tag{3.33}$$

The test statistic proposed by Mantel and Haenszel is expressed by

$$Q = \frac{U^2}{var(U)} \tag{3.34}$$

with an asymptotic χ^2_1 distribution, under H_0.

Peto-Peto test

Another non-parametric test used to compare survivor functions is the Peto-Peto test proposed by Peto and Peto (82). This test is a generalization of the Mann-Whitney-Wilcoxon test in the presence of censored observations, with the statistic:

$$U = \sum_{j=1}^{k} w_j (d_{1j} - e_{1j}) \tag{3.35}$$

The difference between this statistic and the one from the log-rank test (3.32) is in the weight associated with each difference $(d_{1j} - e_{1j})$. While in the log-rank test this weight is equal to 1, the weight in the Peto-Peto test is $w_j = \hat{S}(t_j)$, where $\hat{S}(t_j)$ corresponds to an estimate of the survivor function given by

$$\hat{S}(t_j) = \prod_{t_i \leq t_j} \left(1 - \frac{d_i}{n_i + 1}\right) \tag{3.36}$$

which is similar to the Kaplan-Meier estimator for the joint sample.

3.2.3 Cox Regression Model

A parametric regression model requires a probability distribution for the survival time. If that assumption is not adequate the model can produce incorrect parameter estimates. However, most of the times, the interest centres on the risk or hazard of *death* at any time and not on the parameters of the distribution. In this case, the goal is to determine which combination of potential variables affect the form of the hazard function. The most frequently used model in this case is the proportional hazards model. This model was proposed by Cox (83) and is also known by Cox regression model. The model is referred as a semi-parametric model since no particular form of the distribution is assumed for the survival time; it is based on the assumption of proportional hazards.

Formulation of the proportional hazards model

Let T be a continuous random variable representing the survival time. At time t and for an individual with vector of covariates $\boldsymbol{z} = (z_1, ..., z_p)'$, the hazard function is given by

$$h(t; \boldsymbol{z}) = h_0(t)\exp(\beta' \boldsymbol{z}) \tag{3.37}$$

where β is a vector of regression coefficients that represent the effect of the covariates on survival. $h_0(t)$ represents the baseline hazard function for an individual to whom is associated a vector $\boldsymbol{z} = 0$.

It is a proportional hazards model since the hazard functions corresponding to two individuals with covariates $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ are proportional.

$$\frac{h(t;\boldsymbol{z}_1)}{h(t;\boldsymbol{z}_2)} = \exp\{\beta'(\boldsymbol{z}_1 - \boldsymbol{z}_2)\} \tag{3.38}$$

This formulation implies that the covariates have a multiplicative effect on the hazard function, that is, the hazard ratio is constant over time (76; 77; 79).

Interpretation of parameters estimates

In fact, usually, $\exp(\beta_j)$ is preferred over β_j, since $\exp(\beta_j)$ provides a straightforward interpretation regarding the risk of *death*. $\exp(\beta_j)$ represents the relative risk of ocurrence of the event of interest for two individuals that differ in one unit in the values of the covariate z_j, with the values of the remaining covariates being equal.

Consider a binary covariate defined by $z = 0$ if the individual belongs to group 1 and $z = 1$ if the individual belongs to group 2. When the individual belongs to group 1 then $h(t; z = 0) = h_0(t)$ and when the individual belongs to group 2 then $h(t; z = 1) = h_0(t)\exp^{\beta}$.

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

In the case of a numeric covariate:

$$\exp^{\beta} = \frac{h(t; z = j + 1)}{h(t; z = j)} \tag{3.39}$$

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

Likelihood function

The β_j's coefficients are estimated using the method of maximum likelihood. Consider n individuals with k observed survival times, $t_{(1)} < ... < t_{(k)}, k < n$. The set of individuals who are at risk at time $t_{(i)}$ will be denoted by $R(t_{(i)})$, so that $R(t_{(i)})$ is the group of individuals who are alive and uncensored at a time just prior to $t_{(i)}$ and is called the risk set.

$$R_i = R(t_{(i)}) = \{j : t_j \geq t_{(i)}\} \tag{3.40}$$

The likelihood function, proposed by Cox (83), for the proportional hazards model is given by

$$\mathcal{L}(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' \boldsymbol{z}_{(i)})}{\sum_{l \in R_i} \exp(\beta' \boldsymbol{z}_l)} \tag{3.41}$$

where $\boldsymbol{z}_i$ is the vector of covariates associated with the individual who *dies* at the ith ordered *death* time, $t_{(i)}$. The summation in the denominator of this likelihood function is the sum of the values of $\exp(\beta' \boldsymbol{z})$ over all individuals who are at risk at time $t_{(i)}$. The product is taken over the individuals for whom *death* times have been recorded. Individuals for whom the survival times are censored do not contribute to the numerator of the log-likelihood function but they do enter into the summation over certain risk sets. Moreover, the likelihood function depends only on the ranking of the *death* times. This likelihood function can be seen as a partial likelihood since this function does not depend of the baseline hazard function and allows inference on β, without any restriction regarding the form of $h_0(.)$ At each time t, only the information about the individuals at risk is considered. This formulation is similar to the non-parametric methods but allows an estimation of the effect of the covariates on the survival time. Under general conditions, it verifies the properties of maximum likelihood estimation. $\hat{\beta}$ is consistent, asymptotically normal with mean value β and covariance matrix $I(\beta)^{-1}$, where $I(\beta)$ is the Fisher information matrix:

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

In case of simultaneous *deaths* or when the data is not recorded properly yielding equal values, the function (3.41) cannot be implemented. In this situation, for the n individuals in the study, suppose that the distinct *death* times were observed $t_1 < t_2 < ... < t_k$. Denote d_i as the number of *deaths* occurred at time t_i and $\boldsymbol{z}_{ij}$ the vector of variables associated to individual j, that *dies* in t_i, $j = 1, ..., d_i$, $i = 1, ..., k$. If d_i is small, compared with the number of individuals in the risk set R_i, then the partial likelihood function can be approximated by the function, proposed by Peto and Peto (82) and Breslow (84).

$$\mathcal{L}(\beta) = \prod_{i=1}^{k} \frac{\exp(\beta' \boldsymbol{s}_i)}{[\sum_{l \in R_i} \exp(\beta' \boldsymbol{z}_l)]^{d_i}} \tag{3.42}$$

where $\boldsymbol{s}_i = \sum_{j=1}^{d_i} \boldsymbol{z}_{ij}$, for $i = 1, ..., k$. This is the likelihood usually implemented in software packages. If the observations do not have ties, the function (3.42) reduces to the partial likelihood (3.41) (76; 77; 79).

Variable Selection

The first step in the model selection process is to identify a set of potential explanatory variables, in order to understand which ones have a significant impact on survival of the individuals. As previously discussed, β_j represents the effect of the covariate z_j on the survival of the individual. To evaluate the existence of evidence that the covariate significantly influences the survival time, one can test

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

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

In order to find a model, without unnecessary variables, a range of automatic routines are implemented in several software packages, based on forward selection, backward elimination and a combination of the two, known as stepwise procedure. However, these automatic routines have a number of disadvantages. They lead to the identification of one subset of variables, instead of a set of equally good ones. The subset found often depends on the type of variable selection process used (forward selection, backward elimination or stepwise procedure) and on the stopping rule used to determine whether a term should be included or excluded from a model. Instead of using automatic routines, a strategy for model selection, proposed by Collett (77), was implemented. Collett uses the statistic $-2log\hat{\mathcal{L}}$ to compare alternative models, where $\hat{\mathcal{L}}$ is the maximised likelihood under a certain model.

First, a model with each variable is fitted individually. The values of $-2log\hat{\mathcal{L}}$ are calculated for each model and compared with the value for the null model (without variables) to determine which variables significantly reduce the value of the statistic, on their own.

In the second step, the significant variables from the previous step are fitted together in a model and the value of $-2log\hat{\mathcal{L}}$ is calculated. In the presence of certain variables, others may cease to be important. At each time, a variable is omitted and the value of $-2log\hat{\mathcal{L}}$ is computed. The variables that do not significantly increase the value of $-2log\hat{\mathcal{L}}$, when omitted, are discarded. Every time a variable is dropped the effect of omitting each of the remaining variables should be examined. Only the variables that significantly increased the value of $-2log\hat{\mathcal{L}}$ are kept in the model.

Variables that, when examined on their own, were not important and therefore were not included in the model in the second step, may become important in the presence of others. These variables are included in the model and if any of them significantly reduce the value of $-2log\hat{\mathcal{L}}$ they are retained.

To conclude, a final check is performed to make sure that each of the variables present in the final model, when omitted, significantly increases the value of $-2log\hat{\mathcal{L}}$ and that no variable not included significantly reduces the value of $-2log\hat{\mathcal{L}}$.

When using this procedure, a rigid significance level should be avoided. A level of around 10 % is recommended (76; 77).

3.2.4 Residual Analysis

In linear regression, a residual is the difference between the observed value of the response variable and the predicted value by the model. However, the existence of censored observations and the form of the Cox regression makes the definition of residual harder and less straightforward. A number of residuals have been proposed for use with the Cox regression model. Disadvantages of the Cox-Snell residuals and deviance residuals are that they depend heavily on the observed survival time and require an estimate of the cumulative hazard function. These

disadvantages have been overcome by a residual proposed by Schoenfeld (85). In this thesis, the Schoenfeld and martingale residuals will be used.

Schoenfeld Residuals

This type of residuals was proposed by Schoenfeld (85) and they are very useful to evaluate the assumption of proportional hazards. These residuals differ from the Cox-Snell, martingale and deviance residuals in a way that there is not a single value of the residual for each individual, but a set of values, one for each explanatory variable included in the fitted Cox regression model.

For the ith individual, the Schoenfeld residual corresponding to the covariate $z_j, j = 1, ..., p$, is expressed by

$$r_{ij} = \delta\{z_{ji} - a_{ji}\}$$

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

$$a_{ji} = \frac{\sum_{l \in R_i} z_{jl} \text{exp}(\hat{\beta}' \boldsymbol{z}_l)}{\sum_{l \in R_i} \text{exp}(\hat{\beta}' \boldsymbol{z}_l)} \tag{3.43}$$

For an individual whose survival time was censored, these residuals are always zero. These residuals are usually indicated as missing values to distinguish them from residuals genuinely identical to zero. For an individual whose *death* was observed at t_i, the residual is the difference between the z_j, corresponding to the ith individual, and a weighted average of the values of that variable for all individuals at risk at t_i (76; 77; 79).

Grambsch and Therneau (86) proposed a version of these residuals which is more effective in detecting departures from the assumed model. These residuals are named scaled Schoenfeld residuals and are implemented in the **survival** package of R. Denote $\boldsymbol{r}_i = (r_{1i}, r_{2i}, ..., r_{pi})'$ the vector of Schoenfeld residuals associated to the ith individual. The scaled Schoenfeld residuals, r^*_{ji}, are expressed by

$$\boldsymbol{r}^*_i = k \times var(\hat{\beta}) \boldsymbol{r}_i \tag{3.44}$$

where k is the number of observed *deaths* among the n individuals and $var(\hat{\beta})$ is the covariance matrix of the parameter estimates β_j in the Cox model fitted to the data (76). A plot of the scaled Schoenfeld residuals against the survival times allows to verify if the residuals are equally distributed along the time. If the hypothesis of proportionality of risks is satisfied it should not exist a systematic trend. The interpretation is easier when one adds a line, such as a spline, to better visualize the trend. A spline is a smooth non-parametric function, which draws a line that carries the points density. Besides the visual analysis, it is possible to test the existence of linear correlation between the time and the residuals. Under the null hypothesis, of correlation equal to zero, the test statistic has a χ^2_1 distribution. If the null hypothesis is not rejected, the assumption of proportionality of risks is sustained (79). The test for each covariate is based on a regression

$$\beta_k(t) = \beta_k + \theta_k U_k(t), \qquad k = 1, ..., p \tag{3.45}$$

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

Martingale Residuals

The martingale residuals are very helpful in determining the functional form of the covariates as well as detecting outliers.

When all the variables are fixed, at the beginning of the study, the martingale residual associated to the ith individual, $i = 1, ..., n$, is given by

$$\hat{M}_i = \delta_i - \exp(\hat{\beta}' \boldsymbol{z}_i)\hat{H}_0(t_i) \tag{3.46}$$

where δ_i is the indicator variable, $\hat{M}_i$ represents the difference between the observed number of events for the ith individual in the interval $(0, t_i)$, and the corresponding expected number, estimated based on the fitted model. In fact, the number of expected *deaths* is one if the time t_i is not censored and zero if t_i is censored, i.e., equal to δ_i. r_i is an estimate of $H(t_i)$, which can be interpreted as the expected number of *deaths* in $(0, t_i)$ since only one individual is considered.

These residuals are characterized by great assymetry and values in the interval $(-\infty, 1)$, where the negative values corresponds to the residuals for the censored observations. In large samples, the residuals are not correlated and have an expected value equal to zero, when calculated for the true (unknown) vector of β parameters, $\sum_{i=1}^{n} \hat{M}_i = 0$.

The analysis of the martingale residuals will reveal poorly adjusted individuals, i.e., individuals that *died* too soon or too late, compared with other individuals with similar characteristics. These individuals are designated outliers. To detect its existence, a visual representation of the residuals against the index of each individual is made.

The graphic representation of the residuals against a covariate indicates if the functional form of the covariate is appropriate or if the functional form should be transformed. The simplest approach consists in representing in a plot the martingale residuals of the fitted null model (without covariates) against the values of each one of the covariates included in the model. To facilitate the interpretation of the plot, a smooth curve is added, usually a curve obtained by LOWESS (Locally Weighted Scatterplot Smoother). If the correct model for the covariate z_j is $\exp(f(z_j)\beta_j)$ for some smooth function f, then the LOWESS curve for the covariate z_j will show the form of f. If the curve is linear, no transformation is needed (76).

3.2.5 Collinearity

When at least one of the predictors can be predicted by the remaining, the standard error of the coefficient estimates can be inflated, reducing power to the corresponding tests. On the other hand, collinearity difficults to estimate and interpret parameters, since the data has few

information about the effect of changing one variable keeping other variable, highly correlated, constant.

One way to test the existence of collinearity is through VIF (Variance Inflation Factor). VIF is defined as

$$\frac{1}{(1-R_i^2)}$$

where R_i^2 is the squared coefficient of multiple correlation between the variable i and the remaining variables. VIF provides information about how much larger the standard error is, compared with what it would be if that variable was uncorrelated with the other predictor variables in the model. Values below 10 are an indication of no problem of collinearity (87).

3.2.6 Measures of explained variation

The proportion of variation in a response variable that is explained by a fitted model is often used in statistial modelling to summarise the fit of the model. A number of measures of explained variation have been proposed for use in modelling survival data however, no particular statistic can be recommended for general use. Moreover, few of these options are implemented in software packages.

One global measure that evaluates the fit of a model informs about the proportion of variation explained by the covariates. R^2 can be given by

$$R^2 = 1 - \exp\left(\frac{2(log\mathcal{L}(0) - log\mathcal{L}(\hat{\beta}))}{n}\right) \tag{3.47}$$

where $\mathcal{L}(0)$ is the likelihood function of the null model and $\mathcal{L}(\hat{\beta})$ is the likelihood function of the fitted model. R^2 can be interpreted as the explanatory power of the covariates in the survival time. In survival analysis, the values of R^2 are small. Values for the explained variability below 0.5 are common. Obtaining values of 1 ou close of 1 are unsual since that would mean that the model has predicted the exact time until the ocorrence of the event for each individual.

Another useful measure is the index of concordance C. This measure is used to evaluate the discriminatory power and the predictive accuracy of the Cox model. Concordance implies that, when randomnly selecting two observations, the one with smaller survival time is also the one with smaller estimated risk by the Cox model.

3.3 Software

The statistical analysis was performed using R, version 3.1.0 (88).

An appendix with the relevant code to perform this analysis is provided. The aim is to briefly describe the R code used, including function of libraries available on CRAN (The Comprehensive R Archive Network).

Chapter 4

Evaluation of risk factors for time to recurrence

4.1 Exploratory Analysis

In Portugal, between 1 January 2002 and 31 December 2009, 9882 persons were diagnosed with TB. However, in 1127 cases this was not the first episode of disease and therefore they were eliminated from this analysis. Of the remaining 8755 individuals, 391 died during the first treatment and were excluded since the initial date considered, in this study, is the end of the first treatment. The final dataset consists of 8364 individuals.

All cases

The number of cases of TB per year seems to follow a steady slight increase, with 2008 being the year with more new cases of TB, as seen in figure 4.1. One of the reasons for the decrease in 2009 could be the lack of data. Perhaps the system was not updated on time and therefore, not all individuals diagnosed in 2009 were in the database.

Patients were aged between 0 and 100 years at the end of the first treatment. More than 75% of the individuals have between 20 and 60 years old. Less than 1% are children with less than 10 years or elders with more than 90 years. Figure 4.2 represents the distribution of the age of the individuals.

For the period studied, the mean age is 43 years and the median is 40 years. For males the median age is 42 years and for females is 35 years, as shown in figure 4.3.

The majority of the patients (89%) was born in Portugal, while 10% were born in a high-risk country, such as Angola, Brazil, Cape Verde, China, South Africa, Mozambique, etc. Figure 4.4 represents the number of years that immigrants have been living in Portugal before being diagnosed with TB. Of those, 42% were diagnosed on the first year, which could be a sign that they were infected in their home countries, since the latency period before presenting active disease is around two years.

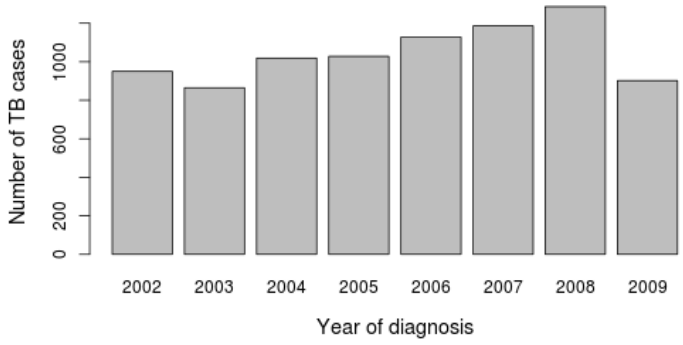

Figure 4.1: Number of cases diagnosed between 2002 and 2009

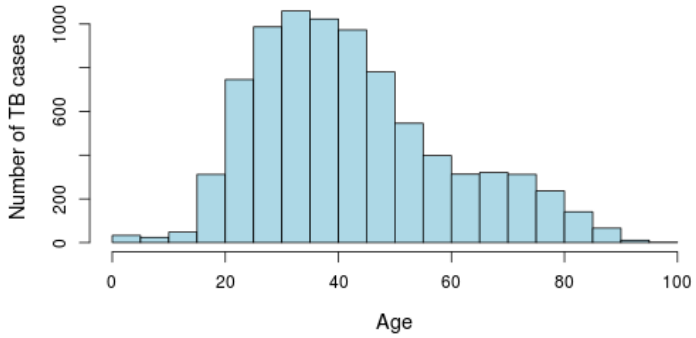

Figure 4.2: Distribution of the age of all the individuals

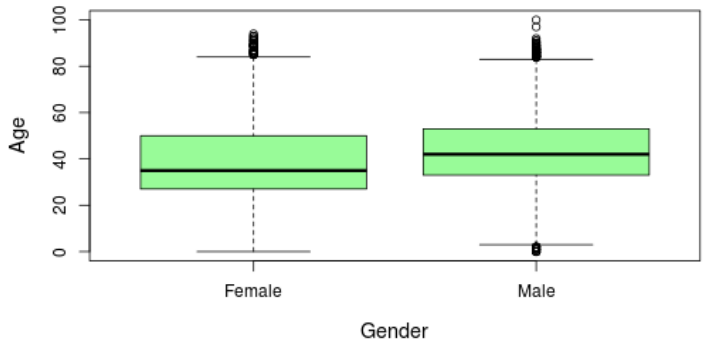

Figure 4.3: Boxplot of age according to gender

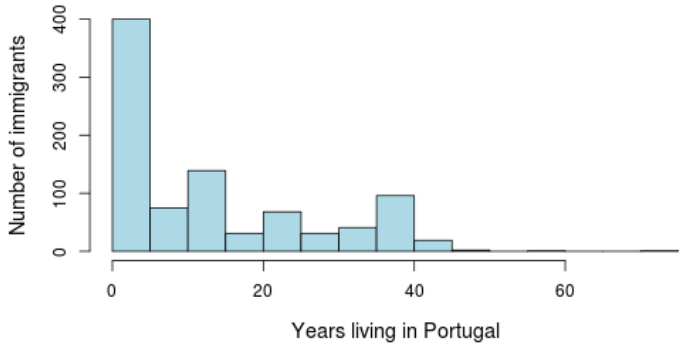

Figure 4.4: Distribution of the number of years immigrants have been living in Portugal

More than half of the individuals (51%) have at least one of the considered risk factors (health care job, alcoholism, smoking, drug problem, in prison or community residence, homeless, unemployed or has HIV or diabetes).

There is a higher proportion of males with TB than females. Table 4.1 explores this difference by showing the number of cases in each category of the variables representing the risk factors that were considered. The amount of missing information is also displayed in this table. For each risk factor considered, the main idea is to compare the proportion of individuals that belong to the category "Yes" with the proportion of individuals with missing information, for each gender.

Less than 3% of the individuals are health care workers, however this information is missing for almost 10%. Through table 4.1 it is possible to see that, among health care workers, almost 60% are female. Around 10% to 15% are either alcoholic, smokers, unemployed or have HIV. Once again, through the table 4.1 it is possible to see that around 76% to 93% of the alcoholic, smokers, unemployed and HIV positive are males. The percentage of drug addicts is 8%. Among the drugs addicts, 87% are males. Only 5% of the individuals have diabetes. Among individuals with diabetes 69% are males. A very small percentage of patients (1% to 2%) are in prison, are homeless or live in a community residence. However, the percentage of missing information for these variables is higher (6%), which can indicate that the real number is actually different. The percentage of men is always superior to the percentage of woman, exept for the variable Job. The proportion of males in the category "Yes" of the variables representing the risk factor is equivalent to proportion of males with missing information.

The median time of follow-up is 1303 days and the mean time is 1383 days. The follow-up time ranges between 1 day and 3049 days. The distribution of the follow-up time is shown in figure 4.5.

Individuals with a recurrent episode

Only a small number of patients (145) had **a recurrent episode** during the time studied. The proportion of men is much higher (77%) than the proportion of women.

Variables		Female	Male	Total
Job	Yes	135 (57.4%)	100 (42.6%)	235 (2.8%)
	No	2228 (30.4%)	5098 (69.6%)	7326 (87.6%)
	Missing	290 (36.1%)	513 (63.9%)	803 (9.6%)
Alc	Yes	94 (7%)	1253 (93%)	1347 (16.1%)
	No	2461 (37.9%)	4032 (62.1%)	6493 (77.6%)
	Missing	98 (18.7%)	426 (81.3%)	524 (6.3%)
Smk	Yes	120 (13%)	802 (87%)	922 (11%)
	No	2442 (34.8%)	4576 (65.2%)	7018 (83.9%)
	Missing	91 (21.5%)	333 (78.5%)	424 (5.1%)
Drugs	Yes	88 (13%)	586 (87%)	674 (8.1%)
	No	2468 (34.4%)	4710 (65.6%)	7178 (85.8%)
	Missing	97 (19%)	415 (81%)	512 (6.1%)
Prison	Yes	7 (13.7%)	44 (86.3%)	51 (0.6%)
	No	2557 (32.6%)	5288 (67.4%)	7845 (93.8%)
	Missing	89 (19%)	379 (81%)	468 (5.6%)
Commu	Yes	45 (24%)	143 (76%)	188 (2.2%)
	No	2511 (32.7%)	5176 (67.3%)	7687 (91.9%)
	Missing	97 (19.8%)	392 (80.2%)	489 (5.9%)
Hmless	Yes	11 (12.1%)	80 (87.9%)	91 (1.1%)
	No	2555 (32.7%)	5256 (67.3%)	7811 (93.4%)
	Missing	87 (18.8%)	375 (81.2%)	462 (5.5%)
Unemp	Yes	260 (23.2%)	863 (76.8%)	1123 (13.4%)
	No	2393 (33%)	4848 (67%)	7241 (86.6%)
HIV	Yes	178 (21.3%)	658 (78.7%)	836 (10%)
	No	2475 (32.9%)	5053 (67.1%)	7528 (90%)
Diabetes	Yes	140 (31%)	312 (69%)	452 (5.4%)
	No	2513 (31.8%)	5399 (68.2%)	7912 (94.6%)

Table 4.1: Distribution of number of cases and missing observations by gender and according to several risk factors (n = 8364)

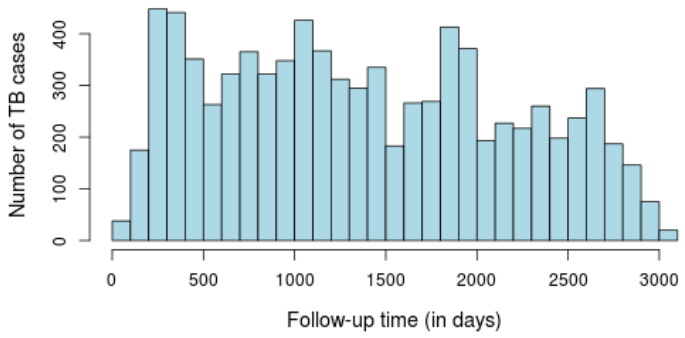

Figure 4.5: Distribution of the follow-up time

Three quarters of the recurrent cases have between 20 and 50 years. Figure 4.6 represents the distribution of the age of recurrent individuals.

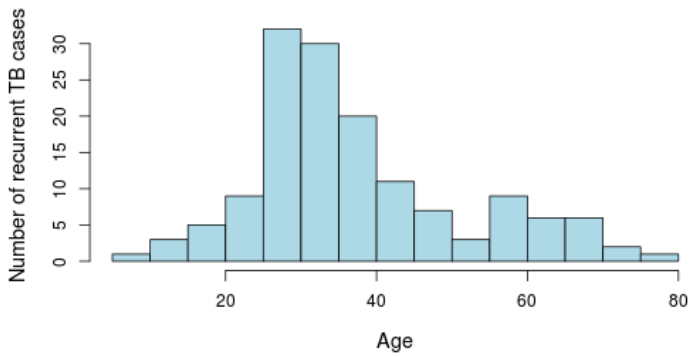

Figure 4.6: Distribution of the age of recurrent cases

The mean age of the patients is 38 years and the median is 34 years. For males the median age is 37 and for females is 29 years. More than 90% of these were born in Portugal.

More than half (67.6%) of those who had a recurrent episode have at least one of the considered risk factors (health care job, alcoholism, smoking, drug problem, in prison, living in a community residence, homeless, unemployed or with HIV or diabetes) (table 4.2).

Among individuals who had a recurrent episode during the time of the study between 15 % to 30 % are either alcoholics, smokers, unemployed, have HIV or drug addicts. Less than 1% have diabetes. Between 1% to 6% are health care workers, homeless, in a prison or in a community residence. All the individuals that are in prison, lives in a community residence or are homeless are males. Between 85% to 98% of the individuals with HIV, smokers, alcoholics, unemployed or drug addicts are males. The variables (health care workers, alcoholic, smoker, drugs, prison, community and homeless) have between 5% to 15% of missing information. The percentage of men is always superior to the percentage of woman, except for the variable Job. The proportion of males in the category "Yes" of the variables representing the risk factor is equivalent to proportion of males with missing information.

Most of the recurrent episodes (61%) happened within the first two years after the end of the first episode. Only around 6% of the episodes happened 5 years or more after the end of the first episode. Figure 4.7 illustrates the distribution of time to a recurrent episode.

Figure 4.8 shows how long it takes for the recurrent episode to occur depending on the previous outcome. Almost all the cases that have defaulted have another episode of TB in the following two years (91%) while only a smaller percentage (49%) of the patients with a previously treated episode have another episode in the same period of time.

Variables		Female	Male	Total
Job	Yes	1 (50%)	1 (50%)	2 (1.4%)
	No	29 (23.6%)	94 (76.4%)	123 (84.8%)
	Missing	4 (20%)	16 (80%)	20 (13.8%)
Alc	Yes	1 (2.4%)	40 (97.6%)	41 (28.3%)
	No	31 (34.4%)	59 (65.6%)	90 (62.1%)
	Missing	2 (14.3%)	12 (85.7%)	14 (9.6%)
Smk	Yes	3 (7.5%)	37 (92.5%)	40 (27.6%)
	No	29 (30.2%)	67 (69.8%)	96 (66.2%)
	Missing	2 (22.2%)	7 (77.8%)	9 (6.2%)
Drugs	Yes	1 (4.2%)	23 (95.8%)	24 (16.6%)
	No	31 (29.5%)	74 (70.5%)	105 (72.4%)
	Missing	2 (12.5%)	14 (87.5)	16 (11%)
Prison	Yes	0 (0%)	4 (100%)	4 (2.7%)
	No	33 (25.6%)	96 (74.4%)	129 (89%)
	Missing	1 (8.3%)	11 (91.7%)	12 (8.3%)
Commu	Yes	0 (0%)	8 (100%)	8 (5.5%)
	No	32 (25.8%)	92 (74.2%)	124 (85.5%)
	Missing	2 (15.4%)	11 (84.6%)	13 (9%)
Hmless	Yes	0 (0%)	6 (100%)	6 (4.1%)
	No	33 (25.8%)	95 (74.2%)	128 (88.3%)
	Missing	1 (9%)	10 (91%)	11 (7.6%)
Unemp	Yes	5 (14.3%)	30 (85.7%)	35 (24.1%)
	No	29 (26.4%)	81 (73.6%)	110 (75.9%)
HIV	Yes	5 (12.2%)	36 (87.8%)	41 (28.3%)
	No	29 (27.9%)	75 (72.1%)	104 (71.7%)
Diabetes	Yes	0 (0%)	1 (100%)	1 (0.7%)
	No	34 (23.6%)	110 76.4%	144 (99.3%)

Table 4.2: Distribution of number of cases and missing observations by gender and according to several risk factors (n = 145)

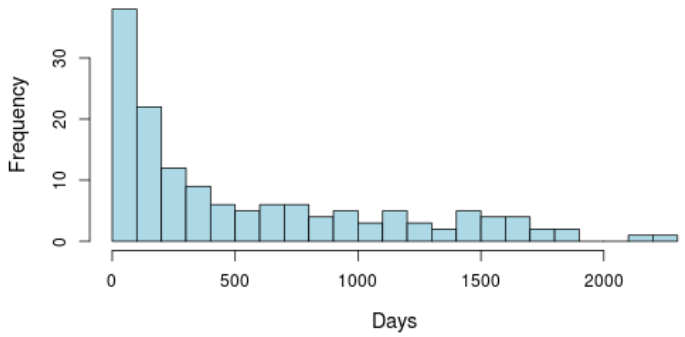

Figure 4.7: Distribution of time to a recurrent episode

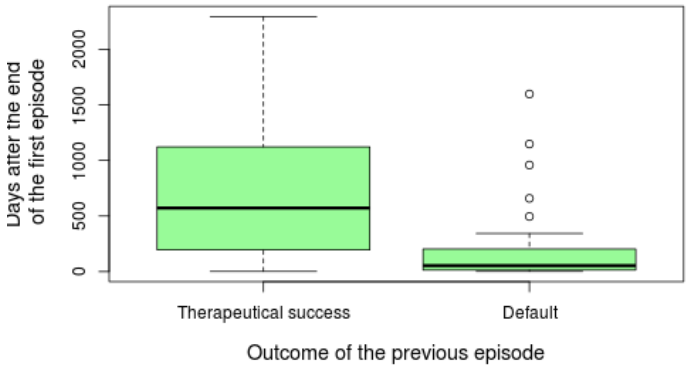

Figure 4.8: Boxplot of time (in years) after the end of the first episode until recurrence, according to the previous outcome

Individuals with censored observations

Most of the patients (8219) **did not have a recurrent episode** during the time studied.

The majority (66%) of the censored observations correspond to individuals that are between 20 and 50 years old. Less than 1% of the cases are less than 10 years or more than 90 years. Figure 4.9 represent the distribution of the age for individuals with censored observations.

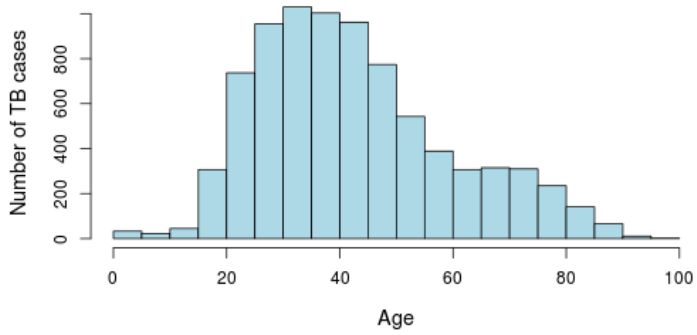

Figure 4.9: Distribution of the age for individuals with censored observations

The mean age is 43 years and the median is 40 years. For males the median age is 42 and for females is 35 years. Around 10% of the censored observations corresponds to individuals that were not born in Portugal.

Half (50%) of the individuals with censored observations have at least one of the considered risk factor (health care job, alcoholic, smoker, drug problem, in prison or community residence, homeless, unemployed or has HIV or diabetes) (table 4.3).

Among individuals that did not have an episode during the time of the study, 8% to 16% are

alcoholics, smokers, unemployed, have HIV or are drug addicts. Less than 5% are health care workers, have diabetes, are homeless or lives in a prison or in a community residence. As seen in table 4.3, around 60% of the health workers are women. Among individuals with diabetes, HIV, unemployed, smokers, alcoholics, drug dependents, homeless, living in a community residence or in prison, more than 69% are males. The variables (health care workers, alcoholic, smoker, drugs, prison, community and homeless) have less than 10% of missing information. The percentage of men is always superior to the percentage of woman, except for the variable Job. The proportion of males in the category "Yes" of the variables representing the risk factor is equivalent to proportion of males with missing information.

Variables		Female	Male	Total
Job	Yes	134 (57.5%)	99 (42.5%)	233 (2.8%)
	No	2199 (30.5%)	5004 (69.5%)	7203 (87.6%)
	Missing	286 (36.5%)	497 (63.5%)	783 (9.6%)
Alc	Yes	93 (7.1%)	1213 (92.9%)	1306 (15.9%)
	No	2430 (38%)	3973 (62%)	6403 (77.9%)
	Missing	96 (18.8%)	414 (81.2%)	510 (6.2%)
Smk	Yes	117 (13.3%)	765 (86.7%)	882 (10.7%)
	No	2413 (34.9%)	4509 (65.1%)	6922 (84.2%)
	Missing	89 (21.4%)	326 (78.6%)	415 (5.1%)
Drugs	Yes	87 (13.4%)	563 (86.6%)	650 (7.9%)
	No	2437 (34.5%)	4636 (65.5%)	7073 (86.1%)
	Missing	95 (19.1%)	401 (80.9)	496 (6%)
Prison	Yes	7 (14.9%)	40 (85.1%)	47 (0.6%)
	No	2524 (32.7%)	5192 (67.3%)	7716 (93.9%)
	Missing	88 (19.3%)	368 (80.7%)	456 (5.5%)
Commu	Yes	45 (25%)	135 (75%)	180 (2.2%)
	No	2479 (32.8%)	5084 (67.2%)	7563 (92%)
	Missing	95 (20%)	381 (80%)	476 (5.8%)
Hmless	Yes	11 (13%)	74 (87%)	85 (1%)
	No	2522 (32.8%)	5161 (67.2%)	7683 (93.5%)
	Missing	86 (19.1%)	365 (80.9%)	451 (5.5%)
Unemp	Yes	255 (23.4%)	833 (76.6%)	1088 (13.2%)
	No	2364 (33.1%)	4767 (66.9%)	7131 (86.8%)
HIV	Yes	173 (21.8%)	622 (78.2%)	795 (9.7%)
	No	2446 (32.9%)	4978 (67.1%)	7424 (90.3%)
Diabetes	Yes	140 (31%)	311 (69%)	451 (5.5%)
	No	2479 (31.9%)	5289 (68.1%)	7768 (94.5%)

Table 4.3: Distribution of number of cases and missing observations by gender and according to several risk factors (n = 8219)

Almost three quarters of the individuals have censored observations 1 year to 5 years after the end of the first episode. Figure 4.10 illustrates the distribution of follow-up time for individuals with censored observations.

Table 4.4 provides a summary of all the variables considered for patients that had a recurrent episode and for those who did not had.

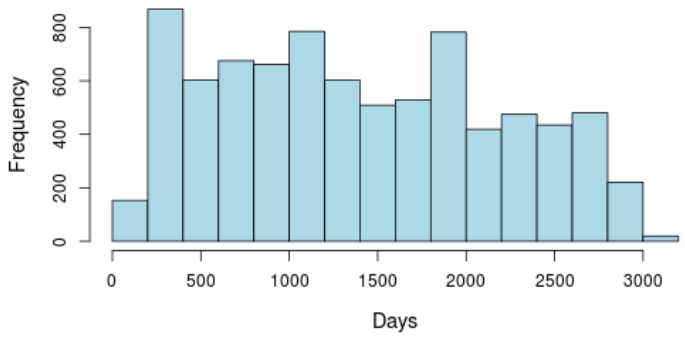

Figure 4.10: Distribution of follow-up time for individuals with censored observations

Table 4.4: Summary of the final dataset

Variable	Category	Recurrent	No-recurrent	Total
Vac	No	32	1317	1349
	Yes	26	2050	2076
	NA	87	4852	4939
CliForm	Pulmonary	122	7533	7655
	Extra pulmonary	23	684	707
	NA	0	2	2
Radio	Normal	10	534	544
	Without Cavitation	60	3198	3258
	With cavitation	59	3950	4009
	NA	16	537	553
Sit	Cured	102	7994	8096
	Default	43	225	268
HIV	No	104	7424	7528
	Yes	41	795	836
Diabetes	No	144	7768	7912
	Yes	1	451	452
NumCo	0	93	6214	6307
	1	47	1736	1783
	+2	5	269	274
Sex	Female	34	2619	2653
	Male	111	5600	5711
age	Mean (IQR)	38.26 (29-44)	43.05 (30-53)	42.97 (30-53)
	Standard Deviation	14.31	17.06	17.02
Origin	Portuguese	131	7313	7444
	Low-risk	1	49	50
	High-risk	13	841	854
	NA	0	16	16

Continue on next page

Table 4.4 – continued from previous page

Variable	Category	Recurrent	No-recurrent	Total
Smk	No	96	6922	7018
	Yes	40	882	922
	NA	9	415	424
Alc	No	90	6403	6493
	Yes	41	1306	1347
	NA	14	510	524
Unemp	No	110	7131	7241
	Yes	35	1088	1123
Drugs	No	105	7073	7178
	Yes	24	650	674
	NA	16	496	512
Prison	No	129	7716	7845
	Yes	4	47	51
	NA	12	456	468
Job	No	123	7203	7326
	Yes	2	233	235
	NA	20	783	803
Commu	No	124	7563	7687
	Yes	8	180	188
	NA	13	476	489
Hmless	No	128	7683	7811
	Yes	6	85	91
	NA	11	451	462
Transf	No	140	7979	8119
	Yes	5	240	245

4.2 Exploratory Analysis of Missing Data

Although most studies ignore missing data and perform only a complete case analysis (even when this is not the most appropriate solution), some studies do not ignore this issue and explore and impute the missing values (48; 89; 90).

As previously shown, the predictor variables have a large amount of missing values. Figure 4.11 shows the frequency of missing data per individual. Only 34% (2890) of the patients had complete data, whereas 66% (5474) of the patients had one or more missing predictors. Most frequently (46%) there was one predictor with missing data per patient.

Despite the large proportion of missing data, this proportion seems to be decreasing over time, as observed in figure 4.12.

This decrease seems to indicate a growing concern with the epidemiology of TB by clinical practitioners. In fact, most of the variables showing this decrease, such as being alcoholic,

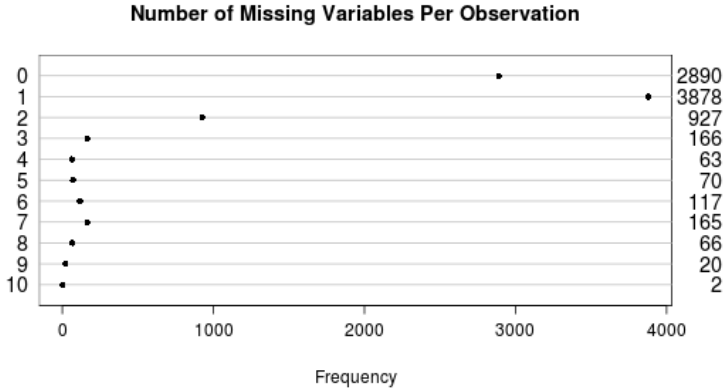

Figure 4.11: Number of missing variables per subject

smoker, drug-addicted or in prison, are known risk factors for the recurrence of TB (29; 30). Pulmonary cavities observed in radiological exams have been associated with recurrence of TB (29; 30), thus, the observed decrease in missing data over time on this variable may be also an indication of an increased concern in understanding the disease. Interestingly, missing data for the variable symptoms does not seem to decrease over time, remaining almost constant. This is not unexpected since this variable depends on retrospective self-diagnosis from the patients, which is not expected to improve over the years.

As discussed in section 3.1, the missingness pattern can be classified as monotone or nonmonotone. The purpose of figure 4.13 is to shed light on the pattern of missingness in this dataset. A monotonic pattern can also be identified by the no return, i.e., once a subject dropped out he will drop out forever, not returning to the study. Whereas, in a non-monotonic pattern the subject may come back or be missing again.

In order to assess the type of missing data of the dataset, a test proposed by Little et al (91) to explore the validity of the MCAR hypothesis was used. The null hypothesis states that the data are MCAR. To implement this test, the function `LittleMCAR`, in R, was used (92). The test results indicate that there are significant differences between the missing values and the observed values, so there is evidence that the data is not MCAR (chi-sq = 11031.29, df= 3777, p-value = 0).

Another interesting analysis is to compare the survival curves of groups of individuals with and without missing information for each predictor separately (figure 4.14).

Through the analysis of figure 4.14, the Kaplan-Meier estimates for the variables radiology and symptoms do not appear to show visual differences. However, some visual differences are observed in the plots for the variables vaccine, alcohol, drugs, smoke, prison and homeless. The log-rank and Peto-Peto tests were used to evaluate the existence of significant differences between the curves. There was a significant difference between the survival distributions of patients with and without missing information for vaccine, alcohol, drugs and the remaining variables. The patients with missing values in vaccine (p-value = 0.06), prison (p-value $<$ 0.001), community residence (p-value $<$ 0.001), smoke (p-value $<$ 0.001) and homeless (p-value $<$ 0.001) had a better prognosis, which could lead to an underestimation of the true

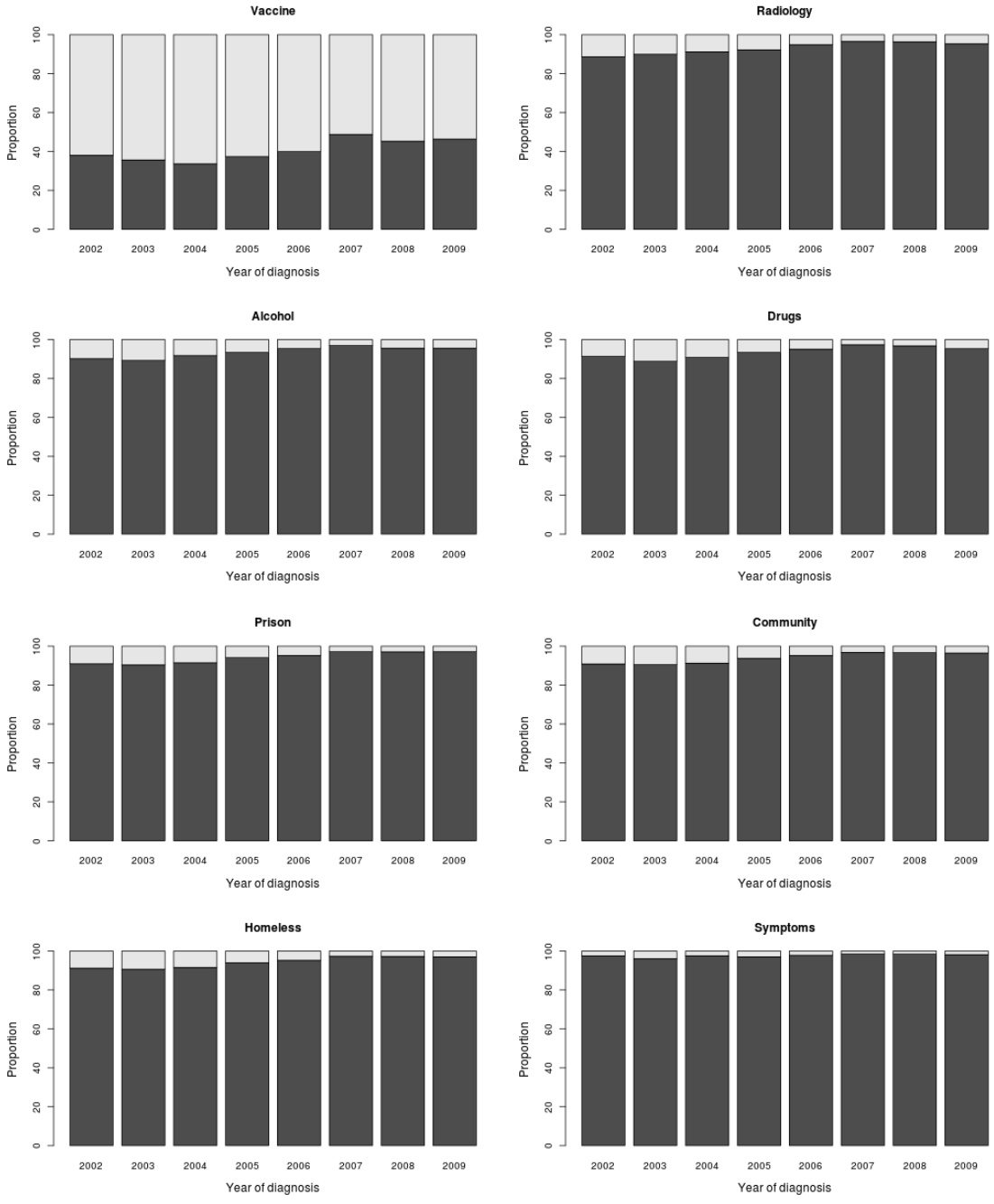

Figure 4.12: Proportion of missing (light gray) and observed data (dark gray) for several predictors

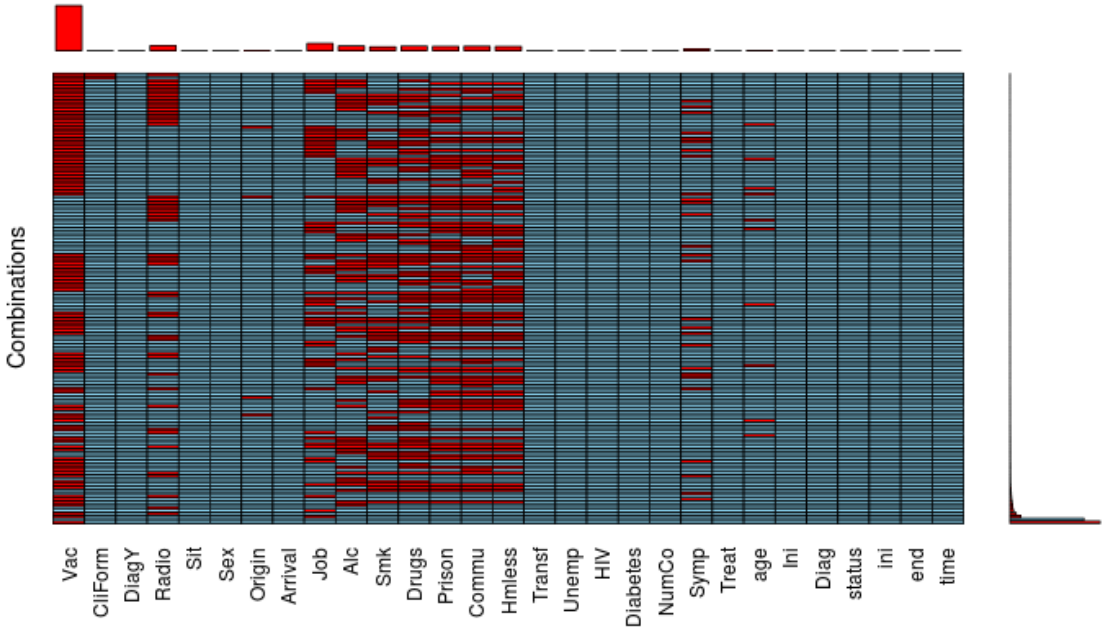

Figure 4.13: Amount of missing data in each variable

survival, if those patients were eliminated from the analysis. The opposite was observed for the variables alcohol (p-value < 0.001) and drugs (p-value < 0.001), with worse prognosis for patients with missing values. However, some caution should be taken when interpreting these effects, since this approach is univariate and, therefore, not adjusted for the other variables.

Unfortunately, there is no standard statistical test to determine if data are MAR. In fact, it is not possible to distinguish between MAR and MNAR using observed data (54). To gain knowledge into the issue of whether data can be considered MAR, a logistic regression model is used for the outcome of missingness (90; 93) (Table 4.5). However, many variables had a problem of complete separation, also known as monotone likelihood. It is observed usually in small samples with highly predictive risk factors. The phenomenon of separation occurs if the observed and missing values can be separated by a single risk factor or by a combination of risk factors. The problem with this situation is the non-existence of the maximum likelihood estimate (94). To address this problem, the package **logistf** (95) that implements Firth's penalized likelihood logistic regression was used. This analysis indicates that HIV, age, number of previous treatment and country of origin are associated with the existence of missing values in several other variables. These results suggest that the missing data could be MAR and the analysis will be done under this assumption.

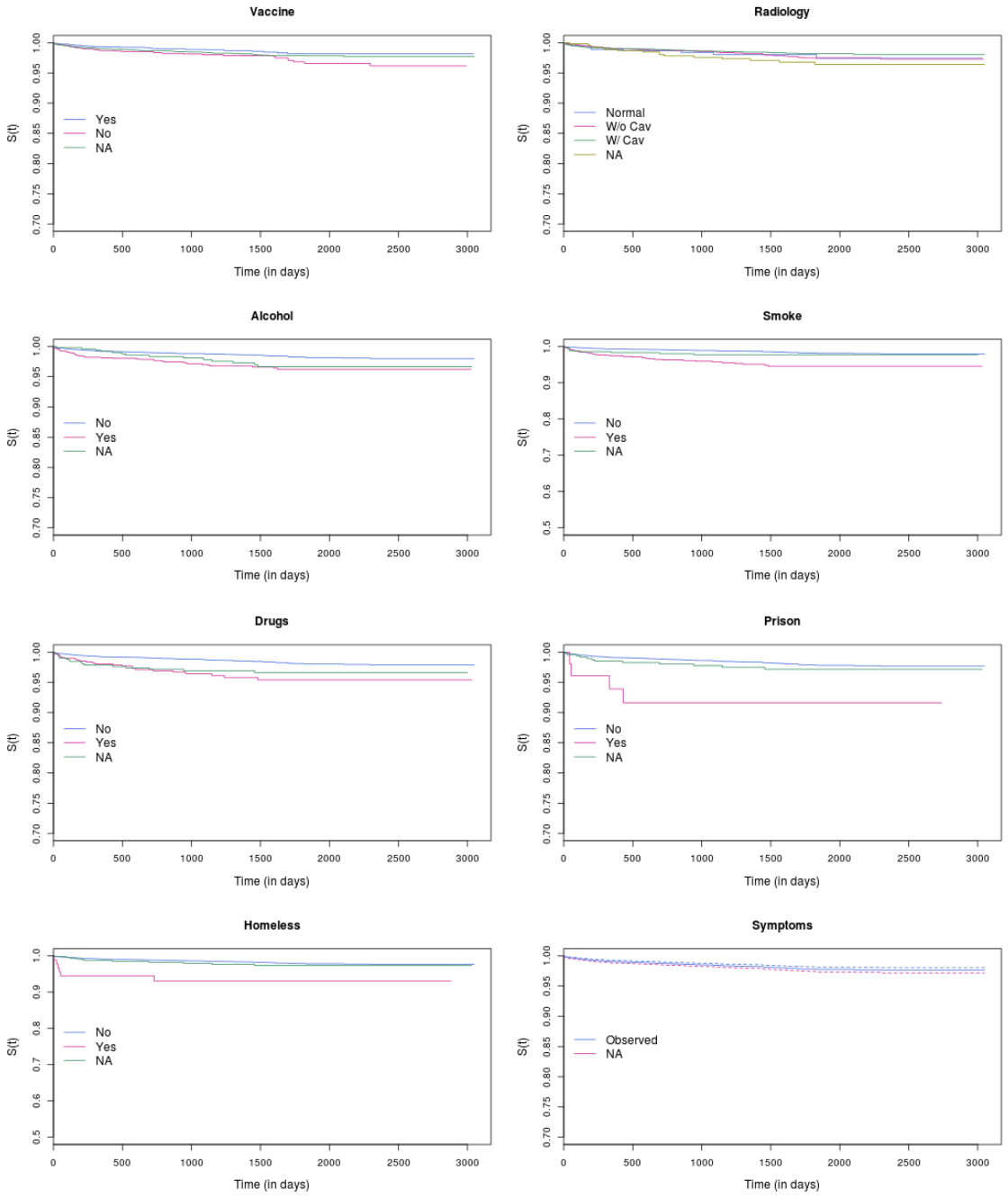

Figure 4.14: Kaplan-Meier estimates of the survivor function for several groups

Outcome	CliForm	Radio	Sit	Sex	Origin	Job	Alc	Drugs	Transf	HIV	NumCo	Symp	Treat	age	Survival
Vac	-	X	-	-	X	-	X	-	X	X	-	X	X	X	>>
Radio	X	-	-	-	-	-	-	-	-	X	-	-	-	-	-
Origin[a]	-	-	-	-	-	X	-	-	-	-	-	-	-	-	-
Job[a]	-	-	-	-	X	-	-	-	-	-	-	-	-	-	-
Alc[a]	-	-	-	-	X	-	-	-	-	-	-	-	-	-	<<
Smk[a]	-	-	-	-	-	-	-	-	-	-	-	-	X	-	>>
Drugs[a]	-	-	-	-	-	-	-	-	-	-	X	-	-	-	<<
Prison[a]	-	-	-	-	-	-	-	-	-	-	-	-	-	X	>>
Commu	Not enough data in the dataset to do the logistic regression														>>
Hmless	Not enough data in the dataset to do the logistic regression														>>
age[a]	-	-	-	-	-	-	-	-	-	-	X	-	-	-	-

Table 4.5: Association between missingness and other potential risk factors. [a] - Due to a separation problem, the package **logistf** was used.

4.3 Non-parametric Inference

For categorical variables time until the second episode of TB was analyzed using the Kaplan-Meier estimator and the log-rank and Peto-Peto tests. The univariate analysis was performed on the complete cases in the original dataset (n=2890).

The Kaplan-Meier estimator was used to estimate the survivor function of the time from the end of the first episode of TB until the beginning of the second episode of TB for all individuals in the dataset.

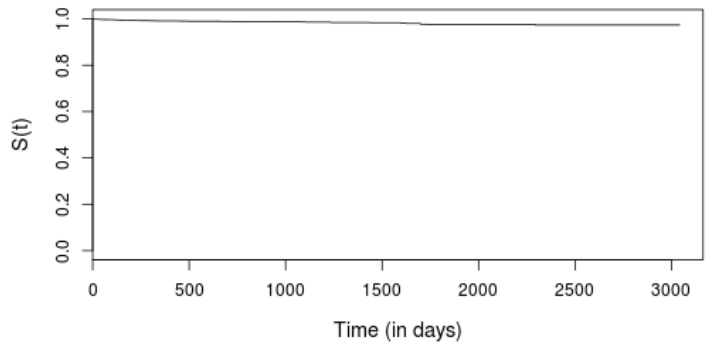

Figure 4.15: Kaplan-Meier estimate of the survivor function of time until the begining of the second episode of TB

Figure 4.15 shows that the survival curve slowly decreases along the years. Seven years after the end of the first treatment, it is estimated that less than 5% will have developed a second episode of disease. Therefore it is not possible to estimate the median or the most used quantiles. Table 4.6 presents a brief summary.

Years	Number at risk	Number of events	Probability of non-recurrence
1	2890	17	0.994
2	2843	11	0.990
3	2250	4	0.989
4	1753	5	0.986
5	1243	4	0.983
6	910	4	0.978
7	598	1	0.977

Table 4.6: Distribution of the number of events since the end of the first treatment

Next, the Kaplan-Meier estimate of the survivor function was obtained for each variable.

A visual analysis of figure 4.16 suggests that individuals without a BCG vaccine have a slightly higher probability to have another episode of TB than a vaccinated individual. It is estimated that individuals that had defaulted the previous episode have a higher probability to have another episode, compared with those who had a successfull treatment. Individuals with

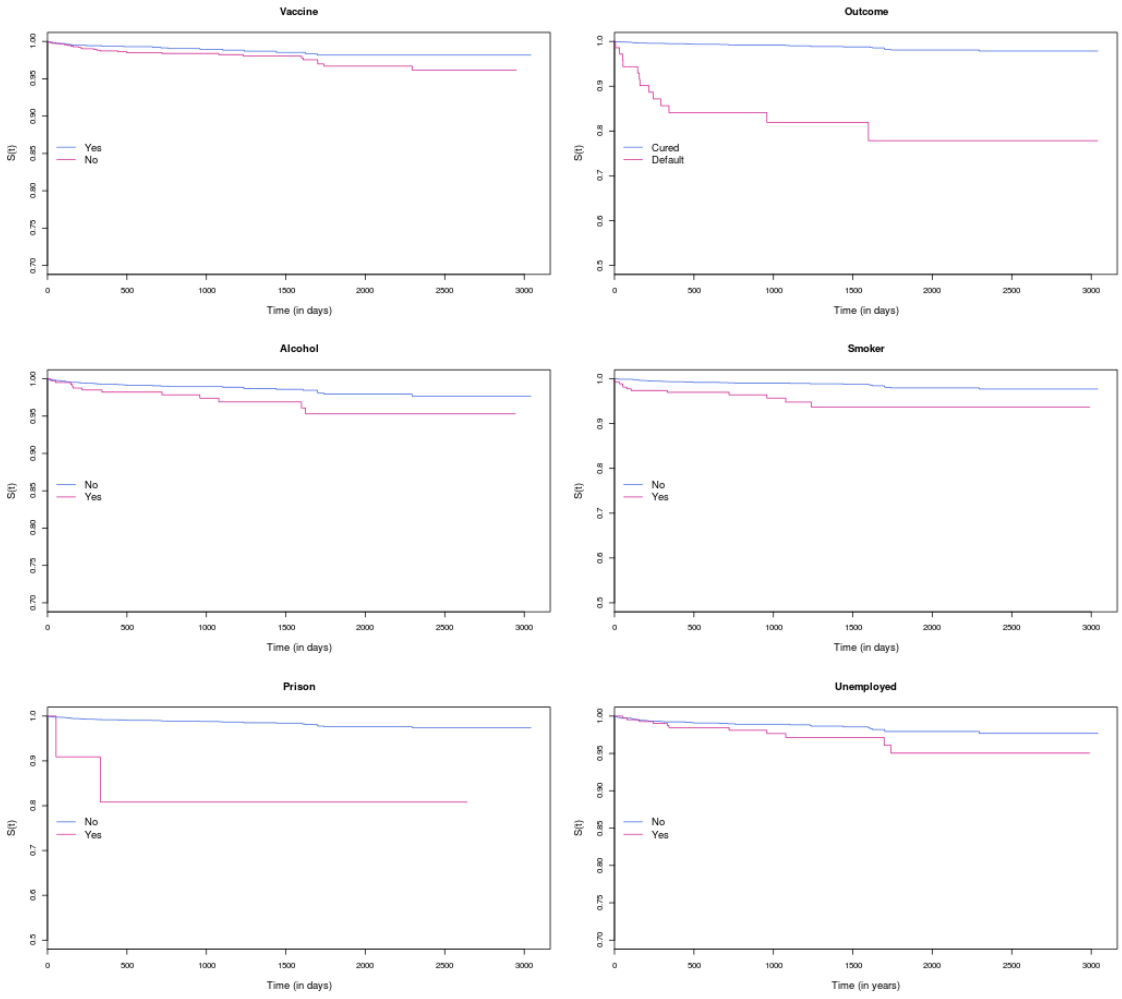

Figure 4.16: Kaplan-Meier estimate of the survivor function for each category of several risk factors, considering complete cases

an alcoholic dependency, smokers, in prison or unemployed also have a higher risk to have another episode than someone who does not drink, smokes, is not in prison nor unemployed, respectively. The analysis was repeated for the remaining variables, although the plots are not presented here. Individuals with an extrapulmonary form of TB have a higher risk to have another episode than individuals with pulmonary form of TB. Individuals born in Portugal have a smaller risk of recurrence than foreigners. No visual difference was found between the survival curves for gender and different types of radiology. Individuals with a drug dependence, living in a community residence, homeless or with HIV are more risk to have another episode of TB. Whereas, individuals that were transfered to another hospital, that have Diabetes and have two more diseases have less risk to have another episode episode of TB.

To analyze how meaningfull were the differences suggested by the graphics the log-rank test and Peto-Peto test were used (table 4.7).

Variables	Test log-rank		Test Peto-Peto		
	$\chi^2(df)$	p-value	$\chi^2(df)$	*p*-value	
Vac	4 (1)	0.04	4 (1)	0.04	**
CliForm	3.7 (1)	0.05	3.7 (1)	0.05	**
Radio	2.7 (2)	0.26	2.7 (2)	0.26	
Sit	141 (1)	0.00	142 (1)	0.00	***
Sex	1.5 (1)	0.22	1.5 (1)	0.22	
Origin	1.8 (2)	0.41	1.8 (2)	0.41	
Job	0.1 (1)	0.70	0.1 (1)	0.71	
Alc	6.3 (1)	0.01	6.3 (1)	0.01	***
Smk	19.4 (1)	0.00	19.5 (1)	0.00	***
Drugs	9.9 (1)	0.00	9.9 (1)	0.00	***
Prison	21.7 (1)	0.00	21.9 (1)	0.00	***
Commu	5.5 (1)	0.02	5.5 (1)	0.02	**
Hmless	0.9 (1)	0.34	0.9 (1)	0.33	
Transf	1.1 (1)	0.29	1.1 (1)	0.29	
Unemp	4.5 (1)	0.03	4.5 (1)	0.03	**
HIV	33.3 (1)	0.00	33.4 (1)	0.00	***
Diabetes	2.1 (1)	0.15	2.1 (1)	0.15	
NumCo	6.4 (2)	0.04	6.4 (2)	0.04	**

Table 4.7: Log-rank test and Peto-Peto test results for each variable. Signif. codes: < 0.01 '***' 0.05 '**' 0.1 '*'

The survival curves for radiology, gender, country of birth, job, homeless, transfered and diabetes did not displayed significant differences, while for all the remaining variables, the survival curves displayed significant differences. These results are somewhat in accordance with the visual inspection of the survival curves.

4.4 Complete Case Analysis

In presence of missing data, one of the most common strategy is to consider only the individuals without missing observations. However, as previously discussed in section 3.1.1 using this

method if the data is not MCAR will lead to biased results. Here, a Cox model was fitted to the complete case data.

4.4.1 Cox Regression Analysis

After removing the missing values, the dataset for analysis has 2890 observations.

To find which variables were significant using the Cox model, a strategy for selecting variables proposed by Collett (77) was used. The first step consists in performing an univariate analysis, in order to assess the individual influence of each variable. The results obtained are in table 4.8.

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	698.35	-	
Vac	694.41	0.05	**
CliForm	695.47	0.09	*
Radio	695.67	0.26	
Sit	652.98	0.00	***
Sex	696.80	0.21	
Origin	697.00	0.51	
Job	698.19	0.69	
Alc	693.22	0.02	**
Smk	685.37	0.00	***
Drugs	691.51	0.01	***
Prison	691.78	0.01	***
Commu	695.01	0.07	*
Hmless	697.70	0.42	
Transf	696.12	0.14	
Unemp	694.54	0.05	**
HIV	679.88	0.00	***
Diabetes	694.26	0.04	**
NumCo	693.05	0.07	*
age	694.56	0.05	**

Table 4.8: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (CC). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

When using this selection procedure, it is recommended to consider a significance level around 10%. For this reason, the variable CliForm is kept in the model for further analysis. If a strict significance level of 5% was used, CliForm, Commu and NumCo would be eliminated from the analysis. The variables Job, Origin, Hmless, Radio, Sex and Transf were discarded.

The variable Diabetes was also excluded although the significance level is below 10 %. The reason to remove this variable is mainly due to a separation problem, commonly named monotone likelihood when fitting a Cox model. The phenomenon of monotone likelihood is not unsual in highly censored samples with strong predictive covariates (96). Monotone likelihood causes parameter estimates to diverge, therefore classical maximum likelihood fails and the confidence intervals cover the whole range of real numbers. The problem is the unability to

obtain the maximum of the likelihood function and it is easily spotted by the value of the Hazard ratio and an insignificant Wald test; in this case, the variable Diabetes is the smallest of all covariate values ($\hat{\beta}$ = -15.82, exp($\hat{\beta}$) = 0.00000013, se($\hat{\beta}$) = 2797, *p*-value = 0.9955, CI (95%) = [0.00; ∞]). To address this issue, the package **coxph** (97) was used since it implements the Firth's penalized maximum likelihood bias reduction for the Cox regression. Although Firth's penalized likelihood method is superior to maximum likelihood analysis, it was still not possible to properly estimate the parameters for the variable Diabetes and for this reason, Diabetes was removed from this analysis.

A Cox model including all the remaining variables was then fitted to the data. Results are presented in table 4.9.

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	619.19	-	
Model without Vac	622.39	0.07	*
Model without CliForm	623.54	0.04	**
Model without Sit	654.79	0.00	***
Model without Alc	621.25	0.15	
Model without Smk	620.12	0.33	
Model without Drugs	620.21	0.31	
Model without Prison	622.72	0.06	*
Model without Commu	620.87	0.19	
Model without Unemp	619.41	0.64	
Model without HIV	624.40	0.02	**
Model without NumCo	620.21	0.60	
Model without age	623.26	0.03	**

Table 4.9: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests (CC). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

The variables that, when omitted did not significantly increased the value of -2 $\log \hat{\mathcal{L}}$, were discarded from the model one by one. This process was repeated every time a variable was removed from the model. The variables that remained in the model were Vac, CliForm, Sit, Prison, HIV and age.

In the third step, the variables that were discarded and not considered in the second step are introduced one by one in the model. If the inclusion of any of the variables reduces significantly the value of -2 $\log \hat{\mathcal{L}}$, the variable is retained in the model. No variable had a significant result. Therefore, the variables Vac, Sit, Prison, HIV, age and CliForm were included in the final Cox model. The results obtained fitting this model to the data can be seen in table 4.10.

The hazard ratio, exp($\hat{\beta}$), was estimated for the variables (table 4.10). Values above 1 indicate higher risk while, value between 0 and 1 point out to a protection effect. Adjusting for the remaining variables, it is estimated that:

- An individual that was not vaccinated has an increase of 80% in the risk of a recurrent episode when compared with a vaccinated individual;
- The risk of recurrence in individuals with an extrapulmonary form of TB is 2.59 times that of individuals with a pulmonary form of TB;

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
Vac1	0.59	1.81	0.32	1.84	0.07	[0.96;3.41]
CliForm1	0.95	2.59	0.45	2.12	0.03	[1.07;6.26]
Sit1	2.56	12.98	0.35	7.41	0.00	[6.59;25.56]
Prison1	2.33	10.25	0.74	3.16	0.00	[2.42;43.36]
HIV1	1.19	3.29	0.36	3.31	0.00	[1.62;6.67]
age	-0.03	0.97	0.01	-2.08	0.04	[0.95;1.00]

Table 4.10: Results using the Cox model (CC)

- The risk of recurrence in individuals that defaulted the previous treatment is 12.98 times that of individuals who completed the treatment;
- The risk of recurrence for an inmate or someone insered in a prison environment is 10.25 times that of someone who is not at the prison;
- The risk of recurrence for someone with HIV is 3.29 times that of someone without HIV;
- Each additional year of age is associated with an estimated 3% decrease in risk. An additional decade corresponds to 2% decrease, exp$(10 * -0.03)$.

The confidence interval for age is quite narrow, capturing only a small range of effect sizes. However, the confidence intervals of Sit, Prison and HIV are wider, capturing a large range of effect sizes. Therefore, the estimates are more imprecise.

4.4.2 Residual Analysis

Schoenfeld Residuals

An important issue when fitting a Cox model is the validity of the proportional risks assumption. Schoenfeld residuals were used to evaluate the proportionality of risks for each covariate. The goal is to assess if the effect of each covariate changes with time.

Besides the visual analysis (figure 4.17), it is possible to test the existence of linear correlation between the residuals and time.

	rho	χ^2	*p*-value
Vac1	0.119	0.687	0.407
CliForm1	-0.024	0.027	0.869
Sit1	-0.215	2.430	0.119
Prison1	-0.022	0.022	0.883
HIV1	0.0423	0.1010	0.751
age	0.163	0.688	0.407
GLOBAL	NA	4.277	0.639

Table 4.11: Test for proportionality of risks (CC)

The results in table 4.11 show the proportionality of the hazard functions, so there is no evidence of non proportional hazards. However, the decision should be based both on the

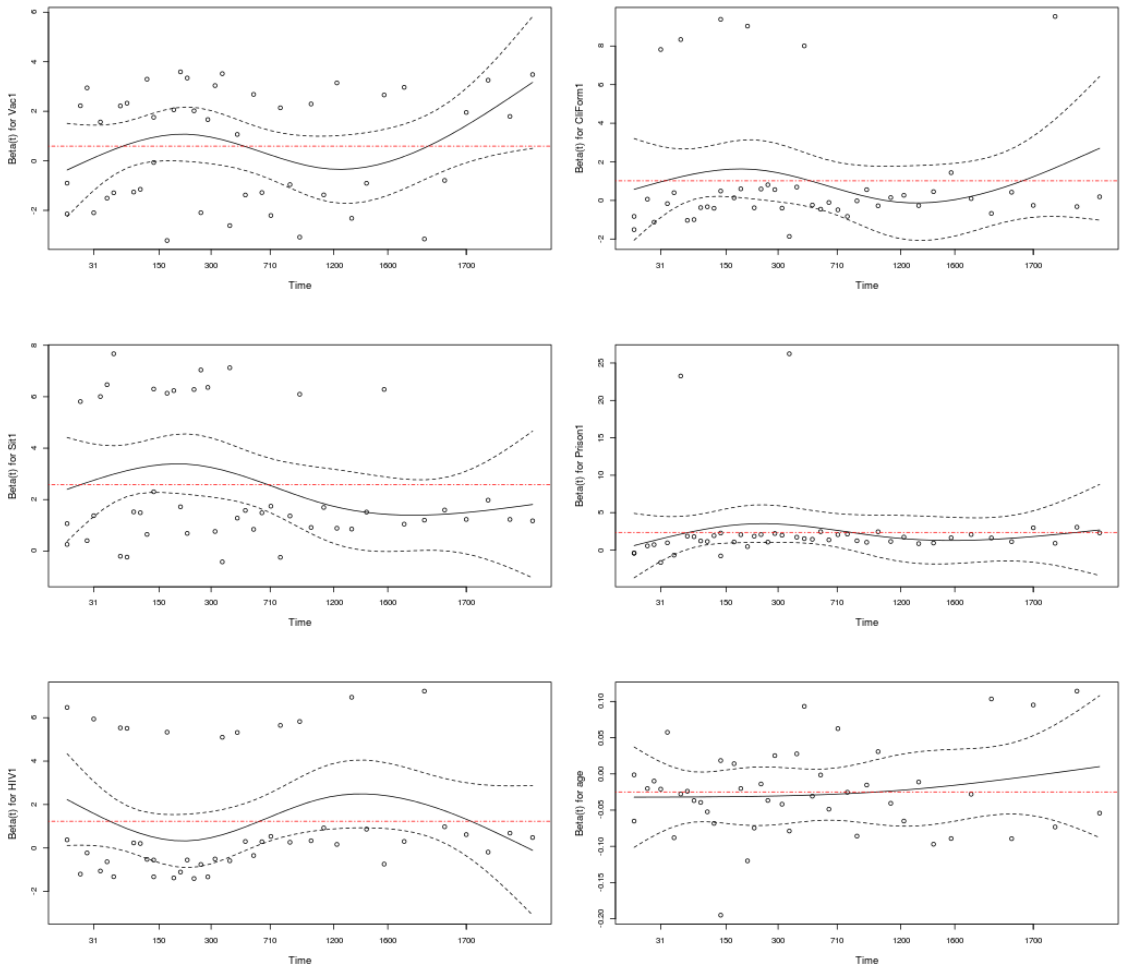

Figure 4.17: Plot of the Schoenfeld residuals (CC)

results of the test and the analysis of the graphics in figure 4.17. A line paralel to the X-axis supports the assumption of proportionality of risks. The smooth line of the variables HIV, Vacine, CliForm and Sit have a oscillatory behaviour, but without a clear trend. Therefore, this is not an indication to reject the hypothesis of proportionality.

Martingale Residuals

The martingale residuals were used in order to find if there are any individuals who were poorly fitted by the model, as well as to investigate the functional form of the continuous variables (figure 4.18).

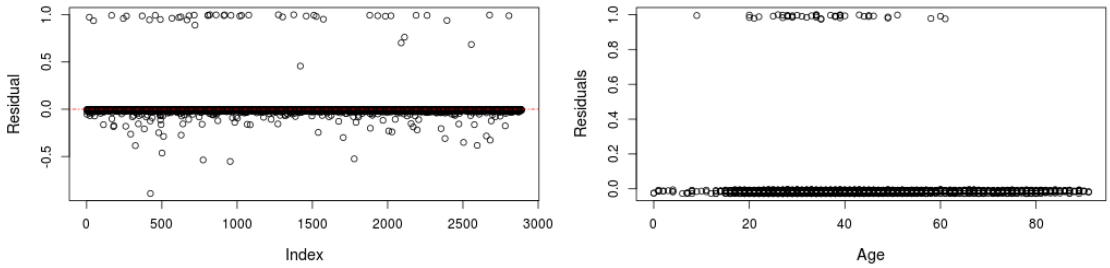

Figure 4.18: Plot of the martingale residuals (CC)

The plot of the martingale residuals versus the index of each individual in figure 4.18, does not present any pattern, corresponding to a good fit since the residuals are evenly distributed above and under zero. Regarding the functional form of the variable age, in figure 4.18, the behaviour of the residuals are linear.

4.4.3 Collinearity

Evaluating the existence of collinearity between the covariates is essential since its existence difficults the estimation of the coefficients. Values of Variance Inflation Factor (VIF) above 10 are considered problematic (87). When the VIF is approximately 1 the covariates are independent (table 4.12).

Vac	CliForm	Sit	Prison	HIV	age
1.180	1.042	1.102	1.021	1.066	1.129

Table 4.12: Values of VIF (CC)

4.5 Complete Case Analysis Without Vaccine

Another approach commonly used in the presence of a variable with a huge amount of missing data, like the variable Vaccine, is to ignore the variable and perform a complete case analysis on the remaining variables. Since the variable Vaccine was removed from this analysis, the complete dataset has 6413 observations.

4.5.1 Cox Regression Analysis

The variables were selecting following the method proposed by Collett (77). The results of the univariate analysis are in table 4.13.

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	1661.823	-	
CliForm	1656.20	0.02	**
Radio	1661.29	0.77	
Sit	1564.31	0.00	***
Sex	1657.79	0.04	**
Origin	1661.47	0.84	
Job	1661.35	0.49	
Alc	1652.05	0.00	***
Smk	1644.12	0.00	***
Drugs	1653.71	0.00	***
Prison	1655.01	0.01	***
Commu	1653.89	0.00	***
Hmless	1658.66	0.08	*
Transf	1661.82	1.00	
Unemp	1657.34	0.03	**
HIV	1644.43	0.00	***
Diabetes	1656.59	0.02	**
NumCo	1657.37	0.11	
age	1656.04	0.02	**

Table 4.13: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (CC without Vaccine). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

The variables Radio, Origin, Job, Transf and NumCo were eliminated from the analysis since they are not significant at a 10% level.

A Cox model including all the remaining variables was fitted to the data. Results are presented in table 4.14.

The variables that, when omitted did not significantly increased the value of -2 $\log \hat{\mathcal{L}}$, were discarded from the model one by one. This process was repeated every time a variable was removed from the model. Afterwards, the variables discarded in the univariate analysis were introduced one by one in the model. Since none of them significantly reduced the value of -2 $\log \hat{\mathcal{L}}$ they were not included in the final model. The results obtained fitting this model to the data can be seen in table 4.15.

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	619.19	-	
Model without CliForm	1532.832	0.043	**
Model without Sit	1609.590	0.000	***
Model without Sex	1529.945	0.2687	
Model without Alc	1532.519	0.051	*
Model without Smk	1529.980	0.262	
Model without Drugs	1530.288	0.211	
Model without Prison	1533.494	0.029	**
Model without Commu	1532.869	0.0417	**
Model without Hmless	1528.726	0.950	
Model without Unemp	1528.786	0.800	
Model without HIV	1530.250	0.216	
Model without Diabetes	1531.846	0.077	*
Model without age	1532.523	0.051	**

Table 4.14: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-value of Likelihood ratio tests (CC without Vaccine). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
CliForm1	0.803	2.233	0.302	2.659	0.008	[0.211;1.395]
Sit1	2.640	14.012	0.226	11.681	0.000	[2.197;3.083]
Alc1	0.532	1.702	0.233	2.283	0.022	[0.075;0.988]
Prison1	1.714	5.550	0.599	2.861	0.004	[0.541;2.887]
Commu1	1.076	2.933	0.395	2.724	0.006	[0.302;1.850]
Diabetes1	-1.425	0.241	1.009	-1.412	0.158	[-3.403;0.554]
age	-0.013	0.987	0.007	-1.857	0.063	[-0.027;0.000]

Table 4.15: Results using the Cox model (CC without Vaccine)

The variable Diabetes is no longer significant at a 10% level. The hazard ratio, $\exp(\hat{\beta})$, was estimated for the variables (table 4.15). Adjusting for the remaining variables, it is estimated that:

- The risk of recurrence in individuals with an extrapulmonary form of TB is 2.23 times that of individuals with a pulmonary form of TB;
- The risk of recurrence in individuals who defaulted the previous episode is 14 times that of individuals who completed the treatment;
- An alcoholic individual has an increase of 70% in the risk of a recurrent episode when compared with someone who does not drink;
- The risk of recurrence for an individual incarcerated is 5.55 times that of someone who is not incarcerated;
- The risk of recurrence for someone living in a community residence is 2.93 times that of someone that does not live in a residence community;
- Each additional year of age is associated with an estimated 1 % decrease in risk. An additional decade corresponds to 12% decrease.

It is important to highlight that all confidence intervals are quite large, indicating more uncertainty about the true estimate of the coefficient. It is also important to note that the variable HIV was not considered significant in this model, which is unexpected due to the vast literature about its significance.

4.5.2 Residual Analysis

Schoenfeld Residuals

An important analysis is to evaluate the validity of the proportional risk assumption, i.e., to observe if the effect of a covariate changes with time. A visual analysis was performed (figure 4.19) and a test to determine the correlation between the residuals and the time (table 4.16).

	rho	χ^2	*p*-value
CliForm1	0.147	2.098	0.147
Sit1	-0.465	22.681	1.9e-06
Alc1	0.145	2.227	0.136
Prison1	-0.096	0.991	0.319
Commu1	-0.054	0.294	0.588
Diabetes1	-0.1309	1.674	0.196
age	-0.188	3.323	0.068
GLOBAL	NA	28.811	1.6e-04

Table 4.16: Test for proportionality of risks (CC without Vaccine)

As seen in table 4.15, the variable Sit have a strong effect on the risk for recurrence. The test suggests that this effect is significantly non proportional. However, if one observes the plot in figure 4.19, the variation in $\hat{\beta}(t)$ is not far from the estimate of $\hat{\beta}$. It is also important to

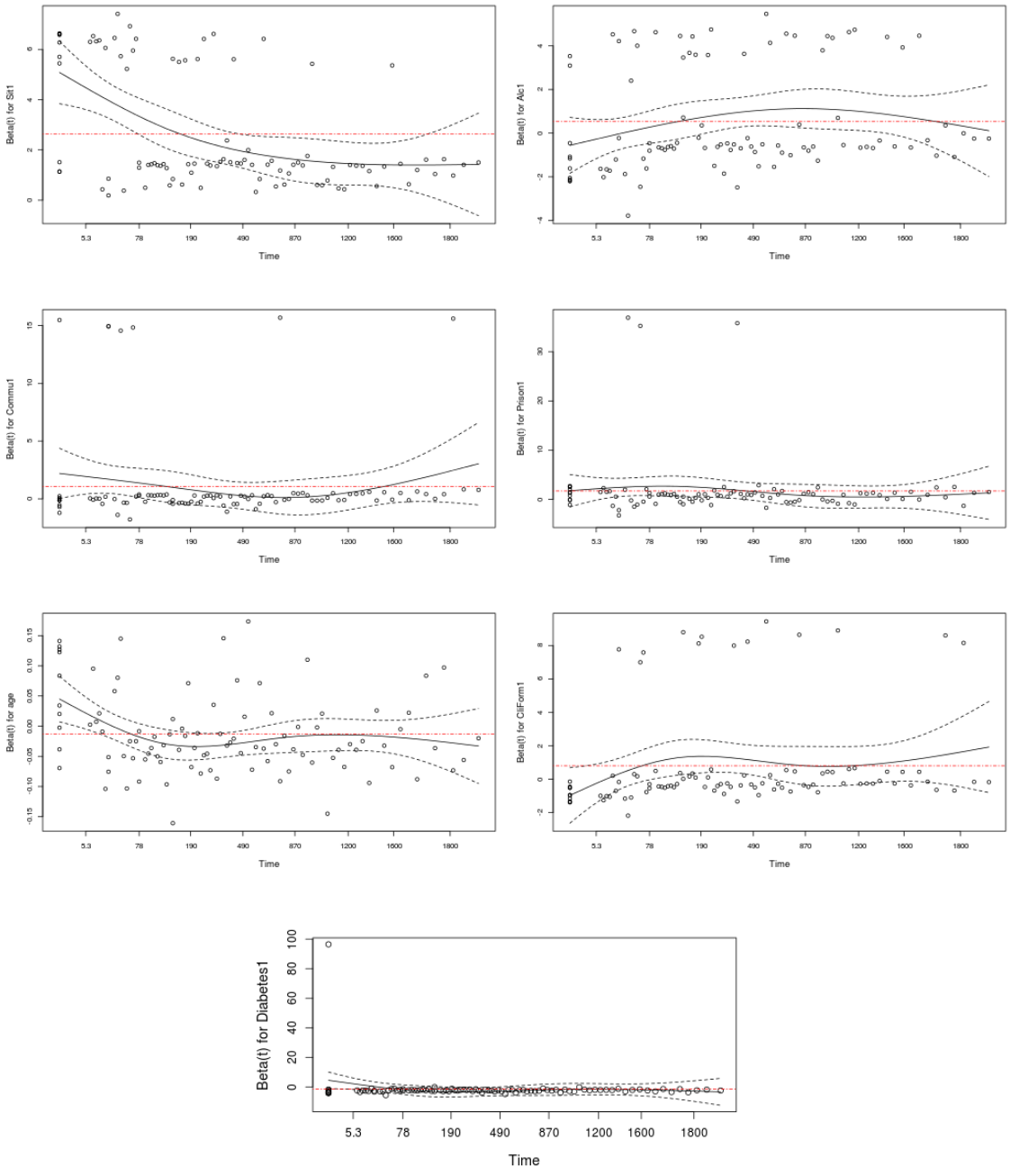

Figure 4.19: Plot of the Schoenfeld residuals (CC without Vaccine)

notice that the sample size has increased due to the exclusion of the variable vaccine. It is known that with very large samples p-values quickly reach 0, therefore, relying exclusively on p-values is not appropriate. In fact, some outliers may cause the hypothesis of proportionality to be rejected. Figure 4.20 shows that the significant test for nonproportionality is mainly due to the existence of very early event times that appears as outliers in the log scale.

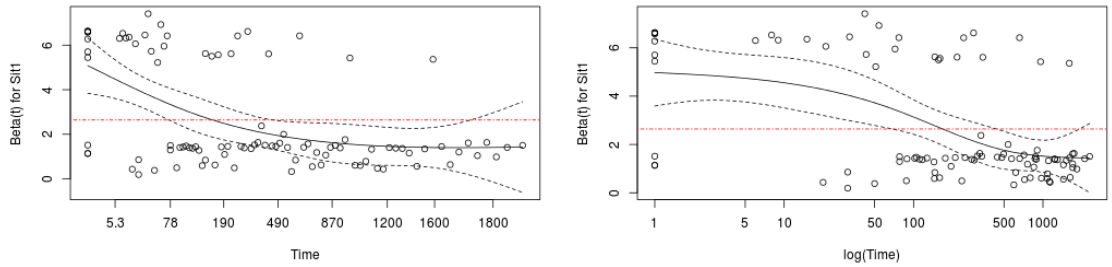

Figure 4.20: Outliers and test for proportionality of risks (CC without Vaccine)

Another possible cause for the nonproportionality is the omission of important covariates, as explained in section 2.2.1. In this case, the Schoenfeld residuals plots may suggest the presence of nonproportionality.

Martingale Residuals

The martingale residuals were used in order to find if there are any individuals who were poorly fitted by the model, as well as to investigate the functional form of continuous variables (figure 4.21).

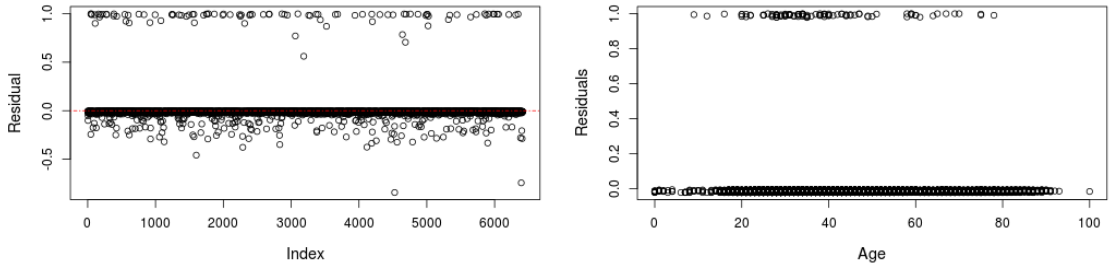

Figure 4.21: Plot of the martingale residuals (CC without Vaccine)

The plot of the martingale residuals versus the index of each individual, in figure 4.21, does not present any pattern, corresponding to a good fit since the residuals are evenly distributed above and under zero. Regarding the functional form of the variable age, in figure 4.21, the behaviour of the residuals are linear.

4.5.3 Collinearity

Evaluating the existence of collinearity between the covariates is essential since its existence difficults the estimation of the coefficients. Values of VIF above 10 are considered problematic (87). When the VIF is approximately 1 the covariates are independent (table 4.17).

CliForm	Sit	Alc	Prison	Commu	Diabetes	age
1.026	1.040	1.084	1.037	1.013	1.008	1.035

Table 4.17: Values of VIF (CC without Vaccine)

4.6 Mean Imputation

Another commonly used method is single mean imputation, although it is not recommend for categorical variables. For numeric variables, this method replace the missing values by the column median while as for categorical variables missing values are replaced by the most frequent value (ties are randomnly atributed).

4.6.1 Cox Regression Analysis

To find significant variables in the dataset obtained with mean imputation the same method of variable selection was used. The results of the univariate analysis are in table 4.18.

The variables Radio, Origin, Job and Transf were discarded. A Cox model with all the remaining variables was fitted to the data. Results are presented in table 4.19.

As before, the variables that, when omitted did not significantly increased the value of -2 $\log \hat{\mathcal{L}}$, were discarded from the model one by one. At the end, the significant variables were Sit, Alc, age, Prison, HIV, CliForm, Diabetes and Vac.

The variables Job, Radio, Transf and Origin did not reduced significantly the value of -2 $\log \hat{\mathcal{L}}$. The results of the final model can be seen in table 4.20.

The hazard ratio, exp($\hat{\beta}$) , was estimated for the variables (table 4.20). Adjusting for the remaining variables, it is estimated that:

- An individual that was not vaccinated has an increase of 70% in the risk of a recurrent episode when compared with a vaccinated individual;
- An individual with an extrapulmonary form of TB has an increase of 95% in the risk of a recurrent episode when compared with an individual with a pulmonary form of TB;
- The risk of recurrence in individuals that defaulted the treatment is 9.99 times that of individuals who completed the treatment;
- An alcoholic individual has an increase of 80% in the risk of a recurrent episode when compared with a non alcoholic individual;

Variable	-2 $\log \hat{\mathcal{L}}$	p-value	
Null model	2543.59	-	
Vac	2539.89	0.05	**
CliForm	2533.18	0.00	***
Radio	2543.07	0.77	
Sit	2409.90	0.00	***
Sex	2538.79	0.03	**
Origin	2543.35	0.89	
Job	2542.30	0.26	
Alc	2529.71	0.00	***
Smk	2510.96	0.00	***
Drugs	2530.71	0.00	***
Prison	2537.23	0.01	***
Commu	2538.02	0.02	**
Hmless	2535.84	0.01	**
Transf	2543.53	0.82	
Unemp	2529.85	0.00	***
HIV	2502.97	0.00	***
Diabetes	2533.65	0.00	***
NumCo	2531.87	0.00	***
age	2533.47	0.00	***

Table 4.18: Values of -2 $\log \hat{\mathcal{L}}$ and p-value of Likelihood ratio tests obtained in the univariate analysis (Mean imputation). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	-2 $\log \hat{\mathcal{L}}$	p-value	
Model with all	2346.04	-	
Model without Vac	2352.36	0.01	***
Model without CliForm	2353.95	0.005	***
Model without Sit	2444.351	0.00	***
Model without Sex	2346.38	0.56	
Model without Alc	2353.97	0.005	***
Model without Smk	2347.99	0.16	
Model without Drugs	2348.18	0.14	
Model without Prison	2350.63	0.03	**
Model without Commu	2346.68	0.42	
Model without Hmless	2347.04	0.32	
Model without Unemp	2346.28	0.62	
Model without HIV	2351.26	0.02	**
Model without Diabetes	2349.11	0.08	*
Model without NumCo	2348.18	0.34	
Model without age	2351.40	0.02	**

Table 4.19: Values of -2 $\log \hat{\mathcal{L}}$ and p-values of Likelihood ratio tests (Mean imputation). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
Vac1	0.544	1.723	0.204	2.667	0.008	[0.144;0.944]
CliForm1	0.668	1.950	0.235	2.843	0.004	[0.208;1.128]
Sit1	2.302	9.998	0.192	11.990	0.000	[1.925;2.680]
Alc1	0.589	1.801	0.191	3.084	0.002	[0.215;0.962]
Prison1	1.433	4.190	0.518	2.766	0.006	[0.417;2.449]
HIV1	0.752	2.121	0.197	3.817	0.000	[0.365;1.138]
Diabetes1	-1.660	0.190	1.007	-1.648	0.099	[-3.635;0.314]
age	-0.014	0.986	0.006	-2.333	0.020	[-0.026;-0.002]

Table 4.20: Results using the Cox model (Mean imputation)

- The risk of recurrence in individuals in prison is 4.19 times that of individuals that are not in prison;
- The risk of recurrence in individuals with HIV is 2.12 times that of individuals without HIV;
- An individual with Diabetes has a decrease of 80% in the risk of a recurrent episode when compared with someone without Diabetes;
- Each additional year of age is associated with an estimated 1 % decrease in risk. An additional decade corresponds to a 13% decrease in risk.

The confidence intervals are not large, however, it was previously discussed that mean imputation leads to overestimation of precision.

4.6.2 Residual Analysis

Schoenfeld Residuals

To assess the proportionality of the hazards the Schoenfeld residuals were calculated.

Table 4.21 shows the *p*-value of the test of proportional hazards for all the variables. As expected, since the sample size increased, the *p*-values are low. Therefore, it is expected that for the variable Sit there is a clear rejection of the null hypothesis of proportionality of risks.

Although the test shows that the residuals of some variables (Sit, Vac and age) are significantly nonproportional, the residual plot (figure 4.22) shows that the variation in $\hat{\beta}(t)$ is small relative to $\hat{\beta}$, suggesting a non rejection of the hypothesis of proportionality of risks. As before, figure 4.23 shows that the rejection of the proportionality is mainly due to the presence of outliers observed in the log scale.

Martingale Residuals

To find if there are any individuals who were not well fitted by the model the martingale residuals were used. These residuals also help to investigate the functional form of continuous variables, i.e., if the variables are linear or need any transformation.

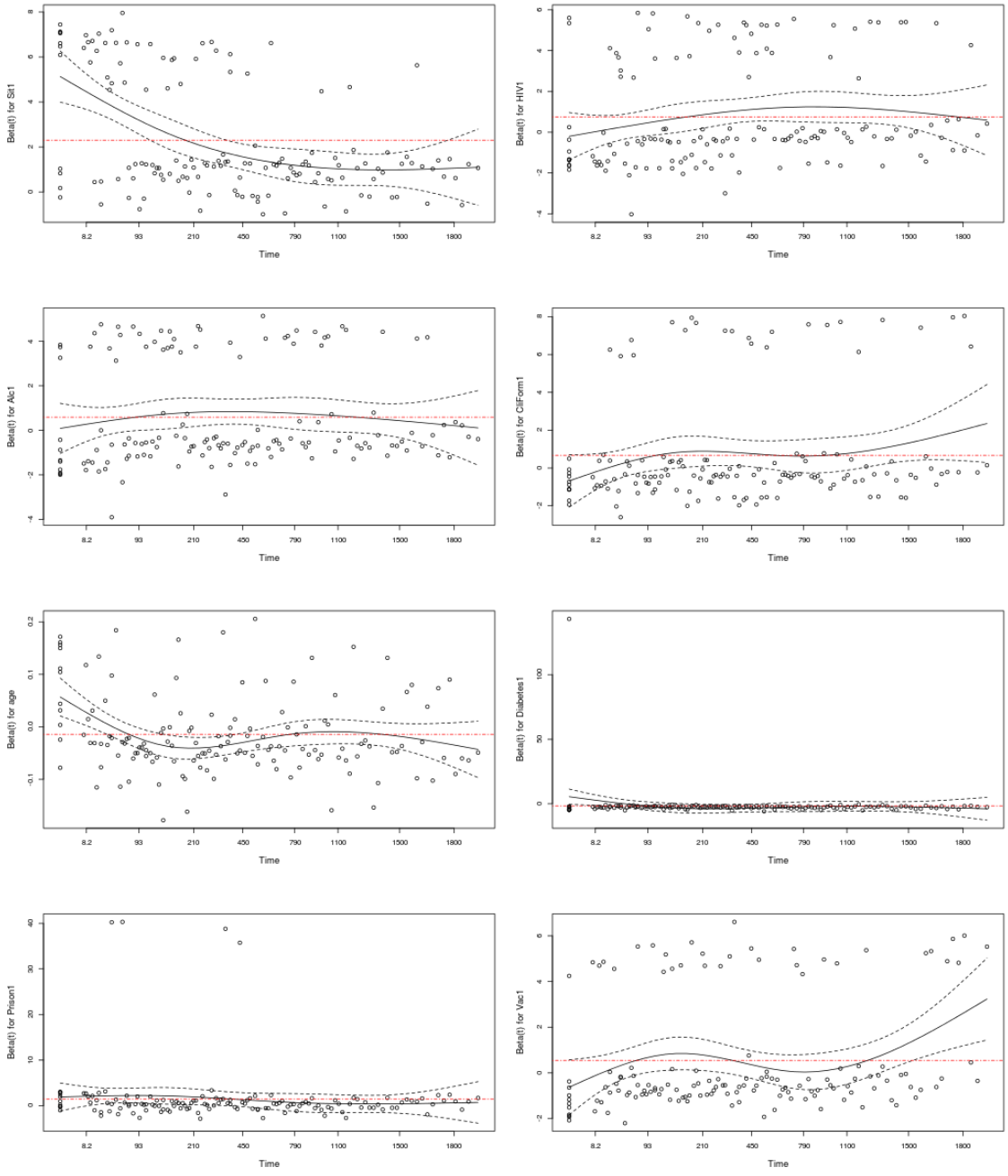

Figure 4.22: Plot of the Schoenfeld residuals (Mean imputation)

	rho	χ^2	*p*-value
Vac1	0.178	4.661	0.031
CliForm1	0.170	4.386	0.036
Sit1	-0.459	37.867	7.6e-10
Alc1	0.013	0.024	0.876
Prison1	-0.100	1.532	0.216
HIV1	0.130	2.812	0.093
Diabetes1	-0.123	2.196	0.138
age	-0.159	3.877	0.049
GLOBAL	NA	49.123	6.0e-08

Table 4.21: Test for proportionality of risks (Mean imputation)

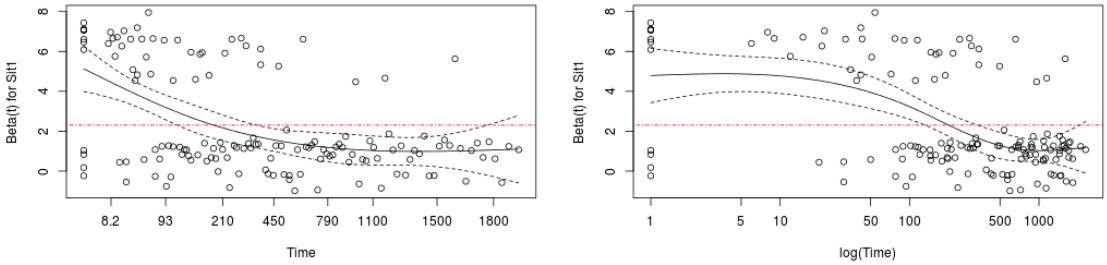

Figure 4.23: Outliers and test for proportionality of risks (Mean imputation)

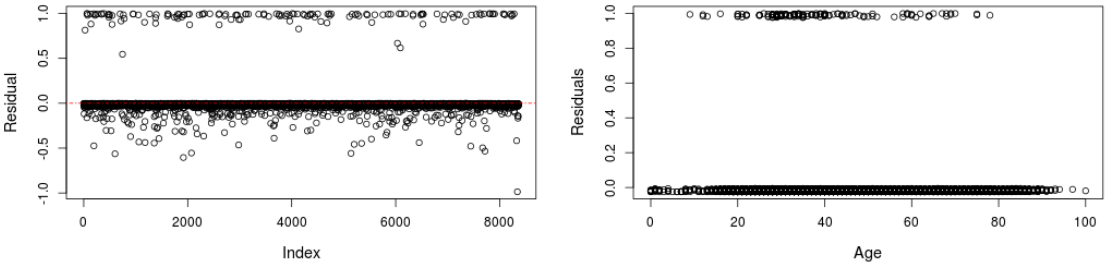

Figure 4.24: Plot of the martingale residuals (Mean imputation)

The plot of the martingale residuals versus the index of each individual, in figure 4.24, does not present any pattern, corresponding to a good fit since the residuals are evenly distributed above and under zero. Regarding the functional form of the variable age, figure 4.24, the behaviour of the residuals are linear.

4.6.3 Collinearity

Evaluating the existence of collinearity between the covariates is essential since its existence difficults the estimation of the coefficients. Values of VIF above 10 are considered problematic (87). When the VIF is approximately 1 the covariates are independent (table 4.22).

Vac	CliForm	Sit	Alc	Prison	HIV	Diabetes	age
1.038	1.066	1.120	1.068	1.044	1.142	1.008	1.061

Table 4.22: Values of VIF (Mean imputation)

4.7 Predictive Mean Matching

4.7.1 Imputation Diagnosis

After the imputation, with PMM, it is necessary to check if the imputed data are plausible. In general, a good imputed value is a value that could have been observed had it not been missing. Differences in the density plots for the observed and imputed values may suggest the existence of a problem that needs to be further checked. Figure 4.25 represents a density plot, for each imputed dataset, of the observed and imputed values for the variable Vacine. All the other variables had similar distributions of imputed and observed values, so there is no indication of any problem during the imputation process.

Another important diagnosis consists in determining if the Gibbs sampling algorithm has converged. Figure 4.26 represent the variables Commu, Hmless and age plotted against the number of iterations. To achieve a good convergence, streams should be freely intermingled with each other, without defining trends. In this case, there is very little trend and the streams mingle well from the start.

4.7.2 Cox Regression Analysis

As mentioned before, the selection of variables should be done in each imputed dataset and not only on the pooled data. Therefore, the selection of variables was performed in each of the imputed dataset. A variable was considered if it was significant in at least half of the models. The results presented in table 4.23 are an average of the imputed analyses.

The variables Vac, Radio, Transf, Origin and Job were discarded from the analyses. The next step consist in including all the significant variables in a model and remove one by one to see its effect. Results are presented in table 4.24.

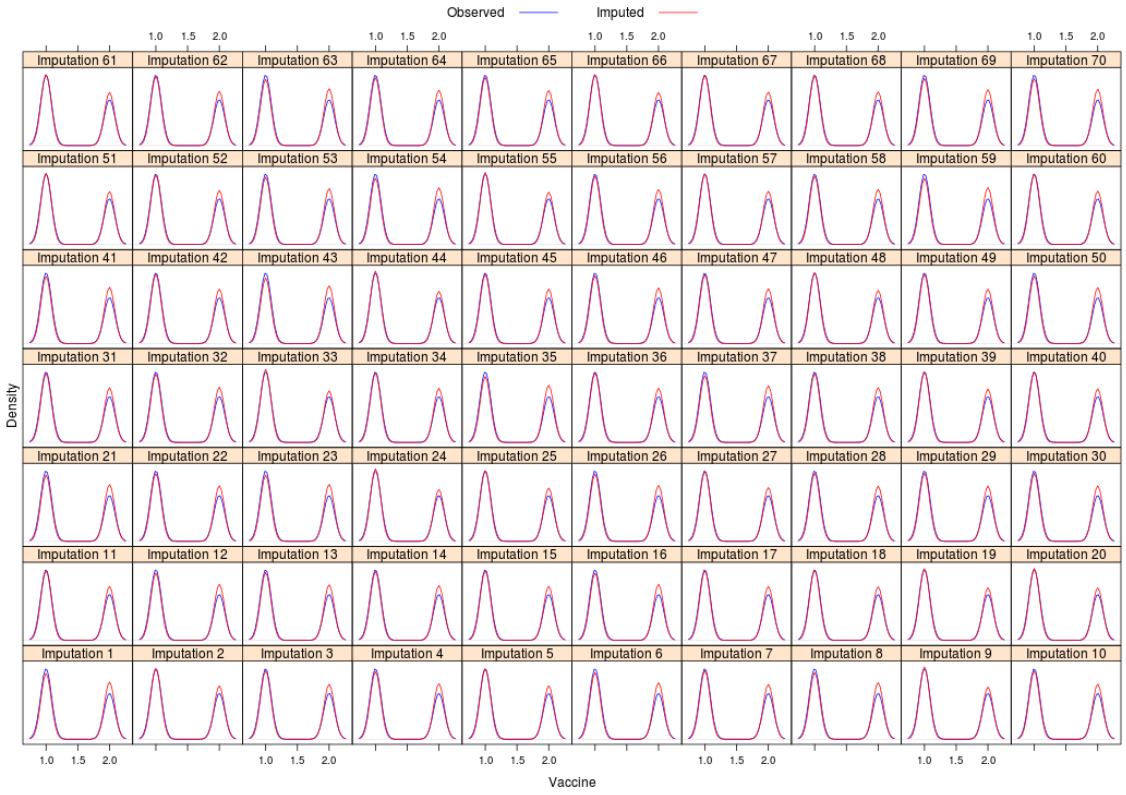

Figure 4.25: Kernel density estimates for the marginal distributions of the observed and imputed values of the Vaccine (PMM)

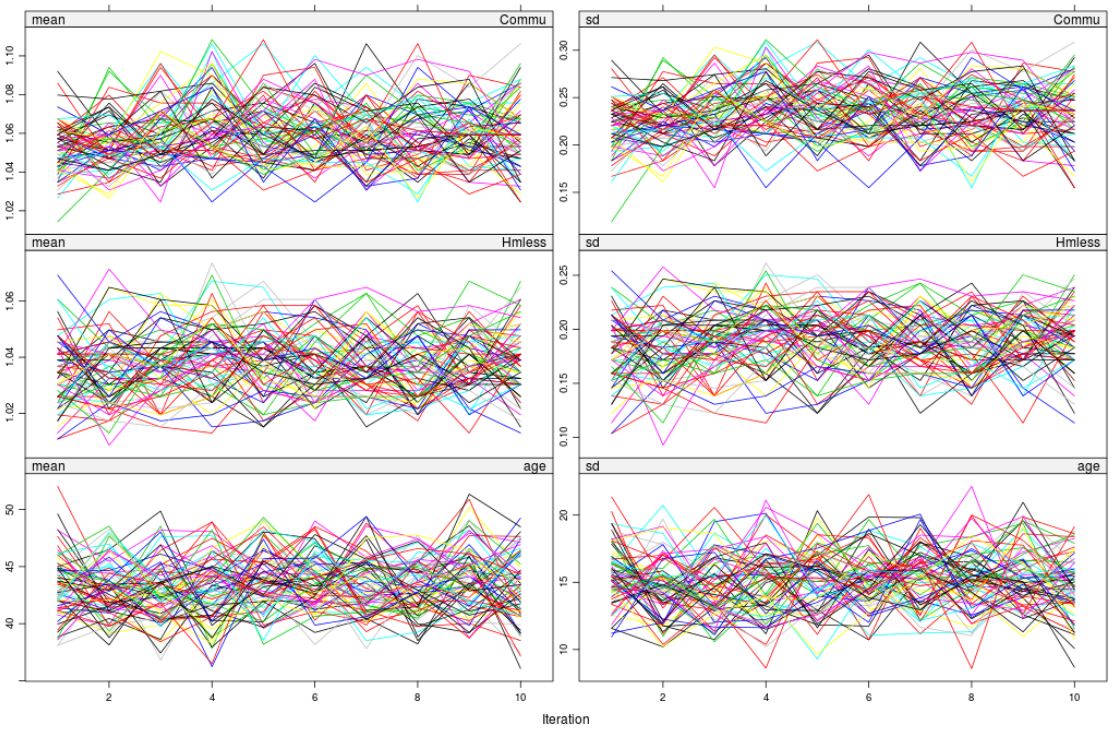

Figure 4.26: Convergence of the Gibbs sampler (PMM)

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	2543.587	-	
Vac	2543.099	0.5885	
CliForm	2533.189	0.0004	***
Radio	2539.76	0.1558	
Sit	2409.896	0	***
Sex	2538.786	0.0344	**
Origin	2543.347	0.8977	
Job	2542.36	0.3847	
Alc	2527.667	6.1e-05	***
Smk	2507.486	1.8e-10	***
Drugs	2522.983	6.5e-06	***
Prison	2537.522	0.0072	***
Commu	2536.977	0.0080	***
Hmless	2536.233	0.0022	***
Transf	2543.535	0.8066	
Unemp	2529.851	6.8e-05	***
HIV	2502.973	9.7e-13	***
Diabetes	2533.647	0.0354	**
NumCo	2531.870	0.0018	***
age	2533.430	0.0021	***

Table 4.23: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (PMM). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	2356.92	-	
Model without CliForm	2363.754	0.0094	***
Model without Sit	2452.399	0	***
Model without Sex	2357.363	0.5115	
Model without Alc	2363.0115	0.0201	**
Model without Smk	2358.418	0.2616	
Model without Drugs	2358.504	0.2976	
Model without Prison	2359.493	0.1958	
Model without Commu	2357.481	0.5599	
Model without Hmless	2357.376	0.6084	
Model without Unemp	2357.450	0.4770	
Model without HIV	2361.735	0.0293	**
Model without Diabetes	2359.953	0.0818	*
Model without NumCo	2358.951	0.0293	**
Model without age	2360.915	0.0462	**

Table 4.24: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests (PMM). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

One by one the variable with the lowest increase of -2 $\log \hat{\mathcal{L}}$ was removed from the model. Sex, Hmless, Unemp, Commu, Drugs, Smk, NumCo and Prison were discarded due to the low value of -2 $\log \hat{\mathcal{L}}$. The next step consists of adding each of the discarded variables in the univariate analysis to the complete model. None of the discarded variable had a significant increase of -2 $\log \hat{\mathcal{L}}$, therefore, the final model consists of CliForm, Sit, Alc, HIV, Diabetes and age. The results presented in table 4.25 are a combination of the multiple analyses through Rubin's rules.

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
CliForm=1	0.634	1.885	0.235	2.693	0.0071	[0.172;1.095]
Sit=1	2.280	9.777	0.193	11.808	0	[1.901;2.658]
Alc=1	0.574	1.775	0.195	2.947	0.0032	[0.192;0.956]
HIV=1	0.706	2.026	0.198	3.565	0.0004	[0.318;1.094]
Diabetes=1	-1.68	0.186	1.007	-1.668	0.0952	[-3.654;0.294]
age	-0.013	0.987	0.006	-2.142	0.0322	[-0.025;-0.001]

Table 4.25: Results using the Cox model (PMM)

The hazard ratio, exp($\hat{\beta}$), was estimated for the variables (table 4.25). Adjusting for the remaining variables, it is estimated that:

- An individual with extrapulmonar TB has an increase of 86% in the risk of a recurrent episode when compared with an individual with a pulmonary form of TB;
- The risk of recurrence in individuals who defaulted the treatment is 9.78 times that of individuals who completed the treatment;
- An alcoholic individual has an increase of 77% in the risk of a recurrent episode when compared with an individual without an alcoholic dependence;
- The risk of recurrence in individuals with HIV is 2.03 times that of individuals without HIV;
- An individual with Diabetes has a decrease of 80% in the risk of a recurrent episode when compared with someone without Diabetes;
- Each additional year of age is associated with an estimated 1% decrease in risk. An additional decade corresponds to a 12% decrease.

In this case, the confidence intervals are narrow.

4.7.3 Residual Analysis

Schoenfeld Residuals

The Schoenfeld residuals were obtained for each one of the imputed datasets. In the previous analyses, the values presented were an average of the values of the 70 imputed datasets. However, it is computationally challenging to average all the values of the Schoenfeld residuals. Since the imputed datasets do not differ much from each other, only one imputed dataset is presented which is representative of the remaining imputations.

	rho	χ^2	*p*-value
CliForm=1	0.167	4.424	0.035
Sit=1	-0.460	36.876	1.3e-09
Alc=1	0.044	0.266	0.606
HIV=1	0.102	1.758	0.185
Diabetes=1	-0.125	2.263	0.133
age	-0.135	2.728	0.099
GLOBAL	NA	43.592	8.9e-08

Table 4.26: Test for proportionality of risks (PMM)

As expected due to the large sample size, the global value rejects the proportionality of hazards (table 4.26). However, analysis of figure 4.27 shows that the variation in $\hat{\beta}(t)$ for the variable Sit is relatively close to $\hat{\beta}$. As previously discussed, the rejection of the proportionality is mainly due to events that appears as outliers in the log scale (figure 4.28).

Martingale Residuals

In order to explore the fit of the models to individuals the martingale residuals were used. The model is well fitted to the data since the residuals are evenly distributed above and under zero (figure 4.29).

4.7.4 Collinearity

Evaluating the existence of collinearity between the covariates is essential since its existence difficults the estimation of the coefficients. The values of VIF in table 4.27 are an average of each imputed dataset. When the VIF is approximately 1 the covariates are independent (table 4.27).

CliForm	Sit	Alc	HIV	Diabetes	age
1.065	1.125	1.068	1.150	1.007	1.033

Table 4.27: Average of the values of VIF (PMM)

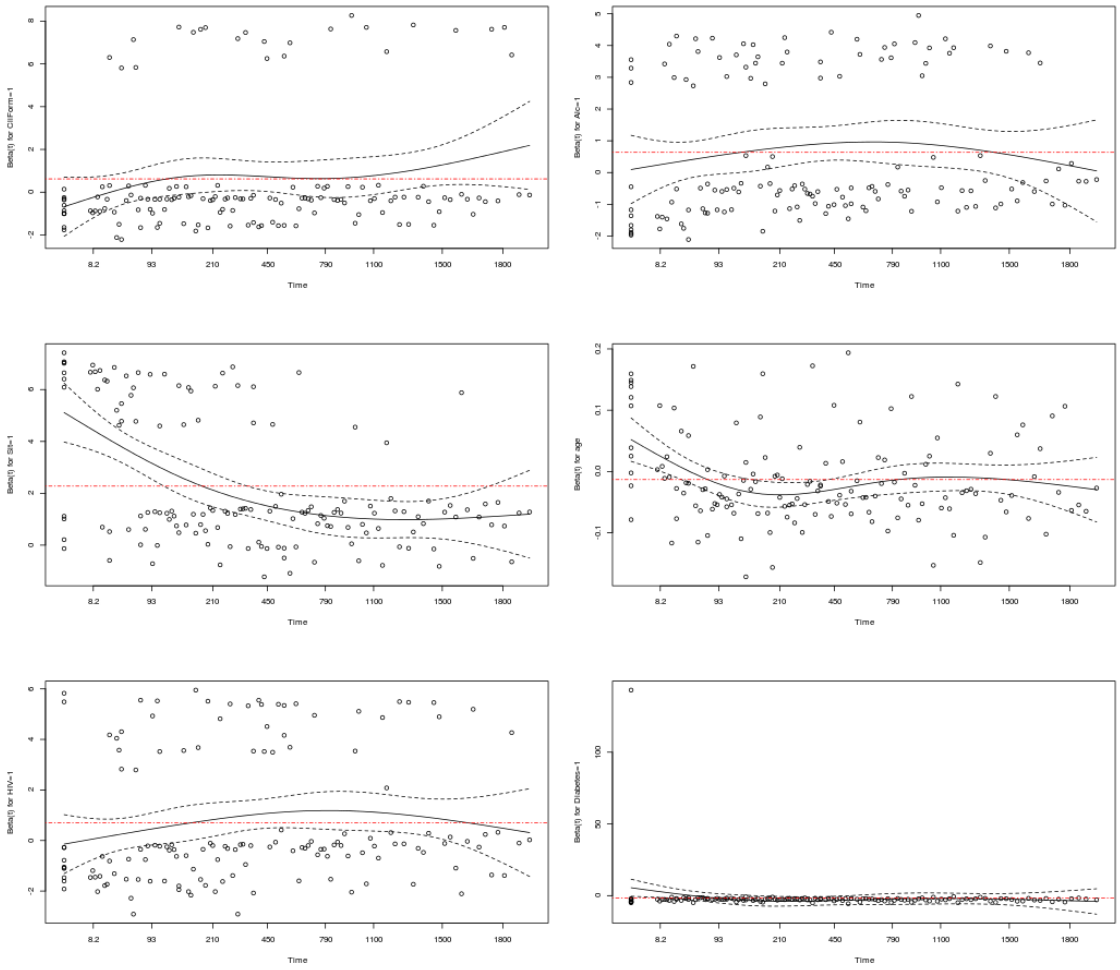

Figure 4.27: Plot of the Schoenfeld residuals (PMM)

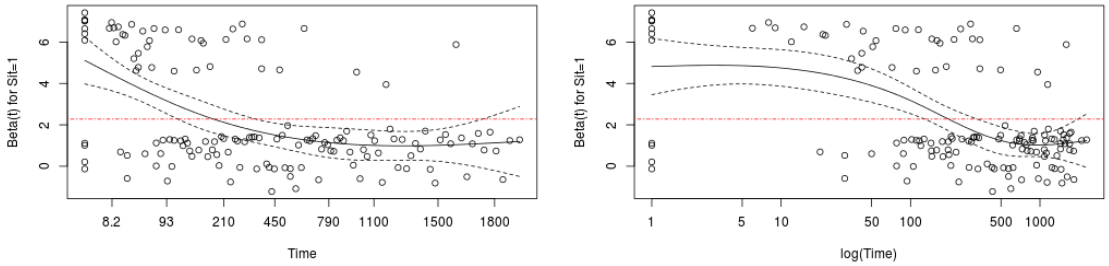

Figure 4.28: Outliers and test for proportionality of risks (PMM)

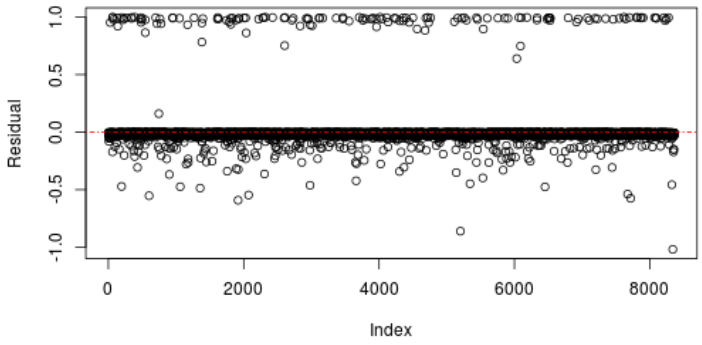

Figure 4.29: Plot of the martingale residuals (PMM)

4.8 Random Forest (mice)

4.8.1 Imputation Diagnosis

After the imputation, with RF, it is necessary to check if the imputed data are plausible. In general, a good imputed value is a value that could have been observed had it not been missing. Differences in the densities between the observed and imputed values may suggest a problem that needs to be further checked. Figure 4.30 represents a density plot, for each imputed dataset, of the observed and imputed values for the variable age. All the other variables had similar distributions of imputed and observed values which does not indicate any problem during the imputation process.

Another important diagnostics relates with the convergence of the Gibbs sampling algorithm. Figure 4.31 represent the variables Origin, Job and Alc against the number of iterations. the figure does not present a clear trend and the streams mingle well from the start, indicating a good convergence.

4.8.2 Cox Regression Analysis

The variables were selected according to Collett's approach (77), in each imputed dataset. A variable was considered if it was significant in at least half of the models. The results presented in table 4.28 are an average of the imputed analyses.

The variables Vac, Radio, Transf, Origin and Job were discarded from the analyses. The next step consists of including all the significant variables in a model and remove one by one to see its effect. Results are in table 4.29.

The variables with the lowest increase of -2 $\log \hat{\mathcal{L}}$ were removed from the model, one by one. Afterwards, each one of the discarded variables in the univariate analysis was added to the final model. However, since none of them significantly increased the value of -2 $\log \hat{\mathcal{L}}$ the final

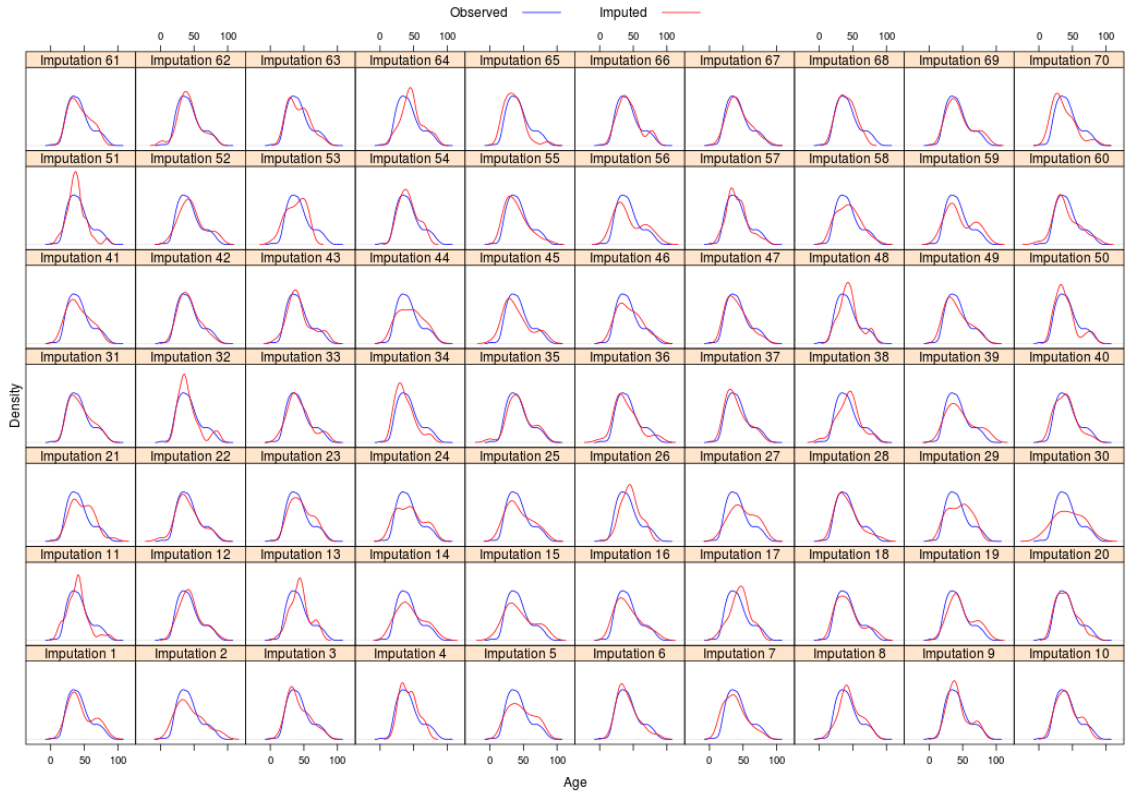

Figure 4.30: Kernel density estimates for the marginal distributions of the observed and imputed values of the age (RF **mice**)

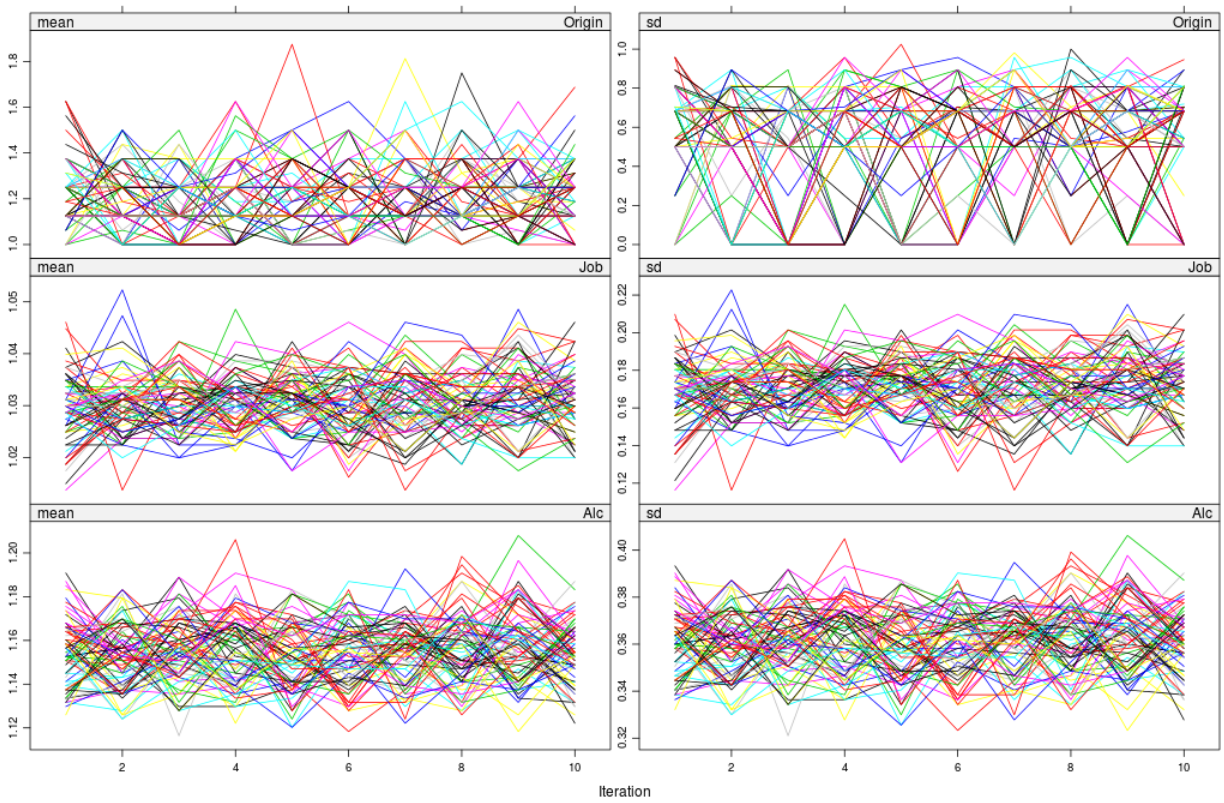

Figure 4.31: Convergence of the Gibbs sampler (RF **mice**)

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	2543.587	-	
Vac	2541.591	0.3123	
CliForm	2533.194	0.0004	***
Radio	2541.382	0.3611	
Sit	2409.896	0	***
Sex	2538.786	0.0344	**
Origin	2543.340	0.8946	
Job	2542.550	0.4365	
Alc	2529.052	0.0001	***
Smk	2511.304	0	***
Drugs	2529.794	0.0001	***
Prison	2537.408	0.002	***
Commu	2537.860	0.0077	***
Hmless	2536.050	0.0010	***
Transf	2543.535	0.8066	
Unemp	2529.851	0.0001	***
HIV	2502.973	0	***
Diabetes	2533.647	0.0354	**
NumCo	2531.870	0.0018	***
age	2533.439	0.0021	***

Table 4.28: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (RF **mice**). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	2353.413	-	
Model without CliForm	2360.429	0.0083	***
Model without Sit	2451.319	0	***
Model without Sex	2353.842	0.5170	
Model without Alc	2360.480	0.0110	**
Model without Smk	2354.939	0.2401	
Model without Drugs	2355.080	0.2375	
Model without Prison	2357.729	0.0400	**
Model without Commu	2354.007	0.4767	
Model without Hmless	2354.402	0.3338	
Model without Unemp	2353.960	0.4648	
Model without HIV	2358.407	0.0258	**
Model without Diabetes	2356.479	0.0800	*
Model without NumCo	2355.64	0.3300	**
Model without age	2357.403	0.0462	**

Table 4.29: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests (RF **mice**). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

model remained equal. The results presented in table 4.30 are a combination of the multiple analyses through Rubin's rules.

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
CliForm=1	0.624	1.866	0.235	2.653	0.0080	[0.163;1.085]
Sit=1	2.319	10.166	0.193	12	0	[1.940;2.697]
Alc=1	0.563	1.756	0.194	2.897	0.0038	[0.182;0.945]
Prison=1	1.362	3.904	0.520	2.619	0.0088	[0.343;2.380]
HIV=1	0.707	2.028	0.197	3.577	0.0003	[0.319;1.094]
Diabetes=1	-1.682	0.186	1.007	-1.671	0.0948	[-3.656;0.291]
age	-0.012	0.988	0.006	-2.054	0.0400	[-0.024;-0.001]

Table 4.30: Results using the Cox model (RF **mice**)

The hazard ratio, exp($\hat{\beta}$), was estimated for the variables (table 4.30). Adjusting for the remaining variables, it is estimated that:

- An individual with extrapulmonar TB has an increase of 87% in the risk of a recurrent episode when compared with an individual with pulmonary TB;
- The risk of recurrence in individuals who defaulted the treatment is 10 times the risk that of individuals who completed the treatment;
- An alcoholic individual has an increase of 76% in the risk of a recurrent episode when compared with an individual without an alcoholic dependence;
- The risk of recurrence in individuals incarcerated or working in a prison is 3.9 times that of individuals who are not in the prison;
- The risk of recurrence in individuals with HIV is 2.03 times that of individuals without HIV;
- An individual with Diabetes has a decrease of 81% in the risk of a recurrent episode when compared with individuals without Diabetes;
- Each additional year of age is associated with an estimated 1% decrease in risk. An additional decade corresponds to a 12% decrease.

4.8.3 Residual Analysis

Schoenfeld Residuals

As discussed in 4.7.3, the Schoenfeld residuals were obtained from a single imputed dataset.

Table 4.31 display the results of the test of proportionality of hazards. Although the global value suggests nonproportionality, a visual analysis of the plots (figure 4.32) suggests that the proportionality of hazards should not be rejected since the variation in $\hat{\beta}(t)$ for the variable Sit is relatively close to $\hat{\beta}$.

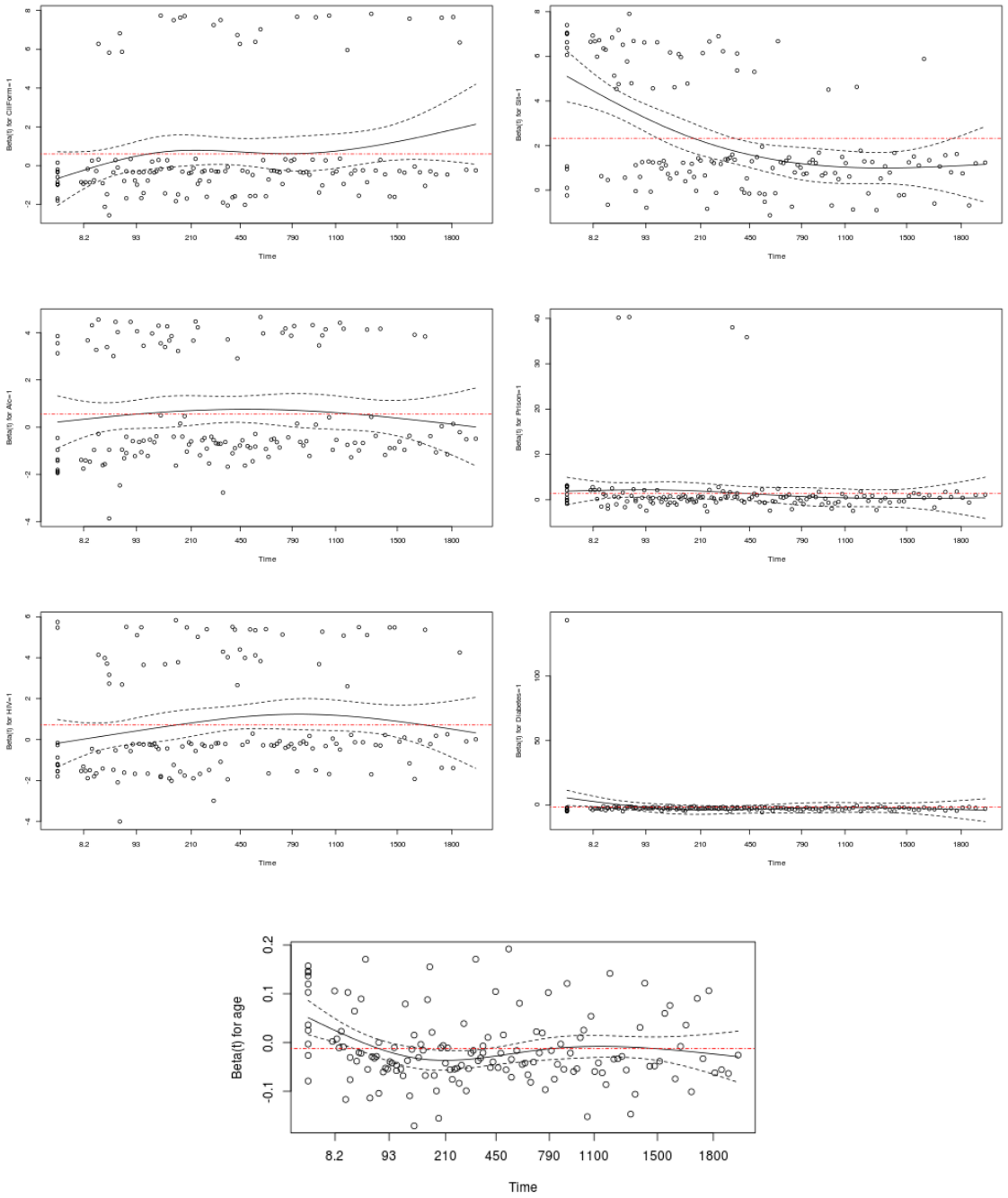

Figure 4.32: Plot of the Schoenfeld residuals (RF **mice**)

	rho	χ^2	*p*-value
CliForm=1	0.1654	4.130	0.042
Sit=1	-0.4574	37.600	8.8e-10
Alc=1	0.0003	1.3e-05	0.997
Prison=1	-0.1054	1.700	0.193
HIV=1	0.1147	2.240	0.135
Diabetes=1	-0.1245	2.250	0.134
age	-0.1327	2.600	0.107
GLOBAL	NA	45.300	1.2e-07

Table 4.31: Test for proportionality of risks (RF **mice**)

Martingale Residuals

In order to explore the fit of the models to individuals, the martingale residuals were used. The model is well fitted to the data since the residuals are evenly distributed above and under zero (figure 4.33).

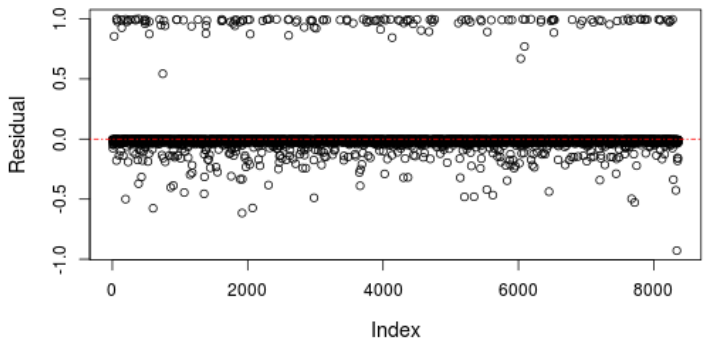

Figure 4.33: Plot of the martingale residuals (RF **mice**)

4.8.4 Collinearity

Evaluating the existence of collinearity between the covariates is essential since its existence difficults the estimation of the coefficients. The values of VIF in table 4.32 are an average of each imputed dataset. When the VIF is approximately 1 the covariates are independent (table 4.32).

CliForm	Sit	Alc	Prison	HIV	Diabetes	age
1.064	1.127	1.066	1.036	1.145	1.007	1.037

Table 4.32: Average of the values of VIF (RF **mice**)

4.9 Random Forest (missForest)

4.9.1 Imputation Diagnosis

The package **missForest** does not provide visualization diagnosis to assess the plausibility of the imputation. Nevertheless, given the closeness of results between this method and previous methods, the imputation is likely to have performed correctly.

4.9.2 Cox Regression Analysis

One of the features of this package, is that it does not return *m* datasets. Instead, it returns one imputed dataset and the analysis is made on this imputed dataset. The results of the univariate analysis are presented in table 4.33.

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	2543.587	-	
Vac	2540.86	0.099	*
CliForm	2533.18	0.001	***
Radio	2539.06	0.104	
Sit	2409.90	0.000	***
Sex	2538.79	0.028	**
Origin	2543.34	0.885	
Job	2542.12	0.225	
Alc	2526.25	0.000	***
Smk	2506.28	0.000	***
Drugs	2526.19	0.000	***
Prison	2537.54	0.014	***
Commu	2536.59	0.008	***
Hmless	2536.21	0.007	***
Transf	2543.53	0.819	
Unemp	2529.85	0.000	***
HIV	2502.97	0.000	***
Diabetes	2533.65	0.002	***
NumCo	2531.87	0.003	***
age	2533.40	0.001	***

Table 4.33: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (RF **missForest**). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

The variables Radio, Origin, Job and Transf were discarded from the analyses. The next step consists of including all the significant variables in a model and remove one by one to see its effect. Results are in table 4.34.

Variables with a small value of -2 $\log \hat{\mathcal{L}}$ were removed from the model individually. The variables discarded in the univariate analysis were introduced in the final model, however, none lead to a significant increase of -2 $\log \hat{\mathcal{L}}$. The results are presented in table 4.35.

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	2344.780	-	
Model without Vac	2352.723	0.0048	***
Model without CliForm	2352.203	0.0060	***
Model without Sit	2439.601	0	***
Model without Sex	2345.095	0.5944	
Model without Alc	2351.124	0.0118	**
Model without Smk	2346.959	0.1398	
Model without Drugs	2347.359	0.1083	
Model without Prison	2349.410	0.0314	**
Model without Commu	2345.859	0.2989	
Model without Hmless	2344.990	0.6468	
Model without Unemp	2345.237	0.4990	
Model without HIV	2349.034	0.0391	**
Model without Diabetes	2348.250	0.0625	**
Model without NumCo	2346.171	0.4987	
Model without age	2353.603	0.0030	**

Table 4.34: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests (RF **missForest**). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
Vac=1	0.560	1.751	0.182	3.077	0.0021	[0.203;0.917]
CliForm=1	0.661	1.937	0.234	2.825	0.0047	[0.202;1.120]
Sit=1	2.271	9.687	0.193	11.767	0	[1.893;2.648]
Alc=1	0.531	1.700	0.183	2.902	0.0037	[0.172;0.890]
Prison=1	1.348	3.848	0.518	2.602	0.0093	[0.333;2.362]
HIV=1	0.685	1.985	0.196	3.495	0.0005	[0.300;1.071]
Diabetes=1	-1.717	0.180	1.007	-1.705	0.0882	[-3.690;0.257]
age	-0.020	0.980	0.007	-2.857	0.0043	[-0.034;-0.007]

Table 4.35: Results using the Cox model (RF **missForest**)

The hazard ratio, $\exp(\hat{\beta})$, was estimated for the variables (table 4.35). Adjusting for the remaining variables, it is estimated that:

- An individual that was not vaccinated has an increase of 75% in the risk of a recurrent episode when compared with a vaccinated individual;
- An individual with extrapulmonar TB has an increase of 94% in the risk of a recurrent episode when compared with an individual with pulmonary TB;
- The risk of recurrence in individuals who defaulted the treatment is 9.69 times that of individuals who completed the treatment;
- An alcoholic individual has an increase of 70% in the risk of a recurrent episode when compared with individuals that do not drink;
- The risk of recurrence in individuals in prison is 3.85 times that of individuals who are not in prison;
- Individuals with HIV has an increase of 99% in the risk of a recurrent episode when compared with individuals without HIV;
- An individual with Diabetes has a decrease of 82% in the risk of a recurrent episode when compared with individuals without Diabetes;
- Each additional year of age is associated with an estimated 2% decrease in the risk. An additional decade corresponds to a 18% decrease.

4.9.3 Residual Analysis

Schoenfeld Residuals

The proportionality of risks was evaluated through a formal test and a visual analysis.

	rho	χ^2	*p*-value
Vac=1	0.051	0.310	0.578
CliForm=1	0.172	4.419	0.035
Sit=1	-0.456	36.780	1.3e-09
Alc=1	0.029	0.121	0.728
Prison=1	-0.115	2.047	0.153
HIV=1	0.106	1.826	0.177
Diabetes=1	-0.124	2.214	0.137
age	-0.163	3.616	0.057
GLOBAL	NA	44.609	4.4e-07

Table 4.36: Test for proportionality of risks (RF **missForest**)

As expected, the global value rejects the proportionality of risks (table 4.36). However, as discussed previously, this is mainly due to the large sample size and to the omission of important covariates. A visual analysis reveals that the difference is not important (figure 4.34).

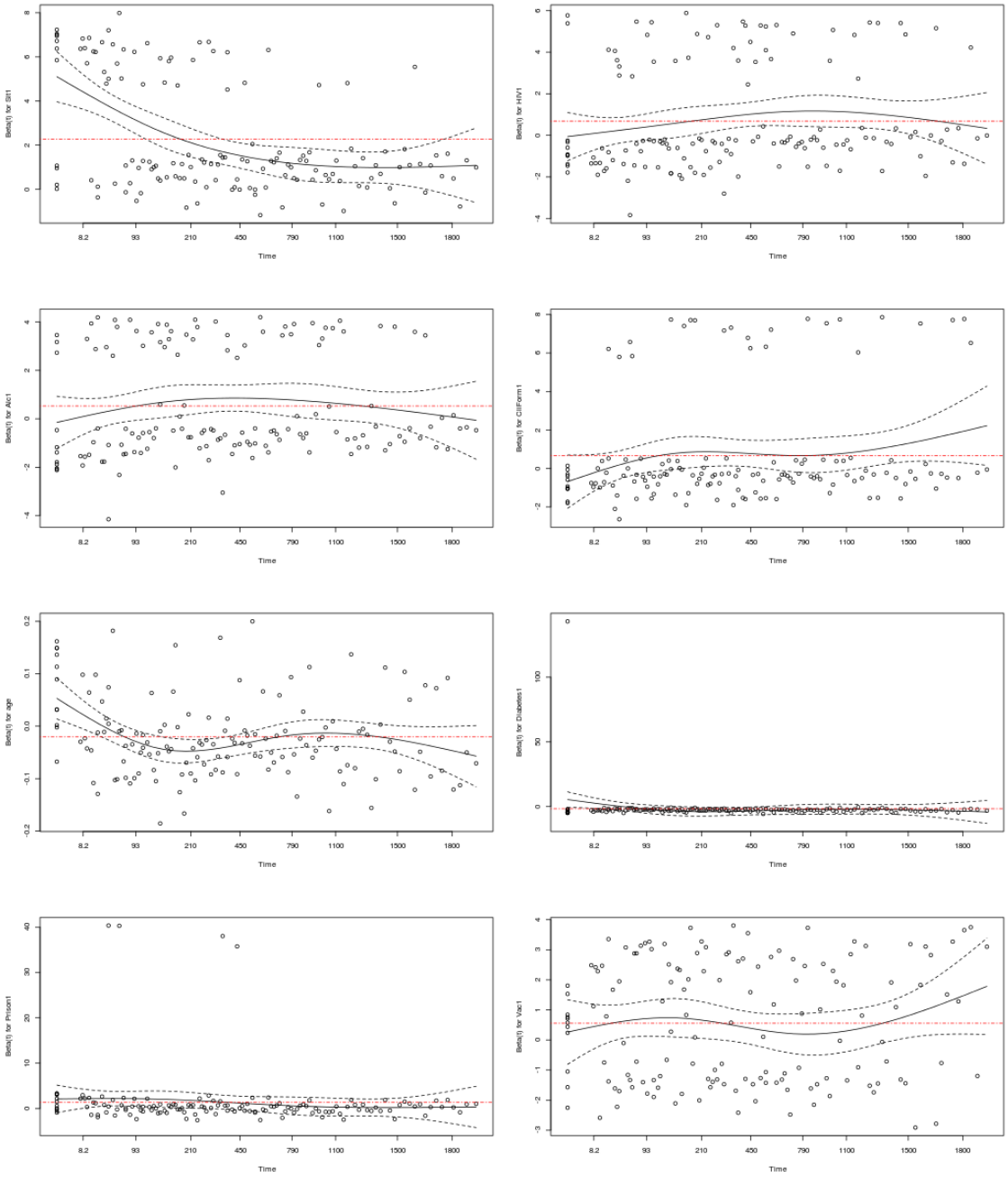

Figure 4.34: Plot of the Schoenfeld residuals (RF **missForest**)

Martingale Residuals

In order to explore the fit of the models to individuals, the martingale residuals were used. The model is well fitted to the data since the residuals are evenly distributed above and under zero (figure 4.35).

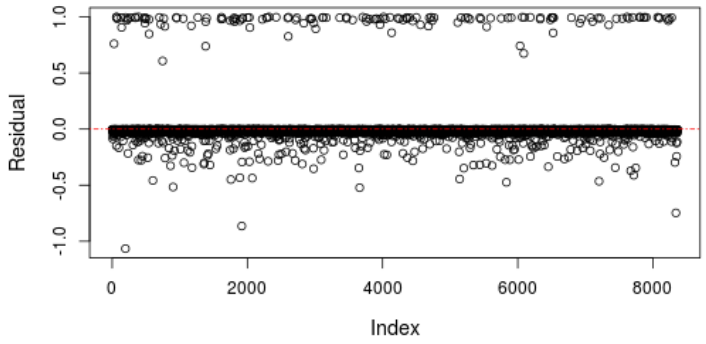

Figure 4.35: Plot of the martingale residuals (RF **missForest**)

4.9.4 Collinearity

The collinearity between covariates was also evaluated. When the VIF is approximately 1 the covariates are independent (table 4.37).

Vac	CliForm	Sit	Alc	Prison	HIV	Diabetes	age
1.191	1.060	1.121	1.077	1.041	1.134	1.007	1.077

Table 4.37: Values of VIF (RF **missForest**)

4.10 Expectation-Maximization with Bootstrapping

4.10.1 Imputation Diagnosis

The package **Amelia II** provides several diagnostic plots, useful to check the plausibility of the imputed data. One diagnostic tool is to compare the distribution of imputed values with the distribution of observed values (figure 4.36). Although no knowledge exists a priori about the distribution of missing values, imputations with different distribution or very distant from the distribution of observed data may indicate that the imputation model needs improvements.

Another useful diagnostic is to check the convergence of the imputation. **Amelia II** provides a function to make sure the imputations do not depend on the starting values. In this sense, the EM algorithm is run from multiple, dispersed starting values. Figure 4.37 shows a well

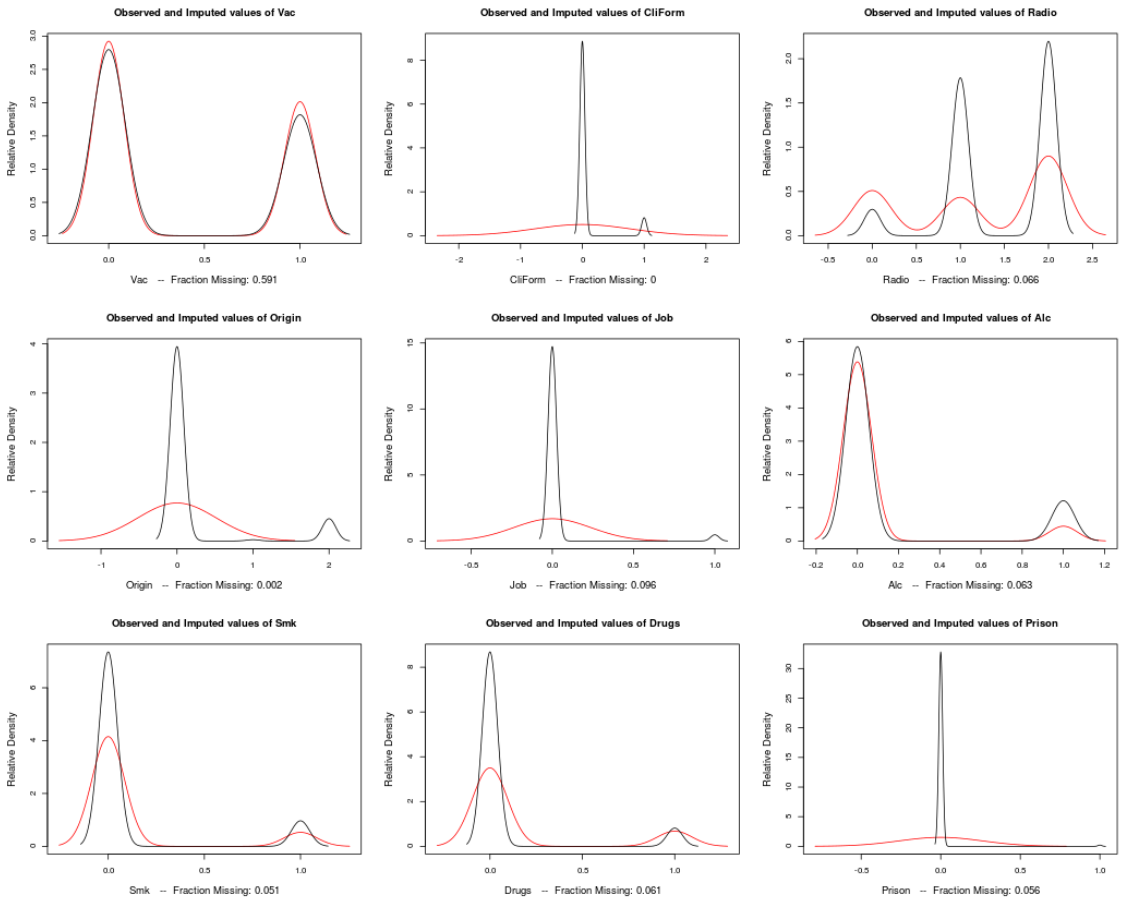

Figure 4.36: Distribution of mean imputations (in red) overlayed on the distribution of observed values (in black) for several variables (EMB)

behaved likelihood, since the starting values converged to the same value, and therefore, it is reasonable to conclude that this is the likely global maximum. Th Y-axis represents movement in the parameter space while the X-axis represents the number of chain iterations.

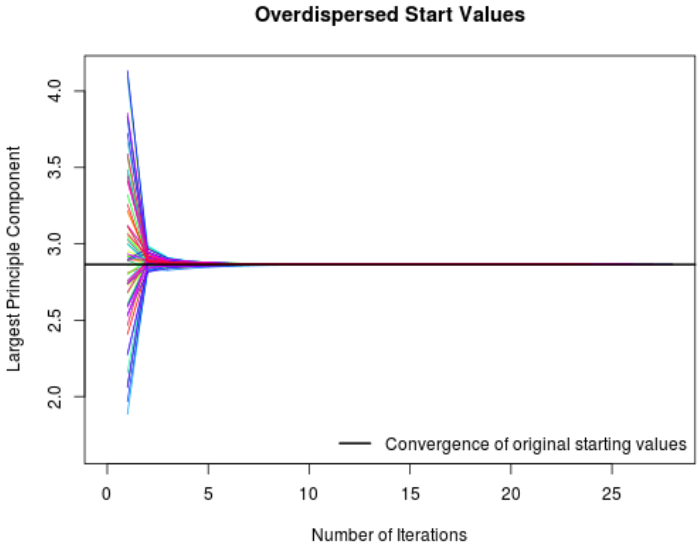

Figure 4.37: Overdispersion diagnostic (EMB)

4.10.2 Cox Regression Analysis

As mentioned in section 3.1.3, the analysis should be done in each imputed dataset and not only on the pooled data. Therefore, the selection of variables was realized in each of the imputed dataset. A variable was considered if it was significant in at least half of the models. The results presented in table 4.38 are an average of the imputed analyses.

The variables Radio, Origin, Job and Transf were discarded from the analyses. The next step consists in including all the significant variables in a model and remove one by one to see its effect. Results are in table 4.39.

One by one the variable with the lowest increase of -2 $\log \hat{\mathcal{L}}$ was removed from the model. Sex, Hmless, Unemp, Commu, Drugs, Smk and NumCo were discarded due to the low value of -2 $\log \hat{\mathcal{L}}$. The next step consists of adding each of the discarded variables in the univariate analysis to the complete model. None of the discarded variables had a significant increase of -2 $\log \hat{\mathcal{L}}$. The results presented in table 4.40 are a combination of the multiple analyses through Rubin's rules.

The hazard ratio, exp($\hat{\beta}$), was estimated for the variables (table 4.40). Adjusting for the remaining variables, it is estimated that:

- An individual that was not vaccinated has an increase of 56% in the risk of a recurrent episode when compared with a vaccinated individual;

Variable	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Null model	2543.587	-	
Vac	2535.981	0.0642	*
CliForm	2533.198	0.0004	***
Radio	2540.687	0.1127	
Sit	2409.896	0	***
Sex	2538.786	0.0344	**
Origin	2543.374	0.6657	
Job	2542.319	0.4325	
Alc	2526.303	4.1e-05	***
Smk	2508.066	3.2e-10	***
Drugs	2524.292	1.6e-05	***
Prison	2524.292	0.0077	***
Commu	2536.725	0.0095	***
Hmless	2537.032	0.0043	***
Transf	2543.535	0.8066	
Unemp	2529.851	6.8e-05	***
HIV	2502.973	9.7e-13	***
Diabetes	2533.647	0.0354	**
NumCo	2535.952	0.0034	***
age	2533.441	0.0021	***

Table 4.38: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests obtained in the univariate analysis (EMB). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	-2 $\log \hat{\mathcal{L}}$	*p*-value	
Model with all	2349.01	-	
Model without Vac	2355.92	0.092	*
Model without CliForm	2355.516	0.011	**
Model without Sit	2440.447	0	***
Model without Sex	2349.409	0.539	
Model without Alc	2354.942	0.027	**
Model without Smk	2350.534	0.282	
Model without Drugs	2350.333	0.353	
Model without Prison	2352.533	0.117	
Model without Commu	2349.982	0.420	
Model without Hmless	2349.454	0.600	
Model without Unemp	2349.327	0.589	
Model without HIV	2355.036	0.015	**
Model without Diabetes	2351.788	0.096	*
Model without NumCo	2350.071	0.309	
Model without age	2355.414	0.015	**

Table 4.39: Values of -2 $\log \hat{\mathcal{L}}$ and *p*-values of Likelihood ratio tests (EMB). Signif. codes: $<$ 0.01 '***' 0.05 '**' 0.1 '*'

	$\hat{\beta}$	exp($\hat{\beta}$)	se($\hat{\beta}$)	Z	*p*-value	CI (95 %)
Vac=1	0.443	1.557	0.255	1.737	0.0836	[-0.059;0.946]
CliForm=1	0.62	1.859	0.237	2.61	0.0091	[0.154;1.085]
Sit=1	1.118	3.059	0.099	11.272	0	[0.924;1.312]
Alc=1	0.531	1.701	0.198	2.681	0.0074	[0.143;0.92]
Prison=1	1.056	2.875	0.55	1.921	0.055	[-0.023;2.135]
HIV=1	0.66	1.935	0.2	3.297	0.001	[0.267;1.052]
Diabetes=1	-1.682	0.186	1.007	-1.669	0.0951	[-3.656;0.293]
age	-0.017	0.983	0.006	-2.547	0.0109	[-0.029;-0.004]

Table 4.40: Results using the Cox model (EMB)

- An individual with an extrapulmonary form of TB has an increase of 86% in the risk of a recurrent episode when compared with an individual with pulmonary form of TB;
- The risk of recurrence in individuals who defaulted the treatment is 30.6 times that of individuals who completed the previous treatment;
- An alcoholic individual has an increase of 70% in the risk of a recurrent episode when compared with an individual that is not alcoholic;
- The risk of recurrence in individuals incarcerated or working in a prison is 2.88 times that of individuals who are not in a prison;
- An individual with HIV has an increase of 94% in the risk of a recurrent episode when compared with an individual without HIV;
- An individual with Diabetes has a decrease of 81% in the risk of a recurrent episode when compared with an individual with Diabetes;
- Each additional year of age is associated with an estimated 2% decrease in risk. An additional decade corresponds to a 16% decrease.

It is important to observe than the confidence intervals are small, capturing a smaller range of effect sizes.

4.10.3 Residual Analyses

Schoenfeld Residuals

As discussed in 4.7.3, the Schoenfeld residuals were obtained for one imputed dataset.

The global value rejects the proportionality of risks (table 4.41), this is not unexpected since the sample size is quite large. Figure 4.38 shows that the variation in $\hat{\beta}(t)$ is relatively close to $\hat{\beta}$. The omission of important covariates can also be responsible for the presence of non-proportionality.

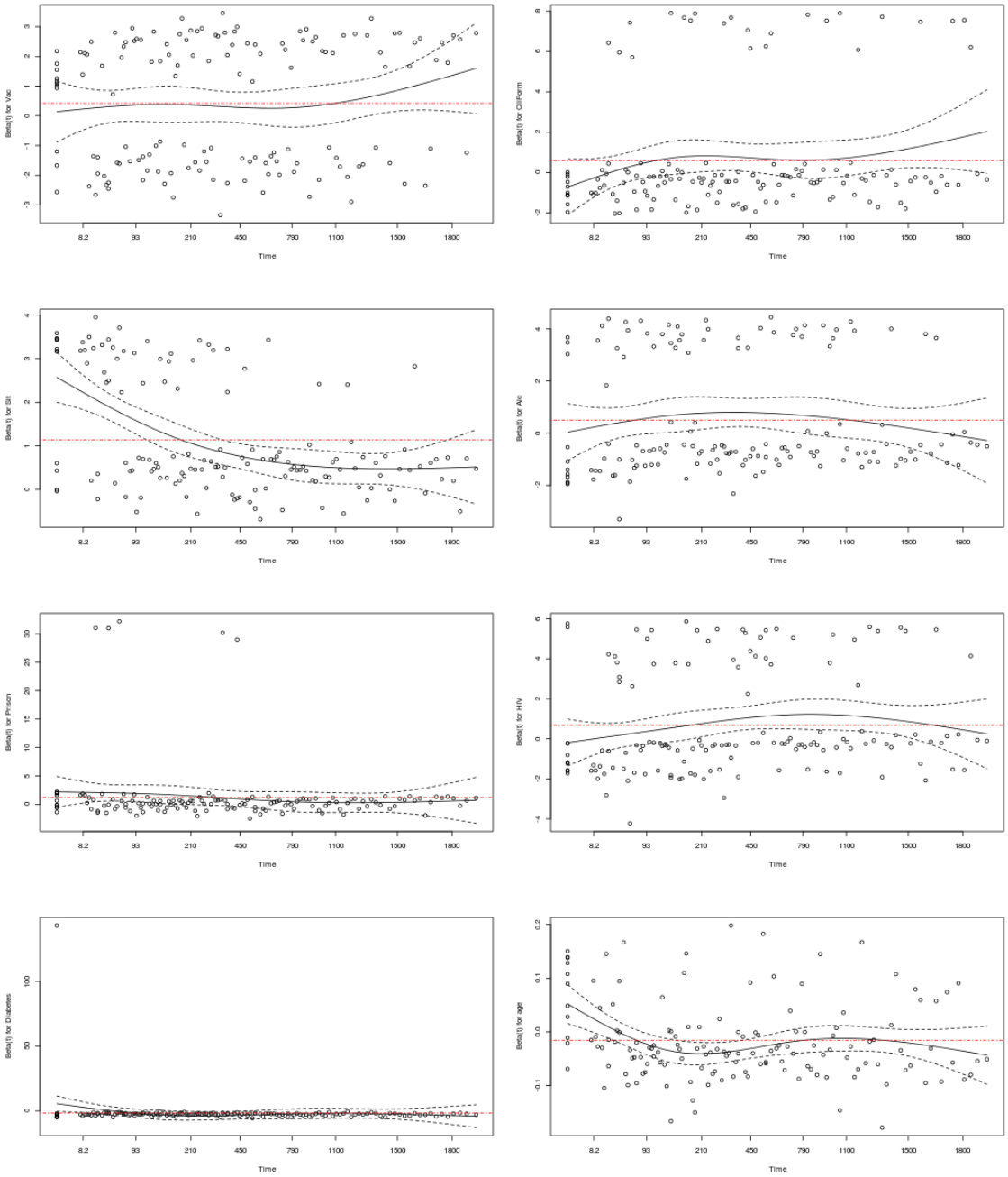

Figure 4.38: Plot of the Schoenfeld residuals (EMB)

	rho	χ^2	*p*-value
Vac=1	0.104	1.552	0.213
CliForm=1	0.159	3.858	0.050
Sit=1	-0.455	38.198	6.4e-10
Alc=1	0.018	0.047	0.828
Prison=1	-0.115	1.990	0.158
HIV=1	0.114	2.212	0.137
Diabetes=1	-0.126	2.280	0.131
age	-0.160	3.723	0.054
GLOBAL	NA	47.067	1.5e-07

Table 4.41: Test for proportionality of risks (EMB)

Martingale Residuals

In order to explore the fit of the models to individuals the martingale residuals were used. Similarly with the Schoenfeld residuals, the plot will be presented based in the same imputed dataset. The model is well fitted to the data since the residuals are evenly distributed above and under zero (figure 4.39).

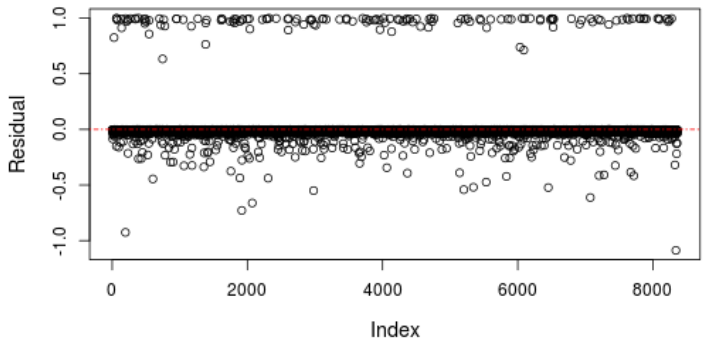

Figure 4.39: Plot of the martingale residuals (EMB)

4.10.4 Collinearity

The collinearity between covariates was also evaluated. When the VIF is approximately 1 the covariates are independent (table 4.42).

Vac	CliForm	Sit	Alc	Prison	HIV	Diabetes	age
1.103	1.069	1.147	1.086	1.041	1.158	1.007	1.112

Table 4.42: Average of the values of VIF (EMB)

4.11 Comparison between models

Table 4.43 summarizes the results obtained in each of the imputation methods that were used. The main results were grouped in a table for a better visualization and interpretation.

Through table 4.43 it can be observed that the use of different imputation tecnhiques did not change the direction of the parameter estimates and showed good consistency. In general, CC showed wider CI and larger standard error of the coefficient estimates than the other methods. CC did not positively associated Alcohol and Diabetes to recurrence, most likely due to the small sample size. Looking at each variable individually, the variable Vaccine was only selected using CC, mean imputation, EMB and RF **missForest**. This variable produced the most controversial results since it was only selected by these methods. CliForm was selected in all methods, the coefficient estimate and standard error is higher for CC and CC without Vaccine. The remaining imputation methods produced outputs that were rather similar. A similar behaviour was observed with the variable Sit. The estimates presented higher values for the methods CC and CC without Vac. Mean imputation, PMM, RF **missForest** and RF **mice** had similar values but EMB had a considerable smaller value for the coefficient estimate and its standard error. Only the CC analysis did not select the variable Alc, the other methods yielded similar values, with a higher value in the case of CC without Vac. Only the method PMM did not select the variable Prison. As previously, similar values were obtained for all the methods but CC has higher values due to the smaller sample. Compared with the other variables, Prison has a wider CI. The variable Commu was only selected for CC without Vac. HIV was selected in all methods, execpt for CC without Vac. Similar values were obtained, with higher values for the estimates in the CC analysis. Due to the issue of monotone likelihood, the variable Diabetes was not included in the CC analysis; however, it was selected in all the remaining methods. Similar values were obtained for this variable. It should be noted that Diabetes has the widest CI and highest standard error obtained in the imputation methods. The variable age was included in all models and the values showed a remarkable similarity between all the methods.

All the methods used yielded a low value of R^2, although in survival analysis values significantly smaller than 0.5 are usual. Values close to 1 are unlikely since it would mean that the model, in this case, would predict the exact day of the recurrence of TB. The values of R^2 are similar, although the value of R^2 obtained in the complete case analysis is slightly higher. Regarding the concordance index, which basically means that by randomly selecting two observations, the one with lower survival time is also the one that has the highest estimated risk, show similar results. All the models showed very good results with CC presenting slightly higher values.

Variables	CC β $(se(\beta))$ width(CI)	CC wt Vac β $(se(\beta))$ width(CI)	Mean β $(se(\beta))$ width(CI)	PMM β $(se(\beta))$ width(CI)	RF **mice** β $(se(\beta))$ width(CI)	RF **missForest** β $(se(\beta))$ width(CI)	EMB β $(se(\beta))$ width(CI)
Vac	0.590 (0.320) 2.450	-	0.544 (0.204) 0.800	-	-	0.560 (0.182) 0.714	0.443 (0.255) 1.005
CliForm	0.950 (0.450) 5.190	0.803 (0.302) 1.184	0.668 (0.235) 0.920	0.634 (0.235) 0.923	0.624 (0.235) 0.922	0.661 (0.234) 0.918	0.620 (0.237) 0.931
Sit	2.560 (12.98) 18.970	2.640 (0.226) 0.886	2.302 (0.192) 0.755	2.280 (0.193) 0.757	2.319 (0.193) 0.757	2.271 (0.193) 0.755	1.118 (0.099) 0.388
Alc	- -	0.532 (0.233) 0.913	0.589 (0.191) 0.747	0.574 (0.195) 0.764	0.563 (0.194) 0.763	0.531 (0.183) 0.718	0.531 (0.198) 0.777
Prison	2.330 (0.740) 40.940	1.714 (0.599) 2.346	1.433 (0.518) 2.032	- -	1.362 (0.520) 2.037	1.348 (0.518) 2.029	1.056 (0.550) 2.158
Commu	- -	1.076 (0.395) 1.548	- -	- -	- -	- -	- -
HIV	1.190 (0.360) 5.050	- -	0.752 (0.197) 0.773	0.706 (0.198) 0.776	0.707 (0.197) 0.775	0.685 (0.196) 0.771	0.660 (0.200) 0.785
Diabetes	- -	-1.425 (1.009) 2.849	-1.660 (1.007) 3.949	-1.680 (1.007) 3.360	-1.682 (1.007) 3.947	-1.717 (1.007) 3.947	-1.682 (1.007) 3.949
age	-0.030 (0.010) 0.050	-0.013 (0.007) 0.027	-0.014 (0.986) 0.006	-0.013 (0.006) 0.024	-0.012 (0.006) 0.023	-0.020 (0.007) 0.027	-0.017 (0.006) 0.025
$\boldsymbol{R^2}$	0.1173	0.0864	0.0854	0.0798	0.0822	0.0863	0.0851
$\boldsymbol{C}$	0.8168	0.7779	0.7952	0.7912	0.7913	0.7945	0.7948
$-2\log\hat{\mathcal{L}}$	624.640	1534.135	2354.097	2366.683	2361.307	2352.176	2354.901
Time	-	-	0m1.228s	27m35.431s	1560m12.561s	5m59.843s	2m18.997s

Table 4.43: Results obtained with the Cox model for all the different methods

Chapter 5

Discussion

The main goal of this thesis was to analyze recurrent cases of TB, in order to identify covariates that affect the time from the end of the first episode until the beginning of the second episode. To achieve this purpose, data from the SVIG-TB database was analyzed. Only patients diagnosed with their first episode of TB between 2002 and 2009, in Portugal, were included. The BCG vaccination was one of the most interesting variables to include in the analysis due to the long-lasting discussion about the protective effect of the vaccine. However, using this variable leads to a problem concerning the amount of missing data in the dataset. In fact, missing data ranged from a low value of 0.03% for Clinical Form to a high value of 59% for Vaccine. One possible explanation for the different amount of missing data is that each health center or health unit ask questions to the patient and uploads the information to the SVIG-TB. Since probably, some centers give more importance to some variables while others disregard them, this has led to many missing values in the current dataset.

Faced with this scenario, two options were possible. Drop all individuals with missing information and perform a complete case analysis only on one third of the data or impute the missing data and use all the available information. However, the literature advises against using complete case analysis in the presence of MAR data since complete case analysis yields biased results (52; 53). Therefore, it is important to explore and understand the missing data.

The exploratory analyses suggested an association between some variables and the missingness of values of other variables. Although it is impossible to prove if the data are MAR or MNAR, this association suggests that the data may be MAR. Hence, the best course of action was to impute the missing data through techniques of multiple imputation. A model was fitted to the complete dataset in order to compare the results with the results obtained by models fitted to imputed datasets. A model was fitted to the complete dataset without the variable Vaccine in order to understand the implications of discarding a variable with a large proportion of missing data. A "complete" dataset was obtained through mean imputation in order to compare the results of single imputation with multiple imputation. Several methods were used in multiple imputation: Predictive Mean Matching, two different models of Random Forest and a model of Expectation-Maximization with Bootstrapping. Predictive Mean Matching presents, in general, consistent results in literature. Recent research (67) argues that there is an advantage in imputing data with Random Forest and Expectation-Maximization with

Bootstrapping imputes data via the maximum likelihood.argues about the value of imputing data with RF and a model of EMB, which imputes data via the maximum likelihood.

The difference in the results between the model fitted to the complete dataset without the Vaccine and all the others models is striking. The model fitted to the complete dataset without the Vaccine was the only model that selected the variable Residence Community and the only one that discarded the variable HIV, which is frequently associated with recurrence in TB in the literature (29; 30; 98). Although the model has a good value of R^2, compared with the models obtained by imputation, this value is obtained only in a subset of the dataset. This subset does not seem to be appropriate to correctly infer about the population, since the results are not consistent with the other models. Extreme attention should be taken when one decides to remove a variable and perform a complete case analysis on the remaining variables, especially if the effect of the variable is unknown or important.

There were some differences between the models regarding the selection of variables. The variable Alcohol and Diabetes were not significant in the model fitted to the complete dataset, the latter was excluded from the model due to a problem of monotone likelihood. The models obtained from the "complete" dataset generated by mean imputation, RF **missForest** and EMB included the same variables (Vac, CliForm, Sit, Alc, Prison, HIV, Diabetes and age). The model used with the dataset obtained by PMM discarded the variable Prison and Vaccine. The results of the variable Vaccine were the most inconsistent. This variable was only significant in CC, mean imputation, RF **missForest** and EMB.

As expected, the coefficient estimates and standard errors of the model fitted to the complete dataset are higher than the ones from imputation. Although the values of R^2 and C are higher than in the models fitted to imputed datasets these values cannot be correctly compared since the models are based on different subset of individuals. The coefficient estimates and standard errors, between single and multiple imputation, are very similar, except for a slight increase of the coefficient estimate for the variable Prison. However, as previously discussed, mean imputation does not introduce variability into the model, ignoring that the values are not all true. Mean imputation results in small standard errors and in unbiased estimates but overestimates the precision (43; 54; 55), which could explain the high value of R^2 and C, compared with the value obtained for the datasets imputed by multiple imputation.

The model fitted to the data imputed by PMM is the only model, compared with the models obtained by multiple imputation techniques, that has not selected the variable Prison. Of all the models fitted to the data obtained from multiple imputation, it has the lower value of R^2 and C, although the results for the other variables are consistent and similar to the remaining models. Comparing the models fitted to the data imputed by techniques of RF, the coefficient estimates and standard errors seem similar and are consistent. Recent research (67) suggests that **mice** RF performs better than **missForest**, since the latter replaces missing values with predicted values rather than draw from a distribution, which leads to biased parameter estimates. However, in this case, the results obtained by these two RF techniques are similar and the major disadvantage of the **mice** package is that the imputation is much slower than with the package **missForest**. The value of $-2\log\hat{\mathcal{L}}$ is lower in the model imputed by RF **missForest**, and the values of R^2 and C are slightly higher in this model. Regarding the measures of explained variation, the model fitted to the data imputed by EMB has a slightly higher value of R^2 and C, compared with the other models obtained by multiple imputation,

and the value of the statistic $-2\log\hat{\mathcal{L}}$ is slightly lower than with the models imputed by PMM and RF **mice**. The fact that the standard errors are smaller in the models fitted to imputed datasets is an indicator of a good performance of the imputation methods. Although the value of R^2 is smaller in the models fitted to the imputed datasets, compared with the value of the model fitted to the complete dataset, these values should not be compared since the subset in analysis is different, which only means that for that subset the complete case analysis performs relatively well.

Overall, the results obtained with the different techniques are similar. In hte case of a variable with more than 30% of missing observations it is recommend to use multiple imputation instead of single imputation. Of the four techniques used, PMM seems to present the least promissing results, since the variable Prison, included in the remaining models, was not included in this model and although, the coefficient estimates and standard errors obtained are not very different from the other models, the values of the measures of explained variation used are worse. The R^2 is the lowest value of R^2 among the imputed models, and the value of the statistic $-2\log\hat{\mathcal{L}}$ is higher compared with the remaining models. Comparing the models fitted to the data imputed by the RF techniques and EMB, they seem to present similar results, except for the model fitted to the dataset imputed by RF **mice** that did not included the variable Vac. Among these three techniques, RF **mice** has a slighlty lower value of R^2 and a higher value of the statistic $-2\log\hat{\mathcal{L}}$. However, the major drawback of this technique is the time to execute the function, which is significantly larger than any of the two other methods. Although the package **missForest** has been previously associated with biased estimates (67), it does not seem to be the case in this study since the results are similar to RF **mice**. The results obtained from a imputed dataset by EMB and RF **missForest** are similar, although the imputation by EMB has produced slight differences in the coefficient estimates and standard errors. The values of C, R^2 and of the statistic $-2\log\hat{\mathcal{L}}$ are similar.

Under these conditions, both EMB and RF **missForest** seem to produce adequate results. It is not possible however, to select the "best" technique to impute missing data. Before imputing data, each dataset should be treated independently. A series of choices (such as the number of imputed datasets, the number of iterations, the method or methods to impute the data, how to incorporate interactions or non-linearities, etc) should be considered and they need to be treated carefully since wrong options could lead to incorrect estimates. Most software packages for multiple imputation have defaults for many of these points. However, these choices are not trivial and the user should carefully consider if the defaults are appropriate for his dataset. The literature has increased on this subject and many articles enlight the user about the most appropriate method for his dataset. However, most of the times, these studies are based on simulation studies. These are based on fully generated data that use models based on limited structures of the population to generate datasets that do not fully reflect a realistic population even if the attributes are based on real datasets (53).

After modelling the data, it is necessary to analyse the results. There are many challenges in managing and controling TB, including prompt diagnosis, effective treatment and successful prevention strategies. One serious problem is the recurrence of tuberculosis. The patient has to go through another round of treatment, which in some countries is more toxic, takes longer to complete and may amplify drug resistance.

Many studies argue about the protection of the BCG vaccine and its efficacy in preventing

pulmonary TB in adults. However, most of them just ignore this issue and do not include information about the BCG vaccine in their analysis. Other studies concer to the presence or absence of a BCG scar; however, the absence of a scar does not confirm the lack of vaccination. Due to this current discussion about the effect of the BCG vaccine, this variable was included in the model to assess its importance and role in recurrence. The estimated effect for the BCG vaccine ranges from 56% (model fitted by EMB imputation) to 80% (complete case analysis). The results obtained are mixed, since Vaccine was significant in some models and discarded in others. Nevertheless, they all point to an increase of more than 50% in the risk of recurrence, for an individual not vaccinated compared with someone vaccinated. Given that the vaccine efficacy declines with time (15 years after the vaccination the effect is negligible) further studies would need to be done to fully understand the efficacy of BCG and its role in recurrence.

The inclusion of the clinical form in a study about recurrence is unsual. The majority of studies discard the cases of extrapulmonary TB and performs the analysis only on the pulmonary cases. A reason for this could be that the proportion of extrapulmonary TB, in developed countries, is low (in this dataset the proportion of extrapulmonary TB is 8.4%) and less infectious than pulmonary TB. However, diagnosing extrapulmonary TB is not straightforward, since the clinical presentations (central nervous system, lymphatic system, genitourinary system, bones and joints, among others) are atypical and simulate other inflammatory and neoplasic conditions. This results in delay in treatment since it is necessary a high level of suspicion to make an early diagnosis. Past history of TB is frequently associated with pulmonary TB, although studies to determine if this is a result of reinfection or relapse are missing. However, a study in Nepal (99), which is a high-burden country, compared pulmonary and extrapulmonary TB and found that, after a primary infection in the lungs, the probability of reactivation at an extrapulmonary site could be higher at younger age, while reactivation of TB in the lungs was more common at older ages.

An individual that defaulted the previous treatment had between 3 times (in the model fitted to the dataset imputed by EMB) to 9 times (similar values for the remaining models) the risk to have a recurrent episode of someone who completed the treatment. This is not unexpected since someone who defaulted the treatment is likely to still be infected and therefore should continue the treatment later on. However, the study from Nepal (99) suggests that the retreatment could be due to extrapulmonary TB since the disease would have spread outside the lungs. Identifying patient characteristics that confer higher risk of default from primary TB treatment may help to establish prevention strategies to reduce the need for retreatment.

The results for the variable Alcohol were consistent for all the implemented models. An alcoholic individual has between 70% and 80% the risk to have a recurrent episode than someone who is not alcoholic. However, alcoholism has been identified as a predictor for treatment noncompliance (100). Treatment noncompliance is responsible for poor results in treatment, death, default and recurrence. Picon (100) studied risk factors for recurrence and found that alcoholism was more common among patients with poor adherence to treatment. In fact, when information about alcohol abuse and treatment noncompliance was included, alcohol abuse was no longer significant. The dataset analysed in this thesis did not had information about treatment noncompliance. Therefore, it seems that the variable alcohol may be a confounding factor, since it is related with treatment noncompliance and treatment noncompliance is a risk factor for recurrence.

Being incarcerated or working in a prison is a known risk factor for infection with TB. This is mainly due to overcrowding, late diagnosis and inadequate treatment (39; 40). The coefficient estimates obtained for this variable are more diverse, the risk of recurrence for someone incarcerated ranged from almost 3 times (model fitted to the dataset imputed by EMB) to 10 times (complete case analysis) that someone who is not in prison. The influence of this variable is mixed in the literature, where some studies argues its significant impact in recurrence while others do not find a significance. More research will need to be done to understand its association or lack of association and understand how the variable Prison relates to treatment noncompliance. Campani (101) studied the variables associated with treatment noncompliance and found a positive association with prison. However, only in cases where the patient escaped from the jail. Although it is expected that individuals in prison are more controlled with their medications than outside prison, our results suggest that this is not the case in Portugal. The problem of overcrowding and lack of good resources could be influential factors for treatment noncompliance and recurrence.

As expected, HIV was positively associated with recurrence of disease. The research focus on HIV patients due to their weak immunity system. In this study, only a variable Yes/No about whether the patient has HIV was included. It would have been interesting to add information about the severity of immunosuppression since it is also a predictor of TB recurrence (102).

The results of the coefficient estimates of Diabetes were not expected. This analysis showed that individuals with Diabetes have a decrease of 80% in the risk of a recurrent episode compared to individuals without Diabetes. This is the same to say that someone without Diabetes has an increase of 20% in the risk of a recurrent episode when compared with someone with Diabetes. Some studies report a positive association with Diabetes while others do not. Those who reported an association between recurrence and Diabetes always found a higher risk to have a recurrent episode of TB between individuals with Diabetes. However, these results should be interpreted bearing in mind the proportion of undiagnosed individuals with Diabetes. Some collaborators, of the Collective Dynamics Group at the Instituto Gulbenkian de Ciência, based in Brazil found a rate of 30% of undiagnosed individuals with Diabetes. In Portugal, Gardete-Correia (103), from the Portuguese Diabetes Association, estimated a proportion of 43% of undiagnosed individuals with Diabetes. This number is high and could lead to an underestimation of the true effect of the variable Diabetes in the recurrence of TB. Furthermore, Diabetes can have a confounding effect with treatment noncompliance, since individuals with Diabetes are used to prolonged daily intake.

Age is a risk factor for TB infection. However, only in some studies a positive association between recurrence and age was found. In this study, an increase of one year in the age of a patient leads to a decrease of around 1% to 2% in the risk of a recurrent episode. Some articles refer older age as a risk factor, which is correlated with a weaker immune system. Younger patients were sometimes associated with relapse. Picon (100) found that age was a confounding factor since it was related with treatment noncompliance. In fact, noncompliance was higher among younger patients. Since information abut treatment noncompliance was not included in the model, age could be acting as a confounding factor.

It is clear from the discussion that the dataset lacks important variables, such as, treatment noncompliance, which has been show to play an important role in recurrence, MDR (Multi Drug Resistance) infection, that was also associated with recurrence, especially among HIV

patients, and stage of the disease when starting treatment.

Some variables were not signicant in this study, but are frequently positively associated with recurrence. Smoking was not significant in this study, although some found a positive association. However, those studies also fail to include important variables such as treatment noncompliance. Male patients were also found associated with recurrence in some studies. However, this exploratory analysis suggests that these results could be due to a confounding factor, since, in this study, there is a higher proportion of males with HIV, alcoholic problems, smoker, etc., than females with HIV, alcoholic problems, etc.

In countries of low incidence, recurrence is usually attributed to relapse, while in countries of high incidence, the recurrence is mainly due to reinfection. Recurrence due to reinfection is a constant risk over time, whereas recurrence due to relapse seem to occur closer to the end of the first treatment. In this study, 55% of the cases of recurrence occurred in the first 12 months. The characteristics associated with recurrence seem to indicated that these cases may be due to relapse. HIV, an extrapulmonary form of TB and younger age are some of the characteristics frequently associated with relapse. Extrapulmonary TB is also less common among diabetic TB patients. Incomplete bacteriological cure, usually caused by irregular medication intake, is also a common cause of relapse. Some collaborators, of the Collective Dynamics Group at the Instituto Gulbenkian de Ciência, based in Brazil have found cases of reinfection 10, 20 and even 36 years after the first episode of TB. Future studies should consider longer follow-up times and inclusion of DNA information in order to properly distinguish between relapse and reinfection.

This study has some limitations, especially due to variables that were not available. Most questions are answered by the patient who may lie about important variables, such as drug and alcohol abuse, smoker, etc, introducing bias in the analysis. A possible solution for the missing variables would be a frailty model. A frailty model is a random effects model and it can be used to describe the influence of unobserved covariates not included in a proportional hazards model. Nevertheless, compared with the majority of studies made on TB, this study has the larger dataset studied. The sample size studied is frequently below 1000 individuals.

Chapter 6

Conclusion

This study reported risk factors associated with recurrence, in Portugal. Overall, not being vaccinated, having extrapulmonary infection, having defaulted treatment, being alcoholic, prison inmate or HIV positive are risk factors for TB recurrence. However, it is likely that some of these variables are correlated to variables that were not measured, particularly, treatmente noncompliance and the stage of the disease when starting treatment. Interestingly, having Diabetes and being of an older age confer protection against TB recurrence. These results suggests that the majority of the recurrence cases may be due to relapse since extrapulmonary TB, younger age and HIV are associated with relapse. Usually, relapse occurs closer to the end of the treatment while reinfection can occur over the time. However, a longer study with genotyping information should be performed to confirm these results. Few studies are performed on portuguese data and more studies would be needed to confirm these results.

Imputation was an important option to obtain the results. Deleting the missing observations would have led to biased estimates and loss of information due to the discarded individuals. Some researchers avoid imputation because of fear of "making up data" but they forget that complete case analyses require stronger assumptions than imputation does.

Before performing the imputation, the user needs to understand the type of missing data and the best methods to impute his dataset. Inadequate handling of the missing data in a statistical analysis can lead to biased or inefficient estimates. As a result, the first time user may get lost in a labyrinth of imputation methods. A proper review of the literature will shed some light into the different techniques to impute data and the most appropriate for his scenario. However, the best imputation approach remains unclear. In this thesis, both techniques of Random Forest performed well, although the package **mice** is very slow. Imputation via maximum likelihood also presented satisfactory results.

Inclusion of information about treatment noncompletion, drug resistance and genotyping data (to distinguish between relapse and reinfection) is essencial. A possible solution for the missing variables would be a frailty model. A frailty model is a random effects model and it can be used to describe the influence of unobserved covariates not included in a proportional hazards model.

Bibliography

[1] Lawn, S. D. & Zumla, A. I. Tuberculosis. *Lancet* **378**, 57–72 (2011).

[2] Thoen, C., Lobue, P. & de Kantor, I. The importance of Mycobacterium bovis as a zoonosis. *Vet. Microbiol.* **112**, 339–345 (2006).

[3] Niemann, S. *et al.* Mycobacterium africanum Subtype II Is Associated with Two Distinct Genotypes and Is a Major Cause of Human Tuberculosis in Kampala , Uganda. *J. Clin. Microbiol.* **40**, 3398–3405 (2002).

[4] Bentley, S. D. *et al.* The genome of Mycobacterium africanum West African 2 reveals a lineage-specific locus and genome erosion common to the M. tuberculosis complex. *PLoS Negl. Trop. Dis.* **6**, e1552 (2012).

[5] Pfyffer, G. E., Auckenthaler, R., van Embden, J. D. A. & van Soolingen, D. Mycobacterium canettii , the Smooth Variant of M . tuberculosis , Isolated from a Swiss Patient Exposed in Africa. *Emerg. Infect. Dis.* **4**, 631–634 (1998).

[6] Kremer, K. *et al.* Mycobacterium microti : More Widespread than Previously Thought. *J. Clin. Microbiol.* **36**, 2793–2795 (1998).

[7] Panteix, G. *et al.* Pulmonary tuberculosis due to Mycobacterium microti: a study of six recent cases in France. *J. Med. Microbiol.* **59**, 984–989 (2010).

[8] Association, A. L. Nontuberculosis mycobacterium. `http://www.lung.org/lung-disease/nontuberculosis-mycobacterium/`. Accessed May 22, 2013.

[9] Trunz, B. B., Fine, P. & Dye, C. Effect of BCG vaccination on childhood tuberculous meningitis and miliary tuberculosis worldwide: a meta-analysis and assessment of cost-effectiveness. *Lancet* **367**, 1173–1180 (2006).

[10] Konstantinos, A. Diagnostic tests Testing for tuberculosis. *Aust. Prescr.* **33**, 12–18 (2010).

[11] Lönnroth, K., Jaramillo, E., Williams, B. G., Dye, C. & Raviglione, M. Drivers of tuberculosis epidemics: the role of risk factors and social determinants. *Soc. Sci. Med.* **68**, 2240–2246 (2009).

[12] World Health Organization. Global Tuberculosis Report 2012. Tech. Rep., World Health Organization, Switzerland (2012).

[13] Brosch, R. *et al.* A new evolutionary scenario for the Mycobacterium tuberculosis complex. *PNAS* **99**, 3684–3689 (2002).

[14] Wirth, T. *et al.* Origin, spread and demography of the Mycobacterium tuberculosis complex. *PLoS Pathog.* **4**, e1000160 (2008).

[15] Daniel, T. M. The history of tuberculosis. *Respir. Med.* **100**, 1862–70 (2006).

[16] Institute, G. T. History of tb. `http://www.umdnj.edu/ntbc/tbhistory.htm`. Accessed May 29, 2013.

[17] Collier, R. Legumes, lemons and streptomycin: a short history of the clinical trial. *Can. Med. Assoc. J.* **180**, 23–4 (2009).

[18] Bhatt, A. Evolution of clinical research: A history before and beyond James Lind. *Perspect. Clin. Res.* **1**, 6–10 (2010).

[19] Storla, D. G., Yimer, S. & Bjune, G. A. A systematic review of delay in the diagnosis and treatment of tuberculosis. *BMC Public Health* **8** (2008).

[20] World Health Organization. Treatment of tuberculosis Guidelines. Tech. Rep., Switzerland (2010).

[21] Dye, C. Global Epidemiology of Tuberculosis. *Lancet* **367**, 938–940 (2006).

[22] Programa Nacional de Luta Contra a Tuberculose. Ponto da Situação Epidemiológica e de Desempenho. Tech. Rep., Direcção Geral de Sáude, Lisboa (2012).

[23] Shen, G. *et al.* Recurrent Tuberculosis and Exogenous Reinfection, Shanghai, China. *Emerg. Infect. Dis.* **12**, 1776–1778 (2006).

[24] Barreto, M. L., Pereira, S. M. & Ferreira, A. A. BCG vaccine: efficacy and indications for vaccination and revaccination. *J. Pediatr. (Rio. J).* **82**, S45–54 (2006).

[25] Soysal, A. *et al.* Effect of BCG vaccination on risk of Mycobacterium tuberculosis infection in children with household tuberculosis contact: a prospective community-based study. *Lancet* **366**, 1443–1451 (2005).

[26] Sreeramareddy, C. T., Panduru, K. V., Menten, J. & Van den Ende, J. Time delays in diagnosis of pulmonary tuberculosis: a systematic review of literature. *BMC Infect. Dis.* **9** (2009).

[27] Lambert, M. L. & Van der Stuyft, P. Delays to tuberculosis treatment: shall we continue to blame the victim? *Trop. Med. Int. Heal.* **10**, 945–946 (2005).

[28] Heller, R. F. *et al.* Prioritising between direct observation of therapy and case-finding interventions for tuberculosis: use of population impact measures. *BMC Med.* **4** (2006).

[29] Panjabi, R., Comstock, G. W. & Golub, J. E. Recurrent tuberculosis and its risk factors: adequately treated patients are still at high risk. *Int. J. Tuberc. Lung Dis.* **11**, 828–837 (2007).

[30] Pascopella, L., Deriemer, K., Watt, J. P. & Flood, J. M. When tuberculosis comes back: who develops recurrent tuberculosis in california? *PLoS One* **6**, e26541 (2011).

[31] Baker, M. A. *et al.* The impact of diabetes on tuberculosis treatment outcomes: a systematic review. *BMC Med.* **9**, 81 (2011).

[32] Balakrishnan, S. *et al.* High diabetes prevalence among tuberculosis cases in Kerala, India. *PLoS One* **7**, e46502 (2012).

[33] Liao, C.-M. *et al.* A probabilistic transmission and population dynamic model to assess tuberculosis infection risk. *Risk Anal.* **32**, 1420–1432 (2012).

[34] Ricks, P. M., Cain, K. P., Oeltmann, J. E., Kammerer, J. S. & Moonan, P. K. Estimating the burden of tuberculosis among foreign-born persons acquired prior to entering the U.S., 2005-2009. *PLoS One* **6**, e27405 (2011).

[35] Garzelli, C. & Rindi, L. Molecular epidemiological approaches to study the epidemiology of tuberculosis in low-incidence settings receiving immigrants. *Infect. Genet. Evol.* **12**, 610–8 (2012).

[36] Davies, P. D. O. *et al.* Smoking and tuberculosis: the epidemiological association and immunopathogenesis. *Trans. R. Soc. Trop. Med. Hyg.* **100**, 291–298 (2006).

[37] Lönnroth, K., Williams, B. G., Stadlin, S., Jaramillo, E. & Dye, C. Alcohol use as a risk factor for tuberculosis - a systematic review. *BMC Public Health* **8** (2008).

[38] Deiss, R. G., Rodwell, T. C. & Garfein, R. S. Tuberculosis and illicit drug use: review and update. *Clin. Infect. Dis.* **48**, 72–82 (2009).

[39] Nogueira, P. A. & Abrahão, R. M. C. d. M. Tuberculosis infection and the length of stay of County Jails prisoners in the western sector of the city of São Paulo A infecção tuberculosa e o tempo de Distritos Policiais da zona oeste da cidade de São Paulo. *Rev. Bras. Epidemiol.* **12**, 1–8 (2009).

[40] Baussano, I. *et al.* Tuberculosis incidence in prisons: a systematic review. *PLoS Med.* **7**, e1000381 (2010).

[41] Moons, K. G. M., Donders, R. A. R. T., Stijnen, T. & Harrell, F. E. Using the outcome for imputation of missing predictor values was preferred. *J. Clin. Epidemiol.* **59**, 1092–1101 (2006).

[42] Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T. & Moons, K. G. M. Review: a gentle introduction to imputation of missing values. *J. Clin. Epidemiol.* **59**, 1087–1091 (2006).

[43] Nur, U., Shack, L. G., Rachet, B., Carpenter, J. R. & Coleman, M. P. Modelling relative survival in the presence of incomplete data: a tutorial. *Int. J. Epidemiol.* **39**, 118–128 (2010).

[44] Vergouw, D. *et al.* Missing data and imputation: a practical illustration in a prognostic study on low back pain. *J. Manipulative Physiol. Ther.* **35**, 464–471 (2012).

[45] Rubin, D. B. Inference and Missing Data. *Biometrika* **63**, 581–592 (1976).

[46] Horton, N. J. & Lipsitz, S. R. Multiple Imputation in Practice : Comparison of Software Packages for Regression Models With Missing Variables. *Stat. Comput. Softw. Rev.* **55**, 244–254 (2001).

[47] Schlomer, G. L., Bauman, S. & Card, N. A. Best practices for missing data management in counseling psychology. *J. Couns. Psychol.* **57**, 1–10 (2010).

[48] van der Heijden, G. J. M. G., Donders, A. R. T., Stijnen, T. & Moons, K. G. M. Imputation of missing values is superior to complete case analysis and the missing-indicator method in multivariable diagnostic research: a clinical example. *J. Clin. Epidemiol.* **59**, 1102–1109 (2006).

[49] Rubin, B. D. *Multiple Imputation For Nonresponse in Surveys* (Willey).

[50] Horton, N. J. & Kleinman, K. P. Much ado about nothing: A comparison of missing data methods and software to fit incomplete data regression models. *Am. Stat.* **61**, 79–90 (2007).

[51] Goulão, B. *Seleção de variáveis na presença de valores omissos: uma aplicação na modelação do Índice de Massa Corporal nos imigrantes africanos e brasileiros residentes em Lisboa e Setúbal*. Master's thesis (2013).

[52] Demissie, S., LaValley, M. P., Horton, N. J., Glynn, R. J. & Cupples, L. A. Bias due to missing exposure data using complete-case analysis in the proportional hazards regression model. *Stat. Med.* **22**, 545–557 (2003).

[53] Marshall, A., Altman, D. G. & Holder, R. L. Comparison of imputation methods for handling missing covariate data when fitting a Cox proportional hazards model: a resampling study. *BMC Med. Res. Methodol.* **10**, 1–10 (2010).

[54] Sterne, J. A. C. *et al.* Multiple imputation for missing data in epidemiological and clinical research : potential and pitfalls. *Res. Methods Report.* **339**, 157–160 (2009).

[55] Stuart, E. A., Azur, M., Frangakis, C. & Leaf, P. Multiple imputation with large data sets: a case study of the Children's Mental Health Initiative. *Am. J. Epidemiol.* **169**, 1133–1139 (2009).

[56] Hippisley-Cox, J. *et al.* Derivation and validation of QRISK, a new cardiovascular disease risk score for the United Kingdom: prospective open cohort study. *BMJ* **335**, 136 (2007).

[57] Bodner, T. E. What Improves with Increased Missing Data Imputations? *Struct. Equ. Model. A Multidiscip. J.* **15**, 651–675 (2008).

[58] Baraldi, A. N. & Enders, C. K. An introduction to modern missing data analyses. *J. Sch. Psychol.* **48**, 5–37 (2010).

[59] Graham, J. W., Olchowski, A. E. & Gilreath, T. D. How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prev. Sci.* **8**, 206–213 (2007).

[60] White, I. R., Royston, P. & Wood, A. M. Multiple imputation using chained equations: Issues and guidance for practice. *Stat. Med.* **30**, 377–399 (2011).

[61] van Buuren, S. & Groothuis-Oudshoorn, K. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software* **45**, 1–67 (2011).

[62] von Hippel, P. T. How to impute interactions, squares, and other transformed variables. *Sociol. Methodol.* **39**, 265–291 (2009).

[63] von Hippel, P. T. Should a Normal Imputation Model be Modified to Impute Skewed Variables? *Sociol. Methods Res.* **42**, 105–138 (2012).

[64] Van Buuren, S. *Flexible Imputation of Missing Data* (Chapman Hall/CRC).

[65] Little, R. J. Missing-data adjustments in large surveys. *Journal of Business Economic Statistics* **6**, 287–96 (1988).

[66] Vink, G., Frank, L. E., Pannekoek, J. & van Buuren, S. Predictive mean matching imputation of semicontinuous variables. *Stat. Neerl.* **68**, 61–90 (2014).

[67] Shah, A. D., Bartlett, J. W., Carpenter, J., Nicholas, O. & Hemingway, H. Comparison of random forest and parametric imputation models for imputing missing data using MICE: a CALIBER study. *Am. J. Epidemiol.* **179**, 764–74 (2014).

[68] Breiman, L. & Cutler, A. Random forests. `http://www.stat.berkeley.edu/~breiman/RandomForests/cc_home.htm`. Accessed July 02, 2014.

[69] Breiman, L. Random Forests. *Mach. Learn.* **45**, 5–32 (2001).

[70] Stekhoven, D. J. & Bühlmann, P. MissForest–non-parametric missing value imputation for mixed-type data. *Bioinformatics* **28**, 112–118 (2012).

[71] Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum Likelihood from Incomplete Data via the EM Algorithm. *J. R. Stat. Soc.* **39**, 1–38 (1977).

[72] Do, C. B. & Batzoglou, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **26**, 897–9 (2008).

[73] Sousa, L. & Turkman, M. A. Complementos de estatistica (2013/2014). Lecture notes of Complementos de Estatistica course from Faculty of Sciences, Lisbon, Portugal.

[74] Honaker, J., King, G. & Blackwell, M. Amelia II: A program for missing data. *Journal of Statistical Software* **45**, 1–47 (2011).

[75] Honaker, J. & King, G. What to Do about Missing Values in Time-Series Cross-Section Data. *Am. J. Pol. Sci.* **54**, 561–581 (2010).

[76] Rocha, C. & Papoila, A. L. *Análise de Sobrevivência* (Sociedade Portuguesa de Estatística).

[77] Collett, D. *Modelling Survival Data in Medical Research* (Chapman Hall/CRC).

[78] Kaplan, E. L. & Meier, P. Nonparametric Estimation from Incomplete Observations. *Am. Stat. Assoc. J.* **53**, 457–481 (1958).

[79] Carvalho, M. *et al. Análise de Sobrevivência: teoria e aplicações em saúde* (Fiocruz).

[80] Mantel, N. & Haenszel, W. Statistical Aspects of the Analysis of Data From Retrospective Studies of Disease. *J. Natl. Cancer Inst.* **22**, 719–748 (1959).

[81] Mantel, N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50**, 163–170 (1966).

[82] Peto, R. & Peto, J. Asymptotically Efficient Rank Invariant Test Procedures. *J. R. Stat. Soc.* **135**, 185–207 (1972).

[83] Cox, D. R. Regression Models and Life-Tables. *J. R. Stat. Soc.* **34**, 187–220 (1972).

[84] Breslow, N. Covariance analysis of censored survival data. *Biometrics* **30**, 89–99 (1974).

[85] Schoenfeld, D. Partial residuals for the proportional hazards regression model. *Biometrika* **69**, 239–241 (1982).

[86] Grambsch, P. M. & Therneau, T. M. Proportional Hazards Tests adn Diagnostics Based on Weighted Residuals. *Biometrika* **81**, 515–526 (1994).

[87] Andreozzi, V. Modelo linear generalizado. `http://curso-glm.wdfiles.com/local--files/introducao/curso.pdf` (2012). Lecture notes of Modelos Lineares Generalizados course from Faculty of Sciences, Lisbon, Portugal - Accessed July 02, 2014.

[88] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria (2014).

[89] van Dijk, M. R., Steyerberg, E. W., Stenning, S. P. & Habbema, J. D. F. Survival estimates of a prognostic classification depended more on year of treatment than on imputation of missing values. *J. Clin. Epidemiol.* **59**, 246–253 (2006).

[90] Clark, T. G. & Altman, D. G. Developing a prognostic model in the presence of missing data: an ovarian cancer case study. *J. Clin. Epidemiol.* **56**, 28–37 (2003).

[91] Little, R. J. A. A Test of Missing Completely at Random for Multivariate Data With Missing Values. *J. Am. Stat. Assoc.* **83**, 1198–1202 (1988).

[92] Beaujean, A. A. *BaylorEdPsych: R Package for Baylor University Educational Psychology Quantitative Courses* (2012). R package version 0.5.

[93] Henry, A. J., Hevelone, N. D., Lipsitz, S. & Nguyen, L. L. Comparative methods for handling missing data in large databases. *J. Vasc. Surg.* **58**, 1353–1359 (2013).

[94] Heinze, G. & Schemper, M. A solution to the problem of separation in logistic regression. *Stat. Med.* **21**, 2409–2419 (2002).

[95] Heinze, G., Ploner, M., Dunkler, D. & Southworth, H. *logistf: Firth's bias reduced logistic regression* (2013). R package version 1.21.

[96] Heinze, G. & Schemper, M. A solution to the problem of monotone likelihood in Cox regression. *Biometrics* **57**, 114–119 (2001).

[97] by Meinhard Ploner, R. & by Georg Heinze, F. *coxphf: Cox regression with Firth's penalized likelihood* (2013). R package version 1.10.

[98] Sonnenberg, P. *et al.* HIV-1 and recurrence, relapse, and reinfection of tuberculosis after cure: a cohort study in South African mineworkers. *Lancet* **358**, 1687–1693 (2001).

[99] Sreeramareddy, C. T., Panduru, K. V., Verma, S. C., Joshi, H. S. & Bates, M. N. Comparison of pulmonary and extrapulmonary tuberculosis in Nepal- a hospital-based retrospective study. *BMC Infect. Dis.* **8** (2008).

[100] Picon, P. D. *et al.* Risk factors for recurrence of tuberculosis. *J. Bras. Pneumol.* **33**, 572–8 (2007).

[101] Campani, S. T. A., Moreira, J. d. S. & Tietbohel, C. N. Pulmonary tuberculosis treatment regimen recommended by the Brazilian National Ministry of Health: predictors of treatment noncompliance in the city of Porto Alegre, Brazil. *J. Bras. Pneumol.* **37**, 776–782 (2011).

[102] Chaisson, R. E. & Churchyard, G. J. Recurrent Tuberculosis - Relapse, Re-infection and HIV. *J. Infect. Dis.* **201**, 653–655 (2010).

[103] Gardete-Correia, L. *et al.* First diabetes prevalence study in Portugal: PREVADIAB study. *DIABETICMedicine* **27**, 879–81 (2010).

[104] Jr, F. E. H., with contributions from Charles Dupont & many others. *Hmisc: Harrell Miscellaneous* (2014). R package version 3.14-4.

[105] Templ, M., Alfons, A., Kowarik, A. & Prantner, B. *VIM: Visualization and Imputation of Missing Values* (2013). R package version 4.0.0.

[106] Therneau, T. M. *A Package for Survival Analysis in S* (2014). R package version 2.37-7.

[107] Jr, F. E. H. *rms: Regression Modeling Strategies* (2014). R package version 4.2-0.

[108] Liaw, A. & Wiener, M. Classification and regression by randomforest. *R News* **2**, 18–22 (2002).

[109] Lumley, T. *mitools: Tools for multiple imputation of missing data* (2012). R package version 2.2.

[110] Stekhoven, D. J. *missForest: Nonparametric Missing Value Imputation using Random Forest* (2013). R package version 1.4.

Appendix A

The aim of this appendix is to display the main functions created. The code necessary to obtain the results are also displayed below. Table A1 state the main R functions used.

Statistical Analysis	Function	Library	Reference
Analysis of missing data	`naclus`	**Hmisc**	(104)
	`naplot`	**Hmisc**	(104)
	`aggr`	**VIM**	(105)
	`LittleMCAR`	**BaylorEdPsych**	(92)
Firth's penalized likelihood	`logistf`	**logistf**	(95)
	`coxphf`	**coxphf**	(97)
Survival Analysis			
Cox models	`coxph`	**survival**	(106)
	`cph`	**rms**	(107)
Residuals	`cox.zph`	**survival**	(106)
	`resid`	**stats**	(88)
Collinearity	`vif`	**rms**	(107)
Predictive ability	`validate`	**rms**	(107)
Discrimination	`validate`	**rms**	(107)
Imputation			
Mean imputation	`na.roughfix`	**randomForest**	(108)
EMB	`amelia`	**Amelia II**	(74)
	`disperse`	**Amelia II**	(74)
	`imputationList`	**mitools**	(109)
MICE	`mice`	**mice**	(61)
	`quickpred`	**mice**	(61)
RF	`mice`	**mice**	(61)
	`missForest`	**missForest**	(110)

Table A1: R functions available in CRAN

Functions developed

```
# This function is used to calculate the value of -2 log
   Likelihood for models without imputation - not of mids
   class
```

```
# To use this function make sure to name your dataset
    completeImp
# This function returns a data frame with all the variables
     and the respective value of - 2 log Likelihood and the
    p-value

firstCollett <- function(x) {
  mod <- coxph(Surv(time, status) ~ completeImp[,x], data=
     completeImp)
  logL <- -2 * mod$loglik[2]
  p <- anova(mod, test = "Chisq")$Pr[2]
  resp <- data.frame(logL, p)
  names(resp) <- c("-2 log L","Pvalue")
  return (resp)
}

# This function is used to calculate the value of -2 log
    Likelihood for models imputed with Amelia - Package
    Mitools was used to convert the data to an appropriate
    format for analysis
# To use this function make sure to name your dataset
    impData
# This function returns a data frame with the value of -2
    log Likelihood for the null model and a model with
    variables, the p-value for the likelihood ratio test is
    also returned - for each m dataset a value is calculated
    .

firstCollettAmelia <- function(mod, anovaImp, m) {
  logLnull <- NULL
  logLcomp <- NULL
  p <- NULL
  for (k in 1:m) {
    fit <- mod[[k]]
    logLnull[k] <- -2 * fit$loglik[1]
    logLcomp[k] <- -2 * fit$loglik[2]
    p[k] <- anovaImp[[k]][5]
  }
  vals <- data.frame(logLnull, logLcomp, p)
  return (vals)
}

  # Function to help compute the values of -2logL for the
     second and third step of Collett method of selection
     of variables

secondCollettAmelia <- function(mod, modf, m) {
  logLcomp <- NULL
  test <- NULL
  p <- NULL
  for (k in 1:m) {
    fit <- mod[[k]]
    fitFull <- modf[[k]]
```

```
    logLcomp[k] <- -2 * fit$loglik[2]
    test <- lrtest(fitFull, fit)
    p[k] <- test$stats[3]
  }
  vals <- data.frame(logLcomp, p)
  return(vals)
}

# This function is used to calculate the value of -2 log
   Likelihood for models imputed with mice
# To use this function make sure to name your dataset imp
# This function returns a data frame with the value of -2
   log Likelihood for the null model and a model with
   variables, the p-value for the likelihood ratio test is
   also returned - for each m dataset a value is calculated
   .

firstCollettMI <- function(mod, anovaImp, m) {
  logLnull <- NULL
  logLcomp <- NULL
  p <- NULL
  for (k in 1:m) {
    fit <- mod$analyses[[k]]
    logLnull[k] <- -2 * fit$loglik[1]
    logLcomp[k] <- -2 * fit$loglik[2]
    p[k] <- anovaImp$analyses[[k]][5]
  }
  vals <- data.frame(logLnull, logLcomp, p)
  return (vals)
}

# Function to help compute the values of -2logL for the
   second and third step of Collett method of selection of
   variables for mice

secondCollettMI <- function (mod, modf, m){
  logLcomp <- NULL
  test <- NULL
  p <- NULL
  for (k in 1:m) {
    fit <- mod$analyses[[k]]
    fitFull <- modf$analyses[[k]]
    logLcomp[k] <- -2 * fit$loglik[2]
    test <- lrtest(fitFull, fit)
    p[k] <- test$stats[3]
  }
  vals <- data.frame(logLcomp, p)
  return (vals)
}

# This function calculates the VIF for each imputed dataset
    and returns a single vector with the average of the VIF
    in all imputed datasets
```

```
collinearity <- function(modf, m) {
  col <- NULL
  for (k in 1:m) {
    col[[k]] <- vif(modf$analyses[[k]])
  }
  vifs <- as.data.frame(do.call("rbind", col))
  return (colMeans(vifs))
}

# This function calculates the R^2 for each imputed dataset
    and returns a single vector with the average of the R^2
    in all imputed datasets

squareR <- function(modf, m) {
  r <- NULL
  for(i in 1:m) {
    r[i] <- modf$analyses[[i]]$stats[8]
  }
  return (mean(r))
}

# This function calculates the concordance index for each
   imputed dataset and returns a single vector with the
   average of the concordance index in all imputed datasets

concordance <- function(modf, m) {
  cHarrell <- NULL
  for (i in 1:m) {
    cHarrell[[i]] <- validate(modf$analyses[[i]], dxy=TRUE,
        B=1)
  }
  #get first element (Dxy) of index.origin
  listC <- lapply(cHarrell, `[[`, 1)
  return (abs(mean(unlist(listC)))/2+0.5)
}
```

Complete case analysis

```
#remove all observations with missing values
completeImp <- data1[complete.cases(data1),]

#calculate the value of -2 log Likelihood for the null
   model
modNul <- coxph(Surv(time, status) ~ 1, data=completeImp)

# index of the columns in the dataframe
vars <- c
   (1,2,4,5,6,7,9,10,11,12,13,14,15,16,17,18,19,20,23)

data.frame(variable = names(completeImp[vars]), t(sapply(
   vars, firstCollett)))
```

To obtain the final model, the value of -2 $\log \mathcal{L}$ of a reduced model was calculated manually and the *p*-value obtained compared the reduced model with the complete model. The code below is an example of the process of selection variables.

```
modComp <- coxph(Surv(time, status) ~ Sit + HIV + Prison +
   Vac + age + CliForm, completeImp)
comp <- -2 * modComp$loglik[2]

modReduce <- coxph(Surv(time, status) ~ Sit + HIV + Prison
   + Vac + age + CliForm + Radio, completeImp)
partial <- -2 * mod1$loglik[2]

anova(modReduce, modComp, test="Chisq")
```

Mean imputation

Only the function used to replace the missing values by the median of each variable is present, since the rest of the analysis is identical to the one presented in the Complete case section.

```
completeImp <- na.roughfix(data1)
```

EMB imputation

```
set.seed(72108)

#create matrix to give information about positive numeric
    variables
#In this case, Symp (data1[,20]) and age (data1[,22])
   have been atributed the lower value to 0 and the
   highest value correspond to the maximum value of the
   respective variable

posNum <- matrix(c(20,22,0,0,624,100), nrow=2, ncol=3)

amelia <- amelia(data1, m=70, noms=c("Vac", "CliForm", "
   Radio", "Sit", "Sex", "Origin", "Job", "Alc", "Smk", "
   Drugs", "Prison", "Commu", "Hmless", "Transf", "Unemp"
   , "HIV", "Diabetes", "NumCo", "status"), bound=posNum)
```

The following code is just a demonstration to use the function firstCollettAmelia. This step was not automated therefore, these lines need to be run for each variable. An example is displayed below.

```
mod <- with(impData, cph(survmod ~ Vac, x=T, y=T))
anovaImp <- with(impData, anova(cph(survmod ~ Vac)))
res <- firstCollettAmelia(mod, anovaImp, 70)
```

The following code is just a demonstration to use the function secondCollettAmelia. This step was not automated therefore, the model need to be updated every time a variable is removed. An example is displayed below.

```
modf <- with(impData, cph(survmod ~ Vac + CliForm + Sit +
    Alc + Prison + HIV + Diabetes + age, x=T, y=T))
logLfull <- NULL
for (k in 1:70) {
    fitFull <- modf[[k]]
    logLfull[k] <- -2 * fitFull$loglik[2]
}

mod <- with(impData, cph(survmod ~ Vac + CliForm + Sit +
   Alc + Prison + HIV, x=T, y=T))
res <- secondCollettAmelia(mod, modf, 70)
```

Through the lines below, the object is converted in order to pool the results.

```
# convert the object to the mids class in order to use
   pool to combine the results using Rubin's rules.
aMids <- datalist2mids(amelia$imputations)
modf <- with(aMids, cph(survmod ~ Vac + CliForm + Sit +
   Alc + Prison + HIV + Diabetes + age, x=T, y=T))
fit <- pool(modf)
```

PMM imputation

```
# Use matrix of predictors
predMatrix <- quickpred(data1, method="spearman")

imp <- mice(data1, m=70, maxit=10, pred=predMatrix, method=
   "pmm", seed=71152, printFlag=TRUE)
```

The following code is just a demonstration to use the function firstCollettMI. This step was not automated therefore, these lines need to be run for each variable. An example is displayed below.

```
mod <- with(imp, cph(survmod ~ Vac, x=T, y=T))
anovaImp <- with(imp, anova(cph(survmod ~ Vac)))
vals <- firstCollettMI(mod, anovaImp, 70)
```

The following code is just a demonstration to use the function secondCollettAmelia. This step was not automated therefore, the model need to be updated every time a variable is removed. An example is displayed below.

```
modf <- with(imp, cph(survmod ~ CliForm + Alc + Sit + age +
    HIV + Diabetes, x=T, y=T))
logLfull <- NULL
for (k in 1:70) {
  fitFull <- modf$analyses[[k]]
  logLfull[k] <- -2 * fitFull$loglik[2]
}
```

```
mod <- with(imp, cph(survmod ~ CliForm + Alc + Sit + age +
   HIV + Diabetes + Job, x=T, y=T))
vals <- secondCollettMI(mod, modf, 70)
```

RF - missForest

```
set.seed(4352)
imp <- missForest(data1, maxiter=10, ntree=15, verbose=TRUE
   )
completeImp <- imp$ximp
```

The analysis is performed similarly to complete case analysis since the `completeImp` is one imputed dataset.

RF - mice

```
# Use matrix of predictors
predMatrix <- quickpred(data1, method="spearman")

imp <- mice(data1, m=70, maxit=10, pred=predMatrix, method=
   "rf", seed=71152, printFlag=TRUE)
```

The analysis is performed with the functions `firstCollettMI` and `secondCollettMI`.