NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Alexandre Miguel Ramos de Sousa

BSc in Electrical and Computer Engineering

Survival outcomes and prognosis in non-small cell lung cancer patients in a tertiary hospital in Spain

MSc in ELECTRICAL AND COMPUTER ENGINEERING

NOVA University Lisbon

November, 2021

NOVA SCHOOL OF SCIENCE & TECHNOLOGY

DEPARTAMENT OF ELECTRICAL AND COMPUTER ENGINEERING

Survival outcomes and prognosis in non-small cell lung cancer patients in a tertiary hospital in Spain

ALEXANDRE MIGUEL RAMOS DE SOUSA
BSc in Eletrical and Computer Engineering

Adviser: Pedro Alexandre da Costa Sousa,
Associate Professor, NOVA University Lisbon

Co-adviser: Gracinda Rita Diogo Guerreiro,
Associate Professor, NOVA University Lisbon

Master of Science in Electrical and Computer Engineering

NOVA University Lisbon
November 2021

Survival outcomes and prognosis in non-small cell lung cancer patients in a tertiary hospital in Spain

Para a minha família.

ACKNOWLEDGEMENTS

First of all, I would like to thank my advisors, Pedro Sousa and Gracinda Guerreiro, and give a special thank you to Dr. Maria Torrente. I am honored to have been guided along this journey by people who love their work. They have always shown a great desire to contribute to this project that aims to help people to overcome one of the most difficult phases of their life. Thanks for all the help and patience to finish this cycle in the best way.
A thank you also to my colleague, Filipa Matos, for all the teamwork over the last months.

A thank you to FCT NOVA, which from the moment I came in made me feel at home and for everything that I was taught in the last few years.

To my family. Especially to my parents, Rui and Ana, who year after year never gave up on me, always with all the patience of the world, never ceased to support me and be there for me, a thank you will never be enough. To my siblings, Alexandra and Telmo, for all the companionship and love. A very special thanks to my grandfather, Celestino, for showing me from an early age that technology can exist in places where we will least expect it, even in the garage of the grandparents' house.

To my girlfriend, Mariana, for all the patience, for never giving up on encouraging me and showing that even in the most difficult moments, with effort and perseverance we can always achieve our goals. Thank you also to Carolina and Diogo for their friendship and patience in these long years, always ready to help me in all I need.

To my group of friends who have been with me for many years, with whom I grew up and learned that it is okay to continue to be children.

"Our greatest weakness lies in giving up. The most certain way to succeed is always to try just one more time"
(Thomas A. Edison)

ABSTRACT

Cancer is one of the main causes of mortality in the world and the number of new diagnosed cases are increasing every year. This number is expected to almost double in the next 20 years which causes health organizations to start taking steps to try to stop this increase in cases and to give the best possible care and treatment to cancer patients.

With the evolution of technology and its solidification and proven evidence in the health world, it is essential to create projects in order to guarantee the best care and monitoring for patients, to try to prevent the evolution of the disease and understand the type of care that patients need. With this, it is expected that cancer survivors will be able to have a better quality of life and an improvement in the survival rates.

The dataset used in this study is from patients diagnosed with lung cancer, one of the most common cancers and with a high mortality rate, specifically non-small cell lung cancer. The aim of the study is to identify risk factors that can affect patient survival. This dissertation discusses how information systems work in the area of health, how data are received, processed and stored. It is also explained how a pre-processing of the data was done in order to adapt the data to the models, a descriptive analysis to better understand our dataset and, lastly, a statistical survival analysis was performed using the Kaplan-Meier estimator, the logrank test and finally, the Cox multivariate proportional-hazard model.

This dissertation was carried out within the scope of the European project CLARIFY [1], with the collaboration of the oncology department of the University Hospital Puerta Hierro de Majadahonda.

Keywords: Non-small cell lung cancer, survival analysis, Kaplan-Meier estimator, logrank test, Cox proportional-hazard model.

RESUMO

O cancro é umas das principais causas de mortalidade no mundo e o número de novos casos diagnosticados tem aumentando todos os anos. É esperado que este número quase que duplique nos próximos 20 anos, o que faz com que as organizações de saúde comecem a tomar medidas para tentar impedir este aumento de casos e para que os pacientes com cancro tenham os melhores cuidados e tratamento possíveis.

Com a evolução da tecnologia e com a sua solidificação e provas dadas no mundo da saúde, é essencial criar projetos de forma a conseguir garantir os melhores cuidados e acompanhamento dos pacientes para tentar prevenir a evolução da doença e perceber quais os cuidados que os pacientes podem vir a necessitar. Com isto, espera-se que os sobreviventes de cancro consigam ter melhor qualidade de vida e melhorar as taxas de sobrevivência.

O dataset usado neste estudo é sobre doentes diagnosticados com cancro do pulmão, um dos cancros mais comuns e com uma grande taxa de mortalidade, mais especificamente com cancro do pulmão de células não pequenas. O objetivo deste estudo identificar fatores de risco que possam afetar a sobrevivência do paciente. Nesta dissertação é abordado como funcionam os sistemas de informação na área da saúde, como os dados são recebidos, processados e armazenados. Também é explicado como se realizou um pré-processamento dos dados para adaptar os dados aos modelos, uma análise descritiva para entender melhor o nosso dataset e, finalmente, uma análise estatística de sobrevivência foi realizada utilizando o estimador de Kaplan-Meier; o teste logrank e por fim, o modelo de risco proporcional multivariado de Cox.

Esta dissertação foi realizada no âmbito do projeto europeu CLARIFY [1], em estreita colaboração com o departamento de oncologia do Hospital Universitário Puerta Hierro de Majadahonda.

Palavras chave: Cancro do pulmão de células não pequenas, análise de sobrevivência, estimador Kaplan-Meier, logrank teste, modelo de risco proporcional de Cox.

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

ACS	American Cancer Society
ALK	Anaplastic Lymphoma Kinase
ASCO	American Society of Clinical Oncology
CLARIFY	Cancer Long Survivor Artificial Intelligence Follow-up
COPD	Chronic Obstructive Pulmonary Decease
CSV	Comma-separated values
CT	Chemotherapy
EGFR	Epidermal Growth Factor Receptor
EHR	Electronic Health Records
ESPEN	European Society for Clinical Nutrition and Metabolism
ETL	Extract, Transform and Load
HUPHM	Hospital Universitario Puerta de Hierro Majadahonda
IO	Immunotherapy
KDD	Knowledge Discovery in Databases
NCCN	National Comprehensive Cancer Network
NLP	Natural Language Process
NSCLC	Non-small cell lung cancer
QoL	Quality of Life
RT	Radiotherapy
SCLC	Small cell lung cancer
TKI	Tyrosine kinase inhibitor

1. INTRODUCTION

The aim of this chapter is to clarify the need and the importance of the study developed in this dissertation. The topic is discussed and the motivation of this project is addressed. Then, the problem under study will be explained, exposing some details that justify it.

1.1. BACKGROUND AND MOTIVATION

Cancer is a leading cause of morbidity and mortality worldwide [2]. According to Globocan, the number of cases diagnosed with cancer has been increasing over the years and in 2018 there were 18,1 million cases diagnosed all over the world, approximately 2,093 million (11,5%) of new cases detected and more than 1.8 million of deaths due to lung cancer [3]. It is expected that the number of cases will rise in the coming years, having an estimated amount of 29,5 million cases by 2040. By 2018, in Europe alone there were approximately 4,3 million new cases, from which 522 thousand were breast cancer and 158 thousand were lung cancer. In 2040 this number is expected to increase to 5,2 million new cases just in that year [4].

A cancer survivor is any patient who has been diagnosed with cancer for the rest of their life [5]. The treatment of cancer depends on the stage at diagnosis, that is, the size, the location and the characteristics of the tumor, among other factors.

In recent decades, due to early detection and effective therapies, the number of cancer survivors has been increasing at a very favorable rate [6]. Nonetheless, a survivor still needs medical care after the cancer has been treated, they often have persistent symptoms such as fatigue, pain, memory difficulties, sexual dysfunction as well as other problems that affect daily lives. For example, studies indicate that some chemotherapy drugs can damage the heart, or that women who have undergone chest radiation to treat childhood cancer have a 20-fold increased risk of developing breast cancer [7]. Other patients may have a genetic predisposition or lifestyle factors that put them at a higher risk of developing secondary

cancer. Follow-up and monitoring can help identify these problems earlier, thus enabling cancer survivors to have better health and quality of life [8].

As more patients leave primary care, there is a greater need for coordination of care and attention to health promotion and disease prevention in survivors [9]. Cancer survivors need follow-up plans and better cares after active cancer treatment. They also need proactive care, such as prevention plans, surveillance based on the patient's personal risks, medical interventions, genetic predisposition and lifestyle [10].

This dissertation was carried out in cooperation with HOLOS, SA in the European project CLARIFY – *Cancer Lung Survivors Artificial Intelligence Follow Up,* project SC1-DHT-01-2019, under the coordination of the *Medical Oncology Department* in the *Hospital Universitario Puerta de Hierro.*

In this data analysis, we will study patients diagnosed with lung cancer, in particular Non-Small Cell Lung Cancer.

1.2. PROBLEM AND PROPOSED SOLUTION

Cancer survivors currently face inadequate follow-up models, where their physical, functional and psychosocial needs are not met. Throughout their lives, they will possibly have new health problems, such as relapses, obesity, insulin resistance, depression, infertility, sexual dysfunction, anxiety, insomnia, heart problems, among many others. These health conditions can lead to psychological and social problems, which cause a reduction in productivity at work and in quality of life [11]. Therefore, these patients need adequate care and follow-up to manage the chronic effects of cancer [12].

The main focus of cancer monitoring has always been the detection of cancer recurrence, and it is frequently identified only after symptoms develop and it may be too late for the patient [13]. This method of monitoring is neither efficient nor sustainable. The follow-up routines in the hospital increase and overload the health services and these issues may compromise the needs of the cancer survivors. Consequently, there is a global need to transform and improve follow-up care to respond to the real needs of the patients [14]. To deal with these problems it is necessary to create methods that create individualized plans, during and after treatment, that help patients to have a better quality of life.

This thesis aims to identify post-treatment cancer survivors who have ongoing health and care needs and determine some factors that help predict poor health status. Relevant factors include relapses and secondary cancers, late and chronic effects of cancer and its

treatment and even patient characteristics. The goal is to collect and analyze the data from the Hospital, stratifying cancer survivors by risk. This information may be useful to the health system to identify the follow-up path to allow the patient together with the medical doctors organize adequate care and encourage healthy lifestyle behaviors to reduce risk in order to get the best assessment of their needs, improve their health status and their quality of life.

The study carried out in this dissertation was a starting point to the development of research papers and metrics that allow for a better understanding of the daily real problems that impact the lives of patients, doctors and relatives.

The use of quantitative statistical analysis brought an insight over the dataset that was performed for the first time in this study.

1.3. CONTRIBUTIONS

At this moment, the models created in this thesis are already being used in the dashboards of the Hospital Universitario Puerta de Hierro Majadahonda. Which allows doctors to have statistical information about the expected survival time for their patients, according to their characteristics.

Through this work, we managed to submit a paper for the RCP annual conference Medicine 2022.

1.4. OUTLINE OF THE DOCUMENT

This section summarizes the structure of the dissertation, which is presented as follows:

• Chapter 2 – State of the Art – This section intends to contextualize the cancer disease and the difficulties that patients go through. This introduction is necessary in order to understand the problem and to identify some of the risk factors that will have an important role in this study. The information systems where the patients' clinical data are kept is explained, as well as how these data are stored and processed until they reach the next stage, which is the application of machine learning models.

• Chapter 3 – Data Engineering - Here is an introduction of the tools used in the development and an explanation of the datasets used in the project, from its source and format to all the pre-processing work.

• Chapter 4 – Survival Analysis: as univariate approach – A descriptive analysis of the data is carried out, as well as a description of survival curves using the Kaplan-Meier estimator as well as the logrank test.

• Chapter 5 – Cox's Proportional-Hazards Model – This is where Cox's multivariable proportional hazard model's development, decision making, evaluation of models and results are presented.

• Chapter 6 – Conclusions and Future Work – In the last section, the work carried out since the beginning of this dissertation, the results and challenges proposed in the introduction are reviewed, as well as the work that still needs to be done.

2. STATE OF THE ART

This section intends to contextualize the cancer disease and the difficulties that patients go through. This introduction is necessary in order to understand the problem and to identify some of the risk factors that will have an important role in this study. The information systems where the patients' clinical data are kept is explained, as well as how these data are stored and processed until they reach the next stage, which is the application of machine learning models.

2.1. LUNG CANCER

Primary lung cancer is very heterogeneous in its clinical presentation, histopathology, and treatment response. Conventionally, lung cancer has been divided into NSCLC (non-small cells lung cancer) and SCLC (small cell lung cancer) [15]. For both groups, the cancer stage is the most significant predictor of survival [16].

There are several factors that can influence the probability of surviving a cancer, such as the stage when the cancer was diagnosed , the country [17], or the health conditions in which the patient is found[18].

2.1.1. RISK FACTORS AND SYMPTOMS

There are some factors that contribute to an increased chance of developing lung cancer [19], [20]:

- Tobacco – is seen as the main risk factor for the development of lung cancer, associated with approximately 90% of diagnoses, smokers have a 10 to 20 times higher risk compared to non-smokers. Approximately 20% of non-smokers who are diagnosed with cancer are passive smokers.

- Harmful substances - people exposed to certain substances such as asbestos, nickel, petroleum and derivatives, among others, can have an increased risk of developing cancer.
- Genetics - A family history of cancer increases the likelihood of developing lung cancer.
- Gender – studies indicate that there is an incidence three times higher in men than in women. Currently, this difference has been decreasing.
- Environmental pollution – There are a greater number of cases diagnosed in urban areas than in rural areas, suggesting that the development of cancer may be related to some air pollutants.
- Other diseases - some diseases such as chronic obstructive pulmonary disease (COPD) or lung fibrosis can lead to the development of lung cancer. There is still a subgroup of patients who developed lung cancer due to genetic alterations at the molecular level, it is estimated that they are between 10 to 20%, being the majority non-smokers.

There are several symptoms that can indicate the development of lung cancer such as persistent cough, chest pain when breathing or coughing, difficulty breathing, lung infections, fatigue, weight loss or poor appetite among other symptoms.

2.1.2 TYPES OF CANCER, DIAGNOSIS AND CLASSIFICATION

Lung cancer is divided into 2 groups:

- Non-small cell lung cancer (NSCLC)
- Small cell lung cancer (SCLC).

As already mentioned, our study is only about patients diagnosed with NSCLC, these are about 80% of the cases of lung cancer diagnosed, and the treatments are different.

The methods used to diagnose lung cancer are imaging techniques and analysis of tissue and cell samples:

- Image Scans - it is possible to verify the existence of the tumor through x-ray, magnetic resonance, computed tomography, among others.
- Biopsy – this is where the analysis of tissues and samples is carried out, to obtain the type of cancer.
- Blood tests – to find the correct number of some cell types in the blood.

After its detection and analysis, the stage of the cancer is determined according to its size, location, dissemination (spread) and if it has already affected other organs. Governed by an international classification system for malignant tumors, developed by the International Cancer Control Union (UICC), called TNM: [16]

- T refers to the size of the tumor, the larger the size of the tumor, the greater the stage:
    - T0, the tumor is microscopic in size (these patients were not included in this dissertation).
    - T1 (stage I), means that the tumor is in an early stage, and can be further divided into two subgroups, T1A and T1B.
    - T2 (stage II), in this stage the tumor is considered in an early stage, but the cancer has grown. It is divided into 2 subgroups, T2A and T2B.
    - T3 (stage III), the tumor is locally in an advanced state and can be divided into three subgroups, T3A, T3B, T3C.
    - T4 (stage IV), the tumor has already spread to other areas, reaching a metastatic stage.
- N refers to its spread to lymph nodes near the tumor site.
- M refers to the existence or absence of metastases, that is, the extension of the tumor located far from the tumor's origin site. [21]

This classification is one of the main factors to decide an individualized treatment that the patient will follow. The survival of the patient declines progressively with increasing clinical stage.

Approximately 70% of patients have been diagnosed in locally advanced stages (stage III) or metastatic disease (stage IV) [22], which contributes to low survival rates. Overall survival at 5 years in NSCLC is around 10–15% [23]. Early-stage NSCLC (stage I-II) patients are typically treated with complete surgical resection of the tumor. Yet, even after the entire resection of the tumor, 30–55% of patients will develop disease recurrence within the first 5 years of surgery [19].

There are two variables that can influence the choice of the type of therapy that should be given to a patient. It is a question of knowing if the patient has EGFR and AKL mutations. Patients whose tumors contain certain epidermal growth factor receptors (EGFR) or anaplastic

lymphoma kinase (ALK) gene mutations are continuously prescribed orally administered therapies.

We analyze the performance status (ECOG PS) of the patient, this variable is a measure developed by *Eastern Cooperative Oncology Group* [25], in order to describe the patient's level of functioning in terms of their ability to care for themself, daily activity, and physical ability:

- 0 - Fully active, the patient is able to carry on with all pre-disease performance without restriction
- 1 - Restricted in physically strenuous activity but ambulatory and able to carry out work of a light or sedentary nature
- 2 - Able to move and capable of all selfcare but unable to carry out any work activities;
- 3 - Only limited self-care; confined to bed or chair more than 50% of waking hours
- 4 - Totally disabled; unable to do any self-care; totally confined to bed or chair
- 5 - Dead

2.1.3 TREATMENTS

There are several treatments depending on the stage of cancer, the patient's clinical status and the performance status:

- Surgery - usually to remove the tumor.
- Radiotherapy - Doses of radiation to damage cancer cells and prevent their growth.
- Chemotherapy - Treatment used to destroy cancer cells. Can be used in combination with other treatments.
- Targeted therapy - Medicines that block the signal from the cells preventing them from growing.
- Immunotherapy - Is a treatment that aims to increase the body's defenses to fight cancer.
- A mix of treatments - Combining treatments is still common, depending on the patient's stage, conditions and comorbidities that the patient may have.

For this type of cancer (NSCLC), treatment is usually as follows:

- For stage I and II, surgical resection is the standard treatment, patients with IB or II are now being offered adjuvant chemotherapy (treatment after surgery).

- Although some stage IIIA tumors can be surgically removed, they often receive pre- or post-operative radiation and/or chemotherapy. Radiation plus chemotherapy is the standard of care for locally advanced, non-resectable stage IIIA tumors.

- Chemotherapy, in conjunction with supportive care, is still the usual treatment for patients diagnosed with stage IIIB, pleural effusion or stage IV.

- Patients who are in a more advanced state may receive more complex treatments, resulting in a greater likelihood and severity of impact on treatment, which can interfere with their quality of life, diet, and physical activity [11].

- Even in stage I or II patients, who are usually treated by surgically resecting the tumor, despite complete tumor removal, there is a possibility of recurrence [19].

Currently, the response rate and toxicity range varies significantly among patients, for a given chemotherapy or radiotherapy regimen, which is prescribed by a standard protocol with a fixed dosage. The survival time varies substantially, even after adjusting for the well-established variables, for instance, disease stage, histology, performance status and treatment.

After treatment, it is necessary to examine how the patient reacted and assess the evolution of the disease.

This assessment is classified as follows:

- Complete - Complete disappearance of the tumor.
- Partial - Tumor shrinkage, but still present.
- Stable - There were no significant changes in the tumor.
- Progressive - The tumor has increased in size.

The patient can have relapse or progression of the disease, making it necessary to discuss new lines of treatment.

2.2. CANCER SURVIVOR

In recent decades, progress in the early detection and treatment of cancer has contributed to an increase in life expectancy and, consequently, in the number of cancer survivors. Along with an increase in the incidence, better treatment, and detection of the disease at an earlier stage, the number of people living with or after cancer treatment grows every year. Besides, medical control of tumor growth and progression in patients with recurrent disease has led to a longer life span with an acceptable quality of life for survivors.[26]

In 1975, the probability of a person diagnosed with cancer surviving over the next 5 years was less than 50 percent, while in 2013 it was already close to 70 percent [27].

The survival of lung cancer patients has two critical attributes: survival time or quantity and quality of life (QoL).

2.2.1. QUALITY OF LIFE

The medical research community did not realize the importance of QoL vs. Survival after a lung cancer diagnosis until recently, mainly due to prevailing battle to prolong patients' live and the need to find a cure. After decades of efforts focusing on reducing lung cancer incidence and mortality, the new challenge for medical doctors is to understand the health conditions and QoL among people who survived lung cancer [28].

One of the more important factors impacting lung cancer patient’s survival and QoL is comorbidities [29]. Developing lung cancer is associated with age and smoking, and both age and smoking are strongly associated with comorbidity. Comorbidities, such as diseases of cardiovascular, pulmonary and other systems may influence prognosis in lung cancer as well as complicate its treatment. Because lung cancer is far more common in smokers and former smokers, these patients are more likely to have tobacco-related illnesses, which are primarily cardiovascular (ischemic or hypertensive heart disease, lower limbs arteriopathy, etc.) and respiratory (COPD, obstructive sleep apnea, usual interstitial fibrosis, etc.). They may also have additional comorbidities, such as diabetes and associated consequences, that are not related to tobacco use but are common in the general population (renal insufficiency, cardiovascular damage). These comorbidities can have a greater impact on a patient's performance status than the tumor itself [30].

Lung cancer is more common in the elderly, as aging is a risk factor in the development of lung cancer. Comorbidities become more common and severe as people age, regardless of the physiological changes that come with it. These comorbidities might have negative consequences for diagnostic techniques and treatment options, therefore they must be thoroughly investigated [29].

Health is a key factor at work [31]. After surgical treatment, the physical and psychological effects of treating cancer survivors can persist or worsen over time, even after 2 or 3 years, most patients still have sequels related to the surgical procedures [32]. Cancer survivors are at significant risk of unemployment [33] and are less likely to be re-employed [34]. Patients diagnosed with cancer may face several struggles, such as the presence of comorbidities, health problems and depression, all of which are heavily linked to job loss. Even so, many of them are willing and able to return to their jobs after treatments [35], [36].

One study investigated the association between dietary habits and mortality among cancer survivors, concluding that higher consumption of fish and vegetables was negatively associated with mortality. Diets based on high amounts of red and processed meat, refined grains, dairy products and desserts or higher alcohol consumption were associated with a higher risk of mortality [37].

Another study indicates that a reduction in the risk of all-cause mortality was associated with an increase in physical activity post-diagnosis [38].

If appropriate to their circumstances and physical condition, cancer survivors who have completed treatment are recommended to follow basic cancer prevention guidelines: be healthy, stay active, and monitor their diet.

Physical activity along with various forms of specific exercises are proposed to create multiple benefits for cancer patients [39]. There is substantial evidence that these benefits include increased blood pressure, reduced depressive symptoms, reduced fatigue, reduced therapeutic toxicity and improved quality of life [40].

According to Fund website guidelines, the people who have been diagnosed with cancer are at a significantly higher rate of developing a second primary cancer or other illnesses such as diabetes and cardiovascular disease [41].

The organizations, ACS, ASCO, NCCN and ESPEN recommend some guidelines and care to be taken to promote the health and quality of life of survivors, shown in FIGURE 2.1.

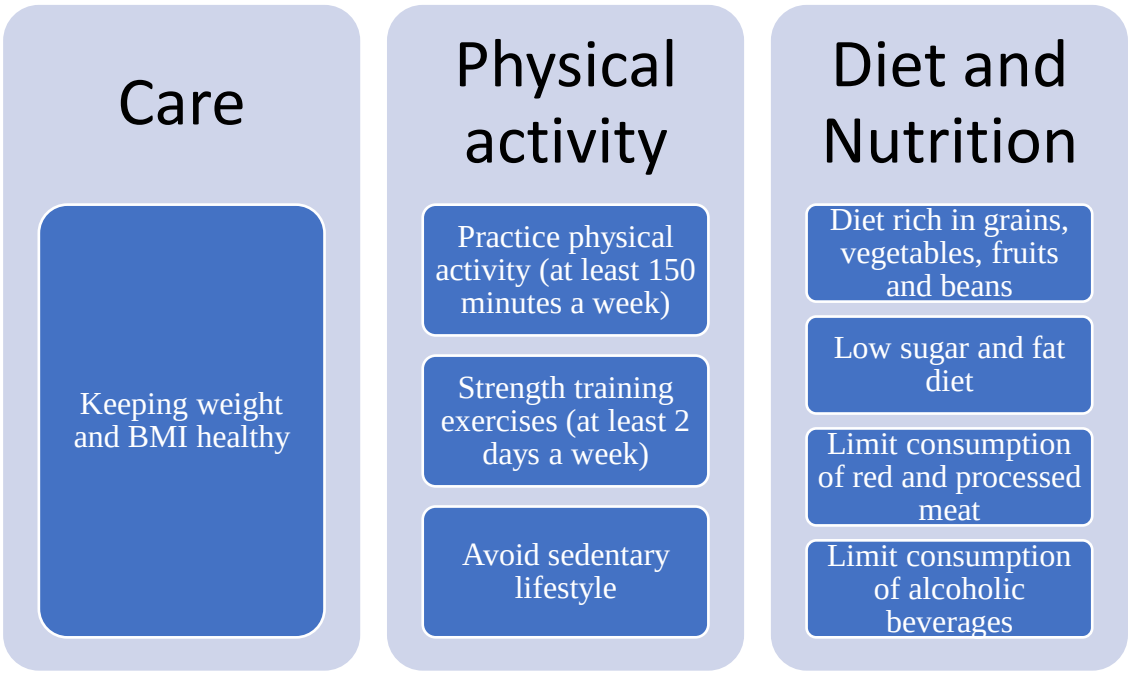


FIGURE 2.1 — GUIDELINES OF WCRF, ACS, ASCO, NCCN AND ESPEN

[12], [41]–[47]

2.2.2. PERSONALIZED PATHWAYS

Patient follow-up is neither efficient nor sustainable, typically, any cancer survivor is supervised for cancer recurrences in hospital appointments that vary in frequency, depending on treatment, illness and local practice. Routine hospital follow-up appointments place an increasing burden in the health service. Given that the primary focus is detecting recurrences, oncologists do not have enough time or opportunities to provide the required support or even make the appropriate recommendation to specialized services [48].

Patient care quality varies according to a number of criteria, including the geographic location, health care services, insurance, access to medications and treatment protocols.

Survivors report a wide range of physical, financial and psychosocial needs, as well as issues such as, for instance, cancer recurrences, cardiopulmonary, pain, fatigue, mobility, sexual dysfunction, depression, anxiety and cognitive problems. These problems can lead to reduced labor productivity [49] and a depletion in the quality of life [50].

2.2.3. TECHNOLOGY IN PERSOLIZED PATHWAYS

In recent years, efforts have been made to develop a tailored care plan with patients after the primary therapy has ended. When appropriate, these care plans motivate the patient to have their own management support, as well as the participation of multidisciplinary teams that are needed in order to focus on the complex needs of cancer survivors [51]. Other chronic disease populations have been benefited by this approach. These benefits include a reduction of the utilization of health services, an improvement in the quality of life and an increase patient satisfaction [52].

Clinical data in the care of cancer patients during treatment can help the identification of physical and psychological difficulties, follow them over time, enhance physician-patient communication, and include patients in decision-making, according to growing evidence from scientific trials [51].

Integrating clinical data with providing algorithm-based advice will assess the late effects of rehabilitation treatments and interventions, helping with the stratification of the risk and with the allocation to the proper care pathway.

Then, a basis for the clinical algorithms that guide the routes of care will be formed for risk stratification with specific criteria and follow-up monitoring recommendations, for instance, symptoms, routine imaging exams, tumor markers [53].

2.3. CLINICAL DATA

In order to perform future data analysis, it is necessary to collect information from patients, computerize the information so that it is possible to process the data and only then analyze and interpret the results.

2.3.1. EHR: ELETRONIC HEALTH RECORDS

In the last years, hospitals have started to adopt electronic health records [54]. EHR systems store the data associated to each patient, including demographic information, diagnoses, laboratory tests and results, prescriptions, radiological images, clinical notes and much more.

Current EHR standards provides a structure information in a format that supports specific data entry for further data analysis [55].

Databases have more and more data, which makes their analysis and consequently knowledge extraction much more complex and stressful. As a result, it was imperative to develop strategies to process this information and to extract useful knowledge for the intended objective, being necessary, in the first instance, to transform the data. In the following sections it will be explained in more detail how this process works.

In this project, as already mentioned, data is obtained through a hospital, which maintains a database of patients. But in order to apply rules and data analysis, this data has to be stored, processed, and transformed. So, for this data to be used there is a long and complex process that will be explained below.

2.3.2. ETL: EXTRACT, TRANSFORM AND LOAD

ETL, which stands for Extract, Transform and Load, is an essential task that is performed on raw data, this process allows you to gather information from various sources to obtain a data structure that can then be used for analysis and processing.

To be able to explore this type of data, it is necessary to define innovative instruments that allow the extraction and transformation of information. This task is complex and crucial, as there are two types of documents: structured documents and unstructured documents. For unstructured documents, it is necessary to use advanced extraction methods, including Natural Language Processing (NLP), because EHRs contain written medical and clinical information, patient notes, medical reports and diagnoses, and this information may not be inserted in a specific format [56].

The NLP process includes the extraction of medical concepts as well as the detection of events in order to structure all the information about each patient. Then comes the loading phase, which loads all the data to a new destination, transformed and ready to be used for analysis.

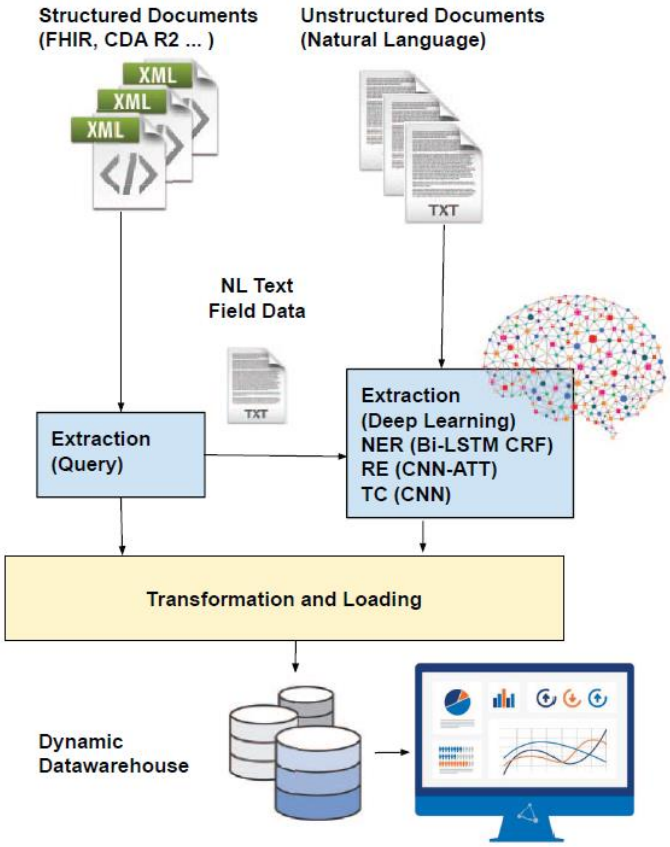


FIGURE 2.2 — ETL PROCESS [56]

The medical exam images contained in the reports are in DICOM (Digital Imaging and Communications in Medicine) format.

2.3.3. NLP: NATURAL LANGUAGE PROCESS

Natural Language Process is the process that will transform documents written by doctors into structured information. This is divided into two levels of tasks.

The low level consists of recognizing the text, that is, detecting abbreviations and titles, identifying the use of punctuation or diacritical marks between words (for example, hyphen), categorizing the words according to the context in which they are used in the sentence (names, verbs, adjectives, etc.), decompose certain words in order to discover their origin and their understanding is easier and group the constituents of the sentence into larger sections.

The high level, in turn, consists of analyzing and categorizing the data that were collected at the low level, that is, correcting spelling or grammatical errors, identifying specific words and categorizing them (diseases, medication, people...), determine the correct meaning of homographic words, identify statements that demonstrate denial of certain symptoms or

uncertainty of their occurrence, determine the relationships that exist between entities and events ("This cures/causes...", "This occurs with/ when…"), extract temporal references, order events chronologically and extract specific information and transform it into structured information [57].

2.3.4. KDD: KNOWLEDGE DISCOVERY IN DATABASES

Knowledge Discovery in Databases is referred to as the general process for extracting useful information from data analysis contained in a database. As we use this method directly from the database, it is necessary to carry out a study to know the data that we will use [58].

As the data we will study have already been submitted to ETL techniques, KDD is applied directly to the database. And we come to the next stage, using algorithms to try to identify patterns and predict and reduce the potential problems that cancer survivors might have.

This method consists of five distinct phases:

- Selection - This step serves to create a data set, where the data that fits the project objective is selected.
- Pre-processing - In this step, the selected data goes through a cleaning process, such as removing "noise" or out-of-the-box values or deciding how to handle empty data fields, to obtain a more subset of data. relevant and structured.
- Transformation - Aims to reduce the number of variables to be considered, through data transformation techniques,
- Data Mining - Of the most important steps, at this stage, it is necessary to choose the purpose of the model and choose the algorithm, depending on the objective we want, for example classification or regression, looking for patterns of interest.
- Interpretation – After all this process, it is necessary to interpret and evaluate the discovered patterns, review the entire process, and if necessary, remove redundant or irrelevant patterns. Then, conclusions about the acquired knowledge are drawn.

As we can see in FIGURE 2.3, it may always be necessary to go back and rectify problems found later.

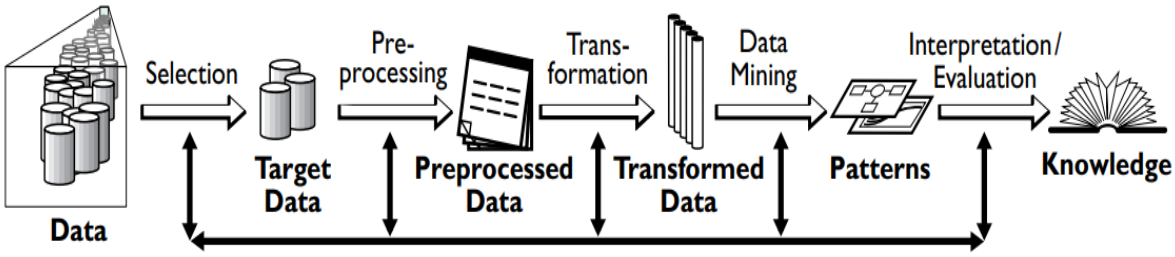


FIGURE 2.3 — DESCRIPTIVE PROCESS ABOUT KDD [58]

2.3.5. CONCERNS IN HEALTH DATA

The most common problems that EHR systems need to solve are security, privacy and confidentiality. Despite the fact that these subjects are closely related, they are not the same. The right to privacy refers to a person's ability to control when, how, and to what extent personal information is exchanged or shared by others. It is also necessary to pay attention to the confidentiality of the data since this type of information must be kept in private. Security, on the other hand, is defined as the degree to which the access to someone's personal information is restricted and only those who are permitted are allowed [59].

With the amount of data stored and increasing, there is a need to ensure that this data is secure, can only be accessed by authorized personnel. All the patient's personal information must be secure and protected so that it cannot be read or modified by unauthorized persons.

2.4. SURVIVAL ANALYSIS

Survival analysis has become one of the most important areas in statistics. The models and technology developed in survival analysis is now used in many fields, such as biology (survival time), engineering (time to failure), medicine (treatment effect or drug efficacy) or quality control (lifetime of components) [52].

Data analysis with medical knowledge, together, are essential for understanding the disease and the evolution of treatments to be developed. The analysis of the factors that influence the survival of patients the most may also prove to be decisive for understanding and improving the chances of survival of patients with cancer.

Not all individuals under study may have observed the occurrence of the event during the period in which these individuals are under observation which means that for those individuals only partial information is available. However, the mathematical methods used in

survival analysis allow these periods to be included in the sample for statistical analysis, by using censorship, this is explained in section 2.4.1.3.

We can also use other variables called covariates. Each covariate represents a risk factor that is supposed to affect the survival time.

There are several different methods that can be used in survival analyses. The FIGURE 2.4 shows some methods divided into non-parametric, semi parametric and parametric models.

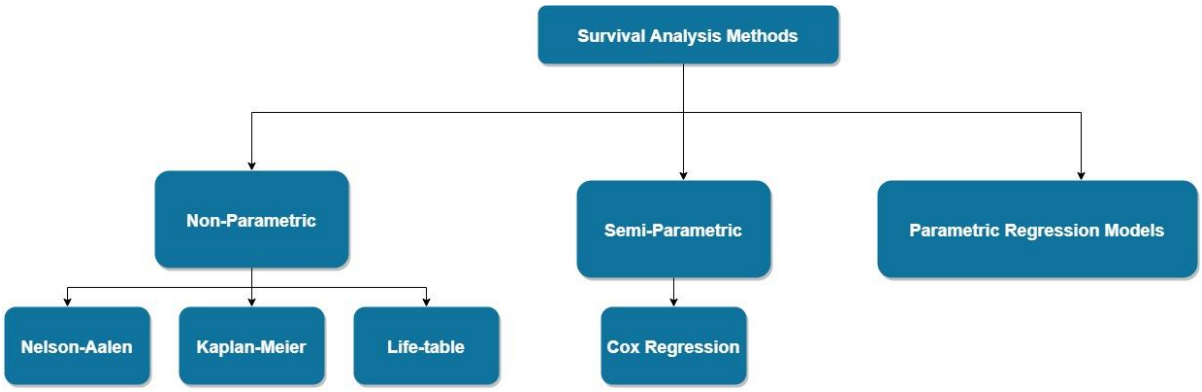


FIGURE 2.4 — SURVIVAL ANALYSIS METHODS.

Regarding non-parametric models, the two most used are the Kaplan-Meier, where we can assess the probability of survival x time, and the Nelson-Aalen estimator, where we estimate the cumulative hazard.

The Cox Model, which is semiparametric, assumes that failure rates are proportional. That is, the risk of failure of the variables is constant over time For example, the model assumes that the risk of dying for women compared to men is constant throughout the follow-up period of the study.

An alternative to the Cox Model is the usage of linear models for survival data. The use of parametric techniques, when well adjusted, present better results than non-parametric techniques, although they are not so flexible. The parametric models assume a probability distribution for the failure time with the Exponential, Gamma, Weibull and Log-Normal distributions being the most common ones.

There are some concepts that are common to the different models. All survival analysis models have as an outcome the estimation of the time until an event of interest occurs. Time can be weeks, months or years from the beginning of the study until the event, or the age of

the individual when the event occurs. Otherwise, the event can be the appearance of a disease or, in the case at study, death.

Once we define the event of interest and the measure of time, survival analysis has three main goals:

1. Estimate and interpret survivor and/or hazard functions from survival data to all the individuals from the population.
2. Compare survivor and/or hazard functions from different individuals. It is common to compare both functions for different values of an explanatory variable.
3. Assess the relationship of explanatory variables to survival time.

2.4.1. BASIC CONCEPTS

2.4.1.1. SURVIVAL FUNCTION AND HAZARD FUNCTION

Let the random variable T denote the survival time, that is the time until the event of interest occurs for an individual from the population in study. In this study, since the event of interest is death of cancer patients, the survival time will be the lifetime of an individual from the moment of diagnosis. T is a non-negative random variable, continuous throughout its domain [56].

We define as survival function and denote by $S(t)$ given by , the function that represents the probability that an individual survives longer than t, where t denotes a specified instant of time. In other words, the survival function gives the probability of the random variable T exceeds t.

$$S(t) = P(T > t), \quad t \geq 0 \tag{2-1}$$

Obtaining survival probabilities for different values of t, is very important to survival analysis because it allows us to obtain summary information from the dataset. This function (2-1) has the following properties:

- Is nonincreasing and continuous, that is, the function decreases as t increases;
- S(0) = 1, which means all patients are alive in the moment of diagnosis.
- $lim_{t\to\infty} S(t) = 0$, meaning that, nobody will survive if the time goes to infinity.

We can also characterize the survival time, T, using the hazard function, defined as

$$h(t) = lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t} \quad (2\text{-}2)$$

This function (2-2) describes the evolution through time of the instantaneous probability of death of an individual.

Both functions ((2-1), (2-2)) are very important for survival analysis. Although the survival function focus on not failing, that is the probability of not dying, the hazard functions focus on the risk of dying, which means that they provide complementary information from the same data set. Therefore, they have different purposes.

The Survival function, $S(t)$, directly describes the survival of the individuals in study and is more naturally and graphically appealing for survival analysis.

The hazard function, $h(t)$, is an instantaneous failure rate and allows us to analyze how the risk of dying will change throughout the time. Moreover, the mathematical modeling of survival data is usually based on the hazard function.

It is important to notice that regardless of which function is used or estimated in survival analysis, there is a relation between both, since it is possible to prove that,

$$S(t) = e^{\left[-\int_0^t h(u)du\right]} \quad (2\text{-}3)$$

and

$$h(t) = -\left[\frac{\frac{d\,S(t)}{dt}}{S(t)}\right] \quad (2\text{-}4)$$

Therefore, if S(t) is known, one can find h(t) and vice versa.

2.4.1.2. EXPLANATORY VARIABLES

In the survival analysis, it is necessary to consider that the survival curve of a patient can be affected by risk factors, such as some individual characteristics, treatments, habits, among others. When we introduce these factors to survival analysis, they are called explanatory variables or covariates.

Explanatory variables can be fixed over time or time dependent. We say that a covariate is constant if its value remains the same throughout the period in which the individual is under observation, e.g., the gender. If a covariate value changes over the observation period, we consider that it is time dependent.

Whenever explanatory variables are introduced in survival analysis, it is important to understand the real effect of that variable in the survival curve of an individual. For instance, it is quite intuitive that the age of diagnosis may have great influence in the survival time of a patient, when the study is about the survival time of an individual with a disease. However, if we do not have any scientific evidence of that fact, it is important that we do not jump into conclusions. There are some statistical approaches that allow us to understand the role of explanatory variables in the survival time, the real effect that they have in the survival function, but also how survival functions change for different categories from the same covariate.

2.4.1.3. CENSORED-OBSERVATIONS

As mentioned before, during the period of the study, it is possible that there are individuals for whom the event of interest does not happen. For those cases, there is incomplete information about the time of life. Therefore, the data from those individuals is censored. There are other reasons under which censoring may occur, e.g., if an individual quits the study or if it is not possible to keep a follow up. Whenever it is not possible to observe the exact moment at which the event of interest occurs, the information becomes censored. There are several types of censorship, and according to the type of limit that we have on the information, we can distinguish right censorship, left censorship or interval censorship.

We say that there is right censorship when it is only known that the lifetime exceeds a certain value, as the individual's observation ends before the event of interest occurs. For example, when the event of interest is death from a disease, patients who survived until the end of study or those whose contact was lost, and those who died from another cause were censored on the right.

Although the information is not complete, we know at least that until the end of the study the individual was alive. Right censorship is the most common type of censorship.

We can also describe some types of censorship on the right, designated by:

• Type I: When the period of the study is fixed in advance, with the number of deaths observed being random, e.g., a study planned to end after 2 years of follow-up.

• Type II: Appears when the study ends when a predefined number of events is observed, in this case the time period is random, for example a study planned for 2 years.

• Random censorship: The most general type of censorship. In this case, individuals enter the study randomly according to the date of diagnosis, if the study ends on a predefined date, the elapsed time of this individual is random.

Left censorship is when it is only known that it is less than a C time that has been recorded. In this case, the event of interest happened before the person went into observation, for example, carrying out a study at which age a child learns to perform a certain task, there are children who may have already learned to do it before the study

In the following subchapters the type of censorship is on the right.

2.4.1.4. TRUNCATED-OBSERVATIONS

Truncation is a mechanism that is used when due to a selection process inherent to the study planning only individuals to whom a certain event occurred are studied.

We can have two types of truncations:

- We say that there is left truncation when only individuals who satisfy a certain condition that was verified before the occurrence of the event of interest are included in the study. For example, exposure to an illness.
- We say that there is right truncation when the individual is observed only if the event of interest has occurred before a specific date, that is, if the survival time is less than a given value. For example, in a study in which the criteria for recruiting individuals are based on the occurrence of the event of interest before the end of the observation period, this type of truncation is commonly used in infectious diseases.

2.4.2. NON-PARAMETRIC MODELS

2.4.2.1. KAPLAN-MEIER ESTIMATOR

When there is no censorship, the survival function empirical estimator at instant t is given by the proportion of individuals who survived beyond that instant, i.e., the proportion of observed lifetimes with a value greater than t. Therefore, for a population with n individuals, the survival function is defined as follows [61]:

$$\hat{S}(t) = \frac{number\ of\ observations > t}{n}, \quad t \geq 0 \tag{2-5}$$

When there are censored observations, Kaplan and Meier, in 1958, proposed a non-parametric estimator of the survival function, known as the Kaplan-Meier estimator.

The Kaplan-Meier estimator is defined as the probability of surviving in a given length of time while considering time as many small intervals. This method is a univariate method i.e., only one variable is considered in the model [62].

Let $t(1), \dots, t(i)$ be the distinct death instants of a population with n individuals , di the number of deaths in $t(i)$ and ni the number of individuals at risk of death in $t(i)$.

The Kaplan-Meier estimator of the survival function is then given by [63]:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \frac{n_i - d_i}{n_i} = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right) \tag{2-6}$$

2.4.2.2. LOGRANK TEST

When we want to compare the lifetime distribution for several groups of individuals with different characteristics, the first approach is obtaining the plot of the Kaplan-Meier survival function for each group. It allows us to understand the impact of one specific characteristic in the survival time and see the differences between each group. However, for a statistical evaluation and to find out if there are significant differences between survival curves, looking for different survival curves is not enough. It is important to use statistical tests.

One of the tests you can use is the logrank test. In this test we tested the null hypothesis that there is no significant difference between the survival curves of two or more groups. Under this hypothesis, the distribution of the logrank statistic for $g \geq 2$ groups is approximately chi-square with $g - 1$ degrees of freedom.

Supposed that we intend to test if there are significant difference between the groups we are studying. The logrank statistic is approximately the sum of the quotient of the square of the total number of observed events minus total number of expected events by the number of expected events for each group, that is:

$$X^2 \approx \sum_{i}^{\#\ of\ groups} \frac{(O_i - E_i)^2}{E_i} \tag{2-7}$$

2.4.3. COX REGRESSION MODEL

The Cox Regression Model extends Kaplan-Meier survival analysis to assess simultaneously the effect of several risk factors on survival time. This is a multivariate model and allows for the inclusion of quantitative (e.g., age) and categorical (e.g., gender, type of treatment and smoking habits) risk factors. It is one of the most important methods used for modelling survival analysis data, especially in medical studies.

With this model we can examine how specified risk factors influence the rate of a particular event happening (death, for instance) at a particular instant of time (this rate is commonly referred to as hazard rate).

There are several risk factors that can influence a patient's lifetime. Cox [64] proposed a model in which at time t, the risk function is of the form:

$$h(t, \boldsymbol{X}) = h_0(t)\, e^{\sum_{i=1}^{p} \beta_i X_i} \tag{2-8}$$

where $\boldsymbol{X}$ = $(X_1, \dots, X_i)$ is a vector of explanatory variables , the coefficients $\beta_1, \dots, \beta_p$ are estimated from data and represent the impact of the covariates in survival time; and $h_0\ (t)$ is the baseline hazard (when all explanatory variables are equal to zero).

The Cox model allows us to estimate the h function (2-8), based on information from the patients under study and to understand the effect of each explanatory variable on the survival curve and test their significance.

Covariates have a multiplicative effect on the risk function, according to the factor $\exp(\beta_i)$ which is called relative risk, so this model assumes that the influence of covariates on the risk function does not change during the period in which individuals are under observation. Therefore, we say that $\boldsymbol{X}$ is time independent [61], [63], [65].

2.4.3.1. HAZARD RATIO

In Cox's regression model is necessary to have one baseline profile and it usually is the most common patient's profile. The risk of the covariate is calculated by comparing each value of the covariate with the value chosen for the baseline profile.

The hazard ratio (HR) is the quotient between the baseline hazard function and the hazard function (2-8) in which only one of the explanatory variables is different from 0, thus allowing to compare the hazard of a baseline patient with the influence of some risk factor (covariate).

According to the hazard ratio we can have different conclusions about the impact of each value of the covariate that we are studying in the risk function:

- HR=1 : The value has no impact on hazard. It means, there is no significant difference in risk when comparing with the baseline profile.
- HR<1 : The value represents a risk reduction when comparing with the baseline.
- HR>1 : The value represents an increase in the hazard function, it means patients with this value have an increased risk comparing with the baseline profile.

For the baseline, the HR = 1.

2.4.3.2. PROPORTIONAL HAZARD ASSUMPTION

In this model, it is necessary to consider an essential hypothesis to its application, the proportional hazard assumption. This hypothesis is characterized by a constant hazard ratio over time for the explanatory variables. So, when we apply the Cox model, after estimating the hazard functions (2-8) we have to verify the hypothesis.

To verify whether the model was effectively well-adjusted and, in this case, to understand if the hypothesis is verified, we use Schoenfeld residuals [66], which are obtained by the

difference between the observed value of the covariate for an individual whose event of interest occurs in t and the expected value of that covariate for all the individuals at risk in the same instant of time t.

Given an individual i, the Schoenfeld residual to the covariate z_j , $j = 1, \dots, p$ is given by:

$$r_{ji} = \delta_i \{ z_{ji} - a_{ji} \} \tag{2-9}$$

where

$$\delta_i = \begin{cases} 1 & if\ t_i\ is\ a\ non\ censored\ observation \\ 0 & if\ t_i\ is\ a\ censored\ observation \end{cases} \tag{2-10}$$

and

$$a_{ji} = \frac{\sum_{l\ \epsilon\ R_i} z_{jl}\ e^{\widehat{\beta}'\ z_i}}{\sum_{l\ \epsilon\ R_i} e^{\widehat{\beta}'\ z_i}} \tag{2-11}$$

The Schoenfeld residuals are defined for each covariate in the model and for every individual who has the event of interest. Once we obtain the Schoenfeld residuals for each covariate, we can verify if the hypothesis of proportionality holds. The idea behind the test is that if the proportional hazard assumptions holds for a covariate, then the Schoenfeld residuals for the same covariate will not be related to survival time.

Graphically we can verify if the proportional hazard assumption holds by checking if the plot of the Schoenfeld residuals (2-9) versus time looks like a random cloud of points around 0. If that happens then the model is well-adjusted.

2.4.3.3. OPTIMIZATION AND INTERPRETATION

To analyze and interpret the results from the model we need to consider following steps:

- Which variables are statistically significant and interpret their values.
- Optimize the model.
- Evaluate the model's performance.
- Checking the proportional hazards assumptions.

To understand which covariates are statistically significant, we perform the Wald test. This test is used in the Cox regression model to verify if the explanatory variables have significant influence in the survival time of the individuals.

The null hypothesis of the Wald test is the coefficient β_j which is associated to the covariate X_j being equal to 0. If the null hypothesis is not rejected the coefficient must not be removed from the model, which means that the covariate is significant to explain the survival time.

The statistic of the test is

$$\frac{\widehat{\beta_j}^2}{Var(\widehat{\beta_j})'} \tag{2-12}$$

and its asymptotic distribution is chi-squared with one degree of freedom.

To verify the hypothesis of the Wald test, we obtain the p-value which is the probability of the statistic of the test be greater or equal than the observed value under the null hypothesis.

For each value of one covariate, lower the p-value, greater the statistical significance.

In our application, as suggested by [63] when the p-value is above 0.10, the value in study is not significant compared to the value of the covariate from the baseline.

To optimize the model, when one of the values of the covariate is not significant, it may be plausible to readjust the covariate, by grouping values or classes of values, and the significance should be tested again. If the covariate cannot be changed, it can be removed from the model. From a clinical point of view, it may make sense to maintain in the model covariates that are not statistically significant, since they can still be relevant for the study.

Once decided which covariates to keep in the model, it can be analyzed the hazard ratio for each covariate, to understand the impact on the survival time, as explain in the section 2.4.3.1.

To evaluate the model, we can analyze the concordance index (c-index). This indicator allows us to quantify the predictive quality of the model. The c-index estimates the probability that a patient will survive longer than another if their risk of dying is lower.
A value of c = 0.5 corresponds to the average performance of a random model, while a value of c = 1 corresponds to a model capable of perfectly separating patients with different outcomes [67].

3. DATA ENGINEERING

In this chapter we present the tools used in the development, an explanation of the datasets used in the project, from its source and format to all the pre-processing work.

3.1. DATASET SOURCE

Before analyzing the data to obtain the expected results, we need to decide which frameworks and tools we will use. First, you need to get and load the data, these files can be in different formats. But, for data analysis and data mining, the most common approach to work on is receiving the data from a CSV or xlsx file or from a relational database.

There are many format files in which the information is encoded for storage in a file. CSV files are one of the most used to export and import data because of their flexibility. The two most used programming languages at this moment are Python and R Project because of their tools, which we use to work and manipulate the type of files mentioned earlier. The other solution is receiving the data in a relational database, this was how I received the data in the first moment.

The dataset studied in this dissertation was provided by HUPHM. These data are exclusively regarding patients diagnosed and treated in this Hospital. Along with the data, came a dictionary to better understand it.

3.2. USED FRAMEWORKS

Since we need to handle and manipulate data, and because the main focus of this dissertation is survival analysis, there are two languages we can think right away, Python and R Project. In this dissertation, the option relied on was Python, because this language has a wide variety of packages, is very popular in the data science and machine learning industries, and we can easily integrate with other languages like JavaScript, to perform web applications.

Anaconda was the platform used to access all Python libraries in this dissertation. This tool is an open-source platform designed to simplify package management, deployment, and perform data science and machine learning.

3.2.1. MySQL AND EXCEL

As already said, the first data received was through MySQL, this tool is an open-source relational database management system. As the name says, with this tool we can manage and maintain databases using SQL queries. The database is just like a table, with columns and rows, creating queries and we can extract the data needed. For example, we have 2 different tables with different information, and they are related with the id of one product, in our case, one patient and by creating one query we can obtain one single table with the information from the two tables.

As FIGURE 3.1 illustrates by using queries, it is possible to filter the database to obtain some variables, instead of getting all of them.

```
SELECT gender,patientStatus,stage_at_dx,age_of_diagnosis,smoker
FROM lung_cancer.patient;
```

FIGURE 3.1 — SQL QUERY - EXAMPLE

The results being the ones represented in FIGURE 3.2:

Result Grid | Filter Rows: | Export: | Wrap Cell Content:

gender	patientStatus	stage_at_dx	age_of_diagnosis	smoker
Male	dead	IV	53	former
Male	dead	IV	70	current
Male	dead	IV	48	no
Female	dead	IV	76	current
Male	dead	IV	55	former
Female	dead	IV	61	former

FIGURE 3.2 — SQL RESULT - EXAMPLE

While doing some treatment and analyzing , we found some errors and inconsistencies in our database, and since these patients were from the Hospital Universitario Puerta Hierro Madrid, we contacted them and they sent us another one in xlsx, and we kept it in xlsx, since our dataset is relatively small instead of importing it into MySQL.

3.2.2. PYTHON

To handle this data, our obvious choice was Python, and to manipulating these data the package chosen was Pandas.

Pandas is an open source library that is created and built on top of NumPy library, this allows us to obtain fast analysis, data cleaning and preparation [68].

Beside Pandas, the libraries we worked to perform the data management, treatment and machine learning were:

- MySql_Connector/Python - Enables Python programs to access MySQL databases, using an API
- *Lifelines* is a complete survival analysis library, written in Python, and we chose this library because it has an easy installation, like all the packages in Python, a wide variety of functions that make the construction and presentation of plots of the intended data and contains several modules such as:
    - Univariate Models, which include the KaplanMeierFitter.
    - Logrank test (logrank_test)
    - Regression Models, which include the SurvivalRegressions (CoxPHFitter) and testing the proportional hazard assumptions (CoxPHFitter.check_assumptions method).

Regarding the type of file, whether from CSV, xlsx or MySQL, Pandas receives the file as an input and converts it to a structure known as Dataframe. This structure is similar to a table from MySQL or Excel, with columns and rows. With this data structure we are able to organize data in order to obtain the data structure needed to apply survival analysis models.

In the FIGURE 3.3 there is an example of how we connect pandas with MySQL and the result.

```
In [1]: import mysql.connector as sql
        import pandas as pd

        db_connection = sql.connect(host='X', database = 'Y' , user = 'Z' , password = 'W')

        df=pd.read_sql('select gender,patientStatus,stage_at_dx,age_of_diagnosis,smoker FROM patient' , con=db_con)

        df
```

FIGURE 3.3 — PYTHON CONNECTING WITH SQL QUERY - EXAMPLE

Out[1]:

	gender	patientStatus	stage_at_dx	age_of_diagnosis	smoker
0	Male	dead	IV	53.0	former
1	Male	dead	IV	70.0	current
2	Male	dead	IV	48.0	no
3	Female	dead	IV	76.0	current
4	Male	dead	IV	55.0	former

FIGURE 3.4 — PYTHON VISUALIZATION OF A DF - EXAMPLE.

Once again, as we can see, Pandas has a similar approach to SQL. Using once again the library Pandas to manipulate Dataframes, we can gather several files by just having an id that relates the 2 files, each one read and converted using Pandas, and by creating another Dataframe with the desired data, illustrated in FIGURE 3.5.

```
In [ ]: df = pd.mrge(data1,df, left_on ='id' , right_on= 'id', how = 'left')
```

FIGURE 3.5 — PYTHON MERGING DF - EXAMPLE.

3.3. DATASET PREPARATION

The first necessary step is data preparation which also allows for an insight and understanding of the data. When we receive the database, it is raw, despite having already been processed by an ETL beforehand.

So, the first step is to study our data, know all the variables that we might use, cross-check data between variables to see if there are inconsistencies or any errors and adapt our data to be analyzed by the model.

3.3.1. DATASET STRUCTURE

As already mentioned, at first, we received the database that we were going to use in MySQL, it consisted of just a table with all the variables that contained all the information regarding each patient. Briefly describing, we had 278 variables that had personal data, demographic, treatment, and cancer characteristics information from 1050 patients.

After a thorough analysis of the data, we had some problems either due to a lack of information or because it was not possible to identify the patients' treatments or their biomolecular tests. Upon this, the Hospital sent us a new database in Excel format, and in this database, the structure was a bit different.

This new database, which was maintained until the end of this dissertation, contained 2133 patients, and was divided into 5 different tables:

- Patient socio-demographic information, such as gender, date of birth, date of diagnosis, age, smoking habits, biomolecular exams and comorbidities, among others that we will go into in more detail in the following subchapters.
- Treatments performed by each patient, for example: number of lines of treatment, type of treatment, medications that were used, response to the treatment, among others.
- Information about the patients' procedures, divided into 2 tables: a table about radiotherapy and another regarding surgery. Both tables have roughly the same information but referring to different procedures: if the patient underwent this treatment, the total amount of patients that did, the responses they had, as well as dates, among others.
- Finally, we have the last table, which is about the progressions and/or relapses of the patient. There we can find information on how many relapses or progressions occurred, the date, the type and even the location.

3.3.2. DATASET VARIABLES DESCRIPTION

The definition of the goals of this dissertation was set according with the aims of CLARIFY project together with the medical team from HUPHM. The understanding of the variables that should be integrated in the analysis of lung cancer patients was always evaluated and discussed with the medical team. Some of the available variables were subject to data transformation in order to better describe the problem and to fit the models.

After we merged some variables in one, according to the clinicians, and created others needed for the model. All the variables, specially to apply the Cox model, must be binary or numeric.

- Regarding the socio-demographics table, the following variables and their values were included in the Kaplan-Meier and Cox regression analysis.

Attributes	Values
Gender	Male (0) Female (1)
Date of diagnosis	Date
Date of birth	Date
Date of death	Date,
Smoking Habits	Non-smoker (0) Previous smoker (1) Current smoker (2) Unknown (3) NA (-1)
Comorbidities (Type of comorbidity)	No (0) Yes (1) Each comorbidity is one column with the value in binary
Patient Performance Status	{0 ,1, 2 , 3, 4}
Previous Cancer	No (0) Yes (1) Unknown (-1)
Family Antecedents Of cancer	No (0) Yes (1) Unknown (-1)
Patient Condition	Dead (0) Alive no disease (1) Alive with disease (2) Lost follow-up (3)

TABLE 3.1 — SOCIO-DEMOGRAPHICS TABLE

The molecular biomarkers are invasive and expensive, but sometimes they are needed because this information impacts the treatment of the patient. In order for the medical doctors to perform these tests, they need to have evidence that point to this diagnostic, otherwise, they will not perform them.

A new variable Mutated was created which TABLE 3.2 and TABLE 3.3 present our decisions. First, we started to create new variables to have the information if the patient is Mutated or Not. After that, we could categorize for each patient if the result of the test is Positive, Negative or Not Tested.

	Variable	Meaning	New Variable
Molecular biomarkers Test	Marmole___1	EGFR performed	
	Marmole___2	ALK Performed	
EGFR Test	egfrresultado___1, egfrresultado___2, egfrresultado___3, egfrresultado___4, egfrresultado___5, egfrresultado___6	Result EGFR: T790M, T790, Exon19, Exon21, NOS, Others	Mutated EGFR No (0) Yes (1)
Alk Test	alkresultado___2, alkresultado___3, alkresultado___5,	Result ALK: IHQ Positive, FISH translocated, RNA detected	Mutated ALK No (0) Yes (1)

TABLE 3.2 — MOLECULAR BIOMARKERS TEST

From the variables Mutated_EGFR, Mutated_ALK, marmole___1 and marmole___2 we are able to create the variable "Mutated":

- If the patient is mutated with ALK or EGFR, this means that the patient is Mutated.
- If the patient has not performed any tests, the result is Not Tested.
- All the other patients are Negative.

Variable	Values
Mutated	Positive
	Not Tested
	Negative

TABLE 3.3 — DECISION TO CREATE "MUTATED" VARIABLE

As we can see in TABLE 3.4, our dataset has 16 values for Stage at Diagnosis, as we already know, the stage is the way how the cancer is categorized and is the main factor to decide the best treatment and the prognosis of the disease.

Variable	Values		New values
Stage at diagnosis	I (24) I A (1) I A1 (25)	I A2 (26) I A3 (28) I B (2)	I
	II A (3) II B (4) II (27)		II
	III A (5) III B (6)	III (15) III C (18)	III
	IV (7) IV A (20) IV B (21)		IV

TABLE 3.4 — STAGE AT DIAGNOSIS

Since we had few patients for each stage, we grouped them by stage I, II, III and IV due to similarities on treatments disease characteristics.

Regarding the demographics table we created other variables such as dead, age at diagnosis and survival in months.

Variable	Condition	Values
Dead	Patient condition = 0	Dead (1)
	All the other patients	Alive (0)

TABLE 3.5 — DECISION TO CREATE VARIABLE DEAD

Variable	Condition
Age at diagnosis	Date of diagnosis - date of birth
Survival in Months	If patient is deceased, (Date of death - date of diagnosis) / 30
	If patient is alive, (Date of database - date of diagnosis) / 30

TABLE 3.6 — DECISION TO CREATE VARIABLE AGE AT DIAGNOSIS AND SURVIVAL IN MONTHS

Regarding the treatments, we have information from several treatment lines, such as the type of treatment, the initial and final date, the drugs used, if the patient is in palliative or active care, the response of the treatment, among others. The treatment lines change when the patient does not respond to the treatment, so a new line with different with different treatments is created. According to the information on the database, a patient can have up to 8 treatment lines, but these cases are very rare (3 in 2133 patients).

Regarding the type of therapy, a variety of attributes are available for this variable and we merged some therapies.

Attribute	New attribute
CT intravenous, Neo adjuvant chemotherapy, Adjuvant chemotherapy, Oral and intravenous chemotherapy, Oral chemotherapy	CT
Concomitant CT-RT, Sequential CT-RT, Adjuvant CT-RT, Neoadjuvant CT-RT	CT + RT

TABLE 3.7 — TREATMENTS MERGED

For this dissertation, we only analyze the first treatment line, and the variables used are presented in TABLE 3.8 and TABLE 3.9.

Variable	Values
Type of therapy – treatment line 1	CT (Chemotherapy) TKI (Oral targeted therapy) CT-RT (Intravenous chemotherapy + radiotherapy) IO (Immunotherapy) CT + IO (Intravenous chemotherapy + Immunotherapy) IO + TKI (Immunotherapy + Oral targeted therapy) No drug therapy

TABLE 3.8 — TREATMENT VARIABLES CREATED (Part I)

Variable	Values
Treatment Response – treatment line 1	1, Complete Response, 2, Partial Response, 3, Stable Disease 4, Progression, 5, No evidence of disease, 6, No evidence of relapse, -1, Unknown
Initial treatment Date – treatment line 1	Date
Patients Care – treatment line 1	Active, Palliative

TABLE 3.9 — TREATMENT VARIABLES CREATED (Part I)

In 2015, a new protocol of treatments was established due to new evidence in cancer treatments. In order to account for this differences, two binary variables were created.

Variable	Condition	Values
Treatment before 2015	Date of first treatment < 2015 or Date of diagnosis < 2015	0, No 1, Yes
Treatment after 2015	Date of first treatment > 2015 or Date of diagnosis > 2015	0, No 1, Yes

TABLE 3.10 — TREATMENT'S DATE VARIABLES

Regarding the radiotherapy procedure, each patient have the number of radiotherapies performed, date, type of radiation, response of the treatment, etc. It was only considered if the patient performed or not radiotherapy.

Variable	Values
Radiotherapy – treatment line 1	0, No 1, Yes

TABLE 3.11 — RADIOTHERAPY

As for surgery, each line has a lot of attributes such as the number of surgeries, the date, the type and the procedure and the response of the treatment. For our study, the only information used is if the patient performed or not a surgery in the first treatment line.

Variable	Values
Surgery – treatment line 1	0, No 1, Yes

TABLE 3.12 — SURGERY

In the progression/relapse table each patient has the information whether or not there was progression or relapse, the number, date, the type of progression and location.

Variable	Values
Progression – treatment line 1	0, No 1, Yes

TABLE 3.13 — PROGRESSION

Variable	Rule	Values
Procedures	If patient performed surgery OR surgery and radiotherapy	Surgery
	If patient performed only radiotherapy	Radiotherapy
	If patient didn't perform any	No Procedures

TABLE 3.14 — PROCEDURES

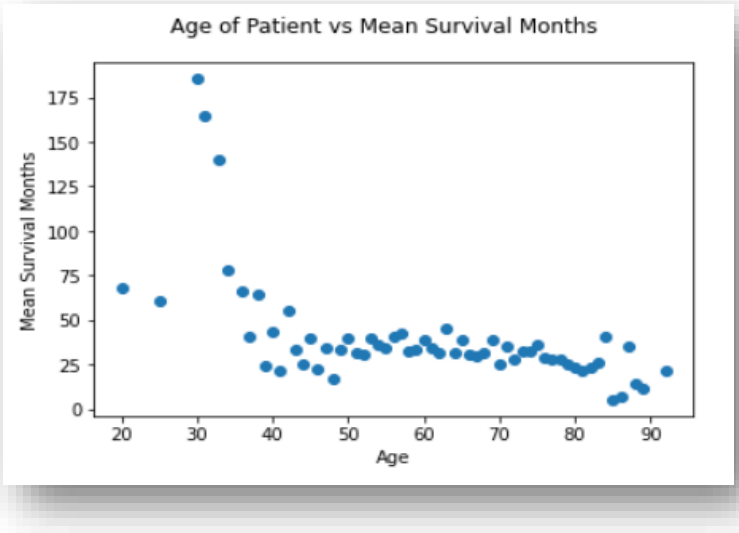


FIGURE 3.6 —AVERAGE SURVIVAL MONTHS FOR EACH AGE.

We may observe that for patients with ages that range between 45 and 75, the average months of survival does not have big fluctuations. Based on this information, as well as on the knowledge and sensibility of the medical team, the age of the patients will be analyzed for three different groups: [20,45], [46,70], [71,92].

4.

SURVIVAL ANALYSIS: AN UNIVARIATE APPROACH

Descriptive analysis, is the type of analysis that helps to describe, show, and summarize our database and helps us to understand the behavior of the data. Kaplan-Meier Estimation and logrank test were performed for each covariate.

After a more detailed analysis of the data, it was possible to verify that the first diagnosis of lung cancer recorded in the database occurred on January 22, 1999, and the last update of the database was 12 February 2021. The database contains a total of 2133 patients. As the data was updated in February of this year and there are lung cancer patients still alive, data is the right censored as previously described in section 2.4.1.3.

First our analysis consists of knowing the number of patients for each variable, their age and their survival time. For these two variables, we calculated the median and mean for age and survival time.

Having made all the decisions about the database, we move on to descriptive analysis and look for each variable. The values for each variable we are considering are the total of patients, the median of the age, the median of the survival time, the average of the age and the average of the survival time.

First, we started to exclude some patients from the database, because these patients were treated in another center, or because their histology was different from what was the purpose of the study, or even if the patients were lost to follow up or if they were treated in palliative care.

Our dataset has 69% of the patients dead and 31% alive.

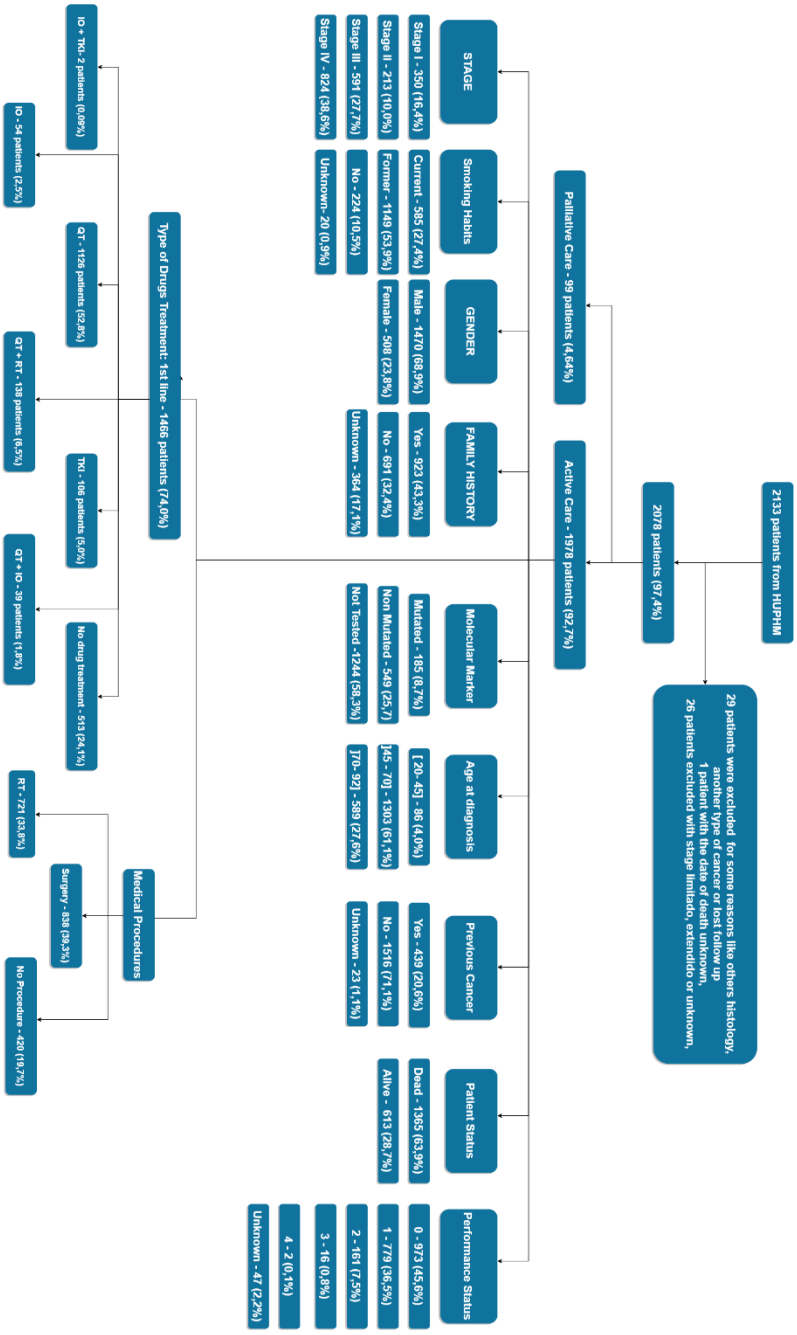


FIGURE 4.1 — DATASET DESCRIPTIVE ANALYSIS

Kaplan-Meier estimator is an example of univariate analysis and is one of the most popular methods used to analyze 'time-to-event' data. Our event is death and the time is the survival timeline of the patient since the cancer diagnose. This method is also one of the best options for measuring the proportion of subjects who survive a certain period of time after treatment. In addition to its estimates, the survival curve created is an estimate for the probability of surviving within a given length of time.

In order for our study to be completed, we will use the logrank test, see section 2.4.2.2, to evaluate if there are any significant differences between the groups being compared.

To declare if a variable is significant, we need to look to the p-value of the logrank test. Please note that our goal here is to find out if there are any differences between the groups we are comparing.

As decision criteria we consider that a p-value less than 10% reflects that there is a significant difference between the groups that we are comparing. We can divide our groups according to, for instance, their gender, age, stage at diagnosis, treatment method.

4.1. SOCIO-DEMOGRAPHICS TABLE

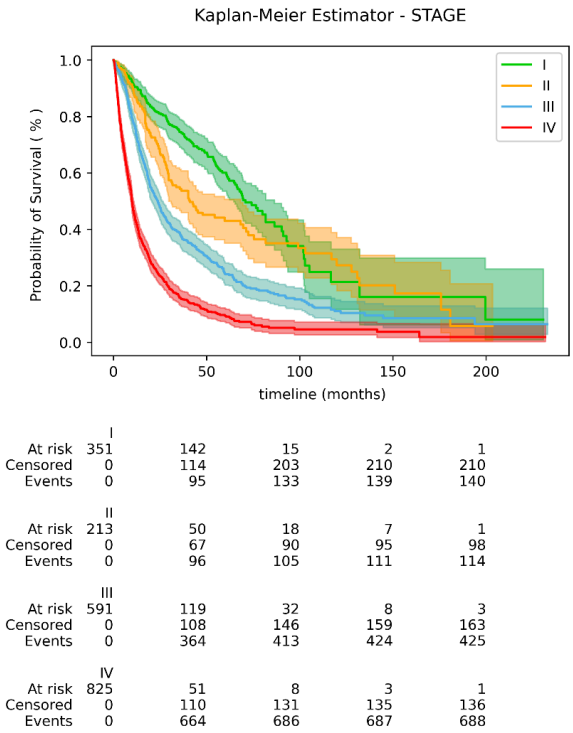
Looking through the FIGURE 4.1, our file has 2133 patients, and after excluding 56 patients, the dataset remains with 2077. After that, since the purpose of this study is just to study patients treated in this department, we also excluded the patients in Palliative Care. Making up a total of 1978 patients. Our descriptive analysis, Kaplan-Meier method and the logrank test will be applied to this dataset of patients.

4.1.1. STAGE AT DIAGNOSIS

Stage at diagnosis	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Stage I	350	17.69	67	66.8	41.8	43.1
Stage II	213	10.77	65	65.6	28.6	39.2
Stage III	591	29.88	66	65.2	19.4	31.6
Stage IV	824	41.66	64	63.5	9.5	16.1
TOTAL	**1978**					

TABLE 4.1 — STAGES AT DIAGNOSIS — DESCRIPTIVE STATISTICS

More than 70% of the patients is in advance stage, 41,7% of the patients in stage IV and almost 30% in stage III. Regarding earlier stages, we have 17,7% of the patients with stage I and 10,8% with stage II. Comparing the age with the diagnosis stage, we can notice that the patients diagnosed in early stages are older than the ones in advanced stages.



Logrank test:
p_value <0.005

FIGURE 4.2 — KAPLAN-MEIER — STAGE

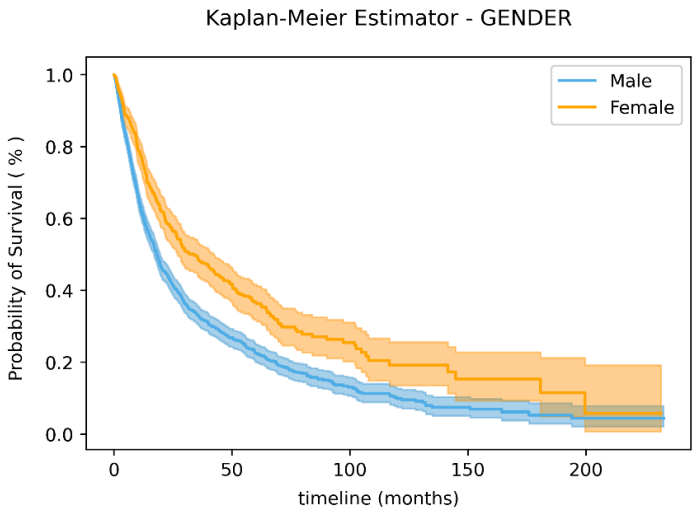
Analyzing the Kaplan-Meier estimates for stage at diagnosis, the results are as expected, patients diagnosed with stage I live longer than the others and patients with stage IV have approximately 50% of surviving over 12 months after diagnosing. Regarding the logrank test, we reject the hypothesis that the survival among the four stages at diagnosis is similar. In other words, the stage at diagnosis, in lung cancer patients, has impact on the expected survival of the patients.

4.1.2. GENDER

Gender	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Male	1470	74.32%	67	65.7	14.8	26
Female	508	25.68%	62	62	22	33.6
TOTAL	**1978**					

TABLE 4.2 — GENDER — DESCRIPTIVE STATISTICS

There is a considerable difference between the survival of men and women diagnosed with lung cancer, almost 75% are men and we can notice that women survive longer than men.



Male					
At risk	1472	244	44	14	5
Censored	0	267	382	396	401
Events	0	961	1046	1062	1066
Female					
At risk	508	118	29	6	1
Censored	0	132	188	203	206
Events	0	258	291	299	301

Logrank test:
p_value <0.005

FIGURE 4.3— KAPLAN-MEIER — GENDER

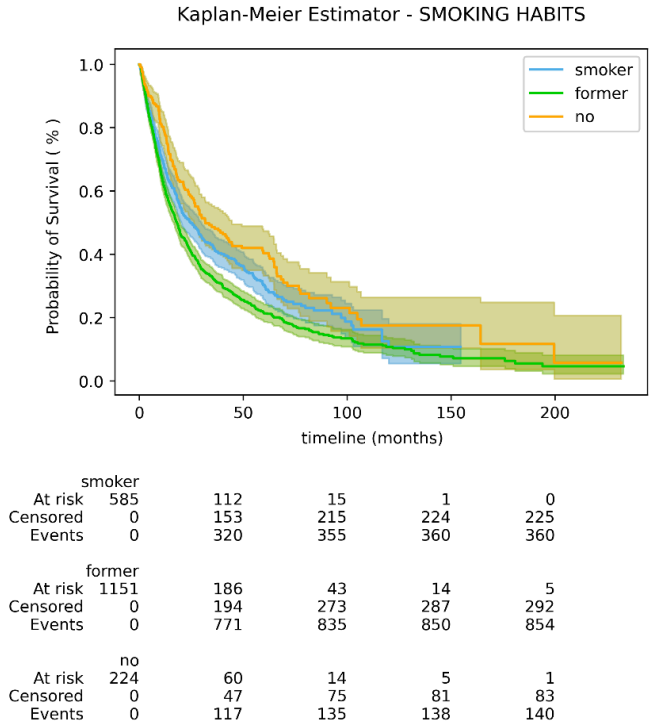
According to logrank test and Kaplan-Meier estimator we may observe significant differences in the survival of men and women.

4.1.3. SMOKING HABITS

Smoking Habits	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Current	585	29.58%	62	62.5	17.9	27.3
Former	1149	58.09%	67	65.7	14.8	26.6
No	224	11.32%	67	65.6	24.9	36.7
Unknown	20	1.01%	71	65.2	9.1	25.6
TOTAL	**1978**					

TABLE 4.3 — SMOKING HABITS — DESCRIPTIVE STATISTICS

Regarding the smoking habits of the patients, according to TABLE 4.3, 87,7% of the patients are or have been smokers in their life and only 11,3% who are non-smokers.



smoker					
At risk	585	112	15	1	0
Censored	0	153	215	224	225
Events	0	320	355	360	360
former					
At risk	1151	186	43	14	5
Censored	0	194	273	287	292
Events	0	771	835	850	854
no					
At risk	224	60	14	5	1
Censored	0	47	75	81	83
Events	0	117	135	138	140

Logrank test:
p_value <0.005

FIGURE 4.4 — KAPLAN-MEIER — SMOKING HABITS

In this table we can observe that the non-smoker patients live longer. Through Kaplan-Meier estimator, in FIGURE 4.4, we can confirm that non-smokers live longer and current smokers live more than former smokers, this might be possible because 40% of the former smokers are stage IV comparing to only 30% of the current smokers, this means that we have former smokers in a more advanced stage of cancer. This observation lead us also to the need of evaluating the survival of lung cancer patients in a multivariate approach.

Once again, the logrank test confirms that there are statistical differences between the survival of patients with difference smoking habits.

4.1.4. FAMILY ANTECEDENTS OF CANCER

Family Antecedents Of Cancer	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Yes	923	46.66%	65	64.4	18.1	28
No	691	34.93%	65	64.5	14	26.6
Unknown	364	18.40%	67	66.2	18.6	30.6
TOTAL	**1978**					

TABLE 4.4 — FAMILY ANTECEDENTS OF CANCER — DESCRIPTIVE STATISTICS

About the family history antecedents, we have almost 20% of the patients who do not know if any of their relatives have had cancer before.

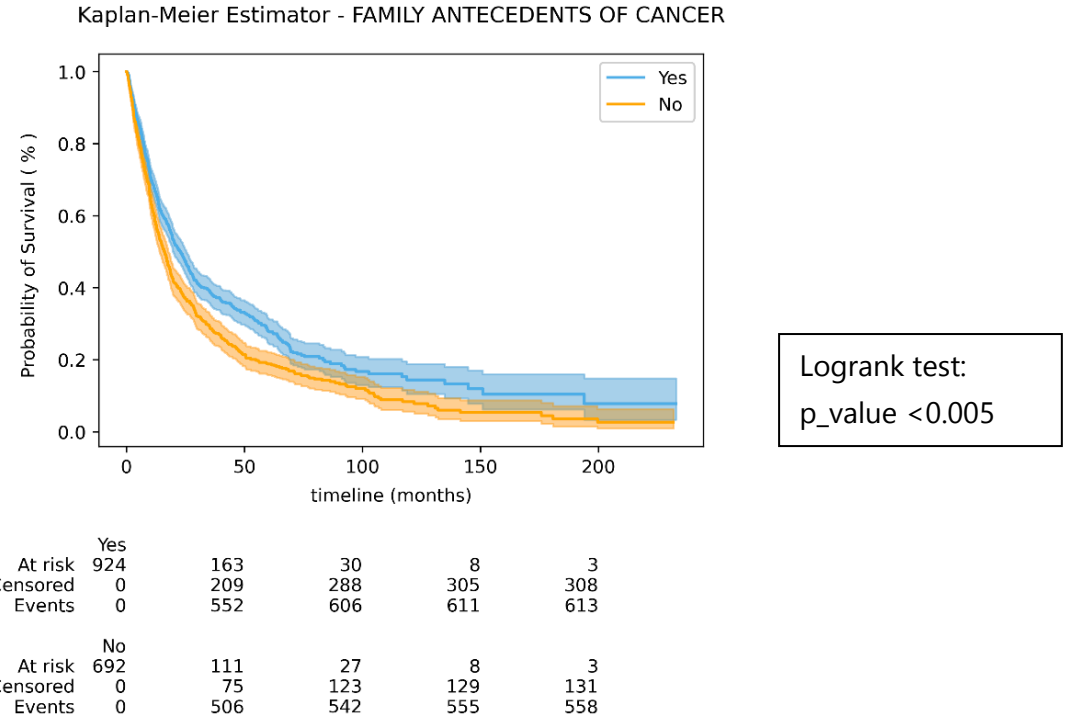


FIGURE 4.5 — KAPLAN-MEIER — FAMILY HISTORY

In the Kaplan-Meier estimator we can see a little difference between them. Logrank test illustrates that there are significant differences in the survival time of patients with or without antecedents of cancer in the family.

4.1.5. MOLECULAR BIOMARKERS

Molecular Biomarkers	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Positive	185	9.35%	64	63.3	24.9	31.9
Not tested	1244	62.89%	66	65.4	15.9	28.5
Negative	549	27.76%	65	63.8	15.3	25.5
TOTAL	**1978**					

TABLE 4.5 — MOLECULAR BIOMARKERS — DESCRIPTIVE STATISTICS

Since the tests for detection of mutations are expensive, not all the patients perform these tests, so almost 63% of the patients are not tested for these mutations, and only 9,4% of the tests are positive.

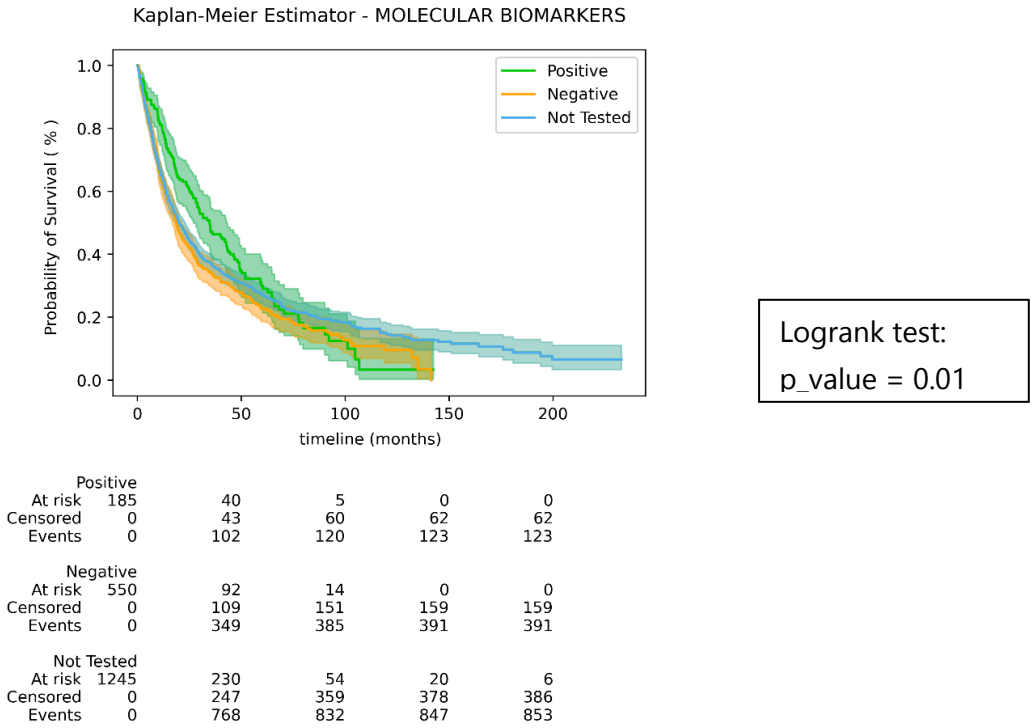


FIGURE 4.6 — KAPLAN-MEIER — MOLECULAR BIOMARKERS

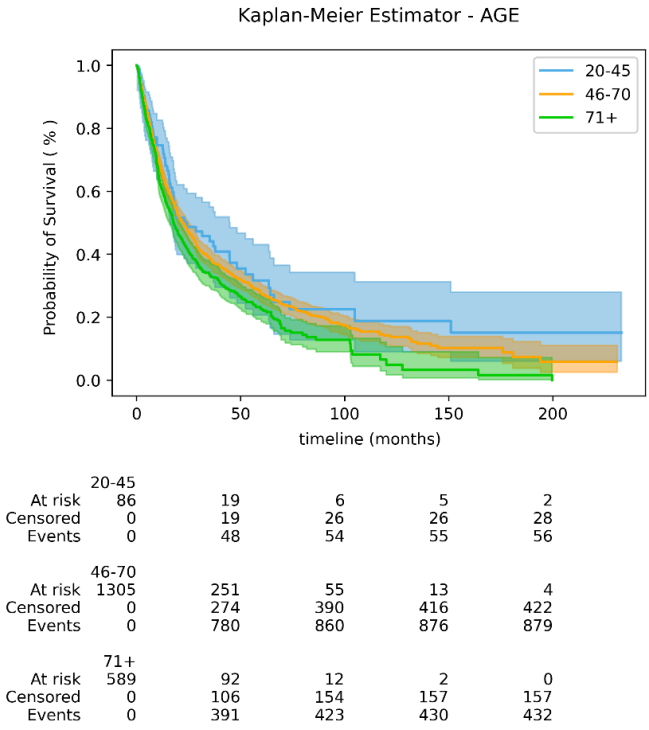
We can observe in FIGURE 4.6 a significant difference in survival between mutated and non-mutated patients, which is confirmed by the logrank test.

4.1.6. AGE

Age	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
[20;45]	86	4.35%	42	41	17.8	35.2
]45;70]	1303	65.87%	62	61.2	17.2	28.9
]70,92]	589	29.78%	75	76.2	14.9	24.8
Total	**1978**					

TABLE 4.6 — GROUP OF AGES — DESCRIPTIVE STATISTICS

We tried to split the patients by group, and we can see that younger patients survive more than older patients. Approximately 66% of the patients have between 46 and 70 years old, and almost 30% of the patients have 71 or more.



Logrank test:
p_value <0.005

FIGURE 4.7 — KAPLAN-MEIER — AGE

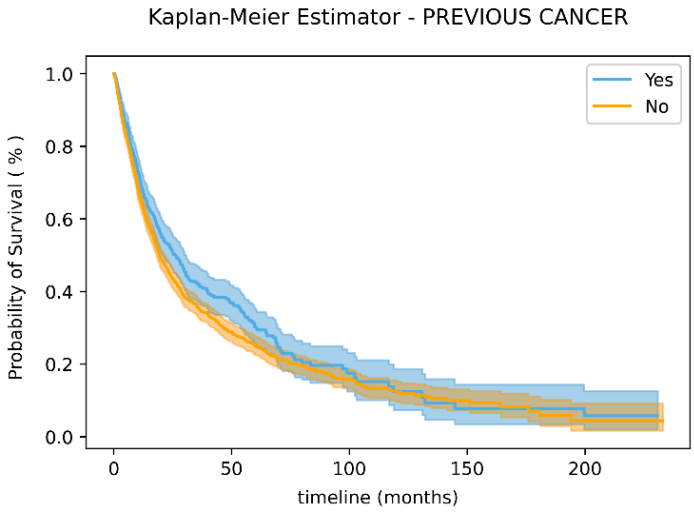
Logrank test and Kaplan-Meier estimator illustrate that there are significant differences in survival time for the three age groups.

4.1.7. PREVIOUS CANCER

Previous Cancer	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Yes	439	22.19%	68	67.5	19.5	31.1
No	1516	76.64%	64	64	15.9	27.2
unknown	23	1.16%	66	66.2	12.7	19.2
TOTAL	**1978**					

TABLE 4.7 — PREVIOUS CANCER — DESCRIPTIVE STATISTICS

In TABLE 4.7, we can observe that 22,2% of the patients had cancer before, and that for 76,6% this is their first cancer. This information is unknown for the remaining 1,2%. In FIGURE 4.8 we can see that there are not many differences between the survival time of patients with or without a previous cancer, but patients who have had cancer before are expected to live a bit more.



	Yes				
At risk	440	97	15	4	3
Censored	0	93	138	142	142
Events	0	250	287	294	295
	No				
At risk	1517	264	58	16	3
Censored	0	298	423	448	456
Events	0	955	1036	1053	1058

Logrank test:
p_value = 0.08

FIGURE 4.8 — KAPLAN-MEIER — PREVIOUS CANCER

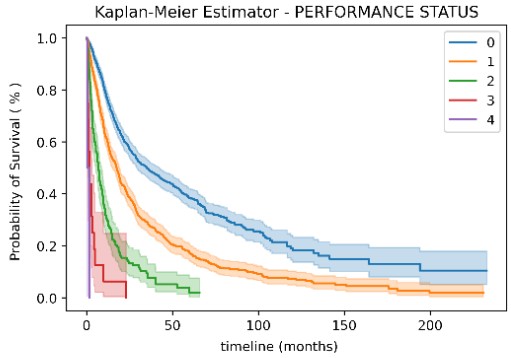
With logrank test we have the expected result: the p-value illustrates that, at a 5% significance level we do not reject that the survival time is identical for patients with and without previous cancer. However, at a 10% significance level, we would reject the equality of survival time. This reflects some instability which may be clarified in a multivariate analysis.

4.1.8. PERFORMANCE STATUS

Previous Cancer	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
0	973	49.19%	65	64.2	23	34.1
1	779	39.38%	66	64.8	14.3	24.6
2	161	8.14%	70	68.2	6.4	10.9
3	16	0.81%	61	64.6	1.9	3.8
4	2	0.10%	66	66	1.1	1.1
unknown	47	2.38%	66	65.2	11.8	25.3
TOTAL	**1978**					

TABLE 4.8 — PERFORMANCE STATUS — DESCRIPTIVE STATISTICS

Most of our patients are classified with 0 and 1 and, as expected, the higher the number, the lower the survival, because these classifications are highly related with the physical and mental status of the patient.



	0				
At risk	974	251	46	10	4
Censored	0	250	387	409	413
Events	0	473	541	555	557
	1				
At risk	780	98	25	9	2
Censored	0	126	156	162	165
Events	0	556	599	609	613
	2				
At risk	161	4	0	0	0
Censored	0	14	16	16	16
Events	0	143	145	145	145
	3				
At risk	16	0	0	0	0
Censored	0	0	0	0	0
Events	0	16	16	16	16
	4				
At risk	2	0	0	0	0
Censored	0	0	0	0	0
Events	0	2	2	2	2

Logrank test:
p value < 0.005

FIGURE 4.9 — KAPLAN-MEIER — PERFORMANCE STATUS

Once more, the Kaplan-Meier estimator confirms the descriptive analysis, we have few patients with the performance status 3 and 4, and they died very shortly after the diagnosis.

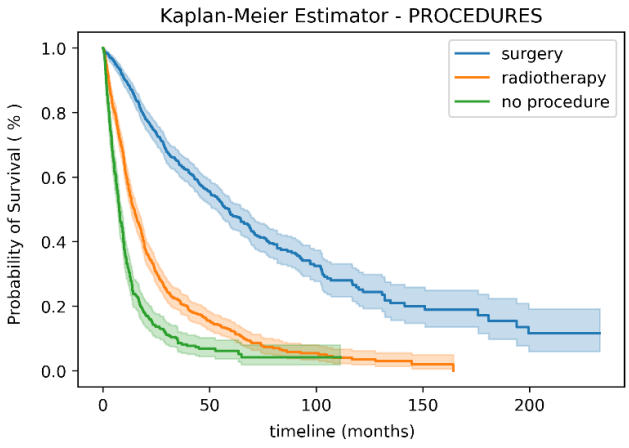
4.2. MEDICAL PROCEDURES

4.2.1. RADIOTHERAPY AND SURGERY

Medical Procedures	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Radiotherapy (RT)	721	36,45	65	64,7	13,9	22,9
Surgery	838	42.37%	65	64.8	33.9	41.7
No Procedures	420	21.23%	66	64.8	7.5	11.8
TOTAL	**1978**					

TABLE 4.9 — PROCEDURES — DESCRIPTIVE STATISTICS

Analyzing this group of patients, we can see that approximately 79% of the patients had been subjects to some procedure, approximately 42% had been subject to surgery, 36% only performed radiotherapy, and 21% did not have any procedure. Obviously, the patients who performed surgery are more likely to survive longer than the others, because the patients who take this procedure are in early stages or stage IIIA.



surgery					
At risk	839	271	59	18	6
Censored	0	261	398	421	428
Events	0	307	382	400	405
radiotherapy					
At risk	721	79	12	2	0
Censored	0	72	100	104	105
Events	0	570	609	615	616
no procedure					
At risk	372	11	2	0	0
Censored	0	31	37	39	39
Events	0	330	333	333	333

Logrank test:
p_value < 0.005

FIGURE 4.10 — KAPLAN-MEIER — PROCEDURES

Regarding the Kaplan-Meier estimator and the logrank, we can confirm the results, and obviously it is very significant.

4.3. TREATMENTS TABLE

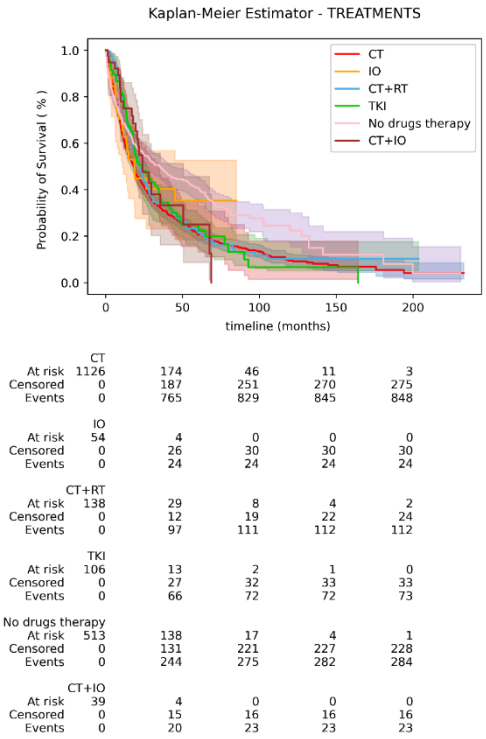
4.3.1. TYPE OF TREATMENTS

Previous Cancer	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Immunotherapy (IO)	54	2.73%	64.5	63.6	10.4	16.7
Chemotherapy (CT)	1126	56.93%	64	63.6	15.4	26.7
TKI	106	5.36%	65	63.9	18.8	26.8
CT+IO	39	1.97%	59	60.6	20.1	22.5
CT+RT	138	6.98%	65	64.5	21.7	34.2
IO+TKI	2	0.10%	59.5	59.5	20.9	20.9
No drug treatment	513	25.94%	69	68.2	17.9	30.9
TOTAL	**1978**					

TABLE 4.10 — TYPE OF TREATMENTS — DESCRIPTIVE STATISTICS

We can see in the TABLE 4.10, that the most common treatment is chemotherapy, 56,9% of the patients have done this therapy, because it is the main treatment to patients with stage IIIB and IV.

In the last years, the hospital started a new drug treatment, immunotherapy. Regarding the survival, patients treated with immunotherapy have the worst results, but that is because this treatment is new, therefore these results will improve given that 60% of the patients are still alive.



Logrank test:
p_value < 0.005

FIGURE 4.11 — KAPLAN-MEIER — TREATMENTS

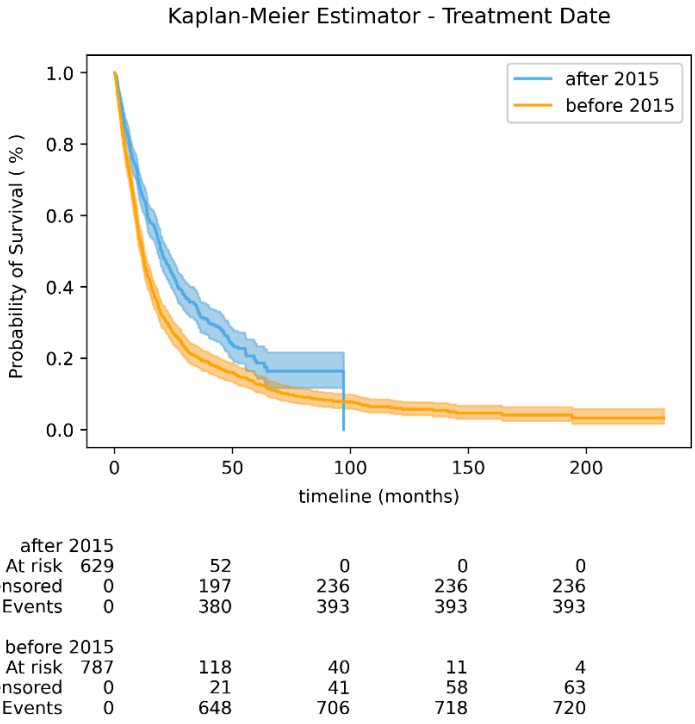
In the Kaplan-Meier estimator, as we can see illustrated in the FIGURE 4.11, we can observe the therapy CT + IO is the one with better results, apart from the "No drug therapy" in which there are mostly patients with stage I and II, whose first procedure is to remove the tumor by surgery.

4.3.2. TREATMENTS DATE

As previously said, here we want to know if the treatments, after 2015, in patients with stage III and IV have improved their survival.

Treatments Date	Count	%	Age		Survival (months)	
			Median	Average	Median	Average
Before 2015	787	55,58	64	63,2	12	27,5
After 2015	629	44,42	66	65,3	15,8	22,1
TOTAL	**1416**					

TABLE 4.11 — TREATMENT'S DATE — DESCRIPTIVE STATISTICS



Logrank test:
p_value < 0.005

FIGURE 4.12 — KAPLAN-MEIER — TREATMENT DATE

Through the descriptive analysis and Kaplan-Meier estimator, we can see that in the last 6 years, the treatments had a significance improvement in the impact of the survival curve of the patients, as we can see it as well in the logrank test.

5.

COX'S PROPORTIONAL-HAZARDS MODEL

As already said, using Cox's method we are able to analyze several variables simultaneously, measuring the impact of each one in a multivariate environment. So, with this model, we can see the impact that each variable, indicating the risk, has on a patient's survival.

We start by choosing which variables we think could impact a patient's survival curve. Using the information obtained in chapter 4 and given that in the individual study the variables chosen are significant, we have good indicators that these variables may be significant in the multivariate model.

We created a model with all stages of diagnosis, but as we know that the stage of diagnosis is the main variable that indicates in which state the cancer is, we will separate the study group in order to understand which variables are significant for different stages of cancer.

When applying Cox regression model with categorical variables, as is the case in this study, it is necessary to define the baseline patient profile, with whom the remaining values for each covariate will be compared. This baseline was defined by analyzing, for each of the covariate, which is the most common value.

In order to be able to compare each model in the same way, we use the same baseline. It is also necessary to consider that for stages I and II, we will not analyze the treatments that patients were submitted to, because if they have had any treatment, that means that, the patient had either a progression or a relapse and therefore, the drug treatments were not the first treatments.

After preparing the dataset, we will apply the model. Next, in the interest of evaluating the model results, we will evaluate which variables are significant. We will optimize the model, and then analyze the hazard risk for each variable.

We will also check if the model complies with the proportional hazard assumption.

The Cox regression model assumes that failure rates are proportional, that is, the risk of failure of the variables is constant over time, for example, the model assumes that the differences of risk between men and women are constant over time. Through the previous chapter, we managed to have some indicators of which variables will not be proportional over time, observing their survival curves.

So, for each model, we define the baseline, create the model and analyze the resulting table to optimize our model. Then, we interpret the results obtained and plot the survival graphs of the significant variables. Finally, we verify if the variables failed the proportionality test.

5.1. DATA PREPARATION

To build the model, all variables must be categorical, that is, each variable was converted to binary, as shown in the figure below.

Variable Name	Values
Smoker	Current
	Former
	No

Variable Name	Values
Smoker_Current	No (0)
	Yes (1)
Smoker_Former	No (0)
	Yes (1)
Smoker_No	No (0)
	Yes (1)

FIGURE 5.1 — DATA PREPARATION (EXAMPLE)

It should be noted that to apply this regression model, none of the variables that we are going to use can have null values. We also have variables where in the dataset the value is "unknown" or "-1". So, there's a need to exclude these patients. For example, we have 26 patients of whom the information whether the patient is a smoker or not is unknown, which means that using the results from these patients would not enable us to apply the regression model.

It is in these covariates that we choose our baseline patient. For example, in model I, as we can see in the section 5.2.1, the dataset has 67 patients who are smokers, 175 former smokers and 39 non-smokers. So, as already explained, in this case we chose former smokers as the characteristic of the baseline patient.

Covariable	Baseline
Gender	Male
Group Age	]45,70]
Smoking Habits	Former
Performance Status	0
Molecular Markers	Negative
Family Antecedents of Cancer	Yes
Progressions or Relapse	Yes
Previous Cancer	No
Comorbidities	Yes
Procedures	No Procedures

TABLE 5.1 — BASELINE PATIENT PROFILE

In a preliminary phase, it was developed a model with all patients and all covariates analyzed in chapter 4. Due to specificities in treatments and medical procedures that are related to the stage at diagnosis, the option for Cox regression model relied on the fitting of three different models: Early Stages (stage I and II), stage III and stage IV.

5.2. MODEL I - EARLY STAGES (STAGE I & II)

In this cohort, there are 281 patients, with 141 observed events and 28 variables. For this case, as mentioned, the types of treatments are not considered, because we are only considering the first treatment, and for these patients, the first treatment was surgery or radiotherapy.

5.2.1. EARLY STAGES - DESCRIPTIVE ANALYSIS

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Stage at diagnosis					
Stage I	158	67	66.6	48.7	51.4
Stage II	123	65	65.3	36.9	48.5
Smoker					
Current	67	65	65.6	45.8	48.5
Former	175	66	65.8	39.4	48.2
Non-smoker	39	69	67.8	63.6	61.9

TABLE 5.2 — MODEL I — STAGE I & II — DESCRIPTIVE STATISTICS (PART I)

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Gender					
Male	204	67	66.7	41.8	48.6
Female	77	64	64.2	48.4	54.2
Group Age					
[20;45]	6	40.5	39.5	65.1	69.6
]45,70]	183	62	61.9	44.4	52.9
]70;92]	92	75	75.9	41.4	43.4
Performance Status					
0	152	66	65.7	50.5	53.3
1	120	66	65.7	33.7	48.1
2	9	76	74.4	14.2	24.9
Family History					
Yes	159	67	66.4	45	50.3
No	122	65.5	65.5	41.5	50
Molecular Markers					
Mutated	15	62	63.7	41.9	52.3
Negative	41	65	64	45.3	51.3
Not tested	225	67	66.5	43.2	49.8
Progressions or Relapse					
Yes	152	66	65.8	35.5	46.6
No	129	67	66.3	53.3	54.3
Previous Cancer					
Yes	81	69	68.3	47.4	54.9
No	200	65	65.1	42.2	48.2
Comorbidities					
Yes	246	67	66.9	44.3	50.1
No	35	60	59.6	41.9	50.8
Surgery					
Yes	251	66	65.3	45.3	52.5
No	30	73	72.4	21.1	30.4
Radiotherapy					
Yes	26	71.5	71.6	23.4	32.5
No	255	66	65.4	45	52

TABLE 5.3 — MODEL I — STAGE I & II — DESCRIPTIVE STATISTICS (PART II)

5.2.2. EARLY STAGES - RESULTS

After choosing the baseline patients, we will apply Cox regression model to the dataset. As mentioned before, we will considerer the covariate significant if the p-value is under 0.10.

	HR (C.I)	p-value	Significance
Smoker			
Former Smoker & Smoker			Value in the baseline
Non-smoker	0.57 (0.32 – 1.01)	0.05	Non-Smoker patients have **43% less risk** of dying comparing with the patients who smoked somewhere in their life.
Group Age			
[20;70]			Value in the baseline
]70;92]	1.44 (0.98-2.13)	0.07	Patients older than 70 years old have **44% more risk** of dying comparing with younger patients.
Performance Status			
0			Value in the baseline
1	1.72 (1.18-2.52)	<0.005	Patients with performanceStatus 1 have **72% more risk** of dying comparing with the patients with performance Status 0.
2	6.63 (2.83-15.56)	<0.005	Patients with performanceStatus 2 have **563% more risk** of dying comparing with the patients with performance Status 0.
Family History			
No			Value in the baseline
Yes	0.56 (0.40-0.79)	0.01	Patients with family history have **44% less risk** of dying than patients with no family history of cancer.
Progressions or Relapse			
No			Value in the baseline
Yes	3.15 (2.04-4.86)	<0.005	Patients with relapses or progressions have **215% more risk** of dying than patients with no progressions.
Previous Cancer			
No			Value in the baseline
Yes	0.65 (0.44-0.97)	0.03	Patients with previous cancer have **35% less risk** of dying than patients with no previous cancer.
Procedures			
None			Value in the baseline
Surgery	0.14 (0.05-0.43)	<0.005	Patients with surgery have **86% less risk** of dying than patients without any of these procedures.
Radiotherapy	0.31 (0.10-0.94)	0.04	Patients with radiotherapy have **69% less risk** of dying than patients who were not treated.

TABLE 5.4 — MODEL I — STAGE I & II — COX REGRESSION MODEL — RESULTS

The performance of the model was evaluated by measuring the c-index of 0.75.

We can also verify the impact that each variable has in the risk by creating a plot shown in FIGURE 5.2. As already mentioned, the variables in the right of the axis have more risk comparing with the baseline, and those at the left have less risk.
The confidence interval, lower and upper bounds, it is possible to visualize in the FIGURE 5.2 as well.

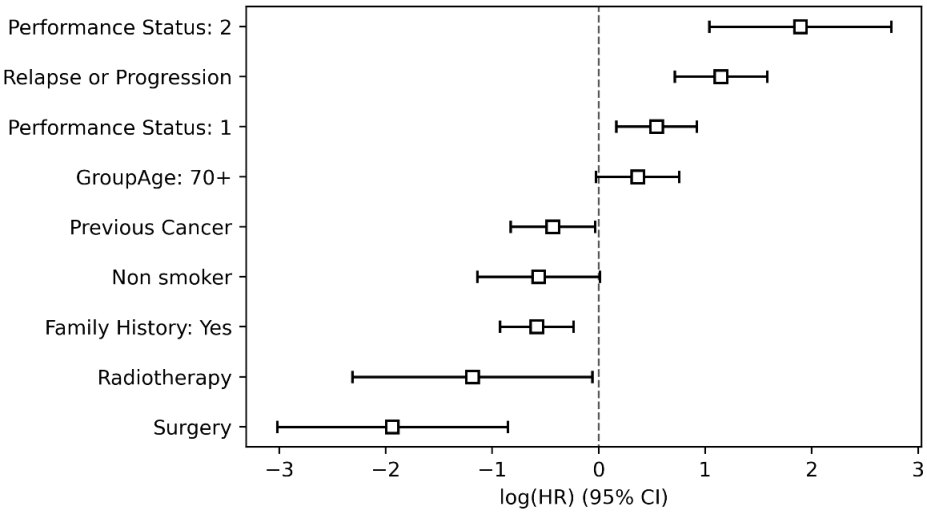


FIGURE 5.2 — MODEL I — STAGE I & II — COX REGRESSION MODEL — RESULTS

Looking at the results in the table and with the plots of the survival curve, we can define a patient profile with:

- Lower risk profile -> Patient younger than 70 years old, non-smoker, performance Status 0, with family history of cancer, with previous cancer, with no relapse and with surgery.

- Higher risk profile -> Patient older than 71 years old, smoker, performance Status 2, with no family history of cancer, without previous cancer, with relapse and no procedures.

5.2.3. EARLY STAGES - TESTING THE PROPORTIONAL HAZARD ASSUMPTIONS

Only the **Non-Smoker** variable fails the proportional hazard assumption. FIGURE 5.3 illustrates the Schoenfeld residuals.

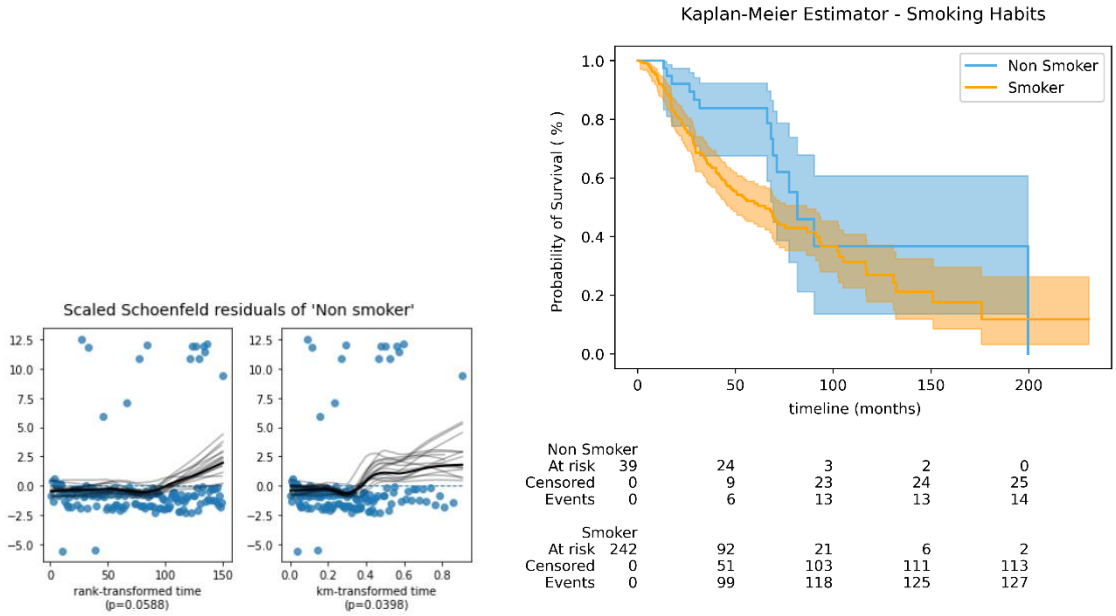


FIGURE 5.3 — MODEL I — STAGE I & II — SCALED SCHOENFELD RESIDUALS & KAPLAN-MEIER — NON-SMOKER

We can see this model only fails in the variable non-smoker, and only after 80 months. As we can see in FIGURE 5.3, the variable fails because we have very few non-smoking patients after 80 months and because the two survival curves intersect.

5.2.4. EARLY SAGES - SURVIVAL CURVES

After fitting and optimizing the model we can estimate the survival curves for each significant variable, as we change a single feature, while all the other covariates are equal. With these plots we can understand the impact of each covariate, given by the model.

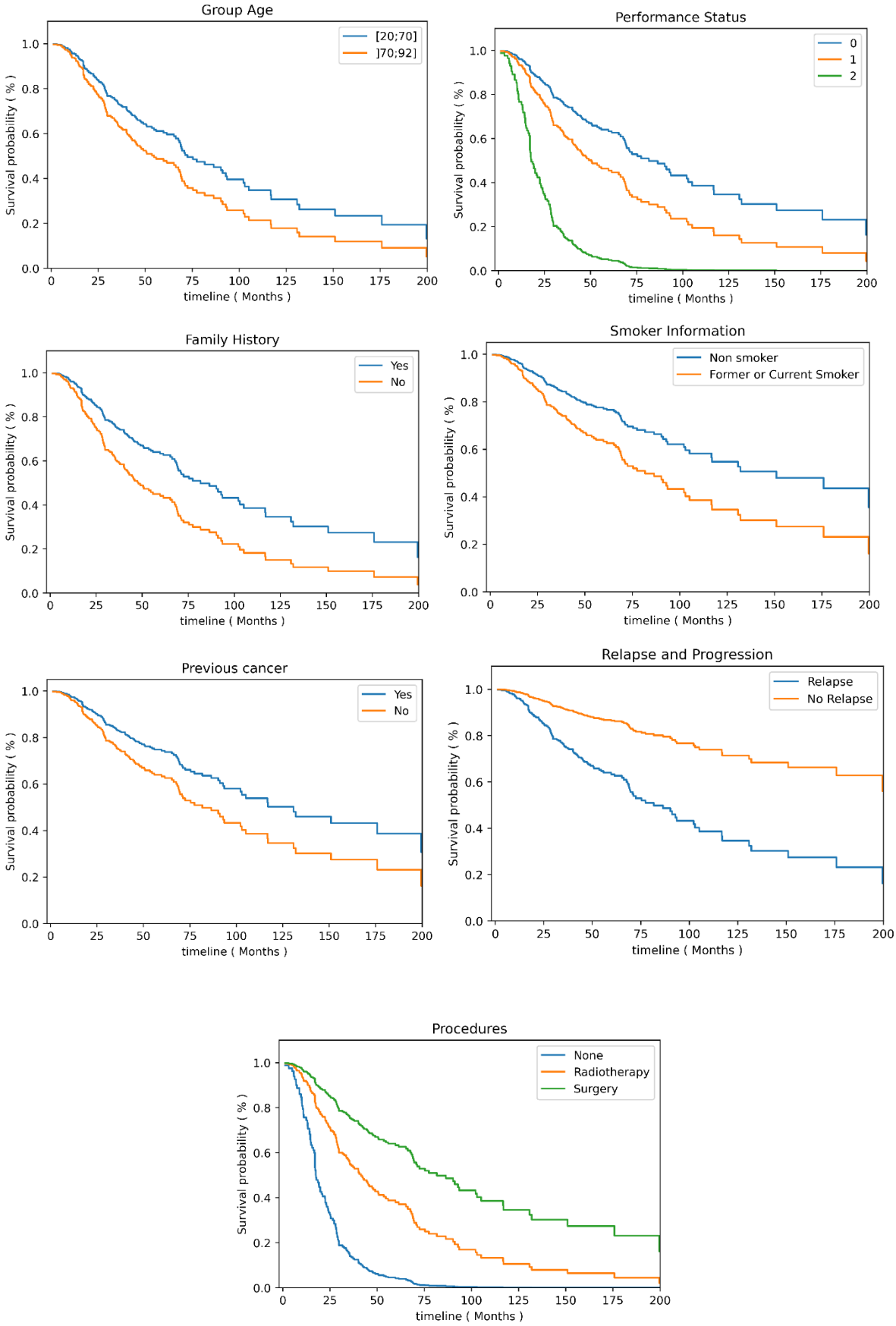


FIGURE 5.4 — MODEL I — STAGE I & II — COX REGRESSION MODEL — SURVIVAL CURVES

In all the plots of FIGURE 5.4, the lines represent the survival for the baseline patient with the impact of the change of the value of one covariate at a time.

This survival curves are a multivariate analysis of survival. For a given set of profile characteristics, Cox regression model allows for the estimation of the survival times and correspondent probabilities.

5.3. MODEL II - STAGE III

In this cohort, there are 470 patients, with 354 observed events and 32 variables. Of note, we have 3 patients treated with immunotherapy, since all the patients are alive, we had to exclude them in this model.

5.3.1. STAGE III - DESCRIPTIVE ANALYSIS

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Smoker					
Current	123	63	63.1	22.8	31.7
Former	306	67	65.9	18.6	31.5
Non-Smoker	41	65	64.4	41.5	45.4
Gender					
Male	353	68	66.3	19	30.9
Female	117	61	61.3	26.3	38.5
Group Age					
[20;45]	16	43	41.8	44.7	54.1
]45,70]	311	62	61.2	22	35.3
]70;92]	143	75	76.1	18.2	25
Performance Status					
0	239	64	63.8	23.6	37.6
1	198	68	65.6	20.1	30.1
2	31	72	70.8	11.9	14.5
3	2	72.5	72.5	13.8	13.8
Family History					
Yes	269	65	64.3	24.6	34.9
No	201	67	66	18.3	29.9

TABLE 5.5 — MODEL II — STAGE III — DESCRIPTIVE STATISTICS (PART I)

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Molecular Markers					
Mutated	31	65	62.3	49.3	50
Negative	123	65	64.7	23.3	32.5
Not tested	316	66	65.5	18.8	31.2
Progressions or Relapse					
Yes	340	65	64.7	19.6	31
No	108	67	65.6	26.5	39.3
Previous Cancer					
Yes	93	67	66.6	22	35
No	377	65	64.7	20.1	32.2
Yes	93	67	66.6	22	35
Yes	389	67	66.2	21.7	32.5
No	81	59	59.8	18.5	34.1
Surgery					
Yes	201	65	64	33.4	44.2
No	269	68	65.9	16.1	24.2
Radiotherapy					
Yes	269	65	64.3	24.6	34.9
No	201	67	66	18.3	29.9
Type of Treatments	143	75	76.1	18.2	25
IO	3	65	58.7	12.2	12.6
CT	318	65	64.1	23.4	35.7
TKI	5	69	66.4	17.9	20.1
CT + IO	4	63	63.5	32	31.8
CT + RT	98	65.5	64.8	19.1	28.9

TABLE 5.6 — MODEL II — STAGE III — DESCRIPTIVE STATISTICS (PART II)

In order to be able to compare all models in the end, we will keep the baseline in most variables, except obviously in the variable stage.

5.3.2. STAGE III - RESULTS

Doing the same process with all non-significant covariates, painted in grey, we are left with the following model presented in the TABLE 5.7 and TABLE 5.8.

	HR (C.I)	p-value	Significance
Smoker			
Former Smoker & Smoker			Value in the baseline
Non-smoker	0.56 (0.37 – 0.85)	<0.005	Non-Smoker patients have **44% less risk** of dying comparing with the patients who smoked sometime in their life
Gender			
Male			Value in the baseline
Female	0.77 (0.58-1.01)	0.06	Women have **23% less risk** of dying comparing with men
Group Age			
[20;70]			Value in the baseline
]70;92]	1.27 (1.00 - 1.62)	0.05	Patients over 71 years old have **27% more risk** of dying than younger patients
Performance Status			
0			Value in the baseline
1	1.27 (1.01-1.58)	0.04	Patients with performanceStatus 1 have **27% more risk** of dying than patients with performanceStatus 0.
2	2.03 (1.32-3.12)	<0.005	Patients with performanceStatus 2 have **103% more risk** of dying than patients with performanceStatus 0.
3	2.83 (0.68-11.76)	0.15	Patients with performanceStatus 3 have **183% more risk** of dying than patients with performanceStatus 0.
Progressions or Relapse			
No			Value in the baseline
Yes	1.70 (1.35-2.15)	<0.005	Patients with relapses or progressions have **70% more risk** of dying than patients with no progressions.
Procedures			
None			Value in the baseline
Surgery	0.11 (0.08-0.16)	<0.005	Patients with surgery have **89% less risk** of dying than patients with no procedures.
Radiotherapy	0.22 (0.15-0.31)	<0.005	Patients with radiotherapy have **78% less risk** of dying than not treated patients .

TABLE 5.7 — MODEL II — STAGE III — COX REGRESSION MODEL — RESULTS (PART I)

	HR (C.I)	p-value	Significance
Treatments			
CT			Value in the baseline
CT + IO	0.36 (0.05-2.61)	0.31	Patients with CT + IO have **64% less risk** of dying than patients treated with CT.
CT + RT	1.18 (0.91-1.55)	0.21	Patients with CT + RT have **18% more risk** of dying than patients treated with CT.
IO	0.00	0.00	All these patients are alive
No Drug Therapy	2.12 (1.46-3.08)	<0.005	Patients with no therapy have **112% more risk** of dying than patients treated with CT.
TKI	1.50 (0.51-4.45)	0.46	Patients with TKI have **50% more risk** of dying than patients treated with CT.

TABLE 5.8 — MODEL II — STAGE III — COX REGRESSION MODEL — RESULTS (PART II)

The performance of the model was evaluated by measuring the c-index of 0.71.

Regarding the treatments, we can have an interpretation of the results, but as we can see in the p_value and in the HR boundaries, the standard error is very high.

We can also verify the impact that each variable has on survival risk by creating a plot, shown in FIGURE 5.5. If one of the bounds is at the left and the other at the right it means that the variable is not significant.

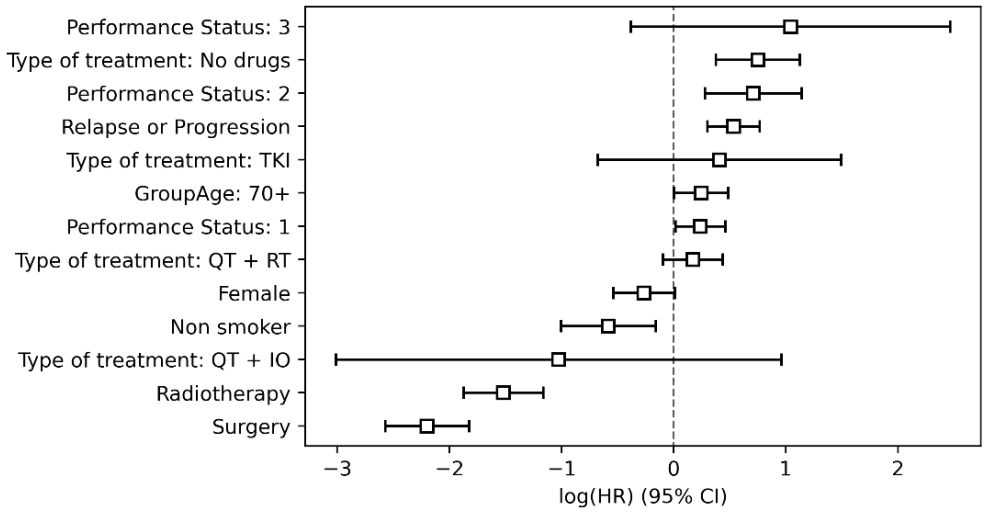


FIGURE 5.5 — MODEL II — STAGE III — COX REGRESSION MODEL — RESULTS

So, looking at the results in TABLE 5.7 and TABLE 5.8., we can define a patient profile with:

- Lower risk profile -> female, non-smoker, under 70 years of age, performance Status 0, no relapse, with surgery and with some chemical therapy, mainly CT + IO.
- Higher risk profile -> male, smoker, performance Status greater than 0, with relapse, no surgery, with no procedures and without therapy.

5.3.3. STAGE III - TESTING THE PROPORTIONAL HAZARD ASSUMPTIONS

Four features failed the proportional hazard assumption:

1. Non-smoker
2. Radiotherapy
3. Surgery
4. Relapse or progression.

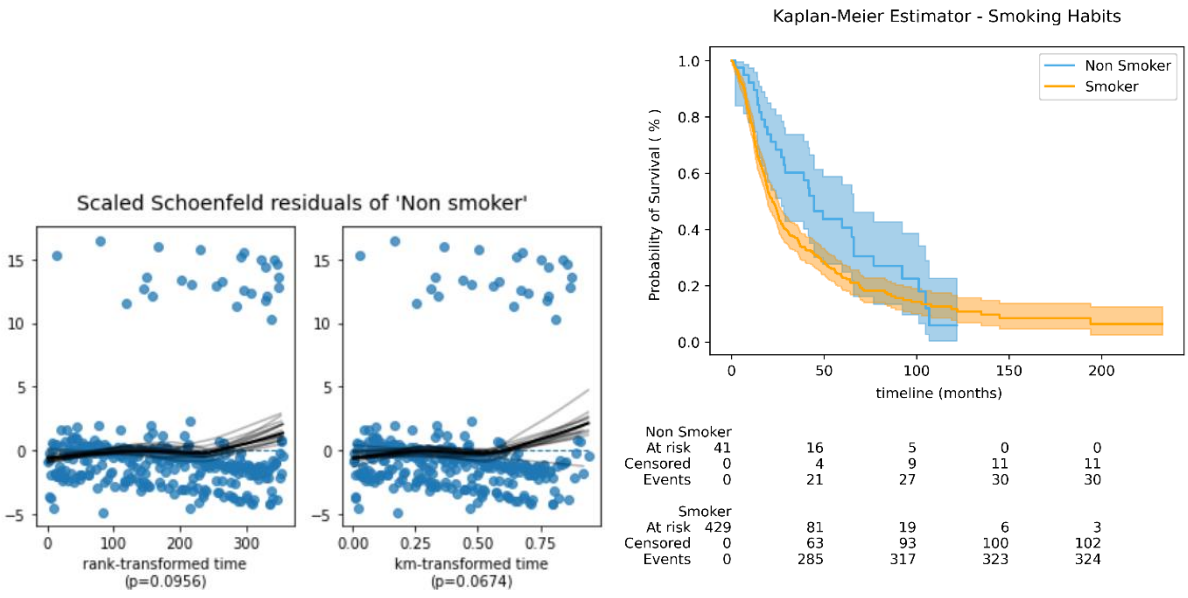


FIGURE 5.6 — MODEL II — STAGE III — SCHOENFELD & KAPLAN-MEIER — NON-SMOKER

Just like in the previous model, the variable non-smoker only fails after 100 months. As we can see in FIGURE 5.6, the variable fails because we have very few non-smoking patients after 80 months and because the two survival curves intersect in that moment.

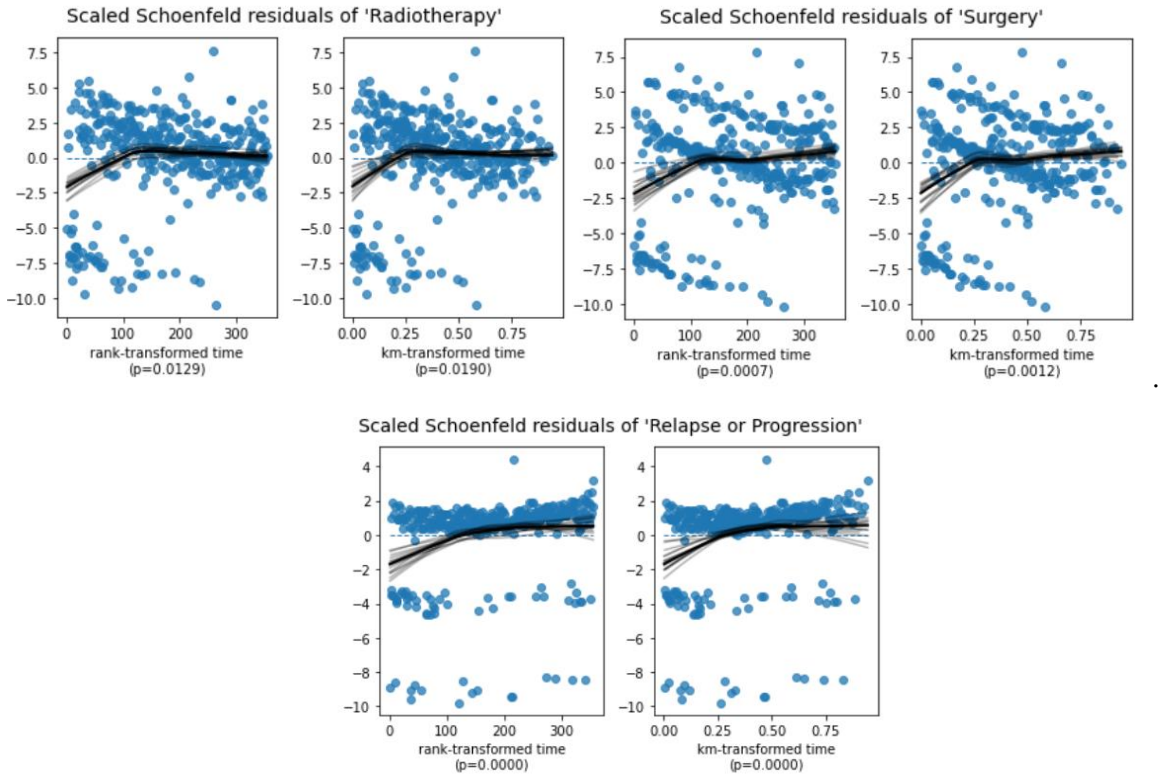
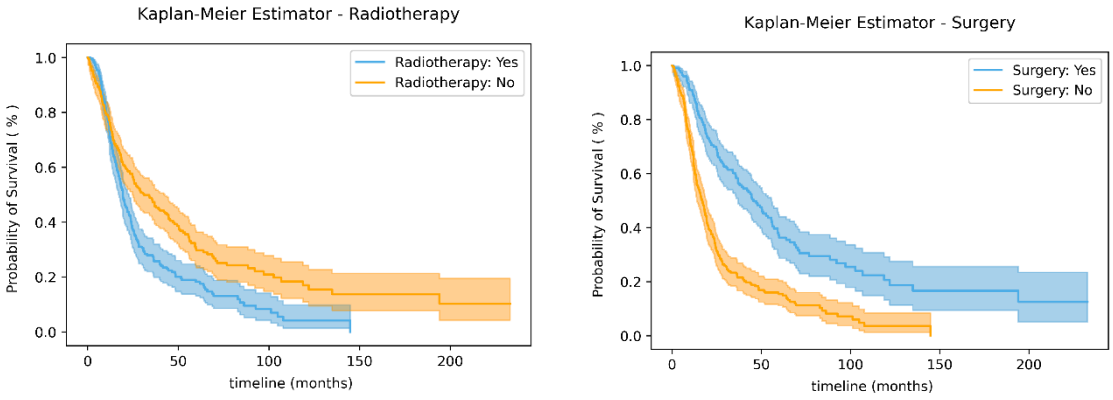


FIGURE 5.7 — MODEL II — STAGE III — SCHOENFELD — RT, SURGERY AND RELAPSE



Radiotherapy: Yes					
At risk	217	34	6	0	0
Censored	0	19	33	35	35
Events	0	164	178	182	182
Radiotherapy: No					
At risk	253	63	18	6	3
Censored	0	48	69	76	78
Events	0	142	166	171	172

Surgery: Yes					
At risk	201	62	18	6	3
Censored	0	45	65	72	74
Events	0	94	118	123	124
Surgery: No					
At risk	269	35	6	0	0
Censored	0	22	37	39	39
Events	0	212	226	230	230

FIGURE 5.8 — MODEL II — STAGE III — KAPLAN-MEIER — RT, SURGERY AND RELAPSE (PART I)

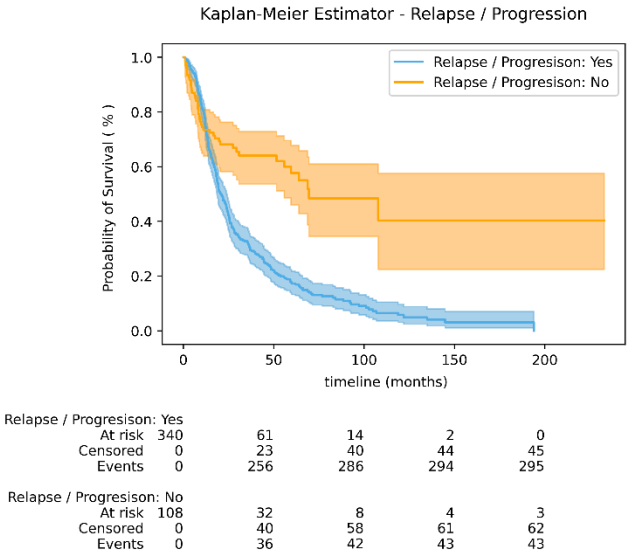


FIGURE 5.9 — MODEL II — STAGE III — KAPLAN-MEIER — RT, SURGERY AND RELAPSE (PART II)

These variables fail because as we can see in the Kaplan-Meier, in the FIGURE 5.8 and FIGURE 5.9, in the beginning the two survival curves have intersected.

5.3.4. STAGE III - SURVIVAL CURVES

It is time to assess the survival curves for each significant variable.

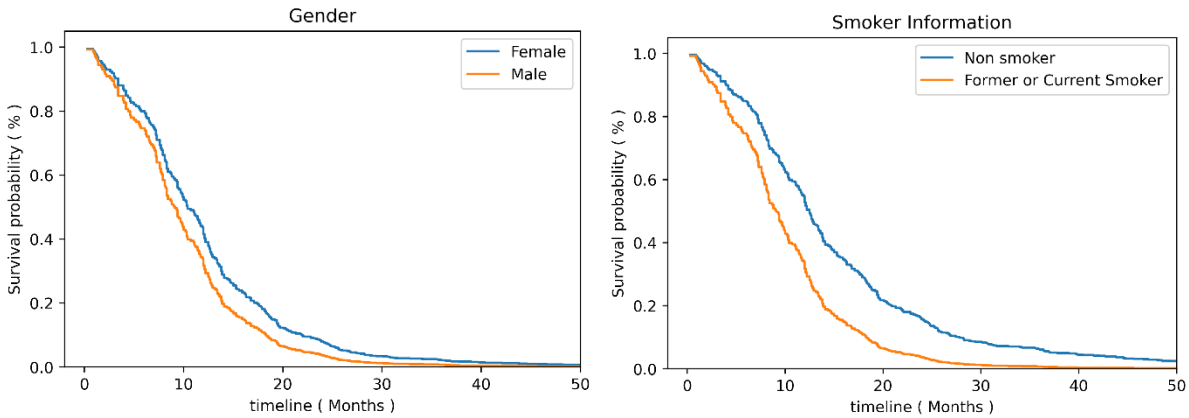


FIGURE 5.10 — MODEL II — STAGE III — COX REGRESSION MODEL — SURVIVAL CURVES (PART I)

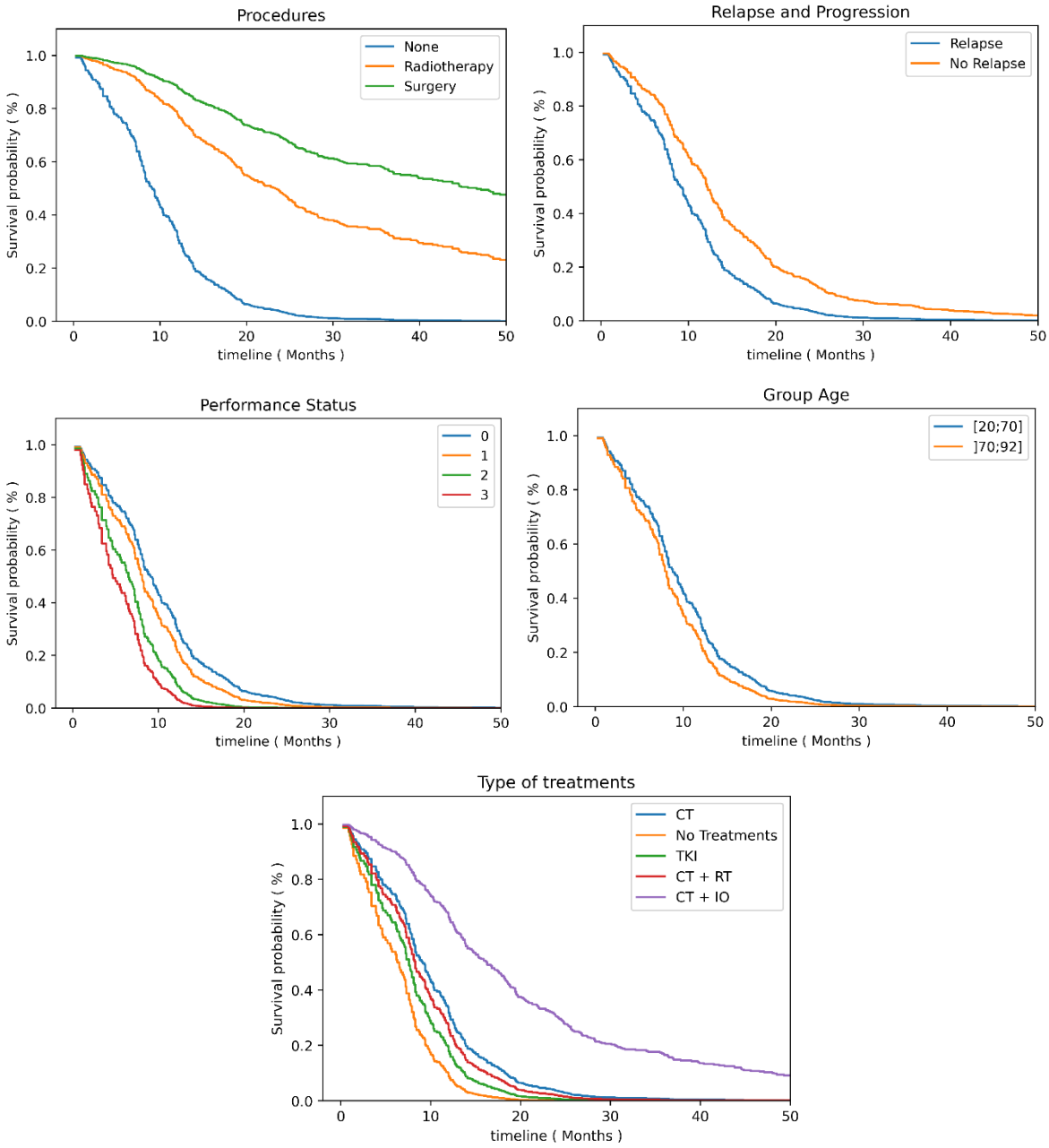


FIGURE 5.11 — MODEL II — STAGE III — COX REGRESSION MODEL — SURVIVAL CURVES (PART II)

Once again, the illustrated survival curves are representing the survival of the baseline patients, changing the value of each covariate one at a time.

5.4. MODEL III - STAGE IV

In this model we only have patients diagnosed with stage IV, it is also where we have almost 50% of the patients. It is expected to have low survival curves here, since the cancer is already in an advance stage and as already presented in the Kaplan-Meier curves in chapter 4.

In this cohort, there are 694 patients, with 611 observed events and 32 variables.

5.4.1. STAGE IV - DESCRIPTIVE ANALYSIS

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Smoker					
Current	194	60	59,8	9	15,8
Former	403	65	64,6	8,7	14,5
No	97	66	65,4	15,1	24,1
Gender					
Male	495	65	63,9	8,2	13,5
Female	199	62	61,9	13	22,7
Group Age					
[20;45]	41	42	42,4	14	23,8
]45,70]	475	61	60,3	9,6	16,4
]70;92]	178	75	76,3	9,5	13,9
Performance Status					
0	218	62	61,8	13,4	23,2
1	356	64	63,4	9,5	14,5
2	108	67	66,5	4,8	9,1
3	12	60	62	1,5	2,4
Family History					
Yes	385	64	63,7	10,2	16,8
No	309	63	63	8,7	15,5
Molecular Biomarkers					
Positive	107	64	63,6	18,9	25,5
Negative	243	63	63,2	10,2	17,7
Not tested	344	64	63,4	7,2	12,2
Progressions or Relapse					
Yes	526	64	63,1	11,1	17,9
No	87	65	63,4	5,5	15,1

TABLE 5.9 — MODEL III — STAGE IV — DESCRIPTIVE ANALYSIS (PART I)

	Count	Age		Survival (months)	
		Median	Average	Median	Average
Previous Cancer					
Yes	130	69,5	67,8	10,1	16,7
No	564	63	62,3	9,6	16,1
Comorbidities					
Yes	534	66	65,1	9,7	16,7
No	160	58	57,6	9	14,5
Surgery					
Yes	49	60	60,9	27	42,8
No	645	64	63,6	9,2	14,2
Radiotherapy					
Yes	354	62	62,7	10,2	15,5
No	340	65	64,1	9	16,9
Treatments					
CT	494	63	62.8	9.3	15.8
CT + IO	25	59	59.9	19.6	21.1
CT + RT	5	62	62	19.4	23.6
IO	31	65	64.5	11	17.5
IO+TKI	1	78	78	12.6	12.6
TKI	75	65	64.1	18.9	25
No Drug Therapy	62	70	67.9	2	5.7

TABLE 5.10 — MODEL III — STAGE IV — DESCRIPTIVE ANALYSIS (PART II)

5.4.2. STAGE IV - RESULTS

Doing the same process with all non-significant covariates, painted in grey, we are left with the following model presented in the TABLE 5.11 and TABLE 5.12.

	HR (C.I)	p-value	Significance
Gender			
Male			Value in the baseline
Female	0.74 (0.60-0.90)	<0.005	Women have **26% less risk** of dying comparing with men
Group Age			
[45,70]			Value in the baseline
[20;45]	0.70 (0.48-1.01)	0.06	Patients younger than 45 years old have **30% less risk** of dying comparing with older patients between 45 and 70.
]70;92]	0.83 (0.68 - 1.01)	0.06	Patients over 71 years old have **17% less risk** of dying comparing with younger patients
Performance Status			
0			Value in the baseline
1	1.42 (1.17-1.71)	<0.005	Patients with performanceStatus 1 have **42% more risk** of dying comparing with the patients with performanceStatus 0.
2	1.95 (1.50-2.54)	<0.005	Patients with performanceStatus 2 have **95% more risk** of dying comparing with the patients with performanceStatus 0.
3	8.22 (4.34-15.60)	<0.005	Patients with performanceStatus 3 have **722% more risk** of dying comparing with the patients with performanceStatus 0.
Molecular Biomarkers			
Negative			Value in the baseline
Positive	0.63 (0.44-0.92)	0.002	Patients who tested positive in these tests have **37% less risk** of dying comparing with the patients that have the test negative.
Not tested	1.21 (1.00 – 1.47)	0.04	Patients who have not been tested have **21 % more risk** of dying comparing with the patients that have tested negative.
Progressions or Relapse			
No			Value in the baseline
Yes	0.77 (0.67-0.88)	<0.005	Patients with relapses or progressions have **23% less risk** of dying than patients with no progressions.
Comorbidities			
No			Value in the baseline
Yes	1.25 (1.03-1.53)	0.03	Patients with comorbidities have **25% more risk** of dying than patients with no comorbidities.

TABLE 5.11 — MODEL IV — STAGE IV — COX REGRESSION MODEL — RESULTS (PART I)

	HR (C.I)	p-value	Significance
Procedures			
None			Value in the baseline
Surgery	0.24 (0.16-0.36)	<0.005	Patients with surgery have **76% less risk** of dying than not treated patients.
Radiotherapy	0.69 (0.58-0.83)	<0.005	Patients with radiotherapy have **31% less risk** of dying than not treated patients.
Treatments			
CT			Value in the baseline
CT + IO	0.62 (0.39-0.99)	0.04	Patients with CT + IO have **38% less risk** of dying than patients treated with CT.
CT + RT	0.95 (0.35-2.57)	0.92	Patients with CT + RT have **5% less risk** of dying than patients treated with CT.
IO	0.61 (0.39-0.95)	0.03	Patients with IO have **39% less risk** of dying than patients treated with CT.
No Drug Therapy	1.92 (1.38-2.68)	<0.005	Patients with no therapy have **92% more risk** of dying than patients treated with CT.
TKI	0.73 (0.49-1.10)	0.14	Patients with TKI have **27% less risk** of dying than patients treated with CT.

TABLE 5.12 — MODEL IV — STAGE IV — COX REGRESSION MODEL — RESULTS (PART II)

The performance of the model was evaluated by measuring the concordance index of 0.72.

Once again, regarding the treatments, we can have an interpretation of the results, but as we can see in the p-value and in the HR boundaries, the standard error is very high in some of them.

We can also verify the impact that each variable has on survival risk by creating a plot, shown in FIGURE 5.12. As mentioned before, if one of the bounds is at the left and the other at the right it means that the variable is not significant.

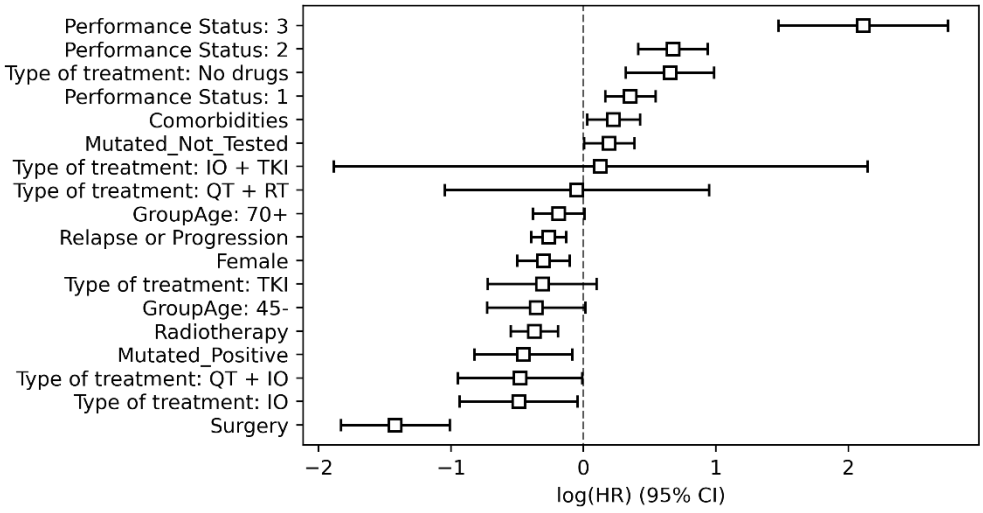


FIGURE 5.12 — MODEL III — STAGE IV — COX REGRESSION MODEL — RESULTS

Looking at the results in the TABLE 5.11, TABLE 5.12 in the FIGURE 5.12, we can define a patient profile with:

- Lower risk profile -> female, under 45 years of age, non-smoker, performance Status 0, with surgery, no comorbidities with molecular biomarker positive and with some chemical therapy, mainly IO or CT+IO.

- Higher risk profile -> male, smoker, performance Status 1 or above, no procedures, with comorbidities and molecular marker negative or not performed and without therapy.

5.4.3. STAGE IV - TESTING THE PROPORTIONAL HAZARD ASSUMPTIONS

Four features failed the proportional hazard assumption:

1. Radiotherapy
2. Surgery
3. Relapse or progression
4. Type of treatment: No drugs therapy

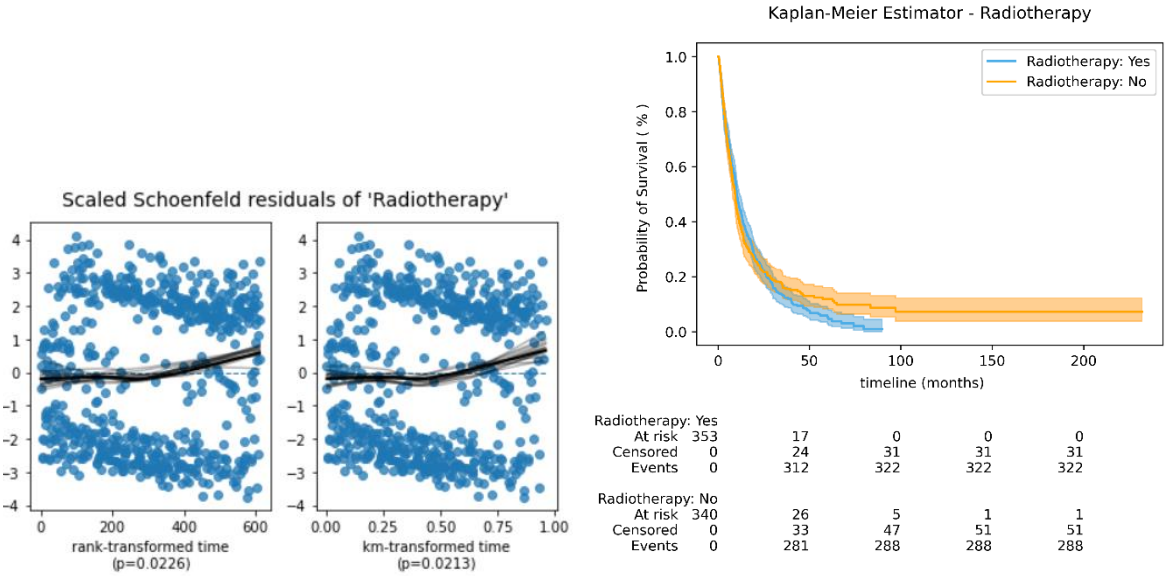


FIGURE 5.13 — MODEL III — STAGE IV — RESIDUALS & KAPLAN-MEIER — RADIOTHERAPY

Regarding Radiotherapy, we can see this variable fails because in the first months there are no differences and both curves intersect.

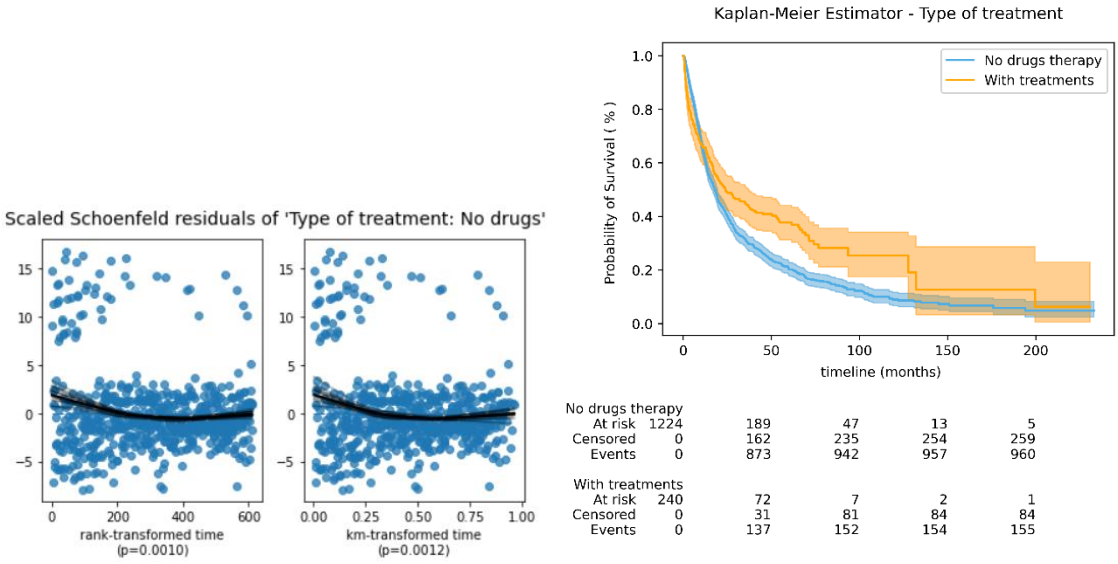


FIGURE 5.14 — MODEL III — STAGE IV — RESIDUALS & KAPLAN-MEIER — TYPE OF TREATMENT: NO DRUGS

Regarding the type of treatment variable: no drugs therapy, it fails because in the first months there are no differences and both curves intersect.

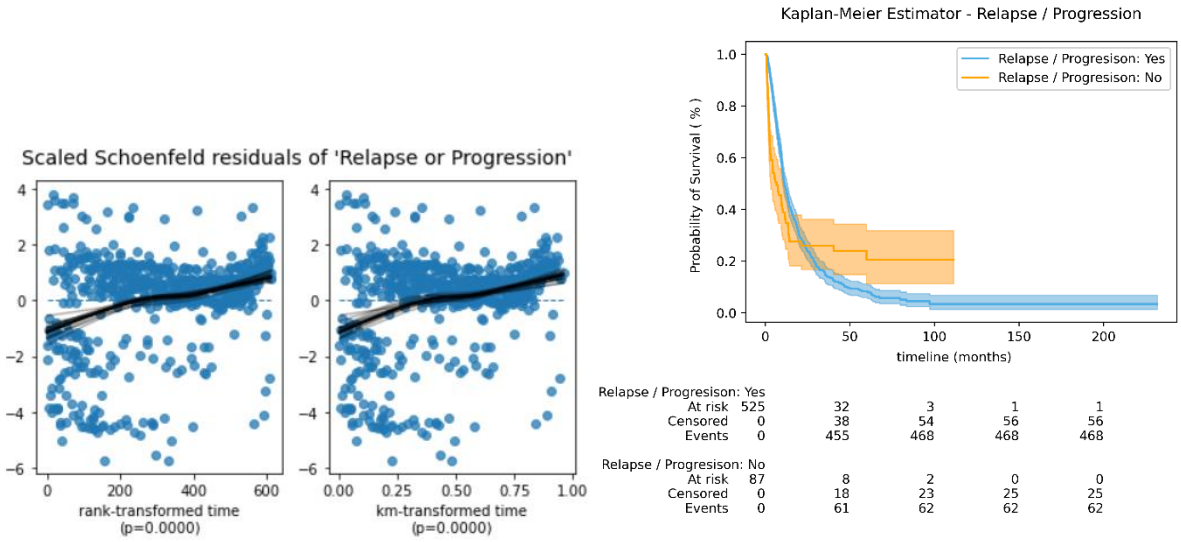


FIGURE 5.15 — MODEL III — STAGE IV — RESIDUALS & KAPLAN-MEIER — RELAPSE

Regarding the variable relapse or progression, this variable fails because both curves intersect.

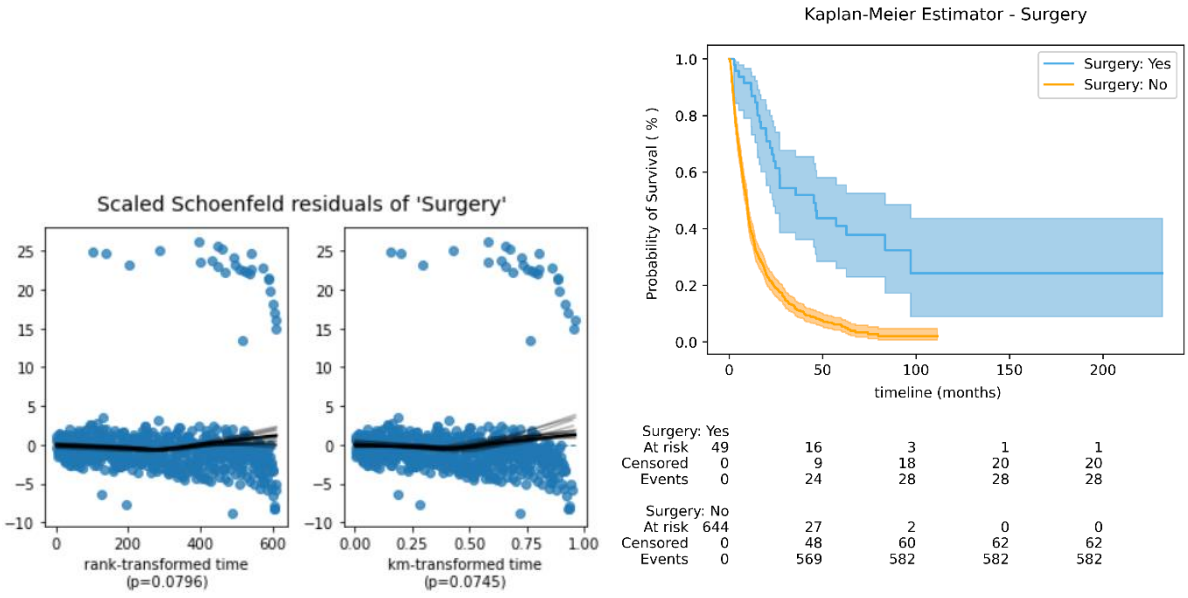


FIGURE 5.16 — MODEL III — STAGE IV — RESIDUALS & KAPLAN-MEIER — SURGERY

The variable surgery fails after the month 500, so we can say that this variable is very significant as we can see in both graphics.

5.4.4. STAGE IV - SURVIVAL CURVES

In the FIGURE 5.17 and FIGURE 5.18 is illustrated the survival curves for each significant covariate.

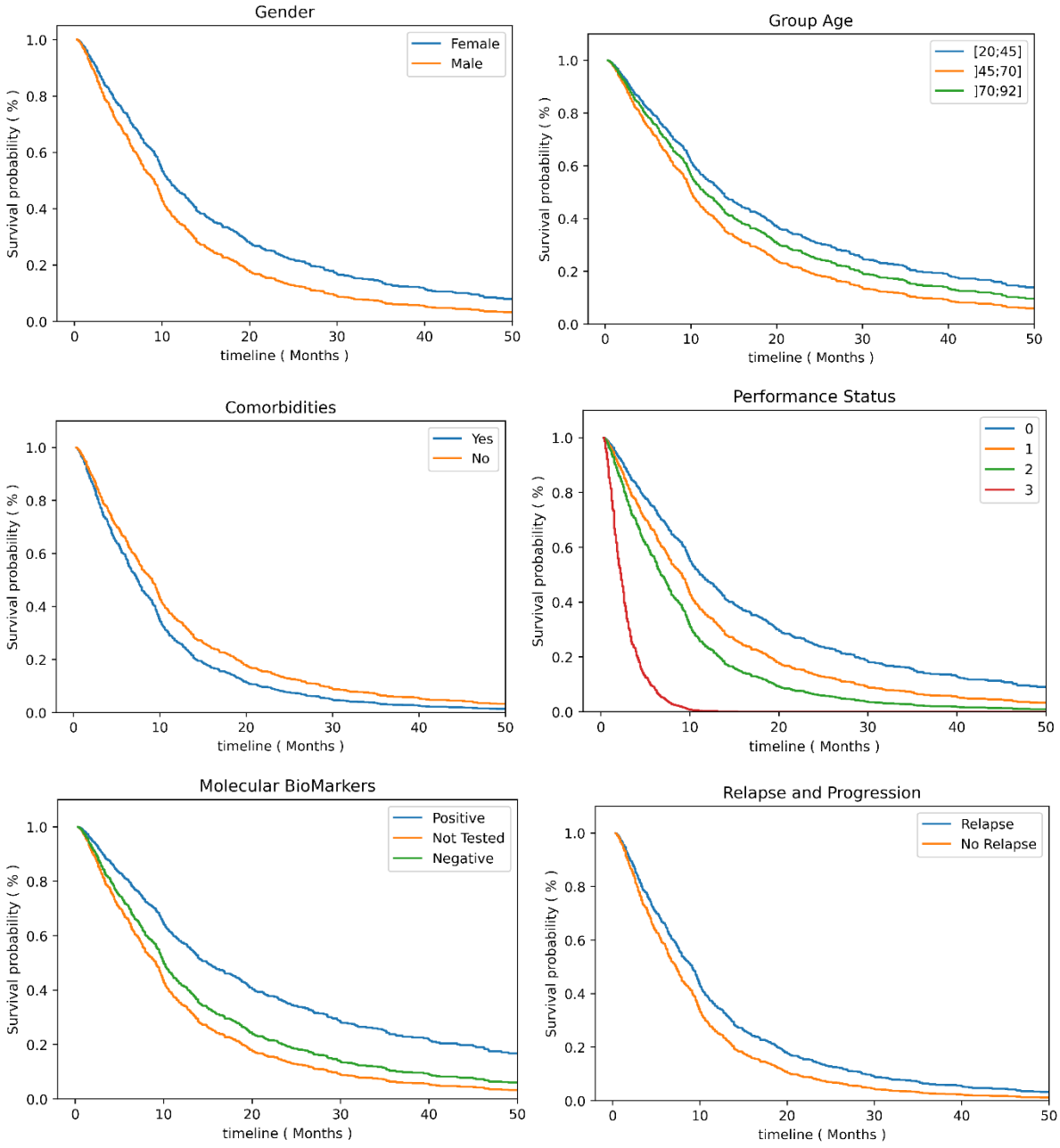


FIGURE 5.17 — MODEL III — STAGE IV — COX REGRESSION MODEL — SURVIVAL CURVES (PART I)

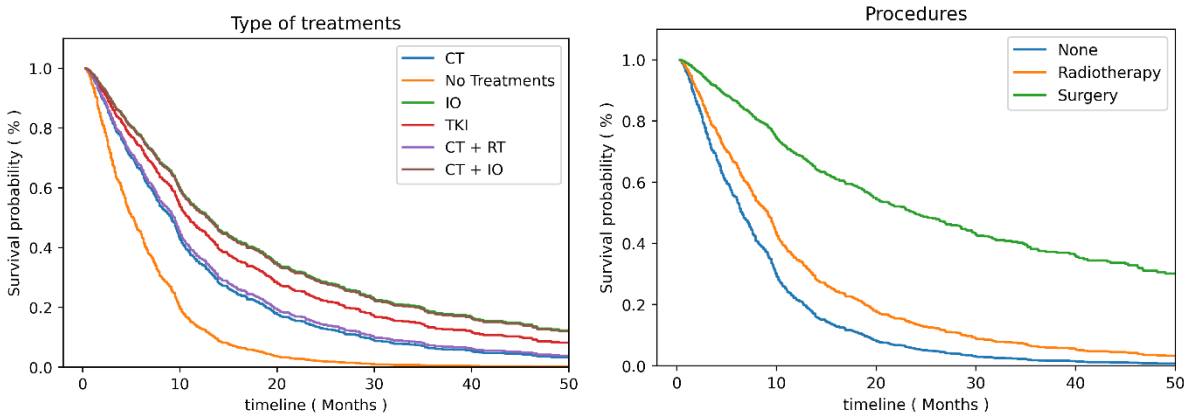


FIGURE 5.18 — MODEL III — STAGE IV — COX REGRESSION MODEL — SURVIVAL CURVES (PART II)

6.

CONCLUSIONS AND FUTURE WORK

6.1. CONCLUSIONS

Technology, with each passing day, has a greater influence and preponderance. With regard to health and especially cancer patients, it is essential that there are tools that can help doctors and health workers so that their patients have the greatest and best medical care they can have to overcome one of the worst phases of their lives.

The aim of this dissertation has been to study patients diagnosed with lung cancer, specifically Non-small Cell Lung Cancer, in order to analyze their survival and identify risk factors so that they can have a better quality of life and greater chances of survival.

It is crucial to understand the problems that cancer survivors face and what are their needs. It is also important to know what are the information systems that exist and how all patients' clinical information is stored and processed. It should be noted that these systems are quite complex, as there are several security and privacy precautions to be considered in this situation, so it is necessary to mention the extreme importance of protecting this type of data.

First, we start by examining which models are most suitable for the study in question and analyzing which algorithms are the most used.

From the moment that we receive the database, we applied the Knowledge Discovery concept. It should be noted that, given the complexity of the disease and the amount of information available, it is essential to have the help of specialists in the field of oncology, in order to better interpret the problem and choose which variables we will use.

Until we had the final database and while having a constant communication with the oncology department, several versions were made available because, as it is an updating database and, incongruent data were detected (during their analysis). After filtering which variables to use and correcting possible errors, the data were cleaned and prepared so that we could apply the models.

The next step was the creation of several models, with different variables, with several optimizations until we reached our final models.

Cox regression model was developed in 3 models, a single model for the early stages at diagnosis, given that there were not any noteworthy differences between stage I and stage II, a second model for stage III and another model for stage IV.

After optimizing the model, we were able to see in detail which variables are the most significant for each stage of cancer, and we can conclude that actually there are differences between the stages and which are the risk factors for each one of them. This made it clear that we had made the best decision by separating the stages by models. We were also able to rank profiles of patients with higher survival probabilities for each stage and the difference in the survival curves between variables in the same group.

We were able to conclude that, for advanced stages, women have a longer survival rate than men and that patients that smoke have a lower survival rate than non-smoking patients. The new treatment with immunotherapy alone or acting together with chemotherapy has very promising results.

As for patients from early stages, we see that the biggest and main factor is whether they are subject to surgery. For more detail on the analysis, we must go to the corresponding chapters.

Given that, according to lifelines, the library used to create the model, in order to have a good fit, the c-index of the models must be between 55% and 75% [69]. The models created were performed successfully and the agreement of the models is between 71% and 75%, which gives confidence in the results obtained as well as in their conclusions.

6.2. FUTURE WORKS

Regarding the models created, different age groups can be studied, or even the inclusion of other variables. As already mentioned, we used the same baseline profile for all models so that we can compare them, but we could have used different baselines for each model.

To apply the models, the variables that are being used must have all the patient's information, otherwise, these patients are excluded from the model. So, the information of the used datasets must contain the most reliable and complete data.

It would also be interesting to understand, as we can see in section 4.1.3 through the Kaplan-Meier estimates and despite in chapter 5, in the Cox regression model results, the difference between smokers and former smokers is not significant, why smoker patients are more likely to survive than former smokers if it is related not only to the stage of cancer but also to the patient's health conditions, such as performance. status or comorbidities.

Regarding the former smoker's variable, it should be noted that we do not have the information if the patients who stopped smoking if it was before or after the diagnosis because the database only have one variable, which prevents us from obtaining information over the patient's lifetime and it could be interesting to analyze the differences between these patients.

Using the same dataset, there are other relevant studies that can be explored within the scope of the presented problem, such as the estimation of the probability of relapse or the impact of treatment lines in the survival.

We started this dissertation with a database of 1050 patients, ending up with a database from February with information on 2133 patients. This means that, this type of database is constantly being updated in real time and, as such, these models will have to be constantly updated to be as correct as possible and perhaps find new risk factors that need to be considered.

With the evolution of machine learning and artificial intelligence techniques proving to be more and more efficient, trying to apply the Cox model together with neural networks may be a good start.

With the end of this dissertation, a dashboard is being developed by Holos S.A, a web application tool, with the main goal of helping doctors to have more and better information to provide the best treatments and support to all cancer patients. Where, this statistical analysis and the models are already available and being used by HUPHM.

BIBLIOGRAFIA

[1] «Clarify 2020», November 15, 2021. https://www.clarify2020.eu/

[2] «Word Health Organization». https://www.who.int/data/gho/data/themes/topics/causes-of-death

[3] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, e A. Jemal, «Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries», *CA: A Cancer Journal for Clinicians*, vol. 68, n. 6, pp. 394–424, Nov. 2018, doi: 10.3322/caac.21492.

[4] «Global Cancer Observatory», 15 de Abril de 2020. https://gco.iarc.fr/

[5] G. Sulik, «MEDICINE AND SOCIETY What Cancer Survivorship Means», p. 7.

[6] K. D. Miller *et al.*, «Cancer treatment and survivorship statistics, 2019», *CA A Cancer J Clin*, vol. 69, n. 5, pp. 363–385, Set. 2019, doi: 10.3322/caac.21565.

[7] C. S. Moskowitz *et al.*, «Breast Cancer After Chest Radiation Therapy for Childhood Cancer», *JCO*, vol. 32, n. 21, pp. 2217–2223, Jul. 2014, doi: 10.1200/JCO.2013.54.4601.

[8] S. E. Allen, «Cancer Survivors: The Success Story That's Straining Health Care», *IEEE Pulse*, vol. 8, n. 1, pp. 14–17, Jan. 2017, doi: 10.1109/MPUL.2016.2629999.

[9] M. S. McCabe *et al.*, «American Society of Clinical Oncology Statement: Achieving High-Quality Cancer Survivorship Care», *JCO*, vol. 31, n. 5, pp. 631–640, Fev. 2013, doi: 10.1200/JCO.2012.46.6854.

[10] S. V. Hudson *et al.*, «Adult Cancer Survivors Discuss Follow-up in Primary Care: "Not What I Want, But Maybe What I Need"», *The Annals of Family Medicine*, vol. 10, n. 5, pp. 418–427, Set. 2012, doi: 10.1370/afm.1379.

[11] L. Nekhlyudov, P. A. Ganz, N. K. Arora, e J. H. Rowland, «Going Beyond Being Lost in Transition: A Decade of Progress in Cancer Survivorship», *JCO*, vol. 35, n. 18, pp. 1978–1981, Jun. 2017, doi: 10.1200/JCO.2016.72.1373.

[12] C. D. Runowicz *et al.*, «American Cancer Society/American Society of Clinical Oncology Breast Cancer Survivorship Care Guideline», *CA: A Cancer Journal for Clinicians*, vol. 66, n. 1, pp. 43–73, 2016, doi: 10.3322/caac.21319.

[13] M. Jefford, J. Rowland, E. Grunfeld, M. Richards, J. Maher, e A. Glaser, «Implementing improved post-treatment care for cancer survivors in England, with reflections

from Australia, Canada and the USA», *Br J Cancer*, vol. 108, n. 1, pp. 14–20, Jan. 2013, doi: 10.1038/bjc.2012.554.

[14] C. M. Alfano, M. Jefford, J. Maher, S. A. Birken, e D. K. Mayer, «Building Personalized Cancer Follow-up Care Pathways in the United States: Lessons Learned From Implementation in England, Northern Ireland, and Australia», *American Society of Clinical Oncology Educational Book*, n. 39, pp. 625–639, Mai. 2019, doi: 10.1200/EDBK_238267.

[15] K. Inamura, «Lung Cancer: Understanding Its Molecular Pathology and the 2015 WHO Classification», *Front. Oncol.*, vol. 7, p. 193, Ago. 2017, doi: 10.3389/fonc.2017.00193.

[16] S. Chheang e K. Brown, «Lung Cancer Staging: Clinical and Radiologic Perspectives», *Semin intervent Radiol*, vol. 30, n. 02, pp. 099–113, Mai. 2013, doi: 10.1055/s-0033-1342950.

[17] C. Allemani *et al.*, «Global surveillance of cancer survival 1995–2009: analysis of individual data for 25 676 887 patients from 279 population-based registries in 67 countries (CONCORD-2)», *The Lancet*, vol. 385, n. 9972, pp. 977–1010, Mar. 2015, doi: 10.1016/S0140-6736(14)62038-9.

[18] H. K. Biesalski *et al.*, «European Consensus Statement on Lung Cancer: risk factors and prevention. Lung Cancer Panel», *CA: A Cancer Journal for Clinicians*, vol. 48, n. 3, pp. 167–176, Mai. 1998, doi: 10.3322/canjclin.48.3.167.

[19] N. Akhtar e J. G. Bansal, «Risk factors of Lung Cancer in nonsmoker», *Current Problems in Cancer*, vol. 41, n. 5, pp. 328–339, Set. 2017, doi: 10.1016/j.currproblcancer.2017.07.002.

[20] J. Malhotra, M. Malvezzi, E. Negri, C. La Vecchia, e P. Boffetta, «Risk factors for lung cancer worldwide», *Eur Respir J*, vol. 48, n. 3, pp. 889–902, Set. 2016, doi: 10.1183/13993003.00359-2016.

[21] «webmd - cancer», October 26, 2021. https://www.webmd.com/cancer/cancer-stages

[22] R. Arriagada *et al.*, «Long-Term Results of the International Adjuvant Lung Cancer Trial Evaluating Adjuvant Cisplatin-Based Chemotherapy in Resected Lung Cancer», *JCO*, vol. 28, n. 1, pp. 35–42, Jan. 2010, doi: 10.1200/JCO.2009.23.2272.

[23] S. Burdett *et al.*, «Adjuvant chemotherapy for resected early-stage non-small cell lung cancer», *Cochrane Database of Systematic Reviews*, Mar. 2015, doi: 10.1002/14651858.CD011430.

[24] A. Aupérin *et al.*, «Meta-Analysis of Concomitant Versus Sequential Radiochemotherapy in Locally Advanced Non–Small-Cell Lung Cancer», *JCO*, vol. 28, n. 13, pp. 2181–2190, Mai. 2010, doi: 10.1200/JCO.2009.26.2543.

[25] «ECOG PS», October 27,2021. https://ecog-acrin.org/resources/ecog-performance-status

[26] World Cancer Research Fund International, *Survivors of breast and other cancers*.

[27] «Howlader N, Noone AM, Krapcho M, Miller D, Brest A, Yu M, Ruhl J, Tatalovich Z, Mariotto A, Lewis DR, Chen HS, Feuer EJ, Cronin KA (eds). SEER Cancer

Statistics Review, 1975-2017, National Cancer Institute. Bethesda, MD, https://seer.cancer.gov/csr/1975_2017/, based on November 2019 SEER data submission, posted to the SEER web site, April 2020.», 9 de Novembro de 2020. https://seer.cancer.gov/csr/1975_2017/

[28] P. Yang, «Epidemiology of Lung Cancer Prognosis: Quantity and Quality of Life», em *Cancer Epidemiology*, vol. 471, M. Verma, Ed. Totowa, NJ: Humana Press, 2009, pp. 469–486. doi: 10.1007/978-1-59745-416-2_24.

[29] C. Leduc, D. Antoni, A. Charloux, P.-E. Falcoz, e E. Quoix, «Comorbidities in the management of patients with lung cancer», *Eur Respir J*, vol. 49, n. 3, p. 1601721, Mar. 2017, doi: 10.1183/13993003.01721-2016.

[30] P. ORYGrINeAvLieNwA, «Comorbidities in lung cancer», *Pneumonologia i Alergologia Polska*, vol. 84, n. 3, p. 7, 2016.

[31] S. Y. Lee, S. J. Kim, J. Shin, K.-T. Han, e E.-C. Park, «The impact of job status on quality of life: general population versus long-term cancer survivors», p. 8, 2015.

[32] P. A. Ganz, A. Coscarelli, C. Fred, B. Kahn, M. L. Polinsky, e L. Petersen, «Breast cancer survivors: Psychosocial concerns and quality of life», *Breast Cancer Res Tr*, vol. 38, n. 2, pp. 183–199, Jun. 1996, doi: 10.1007/BF01806673.

[33] K. Carlsen, S. O. Dalton, F. Diderichsen, e C. Johansen, «Risk for unemployment of cancer survivors: A Danish cohort study», *European Journal of Cancer*, vol. 44, n. 13, pp. 1866–1874, Set. 2008, doi: 10.1016/j.ejca.2008.05.020.

[34] K. S. Choi *et al.*, «Job loss and reemployment after a cancer diagnosis in Koreans—a prospective cohort study», *Psycho-Oncology*, vol. 16, n. 3, pp. 205–213, Mar. 2007, doi: 10.1002/pon.1054.

[35] T. Taskila e M. L. Lindbohm, «Factors affecting cancer survivors' employment and work ability», *Acta Oncologica*, vol. 46, n. 4, pp. 446–451, Jan. 2007, doi: 10.1080/02841860701355048.

[36] C. C. Earle *et al.*, «Employment Among Survivors of Lung Cancer and Colorectal Cancer», *JCO*, vol. 28, n. 10, pp. 1700–1705, Abr. 2010, doi: 10.1200/JCO.2009.24.7411.

[37] C. Schwedhelm, H. Boeing, G. Hoffmann, K. Aleksandrova, e L. Schwingshackl, «Effect of diet on mortality and cancer recurrence among cancer survivors: a systematic review and meta-analysis of cohort studies», *Nutr Rev*, vol. 74, n. 12, pp. 737–748, Dez. 2016, doi: 10.1093/nutrit/nuw045.

[38] S. Zhong *et al.*, «Association between physical activity and mortality in breast cancer: a meta-analysis of cohort studies», *Eur J Epidemiol*, vol. 29, n. 6, pp. 391–404, Jun. 2014, doi: 10.1007/s10654-014-9916-1.

[39] L. M. Buffart, D. A. Galvão, J. Brug, M. J. M. Chinapaw, e R. U. Newton, «Evidence-based physical activity guidelines for cancer survivors: Current guidelines, knowledge gaps and future research directions», *Cancer Treatment Reviews*, vol. 40, n. 2, pp. 327–340, Mar. 2014, doi: 10.1016/j.ctrv.2013.06.007.

[40] L. M. Buffart *et al.*, «Effects and moderators of exercise on quality of life and physical function in patients with cancer: An individual patient data meta-analysis of 34

RCTs», *Cancer Treatment Reviews*, vol. 52, pp. 91–104, Jan. 2017, doi: 10.1016/j.ctrv.2016.11.010.

[41] «https://www.wcrf.org/dietandcancer/cancer-prevention-recommendations», *Cancer Prevention Recommendations*, September 27 ,2020. https://www.wcrf.org/dietandcancer/c

[42] C. L. Rock *et al.*, «Nutrition and physical activity guidelines for cancer survivors», *CA: A Cancer Journal for Clinicians*, vol. 62, n. 4, pp. 242–274, Jul. 2012, doi: 10.3322/caac.21142.

[43] M. J. Resnick, C. Lacchetti, e D. F. Penson, «Prostate Cancer Survivorship Care Guidelines: American Society of Clinical Oncology Practice Guideline Endorsement», *JOP*, vol. 11, n. 3, pp. e445–e449, Mai. 2015, doi: 10.1200/JOP.2015.004606.

[44] T. Sanft *et al.*, «NCCN Guidelines Insights: Survivorship, Version 2.2019», *Journal of the National Comprehensive Cancer Network*, vol. 17, n. 7, pp. 784–794, Jul. 2019, doi: 10.6004/jnccn.2019.0034.

[45] L. H. Kushi *et al.*, «American Cancer Society Guidelines on Nutrition and Physical Activity for Cancer Prevention: Reducing the Risk of Cancer With Healthy Food Choices and Physical Activity», *CA: A Cancer Journal for Clinicians*, vol. 56, n. 5, pp. 254–281, Set. 2006, doi: 10.3322/canjclin.56.5.254.

[46] J. Arends *et al.*, «ESPEN guidelines on nutrition in cancer patients», *Clinical Nutrition*, vol. 36, n. 1, pp. 11–48, Fev. 2017, doi: 10.1016/j.clnu.2016.07.015.

[47] J. A. Ligibel *et al.*, «American Society of Clinical Oncology Position Statement on Obesity and Cancer», *JCO*, vol. 32, n. 31, pp. 3568–3574, Nov. 2014, doi: 10.1200/JCO.2014.58.4680.

[48] W. Y. Cheung, B. A. Neville, e C. C. Earle, «Associations Among Cancer Survivorship Discussions, Patient and Physician Expectations, and Receipt of Follow-Up Care», *JCO*, vol. 28, n. 15, pp. 2577–2583, Mai. 2010, doi: 10.1200/JCO.2009.26.4549.

[49] C. M. Alfano, E. E. Kent, L. S. Padgett, M. Grimes, e J. S. de Moor, «Making Cancer Rehabilitation Services Work for Cancer Patients: Recommendations for Research and Practice to Improve Employment Outcomes», *PM&R*, vol. 9, pp. S398–S406, Set. 2017, doi: 10.1016/j.pmrj.2017.06.019.

[50] H. M. Penttinen *et al.*, «Quality of life and physical performance and activity of breast cancer patients after adjuvant treatments», *Psycho-Oncology*, vol. 20, n. 11, pp. 1211–1220, Nov. 2011, doi: 10.1002/pon.1837.

[51] L. Warrington, K. Absolom, e G. Velikova, «Integrated care pathways for cancer survivors – a role for patient-reported outcome measures and health informatics», *Acta Oncologica*, vol. 54, n. 5, pp. 600–608, Mai. 2015, doi: 10.3109/0284186X.2014.995778.

[52] M. Ouwens, «Integrated care programmes for chronically ill patients: a review of systematic reviews», *International Journal for Quality in Health Care*, vol. 17, n. 2, pp. 141–146, Abr. 2005, doi: 10.1093/intqhc/mzi016.

[53] T. A. Skolarus *et al.*, «American Cancer Society prostate cancer survivorship care guidelines: American Cancer Society Prostate Cancer Survivorship Guidelines», *CA A Cancer Journal for Clinicians*, vol. 64, n. 4, pp. 225–249, Jul. 2014, doi: 10.3322/caac.21234.

[54] G. S. Birkhead, M. Klompas, e N. R. Shah, «Uses of Electronic Health Records for Public Health Surveillance to Advance Public Health», *Annu. Rev. Public Health*, vol. 36, n. 1, pp. 345–359, Mar. 2015, doi: 10.1146/annurev-publhealth-031914-122747.

[55] C. Martínez-Costa e S. Schulz, «Validating EHR clinical models using ontology patterns», *Journal of Biomedical Informatics*, vol. 76, pp. 124–137, Dez. 2017, doi: 10.1016/j.jbi.2017.11.001.

[56] S. Silvestri, A. Esposito, F. Gargiulo, M. Sicuranza, M. Ciampi, e G. De Pietro, «A Big Data Architecture for the Extraction and Analysis of EHR Data», em *2019 IEEE World Congress on Services (SERVICES)*, Milan, Italy, Jul. 2019, pp. 283–288. doi: 10.1109/SERVICES.2019.00082.

[57] P. M. Nadkarni, L. Ohno-Machado, e W. W. Chapman, «Natural language processing: an introduction», *J Am Med Inform Assoc*, vol. 18, n. 5, pp. 544–551, Set. 2011, doi: 10.1136/amiajnl-2011-000464.

[58] U. Fayyad, G. Piatetsky-Shapiro, e P. Smyth, «The KDD process for extracting useful knowledge from volumes of data», *Commun. ACM*, vol. 39, n. 11, pp. 27–34, Nov. 1996, doi: 10.1145/240455.240464.

[59] I. Keshta e A. Odeh, «Security and privacy of electronic health records: Concerns and challenges», *Egyptian Informatics Journal*, vol. 22, n. 2, pp. 177–183, Jul. 2021, doi: 10.1016/j.eij.2020.07.003.

[60] J. Fan e J. Jiang, «Non- and Semi- Parametric Modeling in Survival Analysis», em *Frontiers of Statistics*, vol. 1, CO-PUBLISHED WITH HIGHER EDUCATION PRESS, 2009, pp. 3–33. doi: 10.1142/9789812837448_0001.

[61] E. T. Lee e J. W. Wang, *Statistical methods for survival data analysis*, 3. ed. Hoboken, NJ: Wiley-Interscience, 2003.

[62] J. Kishore, M. Goel, e P. Khanna, «Understanding survival analysis: Kaplan-Meier estimate», *Int J Ayurveda Res*, vol. 1, n. 4, p. 274, 2010, doi: 10.4103/0974-7788.76794.

[63] C. Rocha e A. L. Papoila, *Análise de sobrevivência*. Sociedade Portuguesa de Estatística.

[64] D. R. Cox, «Regression Models and Life-Tables», *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, n. 2, pp. 187–202, Jan. 1972, doi: 10.1111/j.2517-6161.1972.tb00899.x.

[65] D. G. Kleinbaum e M. Klein, *Survival analysis: a self-learning text*, 2nd ed. New York, NY: Springer, 2005.

[66] D. Schoeneeld, «Partial residuals for the proportional hazards regression model», p. 3.

[67] A. Moncada-Torres, M. C. van Maaren, M. P. Hendriks, S. Siesling, e G. Geleijnse, «Explainable machine learning can outperform Cox regression predictions and provide insights in breast cancer survival», *Sci Rep*, vol. 11, n. 1, p. 6968, Dez. 2021, doi: 10.1038/s41598-021-86327-7.

[68] «Quick dive into Pandas for Data Science». https://towardsdatascience.com/quick-dive-into-pandas-for-data-science-cc1c1a80d9c4

[69] «lifelines documentation», October 29,2021. https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html#log-likelihood

⟨2021⟩

Alexandre Miguel Ramos de Sousa

Survival outcomes and prognosis in non-small cell lung cancer patients in a tertiary hospital in Spain

NOVA SCHOOL OF SCIENCE & TECHNOLOGY