Filipa Mendes de Matos

Bachelor's in Electrical and Computer Science

Spanish Non-small Cell Lung Cancer Patients - A Survival Analysis

Dissertation submitted in partial fulfillment of the requirements for the degree of

Master of Science in Electrical and Computer Engineering

Adviser: Pedro Alexandre da Costa Sousa, Associate Professor,
NOVA University of Lisbon
Co-adviser: Gracinda Rita Guerreiro, Associate Professor,
NOVA University of Lisbon

September 2021

FACULDADE DE
CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

To anyone who believes that everything we can imagine can become real.

ACKNOWLEDGEMENTS

It is impossible to thank here to everyone that was part of this phase and helped me on several levels which together resulted in this dissertation.

Firstly I want to thank Prof. Pedro Sousa, Prof. Gracinda, and Dr Maria Torrento that challenged and guided me throughout the steps. More than the knowledge shared, incredible and passionate people who work to find ways to make real what they want to improve in the world. Also, to my colleague Alexandre Sousa for all the sharing and good teamwork during these months.

To my family for the unconditional support, courage and love.

Finally, I would like to thank all my friends and my boyfriend for their patience, encouragement and most of all for being a part of my life.

“Se podes olhar, vê. Se podes ver, repara."

José Saramago

Resumo

A Saúde constitui um bem inestimável que prevalece em qualquer sociedade ao longo da história do ser humano. Sendo um bem inestimável os humanos estão dispostos a todos os sacrifícios de modo a assegurarem este precioso bem.
As técnicas e tecnologias usadas ao longo do tempo, a sua massificação e constante evolução acabam por elevar esta indústria extremamente competitiva e cada vez mais sofisticada numa procura constante e direcionada para a necessidades mais urgentes da população humana. A digitalização dos processos clínicos e a imensidão de dados adquiridos constitui atualmente um enorme capital de conhecimento.

O cancro é a segunda maior causa de morte no mundo, com tendência crescente, e consequentemente uma preocupação para a organização mundial de saúde que está a tentar reverter esta tendência. Neste contexto, há uma necessidade urgente de projetos e investigações, como esta dissertação, que visam antecipar o diagnóstico, prevenir as repercussões da doença, melhorar estilos de vida e aumentar as taxas de sobrevivência em pacientes com cancro. Este trabalho concentra-se explicitamente em pacientes com cancro do pulmão. O cancro de pulmão é um dos cancros mais comuns em todo o mundo, sendo a causa global mais comum de morte por cancro em homens e a terceira mais comum em mulheres. O cancro do pulmão de células não pequenas é responsável por aproximadamente 80% de todas as doenças malignas do pulmão. Além disso, a incidência de cancro do pulmão tem aumentando gradualmente nos últimos 50 anos, tornando-se um problema de saúde pública mundial.

O conjunto de dados analisado é específico para pacientes com cancro de pulmão de células não pequenas e todos os seus atributos foram revistos de uma perspectiva clínica.
Após a compreensão do conteúdo do conjunto de dados, seguiu-se a fase de pré-processamento dos dados, uma análise descritiva de cada atributo e a utilização do método de Kaplan-Meier. Finalmente, este trabalho propõe o uso do modelo de risco proporcional multivariado de Cox.
Além disso, esta dissertação revê o domínio de aplicações, incluindo a estrutura da indústria de Saúde e Sistemas de Informação, tal como o conhecimento técnico necessário para implementar algoritmos de aprendizagem automática.

Esta dissertação é apoiada pela Holos S.A. e envolvida no projeto CLARIFY (European Union Horizon 2020- ao abrigo do acordo da bolsa nº 875160).

Palavras-chave: Saúde, cancro do pulmão de células não pequenas, Kaplan-Meier, modelo de risco proporcional multivariado de Cox, aprendizagem automática.

Abstract

Health is an invaluable asset that prevails in any society throughout human history. As a priceless good, humans are willing to make all sacrifices to ensure this precious good.

The techniques and technologies used over time, their massification and constant evolution, end up raising a highly competitive and increasingly sophisticated industry in a constant search directed to the population's most pressing needs. The digitalization of clinical processes and the immensity data acquired is currently an enormous knowledge capital.

Cancer is the second leading cause of death worldwide, and its increasing number is a concern for world health organisations that are trying to reverse this trend. In this context, there is an urgent need for projects and research such as this dissertation that aims to achieve early diagnosis, predict and prevent disease repercussions, improve quality of life, and increase survivorship rates in cancer patients. This work focuses explicitly on lung cancer patients. Lung cancer is one of the most typical cancers worldwide, being the most common global cause of cancer death in men and the third most common in women. Non-small cell lung cancer (NSCLC) accounts for approximately 80% of all lung malignancies. In addition, the incidence of lung cancer has been gradually increasing over the last 50 years, becoming a worldwide public health issue.

The dataset analysed is specific for non-small cell lung cancer patients, and all its attributes were reviewed from a clinical perspective.

After understanding the dataset's content and the pre-processing data phase followed a descriptive analysis of each attribute, and the use of the Kaplan-Meier method. Finally, this work proposes the use of Cox's Multivariate Proportional Hazard Model.

Additionally, this dissertation reviews the applications domain, including the Healthcare industry structure and Information Systems and the technical knowledge necessary to implement Machine Learning algorithms.

This work was supported by Holos S.A. and engaged with the CLARIFY project (European Union Horizon 2020- under grant agreement nº 875160).

Keywords: Health, Machine Learning, Non-small cell lung cancer, Kaplan-Meier, Cox's Multivariate Proportional Hazard Model

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

CLARIFY	Cancer Long Survivor Artificial Intelligence Follow-up
NSCLC	Non-small cell lung cancer
ML	Machine Learning
AI	Artificial intelligence
HUPHM	Hospital Universitario Puerta de Hierro Majadahonda
WHO	World Health Organization
ICD	International Classification of Diseases
NLM	National Library of Medicine
UMLS	Unified Medical Language System
HIT	Healthcare Information Technology
EHR	Electronic Health Record
EMR	Electronic Medical Record
PHR	Personal Health Record
ISO	International Organization for Standardization
CPOE	Computerized Physician Order Entry
CDSS	Clinical Decision Support System
GDPR	General Data Protection Regulation
RBAC	Role-based Access Control
ABAC	Attribute-based access control
EGFR	Epidermal Growth Factor Receptor
ALK	Anaplastic Lymphoma Kinase

CHAPTER 1

INTRODUCTION

This first chapter provides the context in which this dissertation was developed and the motivation and objectives of the project.

CLARIFY is a project financed by the European Union Horizon 2020 Research and Innovation Programme, comprised by twelve partners, academic and non-academic, with a duration of three years, initiated in January 2020. Holos is one of the consortium partners and the only Portuguese company involved, and have therefore established the link between this dissertation and the project.

The purpose of this dissertation will be explained and structured in the following topics, as well as the contribution that it aims to give.

1.1 BACKGROUND AND MOTIVATION

Being cancer the second cause of death worldwide and an increasing number, world health organisations are trying to turn around the tendency of these numbers [1].

According to the GLOBOCAN (Global Cancer Observatory) [1], the latest statistics show that the 14 million new cancer cases in 2012 increased up to 18.1 million in 2018, and the predictions for the next decades are growing bigger.

In 2020, approximately 1.8 million new cancer cases and 606.520 cancer deaths were projected to occur in the United States [1]. Nevertheless, since 1991, the mortality rate has been declining yearly, resulting in a cancer death rate estimate of 27% in 2019 (2.6 million fewer cancer deaths [2]) and 29% in 2020 (2.9 million fewer cancer deaths).

Considering this information, we can verify that the number of cancer survivors is growing each year, which is one of the reasons this is a subject of enormous interest.

Survivorship Statistics show that 16.9 million Americans with a history of cancer were alive on January 1st, 2019; being the projections for 2030 more than 22.1 million, based on the growth of the population.

From this 16.9 million, 68% were diagnosed five or more years ago, and 18% were diagnosed 20 or more years ago [3].

The majority of the values presented are related to the US since there are no actual cancer survivor official registries in Europe.

In most cases, cancer is not experienced alone, and most of the symptoms experienced after treatment are similar to those experienced by families, friends and caregivers. Therefore, the three groups above mentioned are also included in the definition of cancer survivors, which takes the dimension of the study group to another level.

1.1 RATIONALE

Cancer is a matter of public health worldwide.

In the last decades, numerous discoveries have been made regarding cancer treatments, so the number of survivors has substantially increased.

A cancer survivor is defined as every person who suffers from cancer since the time of the first diagnosis. Long-term cancer survivors are those patients who are still alive 5 years or more post-diagnosis [4]. One of the main problems that these patients face after treatment is mid/long-term health problems for the rest of their lives. These repercussions can be of a physical, psychological or social type, as presented in Table 1.1.

Currently, models for monitoring cancer survivors merely analyse the patient's clinical history and history of disease recurrence. There are no evidence-based guidelines for the follow-up of cancer survivors and their follow-up model is inadequate and poorly protocolized, besides a lack of specialisation of the caregivers.

As the number of this study group is increasing, the Health Care systems will be overburdened, with fewer resources to assist and provide the necessary help until the point that they actually may be unable to deliver treatment and post-treatment follow-up care.

The massive lack of follow-up on sick patients and post-treatment patients after the COVID-19 pandemic has emphasized the urgency of the creation of new follow-up techniques, as the health industry is still not prepared to answer the population´s demands worldwide, and patients are getting sick and dying not because of the disease itself but as a consequence of the lack of treatment and follow-up.

Table 1-1 Cancer survivor's repercussions types and examples.

Physical health	**Psychological health**	**Social health**

Recurrence;
Long-term/ late effects of treatment;
Effects on
co-morbidities;
Death.
	Increased depression;
	Increased anxiety (including fear of recurrence);
	Psychosexual problem;
	Quality of life.
		Financial;
		Employment;
		Education;
		Interpersonal;
		Social integration;
		Disability.

1.2 PROPOSED SOLUTION

The healthcare industry is changing, already addressing technology to its evolution as a way to adapt to the needs of the population and environmental crises.

As a part of its evolutions comes the realisation and application of Machine Learning (ML) techniques and Artificial Intelligence (AI) in Healthcare that is already improving outcomes.

The main applications of AI in Healthcare are [4]:

- Prediction;
- Diagnosis;
- Personalised Treatment and Behavior Modification;
- Drug discovery;
- Follow-up care.

In this line of action, the CLARIFY project aims to create personalised pathways to follow-up cancer survivors, including a part of all applications mentioned above.

Having a predictive methodology, raising awareness, anticipating diagnosis and preventing disease repercussion will reduce the number of deaths, improve lifestyles, and increase survivorship rates.

1.3 CONTRIBUTIONS

The dataset studied in this dissertation was provided by the Medical Oncology Department of the Hospital Universitario Puerta de Hierro Majadahonda (HUPHM) as coordinators of the project. Cases were collected from the Spanish Thoracic Tumour Registry, a nationwide registry sponsored by the Spanish Lung Cancer Group. The final version of the dataset contains one-thousand two-hundred and forty-four (1244) attributes with each variable encoding and the respective description, with both idioms, Spanish and English. The dataset contained the medical register of fifteen thousand three hundred and thirty-seven (15 337) patients.

Since the project is at its early beginning, the first round of objectives and specifically the subjects where this dissertation aims to bring value to, are:

- Identification of the patient and disease factors and characteristics;
- Risk stratification of these same factors and characteristics;
- Predict the best models of follow-up.

1.4 OUTLINE OF THE DOCUMENT

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

- **Chapter 2** – State of the Art – This section presents the Healthcare Industry, going through the Healthcare Data, Analysis and Standards, as well as the Healthcare Information Systems.

- **Chapter 3** – Machine Learning and AI in Healthcare - The necessary knowledge for the practical implementation of the project is reviewed in this section. Definition of the main topics and associated terms, analysis streams, methodologies, and contextualise the project within these subjects.

- **Chapter 4** – Data Engineering – Here are presented the tools used in the development, and the dataset used in the project, going from its sources and formats to all the pre-processing work. The descriptive analysis performed is also presented in this section.

- **Chapter 5** – Cox's Multivariable Proportional hazard Model Results – The development of the models, their results and the proportional hazards assumption tests are presented here.

- **Chapter 6** – Conclusions and Future work – This last section reviews the work developed and analyses the results and challenges since the beginning of the dissertation, as well as the Future work which identifies the optimizations and evolutions of the work developed.

CHAPTER 2

STATE OF THE ART

To create value and add knowledge in any industry or business, it is necessary to understand the sources, lifecycles, meanings, the features involved in its transformation, the standards and regulations that passed through and finally, the purpose of all its data.

This chapter starts by providing the relevant prior knowledge around the Healthcare Industry, going through the Healthcare Data, Analysis and Standards, as well as the Healthcare Information Systems, structures, services and regulations.

2.1 CANCER

Health is one of the areas that will always prevail in any society and throughout history, as it is necessary to all human beings throughout all their lives. With this, health data is extremely sensitive and confidential.

Statistics concerning cancer patients/survivors are compiled on the following patient characterisation factors: sex, ethnicity, age, geographic territory, the status of the first diagnosis, genetics, etc. Over the years of statistics development, some of these factors have developed a notable and consistent discrepancy [5].

In order to understand the cancer statistics, it is crucial to be aware of the real meaning of some definitions, as well as their imprecisions, considering that many factors are not taken into account when developing these numbers. Therefore, before presenting some of the most relevant disparities, we will get into some definitions.

A statistic itself is an approximation of the reality, being excluded factors that can drastically modify them. In this case, the following points are concerns that should not be left aside when looking into the health subject:

- The fact that the data being analysed has an enormous dimension and different sources, which is often a synonym of lack of accuracy and certainty, which leads to misclassifications.
- Factors that can be common to both variables of the equation and are not taking into account. Smoking is an excellent example of this inaccuracy since it is a cause

of lung cancer, and it is also a reason for many other diseases. So, there are factors related to the risk of cancer that may also be related to the risk of dying from other causes.
- Factors that can influence the statistics but are not considered because there are not enough data and scientific reports about it, or even sometimes it is impossible to calculate them.

Relative survival, which estimates the ratio of **observed survival** to **expected survival**, is an excellent example of the concerns stated above.

Observed survival is the actual rate of patients that are still alive for a certain period after diagnosis; on the other hand, expected survival is based on the survival rates of the entire population. Both survivals consider the usual features, such as age, sex, race, and year of patient diagnosis.

Relative survival estimates the probability that a patient will not die of the diagnosed cancer (within a given time interval), assuming that the presence of cancer, in this case, is the only factor that distinguishes the cancer patient from the general population.

If we look at this example, smoking, besides being the primary reason for lung cancer, is also a risk factor for other diseases, so smokers have a shorter life expectancy than nonsmokers. Consequently, expected survival will be unrealistically high, which will be translated into lower relative survival. If the population considered was only smokers instead of the standard life tables, this number would be much more precise [6]. In the same line of thought, having the actual number of the smoking population would be difficult, and even though we may calculate it, it would be necessary to quantify it, which is even more challenging.

2.2 HEALTHCARE DATA

This subchapter aims to review the different Data Categories that Healthcare Information Technologies uses throughout all industry and infrastructure.

There are four data categories, presented in Table 2.1., which are used in every clinical institution or hospital, being used by clinicians, nurses and other medical staff, caregivers and lawyers, hospital administration, and researchers [7]. This data access is ruled by several standards and laws, which are presented in the next chapter.

As well as the access of this data, its creation is also standardised and must follow the different types of imposed classifications to be credible to the world.

Table 2-1 Summary of healthcare data categories.

	Demographic Data	**Socioeconomic Data**	**Financial Data**	**Clinical Data**
Definition	Elements that differentiate patients.	Elements that characterise a patient's personal life.	Information regarding the payer.	All the medical data is recorded during the care process or as a part of a clinical trial.
Examples	Name, birth date, address, etc**.**	Religion/culture, habits, marital status, etc.	Name, address, etc., of the patient's insurance company.	Ordered exam, medical analysis such as narrative reports, comments, etc.

2.2.1 DATA STANDARDS IN HEALTHCARE

The European Committee for Standardization, CEN is an association that brings together the National Standardization Bodies of 34 European countries [8]. On the other hand, ISO, International Organization for Standardization, is an independent, non-governmental international organisation with a membership of 165 national standards bodies [9]. These two entities bring together the most crucial standards for the health industry.

During the International Health Informatics Seminar in 2019 handled by CEN, a resume of the existing Health Informatics Standards was made, which enabled the realisation of the existing dimension of the health informatics standards scope.

Concerning the Health Informatics subject, ISO has currently 187 published standards and CEN 102. Each entity has more than 20 standards in development. Each year, standards suffer updates, and new versions are published [1].

Data standards are the rules by which data is described and recorded, making it easier to create, share, understand and integrate. These standards provide trust in the data presented and allow the flow of information through the healthcare infrastructure.

Data standards can be divided into four categories [10,11,12], and some examples are presented in Table 2.2.:

1. Data elements and Content- The determinations of the data content to be collected and exchanged;
2. Data interchange formats- Standard formats for electronically encoding the data elements, including document architecture for structuring data elements;
3. Terminologies- The medical terms and concepts used to describe, classify and code the data elements and data expression languages and syntax;
4. Knowledge representation- Standard methods for electronically representing medical literature and clinical guidelines.

[1] It is essential to realise that the number of standards in healthcare is enormous. Here we will focus on the ones which are vital to the practical understanding of the data and systems used during this dissertation. For more information on the standards, see Dr Robert A. Stegwee, 12/04/2019 Health-Informatics-Standardization-Seminar.

Table 2-2 Summary of Standards and Developer Organisations.

Standard Category	Standard	Acronym	Description	Developer
Data elements and content	Version 2	V2	Messaging protocol; several of the chapters of this standard cover clinical content	Health Level Seven (HL7)
	Health informatics — Harmonised data types for information interchange	ISO 21090	Offer a practical and valuable contribution to the internal design of health information systems.	International Organization for Standardization (ISO) https://www.iso.org/obp/ui/#iso:std:35646:en
Data interchange formats	Digital Imaging and Communications in Medicine Committee	DICOM	Format for communicating radiology images and data.	National Electronics Manufacturers Association (NEMA)
	Health informatics – Electronic Health Record Communications- Part 1- Reference Model	ISO 13606-1: 2019	Electronic Health Record Communications	International Organization for Standardization (ISO) https://www.iso.org/obp/ui/#iso:std:35646:en
	Fast Health Interop Resources	FHIR®	Interoperability standard intended to facilitate the exchange of healthcare information.	Health Level Seven (HL7)
Terminologies	International Classification of Diseases -10	ICD-10/ ICD-11	Disease Classification	World Health Organization (WHO)
	Systematized Nomenclature of Medicine	SNOMED	Clinical terminology	College of American Pathologists
	Logical Observation Identifiers Names and Codes	LOINC	Laboratory test, data elements.	Regenstrief Institute

	Unified Medical Language System	UMLS		National Library of Medicine (NLM)
	Clinical Document Architecture	(CDA®)	Specifies the structure and semantics of "clinical documents".	Health Level Seven (HL7)
Knowledge representation and Systems requirements	Health informatics — Requirements for an electronic health record architecture	ISO 18308	Standard for Electronic Health Record system requirements.	International Organization for Standardization (ISO) https://www.iso.org/obp/ui/#iso:std:35646:en

2.2.2 CLINICAL DATA AND MEDICAL CODING

Clinical Data is collected by clinicians and other medical staff, to provide the most accurate and complete information to enable them to make the right decisions regarding the patient's health.

Clinical Data is an essential component of distinct Electronics and Clinical Systems, so, before getting into that, this chapter focuses on understanding the different formats and standards at the roots of its creation.

During the healthcare process, much information, with different sources and formats, is generated, divided into three main groups: medical images, clinical notes, and other [13].

Medical images are all the images that come from exams (for example, X-rays, Computed Tomography (CT), Magnetic Resonance Imaging (MRI), microscopy image, Positron Emission Tomography (PET) between others) [14-15]. With the development and application of learning techniques, the utility and value of medical images analysis have been pushing medicine forwards.

Clinical notes are the most abundant data type; in other words, this type of data carries most of the patient's clinical information. Clinical narrative reports such as admission notes, treatment plans, pre-discharge summaries, discharge summaries, and death certificates are highly heterogeneous and complex to analyse.

The category mentioned above as 'other' includes all the remaining data collected during the care process but is not a part of the two first data groups, including lab results, vital signs, demographic information, payment and insurance information, etc. [14].

Focusing on the Clinical notes, which constitute a significant part of the data that will be analysed in this dissertation, it is the data group with a vast number of standards.

Many countries differ in healthcare models, structures and governance, but the following classifications are international and recognised worldwide.

For example, the International Classification of Diseases (ICD), supported by the World Health Organization (WHO), is the global health information standard for mortality and morbidity statistics, translated into 43 languages (International Classification of Diseases (ICD) Information Sheet, 2020) [16] and it is the foundation of the analysis of clinical medical notes.

ICD has had new versions and updates over the years, and the last available version is the ICD-10 [17].

ICD allows easy storage, retrieval and analysis of health information for evidence-based decision-making; sharing and comparing health information between hospitals, regions, settings and countries; and data comparisons in the same location across different periods [18].

There are many other clinical reference terminologies, not just regarding diseases, for example, the Systematized Nomenclature of Medicine Clinical Terms (SNOMED CT) [19], Logical Observation Identifiers Name and Codes (LOINC), which aims to report laboratory and other clinical observations [20] between others, as presented in Table 2.2.

Two-hundred and thirteen (213) vocabularies sources are identified by the US National Library of Medicine (NLM); all of them are used in the Unified Medical Language System (UMLS Metathesaurus Vocabulary Documentation, 2020) [21].

The Unified Medical Language System (UMLS) combines all these terminologies with software, Figure 2.1., providing the different computer systems with the necessary interoperability to change data.

These vocabulary sources are denominated by Metathesaurus, which is one of three sources of knowledge of UMLS.

Considering that clinical notes are narrative text, Semantic and Lexical tools are the second and third sources of knowledge, which have an essential role when treating this type of data, and are heavily correlated within UMLS Metathesaurus.

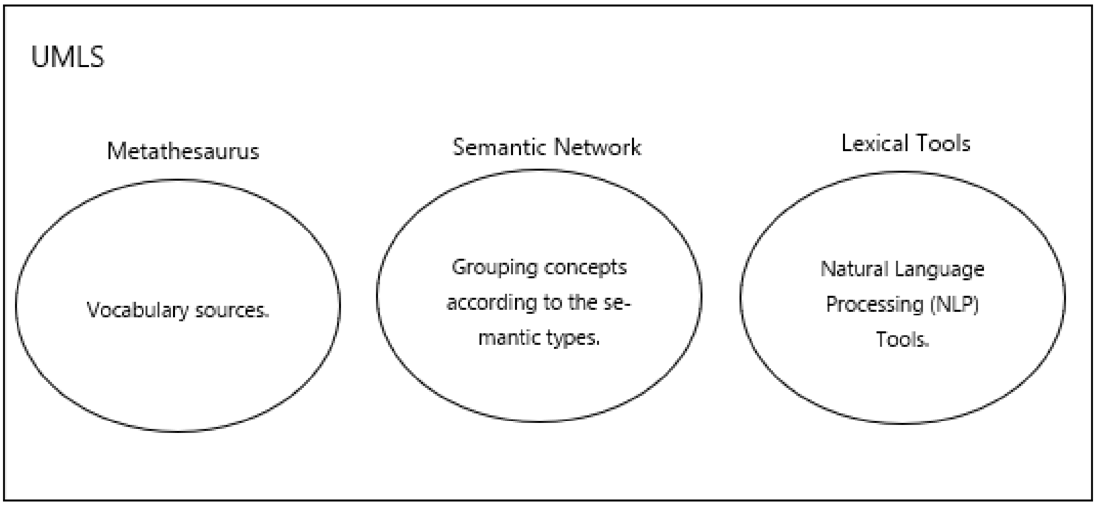

UMLS
Metathesaurus
Semantic Network
Lexical Tools
Vocabulary sources.
Grouping concepts according to the semantic types.
Natural Language Processing (NLP) Tools.

Figure 2-1 Unified Medical Language System (UMLS) sources.

The UMLS semantic network reduces the complexity of the Metathesaurus by grouping concepts according to the semantic types that have been assigned to them [22]. Computable semantic interoperability requires that the meaning of data is ambiguously exchanged from machine to machine, ensuring that the meaning exchange between humans is the same.

On the other hand, the lexical tools focus on the syntactic lexicon of biomedical and general English, using several tools to normalise strings, generate lexical variants and create indexes [21]. Syntactic interoperability enables the exchange of the structure of the data.

The Lexical Systems Group of The Lister Hill National Center of Biomedical Communications developed the SPECIALIST Natural Language Processing (NLP) Tools, which helps application developers with lexical variation and text analysis tasks in the biomedical domain (The SPECIALIST NLP Tools, 2020) [23].

The Natural Language Processing (NLP) techniques are fundamental to the mediation between the user's language (clinicians) and the language of online biomedical information resources. It is due to these techniques that it is possible to extract structured information from narrative clinical text.

2.3 HEALTHCARE INFORMATION SYSTEMS

The previous chapter explained the standards concerning the creation and exchange of medical data. This chapter will review the systems used in the healthcare industry aligned with these standards.

Healthcare Information Technology (HIT) is the intersection of Technology, Health Informatics, Information, and Systems, Figure 2.2.[24]. This section targets the Systems used in HIT, with a particular focus on Electronic Health Records.

In order to have an end-to-end vision of the systems themselves, it is made an overview of the most relevant electronic services, which are strongly correlated with the data analysed in the upcoming chapters.

As expected, in the review of the Healthcare Information Systems, it could not be missing the Security subject. Healthcare is one of the most complex businesses [25], with an immense diversity of interactions and quite sensible data, making security one of the biggest trends in the healthcare industry. Therefore, it will be presented the goals of information security, the core entities responsible for assuring security at the enterprise and international level, and finally, the different types of access controls implemented in the systems.

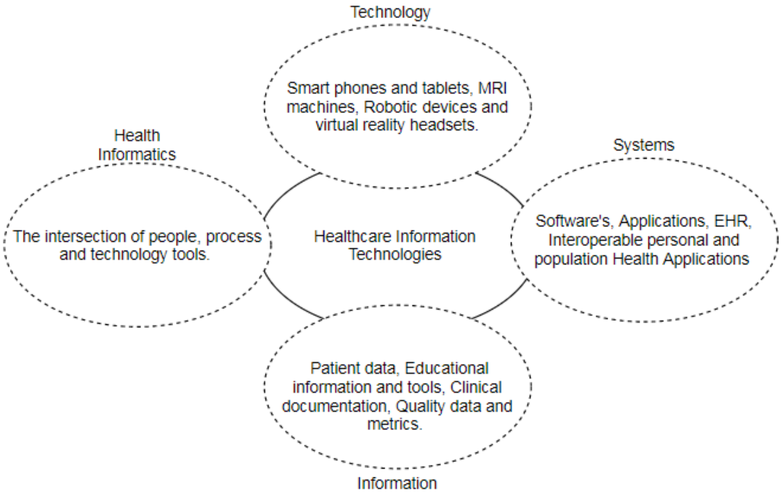

Technology
Smart phones and tablets, MRI machines, Robotic devices and virtual reality headsets.
Health Informatics
The intersection of people, process and technology tools.
Healthcare Information Technologies
Systems
Software's, Applications, EHR, Interoperable personal and population Health Applications
Patient data, Educational information and tools, Clinical documentation, Quality data and metrics.
Information

Figure 2-2 Healthcare Information Technologies components.

2.3.1 CLINICAL INFORMATION SYSTEMS

An Electronic Health Record (EHR) is an electronic version of a patient's paper record. It is a repository of the patient's data. It includes all the medical treatment histories of the patient, diagnosis, medications, laboratory, and test results, among all the information generated before, during and after the care process. It is the record where healthcare providers create, import, store and use comprehensive clinical information for patient care, including all the data types studied in the Healthcare Data chapter [26-27].

A range of other different standard definitions for the EHR lead ISO to categorise these several other names for EHR, including Electronic Medical record (EMR), Electronic Patient Record (EPR), Computerized Patient Record (CPR), Personal Health Record (PHR) between others [28]. The difference between these terms is mainly related to the access control topic.

The Electronic Health Record Architecture (EHRA) definition used in technical specification, as most of the standards, is

> *"The generic structural components from which all EHRs are built, defined in terms of an information model."*

The information in the EHR is hierarchical. The standard ISO 13606-1: 2019- Health Informatics- Electronic Health Record communications is a guide that provides a detailed review of this hierarchy, with the main objective of improving the alignment with other standards within the Health Informatics subject and promote communication.

The EHR Communications Reference Model needs to reflect this hierarchical organisation, meeting the published protocols. The reference model is specified as an information model, representing the global characteristics of the health records and the aggregation between them. This model defines a set of classes that form the generic building blocks of the EHR [29].

The combination of these building blocks for a specific clinical domain is known as Archetype.

An archetype specifies a particular hierarchy of the record components, for example, rules by which the clinical templates should be constructed.

To ensure the meaning and provide a consistent mapping of individual EHR between heterogeneous clinical systems, the standard ISO 13606-1 sub-divides the EHR hierarchy into seven main components [29-30]:

1. EHR is a top-level component for a single subject of care, in other words, one-person electronic record.
2. Folders (hierarchy) are the high-level organisation within an EHR, specifying the care provided to a single subject of care, for a single condition, by a single clinical team (episode of care per speciality).
3. Compositions are a set of information produced from a single episode of care (clinical care session documents/laboratory test results, etc.).
4. Sections are the data within a composition that belong under one clinical heading, reflecting the workflow and consultations processes (Family history, Subjective symptoms, Objective findings, Analysis, Plan, Treatment, etc.).
5. Entries are the information recorded due to a single action/observation/interpretation (clinical statement).
6. Clusters are the ways to organise the data in structures, such as time series and tables.
7. Elements are the nodes of the EHR hierarchy; in other words, single values for one instance.

2.3.2 ELECTRONIC SERVICES

A subset known as Electronic Services within the different Health Information systems mainly focus on improving the internal processes and supporting health professionals.

Besides the electronic services related to administrative functions, billing and financial systems, two services to support clinicians and optimise the care process, which is quite visible in the context of this dissertation, are Computerized Physician Order Entry (CPOE) and Clinical Decision Support System (CDSS).

The CPOE is a computer application integrated into a clinical information system. Its primary function consists in provide an electronic order communication by [31]:

1. Accepting the provider's orders for services.
2. Transmitting the order to the appropriate location.
3. Returning the status of the order.
4. Returning results of order execution.

For example, in the dataset further ahead studied, some attributes specify if a particular exam was requested or not, the date of the request and its results. These attributes are a part of the CPOE application, which, when inserted/requested, are automatically communicated to the endpoint (laboratory or exam centre).

The CDSS is also a computer application that uses pre-established rules and guidelines and integrates clinical data from several sources, generating alerts and treatments suggestions. It assists physicians by matching patient-specific information to a clinical knowledge base (facts, best practices, guidelines, logical rules, and reference information.), providing real-time feedback to support decisions.

2.3.3 HEALTH INFORMATION SECURITY

Over the past decades, information security research has become a well-established area within the information systems discipline. Security is the protection of information and data so that only authorised persons or systems can access them, as well as unauthorised persons or systems, cannot read or modify them [32-33].

The main goals of information security are to prevent Confidentiality, Integrity, and Availability of private health information.

Confidentiality stand by the protection of the information from unauthorised access, use and disclosure. On the other hand, integrity is the protection of information from modification or deletion. Finally, availability protects the information and information systems from disruption or destruction and assures the capability of access from the authorised entities at the time needed [34-35].

The General Data Protection Regulation (GDPR) is a European law regulation on privacy and protection of personal data, applicable in the European Union area, integrating work of the EU's Data

Protection Directive (DPD), the US's Health Insurance Portability and Accountability Act (HIPAA) and various other data protection regimes [36] [2].

Within any organisation, in this case, in a hospital, clinic or health entity, the person who is entitled to assure the correct use of the data protection rules is the Data Protection Officer (DPO), commonly coordinated by the Chief Information Officer (CIO) [37].

Besides the laws, several standards intend to develop organisational security controls and effective security management practices. ISO standards are a starting point, although they do not contain general information on how security measures should be implemented or maintained [38][3].

From IT standards perspective, ISO 27002- 2013 Information technology — Security techniques- Code of practice for information security controls **s**tandard requires that organisations have [39]:

1. Documented information security policy.
2. Allocation of information security responsibilities within the organisation.
3. Information security training for the users.
4. Security incident reporting and response.
5. Virus detection and prevention protocols.
6. Business continuity planning.
7. Control of propriety software copying.
8. Protection of personal data (privacy).
9. Periodic compliance review.

[2] Different countries have different laws for data privacy. See reference number 40, "Big data security and privacy in healthcare: A Review".

[3] In the Portuguese case study presented in reference number 38, "The Adoption of IT Security Standards in a Healthcare Environment", the ISO 27002-2005 was the chose standard. A lattes version of this standard was published in 2013, keeping the main requirements for organisations mentioned above.

2.3.4 ACCESS CONTROL LINES

Electronic systems require access control mechanisms to regulate how they are accessed. Several solutions have been proposed to address security and access control concerns, being Role-based access control (RBAC) and Attribute-based access control (ABAC), the most popular models for EHR [40].

In the RBAC, a role is assigned to each user, by which access to objects is granted. RBAC is defined in five components: subjects, roles, objects, operations, and permissions (Figure 2.3.).

The relationship between the components is:

1. Subjects (individuals) are assigned to roles (doctor, nurse, and staff).
2. Roles are associated with permissions that define which operations can be performed over which objects.

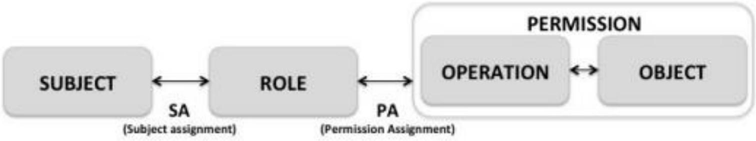


Figure 2-3- The basic components of RBAC. Source: Reference [41].

In the ABAC, each user has access attributed to information objects, which requires experts to assign attributes. ABAC is defined in terms of three components: subjects, environments, and objects (Figure 2.4.).

Subject attributes refer to the individual, environment attributes to the specific department of a hospital or clinical areas, and objects are referent to a single data entry.

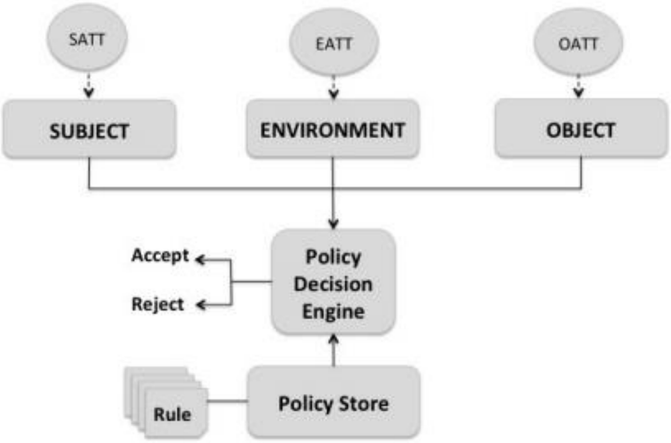


Figure 2-4- The basic components of ABAC. Source: Reference [41].

CHAPTER 3

MACHINE LEARNING AND AI IN HEALTHCARE

This section reviews the necessary knowledge for the practical implementation of the project. Definition of the main topics and associated terms, analysis streams, methodologies, and contextualise the project within these subjects.

Understanding the task is crucial to developing and reporting any research, so guidelines are defined based on the study's content and objectives.

3.1 STREAMS OF ANALYTICS AND REASONING

Big Data, between other characteristics, defines a dataset that is too big to be stored in a conventional relational database system. In other words, it is data whose scale, diversity and complexity require new architectures, techniques, and analysis in order to extract value from it.

In this study, we could analyse large quantities of data that wouldn´t be possible to analyze in any other way besides with AI and big data techniques. More than the quantity of data, the quality of data is a much critical point.

When it comes to velocity, most of the data generated in real-time needs real-time analysis, the major challenge in this area is the ability to process and analyse vast volumes of data in real-time streams [4].

The dataset used in this dissertation is compound by data collected before the study, so there are no concerns related to the real-time analysis for now. This topic is quite significant, as, from a long-term perspective, all types of follow-up will be performed with a significant digitally component, which will require real-time data analysis.

Every project that requires real-time analysis starts from a static dataset and its analysis, which is where this dissertation is inserted.

Data needs to be processed to be valuable, and the classifications of the analysis used are based on the type of output produced. In other words, the classifications are based on the desired results.

The four categories below classify the different streams of analysis, Figure 3.1.:

1. **Descriptive Analytics** uses techniques such as data aggregation and data mining to provide historical understanding. It brings insight to the past, focusing on what happened.
2. **Diagnostic Analytics** is a form of analysis that examines the data to answer why something happened.

3. **Predictive Analytics** allows us to understand the future and predict the likelihood of a future outcome. It uses regression analysis, multivariate statistics, data mining, pattern matching, predictive modelling and machine learning techniques.
4. **Prescriptive Analytics** strives to make decisions for optimal outcomes. In other words, it attempts to quantify the effect of future decisions to advise on possible outcomes before decisions are made.

   The dissertation's practical development includes, first, a descriptive analysis, following by the predictive analysis.

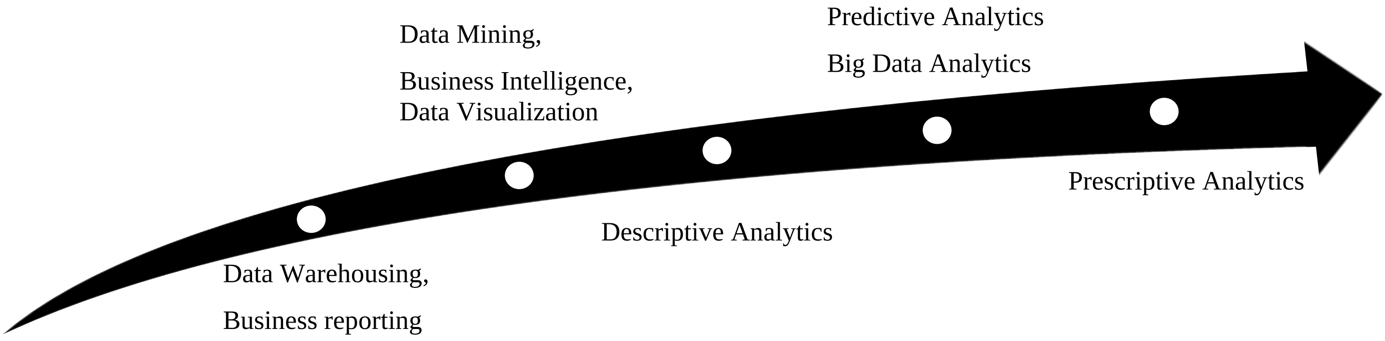


Figure 3-1- Streams of Analytics. Source: Reference [4]

There are different reasoning approaches that a system can use. The three main methods from which the systems can take conclusions are deduction, induction and abduction [4].

The **deduction** is the approach most use in reporting systems that allows the system to determine a true statement. In deduction reasoning, one statement infers a proposition q, which is logically from a permise p.

On the other hand, **induction** reasoning enables us to make statements based on evidence gathered until that moment.

The critical point to understanding these two reasoning methods is that evidence is not the same as fact. Consequently, statements that are determined by the induction method are likely to be true rather than absolute.

By that, it turns clear that Statistical learning is all about inductive reasoning (considering some data, guessing a general hypothesis and making statements or predictions on test data based on these premises).

Finally, **abduction** is an adaptation of induction reasoning. It attempts to use a hypothesis p to explain a proposition q. It is precisely the opposite direction flow of deductive reasoning. The best hypothesis (most effectively explains the data) is inferred to be the most probably correct one.

3.2 COMPONENTS OF AI AND MACHINE LEARNING

Traditional software engineering and machine learning have the common objective of solving a problem or a set of them. However, it is the approaches that distinguish them. Traditional software engineering refers to the development of a function or program, such as that when giving an input, it provides an output. Machine learning differs as, rather than providing instruction about the function, it is provided an input and an output, and it is expected to determine or predict the function.

Getting a system to reason, getting a program to learn, discover and predict, getting a program to communicate with humans, getting a program to have planning and vision capabilities are some of the abilities that an Artificial Intelligence system can comprise.

The core concepts of AI include agents developing traits including knowledge, reasoning, problem-solving, perception, learning, planning and the ability to manipulate and move, Figure 3.2.

Machine Learning can be understood as an AI application [4]. As all AI tasks use some form of data, Data Science or Knowledge Discovery in Databases are terms for the range of techniques used to extract information from the data.

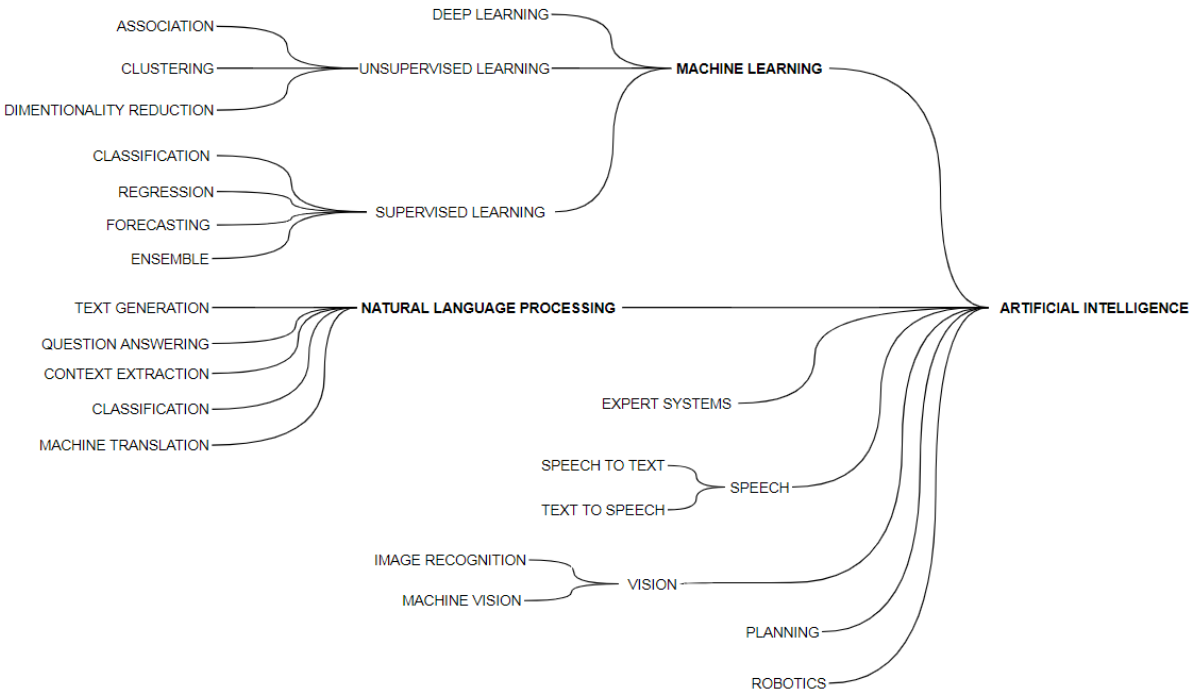


Figure 3-2- AI components.Adapted from reference [4] and [42].

Machine Learning is the collection of algorithms and techniques that can uncover hidden patterns, unknown correlations, trends, preferences, and other information to give meaning and value to the data.

Machine learning algorithms can be categorised as the following:

1. Supervised learning – refers to the process of learning a model from labeled data.
2. Unsupervised learning – refers to the process of learning a model from unlabeled data.
3. Semi-supervised learning – combine both labelled and unlabeled data to generate a function.
4. Reinforcement learning - where the algorithm learns how to act given an observation.

5. Deep-Learning – processes that use deep-layered neural networks architectures. Neural networks can learn how to accomplish a task like a human brain – supervised, unsupervised, and reinforcement learning.

Supervised learning will be the focus from now on, as the regression models used in this dissertation are a part of this category.

However, before focusing on this subject is necessary to review the techniques and methodologies for data preparation and data mining, which is prior and critical work to be developed before starting the model building.

3.3 DATA SCIENCE LOOP

This chapter emphasises the practical techniques and methodologies for data preparation in data-mining applications, and Figure 3.3. presents the six phases of the data science loop [4,10,13].

1. **Data cleaning** – Raw data requires special attention in this stage. Raw data is not treated or processed yet, which is necessary to create value, perform any analysis, and model building. Data cleaning and preparation take between 60% to 80% of the total data engineering effort in practice. It is the process of detecting and correcting or removing incorrect data entries (Data transformation and normalisation), such as missing values, outliers, inaccurate values, and duplicates elimination.
2. **Analyse & Sample** – When analysing the data, two important questions should be asked: Is the data valuable to answer the problem? How is the data distributed?
   In this stage, the data is analysed to determine which information will be the most useful to the machine learning model. In some cases, it might be necessary to sample the data to ensure that the target variable is significantly represented in the data. It can be done through stratified sampling — a method that consists of splitting the data into groups based on the target feature and sampling independently from each group.
3. **Features engineering** – This step consists of a relevance analysis to select the valuable features.
   Based on the analysis performed, the data scientist selects the measurable attributes from the underlying data to be included in the machine learning model. Applying transformations to improve the data is one of the most important and time-consuming parts of developing a machine learning model, for example, creating new features by combining other features.
4. **Model Building –** In this phase, data scientists have to choose the best modelling technique to solve the problem in question. In other words, select the machine learning algorithm(s) and build the model(s).
5. **Hyperparameter Optimization –** In this stage, the objective is to choose the best model parameters. Machine learning models and training algorithms have many settings that can impact how well the resulting model will perform. These settings are called hyperparameters.
6. **Evaluate & Compare -** For every trained model, the data scientists extract performance metrics. Evaluate models by comparing their predictions, pick the best model based on performance and other desired metrics, by looking into how different features and hyperparameter configurations influence the metrics; data scientists can devise new configurations that promise superior performance.

This framework is a loop, which implies passing through each phase many times as necessary. It is continuous work.
The perfect model does not exist. The data scientist and the groups involved in the ML project have to get to a consensus and realise when the goals are met. Otherwise, there will always be space for improvements and modifications.

3.4 SURVIVAL ANALYSIS METHODS

Survival analysis is the area of statistics that include the models and methods developed to analyse survival data.

Survival times are defined as data that measures the time from a defined starting point until the occurrence of a given event of interest.

While binary classifiers typically ignore observations where the event did not happen, one of the main objectives in survival analysis is to account for them. These cases are named 'censored' observations; in other words, they are the cases where the event has not yet occurred or is not known to have occurred [43-45].

For this dataset, the event of interest is the patient's death, so the censored data are those where the outcome is survived or unknown.

Within the Survival Analysis methods, there are Non-parametric, Semi-parametric and Parametric methods, as illustrated in Figure 3.4.

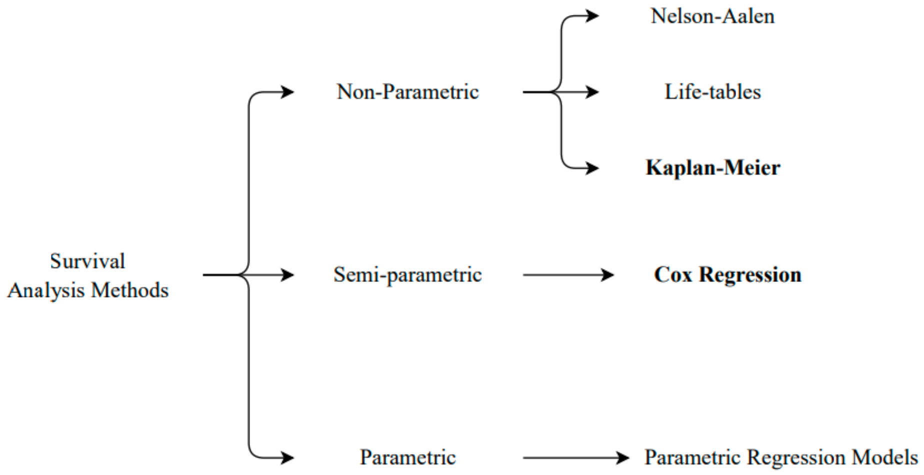


Figure 3-3 - Survival Analysis Methods.

As not all death events have occurred, censoring is a characteristic of survival analysis that does not exclude the patients in this situation.

There are four types of censoring [45]:

- Right censoring – The individuals in a population who have not been subject to the death event are labelled as right-censored. Even though we do not have data further ahead of that time, we can not assume that the event has occurred.
- Left censoring - An individual's lifetime is considered left censored if we only know it is minor than a specific registered time.
- Interval censoring – When it is impossible to observe the exact moment when the event of interest occurs, but we know that it has occurred during some period, it is called interval censoring.
- Double censoring- When both right and left censoring are used, it is called double censoring.

In this dissertation, only right censoring was used in the models developed.

3.4.1 KAPLAN-MEIER METHOD

The Kaplan-Meier method is a univariate model, which means that only one variable is considered in the model. Since it does not consider a parametric distribution for the event variable, it is categorised as a non-Parametric statistical method.

The survival function S(t) is defined as the probability of surviving, at least, up to time t. In this empirical survival function, there are no censoring events. Therefore, the survival function in a specific time t is estimated by the proportion of individuals that survived ahead of the instant t, in a n dimension sample, given by equation (3.1.):

$$S(t) = \frac{Number\ of\ observations > t}{n}, t \geq 0 \tag{3.1}$$

The Kaplan-Meier method proposes a non-parametric estimator of this survival function when there are censored events.

The method is based on the basic idea that the probability of surviving k or more periods after entering the study is the product of the k observed survival rates for each period, as presented below in equation (3.2.) (i.e. the cumulative proportion surviving):

$$\mathrm{S(k)} = p_1 \times p_2 \times p_3 \times \ldots \times p_k , k \in \mathbb{N} \tag{3.2}$$

Here, p_1 is the proportion of individuals surviving the first period, p_2 is the proportion of individuals surviving beyond the second period conditional on having survived up to the second period, and so on. Below, equation (3.3.), the proportion of individuals surviving period i, having survived up to period i:

$$p_i = \frac{n_i - d_i}{n_i} \tag{3.3}$$

where n_i is the number of individuals alive at the beginning of period and d_i the number of deaths within the period.

The Kaplan-Meier estimator of the survival function is then given by equation (3.4.):

$$S(t) = \prod_{i:t_{(i)} \leq t} \frac{n_i - d_i}{n_i} \tag{3.4}$$

where n_i is the number of subjects at risk of death just prior to time t, and d_i are the number of death events at time t.

As one variable can assume different values (e.g., the patient can receive treatment 1 or 2), we may face two different groups of patients. The long-rank test is used to test whether there is a difference between the survival times of different groups, but it does not allow other explanatory variables to be taken into account.

It is used to test the null hypothesis that there is no difference between the population survival curves and the statistic test is given by equation (3.5.):

$$x^2(\log rank) = \frac{O_1 - E_1}{E_1} + \frac{O_2 - E_2}{E_2} \tag{3.5}$$

where the O_1 and O_2 are the total number of observed events in groups 1 and 2, respectively, and $E_1$and E_2 the total number of expected events.

3.4.2 COX REGRESSION

Regression-based systems are one of the main approaches within supervised learning [4], and as previously mentioned, the learning method of these models is inductive.

The algorithm maps the inputs, in other words, represents the input data x within a particular domain. However, x typically represents multiple data points (x1, x2, x3,…, xn). Thus, the goal is to tune a predictor function f(x).

Each record/instance in the dataset is a vector of features x(i). There will also be a target label for each instance in supervised learning, y(i). Therefore, the model is trained with inputs in the form {x(i),y(i)} [4].

Cox’s proportional hazard model is a multiple regression model that enables the difference between survival times of specific groups of patients while allowing for other factors. In other words, it is a multivariate model that involves time and censorship features and additional data as covariates, describing relationships between survival distribution and the covariates.

The Hazard is the probability of dying (experience the event) given that patients have survived up to a given point in time.

Bellow, equations (3.6.) and (3.7.) present the model’s formulation:

$$\ln h(t) = \ln h_0(t) + b_1 x_1 + \cdots + b_p\, x_p \tag{3.6}$$

or

$$\ln \frac{h(t)}{h_0(t)} = b_1 x_1 + \cdots + b_p x_p \tag{3.7}$$

where h(t) is the Hazard at time t; x_1, x_2, …, x_p are the explanatory variables; and $h_0(t)$ is the baseline hazard when all continuous explanatory variables are zero. The coefficients b_1, …, b_p are estimated from data, in other terms, are a unique scaling factor.

Likewise, the Hazard can also be presented as in equation (3.8.), the multiplication between the baseline and the partial hazards.

$$h(t) = h_0(t) \times e^{b_1 x_1 + \cdots + b_p x_p} \tag{3.8}$$

As the Hazard Ratio (HR) does not depend on time comes the proportional hazard assumption.

As mentioned, the proportional hazard assumption is that all individuals have the same hazard function but a unique scaling factor in front. So the shape of the hazard function, equation (3.9.), is the same for all individuals and only a scalar multiple changes per individual.

$$h_i(t) = b_i h(t) \tag{3.9}$$

At the core of the assumption is that b_i is not time-varying, that is, $b_i(t) = b_i$ which leads to equation (3.10.):

$$\frac{h_i(t)}{h_j(t)} = \frac{b_i h(t)}{b_j h(t)} = \frac{b_i}{b_j} \tag{3.10}$$

The baseline profile is the most common patient's profile, and it is the profile to which all variables will be compared, and therefore the risk will be calculated based on that comparison. This baseline always has the HR= 1.

As mentioned, all the other covariates will have an HR that will have values depending on its behaviour. The values of the Hazard Ratio per covariate can be interpreted as explained below, all the remaining covariates equal to the baseline profile defined:

- If HR = 1, it means that there is no significant difference in risk when comparing these variables with the baseline profile.
- If HR<1, it means there is a reduction in the Hazard. In other words, this variable represents a risk reduction comparing with the baseline profile defined.
- HR>1, it means there is an increase in the Hazard. In other words, this variable represents a risk increasing comparing with the baseline profile defined.

As the Hazard measures the instantaneous risk of death, sometimes it can be challenging to illustrate from sample data. Instead, it is commonly used the cumulative hazard function H(t). This function is obtained from the cumulative survival function S(t) as:

$$H(t) = -\ln S(t) \tag{3.11}$$

Further ahead, a practical example of the advantages of this function will be given, mainly when the survival hazards of two groups are close to each other.

CHAPTER 4

DATA ENGINEERING

Data drives learning, so starting from there, the following chapter approaches the tools used in the development, the dataset used in the project, going from its sources and formats to all the pre-processing work.

Likewise, understanding its attributes and content, the descriptive analysis and review was essential to build the models presented in the next chapter.

4.1 DEVELOPMENT TOOLS

Anaconda is an open-source distribution platform that aims to simplify package management, deployment and perform Python/R data science and machine learning on a single machine. Anaconda was the platform used in this dissertation to access all Python libraries.

Python is the programming language used in this dissertation development, and it is increasingly popular within the data science and machine learning industry.

Below is the list of the libraries for Machine Learning used:

- Pandas is a tool to do data aggregation, data manipulation and data visualisation [50]. Within pandas, one-dimensional arrays are referred to as Series, and multidimensional arrays are referred to as DataFrames.
- Matplotlib is a python library for data plotting and visualisation techniques [51].
- Patsy is a Python package describing statistical models and building design matrices (Mainly useful for coding categorical data).
- Lifelines is a complete survival analysis library written in pure Python. Within lifelines, there are several modules[46] :
  - Univariate Models, which include the KaplanMeierFitter
  - Regression Models, which includes the SurvivalRegressions (CoxPHFitter), Time-varying survival regression, Testing the proportional hazard assumptions (CoxPHFitter.check_assumptions method).
- Others – The last version of the data provided was in the Excel format, and to import that format into Python, the Pandas library was enough. Although, before that, a few versions were treated in other formats (For example, SQL, and it was necessary to use a different library that could connect to MySQL -MySQL.connector).

4.2 DATASET DETAILED REVIEW

The sensitivity of this data and the knowledge required to analyse it is pretty unique and too authentic. Hence, it is essential to mention that all the decisions and selection of the attributes were approved and validated by the medical team of Puerta de Hierro-Majadahonda University Hospital (HUPHM).

This subchapter explains the data set's sources, content, and structure and distinguishes the delimitations between what was already defined and what was necessary to define during the pre-processing and implementation of the models.

4.2.1 DATASET SOURCES

The dataset studied was provided by the Medical Oncology Department of HUPHM as coordinators of the project.

Cases were collected from the Spanish Thoracic Tumour Registry, a nationwide registry sponsored by the Spanish Lung Cancer Group. In this registry, more than fifty (50) hospitals collected histologically confirmed lung cancer cases and information from the EHR.

The final version of the dataset contains one-thousand two-hundred and forty-four (1244) attributes with each variable encoding and the respective description, with both idioms, Spanish and English. Also, the dictionary was lately provided organised by attribute groups concerning the nature of the data.

4.2.2 DATASET CONTENT AND STRUCTURE

There were several versions of the dataset, in fact, in different formats. The final format was provided in an Excel file, with each sheet containing a specific data subject.

Tables were created by group subject, and even though the dictionary is organised in the same way, the dataset was not. Before getting into definitions and values details about each attribute used, it is essential to review each table as it was in its original version, and with this, specify how many attributes each one had, mention some of them that could be worthy analysis, and most importantly, the selection and modifications performed.

The **timeline** used in the development is a matter of discussion and an essential point as it would be one of the improvements points in the study. Five out of ten (5/10) tables are recurrent, meaning that they have different information over time and mirror the patient’s performance throughout the different phases of treatment and the different treatments themselves.

Besides the attribute itself, many attributes had other text fields to specify the value ‘other’, or even just complementary information about the value. These text fields were not processed, which is one of the discussion points and optimizations that could add value to the data and the model’s performance.

Also, the attribute's value could be unknown, which was only considered in the descriptive analysis.

Unknow values were discarded as they cannot be considered as attributes of patients, given that the information is missing.

4.3 DATASET PRE-PROCESSING

The studied dataset was originally raw, which required a considerable effort concerning data cleaning, transformation, and normalisation to be then possible to analyse and build the models. This subchapter reviews every step of this stage, specifying all the modifications performed from the raw data until the dataset used in the models, as well as the features selection and clinical meaning of the variables.

Below is a description of each table's general content, along with considerations concerning the attributes used and not used in the models, as well as the modifications performed.

Besides the tables presented below, the dataset also had a table regarding the Clinical Trials, which was not used in the dissertation.

The patient's ID (EHR number) was set as the index of the dataset, as it is a unique number.

The demographics table 4.1. contains all the demographics information of the patient. This table had six (6) attributes from each three (3) were used.

The 'type of patient' attribute was filtered to work only with the non-small cell lung cancer patients. This first filter reduced the original 15.337 patients to 12.981 patients.

The date of birth was used to calculate the age of diagnosis of the patient, along with other date attributes mentioned ahead (calculations are specified in the description of the follow-up table.).

In cases where the date of birth was incomplete (e.g., Unknown day or month), it was replaced with day 15th (middle day of the month) and June for the month (middle month of the year).

Finally, the gender variable which did not suffer any modification.

Table 4-1 Demographic attributes

	Variable	Values
Demographics	**Type Patient**	0, Non-small cell lung cancer 1, Small cell lung cancer 2, Carcinoid tumour 3, Epithelial thymic tumour 4, Mesothelioma 5, Others -, -1
	Specify Type Patient	text
	Date of birth	date
	Gender	0, Male 1, Female
	Race	0, Caucasian 1, Latin 2, Asian 3, African 4, Others -1,-
	Specify race	text

The **smoking habit** ,table 4.2., contains all the attributes related to the smoking history of the patient. This table contained eight (8) attributes from each; only one (1) was used.

The remaining attributes were quantitative specifications such as the number of cigars per day and per year, the number of years active, whether the patient lived with smokers, and how many years.

The attribute 'tobacco history' had four (4) possible values; from each, the 'unknown' was the only one not used in the model (but considered in the descriptive analysis).

Table 4-2 Smoking habit attribute.

	Variable	Values
Smoking Habit	Tabaco history	0, Never smoker (<= 100cigars/lifetime) 1, Former smoker (>= 1 year) 2, Current smoker 3, Unknown

The **diagnosis,** table 4.3., had one hundred and twenty-two (122) attributes from which eighteen (18) were used. The original values can be found in Section A, Attachment 1.

Both dates, data of the first consultation and data of initial diagnosis, were also used to calculate the age at diagnosis [4].

Staging describes cancer based on its size, location, spread, and involvement of other organs. Knowing the stage of the cancer is essential to decide the best treatment and to be able to assess the prognosis of the disease.

The different stages of lung cancer are based on the size, location, and involvement of lymph nodes and other organs, and can be summarized in 5 different categories:

- Stage 0: The tumour only exists at the microscopic level and can be removed by micro-surgery.
- Stage I: The tumour is in an early stage.
- Stage II: The tumour is in the initial phase.
- Stage III: The disease is locally advanced.
- Stage IV: The disease is in the metastatic phase.

[4] Disease Mapping Improvement: The four groups of stages are the general diagnosis groups, but the treatments are specific concerning these 16 different diagnoses. If instead of the general groups of diagnosis that were used in these models, it would be used the more specific diagnosis groups, we would have another hand of patterns along with the treatment lines performed.

The ‘Stage at diagnosis’ variable had originally twenty (20) possible values, from which four (4) were removed, as it did not apply to the study groups, such as Limited and Extended (it does not concern the non-small cell lung cancer), and the value ‘other’.

From the remaining sixteen (16) possible values, four (4) groups were created, each one concerning each stage. These 16 initial values are even more specific groups within each stage, again concerning the location and size of the tumours.

Histopathology diagnoses and studies the tissue’s diseases and involves examining tissues or cells under a microscope. A variety of imaging and biochemical techniques allow characterizing tissue and the presence or absence of specific biomarkers.

Histology variable had originally fourteen (14) values, from which seven (7) were removed as they were extremely rare or not related to lung cancer (the database also contained "thoracic tumours", which are not specific for lung cancer). The remaining seven (7) values (the ones used in the models) can be seen below in Table 4.3.

Another essential diagnosis variable is the **molecular markers**. The molecular markers tests are obtained through tissue biopsy and are therefore invasive techniques, but extremely necessary to choose the most appropriate treatment. Therefore, it is necessary to have shreds of evidence that point in that direction, for instance, patients who never smoked or patients with a family history of cancer.

For patients whose tumours contain specific mutations in the epidermal growth factor receptor (EGFR) or anaplastic lymphoma kinase (ALK) gene, changes determined by molecular testing using a tumour biopsy also impact the treatment choices. Patients with these genes are more likely to be successful than patients without them, as the origin is known, and treatments are better mapped. These 2 mutations are the most frequent in lung cancer.

EGFR and ALK are the two mutations studied, even that the dataset contained others. The first variable concerning the mutations answers whether the molecular markers analyses were performed at diagnoses or not. If not, the value of the mutation in question would be modified to have the value ‘-1’, which was not in the original dataset.

Table 4-3 Diagnosis attributes.

	Variable	Values
Diagnosis	Date of the first consultation (due to disease symptoms before the first consultation in Oncology)	Date
	Date of initial diagnoses Date of diagnoses (Anatomopathological diagnoses)	Date
	Stage at diagnosis	1, Stage I 2, Stage II 3, Stage III 4, Stage IV
	Histology	0, Adenocarcinoma 1, Adenosquamous 3, Large cell carcinoma 7, Neuroendocrine large cell carcinoma 2, Squamous 5, Undifferentiated 4, Sarcomatoid
	EGFR performed	0, No 1, Yes
	ALK performed	0, No 1, Yes
	EGFR result	0, No 1, Yes -1, Not tested
	ALK result	0, No 1, Yes -1, Not tested

The EGFR mutation was reduced to one (1) variable, which can assume three values (‘Yes’, ‘No’ or ‘Not tested’), but initially was distributed in seven (7) different variables.

The ALK mutation was also reduced to one (1) variable, assuming the three same values of EGFR, but originally was distributed into three (3) different variables.

By adding the ‘Not tested’ value to the EGFR and ALK result variables, it was possible to discard the two variables (EGFR and ALK performed) that have the same meaning as the test was not performed.

One of the most critical factors impacting lung cancer patient´s survival is **comorbidities** [52]. Lung cancer is associated with age and smoking, and both age and smoking are strongly associated with comorbidity. Comorbidity, such as cardiovascular diseases, pulmonary and other systems, may influence prognosis in lung cancer and complicate its treatment. With lung cancer being far more frequent in smokers and former smokers, these patients often have tobacco-related illnesses, mainly cardiovascular (ischaemic or hypertensive heart disease, lower limbs arteriopathy, etc.) and respiratory (chronic obstructive pulmonary disease (COPD), obstructive sleep apnea, usual interstitial fibrosis. etc.) in nature. They can also have other comorbidities unrelated to tobacco use but frequent in the general population, e.g. diabetes and its complications (renal insufficiency, cardiovascular damage). These comorbidities can alter the patient’s performance status often more than the tumour development [53].

Lung cancer is also more frequent in elderly patients because ageing is a risk factor for developing lung cancer. Of course, comorbidities are more frequent with ageing and more severe, independent of the physiological alterations inherent to ageing. All these comorbidities can have deleterious effects on the diagnostic procedures and the treatment possibilities and thus must be carefully explored [52].

The dataset had twenty-two (22) initially attributes that described each comorbidity and an open text field to specify others (Section A, Attachment 2).

From these 22 attributes, six variables were merged into 3, as the clinical meaning was similar:

- Dyslipidemia and Hypercholesterolemia
- Liver disease and Hepatitis
- Vascular disease and Cardiopathy

A descriptive analysis was performed considering each one individually, but groups were created concerning the patient's number of comorbidities for the model's development, as presented in table 4.4.

Table 4-4 - Comorbidities attribute values.

	Variable	Values
Comorbidities	Comorbidities	0, No comorbidities 1, 1-3 comorbidities 2, 4-9 comorbidities

Personal and family history of cancer table had sixty (60) attributes, from which two (2) were used (Table 4.5).

Regarding the personal history, the dataset contained information about whether the patient had or not previous tumours and the type of tumour. To consult descriptive analysis regarding the previous tumour type, consult Section A, Attachment 2.

Concerning the family history, if the patient had a family history of cancer, the number of family members with lung cancer/other types of cancer, and specifications about the family member. This table had many attributes as there were fields to input the specifications of several family members.

Table 4-5 - Personal and family cancer history attributes.

	Variable	Values
Personal and family cancer history	Previous cancers	1, Yes 0, No -1, Unknown
	Cancer in family members	1, Yes 0, No -1, Unknown

During treatment, the patient can receive several **treatment lines**. The change of line is performed when the patient does not respond anymore to the current treatment.

The following table specifies the different treatment lines, including the type of therapy, start and end date, the regimen (monotherapy or combination), the drugs, the response of the treatments, the number of cycles, etc.

This table is recurrent; in other words, the same attributes are repeated over time. Each treatment line has its values, but the way of characterising them is the same.

Each treatment line has twelve (12) attributes, times the thirteen (13) lines, which is extremely rare, but the maximum treatment lines that a patient can have led to a total of 156 attributes in this table.

Only the first treatment line was analysed in this study, as the table below shows [5], given that the first line is the most specific and effective for the patient. From the fifteen (15) values that this attribute initially had:

- Five of them, precisely, CT intravenous, Neoadjuvant chemotherapy, Adjuvant chemotherapy, Oral and intravenous chemotherapy, Oral chemotherapy, were merged into one, CT.
- Four of them, specifically, Concomitant CT-RT, Sequential CT-RT, Adjuvant CT-RT, Neoadjuvant CT-RT, were merged into one, CT + RT.
- Hormonal, Treatment 5 (not specified) and the ones with no information were removed from the study, as they were not related to lung cancer or not relevant.

It was then used seven (7) values in the model's deployment (Table 4-6). To see all the original values, consult Section 1- Attachment 3.

[5] Disease Mapping Improvement: If all the treatment lines were used in the models, we would have another set of patterns and a much more extensive view of the entire treatment process. As this is a process over time, it would be better implemented in a model using neural networks, which would enable the creation of layers per specific treatment sequence. Otherwise, using all the treatment lines in a Cox model, and considering their timelines, it would be necessary to create a variable with all the possible sequences, which would be exhausting, and more than that, the accuracy of the model would decrease, as the baseline for this variable would be questionable (considering the considerable number of possible treatment line sequences).

Table 4-6- First treatment line attribute- Type of therapy.

	Variable (English)	Values (English)
Treatment Line 1	Type of therapy (Drug therapy)	1, CT (Chemotherapy) 2, TKI (Oral targeted therapy) 3, CT-RT (Intravenous chemotherapy + radiotherapy) 4, IO (Immunotherapy) 5, CT + IO (Intravenous chemotherapy + Immunotherapy) 6, Others 7, No drug therapy

The **Radiotherapy Treatment** is also a recurrent table. Each line of radiotherapy had ten (10) attributes, times the nine (9) lines of treatment (which is the maximum that a patient can have), give us a total of ninety (90) attributes for this entire table.

Each line of radiotherapy had information such as the start and end date of the treatments, intention of the treatment, type of radiation, the total dosage, among others.

Once again, it was only considered the first line of radiotherapy and the variable that answer the question if it was performed or not (Table 4.7.).

Table 4-7 - Radiotherapy attribute.

	Variable (English)	Values (English)
Radiotherapy Treatment 1	Has the patient received any radiotherapy for the thoracic tumours?	1, Yes 0, No -1, Unknown

The **Surgery table** is also recurrent. Each line of surgery had seventeen (17) attributes, times the eight (8) possible surgeries, give us a total of one-hundred and thirty-six (136) attributes. Each surgery had information such as the date, type of surgery, procedure specifications and response. The analysis only considered if the patient had one surgery (Table 4.8.).

Table 4-8 - Surgery attribute.

	Variable (English)	Values (English)
Surgery 1	Has any surgery been performed for the thoracic tumour?	1, Yes 0, No -1, Unknown

The **Progression/Relapse** table 4.9. had 22 attributes times the eleven (11) recurrences, which give 242 attributes in total. It contained information related to the type of progression and the location specification of the progression itself. This data was not used in the models but used for tests.

Table 4-9 - Progression/Relapse attribute.

	Variable (English)	Values (English)
Progression/Relapse 1	Specify Progression/Relapse	1, Progression 2, Relapse 0, No progression -1, No information

The **follow-up** tables were fifteen (15), each one with thirteen (13) attributes (195 attributes in total).

If the follow-up was performed, we had the date of the last contact and/or the follow-up loss date, the current situation of the patient, and if death, the date of death, as presented in table 4.10.

Besides these attributes, if the patient was dead, it also had information about the reason for death, which was not analysed.

Different dates were used to calculate the age at diagnosis, as there were missing values or incoherences in the dates. The calculations below are presented in the actual order used to calculate the age of diagnosis. If a patient did not have the correct data previously used in the calculation, the next one was necessary.

Firstly, was used the date of the initial diagnosis, which is the most accurate date that can be used to calculate the age at diagnosis [4.1]:

$$Age\ at\ diagnosis = Date\ of\ initial\ diagnosis - Date\ of\ birth \tag{4.1}$$

For the ones in which the date of initial diagnosis was missing or incorrect, it was used the date of the 1st consultation [4.2]:

$$Age\ at\ diagnosis = Date\ of\ 1st\ consulation - Date\ of\ birth \tag{4.2}$$

At this point, there were still patients who weren't possible to calculate it, so it was necessary another date, the date of the 1st treatment [4.3]:

$$Age\ at\ diagnosis = Date\ of\ last\ day\ of\ contact(follow - up\ 1) - Date\ of\ birth \tag{4.3}$$

Finally, the last option [4.4], the date of the 1st treatment, which is the less accurate of all these options, as the patient was already diagnosed by that time.

$$Age\ at\ diagnosis = Date\ of\ 1st\ treatment\ line - Date\ of\ birth \tag{4.4}$$

Still, there were patients without the correct diagnosis age and cases where the age of diagnosis was an outlier. In these cases, the patients were removed from the study.

The age of diagnosis variable (numeric variable), was then converted into groups (categorical variable), based on the distribution of the population:

- Group I – Less than 45 years old.
- Group II - Between 45 and 70.
- Group III – More or equal to 70.

The follow-up table was the only which it was necessary to analyse all the fifteen follow-ups as one of the goals was to obtain the last situation registered of the patient.

A single variable date of death was created in order to join in it all the dates which were distributed in the 15 recurrences in the original table (e.g., A patient could have died in the follow-up 4, so the information of the previous recurrent tables 1-3, was not the current situation of the patient).

The **survival months** were calculated in different ways, first for the dead patients, and once again, using different dates for the diagnosis, as there were missing values [4.5-4.8]:

$$Survival\ months = (Date\ of\ death - Date\ of\ initial\ diagnosis)\ /30 \tag{4.5}$$

$$Survival\ months = (Date\ of\ death - \ Date\ of\ 1st\ consultation)/30 \tag{4.6}$$

$$Survival\ months = (Date\ of\ death - \ Date\ of\ last\ day\ of\ contact(follow - up\ 1))/30 \tag{4.7}$$

$$Survival\ months = (Date\ of\ death - Date\ of\ 1st\ treatment\ line)/30 \tag{4.8}$$

And secondly, for the alive patients[4.9-4.11]:

$$Survival\ months = (Date\ of\ the\ last\ contact - Date\ of\ initial\ diagnosis)\ /30 \tag{4.9}$$

$$Survival\ months = (Date\ of\ the\ last\ contact - Date\ of\ 1st\ consulation)\ /30 \tag{4.10}$$

$$Survival\ months = (Date\ of\ the\ last\ contact - Date\ of\ 1st\ treatment\ line)\ /30 \tag{4.11}$$

The survival analysis was performed firstly in days but changed to months as the graphics interpretation was easier and to comply with most of the survival studies.

The date of the last contact was not, in fact, a part of the follow-up table, as it was a generic attribute that could concern every table.

Table 4-10 - Follow-up attributes and date of the last contact.

Follow-up 1	Follow-up	1, Yes
	Date last contact	Date
	Situation	1, Alive, no disease 2, Alive with disease 0, Dead 3, Lost follow-up
	Follow-up loss date	Date
	Date of death	Date
Last contact date	Date of the last contact	Date

4.4 DESCRIPTIVE ANALYSIS AND KAPLAN-MEIER METHOD RESULTS

After the data curation and, at this point, having defined the meaning of each variable and all the possible values that each of them can assume (reviewed in the previous section), it is necessary to understand their behaviour within the entire group.

In this chapter, a distribution diagram for each variable (containing the number of patients per value of the attribute) is presented, along with a table with helpful information to understand its behaviour (e.g., number of patients, median and mean survival, median and mean age of diagnosis, etc.), and finally the results of the Kaplan-Meier method.

A descriptive analysis was performed with 70,36% of the original dataset, still containing unknown variables in several attributes. Finally, the unknown attributes were removed, and the Kaplan-Meier method was performed with the 8 578 patients. Figure 4.1. illustrates the steps just described.

For a more extensive view of all variables and their distribution, consult Section B, Attachment 1.

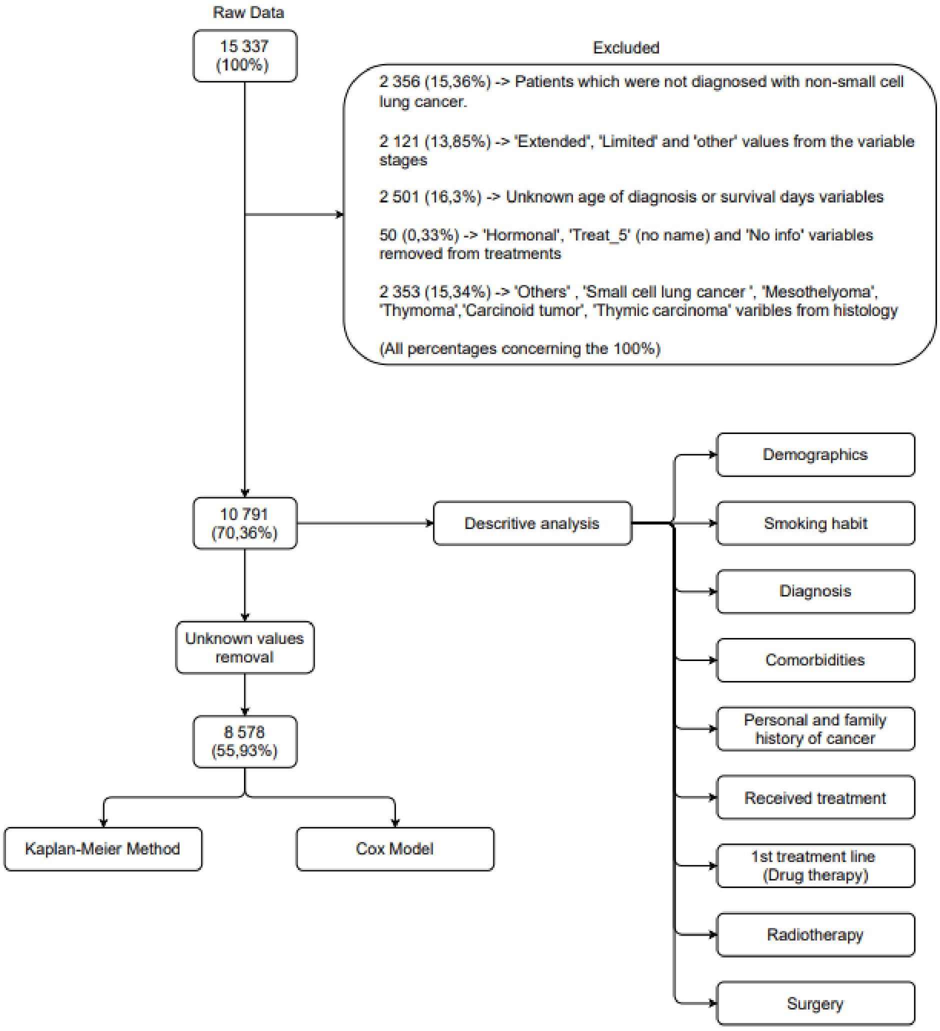

Raw Data
15 337 (100%)
Excluded
2 356 (15,36%) -> Patients which were not diagnosed with non-small cell lung cancer.
2 121 (13,85%) -> 'Extended', 'Limited' and 'other' values from the variable stages
2 501 (16,3%) -> Unknown age of diagnosis or survival days variables
50 (0,33%) -> 'Hormonal', 'Treat_5' (no name) and 'No info' variables removed from treatments
2 353 (15,34%) -> 'Others' , 'Small cell lung cancer ', 'Mesothelyoma', 'Thymoma','Carcinoid tumor', 'Thymic carcinoma' varibles from histology
(All percentages concerning the 100%)
10 791 (70,36%)
Descritive analysis
Demographics
Smoking habit
Diagnosis
Comorbidities
Personal and family history of cancer
Received treatment
1st treatment line (Drug therapy)
Radiotherapy
Surgery
Unknown values removal
8 578 (55,93%)
Kaplan-Meier Method
Cox Model

Figure 4-1 Dataset development during the different phases of the development.

4.4.1 DEMOGRAPHIC DATA

Age at diagnosis and gender were the two demographic variables used in the models.

The diagram in figure 4.2. illustrates both distributions, concerning the 70,36% of the cohort used for the descriptive analysis.

For the variable Age at diagnosis, the group age from 45 until 70 corresponds to 64,43% and based on this percentage, it was the baseline profile defined for the models. In the same line of thought, the males represent 74,4% of the population and will also be a part of the baseline profile.

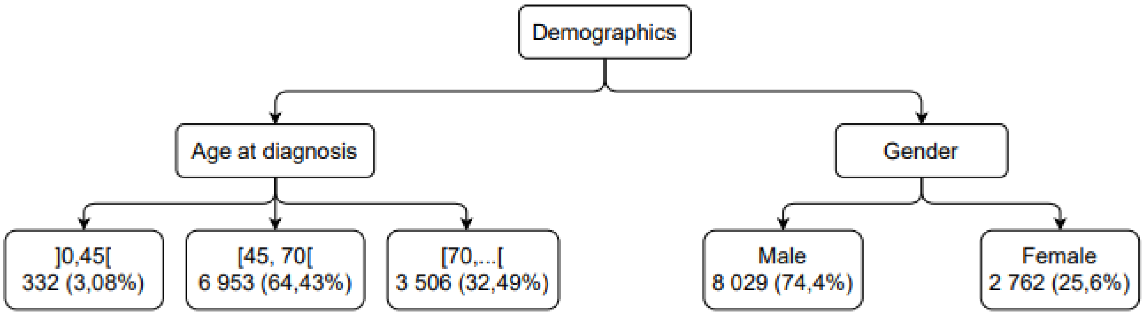


Figure 4-2 - Demographics - Distribution diagram.

In the descriptive table 4.11., the values show that the female's group have higher survival than the males, which is also visible in the KMFcurves in figure 4.3.

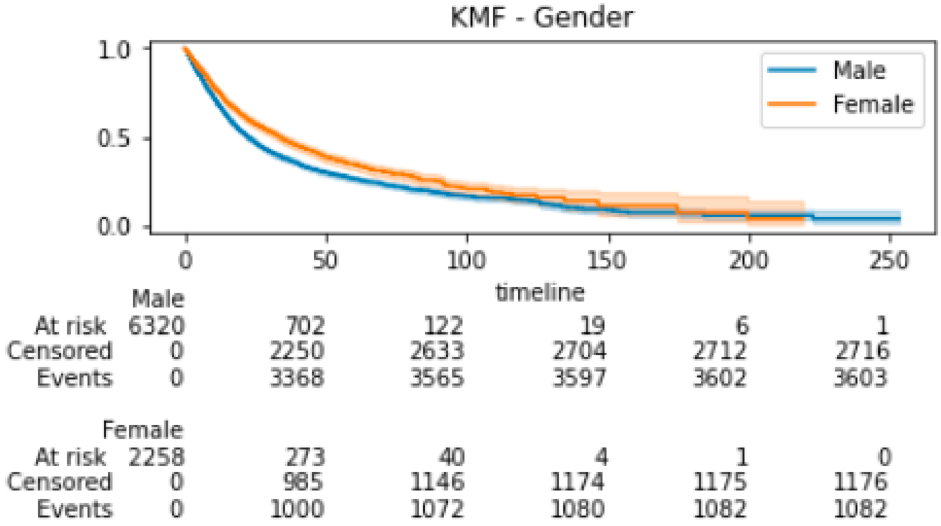
To consult an example of the tables creation code and KMF generation code, consult Section C- Attachment 1 and 2.

Table 4-11 – Gender - Descriptive table.

Gender	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Male	8029	14,1	22,6	66	65,2	859	705	2419	4046
Female	2762	16,6	25,4	62	62,2	272	208	656	1626

In the graphic generated by the Kaplan-Meier method, it is possible to see that the survival lines of both groups cross each other at the end of the time.

This is a flag that it is possible that this variable does not check the proportional hazard assumption when fitting the models. Note that this may be due to the fact that at the end of the studied period, the number of patients is much lower.



Male						
At risk	6320	702	122	19	6	1
Censored	0	2250	2633	2704	2712	2716
Events	0	3368	3565	3597	3602	3603
Female						
At risk	2258	273	40	4	1	0
Censored	0	985	1146	1174	1175	1176
Events	0	1000	1072	1080	1082	1082

Figure 4-3 KMF Gender variable.

The Age at diagnosis variable was a numeric variable at the beginning (created based on the calculations presented in the previous chapter). In order to be converted to categorical, it was necessary to see its behaviour analysing at the same time the survival. Groups were then defined considering the graphic presented in Figure 4.4.

Several range groups were tested, not having significant changes in the risk associated with it or in the model's performance. The final selected Age at diagnosis groups are the ones analysed in the descriptive Table 4.12.

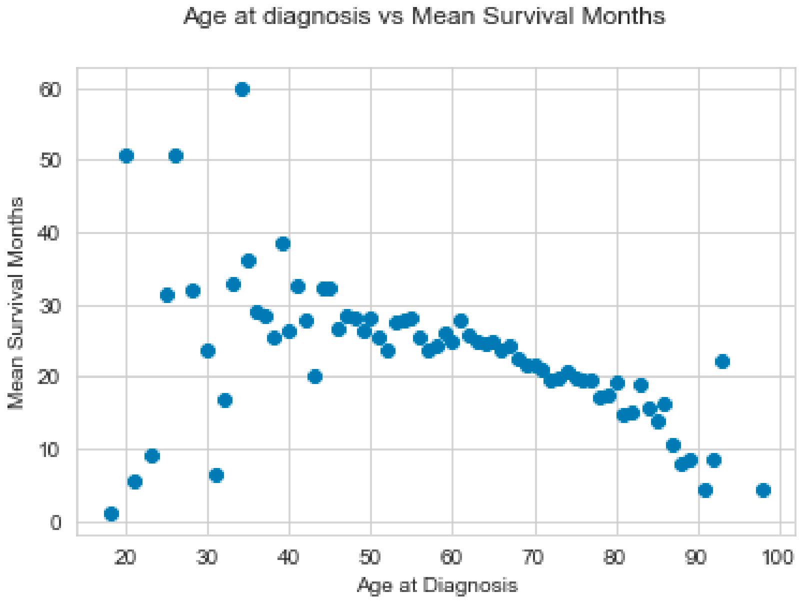


Figure 4-4 - Age at diagnosis vs Mean survival Months.

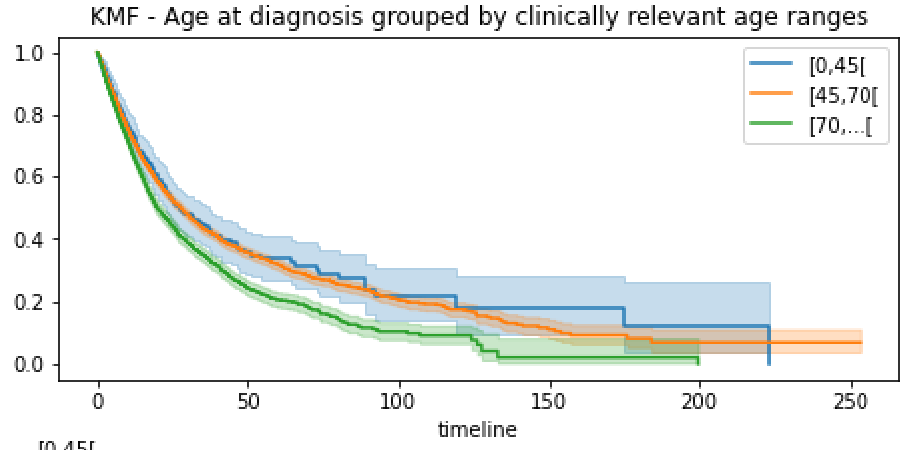
Table 4-12 - Age at diagnosis - Descriptive table.

Age at diagnosis	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
[0,45]	332	18,6	29,0	41	39,8	13	20	80	219
]45,70]	6953	15,6	25,2	61	60,0	694	563	2002	3694
[70,...[	3506	12,8	19,1	75	75,5	424	330	993	1759

It is possible to see that the survival decreases as the age at diagnosis of the patient's increases, which is also mirrored in the KMF curves, Figure 4.5.

The descriptive analysis of the gender variable also stated that males are older at diagnosis than females, which can also influence the survival numbers as the survival reduces with ageing.

A boxplot was generated to display the distribution and see the outliers of the group's defined (Figure 4.6.).



[0,45[						
At risk	278	43	9	3	2	0
Censored	0	91	115	120	120	121
Events	0	144	154	155	156	157
[45,70[						
At risk	5549	742	131	19	5	1
Censored	0	2082	2496	2575	2584	2588
Events	0	2725	2922	2955	2960	2960
[70,...[						
At risk	2751	190	22	1	0	0
Censored	0	1062	1168	1183	1183	1183
Events	0	1499	1561	1567	1568	1568

Figure 4-5 KMF Age at diagnosis grouped by clinically relevant age range.

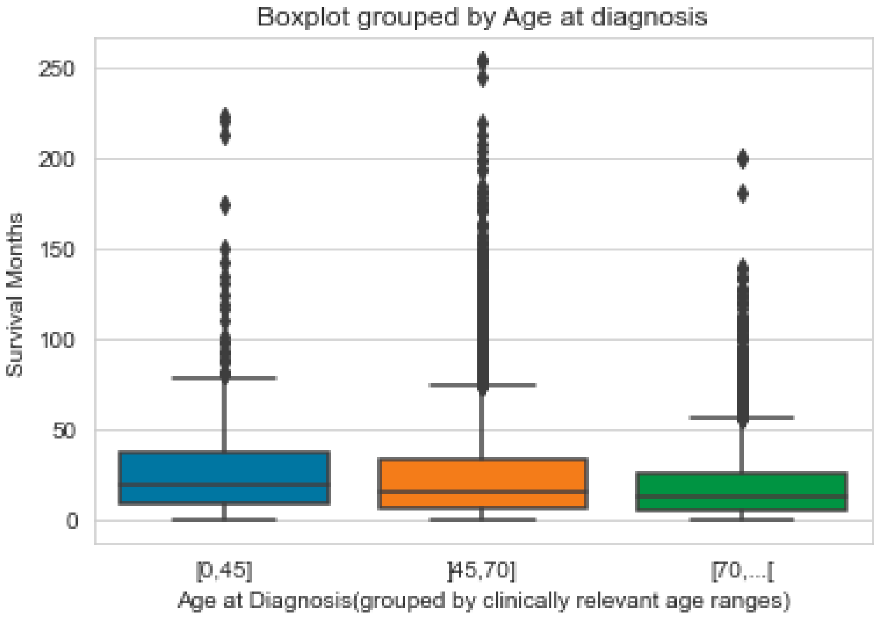


Figure 4-6 Boxplot grouped by Age at diagnosis.

4.4.2 SMOKING HABIT DATA

The diagram in figure 4.7. illustrates the smoking habit distribution, concerning the 70,36% of the cohort used for the descriptive analysis.

For this variable, the former smokers correspond to 47,72% and, based on this percentage, was defined as the baseline profile.

The unknown values were considered in the descriptive analysis but not considered in the KMF method or the models in the next chapter.

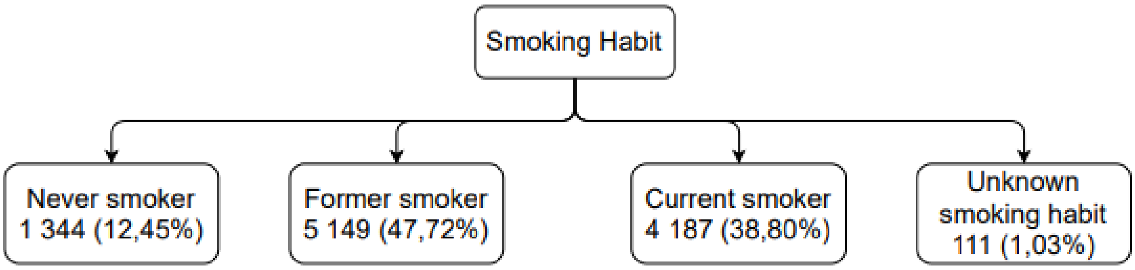

Smoking Habit
Never smoker 1 344 (12,45%)
Former smoker 5 149 (47,72%)
Current smoker 4 187 (38,80%)
Unknown smoking habit 111 (1,03%)

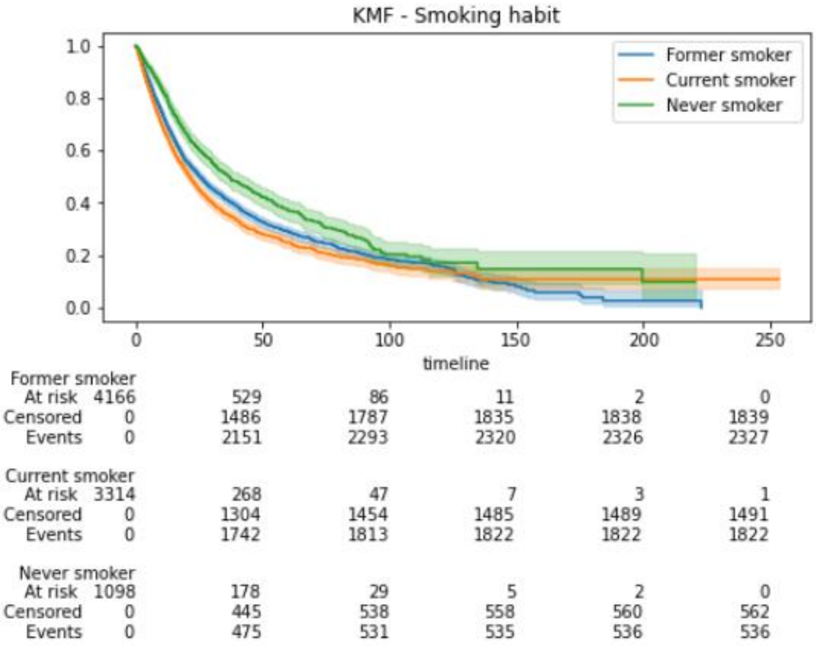
Figure 4-7 Smoking Habit - Distribution diagram.

In Table 4.13. we can see that patients where the smoking habit is unknown, have a median and mean survival vastly different, which indicates this variable's vulnerability.

Table 4-13 Smoking Habit - Descriptive table.

Smokin Habit	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Never smoker	1344	19,5	28,7	68	66,5	141	85	221	897
Former smoker	5149	15,5	24,2	67	66,2	612	511	1581	2445
Current smoker	4187	12,8	20,3	61	61,5	363	307	1256	2261
Unknown smoking habit	111	12,6	31,0	66	66,2	15	10	17	69

The never smoker have much higher survival than the formers or current smokers, also possible to visualise in the KMF curve of the variable, Figure 4.8.



Former smoker						
At risk	4166	529	86	11	2	0
Censored	0	1486	1787	1835	1838	1839
Events	0	2151	2293	2320	2326	2327
Current smoker						
At risk	3314	268	47	7	3	1
Censored	0	1304	1454	1485	1489	1491
Events	0	1742	1813	1822	1822	1822
Never smoker						
At risk	1098	178	29	5	2	0
Censored	0	445	538	558	560	562
Events	0	475	531	535	536	536

Figure 4-8 KMF Smoking Habit.

4.4.3 DIAGNOSIS DATA

The stage at diagnosis, Histology and the Molecular Markers were the diagnosis variables used in the models.

The diagram in figure 4.9. illustrates these variables distributions, concerning the 70,36% of the cohort used for the descriptive analysis.

For the variable Stage, the Stage IV group corresponds to 52,6% and based on this percentage, it was the baseline profile defined for the models. In the same line of thought, the histology Adenocarcinoma represents 60,88% of the population and will also be a part of the baseline profile.

Regarding the Molecular Markers, the baseline was defined as ‘Not tested’ for both ALK and EGFR, with 56,38% and 45,66%, respectively.

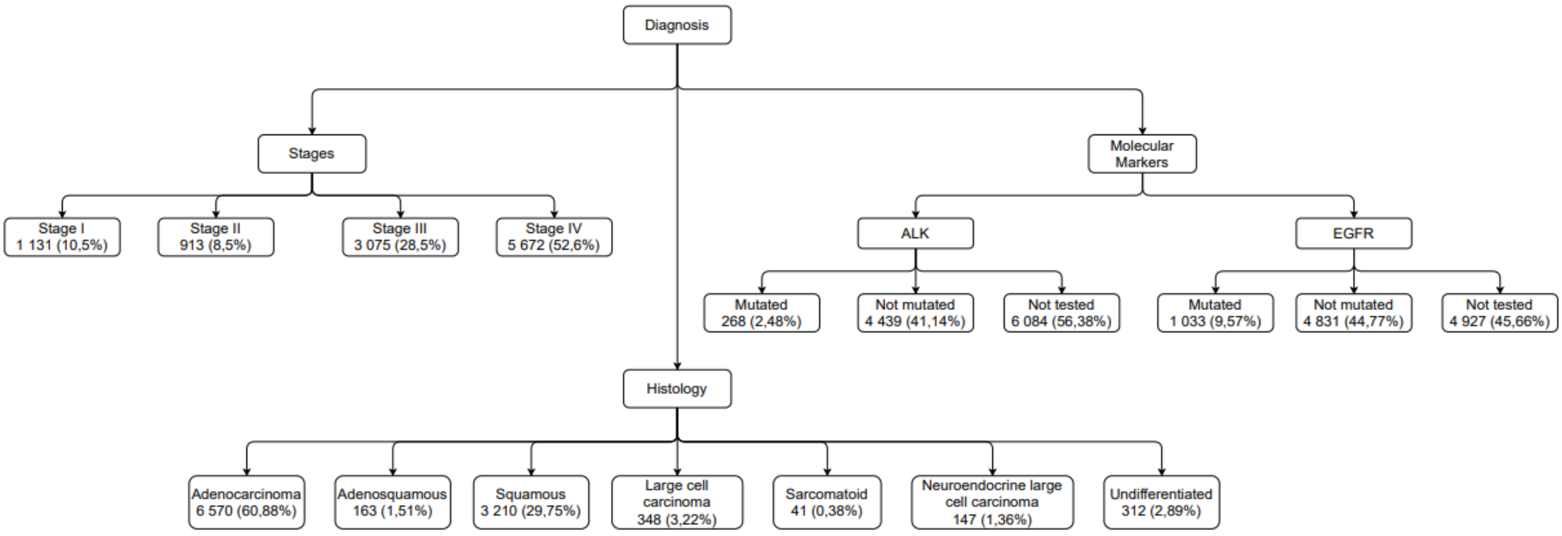


Figure 4-9 Diagnosis - Distribution diagram.

The stage greatly impacts survival at diagnosis. Lung cancer survival remains very low and has hardly improved in recent decades despite important advances in treatments such as immunotherapy and targeted therapies. Even so, over half of the cases are diagnosed in stage IV.

The diagnosis of lung cancer in the early stages remains a challenge and often occurs incidentally in the study of other diseases. The treatment strategy also should take into account factors such as histology and molecular pathology, among others.

The descriptive table 4.14. states the difference in survival depending on the stage at diagnosis, and it is significant the difference between them.

Table 4-14 - Stages - Descriptive table.

Stage	count	median survival	mean survival	median age	mean age	Male	Female
Stage I	1131	33,2	41,0	67	66,3	859	272
Stage II	913	26,8	35,1	66	65,4	705	208
Stage II	3075	17,8	26,4	65	64,5	2419	656
Stage IV	5672	10,5	16,2	64	63,8	4046	1626

A boxplot was generated to display the distribution and see the outlier of the group's defined per stage at diagnosis (Figure 4.10.)

Even looking at the outliers in these groups, it is possible to see a significant decrease in survival for the advanced stages (III and IV).

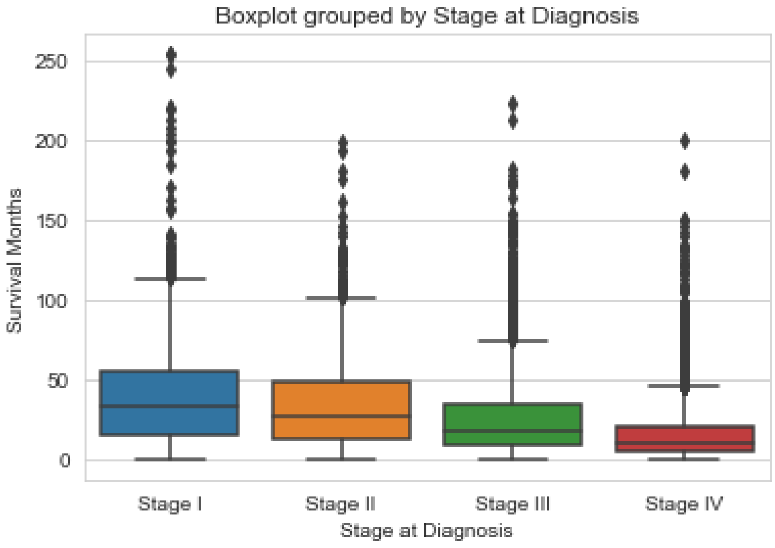
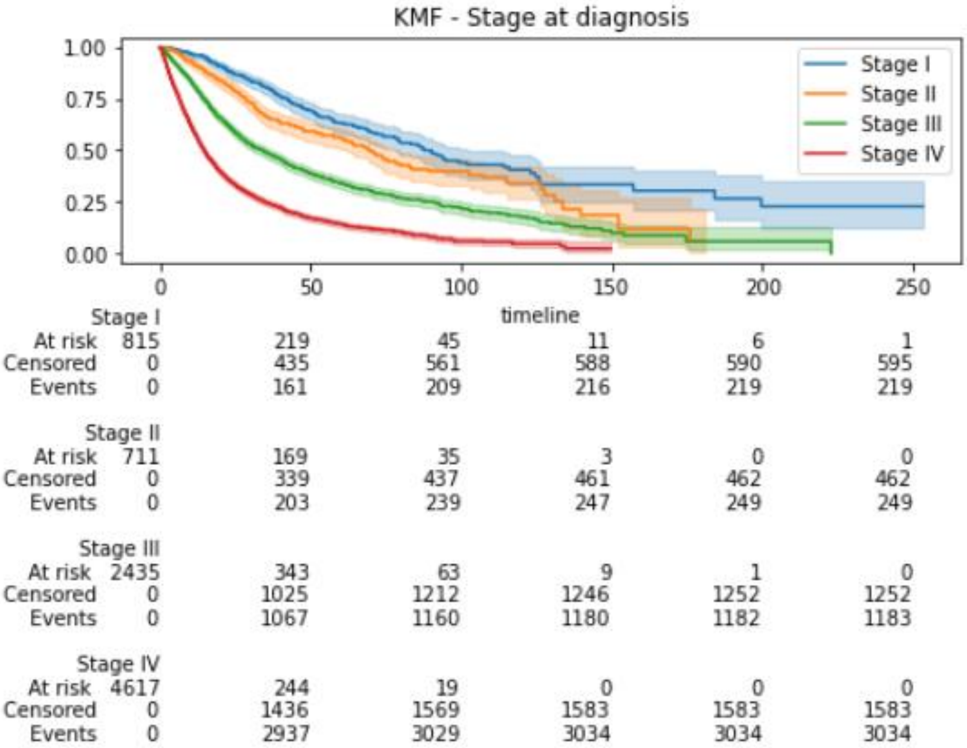


Figure 4-10 Boxplot grouped by Stage at Diagnosis.

In Figure 4.11. the KMF result for the stage at diagnosis variables clearly illustrates the difference in survival of the different groups.

For the stage at diagnosis IV, the survival curve terminates at 150 months due to the low survival of the group, which means that there are no patients alive at that time (even that more than half were censored).



	0	50	100	150	200	250
Stage I						
At risk	815	219	45	11	6	1
Censored	0	435	561	588	590	595
Events	0	161	209	216	219	219
Stage II						
At risk	711	169	35	3	0	0
Censored	0	339	437	461	462	462
Events	0	203	239	247	249	249
Stage III						
At risk	2435	343	63	9	1	0
Censored	0	1025	1212	1246	1252	1252
Events	0	1067	1160	1180	1182	1183
Stage IV						
At risk	4617	244	19	0	0	0
Censored	0	1436	1569	1583	1583	1583
Events	0	2937	3029	3034	3034	3034

Figure 4-11 KMF - Stage at diagnosis.

The histology variable, as mentioned, is a vital diagnosis, Table 4.15. Note that the Undifferentiated value is not the same as the unknown.

To consult the KMF for the histology variable, consult Section B- Attachment 2.

Table 4-15 - Histology - Descriptive table.

Histology	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Adenocarcinoma	6570	15,1	23,6	64	63,3	697	454	1372	4047
Adenosquamous	163	17,1	28,0	66	65,3	23	21	49	70
Squamous	3210	14,8	23,7	67	66,8	341	386	1402	1081
Large cell carcinoma	348	10,25	22,6	64	63,4	28	24	99	197
Sarcomatoid	41	16,3	25,2	65	61,4	7	9	12	13
Undifferentiated	312	9,05	16,0	64	63,9	14	12	94	192
Neuroendocrine large cell carcinoma	147	10,3	17,3	64	63,9	21	7	47	72

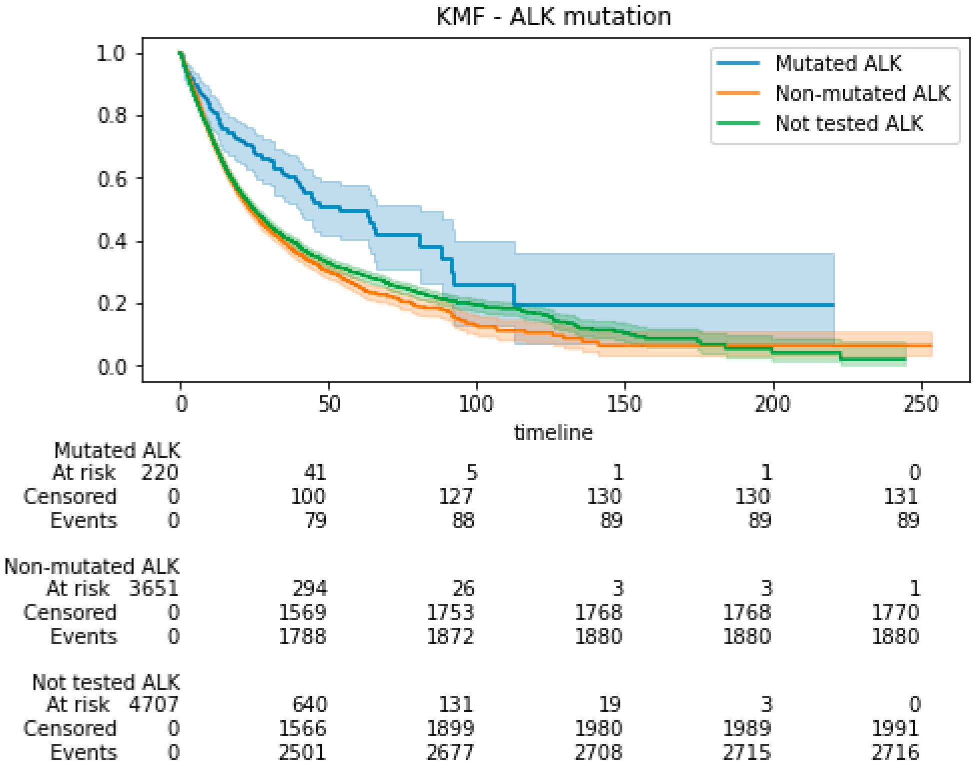
Regarding molecular pathologies, and now just considering the ALK translocation, as the numbers confirm Table 4.16., the mutated patients have higher survival than the non-mutated.

As mentioned, the therapies for patients with molecular markers are specific. Currently, there are new targeted therapies (TKIS or monoclonal antibodies), much more specific and less toxic, whose objective is to block the development and growth of the cells that cause lung cancer and thus prevent the tumour from growing.

Table 4-16 ALK translocation - Descriptive table.

ALK	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Mutated ALK	268	18,85	29,1	59	58,3	10	6	54	198
Not mutated ALK	4439	13,4	20,0	64	63,4	324	265	958	2892
Not tested ALK	6084	15,7	25,5	66	65,4	797	642	2063	2582

The KMF, presented in Figure 4.12. states precisely the number in the descriptive analysis. Also, we can see a difference in survival between the patients whose result was negative and the ones not tested, which indicates that could be a reason to suspect the presence of the mutation since the beginning of the diagnosis.



```
Mutated ALK
    At risk   220     41     5     1     1     0
   Censored     0    100   127   130   130   131
     Events     0     79    88    89    89    89

Non-mutated ALK
    At risk  3651    294    26     3     3     1
   Censored     0   1569  1753  1768  1768  1770
     Events     0   1788  1872  1880  1880  1880

Not tested ALK
    At risk  4707    640   131    19     3     0
   Censored     0   1566  1899  1980  1989  1991
     Events     0   2501  2677  2708  2715  2716
```

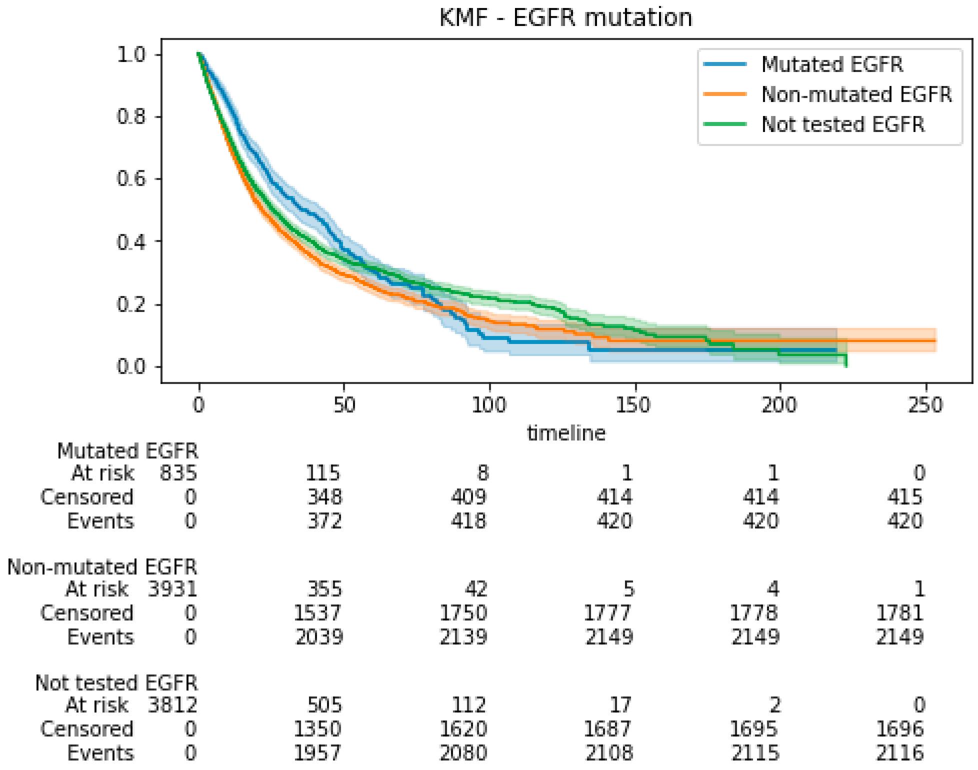
Figure 4-12 KMF - ALK mutation.

The EGFR mutation follows the same reasoning line as the ALK mutated patients, as described above. Table 4.17. presents a descriptive analysis of the EGFR mutation.

Table 4-17 - EGFR mutation - Descriptive table.

EGFR	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Mutated EGFR	1033	19	26,2	67	65,3	82	41	145	765
Not mutated EGFR	4831	13,3	20,8	63	62,9	313	287	1108	3123
Not tested EGFR	4927	15,5	25,3	66	65,7	736	585	1822	1784

Mutated EGFR patients have the longest survival, given that these patients are usually non-smokers, but develop the disease due to the mutation. In these patients, the doctors are able to provide specific targeted therapies that affect only mutated cells, unlike usual chemotherapy that affects both tumour cells and non-tumour cells. The difference in survival between mutated and non-mutated patients is significant during the first 4-5 years of treatment and is higher in mutated patients due to tyrosin kinase inhibitors (TKI).



	0	50	100	150	200	250
Mutated EGFR						
At risk	835	115	8	1	1	0
Censored	0	348	409	414	414	415
Events	0	372	418	420	420	420
Non-mutated EGFR						
At risk	3931	355	42	5	4	1
Censored	0	1537	1750	1777	1778	1781
Events	0	2039	2139	2149	2149	2149
Not tested EGFR						
At risk	3812	505	112	17	2	0
Censored	0	1350	1620	1687	1695	1696
Events	0	1957	2080	2108	2115	2116

Figure 4-13 KMF - EGFR mutation.

4.4.4 COMORBIDITIES DATA

The diagram in figure 4.14. illustrates the comorbidities distribution, concerning the 70,36% of the cohort used for the descriptive analysis.

To see an individual analysis of the comorbidities, consult Section B- Attachment 3.

As mentioned in the pre-processing data chapter, the comorbidities were initially binary, so each comorbidity was an independent variable, which was then converted to numeric.

In order to be converted to categorical, it was necessary to see its behaviour analysing at the same time the survival. Groups were then defined considering the graphic presented in Section B- Attachment 4.

Patients with one to three (1-3) corresponded to 47,72% and were defined as the baseline profile based on this percentage.

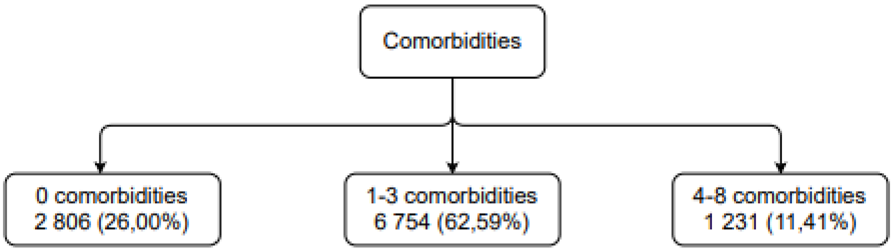


Figure 4-14 - Comorbidities - Distribution diagram.

The descriptive analysis Table 4.18. shows that patients with comorbidities have lower survival than those with no comorbidities.

Table 4-18 - Comorbidities - Grouped by number of comorbidities - Descriptive table.

Groups comorbidities	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
0	2806	15,65	25,04	59	58,91	180	196	786	1644
[1-3]	6754	14,6	23,22	66	65,80	774	584	1930	3466
[4-8]	1231	13,3	20,13	69	69,22	177	133	359	562

A boxplot was generated to display the distribution and see the outlier of the group's defined number of comorbidities (Figure 4.15.).

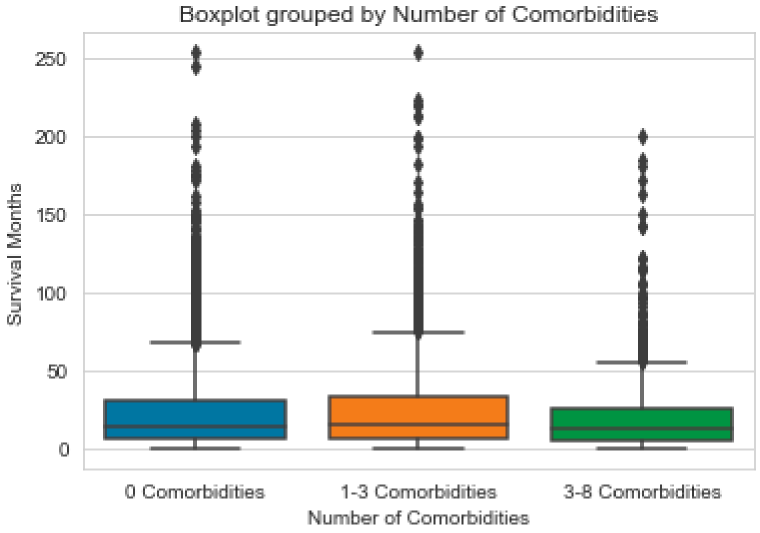
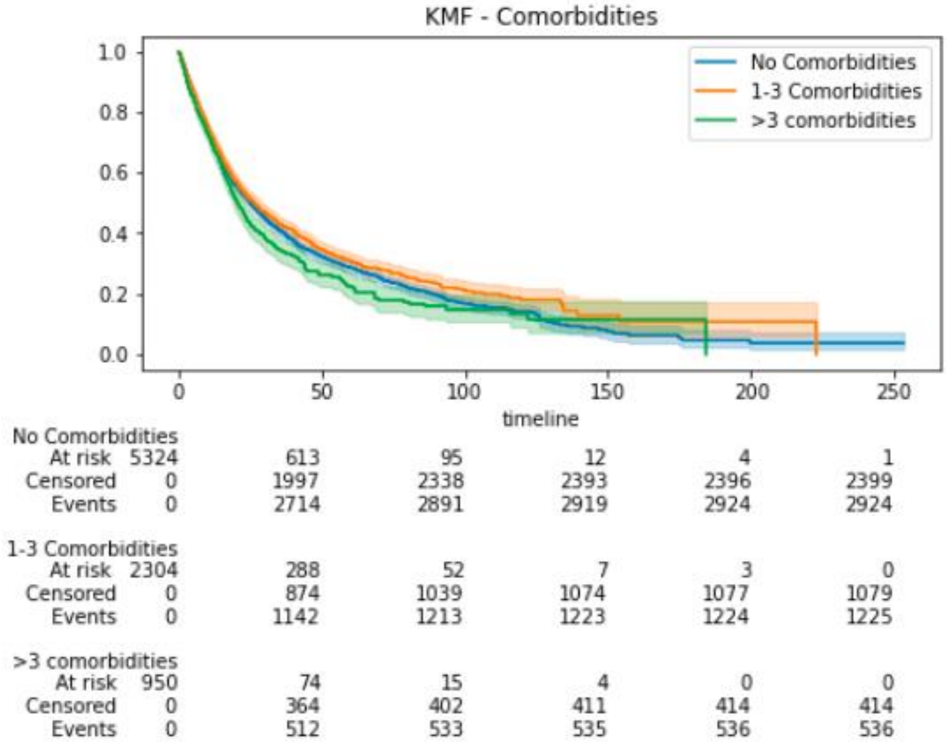


Figure 4-15 - Boxplot grouped by Number of Comorbidities.

The KMF, presented in Figure 4.16. states precisely the number in the descriptive analysis, although it is possible to see that the survival curve for the group with more than three comorbidities is quite irregular since the beginning, and near the 125 months, it crosses over with the group of one to three comorbidities. This is a red flag that there is a chance that this variable will fail the non-proportional test in the cox model.



No Comorbidities						
At risk	5324	613	95	12	4	1
Censored	0	1997	2338	2393	2396	2399
Events	0	2714	2891	2919	2924	2924
1-3 Comorbidities						
At risk	2304	288	52	7	3	0
Censored	0	874	1039	1074	1077	1079
Events	0	1142	1213	1223	1224	1225
>3 comorbidities						
At risk	950	74	15	4	0	0
Censored	0	364	402	411	414	414
Events	0	512	533	535	536	536

Figure 4-16 - KMF Comorbidities.

4.4.5 PATIENTS AND FAMILY HISTORY OF CANCER DATA

The diagram in figure 4.17. illustrates the historical data distribution concerning the 70,36% of the cohort used for the descriptive analysis.

To see an individual descriptive analysis of each patient's previous cancer type, consult Section B-Attachment 5.

As well as the comorbidities, the previous cancers were also independent variables, which were then converted to a binary variable, only considering if the patient had it or not.

Patients with no previous cancer corresponded to 82,25% and were defined as the baseline profile based on this percentage. The value 'unknown' was considered in the descriptive analysis, Table 4.19., but removed from the KMF and the Cox model.

Regarding the Family history of cancer, it was also only considered if the patient's family had previous cancers or not. No previous cancer in the family corresponded to 40,72% of the population, so it was defined as a part of the baseline profile.

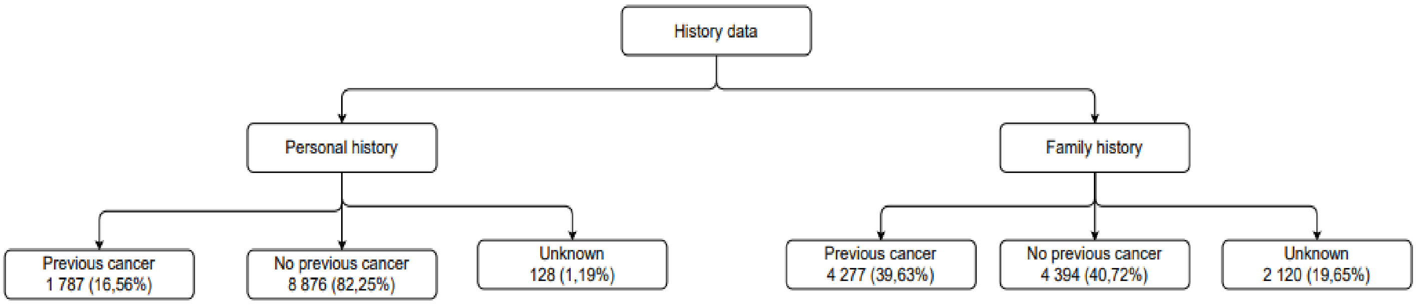


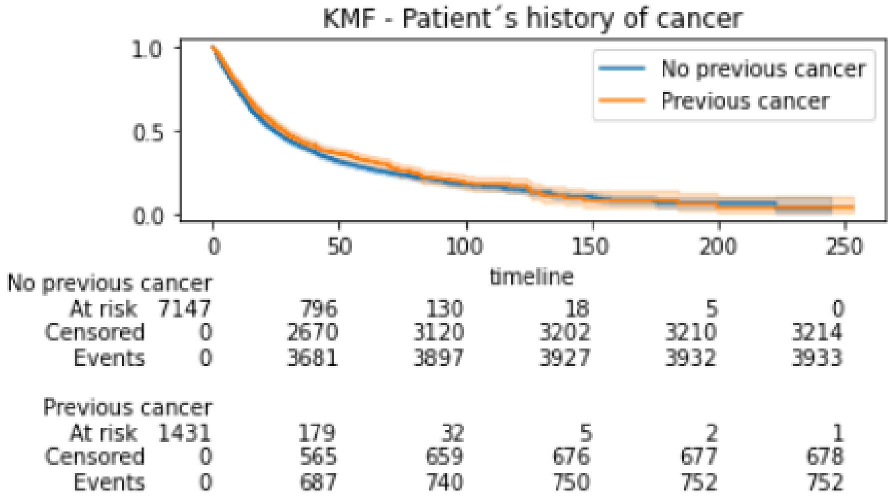
Figure 4-17 - History data - Distribution diagram.

Regarding the patient´s cancer history, the number of patients with no previous cancer is higher than those with previous cancer.

Table 4-19 - Patient history - Descriptive table.

Patient history	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Previous cancer	1787	16,2	25,0	68	67,8	326	192	509	760
No previous cancer	8876	14,5	22,9	64	63,7	785	708	2529	4854
Unknown	128	14,2	28,6	64	64,6	20	13	37	58

In the KMF presented in Figure 4.18. it is possible to see that the difference between the survival curves is not, in fact, very significant and start to cross over each other near to the 100 months.



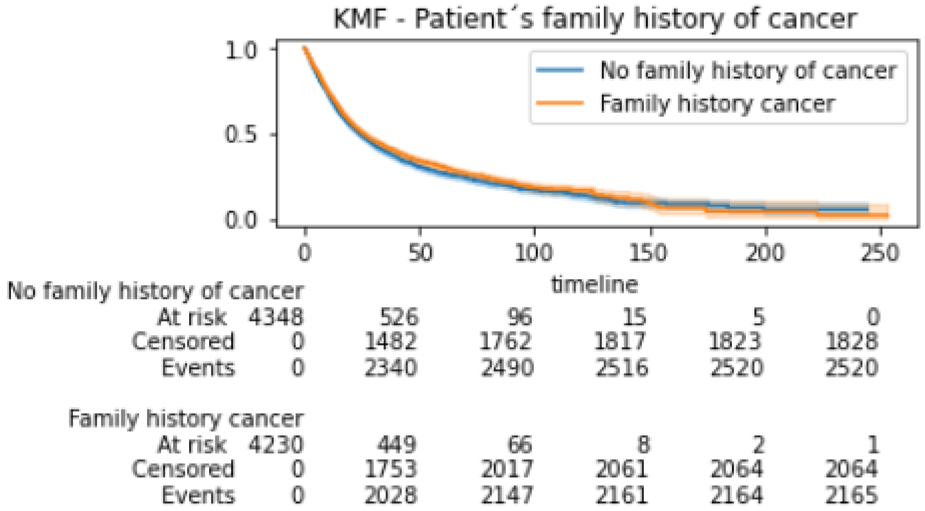
```
No previous cancer
   At risk  7147      796      130       18        5        0
  Censored     0     2670     3120     3202     3210     3214
    Events     0     3681     3897     3927     3932     3933

  Previous cancer
   At risk  1431      179       32        5        2        1
  Censored     0      565      659      676      677      678
    Events     0      687      740      750      752      752
```

Figure 4-18 - KMF - Patient's history of cancer.

Regarding the family history of cancer, there is no significant difference in survival, as shown in Table 4.20. and the KMF in Figure 4.19. Although we can associate this little difference to the fact that the patients aware of its family history may have more precautions and a more frequent follow-up than the others.

Table 4-20 - Family history of cancer - Descriptive analysis.

Family history	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Previous cancer	4277	14,6	22,3	64	63,6	409	356	1225	2287
No previous cancer	4394	14,7	23,6	65	64,8	414	362	1227	2391
Unknown	2120	14,7	25,0	65	65,1	308	195	623	994



```
No family history of cancer
         At risk  4348      526       96       15        5        0
        Censored     0     1482     1762     1817     1823     1828
          Events     0     2340     2490     2516     2520     2520

   Family history cancer
         At risk  4230      449       66        8        2        1
        Censored     0     1753     2017     2061     2064     2064
          Events     0     2028     2147     2161     2164     2165
```

Figure 4-19 - KMF - Patient's family history of cancer.

4.4.6 FIRST TREATMENT LINE (DRUGS THERAPY) DATA

The diagram in figure 4.20. illustrates the first treatment line of drug therapy distribution concerning the 70,36% of the cohort used for the descriptive analysis.

Chemotherapy corresponds to 57,16%, and based on this percentage was the baseline profile defined for the models.

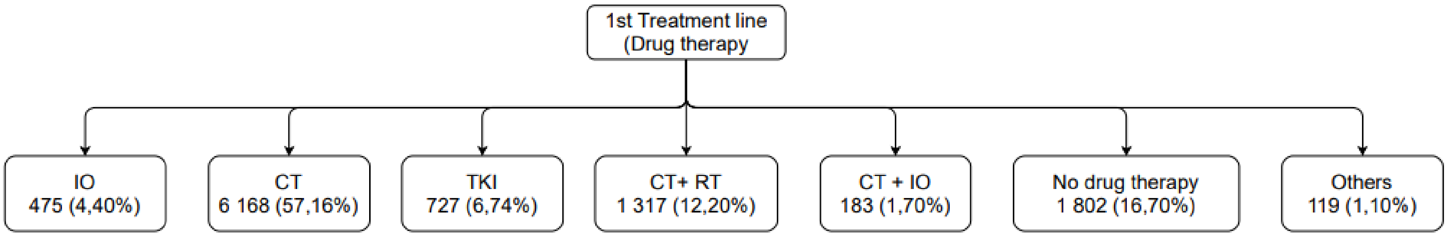


Figure 4-20 - First Treatment line (Drug therapy).

Legend: IO: Immunotherapy; CT: Chemotherapy, TKI: Tyrosine-Kinase Inhibitor; CT + RT: Chemotherapy + Radiotherapy; CT+ IO: Chemotherapy + Immunotherapy.

To see the KMF of the first treatment line of drug therapy, consult Section B- Attachment 6.

As mentioned before in this work, the treatment strategy should take into account factors such as histology, molecular pathology, age, PS, comorbidities and the patient's preferences. Treatment decisions should ideally be discussed within a multidisciplinary tumour board that can evaluate and change management plans, including recommending additional investigations and changes in treatment modality. The best treatment option is selected upon these features to ensure better response and increase survival. Nevertheless, the heterogeneity of the disease and high inter variability of the patients may alter the patient´s prognosis.

Table 4-21 - First treatment line (Drug therapy) - Descriptive table.

Drug therapy	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
IO	475	10,7	14,7	66	65,0	22	9	37	407
CT	6168	15,6	24,6	64	63,2	375	626	1667	3500
TKI	727	16,5	23,3	67	65,2	31	7	45	644
CT + RT	1317	19,9	28,0	64	63,5	63	86	1037	131
CT + IO	183	10,7	16,4	63	63,0	7	1	13	162
No drug treatment	1802	7,9	18,9	70	68,8	626	180	258	738
Others	119	13,4	18,4	63	63,2	7	4	18	90

4.4.7 RADIOTHERAPY DATA

For locally advanced, non-resectable stage IIIA tumours, radiation with chemotherapy remains the standard of care; for selected patients with IIIB (with pleural effusion) or IV disease, chemotherapy remains the standard treatment in conjunction with supportive care.

The diagram in figure 4.21. illustrates the radiotherapy distributions concerning the 70,36% of the cohort used for the descriptive analysis.

Patients who did not have radiotherapy correspond to 53,40% and the baseline profile was defined based on this percentage.

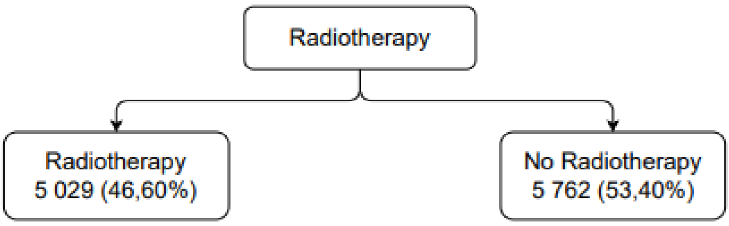


Figure 4-21 - Radiotherapy - Distribution diagram.

Table 4-22 - Radiotherapy - Descriptive table.

	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
Radiotherapy	5029	17,8	26,0	63	63,0	332	337	2085	2275
No radiothe-rapy	5762	12	21,0	66	65,6	799	576	990	3397

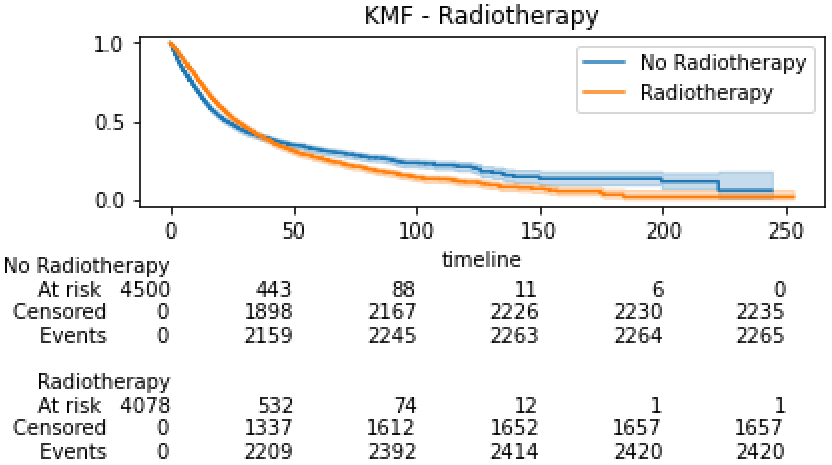


Figure 4-22 - KMF Radiotherapy.

4.4.8 SURGERY DATA

Surgery remains the standard of care in early-stage (I-II) non-small cell lung cancer (NSCLC). Radical radiotherapy or stereotactic ablative radiotherapy (SABR) are alternatives. Also, patients with IB or II disease are now being offered adjuvant chemotherapy. Options for locally advanced (III) NSCLC include surgery with postoperative chemotherapy or chemoradiotherapy. Some stage IIIA tumours are resectable but often receive pre or post-operative radiation and/or chemotherapy.

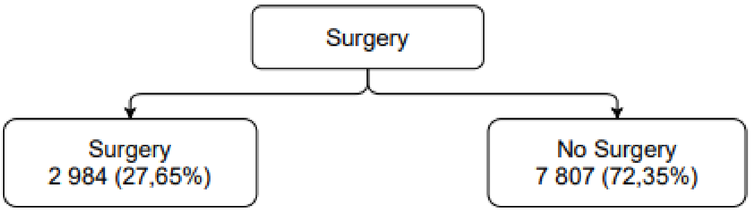
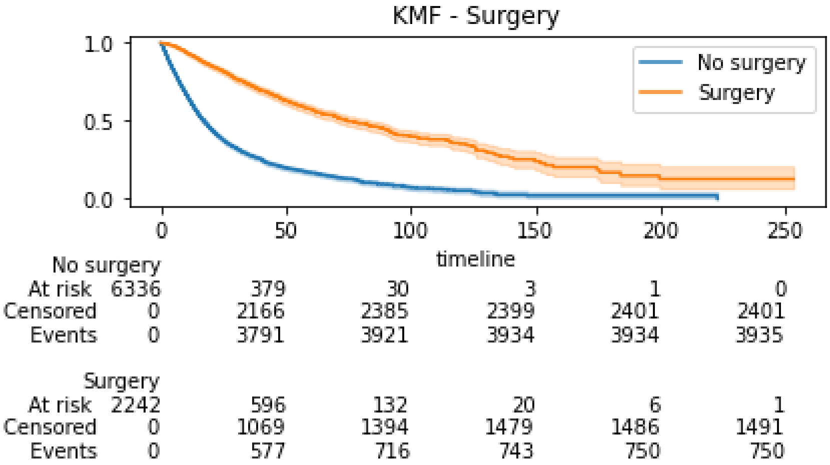


Figure 4-23 - Surgery - Distribution Diagram

Table 4-23 - Surgery - Descriptive table.

	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV	Male	Female
Surgery	2984	29,55	38,8	64	63,7	976	716	925	367	2188	796
No Surgery	7807	11,7	17,4	65	64,7	155	197	2150	5305	5841	1966



```
No surgery
  At risk  6336     379     30      3      1      0
 Censored     0    2166   2385   2399   2401   2401
   Events     0    3791   3921   3934   3934   3935

   Surgery
  At risk  2242     596    132     20      6      1
 Censored     0    1069   1394   1479   1486   1491
   Events     0     577    716    743    750    750
```

Figure 4-24 - KMF Surgery.

CHAPTER 5

COX'S MULTIVARIABLE PROPORTIONAL HAZARD MODELS RESULTS

This section goes through the development of Cox's multivariable proportional hazard model. It includes the baselines defined, the model's results and evaluation metrics.

5.1 MODELS SETUP

The following sections present **five different models**; the first including all variables mentioned in the descriptive analysis, including all the diagnosis stages. This model allows the analysis of the entire group of lung cancer patients, precisely the difference in survival and risk of the patients with a different diagnosis.
The four remaining models concern each stage of diagnosis separately, allowing a more specific analysis of the cohorts. The analysis of each stage model also makes it possible to understand the model's results with all stages in a more detailed view.

Computing the Cox Model requires defining a baseline patient's profile, to which all the other covariates will be compared.
As mentioned in the review of the Cox model, a critical element of the Cox Proportional Hazard equation is that the baseline hazard is a function of time *t*, but not the parameters, whereas the partial hazard is a function of the parameters but not time.
The baseline profile was defined as the most common patient's profile. It does not appear in the analysis tables, as all the other covariates are compared to it. Although, it is possible to see the baseline survival curves in the graphics.
The baseline profile was defined the same for all the tested models, which brings a few points to be taken into consideration.
Below are presented the generic situations where changes were performed in the baseline, and during the following sections, when presenting each model individually, the same changes are detailed individually per attribute.

1. The 'All Stages Model' differs from all the other models as it has the variable Stages;
2. In the models specific for each diagnosis, it was necessary to remove some variables, as there were variables that did not concern every diagnosis;
3. In some situations, the baseline was not defined as the most commons patient profile to comply with previous baselines (and therefore, it is possible to compare models between themselves).

In order to fit a Cox Proportional Hazards model using the Lifelines library, it is necessary to use a one-hot encoding for categorical columns; in other words, all categorical variables must be converted to bin variables.

The function *dmatrices()* was used from the patsy library, Section C- Attachment 3 to perform the binarisation of the categorical variables.

It is at this time that the baseline is selected. One column from each hot-encoded variable must be dropped; otherwise, multi-collinearity issues will be created. It is this dropped column that becomes the baseline characteristics.

There are several approaches when dropping variables (e.g. First column). As mentioned above, the most common patient profile was chosen as the baseline, which affords an intuitive interpretation as the baseline closely resembles the population and induces robustness on estimations.

Now that the baseline is defined and all categorical variables converted to bin variables, it is time to fit the model. To consult an example of the model's generation code, consult Section C- Attachment 4.

The next step consists of assessing the results of the fitted model by looking at its significance and confidence (which are considered in the result's table).

After assessing the fitted model results, verifying whether the model adheres to the proportional hazard assumption is necessary. Of note, a red flag for this scenario is when the survival curves for a given covariate crossover each other when using Kaplan-Meier (cases flagged in the previous chapter).

Proportional hazard is a fundamental assumption in Cox regression, whereby we assume that the hazard ratios do not depend on time, and there are several approaches to diagnose potential violations. It was used *CoxPHFitter.check_assumptions method*, which computes statistics that check the proportional hazard assumption.

The p-value threshold was set at 0.05, and not that even under the null hypothesis of no violations, some covariates will be below the threshold by chance. Similarly, even minor deviances from the proportional hazard assumption will be flagged when there are many observations.

There are two plots for each variable that failed the test. The difference between these two plots is the order of how the residual values are displayed: Rank transformed time and KM-transformed time. When no pattern is present, the black line in the middle will be relatively flat, indicating that the residuals are not correlated with time.

So, for each model, it will be presented the baseline selected and considerations about it (when necessary), a results table (HR, p-value and interpretation), a graphic of the tables results ordered by significance, plot of each variable (survival versus time), and the residual of the variables that failed the non-proportional test.

5.2 COX REGRESSION EVALUATION METRICS

To assess the fitted model results, there are initially three key points to pay attention to, specifically to analyse the variable's behaviour itself [46-47].

Firstly, the statistical significance of each covariate.

The p-value is the probability of obtaining results at least as extreme as the observed results of a statistical hypothesis test, assuming that the null hypothesis is correct.

The lower the p-value, the greater the statistical significance of the observed difference. If the result of this column is below 0.05, it means that the covariate is statistically significant and safe to include.

Secondly, the effect of each covariate on the hazard ratio, referring back to the Cox Proportional Hazards equation, means that a patient's hazard ratio increases or decreases versus the baseline.

Finally, analyse how confident are the coefficients estimated. For this, will be analysed the values of exp(coef) lower 95% and - exp(coef) upper 95%. These bounds can also be viewed visually in the box-and-whisker plot. When the box-and-whisker of a variable cross the value one (1), it means that the variable is not significant, as there are patients in which the variable increases the risk and others where it decreases the risk.

After these variable's analyses, the Cox model's output also includes several metrics that can be used when comparing models and evaluate the overall model's performance, Log-likelihood ratio, Akaike information criterion and the concordance Index [45].

The Akaike information criterion (AIC) is a metric for comparing models as it relies on the log-likelihood. The Akaike information criterion (AIC) is a mathematical method for evaluating how well a model fits the data. In statistics, AIC compares different possible models and determines which one best fits the data. AIC is calculated from:

- The number of independent variables used to build the model.
- The maximum likelihood estimate of the model (how well the model reproduces the data).

The best-fit model, according to AIC, is the one that explains the greatest amount of variation using the fewest possible independent variables.

The Concordance Index is another censoring-sensitive measure, also known as the c-index. This measure evaluates the accuracy of the ranking of the predicted time.

The c-index can assume values between 0 to 1 as:

- 1.0 is a perfect concordance;
- 0.5 is the expected result from random predictions;
- 0.0 is perfect anti-concordance.

Fitted survival models typically have a concordance index between 0.55 and 0.75.

As previously mentioned, a critical assumption of the Cox model is the proportional hazards assumption: when the predictor variables do not vary over time, the hazard ratio comparing any two observations is constant with respect to time. Therefore, to perform credible estimation and inference, after assessing the fitted model results, the next step is to verify whether the model adheres to the proportional hazard assumption.

Schoenfeld proposed a chi-squared goodness-of-fit test statistic for the proportional hazards regression model which utilized a residual of the form Expected – Observed [48].

Shoenfeld residuals represent the difference between the observed covariate and the expected given the risk set at that time [49].

When viewing this type of plot, we do not want to see any sort of pattern in the residuals. When no pattern is present, the black line in the middle will be relatively flat, indicating that the residuals are not correlated with time.

5.3 ALL STAGES MODEL

This model included 8578 patients (number of observations) and 4685 deaths (events observed). As mentioned, it includes all the stages of diagnosis. It was tested with 30 attributes (after the variables binarization) and with the baseline presented in table 5.1.

Table 5-1- All stages Model - Baseline profile.

Attribute	**Baseline profile**
Gender	Male
Age group	]45,70]
Stage	IV
Number of comorbidities	Group 2 – [1,3]
Smoking habit	Former smoker
Patient previous cancer	No previous cancer
Family previous cancer	No previous family cancer
Histology	Adenocarcinoma
1st Treatment line	CT
ALK	Not tested
EGFR	Not tested
Surgery	No surgery
Radiotherapy	No radiotherapy

5.3.1 RESULTS

The model results are presented in Table 5.2., along with its interpretation.

Table 5-2 - Cox Model Results- All stages model (30 attributes).

Covariante	**HR for death**	**p-value**	**Interpretation**
	(95% CI)		
Gender- Female	**0.86** (0.80 - 0.93)	**<0.005**	When compared to men, women have 14% less risk of dying from lung cancer, considering all the other covariates the same.
Smoker habit			
Current smoker	**1.15** (1.08 - 1.22)	**<0.005**	Current smokers have more 15% risk than former smokers, considering all the other covariates.
Never smoker	**0.77** (0.69 -0.86)	**<0.005**	Never smokers have less 23% risk than former smokers, considering all the other covariates the same.
Comorbidities			
No comorbidities	**0.94** (0.87 -1.00)	**0.06**	Patients with no comorbidities have less 6% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
+ 4 comorbidities	**1.06** (0.96 -1.16)	**0.26**	Patients with more than four (4) comorbidities have a more 6% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
Age			
<45	**0.98** (0.83 -1.15)	**0.77**	Patients with <=45 years old have less 4% risk than patients with 45-70 years old, considering all the other covariates the same.
>70	**1.21 (1.13 – 1.29)**	**<0.005**	Patients with >70 years old have more 21% risk than patients with 45-70 years old, considering all the other covariates the same.
Stages			
Stage I	**0.17** (0.15 - 0.20)	**<0.005**	Patients diagnosed with stage I have less 83% risk than patients with stage IV, considering all the other covariates the same.
Stage II	**0.33** (0.29 -0.38)	**<0.005**	Patients diagnosed with stage II have less 67% risk than patients with stage IV, considering all the other covariates the same.
Stage III	**0.53** (0.49 -0.58)	**<0.005**	Patients diagnosed with stage III have less 47% risk than patients with stage IV, considering all the other covariates the same.
History			
Patient with previous cancer	**0.98** (0.91 - 1.06)	**0.65**	Patients with previous cancer have 2% less risk than patients with no previous cancer history.
Patient with family history of cancer	**0.97** (0.92 -1.03)	**0.31**	Patients with a family history of cancer have 3% less risk than patients with no family history of cancer.
Histology			
Adenosquamous	**0.91** (0.72 -1.14)	**0.41**	Patients diagnosed with Adenosquamous have less 9% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Squamous	**1.03** (0.95 -1.12)	**0.49**	Patients diagnosed with Squamous have a more 3% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.

Covariante	HR for death	p-value	Interpretation
	(95% CI)		
Large cell carcinoma	**1.30** (1.12 -1.52)	**<0.005**	Patients diagnosed with Large cell carcinoma have a more 30% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Sarcomatoid	**1.30** (0.79 - 2.14)	**0.30**	Patients diagnosed with Sarcomatoid have a more 30% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Undifferentiated	**1.26** (1.07-1.49)	**0.01**	Patients diagnosed with Undifferentiated have a more 26% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Neuroendocrine large cell carcinoma	**1.50** (1.16 -1.94)	**<0.005**	Patients diagnosed with Neuroendocrine large cell carcinoma have a more 50% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
1s treatment line			
IO	**0.76** (0.66 -0.89)	**<0.005**	Patients with IO as the first treatment line have less 24% risk than patients with CT as the first treatment line, considering all the other covariates the same.
No drug Treatment	**2.72** (2.47 -2.99)	**<0.005**	Patients with no drug treatment in the first treatment line have more 172% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Other Drugs	**0.76** (0.57 -1.01)	**0.06**	Patients with 'others' as the first treatment line have less 24% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+IO	**0.89** (0.71 -1.13)	**0.35**	Patients with CT+IO as the first treatment line have less 11% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+RT	**0.79** (0.71 -0.88)	**<0.005**	Patients with CT+RT as the first treatment line have less 21% risk than patients with CT as the first treatment line, considering all the other covariates the same.
TKI	**0.78** (0.67 -0.92)	**<0.005**	Patients with TKI as the first treatment line have less 22% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Surgery	**0.41** (0.38 – 0.46)	**<0.005**	Patients who did surgery have less 59% risk than patients who did not, considering all the other covariates the same.
Radiotherapy	**0.94** (0.88 – 1.00)	**0.06**	Patients who did radiotherapy have more 6% risk than patients who did not, considering all the other covariates the same.
Molecular Markers			
Not mutated ALK	**0.91** (0.84 -1.00)	**0.04**	Patients without ALK mutation have less 9% risk than non-tested patients, considering all the other covariates the same.
Mutated ALK	**0.57** (0.46 -0.72)	**<0.005**	Patients with ALK mutation have less 43% risk than non-tested patients, considering all the other covariates the same.
Not mutated EGFR	**1.02** (0.92 -1.12)	**0.74**	Patients without EGFR mutation have 2% more risk than nontested patients, considering all the other covariates the same.
Mutated EGFR	**0.87** (0.74 -1.01)	**0.06**	Patients with EGFR mutation have less 13% risk than nonmutated patients, considering all the other covariates the same.

Upper and lower bounds for each coefficient can also be seen visually in Figure 5.1. Note that when viewing the bounds in this way, a variable can be deemed as not significant when the confidence interval includes the value one (1).

Also, to compare the values within variables, a plot of the survival probability against survival time (months) was generated, as well as the cumulative hazard against survival time (in the necessary cases). Figure 5.2. illustrates some of the survival curves. In the case of the variable stages, it is clear the difference of survival between the groups, but there are cases in which the curves are very close to each other, and these are the best scenarios to use the cumulative hazard, as in the variable Histology case.

Even though the considerable number of variables, the model have: C**oncordance** of 0.73, **Partial AIC** of 74648.13 and **log-likelihood ratio** of 2643.17 on 30 df.

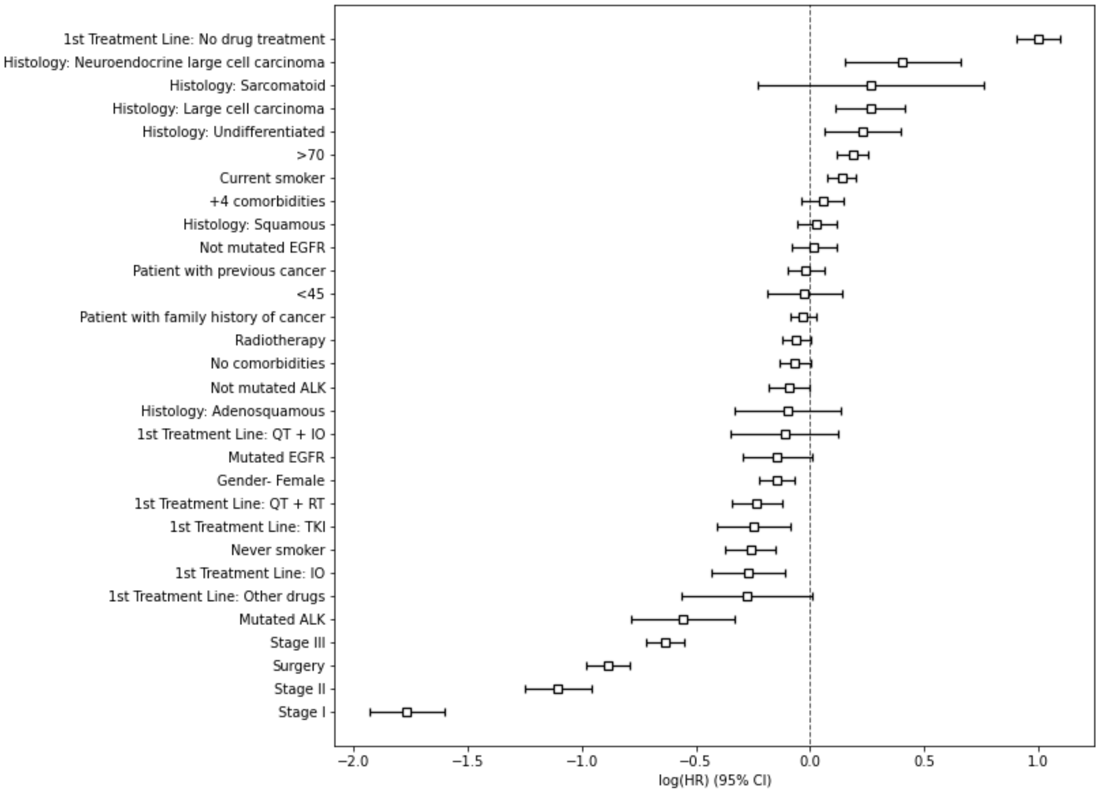


Figure 5-1 Cox Model Results- All stages model (30 attributes).

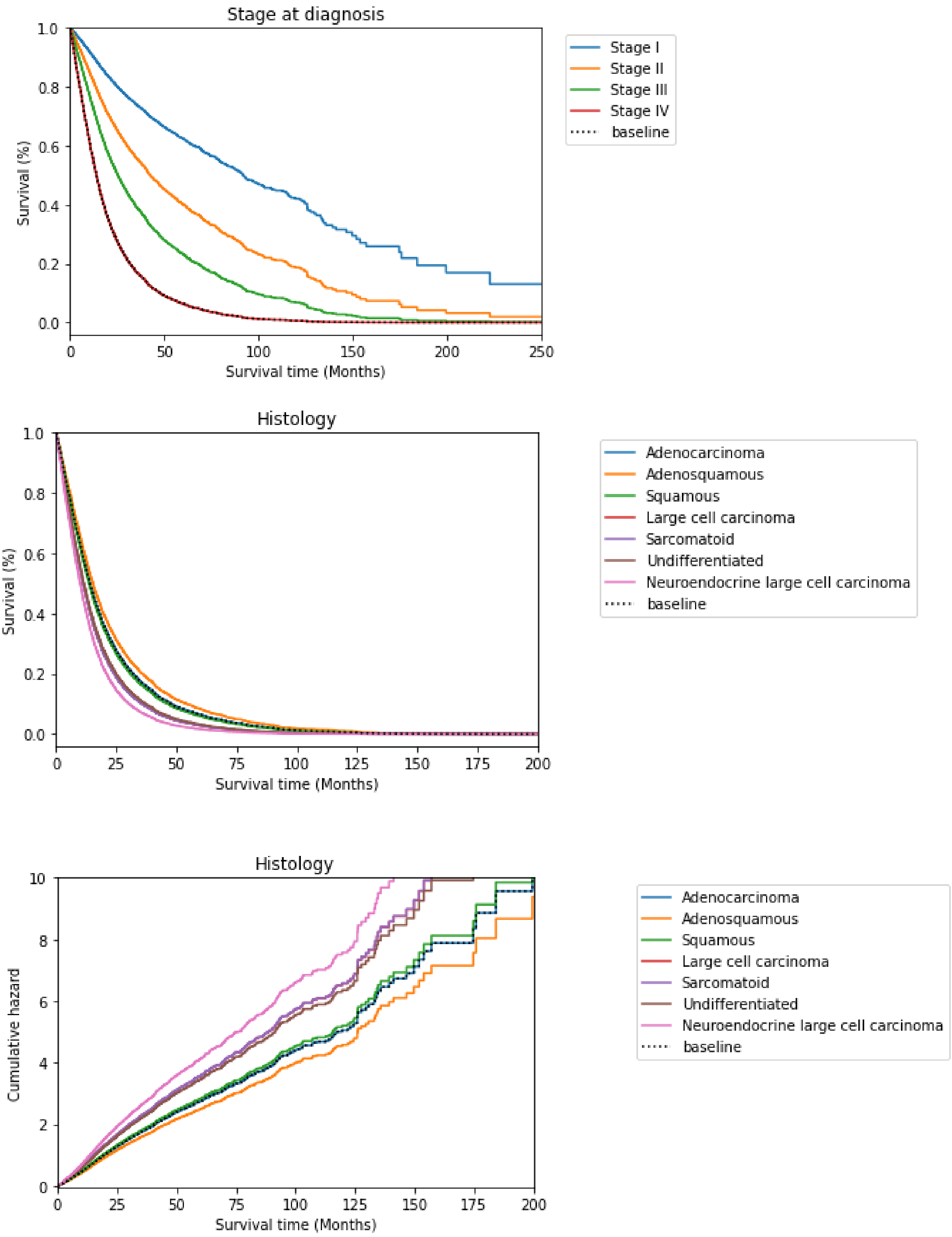


Figure 5-2 - Survival curves of All Stages model. The first image is for the Stages variable, the second and third image is for the variable Histology plotting the survival and the cumulative hazard, respectively.

Before the final model, the one just presented, the variable molecular marker was tested in two different ways:

- Firstly, it was considered that all the patients who were not tested were negative. This way, we discard the variable which indicated if the test was performed or not, ending up with three attributes: Mutated EGFR, Mutated ALK, and Non mutated (baseline). The results of this model simulation are in table 5.3.
- The second way, the version implemented in the final model, uses six attributes ratter then tree: Mutated EGFR, not mutated EGFR, not tested EGFR, Mutated ALK, non-mutated ALK, and Not tested ALK. The results of this model simulation are in Table 5.2.

Table 5-3 - Molecular markers result considering all non-tested patients non-mutated.

Molecular Markers			
ALK	**0.60** (0.48 -0.75)	**<0.005**	Patients with ALK mutation have less 40% risk than non-mutated/non-tested patients.
EGFR	**0.85** (0.74 -0.98)	**0.02**	Patients with EGFR mutation have less 15% risk than non-mutated/non-tested patients.

Looking at the results of both simulations and taking into account that the remaining variables of the model were the same, we can take several conclusions:

Since in the 1st simulation, it was considered that all the non-tested were negative, for ALK mutated patients, the risk was 40% less than the non-mutated/non-tested, comparing with the 2nd simulation, where the non-mutated and non-tested are separated variables, for ALK mutated patients, the risk was 43% less risk than the non-tested patients (both values significant).

Regarding the EGFR mutation, we can also see that the results of both models are similar, being the first one, the patients with EGFR mutation had 15% less risk than patients non-mutated/non-tested, and the second, the patients with EGFR mutation had 16% less risk than the non-tested patients.

Therefore, being these results significant in both models and similar between themselves, it is a mirror to the accuracy of selecting the patients who perform or not the molecular markers test.

5.3.2 TESTING AND INTERPRETING ASSUMPTIONS

Schoenfeld residuals are then used to assess the proportional hazard assumption. As already mentioned, if the proportional hazard assumption holds, we would expect to see a flat smoothed scatterplot of residuals against time.

In this first model, fifteen (15) variables failed the non-proportional test. The Figures 5.3., 5.4. and 5.5., are some examples of the (scaled) Schoenfeld residuals shown for a multivariable Cox regression model fitted to a simulated dataset with 30 covariates.

The remaining variables that failed the proportional hazard assumption can be found in Section D - Attachment 1, although the three cases presented below were explicitly selected to explain and interpreter the assumptions.

Figure 5.3. illustrates the Schoenfeld residuals of the variable Radiotherapy, and it is clear that they are incompatible with the proportional hazards assumption, which was expected since the Kaplan-Meier method, Figure 4.22.

Figure 5.4. illustrates the Schoenfeld residuals of the variable Mutated EGFR, and we can see that it starts to fail at the end of the time, which can be justified as, at that time, the number of patients under observation is low.

Finally, in figure 5.5, the Schoenfeld residuals of the variable Not mutated EGFR show minor changes, which we know is possible to happen based on the immense number of variables in the model. In this last case, we could discard this variable failure.

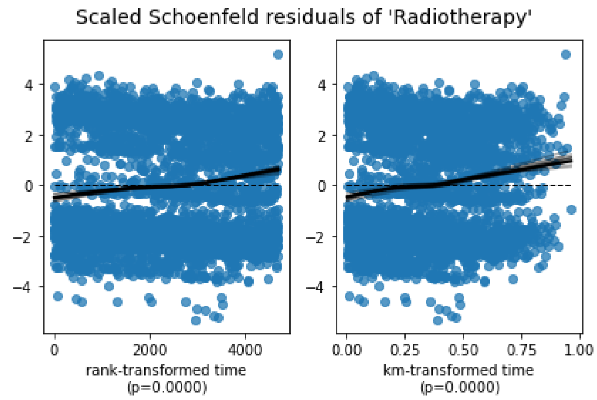


Figure 5-4 – All Stages Model - Scaled Scoenfeld residuals of Radioterapy variable.

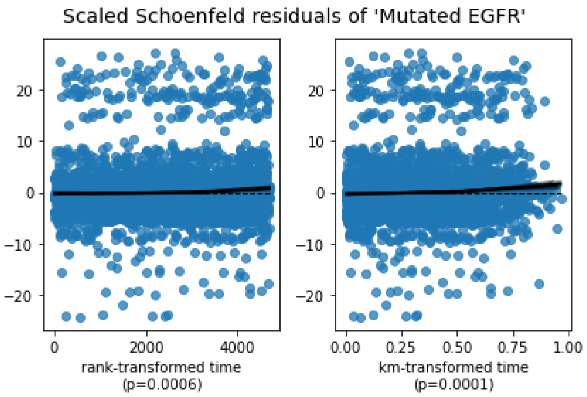


Figure 5-3 - All Stages Model - Scaled Scoenfeld residuals of Mutated EGFR variable.

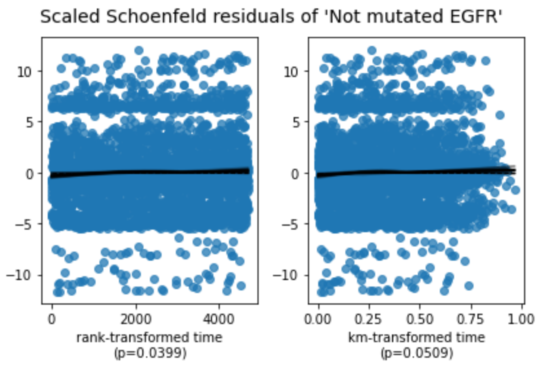


Figure 5-5 - All Stages Model - Scaled Scoenfeld residuals of Not mutated EGFR variable.

5.4 STAGE I MODEL

This model included 807 patients (number of observations) and 216 deaths (events observed). In this model, the variable stage was removed as it only concerns the Stage I diagnosis. It was tested with 27 attributes (after the variables binarization) and with the baseline presented in table 5.4.

Table 5-4 - Stage I Model - Baseline profile.

Attribute	Baseline profile
Gender	Male
Age group	]45,70]
Number of comorbidities	Group 2 – [1,3]
Smoking habit	Former smoker
Patient previous cancer	No previous cancer
Family previous cancer	No previous family cancer
Histology	Adenocarcinoma
1st Treatment line	CT (In fact, the most common patient's profile with stage I, did not have drug treatment, but to comply with the previous pattern, we maintain CT treatment as baseline).
ALK	Not tested
EGFR	Not tested
Surgery	No surgery (In fact, the most common patient's profile with stage I, did the surgery, but in order to comply with the previous pattern, we maintain no surgery performed as the baseline – That way, we can see the risk of the patients who had surgery compared to the ones who did not).
Radiotherapy	No radiotherapy

5.4.1 RESULTS

The model results are presented in Table 5.5., along with its interpretation.

Table 5-5 – Cox Model Results- Stage I model (27 attributes).

Covariante	HR for death	p-value	Interpretation
	(95% CI)		
Gender- Female	**0.96** (0.62-1.51)	**0.87**	Females have less 4% risk than males, considering all the other covariates the same.
Smoker habit			
Current smoker	**0.86** (0.62-1.20)	**0.38**	Current smokers have less 14% risk than former smokers, considering all the other covariates.
Never smoker	**0.35** (0.19-0.66)	**<0.005**	Never smokers have less 65% risk than former smokers, considering all the other covariates the same.
Comorbidities			
No comorbidities	**1.38** (0.89-2.13)	**0.15**	Patients with no comorbidities have more 38% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
+ 4 comorbidities	**2.35** (1.68-3.29)	**<0.005**	Patients with more than four (4) comorbidities have more 135% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
Age			
<45	**0.41** (0.05-3.50)	**0.41**	Patients with <=45 years old have less 59% risk than patients with 45-70 years old, considering all the other covariates the same.
>70	**2.35** (1.70-3.25)	**<0.005**	Patients with >70 years old have more 235% risk than patients with 45-70 years old, considering all the other covariates the same.
History			
Patient with previous cancer	**1.24** (0.92-1.67)	**0.16**	Patients with previous cancer have more 24% risk than patients with no previous cancer history.
Patient with family history of cancer	**0.74** (0.55-0.99)	**0.04**	Patients with a family history of cancer have 26% less risk than patients with no family history of cancer.
Histology			
Adenosquamous	**0.71** (0.27-1.83)	**0.47**	Patients diagnosed with Adenosquamous have less 29% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Squamous	**0.87** (0.62-1.24)	**0.45**	Patients diagnosed with Squamous have less 13% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Large cell carcinoma	**1.27** (0.59-2.74)	**0.54**	Patients diagnosed with large cell carcinoma have a more 27% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Sarcomatoid	**0.47** (0.06-3.66)	**0.47**	Patients diagnosed with Sarcomatoid have less 53% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Undifferentiated	**0.85** (0.32-2.26)	**0.75**	Patients diagnosed with Undifferentiated have less 15% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.

Neuroendocrine large cell carcinoma	**1.26** (0.39-4.07)	**0.70**	Patients diagnosed with Neuroendocrine large cell carcinoma have a more 26% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
1s treatment line			
IO	**2.05** (1.01-4.16)	**0.05**	Patients with IO as the first treatment line have more 105% risk than patients with CT as the first treatment line, considering all the other covariates the same.
No drug Treatment	**0.51** (0.37-0.70)	**<0.005**	Patients with no drug treatment in the first treatment line have less 49% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Other Drugs	**2.09** (0.43-10.20)	**0.36**	Patients with 'others' as the first treatment line have more 109% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+IO	**0.31** (0.04-2.28)	**0.25**	Patients with CT+IO as the first treatment line have less 69% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+RT	**0.68** (0.39-1.18)	**0.17**	Patients with CT+RT as the first treatment line have less 32% risk than patients with CT as the first treatment line, considering all the other covariates the same.
TKI	**0.93** (0.39-2.24)	**0.88**	Patients with TKI as the first treatment line have less 7% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Surgery	**0.49** (0.33-0.72)	**<0.005**	Patients who did surgery have less 51% risk than patients who did not, considering all the other covariates the same.
Radiotherapy	**1.42** (1.02-1.96)	**0.04**	Patients who did radiotherapy have more 42% risk than patients who did not, considering all the other covariates the same.
Molecular Markers			
Not mutated ALK	**0.76** (0.45-1.27)	**0.29**	Patients without ALK mutation have less 24% risk than non-tested patients, considering all the other covariates the same.
Mutated ALK	**1.62** (0.42-6.21)	**0.48**	Patients with ALK mutation have less 62% risk than non-tested patients, considering all the other covariates the same.
Not mutated EGFR	**1.19** (0.73-1.95)	**0.48**	Patients without EGFR mutation have 19% more risk than nontested patients, considering all the other covariates the same.
Mutated EGFR	**0.99** (0.44-2.25)	**0.98**	Patients with EGFR mutation have less 1% risk than nonmutated patients, considering all the other covariates the same.

Upper and lower bounds for each coefficient can also be seen visually in Figure 5.6.

The model has C**oncordance** of 0.72, **Partial AIC** of 2359.84 and **a log-likelihood ratio** of 150.46 on 27 df.

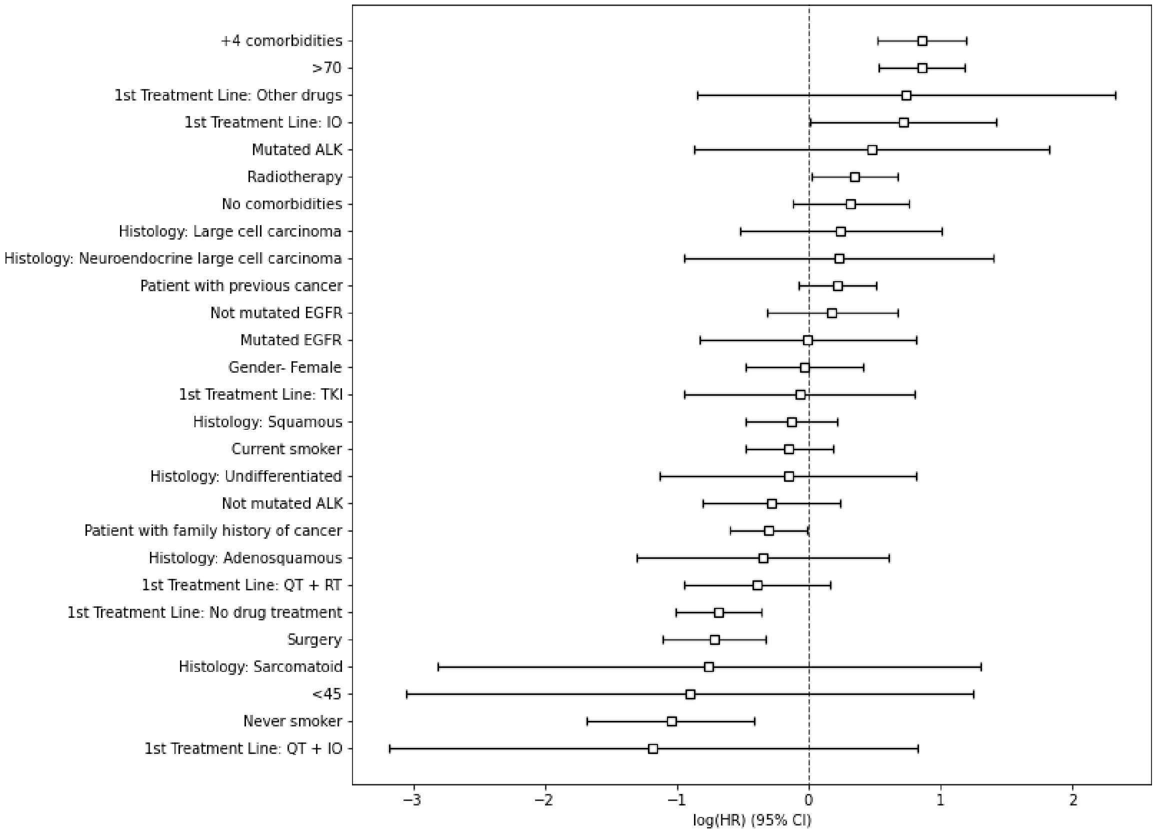


Figure 5-6 Cox Model Results- Stage I model (27 attributes).

5.4.2 TESTING AND INTERPRETING ASSUMPTIONS

In this model, three (3) variables failed the non-proportional test. The (scaled) Schoenfeld residuals below are presented for a multivariable Cox regression model fit to a simulated dataset with 27 covariates.

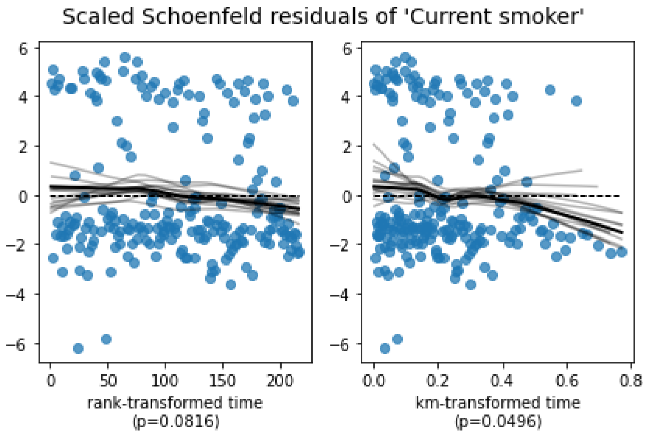


Figure 5-7 Stage I Model - Scaled Scoenfeld residuals of Current smoker variable.

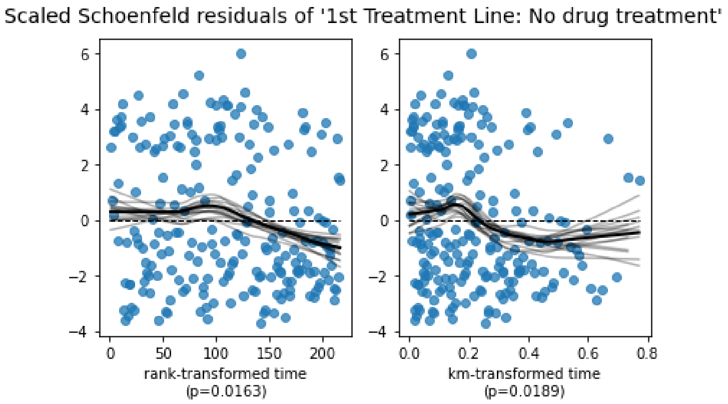


Figure 5-8 Stage I Model - Scaled Scoenfeld residuals of First treatment line- No drug treatment variable.

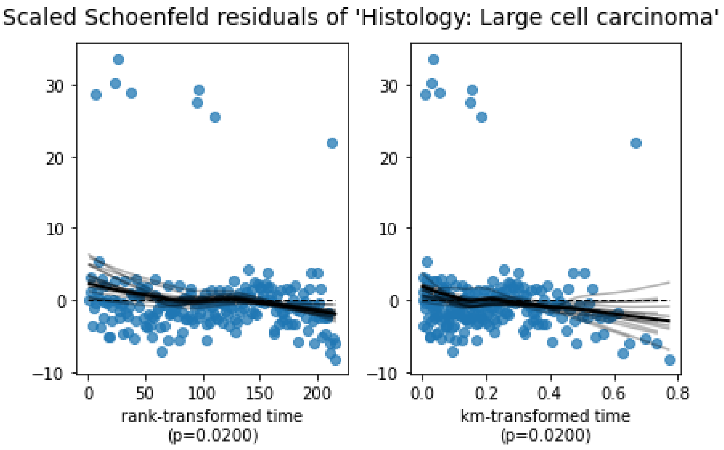


Figure 5-9 Stage I Model - Scaled Scoenfeld residuals of Histology: Large cell carcinoma variable.

5.5 STAGE II MODEL

This model included 707 patients (number of observations) and 248 deaths (events observed).

It was tested with 24 attributes (after the variables binarization) and with the baseline presented in table 5.6.

The treatment line CT+IO was removed in this model, as no patients diagnosed with stage II did drug treatments. The drug treatment designated as 'others' and 'undifferentiated' was also removed as there were just a few patients.

Table 5-6 - Stage II Model - Baseline profile.

Attribute	**Baseline profile**
Gender	Male
Age group	]45,70]
Number of comorbidities	Group 2 – [1,3]
Smoking habit	Former smoker
Patient previous cancer	No previous cancer
Family previous cancer	No previous family cancer
Histology	Adenocarcinoma
1st Treatment line	CT
ALK	Not tested
EGFR	Not tested
Surgery	No surgery (In fact, the most common patients profile with stage II, did the surgery, but in order to comply with the previous pattern, we maintain no surgery performed as the baseline – That way, we can see the risk of the patients who had surgery compared to the ones who did not).
Radiotherapy	No radiotherapy

5.5.1 RESULTS

The model results are presented in Table 5.7., along with its interpretation.

Table 5-7 Cox Model Results- Stage II model (24 attributes).

Covariante	**HR for death**	**p-value**	**Interpretation**
	(95% CI)		
Gender- Female	**0.89** (0.58-1.35)	**0.58**	Females have less 11% risk than males, considering all the other covariates the same.
Smoker habit			
Current smoker	**0.95 (0.71-1.29)**	**0.75**	Current smokers have less 5% risk than former smokers, considering all the other covariates.
Never smoker	**0.84 (0.43-1.61)**	**0.59**	Never smokers have less 16% risk than former smokers, considering all the other covariates the same.
Comorbidities			
No comorbidities	**0.88 (0.63-1.22)**	**0.44**	Patients with no comorbidities have less 12% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
+ 4 comorbidities	**0.68 (0.45-1.03)**	**0.07**	Patients with more than four (4) comorbidities have less 32% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
Age			
<45	**1.52 (0.68-3.38)**	**0.31**	Patients with <=45 years old have more 52% risk than patients with 45-70 years old, considering all the other covariates the same.
>70	**1.21 (0.89-1.64)**	**0.23**	Patients with >70 years old have more 21% risk than patients with 45-70 years old, considering all the other covariates the same.
History			
Patient with previous cancer	**0.94 (0.68-1.30)**	**0.71**	Patients with previous cancer have less 6% risk than patients with no previous cancer history.
Patient with family history of cancer	**0.82 (0.63-1.06)**	**0.12**	Patients with a family history of cancer have 18% less risk than patients with no family history of cancer.
Histology			
Adenosquamous	**0.76 (0.32-1.77)**	**0.52**	Patients diagnosed with Adenosquamous have less 24% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Squamous	**1.59 (1.14-2.22)**	**0.01**	Patients diagnosed with Squamous have more 59% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Large cell carcinoma	**1.63 (0.83-3.18)**	**0.15**	Patients diagnosed with large cell carcinoma have a more 63% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Sarcomatoid	**3.21 (1.24-8.28)**	**0.02**	Patients diagnosed with Sarcomatoid have more 221% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.

Neuroendocrine large cell carcinoma	**1.28 (0.31-5.35)**	**0.73**	Patients diagnosed with Neuroendocrine large cell carcinoma have a more 28% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
1s treatment line			
IO	**0.47 (0.06-3.48)**	**0.46**	Patients with IO as the first treatment line have less 53% risk than patients with CT as the first treatment line, considering all the other covariates the same.
No drug Treatment	**1.73 (1.21-2.48)**	**<0.005**	Patients with no drug treatment in the first treatment line havemore 73% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+RT	**1.24 (0.80-1.92)**	**0.33**	Patients with CT+RT as the first treatment line have more 24% risk than patients with CT as the first treatment line, considering all the other covariates the same.
TKI	**1.68 (0.47-5.95)**	**0.42**	Patients with TKI as the first treatment line have more 68% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Surgery	**0.45 (0.32-0.64)**	**<0.005**	Patients who did surgery have less 55% risk than patients who did not, considering all the other covariates the same.
Radiotherapy	**1.33 (0.98-1.81)**	**0.07**	Patients who did radiotherapy have more 33% risk than patients who did not, considering all the other covariates the same.
Molecular Markers			
Not mutated ALK	**1.12 (0.71-1.76)**	**0.64**	Patients without ALK mutation have more 12% risk than non-tested patients, considering all the other covariates the same.
Mutated ALK	**0.70 (0.09-5.59)**	**0.74**	Patients with ALK mutation have less 30% risk than nontested patients, considering all the other covariates the same.
Not mutated EGFR	**1.54 (0.99-2.42)**	**0.06**	Patients without EGFR mutation have 54% more risk than non-tested patients, considering all the other covariates the same.
Mutated EGFR	**1.00 (0.45-2.22)**	**1.00**	Patients with EGFR mutation have the same risk as nonmutated patients, considering all the other covariates the same.

Upper and lower bounds for each coefficient can also be seen visually in Figure 5.10.
The model has C**oncordance** of 0.70, **Partial AIC** of 2744.09 and **a log-likelihood ratio** of 105.89 on 24 df.

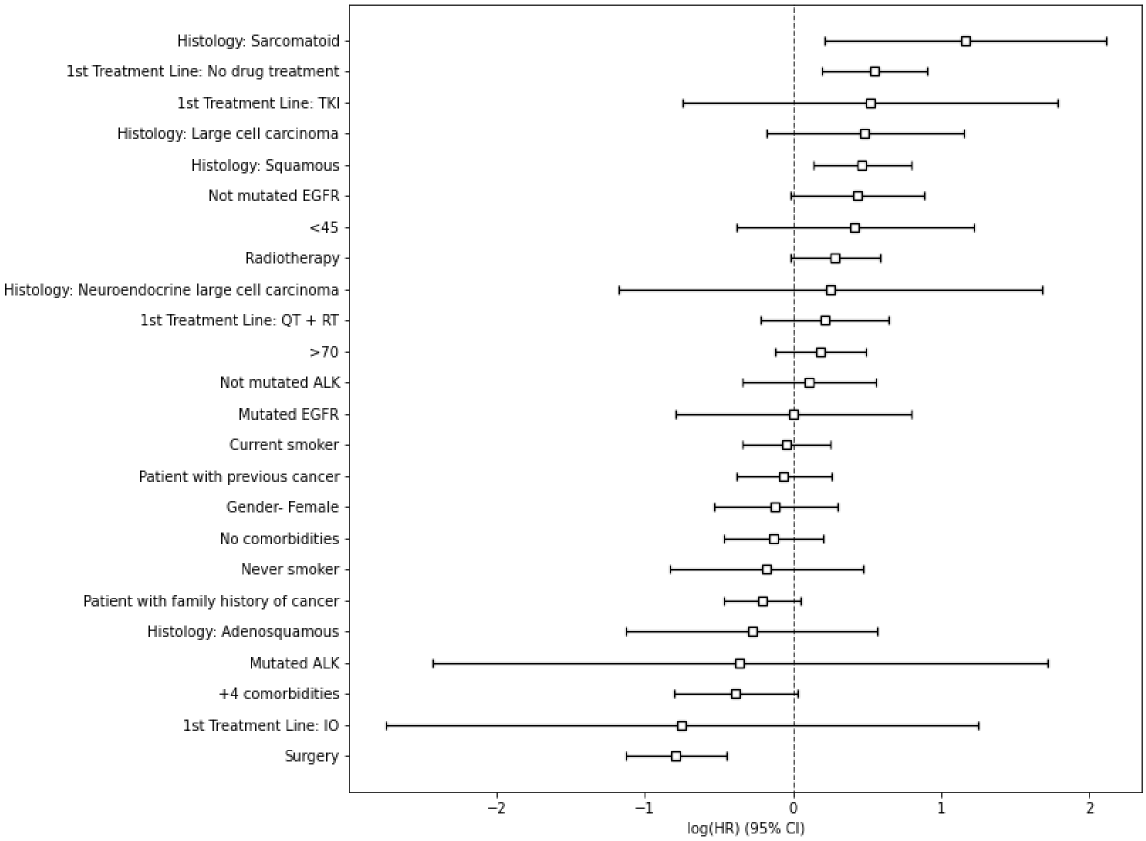


Figure 5-10 - Cox Model Results- Stage II model (24 attributes).

5.5.2 TESTING AND INTERPRETING ASSUMPTIONS

In this model, four (4) variables failed the non-proportional test. The (scaled) Schoenfeld residuals below are presented for a multivariable Cox regression model fit to a simulated dataset with 24 covariates.

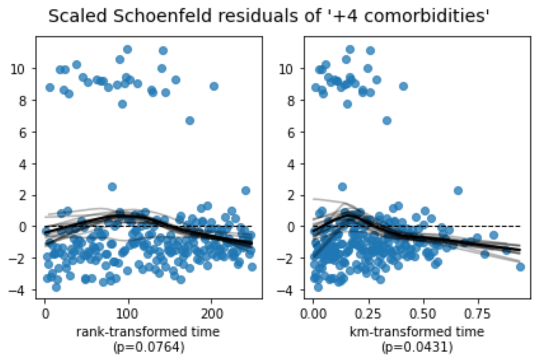


Figure 5-12 Stage II Model - Scaled Scoenfeld residuals more than 4 comorbidities variable.

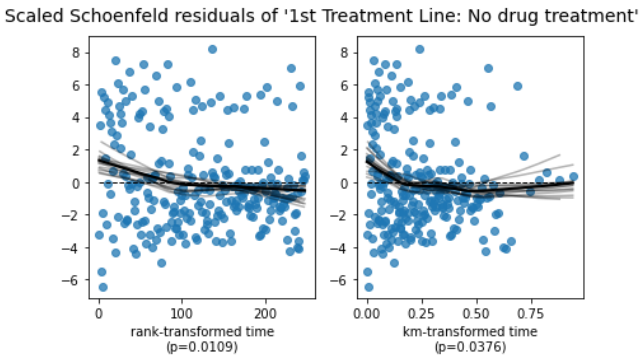


Figure 5-11 Stage II Model - Scaled Scoenfeld residuals of First treatment line- No drug treatment variable.

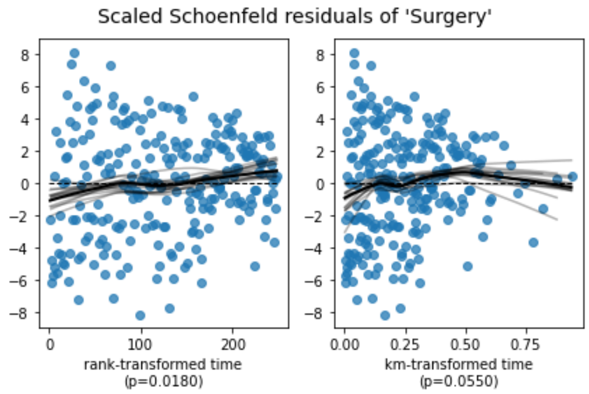


Figure 5-14 Stage II Model - Scaled Scoenfeld residuals of Surgery variable.

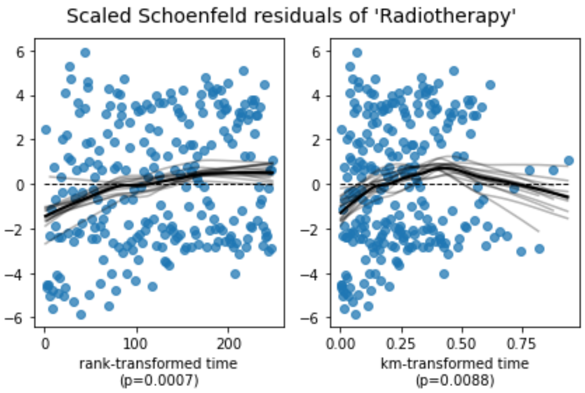


Figure 5-13 Stage II Model - Scaled Scoenfeld residuals of Radiotherapy variable.

5.5 STAGE III MODEL

This model included 2435 patients (number of observations) and 1183 deaths (events observed). In this model, the variable stage was removed as it only concerns the Stage III diagnosis. It was tested with 27 attributes (after the variables binarization) and with the baseline presented in table 5.8.

Table 5-8 - Stage III Model - Baseline profile.

Attribute	**Baseline profile**
Gender	Male
Age group	]45,70]
Number of comorbidities	Group 2 – [1,3]
Smoking habit	Former smoker
Patient previous cancer	No previous cancer
Family previous cancer	No previous family cancer
Histology	Adenocarcinoma
1st Treatment line	CT
ALK	Not tested
EGFR	Not tested
Surgery	No surgery
Radiotherapy	No radiotherapy (In fact, the most common patient's profile with stage III, had radiotherapy, but to comply with the previous pattern, we maintain no radiotherapy performed as the baseline).

5.5.1 RESULTS

The model results are presented in Table 5.9., along with its interpretation.

Table 5-9 Cox Model Results- Stage III model (27 attributes).

Covariante	HR for death	p-value	Interpretation
	(95% CI)		
Gender- Female	0.92 (0.78-1.08)	0.32	Females have less 8% risk than males, considering all the other covariates the same.
Smoker habit			
Current smoker	1.14 (1.01-1.29)	0.04	Current smokers have more 14% risk than former smokers, considering all the other covariates.
Never smoker	0.88 (0.67-1.16)	0.36	Never smokers have less 12% risk than former smokers, considering all the other covariates the same.
Comorbidities			
No comorbidities	0.92 (0.79-1.06)	0.23	Patients with no comorbidities have less 8% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
+ 4 comorbidities	1.09 (0.91-1.30)	0.36	Patients with more than four (4) comorbidities have a more 9% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
Age			
<45	0.67 (0.44-1.02)	0.06	Patients with <=45 years old have less 33% risk than patients with 40-65 years old, considering all the other covariates the same.
>70	1.30 (1.14-1.49)	<0.005	Patients with >70 years old have more 30% risk than patients with 40-65 years old, considering all the other covariates the same.
History			
Patient with previous cancer	1.07 (0.92-1.25)	0.37	Patients with previous cancer have more 7% risk than patients with no previous cancer history.
Patient with family history of cancer	1.04 (0.92-1.17)	0.53	Patients with a family history of cancer have more 4% risk than patients with no family history of cancer.
Histology			
Adenosquamous	1.03 (0.67-1.59)	0.88	Patients diagnosed with Adenosquamous have more 3% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Squamous	1.16 (1.00-1.35)	0.05	Patients diagnosed with Squamous have a more 16% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Large cell carcinoma	1.25 (0.90-1.74)	0.17	Patients diagnosed with Large cell carcinoma have a more 25% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Sarcomatoid	0.85 (0.27-2.67)	0.78	Patients diagnosed with Sarcomatoid have a less 15% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Undifferentiated	1.04 (0.72-1.50)	0.83	Patients diagnosed with Undifferentiated have a more 4% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.

Neuroendocrine large cell carcinoma	**1.55 (0.95-2.54)**	**0.08**	Patients diagnosed with Neuroendocrine large cell carcinoma have a more 55% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
1s treatment line			
IO	**1.19 (0.65-2.18)**	**0.57**	Patients with IO as the first treatment line have more 19% risk than patients with QT as the first treatment line, considering all the other covariates the same.
No drug Treatment	**2.25 (1.82-2.78)**	**<0.005**	Patients with no drug treatment in the first treatment line have more 125% risk than patients with QT as the first treatment line, considering all the other covariates the same.
Other Drugs	**2.17 (0.81-5.84)**	**0.12**	Patients with 'others' as the first treatment line have more 117% risk than patients with QT as the first treatment line, considering all the other covariates the same.
QT+IO	**1.29 (0.48-3.45)**	**0.62**	Patients with QT+IO as the first treatment line have more 29% risk than patients with QT as the first treatment line, considering all the other covariates the same.
QT+RT	**0.83 (0.72-0.95)**	**0.01**	Patients with QT+RT as the first treatment line have less 17% risk than patients with QT as the first treatment line, considering all the other covariates the same.
TKI	**0.81 (0.46-1.44)**	**0.48**	Patients with TKI as the first treatment line have less 19% risk than patients with QT as the first treatment line, considering all the other covariates the same.
Surgery	**0.41 (0.35-0.47)**	**<0.005**	Patients who did surgery have less 59% risk than patients who did not, considering all the other covariates the same.
Radiotherapy	**0.70 (0.60-0.80)**	**<0.005**	Patients who did radiotherapy have less 30% risk than patients who did not, considering all the other covariates the same.
Molecular Markers			
Not mutated ALK	**1.09 (0.88-1.35)**	**0.42**	Patients without ALK mutation have more 9% risk than non-tested patients, considering all the other covariates the same.
Mutated ALK	**0.62 (0.35-1.10)**	**0.10**	Patients with ALK mutation have less 38% risk than non-tested patients, considering all the other covariates the same.
Not mutated EGFR	**1.03 (0.84-1.28)**	**0.75**	Patients without EGFR mutation have more 3% risk than nontested patients, considering all the other covariates the same.
Mutated EGFR	**0.89 (0.63-1.26)**	**0.50**	Patients with EGFR mutation have less 11% risk than nonmutated patients, considering all the other covariates the same.

Upper and lower bounds for each coefficient can also be seen visually in Figure 5.10.

The model has C**oncordance** of 0.68, **Partial AIC** of 16081.12 and **a log-likelihood ratio** of 352.48 on 27 df.

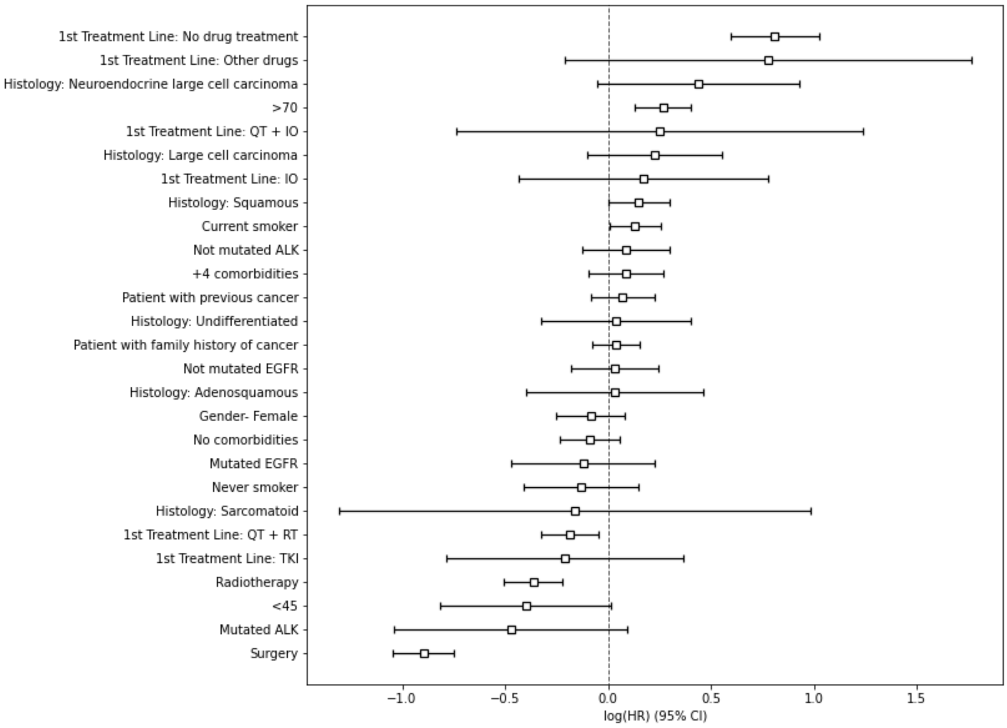


Figure 5-15 Cox Model Results- Stage III model (27 attributes).

5.5.2 TESTING AND INTERPRETING ASSUMPTIONS

In this model, five (5) variables failed the non-proportional test. The (scaled) Schoenfeld residuals below are presented for a multivariable Cox regression model fit to a simulated dataset with 27 covariates.

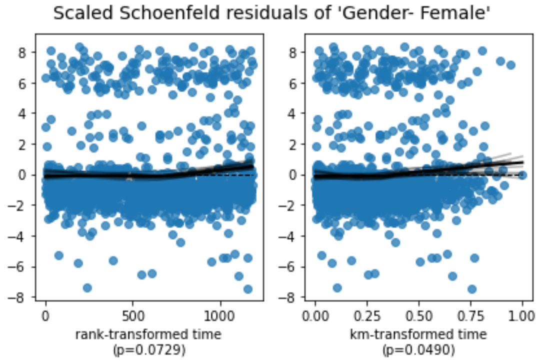


Figure 5-17 Stage III Model - Scaled Scoenfeld residuals of Gender Female variable.

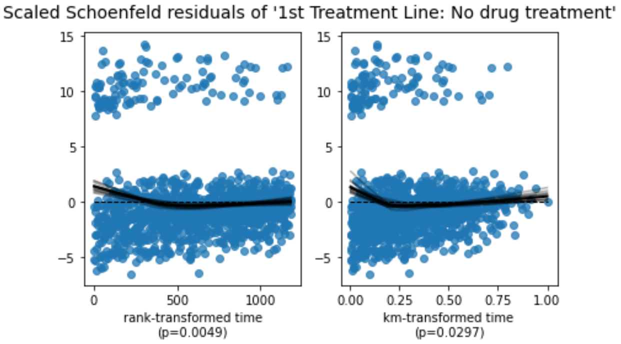


Figure 5-16 Stage III Model - Scaled Scoenfeld residuals of First treatment line- No drug treatment variable.

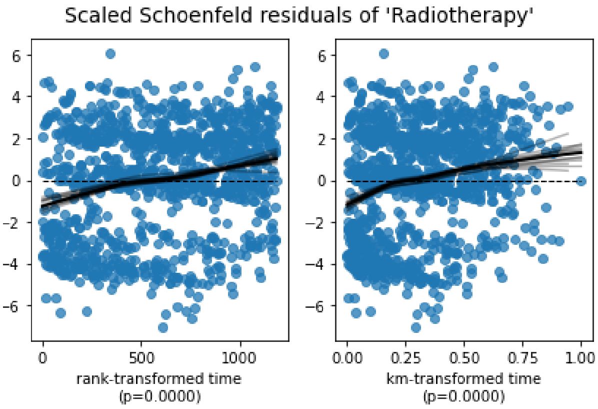


Figure 5-19 Stage III Model - Scaled Scoenfeld residuals of Radiotherapy variable.

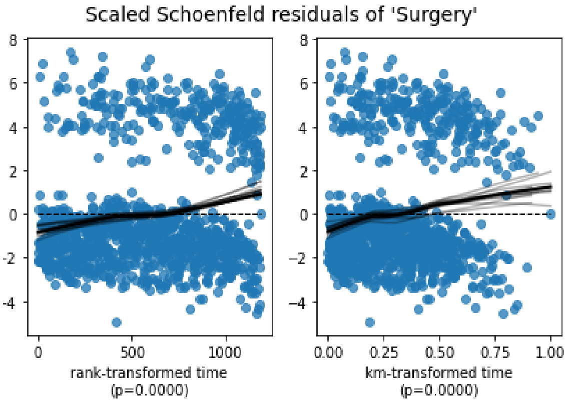


Figure 5-18 Stage III Model - Scaled Scoenfeld residuals of Surgery variable.

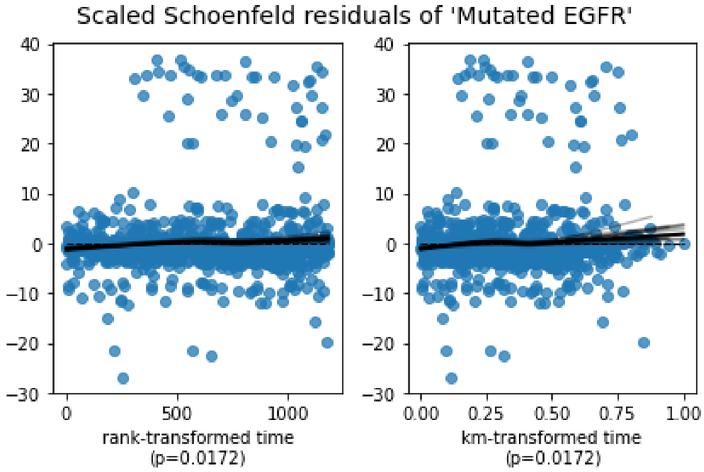


Figure 5-20 Stage III Model - Scaled Scoenfeld residuals of Mutated EGFR variable.

5.6 STAGE IV MODEL

This model included 4617 patients (number of observations) and 3034 deaths (events observed).

In this model, the variable stage was removed as it only concerns the Stage IV diagnosis. It was tested with 27 attributes (after the variables binarization) and with the baseline presented in table 5.10.

Table 5-10 - Stage IV Model - Baseline profile.

Attribute	**Baseline profile**
Gender	Male
Age group	]45,70]
Number of comorbidities	Group 2 – [1,3]
Smoking habit	Former smoker
Patient previous cancer	No previous cancer
Family previous cancer	No previous family cancer
Histology	Adenocarcinoma
1st Treatment line	CT
ALK	Not tested (In fact, the most common patient's profile, did the test for the ALK mutation and tested negative, but in order to comply with the previous models, we maintain 'not tested' as the baseline).
EGFR	Not tested (In fact, the most common patient's profile, did the test for the EGFR mutation and tested negative, but in order to comply with the ALK mutation, we maintain 'not tested' as the baseline).
Surgery	No surgery
Radiotherapy	No radiotherapy

5.6.1 RESULTS

The model results are presented in Table 5.11., along with its interpretation.

Table 5-11 Cox Model Results- Stage IV model (27 attributes).

Covariante	HR for death	p-value	Interpretation
	(95% CI)		
Gender- Female	**0.87 (0.79-0.95)**	**<0.005**	Females have less 13% risk than males, considering all the other covariates the same.
Smoker habit			
Current smoker	**1.16 (1.07-1.26)**	**<0.005**	Current smokers have a more 16% risk than former smokers, considering all the other covariates.
Never smoker	**0.78 (0.69-0.89)**	**<0.005**	Never smokers have less 22% risk than former smokers, considering all the other covariates the same.
Comorbidities			
No comorbidities	**0.94 (0.87-1.03)**	**0.18**	Patients with no comorbidities have less 6% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
+ 4 comorbidities	**1.02 (0.91-1.16)**	**0.70**	Patients with more than four (4) comorbidities have a more 2% risk than patients with 1-3 comorbidities, considering all the other covariates the same.
Age			
<45	**1.06 (0.88-1.27)**	**0.56**	Patients with <=45 years old have more 6% risk than patients with 40-65 years old, considering all the other covariates the same.
>70	**1.11 (1.02-1.21)**	**0.02**	Patients with >70 years old have more 11% risk than patients with 40-65 years old, considering all the other covariates the same.
History			
Patient with previous cancer	**0.93 (0.83-1.03)**	**0.15**	Patients with previous cancer have more 7% risk than patients with no previous cancer history.
Patient with family history of cancer	**1.00 (0.93-1.07)**	**0.96**	Patients with a family history of cancer have the same risk as patients with no family history of cancer.
Histology			
Adenosquamous	**0.88 (0.64-1.19)**	**0.40**	Patients diagnosed with Adenosquamous have less 12% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Squamous	**0.91 (0.81-1.03)**	**0.13**	Patients diagnosed with Squamous have a less 9% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Large cell carcinoma	**1.21 (1.00-1.46)**	**0.05**	Patients diagnosed with Large cell carcinoma have a more 21% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.

Sarcomatoid	**1.24 (0.59-2.62)**	**0.58**	Patients diagnosed with Sarcomatoid have a more 24% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Undifferentiated	**1.26 (1.04-1.53)**	**0.02**	Patients diagnosed with Undifferentiated have a more 26% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
Neuroendocrine large cell carcinoma	**1.37 (1.00-1.89)**	**0.05**	Patients diagnosed with Neuroendocrine large cell carcinoma have a more 37% risk than patients diagnosed with Adenocarcinoma, considering all the other covariates the same.
1s treatment line			
IO	**0.78 (0.66-0.93)**	**0.01**	Patients with IO as the first treatment line have less 22% risk than patients with QT as the first treatment line, considering all the other covariates the same.
No drug Treatment	**3.99 (3.58-4.45)**	**<0.005**	Patients with no drug treatment in the first treatment line have more 299% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Other Drugs	**0.70 (0.52-0.96)**	**0.03**	Patients with 'others' as the first treatment line have less 30% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+IO	**0.90 (0.70-1.15)**	**0.39**	Patients with CT+IO as the first treatment line have less 10% risk than patients with CT as the first treatment line, considering all the other covariates the same.
CT+RT	**0.64 (0.49-0.84)**	**<0.005**	Patients with CT+RT as the first treatment line have less 36% risk than patients with CT as the first treatment line, considering all the other covariates the same.
TKI	**0.78 (0.66-0.93)**	**0.01**	Patients with TKI as the first treatment line have less 22% risk than patients with CT as the first treatment line, considering all the other covariates the same.
Surgery	**0.38 (0.32-0.45)**	**<0.005**	Patients who did surgery have less 62% risk than patients who did not, considering all the other covariates the same.
Radiotherapy	**0.96 (0.89-1.03)**	**0.28**	Patients who did radiotherapy have less 4% risk than patients who did not, considering all the other covariates the same.
Molecular Markers			
Not mutated ALK	**0.86 (0.77-0.95)**	**<0.005**	Patients without ALK mutation have less 14% risk than non-tested patients, considering all the other covariates the same.
Mutated ALK	**0.55 (0.43-0.72)**	**<0.005**	Patients with ALK mutation have less 45% risk than nontested patients, considering all the other covariates the same.
Not mutated EGFR	**0.93 (0.82-1.05)**	**0.22**	Patients without EGFR mutation have less 7% risk than non-tested patients, considering all the other covariates the same.
Mutated EGFR	**0.80 (0.67-0.95)**	**0.01**	Patients with EGFR mutation have less 20% risk than nonmutated patients, considering all the other covariates the same.

Upper and lower bounds for each coefficient can also be seen visually in Figure 5.21.

The model has C**oncordance** of 0.68, **Partial AIC** of 44973.56 and **a log-likelihood ratio** of 1059.00 on 27 df.

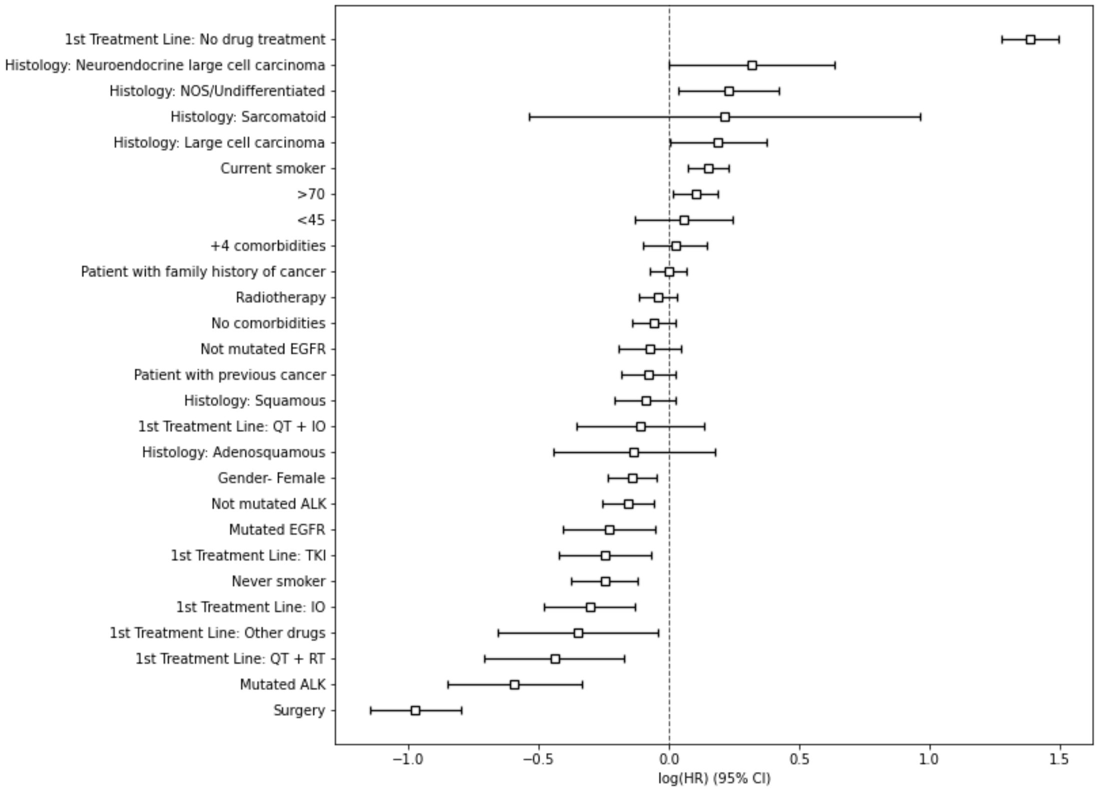


Figure 5-21 Cox Model Results- Stage IV model (27 attributes).

5.6.2 TESTING AND INTERPRETING ASSUMPTIONS

In this model, ten (10) variables failed the non-proportional test. The (scaled) Schoenfeld residuals below are presented for a multivariable Cox regression model fit to a simulated dataset with 27 covariates.

Unlike the previous stage individual models, a lot more variables failed the Stage IV model non-proportional test.

Figure 5.22. illustrates the Schoenfeld residuals of the variable Radiotherapy, and it is clear that they are incompatible with the proportional hazards assumption, which was expected since the Kaplan-Meier method, Figure 4.22.

Figure 5.23. illustrates the Schoenfeld residuals of the variable Mutated EGFR, and we can see that it starts to fail at the end of the time, which can be justified as, at that time, the number of patients under observation is low.

Finally, in figure 5.24. the Schoenfeld residuals of the variable Histology: Large cell carcinoma show minor changes, which we know is possible to happen based on the immense number of variables in the model, but also since the variable histology have seven (7) possible values and the survival curves crossed each other in the Kaplan-Meier method, Section B- Attachment 2.

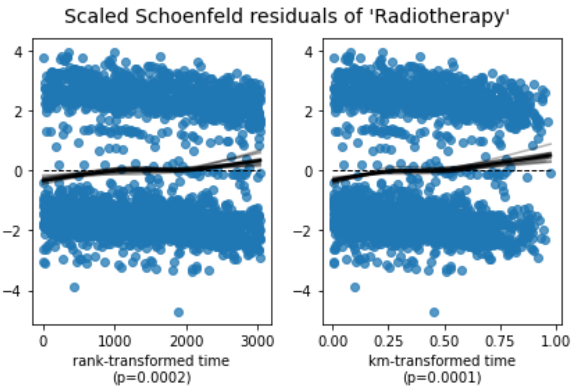


Figure 5-23 - Stage IV Model - Scaled Scoenfeld residuals of Radioterapy variable.

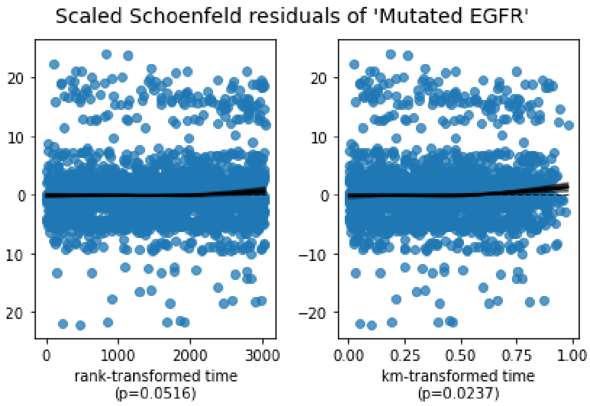


Figure 5-22 - Stage IV Model - Scaled Scoenfeld residuals of Mutated EGFR variable.

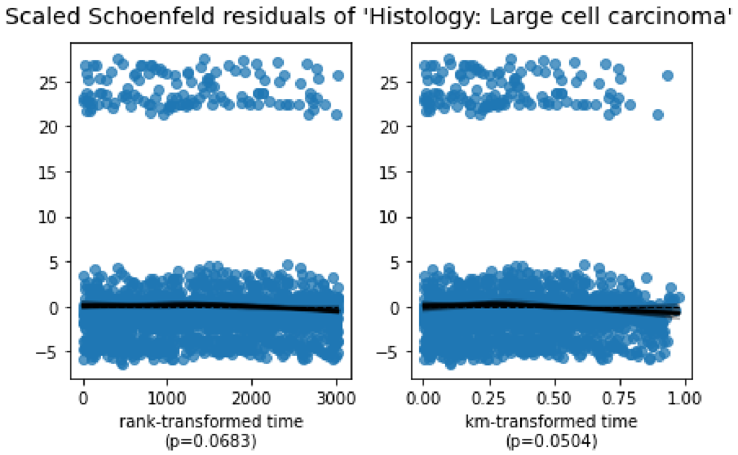


Figure 5-24 Stage IV Model - Scaled Scoenfeld residuals of Histology: Large cell carcinoma variable.

It should be noted that the results presented here are innovative and are already allowing the Spanish physician community to visualise and understand the survival patterns of patients with lung cancer and what is the impact of several explanatory variables on survival.

CHAPTER 6

CONCLUSIONS AND FUTURE WORK

The conclusions section reviews the work developed and analyses the results and challenges since the beginning of the dissertation.

The Future work section identifies the optimizations that could be done in the models, the following steps within the project and evolutions of the work developed.

6.1. CONCLUSIONS

From a technical point of view, knowing and understanding the industry and architecture of the systems where every work is inserted is fundamental. There would be no way of evolution and innovation if the big picture where all the work is inserted would not be known. As a matter of fact, the word 'work' could be replaced by data, as it is the core of every possible knowledge source.

With this in mind, the first step of every data project is learning the applications domain, which includes studying all underlying areas—in this case, understanding the healthcare industry, the core systems, the main processes and standards that the analysed data passed through.

In that way, this dissertation started with an overview of clinical data and medical coding, followed by the healthcare information systems. Here, particular attention was given to the EHR, which was the source of the data used in this dissertation.

The complexity of the medical systems and the interaction between them is enormous. Each topic could be treated and have content to be a dissertation by itself; that is why some topics are only approached at a high level. For example, the security and access controls lines were mentioned but not too detailed, as they are crucial subjects for information systems and services.

It was then studied the different ML algorithms categories, in which it was stated that Supervised learning would be the focus, as the regression models used in this dissertation are a part of this group.

In the next phase, it was necessary to interpret the data from a clinical point of view, and it is at this time that the Data science loop starts.

Before getting into the different phases of the loop which this dissertation passed through, one should emphasize the importance of the interaction and follow-up with the subject experts. In the healthcare case, and here, specifically in the oncology area, the interpretation of the data is highly complex. More than anyone, doctors and specialists handle diagnosis and treatments for years, and without them, it would be impossible to create value. As human beings, our interpretations are conditioned by our experiences and studies and change continuously over time.

So, one of the main challenges, and one of the most interesting, was to see this evolution on both sides, clinical experts/doctors and engineers. As a part of the engineer's team, trying to understand the clinical meaning and interaction between variables, mapping the process of care and the disease itself, it would be unthinkable without the support of the other parts.

That said, it follows the data engineering and pre-processing phase, which was by far the more time-consuming phase.

As previously said, every project that requires real-time analysis starts from a static dataset and its analysis. Several datasets were provided before the final one presented in this dissertation.

The first phase of the data science loop is data cleaning, which required special attention as the dataset was raw. It was the process of detecting and correcting or removing incorrect data entries, such as missing values, outliers, inaccurate values, and duplicates elimination.

Following the second and third phases, the Analysis and Sample, and Feature engineering, which were the phases that required more returning in order to optimize the results and, at the same time, maintaining the clinical meaning. Regarding the features selection, the first filter was clinical relevance, not only based on the other oncology studies; indeed, there are attributes scientifically proved as significant, but also the need to understand the remaining ones, such as comorbidities and mutations.

The fourth and fifth phases of the data science loop, the Model building and Hyperparameter optimization, respectively, were also performed more than once. Each time that a new variable was added or modified, a new model was generated.

The Cox multivariable proportional hazard model was chosen from the beginning because it fills the requirements and goals of this dissertation, specifically the relationship between the risk of an event over time and the features of the sample.

In this phase, five models were built—the first one, including the different diagnosis stages and the remaining models individually by stage at diagnosis.

Finally, the Evaluation and Comparison phases assess the fitted model results and compare the model’s performance.

Assessing the fitted model results included analysing the statistical significance of each covariate, the effect of each covariate on the hazard ratio and analyse how confident are the coefficients estimated.

The “All stages model” allowed the analysis of the entire group of non-small cell lung cancer patients, enable the analysis of the survival and risk of the patients with a different diagnosis. It is clinically established the differences between initial and advanced stages at diagnosis. The model's results show precisely that, so in a statistical point of view, the models state what was supposed. This conclusion can either be seen in the All stages model (in the survival curves of the variables Stage) or by analysing the overall performance of the individual stage models.

Also, it is possible to correlate the failure of some variables in the assumptions tests of the “All stages model” with the failure of those same variables in the individual stages model.

Indeed, these phases are part of a loop, so even though they are presented sequenced in this document, it was not developed that way, and the presented steps, decisions and results are the final ones.

Fitted survival models typically have a concordance index between 0.55 and 0.75 [46], and the concordance indexes obtained were 0.73 and 0.68, respectively the best and the worst. Although, as stated, they were not statistically optimized to keep all the features and have risks associated with every variable.

Note that all the included variables were transformed and optimized, so having that in mind, to improve the model's results, the approach would be to remove the not significant ones.

Although, before this approach, there is another critical point to be considered regarding the baselines defined.

The baselines were defined as the most common patient's profile so that all the other variables would describe the risk comparing with it. The baselines were defined the same for all the models tested. So, regarding the All stages model, the baseline is ‘correct’; in other words, all the attributes are the most common patient’s profile. For the remaining models, some attributes defined as the baseline are not the most common characteristics.

It was selected this way to comply and be possible to analyse the variable’s results between models. This analysis is important and relevant, but the optimization of these models passes, first of all, to completely separate the diagnosis analysis from the predictive analysis. That way, the prediction would exclusively have into consideration the diagnosis.

For either one of these cases, the work developed throughout this dissertation is innovative and relevant as the objective was indeed the identification of factors/characteristics; Risk stratification, and predict the best models of follow-up.

6.2. FUTURE WORK

This section presents the optimizations that could be done in the models, the following steps within the project, and evolutions of the work developed.

So, within the models, some optimizations were already mentioned, but to summarize, to optimize mathematically the model, the approach would be to remove the insignificant features and define the baseline of the individual stage model, as the most common patient's profile. Also, some improvements can be done regarding the patterns, for example, if instead of the general groups of diagnosis (4) that were used in these models, would be used the more specific diagnosis groups (16), we would have another hand of patterns along with the treatment lines performed.

Within the project, the next steps, which are already being developed by Holos S.A. would be to develop an interface where the doctors/clinicians could have access to the statistics and have real-time data to support their decisions.

Indeed, the objective is that the analysis remains autonomous after being deployed, but also having in mind that the data is continuously being updated and requires a certain follow-up. It may and probably will be necessary to perform changes in the models, as new characteristics will be added, new treatments will be created, and as a result of this, new patterns will appear.

Finally, regarding the Cox model itself, other complex models are proved to have better results. So, besides the optimizations mentioned, the evolution of the Cox model passes by extending the Cox proportional hazards model with neural networks [47]. Also, the use of Deep Neural Networks for survival Analysis based on a multi-task framework has proved better results in terms of the model's performance [54].

BIBLIOGRAPHY

[1] F. BRAY, J. FERLAY, I. SOERJOMATARAM, R. L. SIEGEL, L. A. TORRE, AND 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, NO. 6, PP. 394–424, NOV. 2018.

[2] R. L. Siegel, K. D. Miller, and A. Jemal, 'Cancer statistics, 2020', CA A Cancer J Clin, vol. 70, no. 1, pp. 7–30, Jan. 2020.

[3] R. L. Siegel, K. D. Miller, and A. Jemal, 'Cancer statistics, 2019', CA A Cancer J Clin, vol. 69, no. 1, pp. 7–34, Jan. 2019.

[4] Panesar, Arjun. Machine Learning and AI for Healthcare: Big Data for Improved Health Outcomes. 2nd ed. 2021.

[5] K. D. Miller et al., 'Cancer treatment and survivorship statistics, 2019', CA A Cancer J Clin, vol. 69, no. 5, pp. 363–385, Sep. 2019.

[6] SEER CANCER STATISTICS REVIEW 1975-2016: Introduction, Available at: <https://seer.cancer.gov/archive/csr/1975_2016/results_figure/sect_01_intro1_7pgs.pdf>. [Accessed on 16/03/2020]

[7] S. Huotari, M. Jauhiainen, J. Tolonen, and A. Värri, 'Foundational Curricula: Cluster 8: Data Module 15: Data Analytics, Modeling and Reporting Unit 1: Data Analytics FC-C8M15U', p. 18, 2020.

[8] CEN, 2020, Available at: <https://www.cen.eu/about/Pages/default.aspx>. (visited on 21/09/2020)

[9] ISO, 2020, Available at: <https://www.iso.org/about-us.html>. [Accessed on 21/09/2020]

[10] P. Kubben, M. Dumontier, and A. Dekker, Eds., Fundamentals of Clinical Data Science. Cham: Springer International Publishing, 2019.

[11] S. Huotari, M. Jauhiainen, J. Tolonen, and A. Värri, 'Foundational Curriculum: Cluster 6: System Connectivity Module 10: Interoperability, Interfaces and Integration of eHealth Unit 4: Standards and Protocols FC-C6M10U4', p. 29, 2020.

[12] W. E. Hammond, 'The Making And Adoption Of Health Data Standards', Health Affairs, vol. 24, no. 5, pp. 1205–1213, Sep. 2005, doi: 10.1377/hlthaff.24.5.1205.

[13] Y. Yu, M. Li, L. Liu, Y. Li, and J. Wang, 'Clinical big data and deep learning: Applications, challenges, and future outlooks', Big Data Min. Anal., vol. 2, no. 4, pp. 288–305, Dec. 2019.

[14] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. W. M. van der Laak, B. van Ginneken, and C. I. Sanchez, A survey on deep learning ´ in medical image analysis, Med. Image Anal., vol. 42, pp. 60–88, 2017.

[15] G. Litjens et al., 'A survey on deep learning in medical image analysis', Medical Image Analysis, vol. 42, pp. 60–88, Dec. 2017.

[16] World Health Organization. 2020. International Classification Of Diseases (ICD) Information Sheet. [online] Available at: <https://www.who.int/classifications/icd/ICD10Volume2_en_2010.pdf> [Accessed on 21/04/2020].

[17] World Health Organization, Ed., International statistical classification of diseases and related health problems, 10th revision, 2nd edition. Geneva: World Health Organization, 2004.

[18] R. Blake, Angelique Blake, Rachelle Blake, Pauliina, Hulkkonen, Sonja Huotari, Milla Jauhiainen, Johanna Tolonen, and, and Alpo Värri, 'Introduction to Medical Coding', Module 6: Health Information Management, p. 19, 2020.

[19] F. Freitas, S. Schulz, and E. Moraes, 'Survey of current terminologies and ontologies in biology and medicine', RECIIS, vol. 3, no. 1, p. 239/249, Mar. 2009, doi: 10.3395/reciis.v3i1.239en.

[20] C. J. McDonald et al., 'LOINC, a Universal Standard for Identifying Laboratory Observations: A 5-Year Update', Clinical Chemistry, vol. 49, no. 4, pp. 624–633, Apr. 2003.

[21] Nlm.nih.gov. 2020. UMLS Metathesaurus Vocabulary Documentation. [online] Available at: <https://www.nlm.nih.gov/research/umls/sourcereleasedocs/index.html>. [Accessed on 24/04/2020].

[22] A. T. McCray, A. Burgun, and O. Bodenreider, 'Aggregating UMLS Semantic Types for Reducing Conceptual Complexity', p. 10, 2015.

[23] Lexsrv3.nlm.nih.gov. 2020. The SPECIALIST NLP Tools. [online] Available at: <https://lexsrv3.nlm.nih.gov/Specialist/Home/index.html> [Accessed on 24/04/2020].

[24] S. Huotari, M. Jauhiainen, J. Tolonen, and A. Värri, 'Introduction & Overview: Components of Health IT Systems', p. 18, 2020.

[25] Rui GOMES, Luís Velez LAPÃO, The Adoption of IT Security Standards in a Healthcare Environment, eHealth Beyond the Horizon – Get IT There S.K. Andersen et al. (Eds.) IOS Press, 2008.

[26] Knox L, Brach C, Mitchell M, Taylor E. Primary Care Practice Facilitation Curriculum (Module 26). AHRQ Publication No. 15-0060-EF, Rockville, MD: Agency for Healthcare Research and Quality; September 2015.

[27] ISO/TR 20514:2005 Health informatics — Electronic health record — Definition, scope and context, International Standards Organization, ISO/TR 2005.

[28] K. Hayrinen, K. Saranto, and P. Nykanen, 'Definition, structure, content, use and impacts of electronic health records: A review of the research literature', International Journal of Medical Informatics, vol. 77, no. 5, pp. 291–304, May 2008, doi: 10.1016/j.ijmedinf.2007.09.001.

[29] D. Kalra, CHIME, University College London, London, United Kingdom, 'Electronic Health Record Standards', p. 9, 2006.

[30] ISO 13606-1: 2019- Health Informatics- Electronic Health Record communications, International Standards Organization, ISO/TR 2019.

[31] H. Nieminen and M. Jauhiainen, Tampere University 2019, 'BMT-57106 Health Care Processes and Information Systems', p. 112.

[32] Patricia A H Williams, The Role of Standards in Medical Information Security: An Opportunity for Improvement. P. A. H. Williams , School of Computer and Information Science Edith Cowan University Joondalup, Western Australia.

[33] A. Appari and M. E. Johnson, 'Information security and privacy in healthcare: current state of research', IJIEM, vol. 6, no. 4, p. 279, 2010, doi: 10.1504/IJIEM.2010.035624.

[34] Confidentiality, Integrity, & Availability: Basics of Information Security - Smart Eye Technology. (n.d.). Smart Eye Technology. Available at: < https://smarteyetechnology.com/confidentiality-integrity-availability-basics-of-information-security/> [Accessed on 27/09/2020].

[35] R. C. Barrows and P. D. Clayton, ‘Privacy, Confidentiality, and Electronic Medical Records’, Journal of the American Medical Informatics Association, vol. 3, no. 2, pp. 139–148, Mar. 1996, doi: 10.1136/jamia.1996.96236282.

[36] IT Governance Privacy Team, Second edition, UK 2017, EU General Data Protection Regulation (GDPR), “An Implementation and Compliance Guide”.

[37] R. Ahmad, ‘A Review of CIO’s Role In Increasing Competitive Advantage’, IJET, vol. 11, no. 2, pp. 312–317, Apr. 2019, doi: 10.21817/ijet/2019/v11i2/191102053.

[38] R. Gomes and L. V. Lapão, Organizing Committee of MIE 2008, ‘The Adoption of IT Security Standards in a Healthcare Environment’, p. 7.

[39] ISO/IEC 27002:2013 Information technology -Security techniques - Code of practice for information security management, International Standards Organization, 2013.

[40] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, ‘Big data security and privacy in healthcare: A Review’, Procedia Computer Science, vol. 113, pp. 73–80, 2017, doi: 10.1016/j.procs.2017.08.292.

[41] S. Alshehri and R. K. Raj, ‘Secure Access Control for Health Information Sharing Systems’, in 2013 IEEE International Conference on Healthcare Informatics, Philadelphia, PA, USA, Sep. 2013, pp. 277–286, doi: 10.1109/ICHI.2013.40.

[42] Components Of Artificial Intelligence - How It Works? (n.d.). Blogs & Updates on Data Science, Business Analytics, AI Machine Learning. Available at: <https://www.analytixlabs.co.in/blog/components-of-artificial-intelligence/> [Accessed on 01/10/2020].

[43] Viv Bewick, Liz Cheek and Jonathan Ball, Review Statistics review 12: Survival analysis, 2004 BioMed Central Ltd, Available online http://ccforum.com/content/8/5/389.

[44] : Taylor & Francis, Ltd. on behalf of the American Statistical Association, Source: Journal of the American Statistical Association , Jun., 1988, Vol. 83, No. 402 (Jun., 1988), pp. 414-425, ‘Logistic Regression, Survival Analysis, and the Kaplan-Meier Curve’, Stable URL: https://www.jstor.org/stable/2288857.

[45] Cristina Rocha and Ana Luísa Papoila, XVII Congresso da Sociedade Portuguesa de Estatística 2009,‘Análise de Sobrevivência’.

[46] Lifelines.readthedocs.io. 2021. Survival regression - lifelines 0.26.2 documentation. [online] Available at: <https://lifelines.readthedocs.io/en/latest/Survival%20Regression.html#log-likelihood> [Accessed on 15/08/2021].

[47] Havard Kvamme, Ørnulf Borgan, Ida Scheel, ‘Time-to-Event Prediction with Neural Networks and Cox Regression’, Journal of Machine Learning Research 20 (2019), Department of Mathematics University of Oslo.

[48] Yishu Xue , Elizabeth D. Schifano, Diagnostics for the Cox model, Communications for Statistical Applications and Methods 2017, Vol. 24, No. 6, 583–604.

[49] Partial Residuals for The Proportional Hazards Regression Model David Schoenfeld Biometrika, Vol. 69, No. 1. (Apr., 1982), pp. 239-241.

[50] Package overview — pandas 1.3.3 documentation. (n.d.). pandas - Python Data Analysis Library. Available at: <https://pandas.pydata.org/docs/getting_started/overview.html> [Accessed on 15/08/2021].

[51] DevDocs — Matplotlib 3.1 documentation. (n.d.). DevDocs API Documentation. Available at: <https://devdocs.io/matplotlib~3.1/> [Accessed on 15/08/2021].

[52] Leduc C, Antoni D, Charloux A, Falcoz PE, Quoix E. Comorbidities in the management of

patients with lung cancer. Eur Respir J. 2017 Mar 29;49(3):1601721. doi: 10.1183/13993003.01721-2016. PMID: 28356370.

[53] Dutkowska AE, Antczak A. Comorbidities in lung cancer. Pneumonol Alergol Pol. 2016; 84(3): 186-92. doi: 10.5603/PiAP.2016.0022. PMID: 27238182.

[54] Stephane Fotso, Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework' San Francisco, CA, U.S.A.,doi: arXiv:1801.05512 [stat.ML]

ATTACHMENTS

SECTION A

This attachment section includes the original values of some attributes studied.

Section A- Attachment 1 - Diagnosis table- Original values.

	Variable	Values
Diagnosis	Stage at diagnosis	24, I 1, IA 25, IA1 26, IA2 28, IA3 2, IB 27, II 3, IIA 4, IIB 15, III 5, IIIA 6, IIIB 18, IIIC 7, IV 20, IVA 21, IVB 30, Limited 31, Extended 888, Others -1, -
	Histology	0, Adenocarcinoma 1, Adenosquamous 3, Large cell carcinoma 6, Small cell lung cancer (microcytic) 7, Neuroendocrine large cell carcinoma 11, Thymic carcinoma 2, Squamous 9, Mesothelioma 5, Undifferentiated 4, Sarcomatoid 10, Thymoma 8, Carcinoidtumour 12,Others -1, -
	Were molecular markers analyses performed at diagnoses?	1, Yes 0, No

	EGFR performed	1, Yes
	ALK performed	1, Yes
	Result EGFR: Negative	1, Yes
	Result EGFR: T790M	1, Yes
	Result EGFR: T790	1, Yes
	Result EGFR: Exon19	1, Yes
	Result EGFR: Exon21	1, Yes
	Result EGFR: NOS	1, Yes
	Result EGFR: Exon 20	1, Yes
	Result EGFR: Others	1, Yes
	Result ALK IHQ	0, Negative 1, Positive
	Result ALK FISH	1, Translocated 0, Non-Translocated
	Result ALK RNA	1, Detected 0, Non detected

Section A- Attachment 2 - Comorbidities table - Original values.

	Variable	Values
Comorbidities	No comorbidities	1, Yes
	Comorbidity: Asthma	1, Yes
	Comorbidity: Cardiopathy	1, Yes
	Comorbidity: Diabetes Mellitus (DM)	1, Yes
	Comorbidity: Dyslipidemia	1, Yes
	Comorbidity: Chronic obstructive pulmonary disease	1, Yes
	Comorbidity: Alcoholism/Ex Alcoholism	1, Yes
	Comorbidity: Hepatitis	1, Yes
	Comorbidity: Hypercholesterolemia	1, Yes
	Comorbidity: HT	1, Yes
	Comorbidity: Renal disease	1, Yes
	Comorbidity: Obesity	1, Yes
	Comorbidity: Depressive syndrome / Anxiety	1, Yes
	Comorbidity: Tuberculosis	1, Yes

	Comorbidity: Vascular disease	1, Yes
	Comorbidity: Others	1, Yes
	Comorbidity: Liver disease	1, Yes
	Comorbidity: Gastrointestinal	1, Yes
	Comorbidity; Neurodegenerative disorder	1, Yes
	Comorbidity: Benign prostatic Hyperplasia	1, Yes
	Comorbidity: Obstructive sleep Apnea	1, Yes

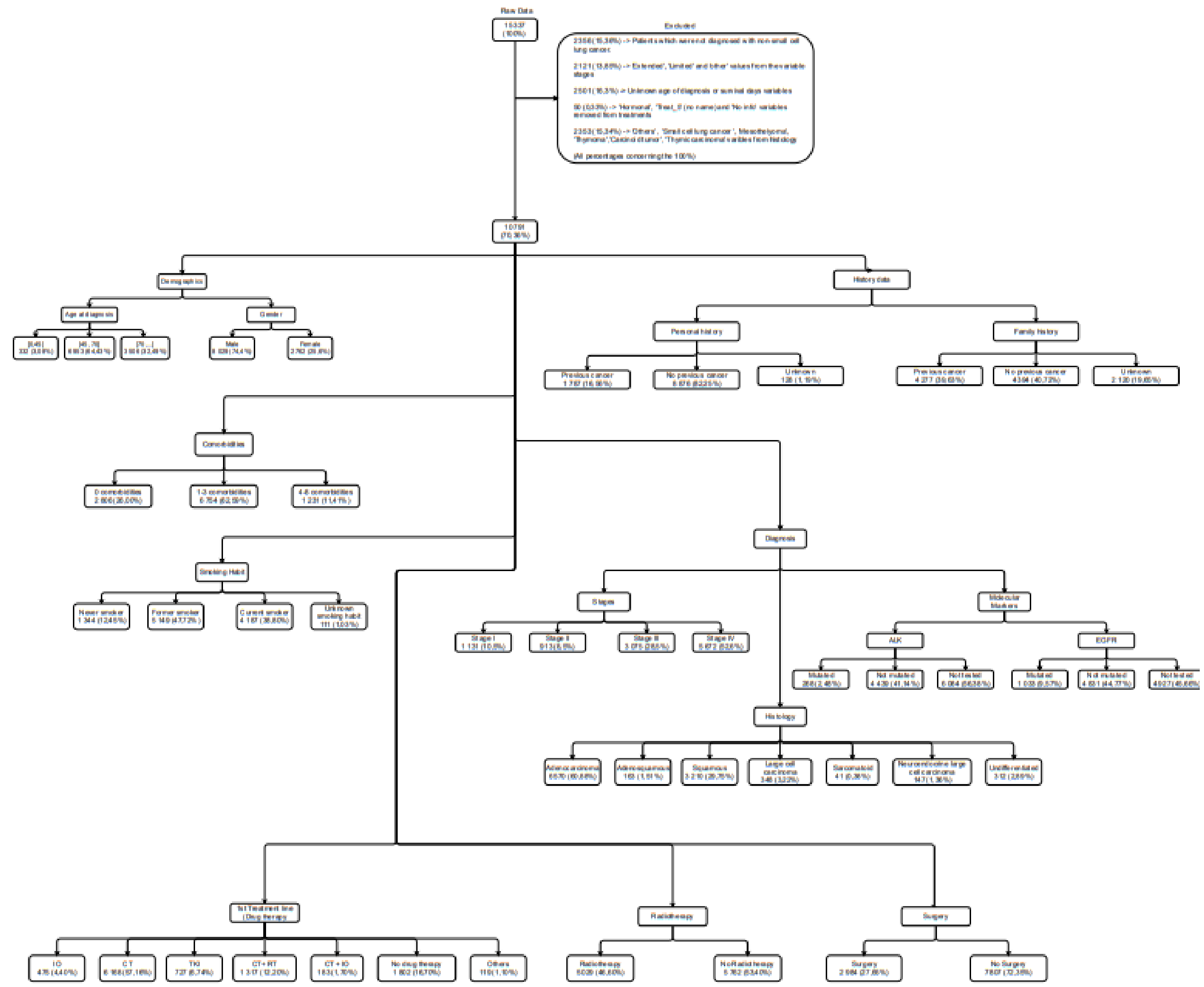
Section A- Attachment 3 - Treatment line 1 - Original values.

	Variable	Values
Treatment Line 1	Type of therapy	1, CT intravenous 2, Oral targeted therapy 3, Neoadjuvant chemotherapy 4, Adjuvant chemotherapy 6, Concomitant CT-RT 7, Sequential CT-RT 8, Adjuvant CT-RT 9, Neoadjuvant CT-RT 10, Immunotherapy 12, Hormonal 13, Oral and intravenous chemotherapy 14, Oral chemotherapy 15, Intravenous chemotherapy + immunotherapy 11, Others -1, -

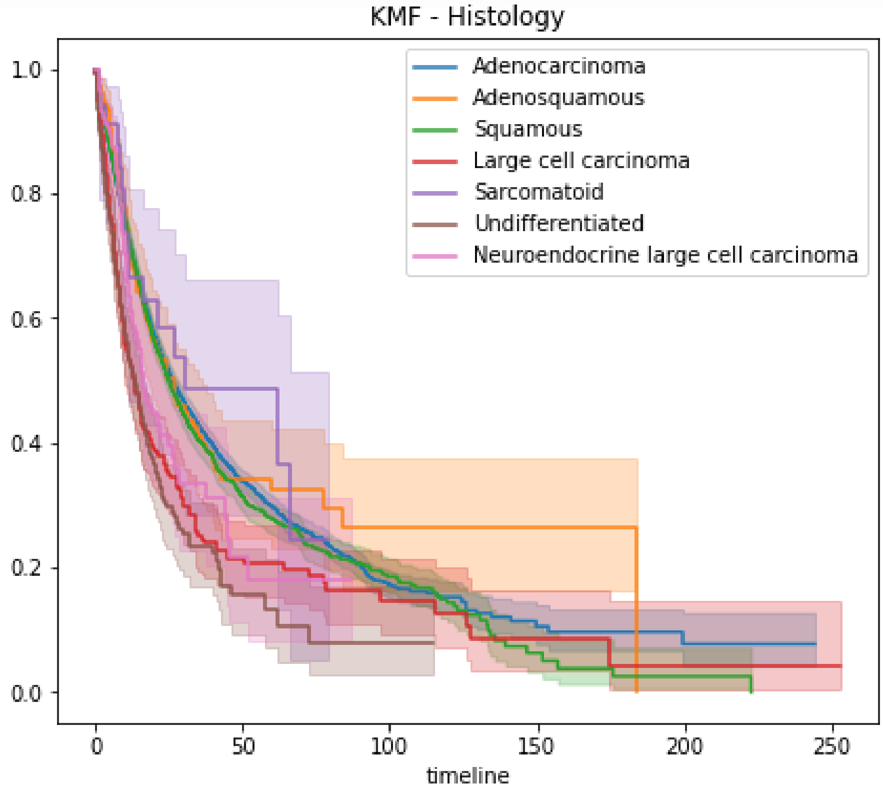
SECTION B

This attachment section contains the complete diagram with all variables and a few tables concerning the descriptive analysis of the variables.

Section B- Attachment 1 – Complete diagram -All variables and their distribution



Section B- Attachment 2 – KMF Histology.

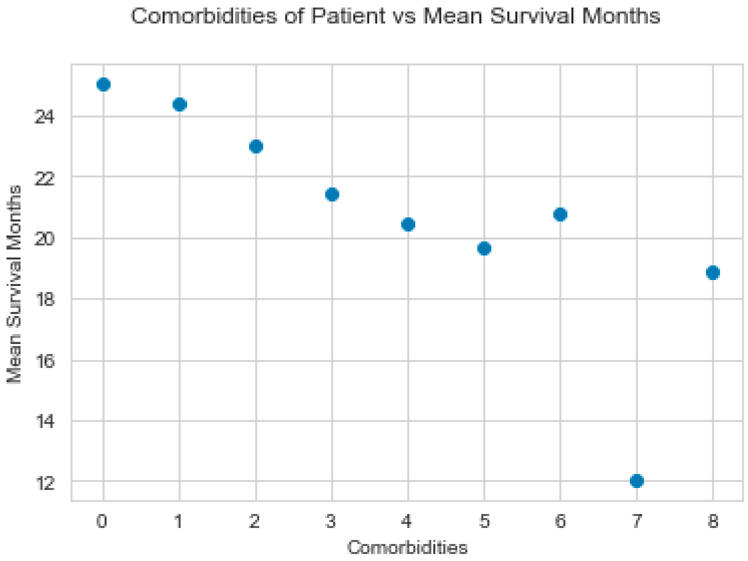


```
                     Adenocarcinoma
                   At risk   5277       610       87       12        4        0
                  Censored      0      2054     2401     2459     2465     2469
                    Events      0      2613     2789     2806     2808     2808

                     Adenosquamous
                   At risk    141        23        8        3        0        0
                  Censored      0        45       57       62       64       64
                    Events      0        73       76       76       77       77

                          Squamous
                   At risk   2498       289       56        5        2        0
                  Censored      0       931     1086     1117     1117     1118
                    Events      0      1278     1356     1376     1379     1380

              Large cell carcinoma
                   At risk    272        29        8        3        1        1
                  Censored      0        64       79       81       82       82
                    Events      0       179      185      188      189      189

                       Sarcomatoid
                   At risk     35         8        0        0        0        0
                  Censored      0        13       19       19       19       19
                    Events      0        14       16       16       16       16

                  Undifferentiated
                   At risk    235        10        3        0        0        0
                  Censored      0        75       79       82       82       82
                    Events      0       150      153      153      153      153

Neuroendocrine large cell carcinoma
                   At risk    120         6        0        0        0        0
                  Censored      0        53       58       58       58       58
                    Events      0        61       62       62       62       62
```

Comorbidity	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV	Male	Female
No comorbidities	1739	16,2	26,02	58	58,04	105	127	479	1028	1110	629
Asthma	216	13,5	21,48	64	63,29	30	14	55	117	102	114
Vascular disease	2114	13,7	21,61	69	69,26	295	212	614	993	1813	301
Diabetes mellitus	2095	13,7	21,05	69	68,18	221	192	614	1068	6252	2444
Dyslipidemia	3552	14,5	22,47	68	67,35	435	336	964	1817	2738	814
COPD	2412	14,65	22,96	67	67,22	381	262	801	968	2140	272
Alcoholism/Ex-alcoholism	778	9,8	18,33	64	63,83	71	75	235	397	739	39
Renal disease	295	10,8	18,99	72	71,03	39	37	88	131	246	49
Obesity	469	16,6	23,31	66	65,29	63	55	137	214	352	117
Depressive syndrome anxiety	759	14,7	23,10	63	63,01	76	66	203	414	384	375
Tuberculosis	190	13,35	21,82	65,5	65,18	24	20	48	98	148	42
Liver disease	222	13,35	19,29	62	60,97	25	14	63	120	186	36
Gastrointestinal	62	11,95	19,12	65	65,89	4	6	14	38	50	12
Neurodegenerative disorder	8	5,85	8,36	81,5	74,13	0	1	2	5	7	1
Benign prostatic hyperplasia	104	7,9	13,45	74	73,83	7	10	38	49	104	0
Obstructive sleep apnea	40	8,1	16,81	63,5	64,43	2	9	10	19	33	7
HT	4515	14,1	22,25	69	68,28	575	414	1247	2279	3549	966

Section B- Attachment 3 - Comorbidities - Descriptive table.

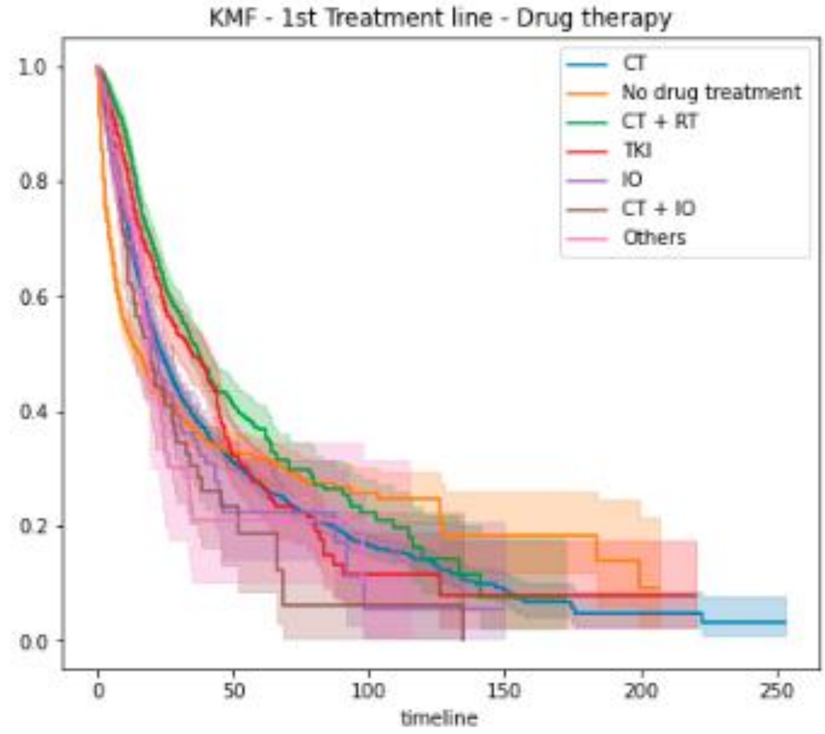
Section B- Attachment 4 - Comorbidities of patients vs Mena Survival Months.



Section B- Attachment 5 - Patient’s previous cancer - Descriptive table.

Previous cancer	count	median survival	mean survival	median age	mean age	Stage I	Stage II	Stage III	Stage IV
No previous tumors	9004	14,45	23,02	64	63,71	805	721	2566	4912
Breast	115	19,6	29,52	64	64,50	9	7	36	63
Head and neck	235	16,1	21,80	66	65,56	59	26	77	73
Germinal tumors	10	15,95	33,73	68	65,90	1	2	2	5
Sarcoma	10	27,2	40,88	63	64,00	2	0	1	7
Central nervous system	8	9,6	23,09	66	68,50	1	1	2	4
Unknown origin carcinoma	2	37,9	37,90	66,5	66,50	0	0	1	1
Colorectal	164	19	27,28	69	68,60	35	15	43	71
Esophagogastric	29	12,6	27,95	69	67,07	8	7	6	8
Pancreatic	7	5,5	15,40	57	60,71	3	0	2	2
Gallbladder	9	16,6	16,68	68	65,67	2	0	2	5
Liver	14	12,4	18,69	63	62,93	3	1	4	6
Melanoma	33	14,4	20,42	68	65,39	7	4	8	14
Skin no melanoma	90	13,9	19,65	70,5	68,99	7	4	34	45
Bladder/urinary tract	275	16,1	23,76	70	70,04	43	36	89	107
Renal	37	15,6	30,10	67	68,14	7	5	10	15
Prostate	292	14,95	24,00	71	71,25	44	32	71	145
Uterus/Cervical	40	27,55	32,00	63,5	64,10	7	8	9	16
Lymphoma	62	17,7	22,00	65	64,19	13	12	14	23
Leukemia	17	9,2	14,48	68	66,24	1	0	7	9
Lung	62	20,1	32,30	67	66,58	24	7	17	14
Ovarian	5	6,4	25,70	66	65,00	1	0	1	3
Others	271	15,7	26,38	68	67,37	49	25	73	124

Section B- Attachment 6 - KMF – First treatment line (Drug therapy).



CT						
At risk	5067	631	107	14	3	1
Censored	0	1711	2048	2114	2120	2121
Events	0	2725	2912	2939	2944	2945
No drug treatment						
At risk	1263	113	27	5	2	0
Censored	0	472	543	561	562	564
Events	0	678	693	697	699	699
CT + RT						
At risk	1017	147	18	2	0	0
Censored	0	446	537	546	548	548
Events	0	424	462	469	469	469
TKI						
At risk	607	60	7	2	2	0
Censored	0	279	310	314	314	316
Events	0	268	290	291	291	291
IO						
At risk	384	12	1	0	0	0
Censored	0	215	223	224	224	224
Events	0	157	160	160	160	160
CT + IO						
At risk	140	7	1	0	0	0
Censored	0	65	68	68	68	68
Events	0	68	71	72	72	72
Others						
At risk	100	5	1	0	0	0
Censored	0	47	50	51	51	51
Events	0	48	49	49	49	49

SECTION C

This attachment section contains pieces of the code in order to clarify some functions used during development.

Section C- Attachment 1 - Descriptive table examples.

```
#######################
#!/usr/bin/env python
#-*- coding: utf-8 -*-
#@author: Filipa Matos
#######################

import pandas as pd
import matplotlib.pyplot as plt

#Gender table - Descriptive Analysis

gender_info=[
   ['Male' , analysis_1.survival_days[analysis_1['gender']==0].count(),
    analysis_1.survival_days[analysis_1['gender']==0].median(),
    analysis_1.survival_days[analysis_1['gender']==0].mean(),
    analysis_1.Age_diagnosis[analysis_1['gender']==0].median(),
    analysis_1.Age_diagnosis[analysis_1['gender']==0].mean(),
    analysis_1.survival_days[(analysis_1['gender']==0) &
               (analysis_1['stage_at_diagnosis_groups']== 1.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==0) &
               (analysis_1['stage_at_diagnosis_groups']== 2.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==0) &
               (analysis_1['stage_at_diagnosis_groups']== 3.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==0) &
               (analysis_1['stage_at_diagnosis_groups']== 0.0)].count(),
    (analysis_1.survival_days[(analysis_1['gender']==0) & ((analysis_1[
'tabac_info']== 1)|(analysis_1['tabac_info']== 2))].count()),
    analysis_1.survival_days[(analysis_1['gender']==0) & (analysis_1['t
abac_info']== 0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==0) & (analysis_1['t
abac_info']== 3)].count()],

    ['Female' , analysis_1.survival_days[analysis_1['gender']==1].count
(),
    analysis_1.survival_days[analysis_1['gender']==1].median(),
    analysis_1.survival_days[analysis_1['gender']==1].mean(),
    analysis_1.Age_diagnosis[analysis_1['gender']==1].median(),
    analysis_1.Age_diagnosis[analysis_1['gender']==1].mean(),
    analysis_1.survival_days[(analysis_1['gender']==1) &
               (analysis_1['stage_at_diagnosis_groups']== 1.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==1) &
               (analysis_1['stage_at_diagnosis_groups']== 2.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==1) &
               (analysis_1['stage_at_diagnosis_groups']== 3.0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==1) &
               (analysis_1['stage_at_diagnosis_groups']== 0.0)].count(),
    (analysis_1.survival_days[(analysis_1['gender']==1) &
        ((analysis_1['tabac_info']== 1)|(analysis_1['tabac_info']== 2))]
.count()),
    analysis_1.survival_days[(analysis_1['gender']==1) &
         (analysis_1['tabac_info']== 0)].count(),
    analysis_1.survival_days[(analysis_1['gender']==1) &
        (analysis_1['tabac_info']== 3)].count()]]


table_gender_info = pd.DataFrame(gender_info, columns = ['Gender',
'count', 'median survival','mean survival','median age','mean age','sta
geI','stageII','stageIII','stageIV','Former Or Current','No smoker',
'Unknonw smoking habit'])
table_gender_info.round(2)

#EXPORT TO EXCEL
with pd.ExcelWriter(r'C:\ Desktop\project\result_tables_excel\Descri
ptive analysis\gender.xlsx')
        as writer:table_age_info.to_excel(writer, sheet_name = 'gender
')
```

Section C- Attachment 2 – Kaplan Meier examples.

```
#######################
#!/usr/bin/env python
#-*- coding: utf-8 -*-
#@author: Filipa Matos
#######################

import pandas as pd
import matplotlib.pyplot as plt
from lifelines import KaplanMeierFitter
from lifelines.statistics import logrank_test
from lifelines.plotting import add_at_risk_counts


#KM Gender curve

male=df_original[df_original['gender']==0]
female=df_original[df_original['gender']==1]
T=male['survival_days']
E=male['dead_alive']
T1=female['survival_days']
E1=female['dead_alive']

kmf_male = KaplanMeierFitter()

ax = plt.subplot(111)
ax = kmf_male.fit(T, E, label="Male").plot(ax=ax)

kmf_female = KaplanMeierFitter()
ax = kmf_female.fit(T1, E1, label="Female",).plot(ax=ax)
```

```
#logrank_test
results=logrank_test(T,T1,event_observed_A=E, event_observed_B=E1)
results.print_summary()

add_at_risk_counts(kmf_male, kmf_female, ax=ax)
plt.title('KMF - Gender')
plt.tight_layout()

#KM Comorbidities curve

no_comorb=df_original[df_original['number_comorb']==0]
one_3_comorb=df_original[df_original['number_comorb']==1]
more_3_comorb=df_original[df_original['number_comorb']==3]

T=no_comorb['survival_days']
E=no_comorb['dead_alive']

T1=one_3_comorb['survival_days']
E1=one_3_comorb['dead_alive']

T2=more_3_comorb['survival_days']
E2=more_3_comorb['dead_alive']

kmf_zero = KaplanMeierFitter()

ax = plt.subplot(111)
ax = kmf_zero.fit(T, E, label="No Comorbidities").plot(ax=ax,figsize=(7
, 6))

kmf_one_3 = KaplanMeierFitter()
ax = kmf_one_3.fit(T1, E1, label="1-3 Comorbidities",).plot(ax=ax,figsi
ze=(7, 6))

kmf_more_3 = KaplanMeierFitter()
ax = kmf_more_3.fit(T2, E2, label=">3 comorbidities",).plot(ax=ax,figsi
ze=(7, 6))

#logrank_test
results=logrank_test(T,T2,event_observed_A=E, event_observed_B=E2)
results.print_summary()

add_at_risk_counts(kmf_zero, kmf_one_3,kmf_more_3, ax=ax)

L=ax.legend()
L=ax.legend(bbox_to_anchor=(1,1))
#plt.xlabel('Survival time (Months)',axes=ax)
plt.ylabel('Survival (%)',axes=ax)
plt.title('KMF - Comorbidities')
plt.tight_layout()
```

Section C- Attachment 3 - Categorical variable into bin variables funtion.

```
from patsy import dmatrices

#Define varibles to be included in the expression
model_expr = 'survival_days ~ gender + tabac_info + number_comorb +
group_age+ stage_at_diagnosis_groups + patient_previous_cancer +
family_cancer+ histology +treat_line_type_therapy+ radio_therapy +
surgery + mutated_alk + mutated_egfr +survival_days + dead_alive'

#Use the model expression to break out the CELL_TYPE categorical variab
le into 1-0 type columns
y, X = dmatrices(model_expr, df_original ,eval_env=1, return_type='data
frame')

#Print out the first few rows
X.head()
```

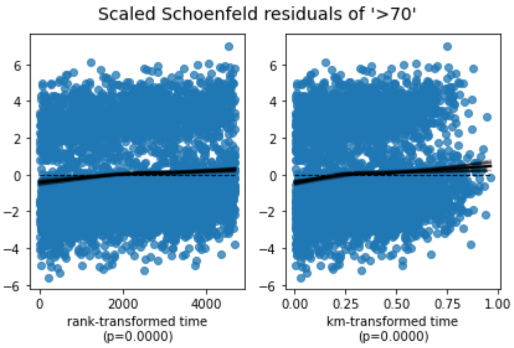
Section C- Attachment 4 -Code Cox Model example.

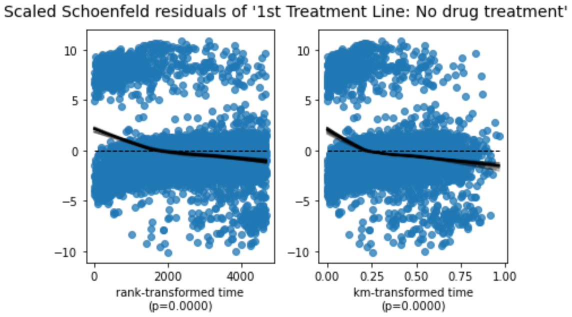
```
import pandas as pd
import matplotlib.pyplot as plt
from lifelines import CoxPHFitter

#Run the Cox model
cph = CoxPHFitter()

# X is the input data
cph.fit(X, duration_col='survival_days', event_col='dead_alive')

#Printing the summary table
cph.print_summary()

#Plot results
plt.figure(figsize=(10, 10))
cph.plot()

#Survival curves from a variable (Example: Stage variable)

cph.plot_partial_effects_on_outcome(covariates=['stage_at_diagnosis_gro
ups[T.1]','stage_at_diagnosis_groups[T.2]','stage_at_diagnosis_groups[T
.3]',], values=[[1,0,0],[0,1,0],[0,0,1],[0,0,0]])
plt.xlim(right=250, left=0)
plt.ylim(-0.04,1)
plt.xlabel('Survival time (Months)')
plt.ylabel('Survival (%)')
plt.title('Stage at diagnosis')

L=plt.legend()
L=plt.legend(bbox_to_anchor=(1.3,1.0))
L.get_texts()[0].set_text('Stage I')
L.get_texts()[1].set_text('Stage II')
```

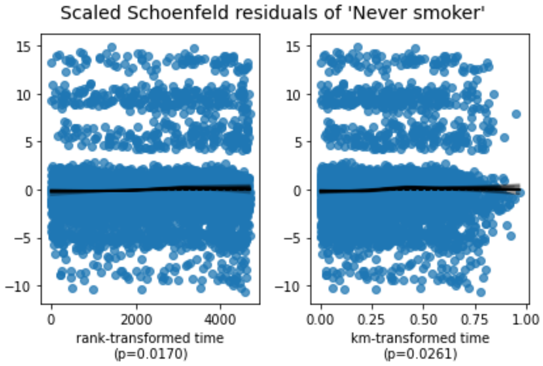
```
L.get_texts()[2].set_text('Stage III')
L.get_texts()[3].set_text('Stage IV')
```

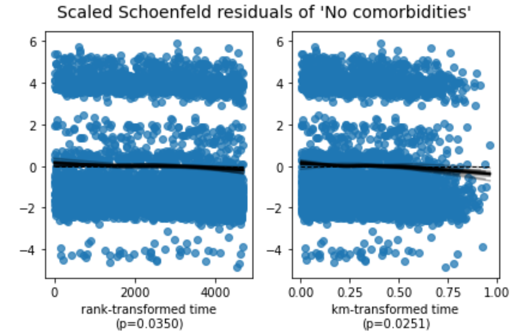
Section D

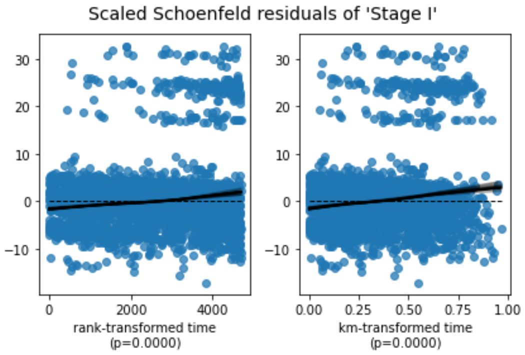
Section D - Attachment 1 – All stages model

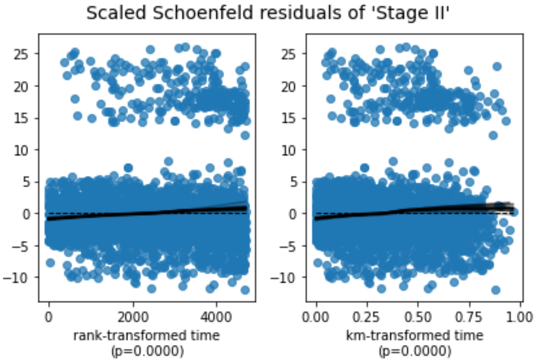


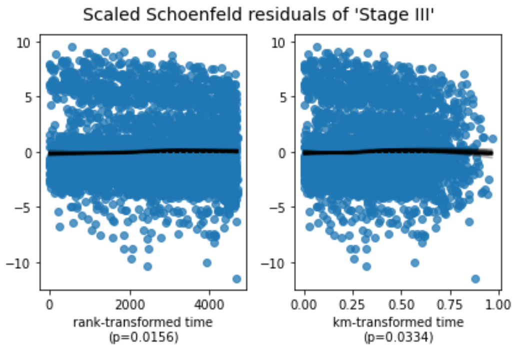


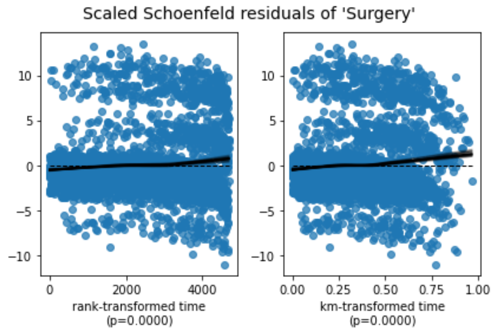


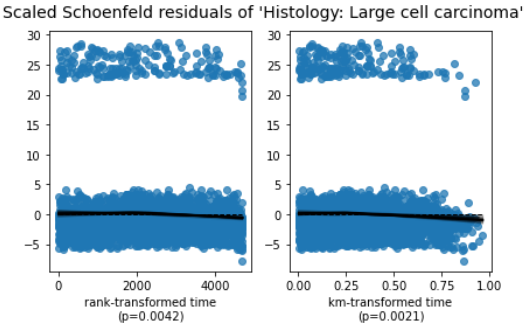


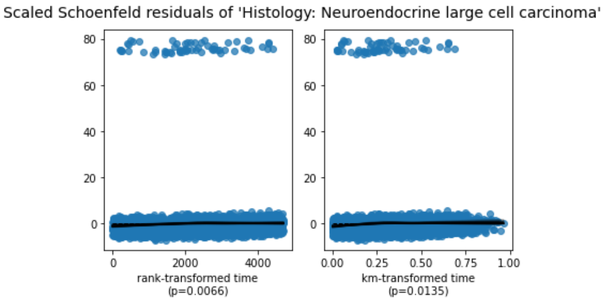


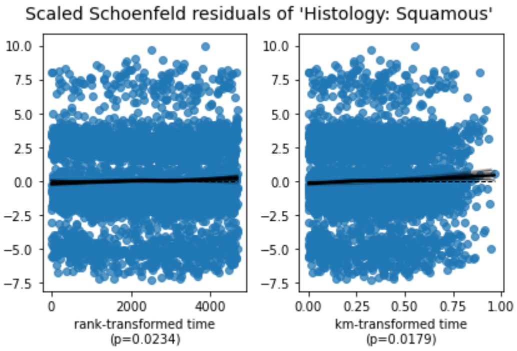


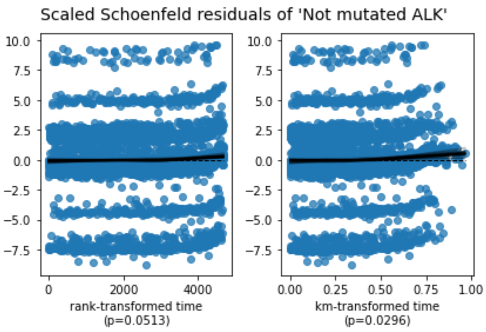




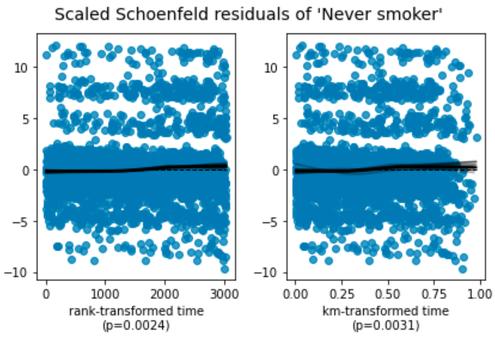


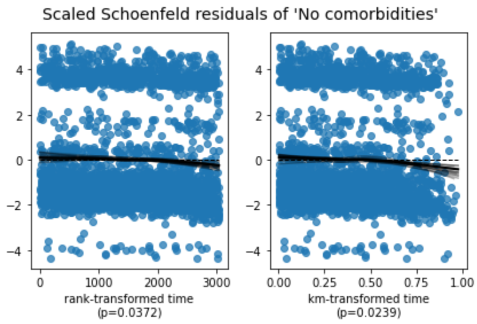


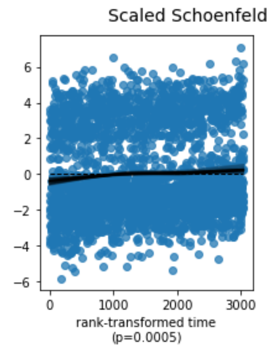


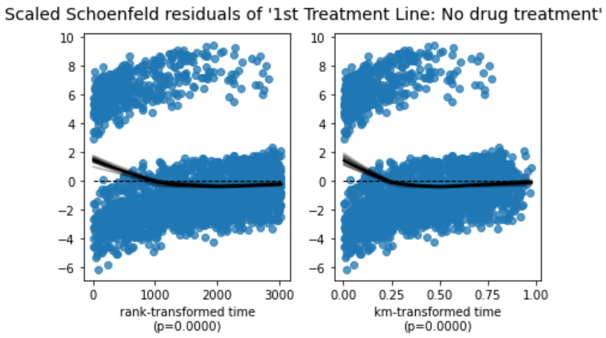


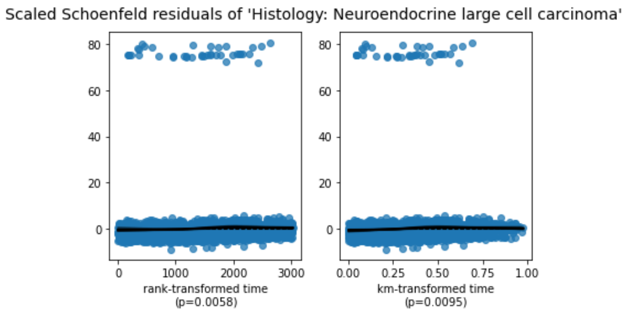
Section D - Attachment 2 - Stage IV Model

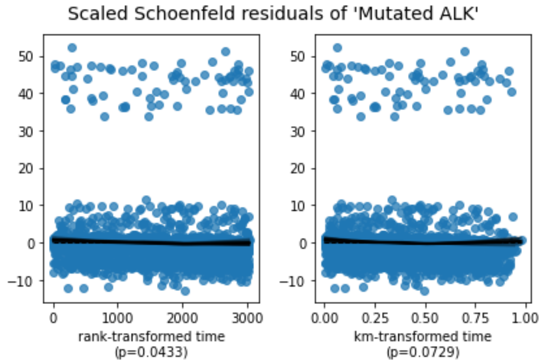


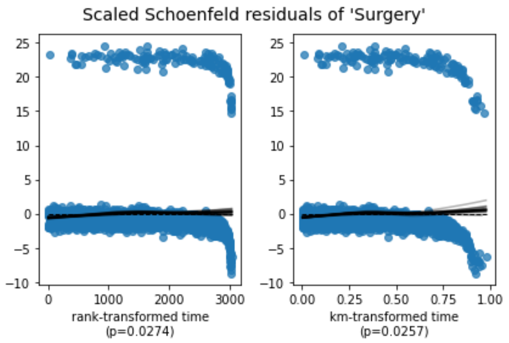



Scaled Schoenfeld residuals of '>70'
6
4
2
0
−2
−4
−6
0
1000
2000
3000
rank-transformed time
(p=0.0005)
0.00
0.25
0.50
0.75
1.00
km-transformed time
(p=0.0005)


Scaled Schoenfeld residuals of '1st Treatment Line: No drug treatment'
10
8
6
4
2
0
−2
−4
−6
0
1000
2000
3000
rank-transformed time
(p=0.0000)
0.00
0.25
0.50
0.75
1.00
km-transformed time
(p=0.0000)


Scaled Schoenfeld residuals of 'Histology: Neuroendocrine large cell carcinoma'
80
60
40
20
0
0
1000
2000
3000
rank-transformed time
(p=0.0058)
0.00
0.25
0.50
0.75
1.00
km-transformed time
(p=0.0095)


Scaled Schoenfeld residuals of 'Mutated ALK'
50
40
30
20
10
0
−10
0
1000
2000
3000
rank-transformed time
(p=0.0433)
0.00
0.25
0.50
0.75
1.00
km-transformed time
(p=0.0729)


Scaled Schoenfeld residuals of 'Surgery'
25
20
15
10
5
0
−5
−10
0
1000
2000
3000
rank-transformed time
(p=0.0274)
0.00
0.25
0.50
0.75
1.00
km-transformed time
(p=0.0257)